Fanshawe College Astronomy

FANSHAWE COLLEGE ASTRONOMY

DR. IFTEKHAR HAQUE

Fanshawe College Pressbooks
London Ontario

CONTENTS

Chapter 1: Astronomy and the Nature of Science

Chapter 2: Numbers and Mathematics in Astronomy

Chapter 3: An Overview of the Universe

Chapter 4: Imaging the Universe

Chapter 5: An Overview of the Solar System

Chapter 6: The Terrestrial Planets

Chapter 7: The Gas and Ice Giants

Chapter 8: Other Objects in the Solar System

Chapter 9: Stars: An Introductory Survey

Chapter 10: Stars: All About Sol

Chapter 11: Stars: Classification, Evolution, and Death

Chapter 12: Galaxies: The Milky Way and Its Neighbourhood

Chapter 13: Galaxies: Collisions, Mergers, and Evolution

Chapter 14: The Structure of the Universe

Chapter 15: The Origin and Evolution of the Universe

Chapter 16: Life in the Universe

ACKNOWLEDGEMENTS

This open textbook has been complied and adapted by Iftekhar Haque, in partnership with the OER Design Studio and the Library Learning Commons at Fanshawe College in London, Ontario This work is part of the FanshaweOpen learning initiative and is made available through a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License unless otherwise noted.

Attribution

We would like to acknowledge and thank the following authors/entities who have graciously made their work available for the remixing, reusing, and adapting of this text:

- Adpative Optics by Michael Richmond is licensed under a Creative Commons Attribution Non-Commercial ShareAlike 4.0 International License, except where otherwise noted.
- Astronomy by Andrew Fraknoi, David Morrison, Sidney C. Wolff, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.
- Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.
- Contemporary Mathematics by Donna Kirk, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.
- Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.
- Introduction to Astronomy by Lumen Learning is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.
- Introductory Chemistry – 1st Canadian Edition by David W. Ball and Jessie A. Key is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0, except where otherwise noted.

- Physical Geography and Natural Disasters by R. Adam Dastrup, MA, GISP is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.
- Physics by OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.
- Prealgebra 2e by Lynn Marecek, MaryAnne Anthony-Smith, and Andrea Honeycutt Mathis, © OpenStax are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

This book is a compilation of the resources listed above. Overall formatting changes to ensure ease of use and consistency were made. For more information see the list of changes made.

Cover Photo by the Space Telescope Science Institute and NASA is in the Public Domain.

Collaborators

- Jason Benoit, *Instructional Design Student*
- Robert Armstrong, *Quality Assurance*
- Freddy Vale, *Graphic Designer*
- Wilson Poulter, *Copyright Officer*
- Shauna Roch, *Project Lead*

ABOUT THIS BOOK

The content of this book is organized to align with and supplement the syllabus of the Introduction to Astronomy course at Fanshawe College. The course presents an exclusively scientific survey of modern astronomy, from cosmology and galaxies, to stars, planets, and atoms. It also highlights the impact of astronomical developments on history and culture and encourages an appreciation for astronomy in the context of overall human existence.

The outline and the in-depth syllabus for the course, in its present form, have been developed over several semesters of first-hand delivery by Dr. Iftekhar Haque, professor of science and mathematics at Fanshawe College. Dr. Haque would like to acknowledge the valuable roadmap through the vast field of astronomical knowledge provided by the original outline for the course, developed by Peter Jedicke, former Fanshawe College faculty.

Accessibility Statement

We are actively committed to increasing the accessibility and usability of the textbooks we produce. Every attempt has been made to make this OER accessible to all learners and is compatible with assistive and adaptive technologies. We have attempted to provide closed captions, alternative text, or multiple formats for on-screen and off-line access.

The web version of this resource has been designed to meet Web Content Accessibility Guidelines 2.0, level AA. In addition, it follows all guidelines in Appendix A: Checklist for Accessibility of the *Accessibility Toolkit – 2nd Edition*.

In addition to the web version, additional files are available in a number of file formats including PDF, EPUB (for eReaders), and MOBI (for Kindles).

If you are having problems accessing this resource, please contact us at oer@fanshawec.ca.

Please include the following information:

- The location of the problem by providing a web address or page description
- A description of the problem
- The computer, software, browser, and any assistive technology you are using that can

help us diagnose and solve your issue (e.g., Windows 10, Google Chrome (Version 65.0.3325.181), NVDA screen reader)

Your feedback is welcomed. Please share your adoption, and any feedback you have about the book with us at oer@fanshawec.ca

CHANGES FROM ADAPTED SOURCE

This book is a compilation of a number of open resources. Specific changes to each chapter can be found in the table below. As well some overall changes were made to this version.

Overall Changes

- Some images replaced with high-resolution versions.
- Some images replaced where links to originals could not be located.
- Attributions with links added for all images.
- Some image captions modified.
- Added new learning outcomes to each chapter
- Added key terms to each chapter.
- Removed Youtube videos embedded or linked.
- Some sections were removed or reformatted to make all chapters consistent.
- Removed key concepts and summaries.
- Removed references to other chapters.

Chapter 1	• 1.1 from Astronomy • 1.2 & 1.3 from Douglas College Astronomy 1105 • 1.4 – 1.6 from Physical Geography and Natural Disasters ◦ Omitted the sentence: "Among these, denial of climate change is strongly connected with geography." ◦ Replaced the sentence "Geologists, scientists, or anyone exploring scientific inquiry must discern valid sources of information from pseudoscience and misinformation." with "Anyone exploring scientific inquiry must discern valid sources of information from pseudoscience and misinformation."
Chapter 2	• 2.1, 2.3, 2.4, 2.6, 2.7 from Douglas College Astronomy 1105 ◦ A few sentences changed, and only some of the original source used. • 2.2 from Introductory Chemistry – 1st Canadian Edition • 2.5 from Physics • 2.8 from Contemporary Mathematics • 2.9 from Prealgebra 2e
Chapter 3	• All sections from Douglas College Astronomy 1105. ◦ Omitted Canadians in Space boxed section
Chapter 4	• All sections from Douglas College Astronomy 1105. ◦ Not all content from original source was included. Some sections were omitted. • 4.8 includes content from Adpative Optics
Chapter 5	• 5.1, 5.2 from Douglas College Astronomy 1105. ◦ Omitted Canadians in Space boxed section. • 5.3 from Physical Geography and Natural Disasters
Chapter 6	• 6.1 – 6.3 from Douglas College Astronomy 1105. • 6.4 from Astronomy 2e and Introduction to Astronomy • 6.5 & 6.6 from Astronomy 2e ◦ Omitted subsection The Origin of Mercury
Chapter 7	• All sections from Astronomy 2e ◦ Omitted subsection Winds and Weather

Chapter 8	• 8.1 – 8.4 & 8.8 from Astronomy 2e ◦ Not all sections included. For example, omitted subsection What Causes Rings? • 8.5 from Introduction to Astronomy ◦ Included subsections The Nature of Pluto, Geology of Pluto and A Quick Look at Charon. • 8.6 – 8.7 from Physical Geography and Natural Disasters
Chapter 9	• 9.1 from Physical Geography and Natural Disasters • 9.2 – 9.6 from Douglas College Astronomy 1105 ◦ Some content re-organized. Some sections omitted.
Chapter 10	• All sections from from Douglas College Astronomy 1105 ◦ Omitted rows about Mean Angular Diameter and Inclination of equator to ecliptic from Figure. As well only some sections of content included.
Chapter 11	• 11.1 – 11.6 and 11.8 – 11.13 from from Douglas College Astronomy 1105 • 11.7 from Introduction to Astronomy ◦ some sections omitted. 11.11 – The Synthesis of Heavy Elements, Neutrinos from SN 1987A ◦ 11.12 removed references to spinning ice skater and omitted subsection Tests of the Model
Chapter 12	• All sections from from Douglas College Astronomy 1105 ◦ Omitted The Milky Way Galaxy in Myth and Legend boxed item, and the Canadians in Space boxed item
Chapter 13	• All sections from from Douglas College Astronomy 1105 ◦ Omitted Recession Speed of a Quasar boxed item.
Chapter 14	• All sections from from Douglas College Astronomy 1105 ◦ Omitted boxed item Hubble-Lemaitre Law and distant galaxies example, Mass-to-Light Ratio section, references to Astronomy Basics feature box Gravitational Lensing.
Chapter 15	• All sections from from Douglas College Astronomy 1105 ◦ Omitted boxed items

Chapter 16	- All sections from from Douglas College Astronomy 1105

CHAPTER 1: ASTRONOMY AND THE NATURE OF SCIENCE

Chapter Overview

1.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn to:

- Define astronomy.
- Critically evaluate sources of information related to scientific research, distinguishing between reputable scientific works and pseudoscience, and recognizing the importance of peer review in scientific inquiry.
- Describe the scientific approach by which astronomers have come to understand the realms beyond Earth.

1.1 AN INVITATION TO INQUIRY

We invite you to come along on a series of voyages to explore the universe as astronomers understand it today. Beyond Earth are vast and magnificent realms full of objects that have no counterpart on our home planet. Nevertheless, we hope to show you that the evolution of the universe has been directly responsible for your presence on Earth today.

Distant Galaxies

Figure 1.1. These two interacting islands of stars (galaxies) are so far away that their light takes hundreds of millions of years to reach us on Earth (photographed with the Hubble Space Telescope). Interacting Galaxies Arp 273 by NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and K. Noll (STScI), NASA Media Licence.

Along your journey, you will encounter:

- a canyon system so large that, on Earth, it would stretch from Los Angeles to Washington, DC, USA.
- a crater and other evidence on Earth that tell us that the dinosaurs (and many other creatures) died because of a cosmic collision.

- a tiny moon whose gravity is so weak that one good throw from its surface could put a baseball into orbit.
- a collapsed star so dense that to duplicate its interior we would have to squeeze every human being on Earth into a single raindrop.
- exploding stars whose violent end could wipe clean all of the life-forms on a planet orbiting a neighbouring star.
- a “cannibal galaxy” that has already consumed a number of its smaller galaxy neighbours and is not yet finished finding new victims.
- a radio echo that is the faint but unmistakable signal of the creation event for our universe.

Mars Mosaic

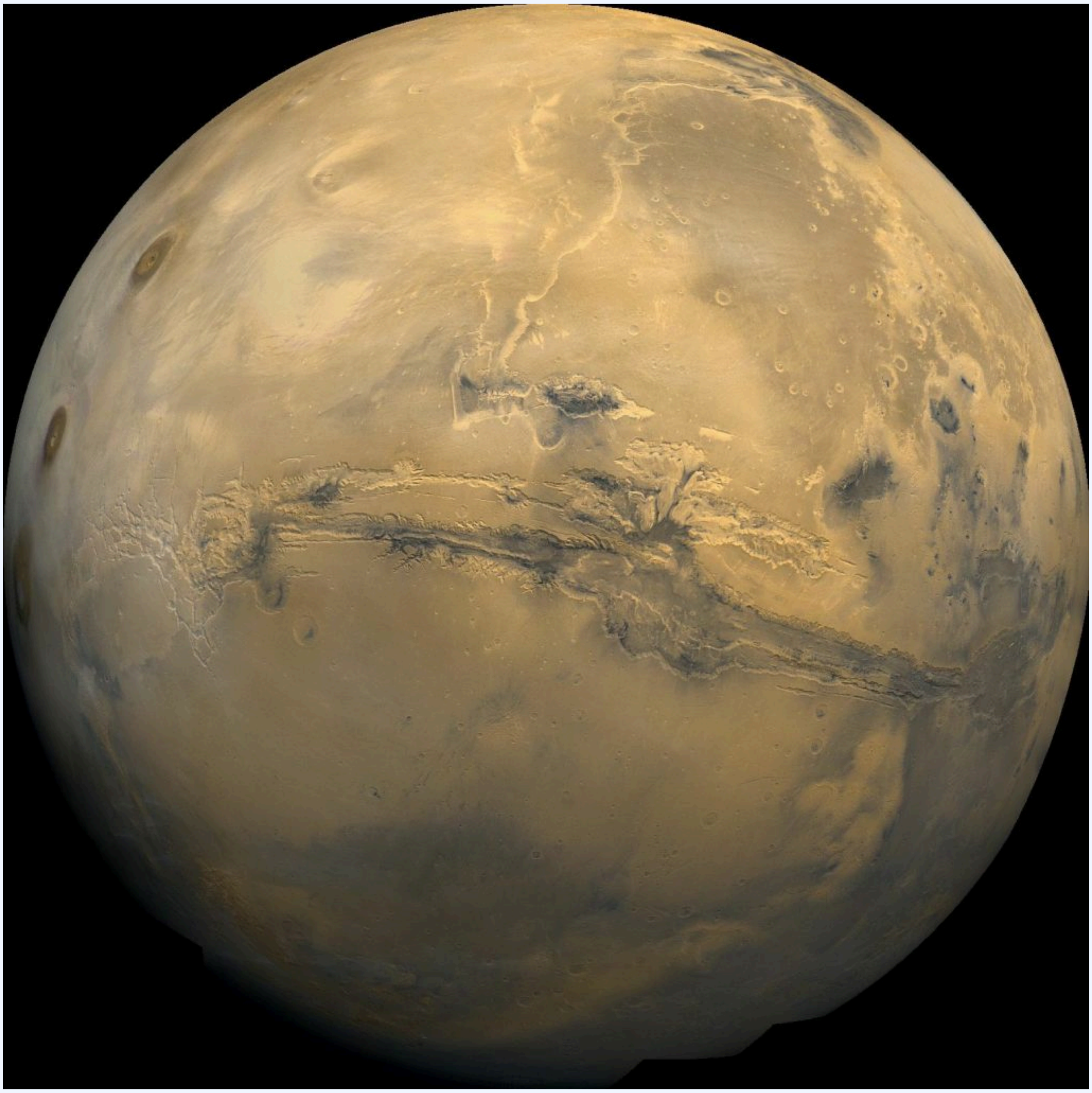


Figure 1.2. This image of Mars is centred on the Valles Marineris (Mariner Valley) complex of canyons, which is as long as the United States is wide.
Valles Marineris: The Grand Canyon of Mars by NASA, NASA Media Licence.

Such discoveries are what make astronomy such an exciting field for scientists and many others—but you will explore much more than just the objects in our universe and the latest discoveries about them. We will pay equal attention to the *process* by which we have come to understand the realms beyond Earth and the tools we use to increase that understanding. We gather information about the cosmos from the messages the universe sends our way. Because the stars are the fundamental building blocks of the universe, decoding the message of starlight has been a central challenge and triumph of modern astronomy. By the time you have finished reading this text, you will know a bit about how to read that message and how to understand what it is telling us.

Stellar Corpse

Figure 1.3. We observe the remains of a star that was seen to explode in our skies in 1054 (and was, briefly, bright enough to be visible during the daytime). Today, the remnant is called the Crab Nebula and its central region is seen here. Such exploding stars are crucial to the development of life in the universe. Messier 1 (The Crab Nebula) by NASA, ESA, J. Hester (Arizona State University), NASA Media Licence.

Attribution

1.2 THE NATURE OF ASTRONOMY

Astronomy is defined as the study of the objects that lie beyond our planet Earth and the processes by which these objects interact with one another. We will see, though, that it is much more. It is also humanity's attempt to organize what we learn into a clear history of the universe, from the instant of its birth in the Big Bang to the present moment. Throughout this book, we emphasize that science is a **progress report**—one that changes constantly as new techniques and instruments allow us to probe the universe more deeply.

In considering the history of the universe, we will see again and again that the cosmos evolves; it changes in profound ways over long periods of time. For example, the universe made the carbon, the calcium, and the oxygen necessary to construct something as interesting and complicated as you. Today, many billions of years later, the universe has evolved into a more hospitable place for life. Tracing the evolutionary processes that continue to shape the universe is one of the most important (and satisfying) parts of modern astronomy.

Attribution

"1.1 The Nature of Astronomy" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

1.3 THE NATURE OF SCIENCE

The ultimate judge in science is always what nature itself reveals based on observations, experiments, models, and testing. Science is not merely a body of knowledge, but a **method** by which we attempt to understand nature and how it behaves. This method begins with many observations over a period of time. From the trends found through observations, scientists can **model** the particular phenomena we want to understand. Such models are always approximations of nature, subject to further testing.

As a concrete astronomical example, ancient astronomers constructed a model (partly from observations and partly from philosophical beliefs) that Earth was the center of the universe and everything moved around it in circular orbits. At first, our available observations of the Sun, Moon, and planets did fit this model; however, after further observations, the model had to be updated by adding circle after circle to represent the movements of the planets around Earth at the center. As the centuries passed and improved instruments were developed for keeping track of objects in the sky, the old model (even with a huge number of circles) could no longer explain all the observed facts. As we will see in a later chapter, a new model, with the Sun at the center, fit the experimental evidence better. After a period of philosophical struggle, it became accepted as our view of the universe.

When they are first proposed, new models or ideas are sometimes called **hypotheses**. You may think there can be no new hypotheses in a science such as astronomy—that everything important has already been learned. Nothing could be further from the truth. Throughout this textbook you will find discussions of recent, and occasionally still controversial, hypotheses in astronomy. For example, the significance that the huge chunks of rock and ice that hit Earth have for life on Earth itself is still debated. And while the evidence is strong that vast quantities of invisible "dark energy" make up the bulk of the universe, scientists have no convincing explanation for what the dark energy actually is. Resolving these issues will require difficult observations done at the forefront of our technology, and all such hypotheses need further testing before we incorporate them fully into our standard astronomical models.

This last point is crucial: a hypothesis must be a proposed explanation that can be *tested*. The most straightforward approach to such testing in science is to perform an experiment. If the experiment is conducted properly, its results either will agree with the predictions of the hypothesis or they will contradict it. If the experimental result is truly inconsistent with the hypothesis, a scientist must discard the hypothesis and try to develop an alternative. If the experimental result agrees with predictions, this does not necessarily prove that the hypothesis is absolutely correct; perhaps later experiments will contradict crucial parts of the hypothesis. But, the more experiments that agree with the hypothesis, the more likely we are to accept the hypothesis as a useful description of nature.

One way to think about this is to consider a scientist who was born and lives on an island where only black sheep live. Day after day the scientist encounters black sheep only, so he or she hypothesizes that all sheep are black. Although every observed sheep adds confidence to the theory, the scientist only has to visit the mainland and observe one white sheep to prove the hypothesis wrong.

When you read about experiments, you probably have a mental picture of a scientist in a laboratory conducting tests or taking careful measurements. This is certainly the case for a biologist or a chemist, but what can astronomers do when our laboratory is the universe? It's impossible to put a group of stars into a test tube or to order another comet from a scientific supply company.

As a result, astronomy is sometimes called an **observational science**; we often make our tests by observing many samples of the kind of object we want to study and noting carefully how different samples vary. New instruments and technology can let us look at astronomical objects from new perspectives and in greater detail. Our hypotheses are then judged in the light of this new information, and they pass or fail in the same way we would evaluate the result of a laboratory experiment.

Much of astronomy is also a **historical science**—meaning that what we observe has already happened in the universe and we can do nothing to change it. In the same way, a geologist cannot alter what has happened to our planet, and a paleontologist cannot bring an ancient animal back to life. While this can make astronomy challenging, it also gives us fascinating opportunities to discover the secrets of our cosmic past.

You might compare an astronomer to a detective trying to solve a crime that occurred before the detective arrived at the scene. There is lots of evidence, but both the detective and the scientist must sift through and organize the evidence to test various hypotheses about what actually happened. And there is another way in which the scientist is like a detective: they both must prove their case. The detective must convince the district attorney, the judge, and perhaps ultimately the jury that his hypothesis is correct. Similarly, the scientist must convince colleagues, editors of journals, and ultimately a broad cross-section of other scientists that her hypothesis is provisionally correct. In both cases, one can only ask for evidence "beyond a reasonable doubt." And sometimes new evidence will force both the detective and the scientist to revise their last hypothesis.

This self-correcting aspect of science sets it off from most human activities. Scientists spend a great deal of time questioning and challenging one another, which is why applications for project funding—as well as reports for publication in academic journals—go through an extensive process of **peer review**, which is a careful examination by other scientists in the same field. In science (after formal education and training), everyone is encouraged to improve upon experiments and to challenge any and all hypotheses. New scientists know that one of the best ways to advance their careers is to find a weakness in our current understanding of something and to correct it with a new or modified hypothesis.

This is one of the reasons science has made such dramatic progress. An undergraduate science major today knows more about science and math than did Sir Isaac Newton, one of the most renowned scientists who ever lived. Even in this introductory astronomy course, you will learn about objects and processes that no one a few generations ago even dreamed existed.

Attribution

1.4 UNDERSTANDING SCIENCE

Scientists seek to understand the fundamental principles that explain natural patterns and processes. Science is more than just a body of knowledge; *science provides a means to evaluate and create new knowledge without bias.* Scientists use objective evidence over subjective evidence to reach sound and logical conclusions.

Objective observation is without personal bias and the same by all individuals. Humans are biased by nature, so they cannot be completely objective; the goal is to be unbiased. A **subjective observation** is based on a person's feelings and beliefs and is unique to that individual.

Another way scientists avoid bias is by using quantitative over qualitative measurements whenever possible. **Quantitative measurement** is expressed with a specific numerical value. **Qualitative observations** are general or relative descriptions. For example, describing a rock as red or heavy is a qualitative observation. Determining a rock's colour by measuring wavelengths of reflected light or its density by measuring the proportions of minerals it contains is quantitative. Numerical values are more precise than general descriptions, and they can be analyzed using statistical calculations. This is why quantitative measurements are much more useful to scientists than qualitative observations.

It is challenging to establish truth in science because all scientific claims are falsifiable, which means any initial hypothesis may be tested and proven false. Only after exhaustively eliminating false results, competing ideas, and possible variations does a hypothesis become regarded as a reliable scientific theory. This meticulous scrutiny reveals weaknesses or flaws in a hypothesis and is the strength that supports all scientific ideas and procedures. Proving current ideas are wrong has been the driving force behind many scientific careers.

Falsifiability separates science from pseudoscience. Scientists are wary of explanations of natural phenomena that discourage or avoid falsifiability. An explanation that cannot be tested or does not meet scientific standards is not considered science, but pseudoscience. **Pseudoscience** is a collection of ideas that may appear scientific but does not use the scientific method. Astrology is an example of pseudoscience. It is a belief system that attributes the movement of celestial bodies to influencing human behavior. Astrologers rely on celestial observations, but their conclusions are not based on experimental evidence, and their statements are not falsifiable. This is not to be confused with astronomy, which is the scientific study of celestial bodies and the cosmos.

Science is also a social process. Scientists share their ideas with peers at conferences, seeking guidance and feedback. Research papers and data submitted for publication are rigorously reviewed by qualified peers, scientists who are experts in the same field. The scientific review process aims to weed out misinformation, invalid research results, and wild speculation. Thus, it is slow, cautious, and conservative. Scientists tend to wait until a hypothesis is supported by an overwhelming amount of evidence from many independent researchers before accepting it as a scientific theory.

Attribution

1.5 GOALS OF SCIENCE

The broad goals of science are to understand natural phenomena and explain how they may be changing over time. To achieve those goals, scientists undertake investigations based on information, inferences, and conclusions developed through a systematic application of logic, usually of the inductive sort. As such, scientists carefully observe natural phenomena and conduct experiments.

A higher goal of scientific research is to formulate laws that describe the workings of the universe in general terms. Universal laws, along with theories and hypotheses, are used to understand and explain natural phenomena. However, many natural phenomena are incredibly complicated and may never be fully understood in terms of physical laws. This is particularly true of the ways that organisms and ecosystems are organized and function.

Scientific investigations may be pure or applied. Pure science is driven by intellectual curiosity – it is the unfettered search for knowledge and understanding, without regard for its usefulness in human welfare. Applied science is more goal-oriented and deals with practical difficulties and problems of one sort or another. Applied science might examine how to improve technology, advance the management of natural resources, or reduce pollution or other environmental damages associated with human activities.

Attribution

1.6 SCIENCE DENIAL AND EVALUATING SOURCES

Introductory science courses usually deal with accepted scientific theory and do not include opposing ideas, even though these alternate ideas may be credible. This makes it easier for students to understand complex material. Advanced students will encounter more controversies as they continue to study their discipline.

Some groups argue that some established scientific theories are wrong, not based on their scientific merit but the group's ideology. This section focuses on how to identify evidence-based information and differentiate it from pseudoscience.

Science Denial

Science denial happens when people argue that established scientific theories are wrong, not based on scientific merit but rather on subjective ideology – such as for social, political, or economic reasons. Organizations and people use science denial as a rhetorical argument against issues or ideas they oppose. Three examples of science denial versus science are:

- Teaching evolution in public schools
- Linking tobacco smoke to cancer
- Linking human activity to climate change.

A climate denier denies explicitly or doubts the objective conclusions of geologists and climate scientists. Science denial generally uses three false arguments. The first argument tries to undermine the scientific conclusion's credibility by claiming the research methods are flawed, or the theory is not universally accepted—the science is unsettled. The notion that scientific ideas are not absolute creates doubt for non-scientists; however, a lack of universal truths should not be confused with scientific uncertainty. Because science is based on falsifiability, scientists avoid claiming universal truths and use language that conveys uncertainty. This allows scientific ideas to change and evolve as more evidence is uncovered.

The second argument claims the researchers are not objective and motivated by ideology or economic agenda. This is an ***ad hominem* argument** in which a person's character is attacked instead of the merit of their argument. They claim results have been manipulated so researchers can justify asking for more funding. They claim that because a federal grant funds the researchers, they are using their results to lobby for expanded government regulation.

The third argument is to demand a balanced view, equal time in media coverage, and educational curricula to engender the false illusion of two equally valid arguments. Science deniers frequently demand equal coverage of their proposals, even when there is little scientific evidence supporting their ideology. For example, science deniers might demand religious explanations to be taught as an alternative to the well-established theory of evolution. Alternatively, all possible causes of climate change are discussed as equally probable, regardless of the body of evidence. Conclusions derived using the scientific method should not be confused with those based on ideologies.

Furthermore, conclusions about nature derived from ideologies have no place in science research and education. For example, it would be inappropriate to teach the flat earth model in modern geography or earth science courses because this idea has been disproved by the scientific method. Unfortunately, widespread scientific illiteracy allows these arguments to be used to suppress scientific knowledge and spread misinformation.

The formation of new conclusions based on the scientific method is the only way to change scientific conclusions. We would not teach Flat Earth geology and plate tectonics because Flat Earthers do not follow the scientific method. The fact that scientists avoid universal truths and change their ideas as more evidence is uncovered should not be seen as meaning that the science is unsettled. Because of widespread scientific illiteracy, these arguments are used by those who wish to suppress science and misinform the general public.

In a classic case of science denial, beginning in the 1960s and for the next three decades, the tobacco industry and their scientists used rhetorical arguments to deny a connection between tobacco usage and cancer. Once it became clear scientific studies overwhelmingly found that using tobacco dramatically increased a person's likelihood of getting cancer, their next strategy was to create a sense of doubt about the science. The tobacco industry suggested the results were not yet fully understood, and more study was needed. They used this doubt to lobby for delaying legislative action to warn consumers of the potential health hazards. This tactic is currently employed by those who deny the significance of human involvement in climate change.

Evaluating Sources of Information

In the age of the internet, information is plentiful. Anyone exploring scientific inquiry must discern valid sources of information from pseudoscience and misinformation. This evaluation is especially critical in scientific research because scientific knowledge is respected for its reliability. Textbooks such as this one can aid this complex and crucial task. At its roots, quality information comes from the scientific method, beginning with the empirical thinking of Aristotle. The application of the scientific method helps produce unbiased results. A valid inference or interpretation is based on objective evidence or data. Credible data and inferences are clearly labelled, separated, and differentiated. Anyone looking over the data can understand how the author's conclusion was derived or come to an alternative conclusion.

Scientific procedures are clearly defined, so the investigation can be replicated to confirm the original results or expanded further to produce new results. These measures make a scientific inquiry valid and its use as a source reputable. Of course, substandard work occasionally slips through, and retractions are published from time to time. An infamous article linking the MMR vaccine to autism appeared in the highly reputable journal *Lancet* in 1998. Journalists discovered that the author had multiple conflicts of interest and fabricated data, and the article was retracted in 2010.

In addition to methodology, data, and results, the authors of a study should be investigated. When looking into any research, the author(s) should be investigated. An author's credibility is based on multiple factors, such as having a degree in a relevant topic or being funded from an unbiased source.

The same rigor should be applied to evaluating the publisher, ensuring the results reported come from an unbiased process. The publisher should be easy to discover. Good publishers will show the latest papers in the journal and make their contact information and identification clear. Reputable journals show their peer review style. Some journals are predatory, where they use unexplained and unnecessary fees to submit and access journals. Reputable journals have recognizable editorial boards. Often, a reputable journal will associate with a trade, association, or recognized open-source initiative.

One of the hallmarks of scientific research is peer review. Research should be transparent to peer review. This allows the scientific community to reproduce experimental results, correct and retract errors, and validate theories. This allows the reproduction of experimental results, corrections of errors, and proper justification of the research to experts.

Citation is imperative to avoid plagiarism, and also allows readers to investigate an author's line of thought and conclusions. When reading scientific works, it is essential to confirm that the citations are from reputable scientific research. Most often, scientific citations are used to reference paraphrasing rather than quotes. The number of times a work is cited is said to measure the investigation has within the scientific community, although this technique is inherently biased.

Critical Evaluation of an Overload of Information

More so than any previous society, we live today in a world of accessible and abundant information. It has become remarkably easy for people to communicate with others over vast distances, turning the world into a "global village" (a phrase coined by Marshall McLuhan (1911-1980), a Canadian philosopher, to describe the phenomenon of universal networking). Technologies have facilitated this global connectedness for transferring ideas and knowledge – mainly electronic communication devices, such as radio, television, computers, and their networks. Today, these technologies compress space and time to achieve a virtually instantaneous communication. So much information is now available that the situation is often referred to as an "information overload" that must be analyzed critically. **Critical analysis** is the process of sorting information and making scientific inquiries about data. Involved in all aspects of the scientific process, critical analysis scrutinizes information, and research by posing sensible questions such as the following:

- Is the information derived from a scientific framework consisting of a hypothesis that has been developed and tested within the context of an existing body of knowledge and theory in the field?

- Were the methodologies used likely to provide data that are objective, accurate, and precise? Were the data analyzed using statistical methods appropriate to the data structure and the questions being asked?
- Were the results of the research compared with other pertinent work that has been previously published? Were key similarities and differences discussed and a conclusion deduced about what the new work reveals about the issue being investigated?
- Is the information based on research published in a refereed journal that requires highly qualified reviewers in the subject area to scrutinize the work, followed by an editorial decision about whether it warrants publication?
- If the analysis of an issue was based on incomplete or possibly inaccurate information, was a precautionary approach used to accommodate the uncertainty inherent in the recommendations? All users of published research have an obligation to critically evaluate what they are reading in these ways in order to decide whether the theory is appropriate, the methodologies reliable, and the conclusions sufficiently robust. Because so many environmental issues are controversial, with data and information presented on both sides of the debate, people need to formulate objectively critical judgments. Thus, people need a high degree of environmental literacy – an informed understanding of the causes and consequences of environmental damage. Being able to analyze information critically is a key personal benefit of studying environmental science.

Attribution

"1.1 Science as a Way of Knowing" from Physical Geography and Natural Disasters by R. Adam Dastrup, MA, GISP is licensed under a Creative Commons Attribution Non-Commerical Share-Alike 4.0 International License, except where otherwise noted.

1.7 KEY TERMS

***Ad hominem* argument:** argument that claims the researchers are not objective and motivated by ideology or economic agenda. 1.6

Astronomy is defined as the study of the objects that lie beyond our planet Earth and the processes by which these objects interact with one another. 1.2

Critical analysis is the process of sorting information and making scientific inquiries about data. 1.6

Falsifiability: any initial hypothesis may be tested and proven false. 1.4

Historical science: what we observe has already happened in the universe and we can do nothing to change it. 1.3

Hypothesis: a newly proposed model or idea that can be tested. 1.3

Method: In science, a method begins with many observations over a period of time. 1.3

Model: After obtaining trends from observations, a model provides an approximation of the particular phenomena we want to understand, subject to further testing. 1.3

Objective observation is without personal bias and the same by all individuals. 1.4

Observational science: conducting tests by observing many samples of the kind of object we want to study and noting carefully how different samples vary. 1.3

Peer review: a careful examination by other scientists in the same field conducted in applications for project funding, as well as reports for publication in academic journals. 1.3

Progress report: a report that changes constantly as new techniques and instruments become available. In this book science is emphasized as an example. 1.2

Pseudoscience is a collection of ideas that may appear scientific but does not use the scientific method. 1.4

Qualitative observations are general or relative descriptions. 1.4

Quantitative measurement is expressed with a specific numerical value. 1.4

Science denial happens when people argue that established scientific theories are wrong, not based on scientific merit but rather on subjective ideology – such as for social, political, or economic reasons. 1.6

Subjective observation: based on a person's feelings and beliefs and is unique to that individual. 1.4

CHAPTER 2: NUMBERS AND MATHEMATICS IN ASTRONOMY

Chapter Overview

2.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn to:

- Explain how scientific laws apply consistently throughout the universe and how this consistency benefits fields like astronomy.
- Describe the process of expressing numbers in both standard notation and scientific notation, and demonstrate the ability to convert between the two forms.
- Use scientific notation to describe distances in the universe, such as using light-years as a unit of measurement.
- Analyze the concept of speed, including average speed and instantaneous speed, and its relationship to distance and time.
- Examine the significance of the speed of light as a fundamental limit in receiving information from distant celestial objects.
- Create and solve ratios and proportions to express and solve problems involving comparisons of quantities, such as converting between weights on different planetary bodies.
- Apply properties of circles to calculate circumference and area, and explore the volume and surface area of a sphere based on its radius.

2.1 THE LAWS OF NATURE

Over centuries scientists have extracted various scientific laws from countless observations, hypotheses, and experiments. These scientific laws are, in a sense, the "rules" of the game that nature plays. One remarkable discovery about nature—one that underlies everything you will read about in this text—is that the same laws apply everywhere in the universe. The rules that determine the motion of stars so far away that your eye cannot see them are the same laws that determine the arc of a baseball after a batter has hit it out of the park.

Note that without the existence of such universal laws, we could not make much headway in astronomy. If each pocket of the universe had different rules, we would have little chance of interpreting what happened in other "neighbourhoods." But, the consistency of the laws of nature gives us enormous power to understand distant objects without travelling to them and learning the local laws. In the same way, if every region of a country had completely different laws, it would be very difficult to carry out commerce or even to understand the behaviour of people in those different regions. A consistent set of laws, though, allows us to apply what we learn or practice in one state to any other state.

This is not to say that our current scientific models and laws cannot change. New experiments and observations can lead to new, more sophisticated models—models that can include new phenomena and laws about their behaviour. The general theory of relativity proposed by Albert Einstein is a perfect example of such a transformation that took place about a century ago; it led us to predict, and eventually to observe, a strange new class of objects that astronomers call *black holes*. Only the patient process of observing nature ever more carefully and precisely can demonstrate the validity of such new scientific models.

One important problem in describing scientific models has to do with the limitations of language. When we try to describe complex phenomena in everyday terms, the words themselves may not be adequate to do the job. For example, you may have heard the structure of the atom likened to a miniature solar system. While some aspects of our modern model of the atom do remind us of planetary orbits, many other of its aspects are fundamentally different.

This problem is the reason scientists often prefer to describe their models using equations rather than words. In this book, which is designed to introduce the field of astronomy, we use mainly words to discuss what scientists have learned. We avoid complex math, but if this course piques your interest and you go on in science, more and more of your studies will involve the precise language of mathematics.

Attribution

2.2 EXPRESSING NUMBERS

Quantities have two parts: the number and the unit. The number tells “how many.” It is important to be able to express numbers properly so that the quantities can be communicated properly.

Standard notation is the straightforward expression of a number. Numbers such as 17, 101.5, and 0.00446 are expressed in standard notation. For relatively small numbers, standard notation is fine. However, for very large numbers, such as 306,000,000, or for very small numbers, such as 0.000000419, standard notation can be cumbersome because of the number of zeros needed to place nonzero numbers in the proper position.

Scientific notation is an expression of a number using powers of 10. Powers of 10 are used to express numbers that have many zeros:

Table 2.1 Powers of 10

10^0	= 1
10^1	= 10
10^2	= $100 = 10 \times 10$
10^3	= $1,000 = 10 \times 10 \times 10$
10^4	= $10,000 = 10 \times 10 \times 10 \times 10$

and so forth. The raised number to the right of the 10 indicating the number of factors of 10 in the original number is the **exponent**. (Scientific notation is sometimes called *exponential notation*.) The exponent’s value is equal to the number of zeros in the number expressed in standard notation.

Small numbers can also be expressed in scientific notation but with negative exponents:

Table 2.2 Powers of Negative 10

10^{-1}	$= 0.1 = \dfrac{1}{10}$
10^{-2}	$= 0.01 = \dfrac{1}{100}$
10^{-3}	$= 0.001 = \dfrac{1}{1,000}$
10^{-4}	$= 0.0001 = \dfrac{1}{10,000}$

and so forth. Again, the value of the exponent is equal to the number of zeros in the denominator of the associated fraction. A negative exponent implies a decimal number less than one.

A number is expressed in scientific notation by writing the first nonzero digit, then a decimal point, and then the rest of the digits. The part of a number in scientific notation that is multiplied by a power of 10 is called the **coefficient**. Then determine the power of 10 needed to make that number into the original number and multiply the written number by the proper power of 10. For example, to write 79,345 in scientific notation,

$$79,345 = 7.9345 \times 10,000 = 7.9345 \times 10^4$$

Thus, the number in scientific notation is 7.9345×10^4. For small numbers, the same process is used, but the exponent for the power of 10 is negative:

$$0.000411 = 4.11 \times \frac{1}{10,000} = 4.11 \times 10^{-4}$$

Typically, the extra zero digits at the end or the beginning of a number are not included.

Example 2.1

Problems

Express these numbers in scientific notation.

1. $306,000$
2. 0.00884

3. $2,760,000$
4. 0.000000559

Solutions

1. The number $306,000$ is 3.06 times $100,000$, or 3.06 times 10^5. In scientific notation, the number is 3.06×10^5.
2. The number 0.00884 is 8.84 times $\frac{1}{1,000}$, which is 8.84 times 10^{-3}. In scientific notation, the number is 8.84×10^{-3}.
3. The number $2,760,000$ is 2.76 times $1,000,000$, which is the same as 2.76 times 10^6. In scientific notation, the number is written as 2.76×10^6. Note that we omit the zeros at the end of the original number.
4. The number 0.000000559 is 5.59 times $\frac{1}{10,000,000}$, which is 5.59 times 10^{-7}. In scientific notation, the number is written as 5.59×10^{-7}.

Exercise 2.1

Express these numbers in scientific notation.

1. $23,070$
2. 0.0009706

Solutions

1. 2.307×10^4
2. 9.706×10^{-4}

Another way to determine the power of 10 in scientific notation is to count the number of places you need to move the decimal point to get a numerical value between 1 and 10. The number of places equals the power of

10. This number is positive if you move the decimal point to the right and negative if you move the decimal point to the left.

Many quantities in chemistry are expressed in scientific notation. When performing calculations, you may have to enter a number in scientific notation into a calculator. Be sure you know how to correctly enter a number in scientific notation into your calculator. Different models of calculators require different actions for properly entering scientific notation. If in doubt, consult your instructor immediately.

Exercise 2.2

1. Express these numbers in scientific notation.
    a. 56.9
    b. $563,100$
    c. 0.0804
    d. 0.00000667
2. Express these numbers in scientific notation.
    a. $-890,000$
    b. $602,000,000,000$
    c. 0.0000004099
    d. 0.0000000000000011
3. Express these numbers in scientific notation.
    a. 0.00656
    b. $65,600$
    c. $4,567,000$
    d. 0.000005507
4. Express these numbers in scientific notation.
    a. 65
    b. -321.09
    c. 0.000077099
    d. 0.000000000218

5. Express these numbers in standard notation.
    a. 1.381×10^5
    b. 5.22×10^{-7}
    c. 9.998×10^4
6. Express these numbers in standard notation.
    a. 7.11×10^{-2}
    b. 9.18×10^2
    c. 3.09×10^{-10}
7. Express these numbers in standard notation.
    a. 8.09×10^0
    b. 3.088×10^{-5}
    c. -4.239×10^2
8. Express these numbers in standard notation.
    a. 2.87×10^{-8}
    b. 1.78×10^{11}
    c. 1.381×10^{-23}
9. These numbers are not written in proper scientific notation. Rewrite them so that they are in proper scientific notation.
    a. 72.44×10^3
    b. $9,943 \times 10^{-5}$
    c. $588,399 \times 10^2$
10. These numbers are not written in proper scientific notation. Rewrite them so that they are in proper scientific notation.
    a. 0.000077×10^{-7}
    b. 0.000111×10^8
    c. $602,000 \times 10^1 8$
11. These numbers are not written in proper scientific notation. Rewrite them so that they are in proper scientific notation.

   a. 345.1×10^2
   b. 0.234×10^{-3}
   c. $1,800 \times 10^{-2}$

12. These numbers are not written in proper scientific notation. Rewrite them so that they are in proper scientific notation.

    a. $8,099 \times 10^{-8}$
    b. 34.5×10^0
    c. 0.000332×10^4

13. Write these numbers in scientific notation by counting the number of places the decimal point is moved.

    a. $123,456.78$
    b. $98,490$
    c. 0.000000445

14. Write these numbers in scientific notation by counting the number of places the decimal point is moved.

    a. 0.000552
    b. $1,987$
    c. 0.00000000887

15. Use your calculator to evaluate these expressions. Express the final answer in proper scientific notation.

    a. $456 \times (7.4 \times 10^8) =?$
    b. $(3.02 \times 10^5) \div (9.04 \times 10^1 5) =?$
    c. $0.0044 \times 0.000833 =?$

16. Use your calculator to evaluate these expressions. Express the final answer in proper scientific notation.

    a. $98,000 \times 23,000 =?$
    b. $98,000 \div 23,000 =?$
    c. $(4.6 \times 10^{-5}) \times (2.09 \times 10^3) =?$

17. Use your calculator to evaluate these expressions. Express the final answer in proper scientific notation.
    a. $45 \times 132 \div 882 = ?$
    b. $[(6.37 \times 10^4) \times (8.44 \times 10^{-4})] \div (3.2209 \times 10^{15}) = ?$
18. Use your calculator to evaluate these expressions. Express the final answer in proper scientific notation.
    a. $(9.09 \times 10^8) \div [(6.33 \times 10^9) \times (4.066 \times 10^{-7})] = ?$
    b. $9,345 \times 34.866 \div 0.00665 = ?$

Solutions to Odd-Numbered Questions

1.
    a. 5.69×10^1
    b. 5.631×10^5
    c. 8.04×10^{-2}
    d. 6.67×10^{-6}
3.
    a. 6.56×10^{-3}
    b. 6.56×10^4
    c. 4.567×10^6
    d. 5.507×10^{-6}
5.
    a. $138,100$
    b. 0.000000522
    c. $99,980$
7.
    a. 8.09
    b. 0.00003088
    c. -423.9
9.
    a. 7.244×10^4
    b. 9.943×10^{-2}
    c. 5.88399×10^7
11.
    a. 3.451×10^4
    b. 2.34×10^{-4}

   c. 1.8×10^{1}

13.
   a. 1.2345678×10^{5}
   b. 9.849×10^{4}
   c. 4.45×10^{-7}

15.
   a. $3.3744 \times 10^{1}1$
   b. 3.3407×10^{-11}
   c. 3.665×10^{-6}

17.
   a. 6.7346×10^{0}
   b. 1.6691×10^{-14}

Attribution

2.3 MULTIPLICATION AND DIVISION

Scientific notation is not only compact and convenient, it also simplifies arithmetic. To multiply two numbers expressed as powers of ten, you need only multiply the numbers out front and then *add* the exponents. If there are no numbers out front, as in $100 \times 100,000$, then you just add the exponents (in our notation, $10^2 \times 10^5 = 10^7$). When there are numbers out front, you have to multiply them, but they are much easier to deal with than numbers with many zeros in them.

Here's an example:

$$(3 \times 10^5) \times (2 \times 10^9) = (3 \times 2) \times (10^{5+9}) = 6 \times 10^{14}$$

And here's another example:

$$\begin{aligned} 0.04 \times 6,000,000 &= (4 \times 10^{-2}) \times (6 \times 10^6) \\ &= (4 \times 6) \times (10^{-2+6}) \\ &= 24 \times 10^4 \\ &= 2.4 \times 10^5 \end{aligned}$$

Note in the second example that when we added the exponents, we treated negative exponents as we do in regular arithmetic (-2 plus 6 equals 4). Also, notice that our first result had a 24 in it, which was not in the acceptable form, having two places to the left of the decimal point, and we therefore changed it to 2.4 and changed the exponent accordingly.

To divide, you divide the numbers out front and *subtract* the exponents. Here are several examples:

- $\dfrac{1,000,000}{1000} = \dfrac{10^6}{10^3} = 10^{6-3} = 10^3$
- $\dfrac{9 \times 10^{12}}{2 \times 10^3} = 4.5 \times 10^{12-3} = 4.5 \times 10^9$
- $\dfrac{2.8 \times 10^2}{6.2 \times 10^5} = .452 \times 10^{2-5} = .452 \times 10^{-3} = 4.52 \times 10^{-4}$

In the last example, our first result was not in the standard form, so we had to change 0.452 into 4.52, and change the exponent accordingly.

If this is the first time that you have met scientific notation, we urge you to practice many examples using it. You might start by solving the exercises below. Like any new language, the notation looks complicated at first but gets easier as you practice it.

Example 2.2

1. At the end of September 2015, the New Horizons spacecraft (which encountered Pluto for the first time in July 2015) was 4.898 billion km from Earth. Convert this number to scientific notation. How many astronomical units is this? (An astronomical unit is the distance from Earth to the Sun, or about 150 million km.)
   Solution
   4.898 billion is 4.898×10^9 km. One astronomical unit (AU) is 150 million km = 1.5×10^8 km. Dividing the first number by the second, we get $3.27 \times 10^{(9-8)} = 3.27 \times 10^1$ AU.
2. During the first six years of its operation, the Hubble Space Telescope circled Earth $37,000$ times, for a total of $1,280,000,000$ km. Use scientific notation to find the number of km in one orbit.
   Solution
   $$\frac{(1.28 \times 10^9\ \text{km})}{(3.7 \times 10^4\ \text{orbits})} = 0.346 \times 10^{(9-4)} = 0.346 \times 10^5 = 3.46 \times 10^4\ \text{km per orbit}$$
3. In a large university cafeteria, a soybean-vegetable burger is offered as an alternative to regular hamburgers. If $889,875$ burgers were eaten during the course of a school year, and 997 of them were veggie-burgers, what fraction and what percent of the burgers does this represent?
   Solution
   $$\frac{(9.97 \times 10^2\ \text{veggie burgers})}{(8.90 \times 10^5\ \text{total burgers})} = 1.12 \times 10^{(2-5)} = 1.12 \times 10^{-3}$$
   (or roughly about one thousandth) of the burgers were vegetarian. Percent means per hundred. So
   $$1.12 \times 10^{-3} \times 10^2 = 1.12 \times 10^{(-3+2)} = 1.12 \times 10^{-1}\ \text{percent}$$
4. In a 2012 Kelton Research poll, 36 percent of adult Americans thought that alien beings have actually landed on Earth. The number of adults in the United States in 2012 was about $222,000,000$. Use scientific notation to determine how many adults believe aliens have visited Earth.

Solution

36% is 36 hundredths or 0.36 or 3.6×10^{-1}. Multiply that by 2.22×10^{8} and you get about $7.99 \times 10^{(-1+8)} = 7.99 \times 10^{7}$ or almost 80 million people who believe that aliens have landed on our planet. We need more astronomy courses to educate all those people.

5. In the school year 2009–2010, American colleges and universities awarded $2,354,678$ degrees. Among these were $48,069$ PhD degrees. What fraction of the degrees were PhDs? Express this number as a percent. (Now go and find a job for all those PhDs!)

   Solution

   $$\frac{(4.81 \times 10^{4})}{(2.35 \times 10^{6})} = 2.05 \times 10^{(4-6)} = 2.05 \times 10^{-2} = \text{about } 2\%$$

   (Note that in these examples we are rounding off some of the numbers so that we don't have more than 2 places after the decimal point.)

6. A star 60 light-years away has been found to have a large planet orbiting it. Your uncle wants to know the distance to this planet in old-fashioned miles. Assume light travels $186,000$ miles per second, and there are 60 seconds in a minute, 60 minutes in an hour, 24 hours in a day, and 365 days in a year. How many miles away is that star?

   Solution

   One light-year is the distance that light travels in one year. (Usually, we use metric units and not the old British system that the United States is still using, but we are going to humor your uncle and stick with miles.) If light travels $186,000$ miles every second, then it will travel 60 times that in a minute, and 60 times that in an hour, and 24 times that in a day, and 365 times that in a year. So we have

   $$1.86 \times 10^{5} \times 6.0 \times 10^{1} \times 6.0 \times 10^{1} \times 2.4 \times 10^{1} \times 3.65 \times 10^{2}.$$

   So we multiply all the numbers out front together and add all the exponents. We get $586.57 \times 10^{10} = 5.86 \times 10^{12}$ miles in a light year (which is roughly 6 trillion miles—a heck of a lot of miles). So if the star is 60 light-years away, its distance in miles is

   $$6 \times 10^{1} \times 5.86 \times 10^{12} = 35.16 \times 10^{13} = 3.516 \times 10^{14} \text{ miles.}$$

Attribution

2.4 UNITS USED IN SCIENCE

In the American system of measurement (originally developed in England), the fundamental units of length, weight, and time are the foot, pound, and second, respectively. There are also larger and smaller units, which include the ton (2240 lb), the mile (5280 ft), the rod ($16\frac{1}{2}$ ft), the yard (3 ft), the inch ($1/12$ ft), the ounce ($1/16$ lb), and so on. Such units, whose origins in decisions by British royalty have been forgotten by most people, are quite inconvenient for conversion or doing calculations.

In science, therefore, it is more usual to use the **metric system**, which has been adopted in virtually all countries except the United States. Its great advantage is that every unit increases by a factor of ten, instead of the strange factors in the American system. The fundamental units of the metric system are:

- length: 1 meter (m)
- mass: 1 kilogram (kg)
- time: 1 second (s)

A meter was originally intended to be 1 ten-millionth of the distance from the equator to the North Pole along the surface of Earth. It is about 1.1 yd. A kilogram is the mass that on Earth results in a weight of about 2.2 lb. The second is the same in metric and American units.

Length

The most commonly used quantities of length of the metric system are the following.

Table 2.3 Length

Conversions
1 kilometer (km) = 1000 metres = 0.6214 mile
1 meter (m) = 0.001 km = 1.094 yards = 39.37 inches
1 centimeter (cm) = 0.01 meter = 0.3937 inch
1 millimeter (mm) = 0.001 meter = 0.1 cm
1 micrometer (μm) = 0.000001 meter = 0.0001 cm
1 nanometer (nm) = 10^{-9} meter = 10^{-7} cm

To convert from the American system, here are a few helpful factors:

- 1 mile = 1.61 km
- 1 inch = 2.54 cm

Mass

Although we don't make the distinction very carefully in everyday life on Earth, strictly speaking the kilogram is a unit of mass (measuring the quantity of matter in a body, roughly how many atoms it has,) while the pound is a unit of weight (measuring how strongly Earth's gravity pulls on a body).

The most commonly used quantities of mass of the metric system are the following.

Table 2.4 Mass

Conversions
1 metric ton = 10^6 grams = 1000 kg (and it produces a weight of 2.205×10^3 lb on Earth)
1 kg = 1000 grams (and it produces a weight of 2.2046 lb on Earth)
1 gram (g) = 0.0353 oz (and the equivalent weight is 0.002205 lb)
1 milligram (mg) = 0.001 g

A weight of 1 lb is equivalent on Earth to a mass of 0.4536 kg, while a weight of 1 oz is produced by a mass of 28.35 g.

Temperature

Three temperature scales are in general use:

- Fahrenheit (F); water freezes at 32 °F and boils at 212 °F.
- Celsius or centigrade[1] (C); water freezes at 0 °C and boils at 100 °C.
- Kelvin or absolute (K); water freezes at 273 K and boils at 373 K.

All molecular motion ceases at about -459 ºF =-273 ºC=0 K, a temperature called **absolute zero**. Kelvin

---

[1] Celsius is now the name used for centigrade temperature; it has a more modern standardization but differs from the old centigrade scale by less than [latex]0.1[/latex]°.

temperature is measured from this lowest possible temperature, and it is the temperature scale most often used in astronomy. Kelvins have the same value as centigrade or Celsius degrees, since the difference between the freezing and boiling points of water is 100 degrees in each. (Note that we just say "kelvins," not kelvin degrees.)

On the Fahrenheit scale, the difference between the freezing and boiling points of water is 180 degrees. Thus, to convert Celsius degrees or kelvins to Fahrenheit degrees, it is necessary to multiply by

$$\frac{180}{100}=\frac{9}{5}.$$

To convert from Fahrenheit degrees to Celsius degrees or kelvins, it is necessary to multiply by

$$\frac{100}{180}=\frac{5}{9}.$$

The full conversion formulas are:

- K = °C $+273$
- °C = $0.555\times$ (°F -32)
- °F = $(1.8\times$ °C$)+32$

Attribution

2.5 AVERAGE SPEED

A description of how fast or slow an object moves is its speed. **Speed** is the rate at which an object changes its location. The SI unit of time is the second (s), and the SI unit of speed is metres per second (m/s), but sometimes kilometres per hour (km/h), miles per hour (mph) or other units of speed are used.

When you describe an object's speed, you often describe the average over a time period. **Average speed**, V_{avg}, is the distance traveled divided by the time during which the motion occurs.

$$V_{avg} = \frac{\text{distance}}{\text{time}}$$

You can, of course, rearrange the equation to solve for either distance or time:

$$\text{time} = \frac{\text{distance}}{V_{avg}}$$

$$\text{distance} = V_{avg} \times \text{time}$$

Suppose, for example, a car travels 150 kilometres in 3.2 hours. Its average speed for the trip is

$$\begin{aligned} V_{avg} &= \frac{\text{distance}}{\text{time}} \\ &= \frac{150\text{ km}}{3.2\text{ h}} \\ &= 47\text{ km/h} \end{aligned}$$

A car's speed would likely increase and decrease many times over a 3.2 hour trip. Its speed at a specific instant in time, however, is its **instantaneous speed**. A car's speedometer describes its instantaneous speed.

Attribution

2.6 USEFUL MEASUREMENTS IN ASTRONOMY

In astronomy we deal with distances on a scale you may never have thought about before, with numbers larger than any you may have encountered. We adopt two approaches that make dealing with astronomical numbers a little bit easier. First, we use a system for writing large and small numbers called *scientific notation* (or sometimes *powers-of-ten notation*). This system is very appealing because it eliminates the many zeros that can seem overwhelming to the reader. In scientific notation, if you want to write a number such as $500,000,000$, you express it as 5×10^8. The small raised number after the 10, called an *exponent*, keeps track of the number of places we had to move the decimal point to the left to convert $500,000,000$ to 5. The second way we try to keep numbers simple is to use a consistent set of units—the metric International System of Units, or SI (from the French *Système International d'Unités*).

Watch this brief PBS animation that explains how scientific notation works and why it's useful.

A common unit astronomers use to describe distances in the universe is a **light-year**, which is the distance light travels during one year. Because light always travels at the same speed, and because its speed turns out to be the fastest possible speed in the universe, it makes a good standard for keeping track of distances. You might be confused because a "light-year" seems to imply that we are measuring time, but this mix-up of time and distance is common in everyday life as well. For example, when your friend asks where the movie theater is located, you might say "about 20 minutes from downtown."

So, how many kilometres are there in a light-year? Light travels at the amazing pace of 3×10^5 kilometres per second (km/s), which makes a light-year 9.46×10^{12} kilometres. You might think that such a large unit would reach the nearest star easily, but the stars are far more remote than our imaginations might lead us to believe. Even the nearest star is 4.3 light-years away—more than 40 trillion kilometres. Other stars visible to the unaided eye are hundreds to thousands of light-years away as seen in Figure 2.1.

ORION NEBULA

Figure 2.1. This beautiful cloud of cosmic raw material (gas and dust from which new stars and planets are being made) called the Orion Nebula is about 1400 light-years away. That's a distance of roughly 1.34 × 1016 kilometres—a pretty big number. The gas and dust in this region are illuminated by the intense light from a few extremely energetic adolescent stars.
Hubble's Sharpest View of the Orion Nebula by NASA, ESA, M. Robberto (Space Telescope Science Institute/ESA) and the Hubble Space Telescope Orion Treasury Project Team, NASA Media Licence.

Table 2.4 Astronomical Constants

Name	Value
speed of Light (**c**)	$3 \times 10^5 \text{ km/s}$
astronomical unit (**AU**)	$1.496 \times 10^{11} \text{ m}$
light-year (**ly**)	$9.461 \times 10^{15} \text{ m}$
parsec (**pc**)	$3.086 \times 10^{16} \text{ m} = 3.262 \text{ light-years}$
sidereal year (**y**)	$3.156 \times 10^7 \text{ s}$
mass of Earth (M_{Earth})	$5.974 \times 10^{24} \text{ kg}$
equatorial radius of Earth (R_{Earth})	$6.378 \times 10^6 \text{ m}$
obliquity of ecliptic	$23.4^\circ 26'$
surface gravity of Earth (g)	9.807 m/s^2
escape velocity of Earth (V_{Earth})	$1.119 \times 10^4 \text{ m/s}$
mass of Sun (M_{Sun})	$1.989 \times 10^{30} \text{ kg}$
equatorial radius of Sun (R_{Sun})	$6.960 \times 10^8 \text{ m}$
luminosity of Sun (L_{Sun})	$3.85 \times 10^{26} \text{ W}$
solar constant (flux of energy received at Earth) (**S**)	$1.368 \times 10^3 \text{ W/m}^2$
Hubble constant (**H₀**)	approximately 20 km/s per million light-years, or approximately 70 km/s per megaparsec

---

Attribution

"1.4 Numbers in Astronomy and Scientific Notation or Powers of Ten" and "A.5 Some Useful Constants for Astronomy" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

2.7 CONSEQUENCES OF THE FINITE SPEED OF LIGHT

There is another reason the speed of light is such a natural unit of distance for astronomers. Information about the universe comes to us almost exclusively through various forms of light, and all such light travels at the speed of light-that is, 1 light-year every year. This sets a limit on how quickly we can learn about events in the universe. If a star is 100 light-years away, the light we see from it tonight left that star 100 years ago and is just now arriving in our neighborhood. The soonest we can learn about any changes in that star is 100 years after the fact. For a star 500 light-years away, the light we detect tonight left 500 years ago and is carrying 500-year-old news.

Because many of us are accustomed to instant news from the Internet, some might find this frustrating.

“You mean, when I see that star up there,” you ask, “I won’t know what’s actually happening there for another 500 years?”

But this isn’t the most helpful way to think about the situation. For astronomers, *now* is when the light reaches us here on Earth. There is no way for us to know anything about that star (or other object) until its light reaches us.

But what at first may seem a great frustration is actually a tremendous benefit in disguise. If astronomers really want to piece together what has happened in the universe since its beginning, they must find evidence about each epoch (or period of time) of the past. Where can we find evidence today about cosmic events that occurred billions of years ago?

The delay in the arrival of light provides an answer to this question. The farther out in space we look, the longer the light has taken to get here, and the longer ago it left its place of origin. By looking billions of light-years out into space, astronomers are actually seeing billions of years into the past. In this way, we can reconstruct the history of the cosmos and get a sense of how it has evolved over time.

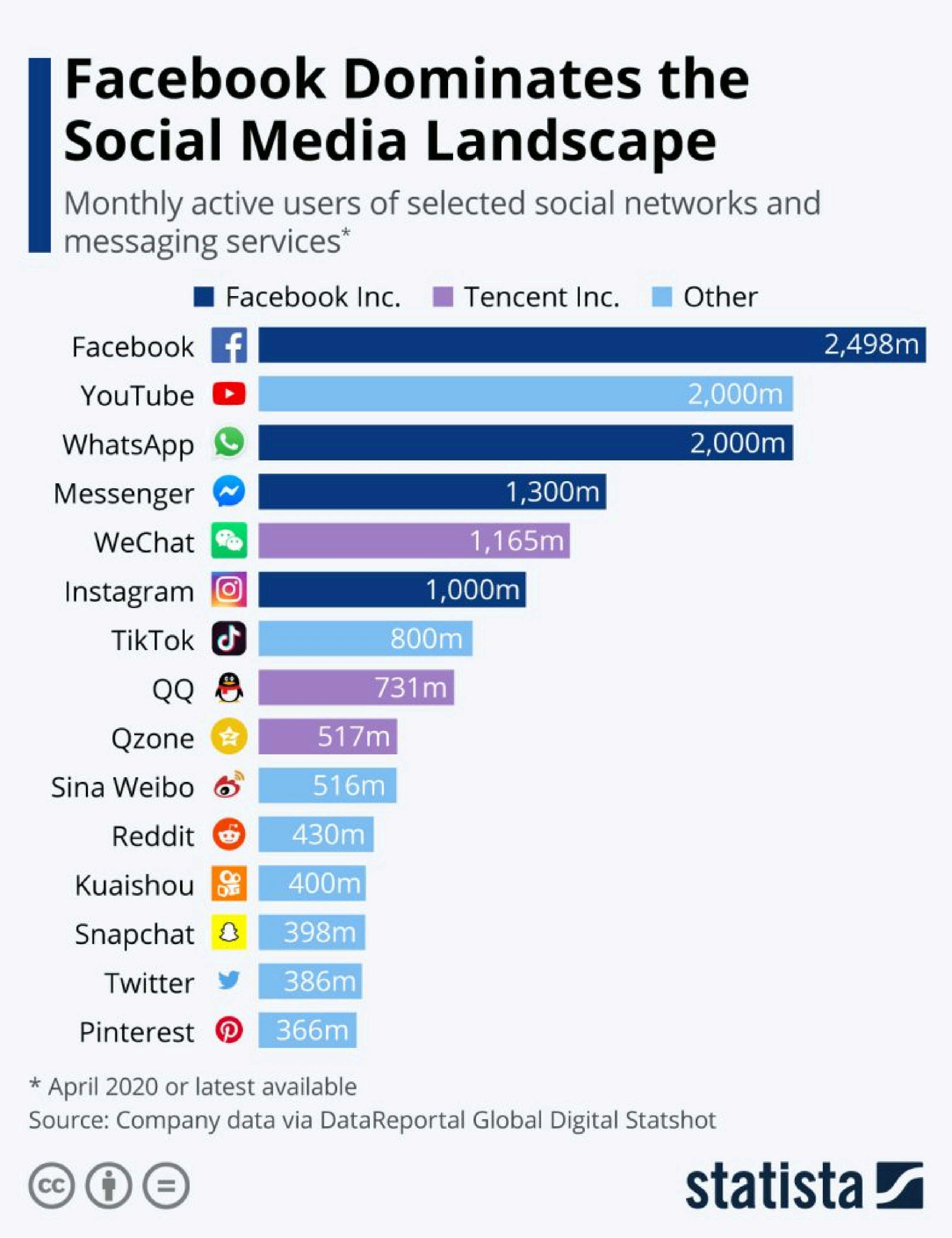
This is one reason why astronomers strive to build telescopes that can collect more and more of the faint light in the universe. The more light we collect, the fainter the objects we can observe. On average, fainter objects are farther away and can, therefore, tell us about periods of time even deeper in the past.

Attribution

“1.5 Consequences of Light Travel Time” from Douglas College Astronomy 1105 by Douglas College

2.8 RATIOS

Ratios and proportions are used in a wide variety of situations to make comparisons. For example, using the information from Figure 2.2, we can see that the number of Facebook users compared to the number of Twitter users is 2,498 m to 386 m. Note that the "m" stands for million, so 2,498 million is actually 2,498,000,000 and 386 million is 386,000,000. Similarly, the number of Qzone users compared to the number of Pinterest users is in a ratio of 517 million to 366 million. These types of comparisons are ratios.



* April 2020 or latest available

Source: Company data via DataReportal Global Digital Statshot

Figure 2.2 This bar graph shows popular social media app usage.

Chart: Facebook Inc. Dominates the Social Media Landscape by Statista, CC BY-ND 3.0.

Constructing Ratios to Express Comparison of Two Quantities

Note there are three different ways to write a **ratio**, which is a comparison of two numbers that can be written as: a to b OR $a:b$ OR the fraction a/b. Which method you use often depends upon the situation. For the most part, we will want to write our ratios using the fraction notation. Note that, while all ratios are fractions, not all fractions are ratios. Ratios make part to part, part to whole, and whole to part comparisons. Fractions make part to whole comparisons only.

Example 2.3

Expressing the Relationship between Two Currencies as a Ratio

The Euro (€) is the most common currency used in Europe. Twenty-two nations, including Italy, France, Germany, Spain, Portugal, and the Netherlands use it. On June 9, 2021, 1 U.S. dollar was worth 0.82 Euros. Write this comparison as a ratio.

Solution

Using the definition of ratio, let $a=1$ U.S. dollar and let $b=0.82$ Euros. Then the ratio can be written as either 1 to 0.82; or $1:0.82$; or $\frac{1}{0.82}$.

Formula does not parse

Exercise 2.3

On June 9, 2021, 1 U.S. dollar was worth 1.21 Canadian dollars. Write this comparison as a ratio.

Solution

$a = 1$ U.S. dollar, and $b = 1.21$ Canadian dollars, the ratio is 1 to 1.21; or $1 : 1.21$; or $\frac{1}{1.21}$.

Example 2.4

Expressing the Relationship between Two Weights as a Ratio

The gravitational pull on various planetary bodies in our solar system varies. Because weight is the force of gravity acting upon a mass, the weights of objects is different on various planetary bodies than they are on Earth. For example, a person who weighs 200 pounds on Earth would weigh only 33 pounds on the moon! Write this comparison as a ratio.

pounds on Earth and let

Solution
Using the definition of ratio, let $a = 200$ pounds on Earth and let $b = 33$ pounds on the moon. Then the ratio can be written as either 200 to 33; or $200 : 33$; or $\frac{200}{33}$.

Exercise 2.4

A person who weighs 170 pounds on Earth would weigh 64 pounds on Mars. Write this comparison as a ratio.

Solution
With $a = 170$ pounds on Earth, and $b = 64$ pounds on Mars, the ratio is 170 to 64; or $170 : 64$; or $\frac{170}{64}$.

Using and Applying Proportional Relationships to Solve Problems

Using proportions to solve problems is a very useful method. It is usually used when you know three parts of the proportion, and one part is unknown. Proportions are often solved by setting up like ratios. If $\frac{a}{b}$ and $\frac{c}{d}$ are two ratios such that $\frac{a}{b} = \frac{c}{d}$ then the fractions are said to be **proportional**. Also, two fractions $\frac{a}{b}$ and $\frac{c}{d}$ are proportional ($\frac{a}{b} = \frac{c}{d}$) if and only if $a \times d = b \times c$.

Example 2.5

Solving a Proportion Involving Two Currencies

You are going to take a trip to France. You have $\$520$ U.S. dollars that you wish to convert to Euros. You know that 1 U.S. dollar is worth 0.82 Euros. How much money in Euros can you get in exchange for $\$520$?

be the variable that represents the unknown. Notice that U.S. dollar amounts are in both numerators and Euro amounts are in both denominators.

Solution

Step 1: Set up the two ratios into a proportion; let x be the variable that represents the unknown. Notice that U.S. dollar amounts are in both numerators and Euro amounts are in both denominators.

$$\frac{1}{0.82} = \frac{520}{x}$$

Step 2: Cross multiply, since the ratios $\frac{a}{b}$ and $\frac{c}{d}$ are proportional, then

$$520(0.82) = 1(x)$$
$$426.4 = x$$

You should receive 426.4 Euros.

Exercise 2.5

After your trip to France, you have 180 Euros remaining. You wish to convert them back into U.S. dollars. Assuming the exchange rate is the same ($\$1 = 0.82$ €), how many dollars should you receive? Round to the nearest cent if necessary.

Solution

$\$219.51$

Example 2.6

Solving a Proportion Involving Weights on Different Planets

A person who weighs 170 pounds on Earth would weigh 64 pounds on Mars. How much would a typical racehorse ($1,000$ pounds) weigh on Mars? Round your answer to the nearest tenth.

Solution

Step 1: Set up the two ratios into a proportion. Notice the Earth weights are both in the numerator and the Mars weights are both in the denominator.

$$\frac{170}{64} = \frac{1,000}{x}$$

Step 2: Cross multiply, and then divide to solve.

$$\begin{aligned} 170x &= 1,000(64) \\ 170x &= 64,000 \\ \frac{170x}{170} &= \frac{64,000}{170} \\ x &= 376.5 \end{aligned}$$

So the $1,000$-pound horse would weigh about 376.5 pounds on Mars.

Exercise 2.6

A person who weighs 200 pounds on Earth would weigh only 33 pounds on the moon. A 2021 Toyota Prius weighs $3,040$ pounds on Earth; how much would it weigh on the moon? Round to the nearest tenth if necessary.

Solution

501.6 pounds

Example 2.7

Solving a Proportion Involving Baking

A cookie recipe needs $2\frac{1}{4}$ cups of flour to make 60 cookies. Jackie is baking cookies for a large fundraiser; she is told she needs to bake $1,020$ cookies! How many cups of flour will she need?

Solution

Step 1: Set up the two ratios into a proportion. Notice that the cups of flour are both in the numerator and the amounts of cookies are both in the denominator. To make the calculations more efficient, the cups of flour $(2\frac{1}{4})$ is converted to a decimal number (2.25).

Formula does not parse

Step 2: Cross multiply, and then simplify to solve.

$$\begin{aligned}2.25(1,020) &= 60x\\ 2,295 &= 60x\\ 38.25 &= x\end{aligned}$$

Jackie will need 38.25, or $38\frac{1}{4}$, cups of flour to bake $1,020$ cookies.

Attribution

2.9 GEOMETRY

Properties of Circles

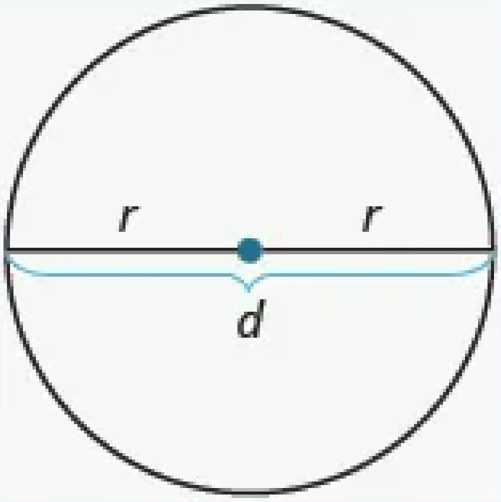


Figure 2.3. Properties of a circle.

- r is the length of the radius
- d is the length of the diameter
- $d = 2r$
- Circumference is the perimeter of a circle. The formula for circumference is

$$C = \frac{2}{\pi r}$$

- The formula for area of a circle is

$$A = \pi r^2$$

Remember, that we approximate π with 3.14 or $\frac{22}{7}$ depending on whether the radius of the circle is given as a decimal or a fraction. If you use the π key on your calculator to do the calculations in this chapter, your answers will be slightly different from the answers shown. That is because the π key uses more than two decimal places.

Volume and Surface Area of a Sphere

For a sphere with radius r:

Figure 2.4. Volume and Surface Area of a Sphere.

Attribution

"9.5 Solve Geometry Applications: Circles and Irregular Figures" and "9.6 Solve Geometry Applications: Volume and Surface Area" from Prealgebra 2e by Lynn Marecek, MaryAnne Anthony-Smith, and Andrea Honeycutt Mathis, © OpenStax are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

2.10 KEY TERMS

Absolute zero: temperature at which all molecular motion ceases (at about $-459\ {}^{\circ}\mathrm{F} = -273\ {}^{\circ}\mathrm{C} = 0\ \mathrm{K}$) 2.4

Average speed (V_{avg}): is the distance traveled divided by the time during which the motion occurs. 2.5

Coefficient: the part of a number in scientific notation that is multiplied by a power of 10. 2.2

Exponent: the raised number to the right of the 10 indicating the number of factors of 10 in the original number. 2.2

Instantaneous speed: the speed at a specific instant in time. 2.5

Light-year: the distance that light travels during one year. 2.6

Metric system: system of measurement that has been adopted in virtually all countries except the United States and where every unit increases by a factor of ten. 2.4

Proportional: for ratios, if $\frac{a}{b}$ and $\frac{c}{d}$ are two ratios such that $\frac{a}{b} = \frac{c}{d}$ then the fractions are said to be proportional. 2.8

Ratio: a comparison of two numbers that can be written as: a to b OR $a : b$ OR the fraction a/b. 2.8

Scientific notation is an expression of a number using powers of 10. 2.2

Standard notation is the straightforward expression of a number. 2.2

Speed is the rate at which an object changes its location. 2.5

CHAPTER 3: AN OVERVIEW OF THE UNIVERSE

Chapter Overview

3.0 LEARNING OBJECTIVES

Learning Objectives

By the end of this chapter, learners will be able to:

- Identify the order of celestial bodies encountered during a space traveler's journey from Earth to distant galaxies.
- Explain the concept of an astronomical unit (AU).
- Assess the significance of the discovery of dark matter and its role in shaping our understanding of the universe, including its implications on the formation and evolution of galaxies.
- Discuss the fundamental forces that explain all observable phenomena in the universe.

3.1 A TOUR OF THE UNIVERSE

We can now take a brief introductory tour of the universe as astronomers understand it today to get acquainted with the types of objects and distances you will encounter throughout the text. We begin at home with Earth, a nearly spherical planet about 13,000 kilometres in diameter as shown in Figure 3.1. A space traveller entering our planetary system would easily distinguish Earth from the other planets in our solar system by the large amount of liquid water that covers some two thirds of its crust. If the traveller had equipment to receive radio or television signals, or came close enough to see the lights of our cities at night, she would soon find signs that this watery planet has sentient life.

HUMANITY’S HOME BASE

Figure 3.1. This image shows the Western hemisphere as viewed from space 35,400 kilometres (about 22,000 miles) above Earth. Data about the land surface from one satellite was combined with another satellite's data about the clouds to create the image.
Blue Marble 2000 by R. Stockli, A. Nelson, F. Hasler, NASA/ GSFC/ NOAA/ USGS, CC BY 2.0.

Our nearest astronomical neighbour is Earth's satellite, commonly called the **Moon**. Figure 2 shows Earth and the Moon drawn to scale on the same diagram. Notice how small we have to make these bodies to fit them on the page with the right scale. The Moon's distance from Earth is about 30 times Earth's diameter, or approximately 384,000 kilometres, and it takes about a month for the Moon to revolve around Earth. The Moon's diameter is 3476 kilometres, about one fourth the size of Earth.

Earth and Moon, Drawn to Scale

Figure 3.2. This image shows Earth and the Moon shown to scale for both size and distance. Earth-Moon by Nickshanks, CC BY 2.5.

Light (or radio waves) takes 1.3 seconds to travel between Earth and the Moon. If you've seen videos of the Apollo flights to the Moon, you may recall that there was a delay of about 3 seconds between the time Mission Control asked a question and the time the astronauts responded. This was not because the astronomers were thinking slowly, but rather because it took the radio waves almost 3 seconds to make the round trip.

Earth revolves around our star, the Sun, which is about 150 million kilometres away—approximately 400 times as far away from us as the Moon. We call the average Earth–Sun distance an **astronomical unit (AU)** because, in the early days of astronomy, it was the most important measuring standard. Light takes slightly more than 8 minutes to travel 1 astronomical unit, which means the latest news we receive from the Sun is always 8 minutes old. The diameter of the Sun is about 1.5 million kilometres; Earth could fit comfortably inside one of the minor eruptions that occurs on the surface of our star. If the Sun were reduced to the size of a basketball, Earth would be a small apple seed about 30 metres from the ball.

It takes Earth 1 year (3×10^7 seconds) to go around the Sun at our distance; to make it around, we must travel at approximately 110,000 kilometres per hour. (If you, like many students, still prefer miles to kilometres, you might find the following trick helpful. To convert kilometres to miles, just multiply kilometres by 0.6. Thus, 110,000 kilometres per hour becomes 66,000 miles per hour.) Because gravity holds us firmly to Earth and there is no resistance to Earth's motion in the vacuum of space, we participate in this extremely fast-moving trip without being aware of it day to day.

Earth is only one of eight planets that revolve around the Sun. These planets, along with their moons and swarms of smaller bodies such as dwarf planets, make up the solar system as shown Figure 3.3. A **planet** is defined as a body of significant size that orbits a star and does not produce its own light. (If a large body consistently produces its own light, it is then called a **star**.) Later in the book this definition will be modified a bit, but it is perfectly fine for now as you begin your voyage.

Our Solar Family

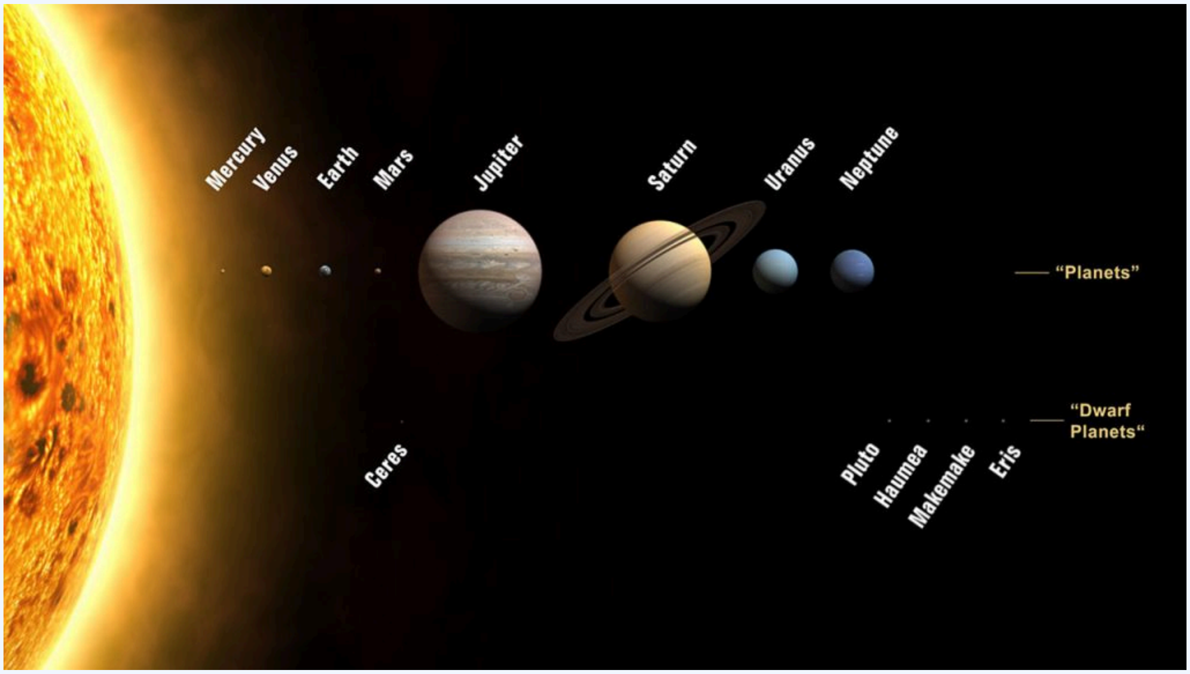


Figure 3.3. The Sun, the planets, and some dwarf planets are shown with their sizes drawn to scale. The orbits of the planets are much more widely separated than shown in this drawing. Notice the size of Earth compared to the giant planets.
Planets 2008 by Farry, NASA, NASA Media Licence.

We are able to see the nearby planets in our skies only because they reflect the light of our local star, the Sun. If the planets were much farther away, the tiny amount of light they reflect would usually not be visible to us. The planets we have so far discovered orbiting other stars were found from the pull their gravity exerts on their parent stars, or from the light they block from their stars when they pass in front of them. We can't see most of these planets directly, although a few are now being imaged directly.

The Sun is our local star, and all the other stars are also enormous balls of glowing gas that generate vast amounts of energy by nuclear reactions deep within. We will discuss the processes that cause stars to shine in more detail later in the book. The other stars look faint only because they are so very far away. If we continue our basketball analogy, Proxima Centauri, the nearest star beyond the Sun, which is 4.3 light-years away, would be almost 7000 kilometres from the basketball.

When you look up at a star-filled sky on a clear night, all the stars visible to the unaided eye are part of a single collection of stars we call the **Milky Way Galaxy**, or simply the *Galaxy*. (When referring to the Milky Way, we capitalize *Galaxy*; when talking about other galaxies of stars, we use lowercase *galaxy*.) The Sun is one of hundreds of billions of stars that make up the Galaxy; its extent, as we will see, staggers the human imagination. Within a sphere 10 light-years in radius centred on the Sun, we find roughly ten stars. Within a sphere 100 light-years in radius, there are roughly 10,000 (10^4) stars—far too many to count or name—but we have still traversed only a tiny part of the Milky Way Galaxy. Within a 1000-light-year sphere, we find some ten million (10^7) stars; within a sphere of 100,000 light-years, we finally encompass the entire Milky Way Galaxy.

Our Galaxy looks like a giant disk with a small ball in the middle. If we could move outside our Galaxy and look down on the disk of the Milky Way from above, it would probably resemble the galaxy in Figure 3.4, with its spiral structure outlined by the blue light of hot adolescent stars.

Spiral Galaxy

Figure 3.4. This galaxy of billions of stars, called by its catalogue number NGC 1073, is thought to be similar to our own Milky Way Galaxy. Here we see the giant wheel-shaped system with a bar of stars across its middle.
Hubble Observes NGC 1073 by NASA, ESA, NASA Media Licence.

The Sun is somewhat less than 30,000 light-years from the centre of the Galaxy, in a location with nothing much to distinguish it. From our position inside the Milky Way Galaxy, we cannot see through to its far rim (at least not with ordinary light) because the space between the stars is not completely empty. It contains a sparse distribution of gas (mostly the simplest element, hydrogen) intermixed with tiny solid particles that we call **interstellar dust**. This gas and dust collect into enormous clouds in many places in the Galaxy, becoming the raw material for future generations of stars. Figure 3.5 shows an image of the disk of the Galaxy as seen from our vantage point.

Milky Way Galaxy

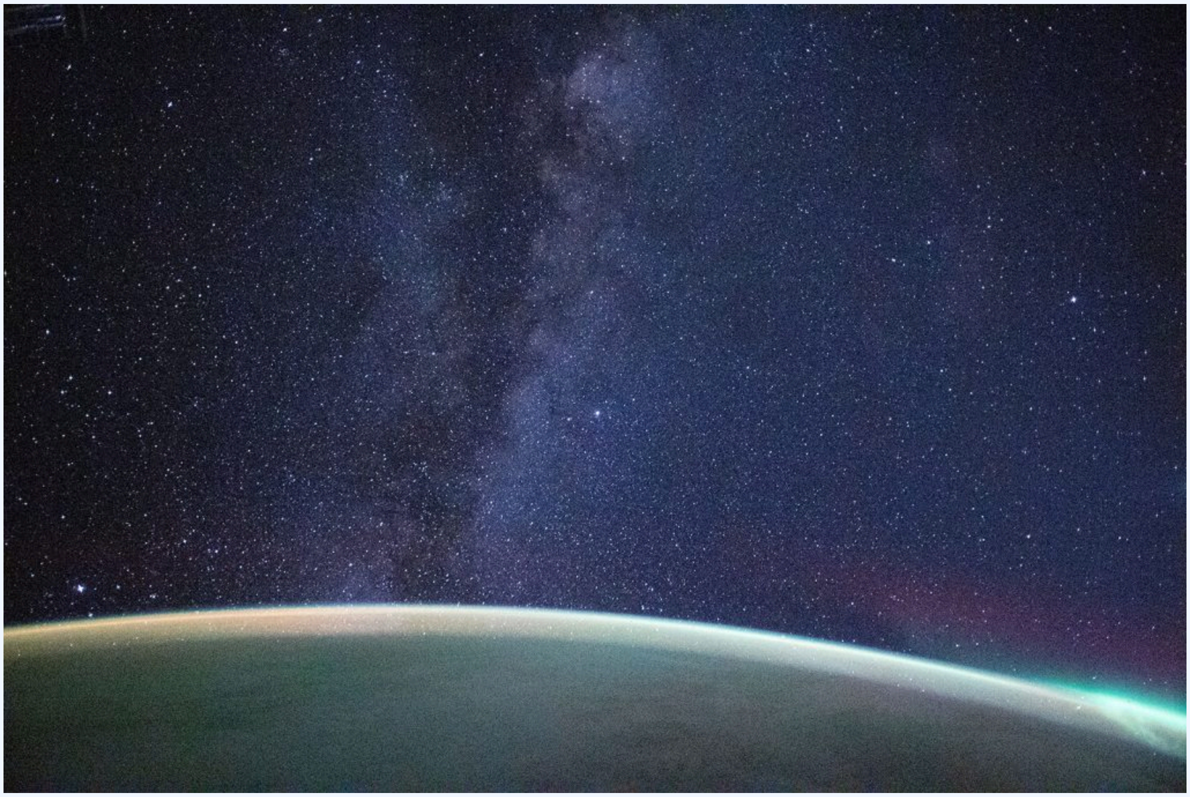


Figure 3.5. Because we are inside the Milky Way Galaxy, we see its disk in cross-section flung across the sky like a great milky white avenue of stars with dark "rifts" of dust.
The Milky Way extends above the Earth's horizon by NASA, NASA Media Licence.

Typically, the interstellar material is so extremely sparse that the space between stars is a much better vacuum than anything we can produce in terrestrial laboratories. Yet, the dust in space, building up over thousands of light-years, can block the light of more distant stars. Like the distant buildings that disappear from our view on a smoggy day in Los Angeles, the more distant regions of the Milky Way cannot be seen behind the layers of interstellar smog. Luckily, astronomers have found that stars and raw material shine with various forms of light, some of which do penetrate the smog, and so we have been able to develop a pretty good map of the Galaxy.

Recent observations, however, have also revealed a rather surprising and disturbing fact. There appears to be more—much more—to the Galaxy than meets the eye (or the telescope). From various investigations, we have evidence that much of our Galaxy is made of material we cannot currently observe directly with our instruments. We therefore call this component of the Galaxy *dark matter*. We know the dark matter is there by the pull its gravity exerts on the stars and raw material we can observe, but what this dark matter is made of and how much of it exists remain a mystery. Furthermore, this dark matter is not confined to our Galaxy; it appears to be an important part of other star groupings as well.

By the way, not all stars live by themselves, as the Sun does. Many are born in double or triple systems with two, three, or more stars revolving about each other. Because the stars influence each other in such close systems, multiple stars allow us to measure characteristics that we cannot discern from observing single stars. In a number of places, enough stars have formed together that we recognized them as star clusters as seen in Figure 3.6. Some of the largest of the star clusters that astronomers have catalogued contain hundreds of thousands of stars and take up volumes of space hundreds of light-years across.

Star Cluster

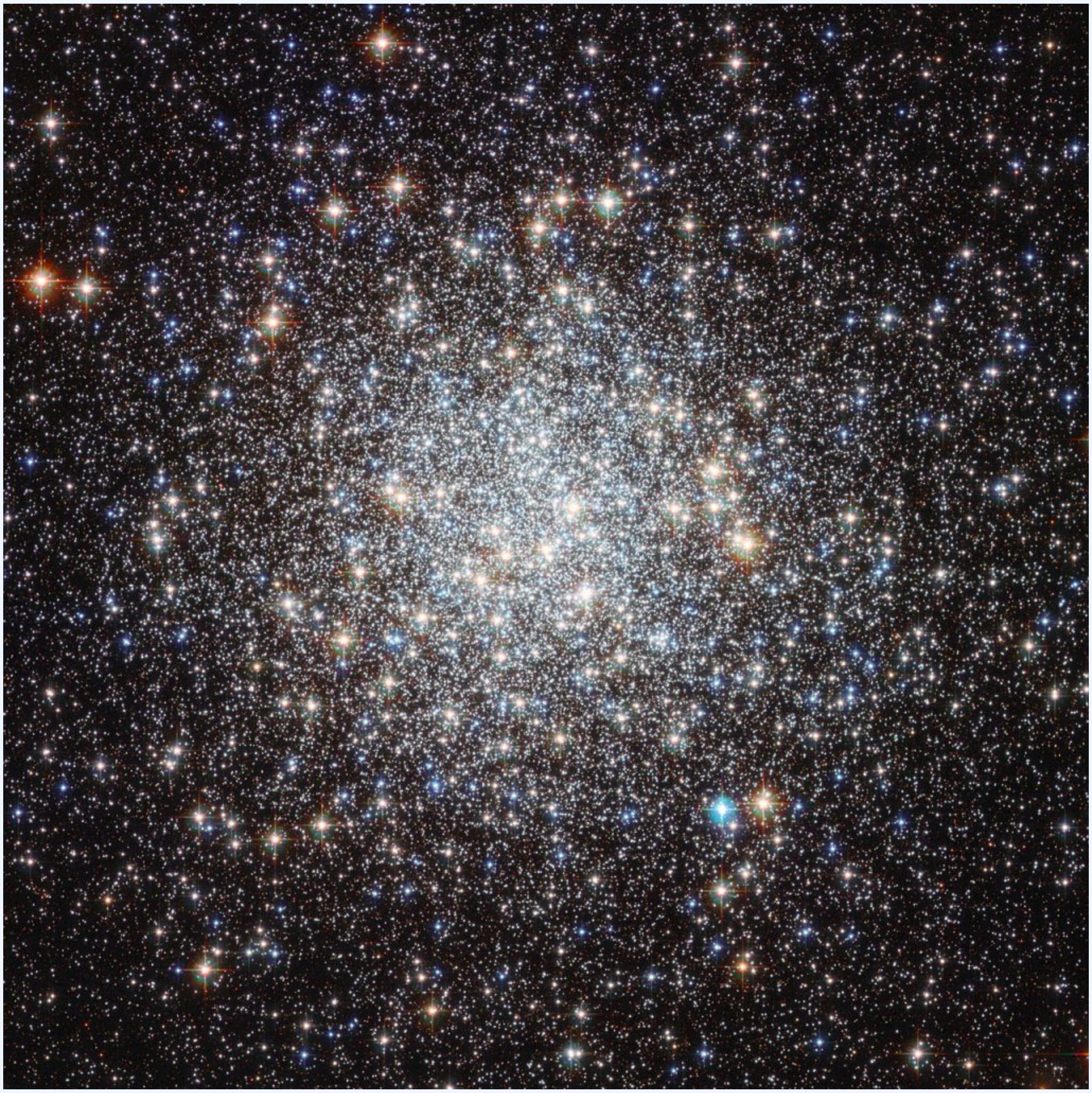


Figure 3.6 This large star cluster is known by its catalogue number, M9. It contains some 250,000 stars and is seen more clearly from space using the Hubble Space Telescope. It is located roughly 25,000 light-years away.
Messier 9 by NASA, ESA, NASA Media Licence.

You may hear stars referred to as "eternal," but in fact no star can last forever. Since the "business" of stars is making energy, and energy production requires some sort of fuel to be used up, eventually all stars run out of fuel. This news should not cause you to panic, though, because our Sun still has at least 5 or 6 billion years to go. Ultimately, the Sun and all stars will die, and it is in their death throes that some of the most intriguing and important processes of the universe are revealed. For example, we now know that many of the atoms in our bodies were once inside stars. These stars exploded at the ends of their lives, recycling their material back into the reservoir of the Galaxy. In this sense, all of us are literally made of recycled "star dust."

Attribution

3.2 THE UNIVERSE ON THE LARGE SCALE

In a very rough sense, you could think of the solar system as your house or apartment and the Galaxy as your town, made up of many houses and buildings. In the twentieth century, astronomers were able to show that, just as our world is made up of many, many towns, so the universe is made up of enormous numbers of galaxies. (We define the universe to be everything that exists that is accessible to our observations.) Galaxies stretch as far into space as our telescopes can see, many billions of them within the reach of modern instruments. When they were first discovered, some astronomers called galaxies *island universes*, and the term is aptly descriptive; galaxies do look like islands of stars in the vast, dark seas of intergalactic space.

The nearest galaxy, discovered in 1993, is a small one that lies 75,000 light-years from the Sun in the direction of the constellation Sagittarius, where the smog in our own Galaxy makes it especially difficult to discern. (A **constellation**, we should note, is one of the 88 sections into which astronomers divide the sky, each named after a prominent star pattern within it.) Beyond this Sagittarius dwarf galaxy lie two other small galaxies, about 160,000 light-years away. First recorded by Magellan's crew as he sailed around the world, these are called the *Magellanic Clouds*. They are shown in Figure 3.7. All three of these small galaxies are satellites of the Milky Way Galaxy, interacting with it through the force of gravity. Ultimately, all three may even be swallowed by our much larger Galaxy, as other small galaxies have been over the course of cosmic time.

Neighbour Galaxies

Figure 3.7. This image shows both the Large Magellanic Cloud and the Small Magellanic Cloud above the telescopes of the Atacama Large Millimeter/Submillimeter Array (ALMA) in the Atacama Desert of northern Chile.
Under the spell of the Magellanic Clouds by ESO/C. Malin, CC BY 4.0.

The nearest large galaxy is a spiral quite similar to our own, located in the constellation of Andromeda, and is thus called the Andromeda galaxy; it is also known by one of its catalog numbers, M31, shown in Figure 3.8. M31 is a little more than 2 million light-years away and, along with the Milky Way, is part of a small cluster of more than 50 galaxies referred to as the *Local Group*.

Closest Spiral Galaxy

Figure 3.8. The Andromeda galaxy (M31) is a spiral-shaped collection of stars similar to our own Milky Way.
The Galaxy Next Door by NASA/JPL-Caltech, NASA Media Licence.

At distances of 10 to 15 million light-years, we find other small galaxy groups, and then at about 50 million light-years there are more impressive systems with thousands of member galaxies. We have discovered that galaxies occur mostly in clusters, both large and small as shown in Figure 3.9.

Fornax Cluster of Galaxies

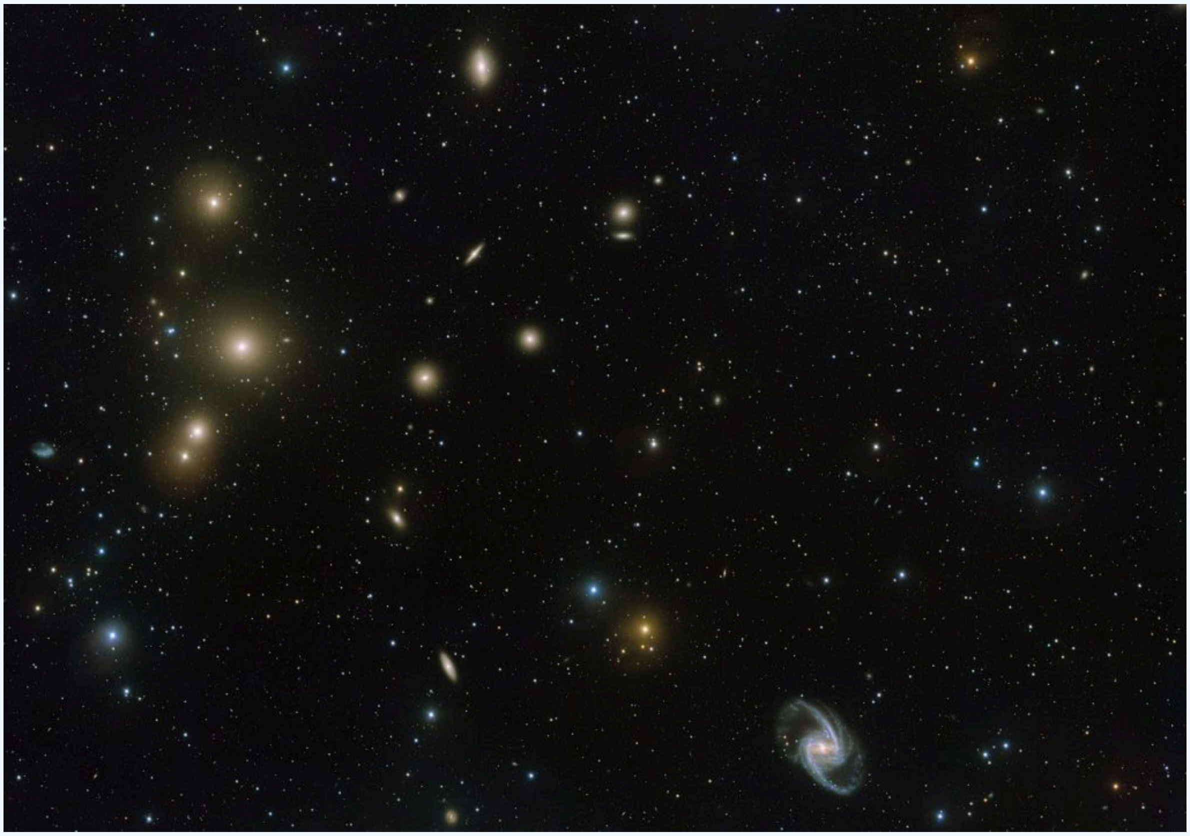


Figure 3.9. In this image, you can see part of a cluster of galaxies located about 60 million light-years away in the constellation of Fornax. All the objects that are not pinpoints of light in the picture are galaxies of billions of stars.
Inside the Fiery Furnace by ESO, J. Emerson, VISTA. Acknowledgement: Cambridge Astronomical Survey Unit, CC BY 4.0.

Some of the clusters themselves form into larger groups called **superclusters**. The Local Group is part of a supercluster of galaxies, called the Virgo Supercluster, which stretches over a diameter of 110 million light-years. We are just beginning to explore the structure of the universe at these enormous scales and are already encountering some unexpected findings.

At even greater distances, where many ordinary galaxies are too dim to see, we find **quasars**. These are brilliant centres of galaxies, glowing with the light of an extraordinarily energetic process. The enormous energy of the quasars is produced by gas that is heated to a temperature of millions of degrees as it falls toward a massive black hole and swirls around it. The brilliance of quasars makes them the most distant beacons we can see in the dark oceans of space. They allow us to probe the universe 10 billion light-years away or more, and thus 10 billion years or more in the past.

With quasars we can see way back close to the Big Bang explosion that marks the beginning of time. Beyond the quasars and the most distant visible galaxies, we have detected the feeble glow of the explosion itself, filling the universe and thus coming to us from all directions in space. The discovery of this "afterglow of creation" is considered to be one of the most significant events in twentieth-century science, and we are still exploring the many things it has to tell us about the earliest times of the universe.

Measurements of the properties of galaxies and quasars in remote locations require large telescopes, sophisticated light-amplifying devices, and painstaking labour. Every clear night, at observatories around the world, astronomers and students are at work on such mysteries as the birth of new stars and the large-scale structure of the universe, fitting their results into the tapestry of our understanding.

Attribution

3.3 THE UNIVERSE OF THE VERY SMALL

The foregoing discussion has likely impressed on you that the universe is extraordinarily large and extraordinarily empty. On average, it is 10,000 times more empty than our Galaxy. Yet, as we have seen, even the Galaxy is mostly empty space. The air we breathe has about 10^{19} atoms in each cubic centimeter—and we usually think of air as empty space. In the interstellar gas of the Galaxy, there is about one atom in every cubic centimeter. Intergalactic space is filled so sparsely that to find one atom, on average, we must search through a cubic meter of space. Most of the universe is fantastically empty; places that are dense, such as the human body, are tremendously rare.

Even our most familiar solids are mostly space. If we could take apart such a solid, piece by piece, we would eventually reach the tiny molecules from which it is formed. Molecules are the smallest particles into which any matter can be divided while still retaining its chemical properties. A molecule of water (H_2O), for example, consists of two hydrogen atoms and one oxygen atom bonded together.

Molecules, in turn, are built of atoms, which are the smallest particles of an element that can still be identified as that element. For example, an atom of gold is the smallest possible piece of gold. Nearly 100 different kinds of atoms (elements) exist in nature. Most of them are rare, and only a handful account for more than 99% of everything with which we come in contact. The most abundant elements in the cosmos today are listed in Table 3.1; think of this table as the “greatest hits” of the universe when it comes to elements.

Table 3.1. The Cosmically Abundant Elements

Element[1]	Symbol	Number of Atoms per Million Hydrogen Atoms
Hydrogen	H	1,000,000
Helium	He	80,000
Carbon	C	450
Nitrogen	N	92
Oxygen	O	740
Neon	Ne	130
Magnesium	Mg	40
Silicon	Si	37
Sulfur	S	19
Iron	Fe	32

All atoms consist of a central, positively charged nucleus surrounded by negatively charged electrons. The bulk of the matter in each atom is found in the nucleus, which consists of positive protons and electrically neutral neutrons all bound tightly together in a very small space. Each element is defined by the number of protons in its atoms. Thus, any atom with 6 protons in its nucleus is called *carbon*, any with 50 protons is called *tin*, and any with 70 protons is called *ytterbium*.

The distance from an atomic nucleus to its electrons is typically 100,000 times the size of the nucleus itself. This is why we say that even solid matter is mostly space. The typical atom is far emptier than the solar system out to Neptune. (The distance from Earth to the Sun, for example, is only 100 times the size of the Sun.) This is one reason atoms are not like miniature solar systems.

Remarkably, physicists have discovered that everything that happens in the universe, from the smallest atomic nucleus to the largest superclusters of galaxies, can be explained through the action of only four forces: gravity, electromagnetism (which combines the actions of electricity and magnetism), and two forces that act at the nuclear level. The fact that there are four forces (and not a million, or just one) has puzzled physicists and astronomers for many years and has led to a quest for a unified picture of nature.

Attribution

"1.8 The Universe of the Very Small" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

---

1. This list of elements is arranged in order of the atomic number, which is the number of protons in each nucleus.

3.4 A CONCLUSION AND A BEGINNING

If you are new to astronomy, you have probably reached the end of our brief tour in this chapter with mixed emotions. On the one hand, you may be fascinated by some of the new ideas you've read about and you may be eager to learn more. On the other hand, you may be feeling a bit overwhelmed by the number of topics we have covered, and the number of new words and ideas we have introduced. Learning astronomy is a little like learning a new language: at first it seems there are so many new expressions that you'll never master them all, but with practice, you soon develop facility with them.

At this point you may also feel a bit small and insignificant, dwarfed by the cosmic scales of distance and time. But, there is another way to look at what you have learned from our first glimpses of the cosmos. Let us consider the history of the universe from the Big Bang to today and compress it, for easy reference, into a single year. (We have borrowed this idea from Carl Sagan's 1997 Pulitzer Prize-winning book, *The Dragons of Eden*.)

On this scale, the Big Bang happened at the first moment of January 1, and this moment, when you are reading this chapter would be the end of the very last second of December 31. When did other events in the development of the universe happen in this "cosmic year?" Our solar system formed around September 2, and the first life appeared on Earth on September 21 as shown in Table 3.2.

Table 3.2. On a cosmic calendar, where the time since the Big Bang is compressed into 1 year, creatures we would call human do not emerge on the scene until the evening of December 31 (Data Source: Wikipedia).

Date	Event
January 1	Big Bang Occurs
January 26	First Galaxies Form
March 16	Milky Way Galaxy Formed
May 13	Milky Way Galaxy Disk Formed
September 2	Formation of the Solar System
September 21	First Life on Earth
October 29	Oxygenation of Earth's Atmosphere
December 5	First Multicellular Life
December 19	Fish and Proto-amphibians Appear
December 20	Land Plants Appear
December 25	Dinosaurs Appear
December 26	Mammals Appear
December 27	Birds (Avian Dinosaurs)
December 30	Dinosaurs Become Extinct
December 31, 10:24PM	Humans Appear

Where does the origin of human beings fall during the course of this cosmic year? The answer turns out to be the evening of December 31. The invention of the alphabet doesn't occur until the fiftieth second of 11:59 p.m. on December 31. And the beginnings of modern astronomy are a mere fraction of a second before the New Year. Seen in a cosmic context, the amount of time we have had to study the stars is minute, and our success in piecing together as much of the story as we have is remarkable.

Certainly our attempts to understand the universe are not complete. As new technologies and new ideas allow us to gather more and better data about the cosmos, our present picture of astronomy will very likely undergo many changes. Still, as you read our current progress report on the exploration of the universe, take a few minutes every once in a while just to savor how much you have already learned.

Attribution

3.5 KEY TERMS

Astronomical unit (AU): the average Earth-Sun distance. 3.1

Constellation: one of the 88 sections into which astronomers divide the sky, each named after a prominent star pattern within it. 3.2

Interstellar dust: dust that, together with a sparse distribution of gas, collects into enormous clouds in many places in the Galaxy, becoming the raw material for future generations of stars. 3.1

Milky Way Galaxy: a single collection of stars which contains the Sun, the stars visible to the unaided eye from Earth, as well as hundreds of billions of other stars. 3.1

Moon: our nearest astronomical neighbour which is also Earth's satellite. 3.1

Planet: a body of significant size that orbits a star and does not produce its own light. 3.1

Quasars: brilliant centres of galaxies, glowing with the light of an extraordinarily energetic process. 3.2

Star: a large body that consistently produces its own light. 3.1

Superclusters: when clusters of galaxies themselves form into larger groups. 3.2

CHAPTER 4: IMAGING THE UNIVERSE

Chapter Overview

4.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify and define the three fundamental components of a modern system for measuring radiation from astronomical sources: telescope, wavelength-sorting instrument, and detector.
- Summarize the historical development of astronomical telescopes, including their ancient observatory origins and the impact of Galileo's telescope on our understanding of celestial objects.
- Calculate the light-collecting area of telescopes with different mirror diametres and demonstrate the concept of increasing light collection with larger apertures.
- Identify the technological advancements that enabled the construction of larger telescopes in the decades starting from 1990.
- Assess the effectiveness of adaptive optics in correcting atmospheric distortions and improving image quality in ground-based telescopes.
- Summarize the importance of using electronic equipment to detect radio waves from space instead of relying on human senses.

4.1 INTRODUCTION

There are three basic components of a modern system for measuring radiation from astronomical sources. First, there is a telescope, which serves as a “bucket” for collecting visible light (or radiation at other wavelengths, as shown in Figure 4.1. Just as you can catch more rain with a garbage can than with a coffee cup, large telescopes gather much more light than your eye can. Second, there is an instrument attached to the telescope that sorts the incoming radiation by wavelength. Sometimes the sorting is fairly crude. For example, we might simply want to separate blue light from red light so that we can determine the temperature of a star. But at other times, we want to see individual spectral lines to determine what an object is made of, or to measure its speed . Third, we need some type of detector, a device that senses the radiation in the wavelength regions we have chosen and permanently records the observations.

Orion Region at Different Wavelengths

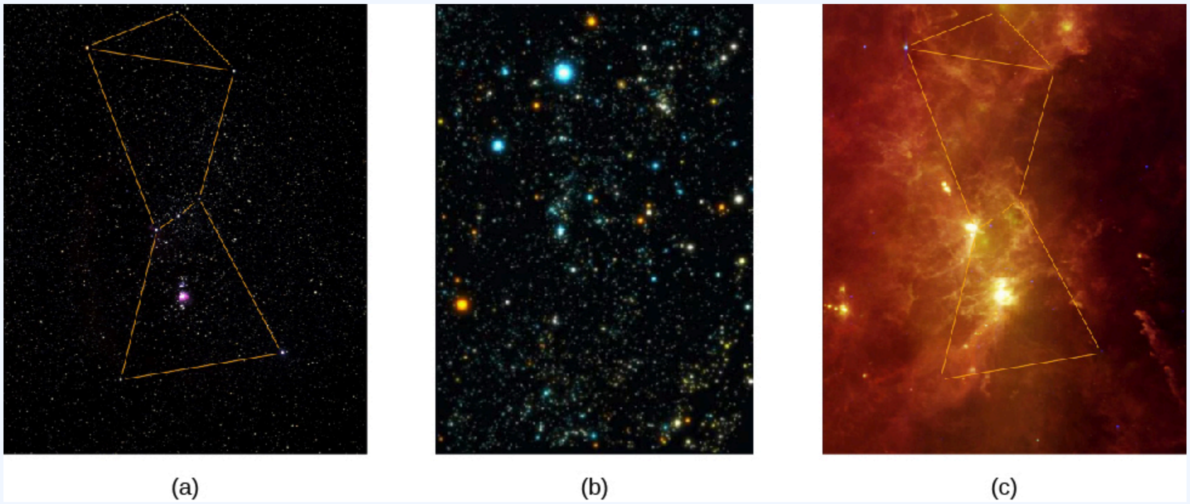


Figure 4.1. The same part of the sky looks different when observed with instruments that are sensitive to different bands of the spectrum. (a) Visible light: this shows part of the Orion region as the human eye sees it, with dotted lines added to show the figure of the mythical hunter, Orion. (b) X-rays: here, the view emphasizes the point-like X-ray sources nearby. The colours are artificial, changing from yellow to white to blue with increasing energy of the X-rays. The bright, hot stars in Orion are still seen in this image, but so are many other objects located at very different distances, including other stars, star corpses, and galaxies at the edge of the observable universe. (c) Infrared radiation: here, we mainly see the glowing dust in this region. (credit a: modification of work by Howard McCallon/NASA/IRAS; credit b: modification of work by Howard McCallon/NASA/IRAS; credit c: modification of work by Michael F. Corcoran; CC-BY-4.0)

Attribution

4.2 HISTORY OF TELESCOPES

The history of the development of astronomical telescopes is about how new technologies have been applied to improve the efficiency of these three basic components: the telescopes, the wavelength-sorting device, and the detectors. Let's first look at the development of the telescope.

Many ancient cultures built special sites for observing the sky as shown in Figure 4.2. At these **ancient *observatories***, they could measure the positions of celestial objects, mostly to keep track of time and date. Many of these ancient observatories had religious and ritual functions as well. The eye was the only device available to gather light, all of the colours in the light were observed at once, and the only permanent record of the observations was made by human beings writing down or sketching what they saw.

Two Pre-Telescopic Observatories

Figure 4.2. (a) Machu Picchu is a fifteenth century Incan site located in Peru. (b) Stonehenge, a prehistoric site (3000–2000 BCE), is located in England.
(a): Peru Machu Picchu Sunrise 2 by Allard Schmidt, CC0 1.0.
(b): Stonehenge Closeup by Daveahern, CC0 1.0.

While Hans Lippershey, Zaccharias Janssen, and Jacob Metius are all credited with the invention of the telescope around 1608—applying for patents within weeks of each other—it was Galileo who, in 1610, used this simple tube with lenses (which he called a spyglass) to observe the sky and gather more light than his eyes alone could. Even his small telescope—used over many nights—revolutionized ideas about the nature of the planets and the position of Earth.

Attribution

4.3 HOW TELESCOPES WORK

Telescopes have come a long way since Galileo's time. Now they tend to be huge devices; the most expensive cost hundreds of millions to billions of dollars. (To provide some reference point, however, keep in mind that just renovating college football stadiums typically costs hundreds of millions of dollars—with the most expensive recent renovation, at Texas A&M University's Kyle Field, costing $450 million.) The reason astronomers keep building bigger and bigger telescopes is that celestial objects—such as planets, stars, and galaxies—send much more light to Earth than any human eye (with its tiny opening) can catch, and bigger telescopes can detect fainter objects. If you have ever watched the stars with a group of friends, you know that there's plenty of starlight to go around; each of you can see each of the stars. If a thousand more people were watching, each of them would also catch a bit of each star's light. Yet, as far as you are concerned, the light not shining into your eye is wasted. It would be great if some of this "wasted" light could also be captured and brought to your eye. This is precisely what a telescope does.

The most important functions of a telescope are (1) to *collect* the faint light from an astronomical source and (2) to *focus* all the light into a point or an image. Most objects of interest to astronomers are extremely faint: the more light we can collect, the better we can study such objects. (And remember, even though we are focusing on visible light first, there are many telescopes that collect other kinds of electromagnetic radiation.)

Telescopes that collect visible radiation use a lens or mirror to gather the light. Other types of telescopes may use collecting devices that look very different from the lenses and mirrors with which we are familiar, but they serve the same function. In all types of telescopes, the light-gathering ability is determined by the area of the device acting as the light-gathering "bucket." Since most telescopes have mirrors or lenses, we can compare their light-gathering power by comparing the apertures, or diametres, of the opening through which light travels or reflects.

The amount of light a telescope can collect increases with the size of the aperture. A telescope with a mirror that is 4 metres in diameter can collect 16 times as much light as a telescope that is 1 meter in diameter. (The diameter is squared because the area of a circle equals $\pi d^2/4$, where d is the diameter of the circle.)

Example 4.1

Calculating the Light-Collecting Area

What is the area of a 1 m diameter telescope? A 4 m diameter one?

Solution

Using the equation for the area of a circle,

$$A = \frac{\pi d^2}{4}$$

the area of a 1 m telescope is

$$\frac{\pi d^2}{4} = \frac{\pi(1\,\mathrm{m})^2}{4} = 0.79\,\mathrm{m}^2$$

and the area of a 4 m telescope is

$$\frac{\pi d^2}{4} = \frac{\pi(4\,\mathrm{m})^2}{4} = 12.6\,\mathrm{m}^2$$

Exercise 4.1

Show that the ratio of the two areas is $16 : 1$.

Solution

$\frac{12.6\,\mathrm{m}^2}{0.79\,\mathrm{m}^2} = 16$. Therefore, with 16 times the area, a 4 m telescope collects 16 times the light of a 1 m telescope.

After the telescope forms an image, we need some way to detect and record it so that we can measure, reproduce, and analyze the image in various ways. Before the nineteenth century, astronomers simply viewed images with their eyes and wrote descriptions of what they saw. This was very inefficient and did not lead to a very reliable long-term record; you know from crime shows on television that eyewitness accounts are often inaccurate.

In the nineteenth century, the use of photography became widespread. In those days, photographs were a chemical record of an image on a specially treated glass plate. Today, the image is generally detected with sensors similar to those in digital cameras, recorded electronically, and stored in computers. This permanent record can then be used for detailed and quantitative studies. Professional astronomers rarely look through the large telescopes that they use for their research.

Attribution

"6.1 Telescopes" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

4.4 REFRACTION VS. REFLECTION

Whether or not you wear glasses, you see the world through lenses; they are key elements of your eyes. A lens is a transparent piece of material that bends the rays of light passing through it. If the light rays are parallel as they enter, the lens brings them together in one place to form an image as shown in Figure 4.3. If the curvatures of the lens' surfaces are just right, all parallel rays of light (say, from a star) are bent, or *refracted*, in such a way that they converge toward a point, called the **focus** of the lens. At the focus, an image of the light source appears. In the case of parallel light rays, the distance from the lens to the location where the light rays focus, or image, behind the lens is called the **focal length** of the lens.

Formation of an Image by a Simple Lens

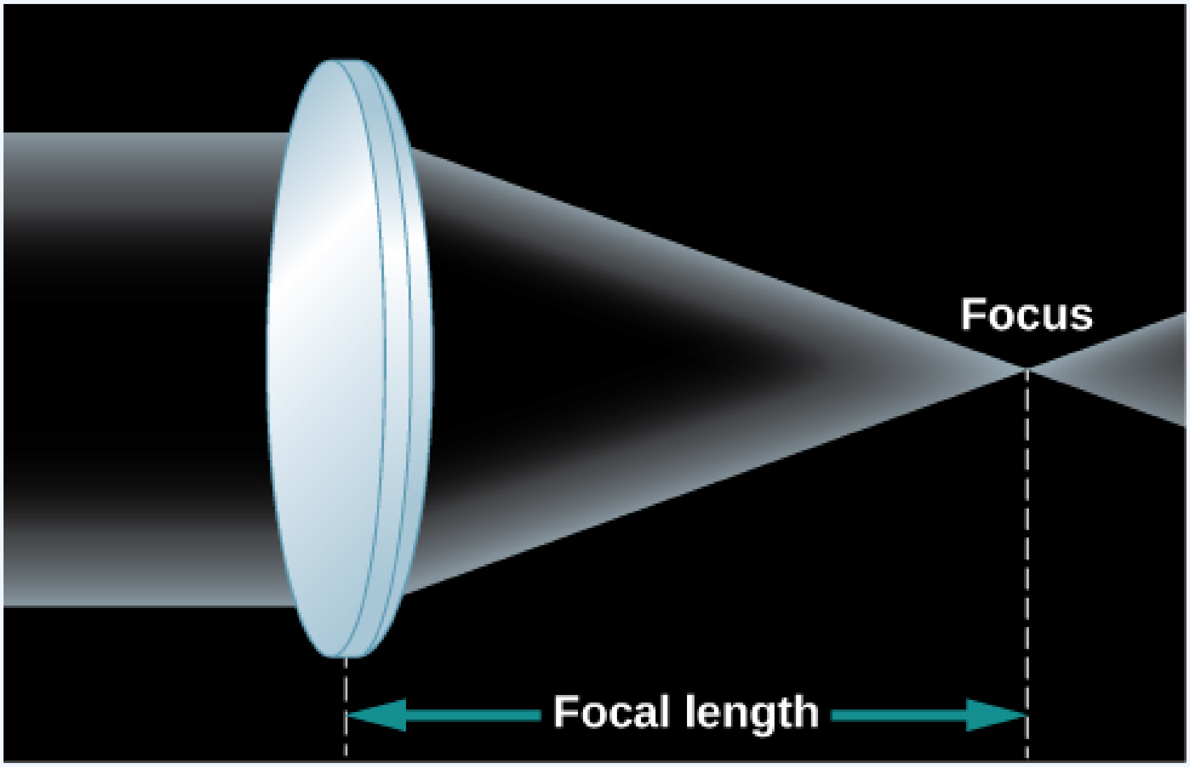


Figure 4.3. Parallel rays from a distant source are bent by the convex lens so that they all come together in a single place (the focus) to form an image. Image by Open Stax, CC BY 4.0

As you look at Figure 4.3, you may ask why two rays of light from the same star would be parallel to each other. After all, if you draw a picture of star shining in all directions, the rays of light coming from the star don't look parallel at all. But remember that the stars (and other astronomical objects) are all extremely far away. By the time the few rays of light pointed toward us actually arrive at Earth, they are, for all practical purposes, parallel to each other. Put another way, any rays that were *not* parallel to the ones pointed at Earth are now heading in some very different direction in the universe.

To view the image formed by the lens in a telescope, we use an additional lens called an eyepiece. The eyepiece focuses the image at a distance that is either directly viewable by a human or at a convenient place for a detector. Using different eyepieces, we can change the **magnification** (or size) of the image and also redirect the light to a more accessible location. Stars look like points of light, and magnifying them makes little difference, but the image of a planet or a galaxy, which has structure, can often benefit from being magnified.

Many people, when thinking of a telescope, picture a long tube with a large glass lens at one end. This design, which uses a lens as its main optical element to form an image, as we have been discussing, is known as a **refractor**, as shown in Figure 4.4, and a telescope based on this design is called a **refracting telescope**. Galileo's telescopes were refractors, as are today's binoculars and field glasses. However, there is a limit to the size of a refracting telescope. The largest one ever built was a 49-inch refractor built for the Paris 1900 Exposition, and it was dismantled after the Exposition. Currently, the largest refracting telescope is the 40-inch refractor at Yerkes Observatory in Wisconsin.

Refracting and Reflecting Telescopes

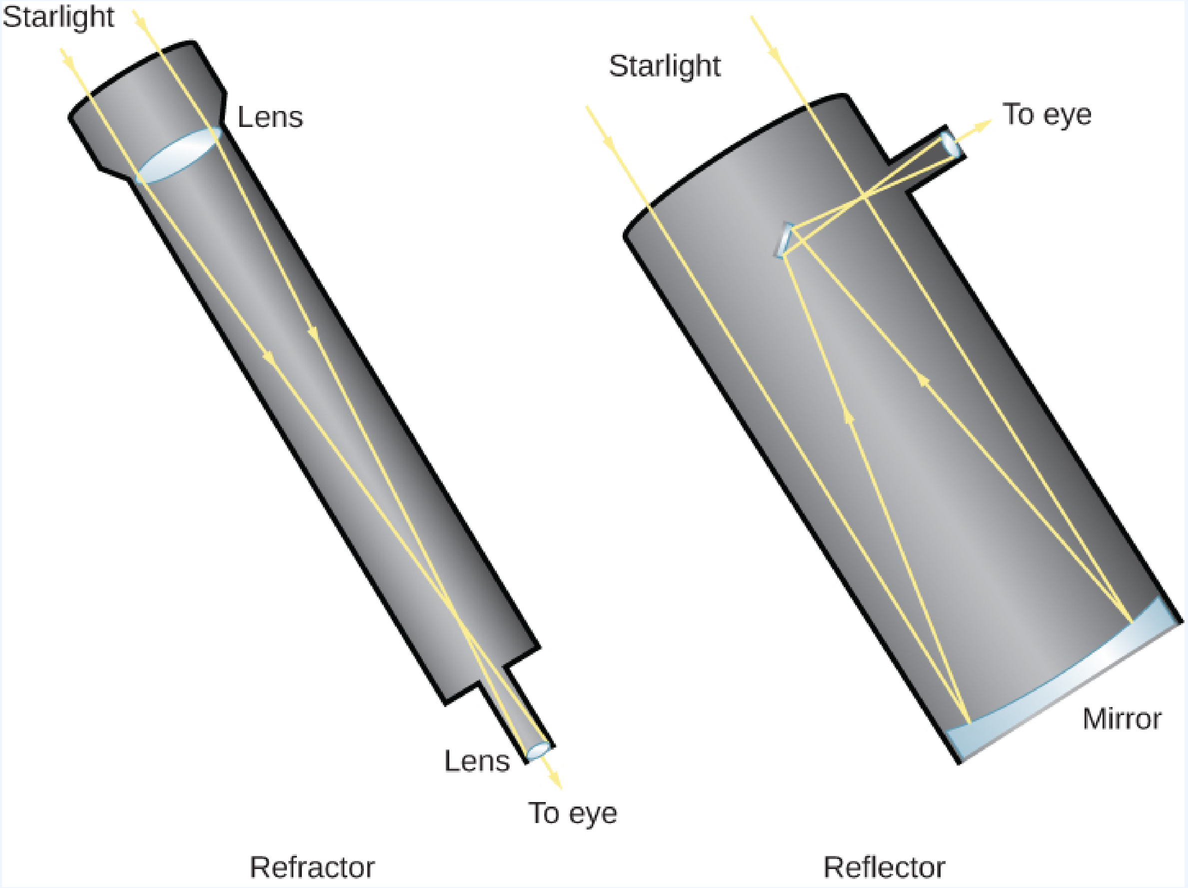


Figure 4.4. Light enters a refracting telescope through a lens at the upper end, which focuses the light near the bottom of the telescope. An eyepiece then magnifies the image so that it can be viewed by the eye, or a detector like a photographic plate can be placed at the focus. The upper end of a reflecting telescope is open, and the light passes through to the mirror located at the bottom of the telescope. The mirror then focuses the light at the top end, where it can be detected. Alternatively, as in this sketch, a second mirror may reflect the light to a position outside the telescope structure, where an observer can have easier access to it. Professional astronomers' telescopes are more complicated than this, but they follow the same principles of reflection and refraction. Image by Open Stax, CC BY 4.0

One problem with a refracting telescope is that the light must pass *through* the lens of a refractor. That means the glass must be perfect all the way through, and it has proven very difficult to make large pieces of glass without flaws and bubbles in them. Also, optical properties of transparent materials change a little bit with the wavelengths (or colours) of light, so there is some additional distortion, known as chromatic aberration. Each wavelength focuses at a slightly different spot, causing the image to appear blurry.

In addition, since the light must pass through the lens, the lens can only be supported around its edges (just like the frames of our eyeglasses). The force of gravity will cause a large lens to sag and distort the path of the light rays as they pass through it. Finally, because the light passes through it, both sides of the lens must be manufactured to precisely the right shape in order to produce a sharp image.

A different type of telescope uses a concave **primary mirror** as its main optical element. The mirror is curved like the inner surface of a sphere, and it reflects light in order to form an image as shown in Figure 4.4. Telescope mirrors are coated with a shiny metal, usually silver, aluminum, or, occasionally, gold, to make them highly reflective. If the mirror has the correct shape, all parallel rays are reflected back to the same point, the focus of the mirror. Thus, images are produced by a mirror exactly as they are by a lens.

Telescopes designed with mirrors avoid the problems of refracting telescopes. Because the light is reflected from the front surface only, flaws and bubbles within the glass do not affect the path of the light. In a telescope designed with mirrors, only the front surface has to be manufactured to a precise shape, and the mirror can be supported from the back. For these reasons, most astronomical telescopes today (both amateur and professional) use a mirror rather than a lens to form an image; this type of telescope is called a reflecting telescope. The first successful reflecting telescope was built by Isaac Newton in 1668.

In a reflecting telescope, the concave mirror is placed at the bottom of a tube or open framework. The mirror reflects the light back up the tube to form an image near the front end at a location called the **prime focus**. The image can be observed at the prime focus, or additional mirrors can intercept the light and redirect it to a position where the observer can view it more easily as shown in Figure 4.5. Since an astronomer at the prime focus can block much of the light coming to the main mirror, the use of a small **secondary mirror** allows more light to get through the system.

Focus Arrangements for Reflecting Telescopes

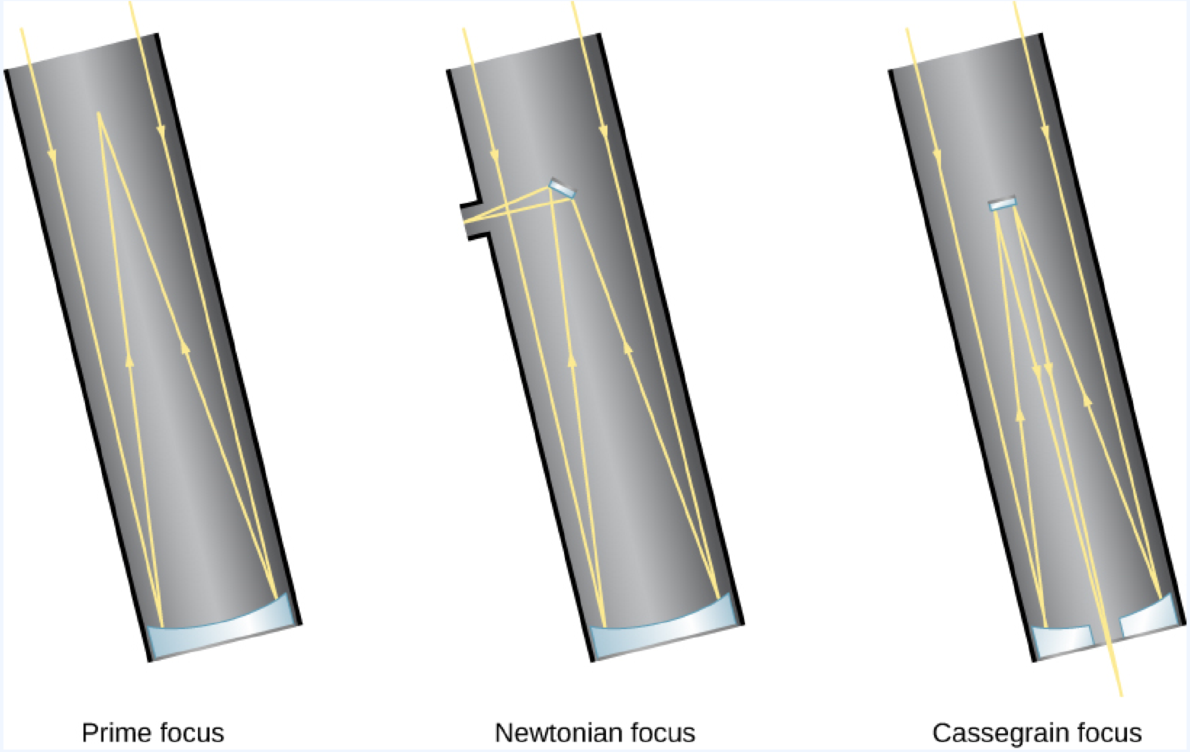


Figure 4.5. Reflecting telescopes have different options for where the light is brought to a focus. With prime focus, light is detected where it comes to a focus after reflecting from the primary mirror. With Newtonian focus, light is reflected by a small secondary mirror off to one side, where it can be detected (see also [link]). Most large professional telescopes have a Cassegrain focus in which light is reflected by the secondary mirror down through a hole in the primary mirror to an observing station below the telescope. Image by Open Stax, CC BY 4.0

Choosing Your Own Telescope

If the astronomy course you are taking whets your appetite for exploring the sky further, you may be thinking about buying your own telescope. Many excellent amateur telescopes are available, and some research is required to find the best model for your needs. Some good sources of information about personal telescopes are the two popular US magazines aimed at amateur astronomers: *Sky & Telescope* and *Astronomy*. Both carry regular articles with advice, reviews, and advertisements from reputable telescope dealers.

Some of the factors that determine which telescope is right for you depend upon your preferences:

- Will you be setting up the telescope in one place and leaving it there, or do you want an instrument that is portable and can come with you on outdoor excursions? How portable should it be, in terms of size and weight?
- Do you want to observe the sky with your eyes only, or do you want to take photographs? (Long-exposure photography, for example, requires a good clock drive to turn your telescope to compensate for Earth's rotation.)
- What types of objects will you be observing? Are you interested primarily in comets, planets, star clusters, or galaxies, or do you want to observe all kinds of celestial sights?

You may not know the answers to some of these questions yet. For this reason, you may want to "test-drive" some telescopes first. Most communities have amateur astronomy clubs that sponsor star parties open to the public. The members of those clubs often know a lot about telescopes and can share their ideas with you. Your instructor may know where the nearest amateur astronomy club meets; or, to find a club near you, use a good search engine. In Canada, the Royal Astronomical Society of Canada keeps an up to date list. The Vancouver chapter is Jennifer Kirkey's favourite. https://www.rasc.ca/

Furthermore, you may already have an instrument like a telescope at home (or have access to one through a relative or friend). Many amateur astronomers recommend starting your survey of the sky with a good pair of binoculars. These are easily carried around and can show you many objects not visible (or clear) to the unaided eye.

When you are ready to purchase a telescope, you might find the following ideas useful:

- The key characteristic of a telescope is the aperture of the main mirror or lens; when someone says they have a 6-inch or 8-inch telescope, they mean the diameter of the collecting surface. The larger the aperture, the more light you can gather, and the fainter the objects you can see or photograph.
- Telescopes of a given aperture that use lenses (refractors) are typically more expensive than those using mirrors (reflectors) because both sides of a lens must be polished to great accuracy. And, because the light passes through it, the lens must be made of high-

quality glass throughout. In contrast, only the front surface of a mirror must be accurately polished.

- Magnification is not one of the criteria on which to base your choice of a telescope. As we discussed, the magnification of the image is done by a smaller eyepiece, so the magnification can be adjusted by changing eyepieces. However, a telescope will magnify not only the astronomical object you are viewing but also the turbulence of Earth's atmosphere. If the magnification is too high, your image will shimmer and shake and be difficult to view. A good telescope will come with a variety of eyepieces that stay within the range of useful magnification.
- The mount of a telescope (the structure on which it rests) is one of its most critical elements. Because a telescope shows a tiny field of view, which is magnified significantly, even the smallest vibration or jarring of the telescope can move the object you are viewing around or out of your field of view. A sturdy and stable mount is essential for serious viewing or photography (although it clearly affects how portable your telescope can be).
- A telescope requires some practice to set up and use effectively. Don't expect everything to go perfectly on your first try. Take some time to read the instructions. If a local amateur astronomy club is nearby, use it as a resource.

Attribution

4.5 MODERN OBSERVATORIES

Since Newton's time, when the sizes of the mirrors in telescopes were measured in inches, reflecting telescopes have grown ever larger. In 1948, US astronomers built a telescope with a 5-meter (200-inch) diameter mirror on Palomar Mountain in Southern California. It remained the largest visible-light telescope in the world for several decades. The giants of today, however, have primary mirrors (the largest mirrors in the telescope) that are 8- to 10-metres in diameter, and larger ones are being built as shown in Figure 4.6.

Large Telescope Mirror

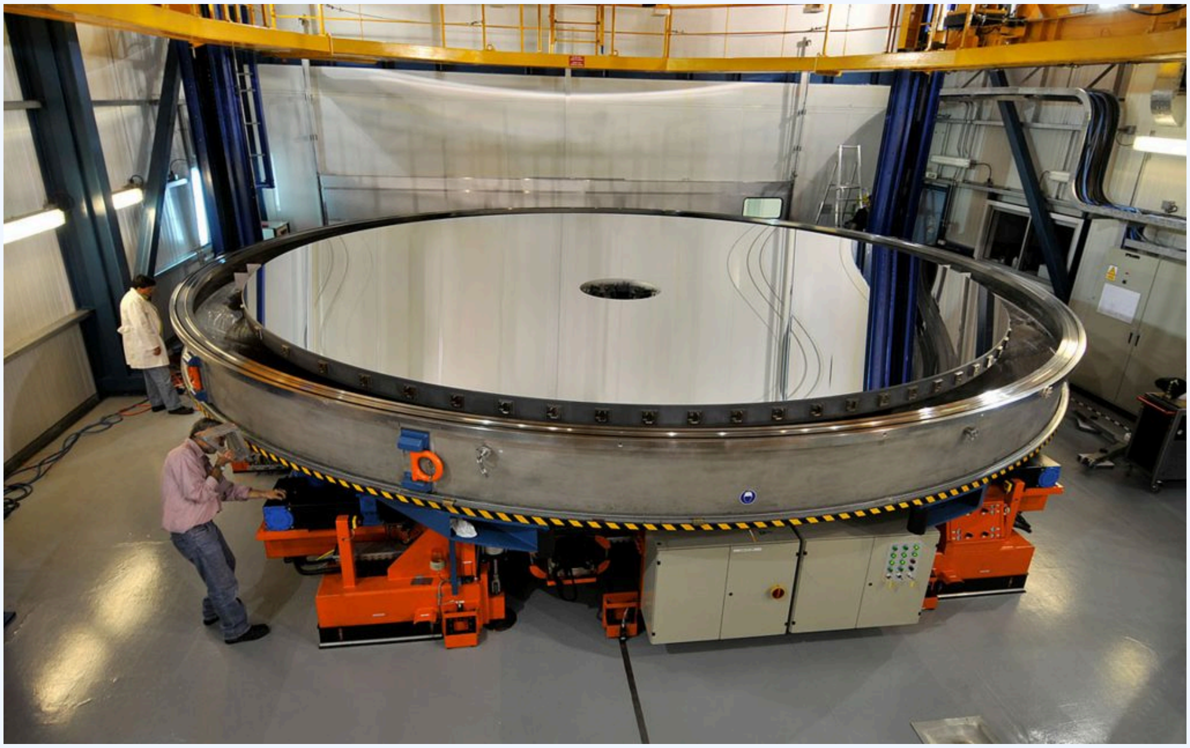


Figure 4.6. This image shows one of the primary mirrors of the European Southern Observatory's Very Large Telescope, named Yepun, just after it was recoated with aluminum. The mirror is a little over 8 metres in diameter.
Recoating Yepun's mirror by ESO/G. Huedepohl, CC BY 4.0.

Modern Visible-Light and Infrared Telescopes

The decades starting in 1990 saw telescope building around the globe grow at an unprecedented rate. (See Table 4.1, which also includes websites for each telescope in case you want to visit or learn more about them.) Technological advancements had finally made it possible to build telescopes significantly larger than the 5-meter telescope at Palomar at a reasonable cost. New technologies have also been designed to work well in the infrared, and not just visible, wavelengths.

Table 4.1. Large Single-Dish Visible-Light and Infrared Telescopes

Aperture (m)	Telescope Name	Location	Status	Website
39	European Extremely Large Telescope (E-ELT)	Cerro Armazonas, Chile	First light 2025 (estimated)	www.eso.org/sci/facilities/eelt
30	Thirty-Meter Telescope (TMT)	Mauna Kea, HI	First light 2025 (estimated)	www.tmt.org
24.5	Giant Magellan Telescope (GMT)	Las Campanas Observatory, Chile	First light 2025 (estimated)	www.gmto.org
11.1 × 9.9	Southern African Large Telescope (SALT)	Sutherland, South Africa	2005	www.salt.ac.za
10.4	Gran Telescopio Canarias (GTC)	La Palma, Canary Islands	First light 2007	http://www.gtc.iac.es
10.0	Keck I and II (two telescopes)	Mauna Kea, HI	Completed 1993–96	www.keckobservatory.org
9.1	Hobby–Eberly Telescope (HET)	Mount Locke, TX	Completed 1997	www.as.utexas.edu/mcdonald/het
8.4	Large Binocular Telescope (LBT) (two telescopes)	Mount Graham, AZ	First light 2004	www.lbto.org
8.4	Large Synoptic Survey Telescope (LSST)	The Cerro Pachón, Chile	First light 2021	www.lsst.org

Aperture (m)	**Telescope Name**	**Location**	**Status**	**Website**
8.3	Subaru Telescope	Mauna Kea, HI	First light 1998	www.naoj.org
8.2	Very Large Telescope (VLT)	Cerro Paranal, Chile	All four telescopes completed 2000	www.eso.org/public/teles-instr/paranal
8.1	Gemini North and Gemini South	Mauna Kea, HI (North) and Cerro Pachón, Chile (South)	First light 1999 (North), First light 2000 (South)	www.gemini.edu
6.5	Magellan Telescopes (two telescopes: Baade and Landon Clay)	Las Campanas, Chile	First light 2000 and 2002	obs.carnegiescience.edu/Magellan
6.5	Multi-Mirror Telescope (MMT)	Mount Hopkins, AZ	Completed 1979	www.mmto.org
6.0	Big Telescope Altazimuth (BTA-6)	Mount Pastukhov, Russia	Completed 1976	w0.sao.ru/Doc-en/Telescopes/bta/descrip.html
5.1	Hale Telescope	Mount Palomar, CA	Completed 1948	www.astro.caltech.edu/palomar/about/telescopes/hale.html

The differences between the Palomar telescope and the modern Gemini North telescope (to take an example) are easily seen in Figure 4.7. The Palomar telescope is a massive steel structure designed to hold the 14.5-ton primary mirror with a 5-meter diameter. Glass tends to sag under its own weight; hence, a huge steel structure is needed to hold the mirror. A mirror 8 metres in diameter, the size of the Gemini North telescope, if it were built using the same technology as the Palomar telescope, would have to weigh at least eight times as much and would require an enormous steel structure to support it.

Modern Reflecting Telescopes – Hale and Gemini

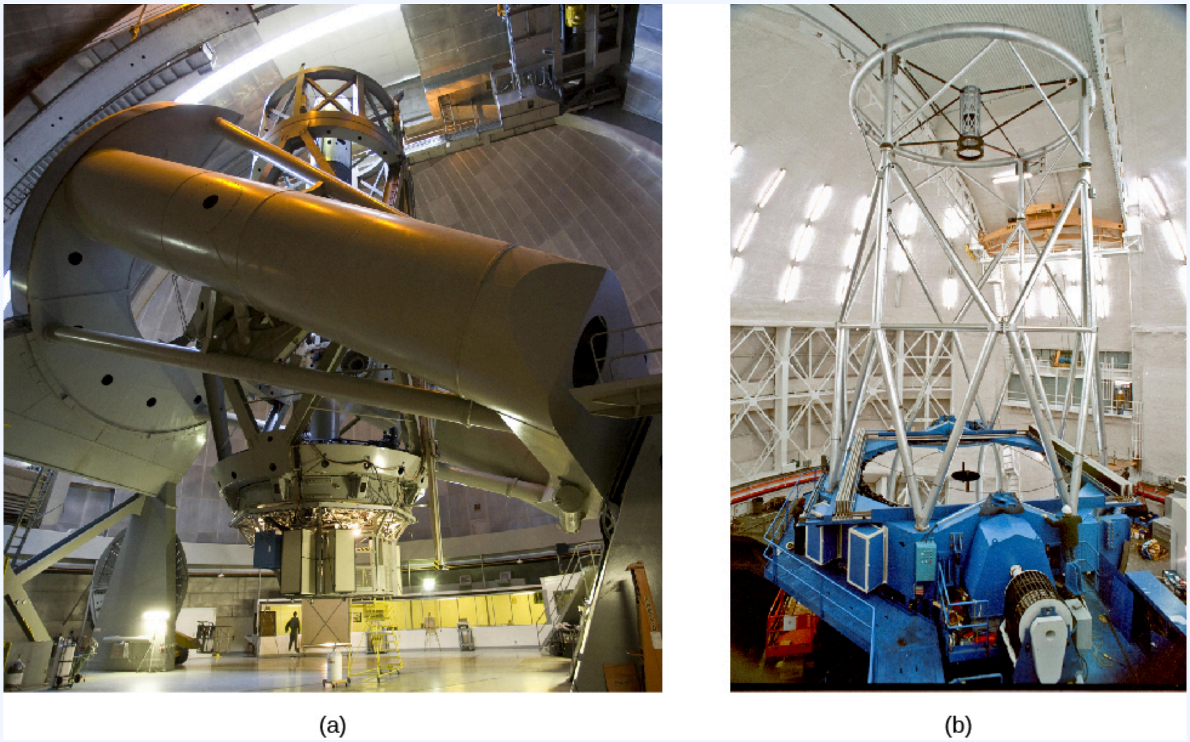


Figure 4.7. (a) The Palomar 5-meter reflector: The Hale telescope on Palomar Mountain has a complex mounting structure that enables the telescope (in the open "tube" pointing upward in this photo) to swing easily into any position. (b) The Gemini North 8-meter telescope: The Gemini North mirror has a larger area than the Palomar mirror, but note how much less massive the whole instrument seems.
(a): Hale Telescope, Palomar Observatory by NASA/JPL, NASA Media Licence.
(b:) Gemini North Telescope by Gemini Observatory/AURA, CC BY 4.0.

The 8-meter Gemini North telescope looks like a featherweight by contrast, and indeed it is. The mirror is only about 8 inches thick and weighs 24.5 tons, less than twice as much as the Palomar mirror. The Gemini North telescope was completed about 50 years after the Palomar telescope. Engineers took advantage of new technologies to build a telescope that is much lighter in weight relative to the size of the primary mirror. The Gemini mirror does sag, but with modern computers, it is possible to measure that sag many times each second and apply forces at 120 different locations to the back of the mirror to correct the sag, a process called **active control**. Seventeen telescopes with mirrors 6.5 metres in diameter and larger have been constructed since 1990.

The twin 10-meter Keck telescopes on Mauna Kea, which were the first of these new-technology instruments, use precision control in an entirely novel way. Instead of a single primary mirror 10 metres in diameter, each Keck telescope achieves its larger aperture by combining the light from 36 separate hexagonal mirrors, each 1.8 metres wide as shown in Figure 4.8. Computer-controlled actuators (motors) constantly adjust these 36 mirrors so that the overall reflecting surface acts like a single mirror with just the right shape to collect and focus the light into a sharp image.

Thirty-Six Eyes Are Better Than One

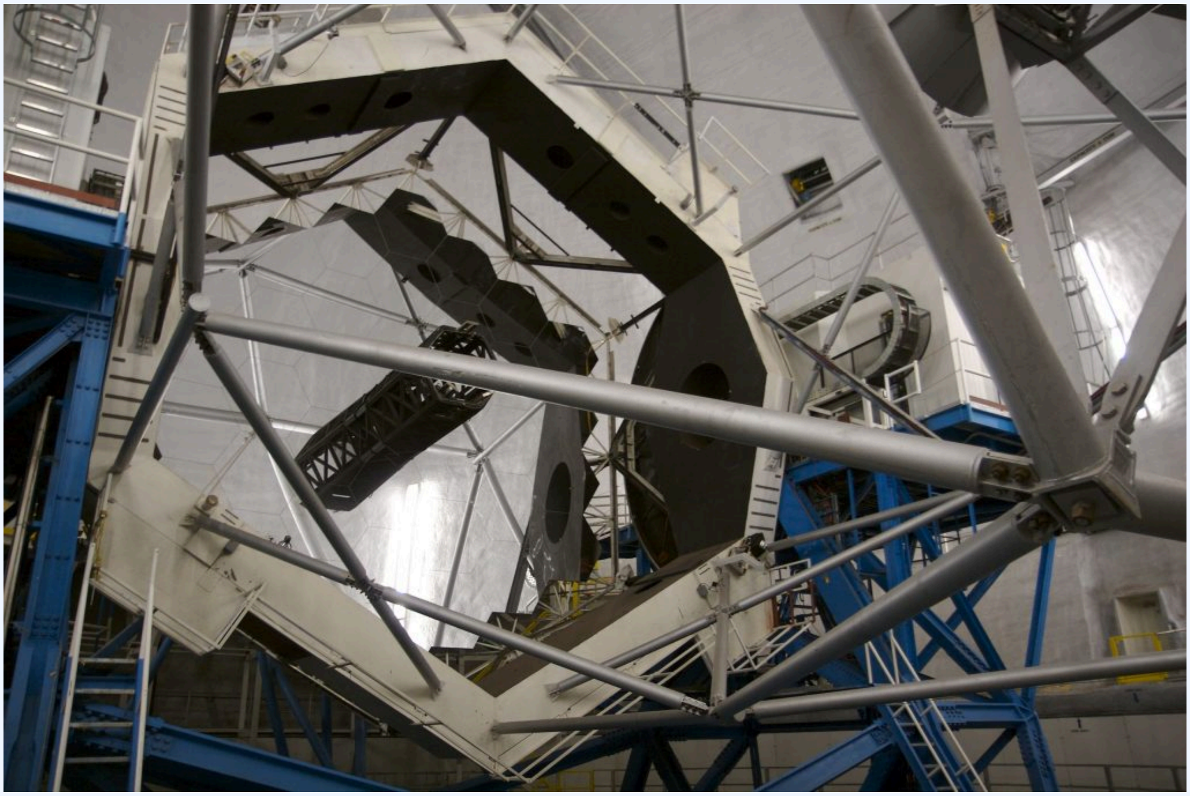


Figure 4.8. The mirror of the 10-meter Keck telescope is composed of 36 hexagonal sections. Primary Mirror of Keck Telescope by z2amiller, CC BY-SA 2.0.

Learn more about the Keck Observatory on Mauna Kea through this link.

In addition to holding the mirror, the steel structure of a telescope is designed so that the entire telescope can be pointed quickly toward any object in the sky. Since Earth is rotating, the telescope must have a motorized drive system that moves it very smoothly from east to west at exactly the same rate that Earth is rotating from west to east, so it can continue to point at the object being observed. All this machinery must be housed in a dome to protect the telescope from the elements. The dome has an opening in it that can be positioned in front of the telescope and moved along with it, so that the light from the objects being observed is not blocked.

George Ellery Hale: Master Telescope Builder

George Ellery Hale, shown in Figure 4.9, was a giant among early telescope builders. Not once, but four times, he initiated projects that led to the construction of what was the world's largest telescope at the time. And he was a master at winning over wealthy benefactors to underwrite the construction of these new instruments.

George Ellery Hale (1868–1938)

Figure 4.9. Hale's work led to the construction of several major telescopes, including the 40-inch refracting telescope at Yerkes Observatory, and three reflecting telescopes: the 60-inch Hale and 100-inch Hooker telescopes at Mount Wilson Observatory, and the 200-inch Hale Telescope at Palomar Observatory.

Photo in the Public Domain

Hale's training and early research were in solar physics. In 1892, at age 24, he was named associate professor of astral physics and director of the astronomical observatory at the University of Chicago. At the time, the largest telescope in the world was the 36-inch refractor at the Lick Observatory near San Jose, California. Taking advantage of an existing glass blank for a 40-inch telescope, Hale set out to raise money for a larger telescope than the one at Lick. One prospective donor was Charles T. Yerkes, who, among other things, ran the trolley system in Chicago.

Hale wrote to Yerkes, encouraging him to support the construction of the giant telescope by saying that "the donor could have no more enduring monument. It is certain that Mr. Lick's name would not have been nearly so widely known today were it not for the famous observatory established as a result of his munificence." Yerkes agreed, and the new telescope was completed in May 1897; it remains the largest refractor in the world, it is shown in Figure 4.10.

World's Largest Refractor

Figure 4.10. The Yerkes 40-inch (1-meter) telescope.
Photo by Kb9vrg, CC0 1.0.

Even before the completion of the Yerkes refractor, Hale was not only dreaming of building a still larger telescope but was also taking concrete steps to achieve that goal. In the 1890s, there was a major controversy about the relative quality of refracting and reflecting telescopes. Hale realized that 40 inches was close to the maximum feasible aperture for refracting telescopes. If telescopes with significantly larger apertures were to be built, they would have to be reflecting telescopes.

Using funds borrowed from his own family, Hale set out to construct a 60-inch reflector. For a site, he left the Midwest for the much better conditions on Mount Wilson—at the time, a wilderness peak above the small city of Los Angeles. In 1904, at the age of 36, Hale received funds from the Carnegie Foundation to establish the Mount Wilson Observatory. The 60-inch mirror was placed in its mount in December 1908.

Two years earlier, in 1906, Hale had already approached John D. Hooker, who had made his fortune in hardware and steel pipe, with a proposal to build a 100-inch telescope. The technological risks were substantial. The 60-inch telescope was not yet complete, and the usefulness of large reflectors for astronomy had yet to be demonstrated. George Ellery Hale's brother called him "the greatest gambler in the world." Once again, Hale successfully obtained funds, and the 100-inch telescope was completed in November 1917. (It was with this telescope that Edwin Hubble was able to establish that the spiral nebulae were separate islands of stars—or galaxies—quite removed from our own Milky Way.)

Hale was not through dreaming. In 1926, he wrote an article in *Harper's Magazine* about the scientific value of a still larger telescope. This article came to the attention of the Rockefeller Foundation, which granted $6 million for the construction of a 200-inch telescope. Hale died in 1938, but the 200-inch (5-meter) telescope on Palomar Mountain was dedicated 10 years later and is now named in Hale's honour.

Picking the Best Observing Sites

A telescope like the Gemini or Keck telescope costs about $100 million to build. That kind of investment demands that the telescope be placed in the best possible site. Since the end of the nineteenth century, astronomers have realized that the best observatory sites are on mountains, far from the lights and pollution of cities. Although a number of urban observatories remain, especially in the large cities of Europe, they have become administrative centres or museums. The real action takes place far away, often on desert mountains or isolated peaks in the Atlantic and Pacific Oceans, where we find the staff's living quarters, computers, electronic and machine shops, and of course the telescopes themselves. A large observatory today requires a supporting staff of 20 to 100 people in addition to the astronomers.

The performance of a telescope is determined not only by the size of its mirror but also by its location. Earth's atmosphere, so vital to life, presents challenges for the observational astronomer. In at least four ways, our air imposes limitations on the usefulness of telescopes:

1. The most obvious limitation is weather conditions such as clouds, wind, and rain. At the best sites, the weather is clear as much as 75% of the time.
2. Even on a clear night, the atmosphere filters out a certain amount of starlight, especially in the infrared, where the absorption is due primarily to water vapour. Astronomers therefore prefer dry sites, generally found at high altitudes.

3. The sky above the telescope should be dark. Near cities, the air scatters the glare from lights, producing an illumination that hides the faintest stars and limits the distances that can be probed by telescopes. (Astronomers call this effect **light pollution**.) Observatories are best located at least 100 miles from the nearest large city.
4. Finally, the air is often unsteady; light passing through this turbulent air is disturbed, resulting in blurred star images. Astronomers call these effects "bad seeing." When **seeing** is bad, images of celestial objects are distorted by the constant twisting and bending of light rays by turbulent air.

The best observatory sites are therefore high, dark, and dry. The world's largest telescopes are found in such remote mountain locations as the Andes Mountains of Chile as shown in Figure 4.11, the desert peaks of Arizona, the Canary Islands in the Atlantic Ocean, and Mauna Kea in Hawaii, a dormant volcano with an altitude of 13,700 feet (4200 metres).

Light pollution is a problem not just for professional astronomers but for everyone who wants to enjoy the beauty of the night sky. In addition research is now showing that it can disrupt the life cycle of animals with whom we share the urban and suburban landscape. And the light wasted shining into the sky leads to unnecessary municipal expenses and use of fossil fuels. Concerned people have formed an organization, the International Dark-Sky Association, whose website is full of good information. A citizen science project called *Globe at Night* allows you to measure the light levels in your community by counting stars and to compare it to others around the world.

High and Dry Site

Figure 4.11. Cerro Paranal, a mountain summit 2.7 kilometres above sea level in Chile's Atacama Desert, is the site of the European Southern Observatory's Very Large Telescope. This photograph shows the four 8-meter telescope buildings on the site and vividly illustrates that astronomers prefer high, dry sites for their instruments. The 4.1-meter Visible and Infrared Survey Telescope for Astronomy (VISTA) can be seen in the distance on the next mountain peak.
An aerial view of the Paranal Observatory in Chile by ESO/G.Hüdepohl, CC BY 4.0.

Attribution

"6.2 Telescopes Today" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

4.6 CANADIAN TELESCOPES OF HISTORICAL IMPORTANCE

Dominion Astrophysical Observatory in Victoria, British Columbia, Canada

Since becoming operational in 1918, the 1.8 metre Plaskett has accumulated nearly a century of upgrades that make it 10,000 times more sensitive than when it was first built. The 1.8-m Plaskett Telescope can perform optical imaging as well as spectroscopy. The Observatory has been designated a national historic site because of its important role in establishing Canada's international scientific reputation in astronomy. Over the decades, these telescopes have contributed significantly to our understanding of the rotation, size and mass of the Milky Way, and of the rarefied interstellar medium between the stars. Recent projects include orbital determination of comets and asteroids, spectroscopy of magnetic stars, and ongoing studies of distant quasars and galaxies. A group called Friends of the Observatory is working to restore the glory of this telescope and conducts tours and star parties in the summer (https://centreoftheuniverse.org/).

Dominion Astrophysical Observatory

Figure 4.12. Dominion Astrophysical Observatory in Victoria, British Columbia, Canada
Dominion Astrophysical Observatory by Government of Canada, Public Domain.

David Dunlap Observatory in Richmond Hill, Ontario, Canada

This opened in 1935. The David Dunlap Observatory (DDO) is the largest telescope in Canada. A number of important studies have taken place here, including providing the first direct evidence that Cygnus X-1 was a black hole, pioneering measurements of the distance to globular clusters and the discovery that Polaris was stabilizing. The DDO property was 76.5 hectares (189 acres) bordered by Hillsview Drive to the north, Bayview Avenue to the east, 16th Avenue to the south and the Canadian National Railway Bala Line to the west. The property was the site of a 19th century farmstead owned by Alexander Marsh, comprised of a brick farmhouse, a lane from Yonge Street, agricultural fields with hedgerows and an orchard.

The University of Toronto constructed the Observatory on the site. It included a dome, housing a 74-inch (1.88m) reflector telescope, and an Administration Building, with three smaller telescope domes. When construction was complete in 1935, the main telescope was the second largest in the world and the largest in Canada.

On September 29, 2009, Richmond Hill Council designated the DDO Property as a “property of cultural heritage value or interest” under Part IV, Section 29 of the Ontario Heritage Act.

David Dunlap Observatory

Figure 4.13. The 74 inch telescope at the David Dunlap Observatory in Ontario, Canada.
Dunlap Observatory by John H. Martin, CC BY-SA 3.0.

Attribution

"6.1 Telescopes" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

4.7 THE RESOLUTION OF A TELESCOPE

In addition to gathering as much light as they can, astronomers also want to have the sharpest images possible. **Resolution** refers to the precision of detail present in an image: that is, the smallest features that can be distinguished. Astronomers are always eager to make out more detail in the images they study, whether they are following the weather on Jupiter or trying to peer into the violent heart of a "cannibal galaxy" that recently ate its neighbour for lunch.

One factor that determines how good the resolution will be is the size of the telescope. Larger apertures produce sharper images. Until very recently, however, visible-light and infrared telescopes on Earth's surface could not produce images as sharp as the theory of light said they should.

The problem—as we saw earlier in this chapter—is our planet's atmosphere, which is turbulent. It contains many small-scale blobs or cells of gas that range in size from inches to several feet. Each cell has a slightly different temperature from its neighbour, and each cell acts like a lens, bending (refracting) the path of the light by a small amount. This bending slightly changes the position where each light ray finally reaches the detector in a telescope. The cells of air are in motion, constantly being blown through the light path of the telescope by winds, often in different directions at different altitudes. As a result, the path followed by the light is constantly changing.

For an analogy, think about watching a parade from a window high up in a skyscraper. You decide to throw some confetti down toward the marchers. Even if you drop a handful all at the same time and in the same direction, air currents will toss the pieces around, and they will reach the ground at different places. As we described earlier, we can think of the light from the stars as a series of parallel beams, each making its way through the atmosphere. Each path will be slightly different, and each will reach the detector of the telescope at a slightly different place. The result is a blurred image, and because the cells are being blown by the wind, the nature of the blur will change many times each second. You have probably noticed this effect as the "twinkling" of stars seen from Earth. The light beams are bent enough that part of the time they reach your eye, and part of the time some of them miss, thereby making the star seem to vary in brightness. In space, however, the light of the stars is steady.

Astronomers search the world for locations where the amount of atmospheric blurring, or turbulence, is as small as possible. It turns out that the best sites are in coastal mountain ranges and on isolated volcanic peaks in the middle of an ocean. Air that has flowed long distances over water before it encounters land is especially stable.

The resolution of an image is measured in units of angle on the sky, typically in units of arcseconds. One arcsecond is 1/3600 degree, and there are 360 degrees in a full circle. So we are talking about tiny angles on the sky. To give you a sense of just how tiny, we might note that 1 arcsecond is how big a quarter would look when seen from a distance of 5 kilometres. The best images obtained from the ground with traditional techniques reveal details as small as several tenths of an arcsecond across. This image size is remarkably good. One of the main reasons for launching the Hubble Space Telescope was to escape Earth's atmosphere and obtain even sharper images.

Attribution

"6.2 Telescopes Today" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

4.8 OPTICAL DETECTORS AND ADAPTIVE OPTICS

Throughout most of the twentieth century, photographic film or **glass plates** served as the prime astronomical detectors, whether for photographing spectra or direct images of celestial objects. In a photographic plate, a light-sensitive chemical coating is applied to a piece of glass that, when developed, provides a lasting record of the image. At observatories around the world, vast collections of photographs preserve what the sky has looked like during the past 100 years. Photography represents a huge improvement over the human eye, but it still has limitations. Photographic films are inefficient: only about 1% of the light that actually falls on the film contributes to the chemical change that makes the image; the rest is wasted.

Astronomers today have much more efficient electronic detectors to record astronomical images. Most often, these are **charge-coupled devices (CCDs)**, which are similar to the detectors used in video camcorders or in digital cameras (like the one more and more students have on their cell phones). In a CCD, photons of radiation hitting any part of the detector generate a stream of charged particles (electrons) that are stored and counted at the end of the exposure. Each place where the radiation is counted is called a pixel (picture element), and modern detectors can count the photons in millions of pixels (megapixels, or MPs).

Charge-Coupled Devices (CCDs)

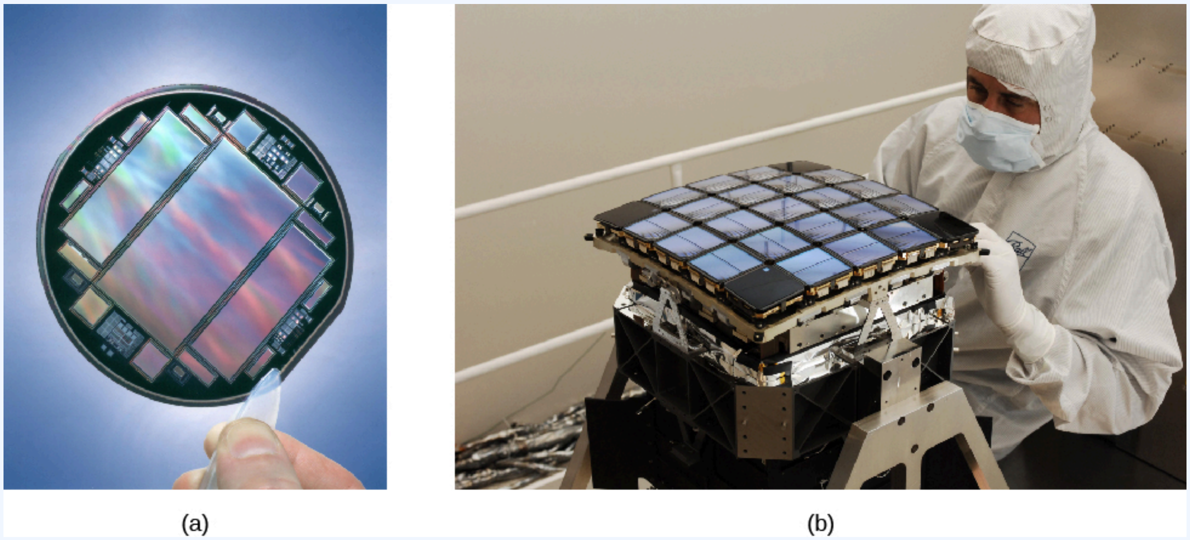


Figure 4.14. (a) This CCD is a mere 300-micrometres thick (thinner than a human hair) yet holds more than 21 million pixels. (b) This matrix of 42 CCDs serves the Kepler telescope.
(a): Fabrication run 3 (100 mm wafers) by Berkeley Lab/US Department of Energy, Berkeley Lab Media Licence.
(b): Kepler Focal Plane Array by NASA/Ball Aerospace, NASA Media Licence.

Because CCDs typically record as much as 60–70% of all the photons that strike them, and the best silicon and infrared CCDs exceed 90% sensitivity, we can detect much fainter objects. Among these are many small moons around the outer planets, icy dwarf planets beyond Pluto, and dwarf galaxies of stars. CCDs also provide more accurate measurements of the brightness of astronomical objects than photography, and their output is digital—in the form of numbers that can go directly into a computer for analysis.

Charge-coupled devices are not only used for detailed astronomical observations but are also the basis for most digital photography. Canadian physicist Willard S. Boyle (and American George E. Smith) created the first CCD in 1969 at Bell Labs. They shared the 2009 Nobel Prize in Physics for this new “electronic eye” along with Charles K. Kao for his work on fibre optic technology.

Figure 4.15. Willard S. Boyle.
Willard Boyle by Prolineserver, CC BY-SA 3.0.

Willard Boyle was born in 1924 in Amherst, Nova Scotia. He completed his education at McGill University with a Doctorate in Physics in 1950 after flying Spitfire fighter planes for the Royal Canadian Navy. Besides inventing the CCD, Willard was a co-creator of the first continuously operating ruby laser. He also worked as the Director of Space Science and Exploratory Studies at Bellcom (a division of Bell Labs) and supported the Apollo space program. Specifically, Willard aided in the selection of certain lunar landing sites during the 1960s. Willard worked on 18 patents during his career at Bell Labs and won several other awards for his work. In 2010, he was appointed a Companion to the Order of Canada before passing in 2011.

When fast-acting cameras, and faster-acting computers, were developed in the 1970s and 1980s, some astronomers began to think about ways to get around the blurring effects of the atmosphere. They have developed several approaches to the problem, but all are given the name **adaptive optics**.

The basic idea sounds pretty simple:

A. take a quick picture of a bright object with a “fast camera”
B. measure quantitatively how imperfect the image is
C. figure out how the atmosphere must be distorting the light rays right now
D. using an *active optical element*, distort the light rays in exactly the OPPOSITE manner
E. focus the un-distorted light rays on a “slow camera”

The first thing one needs is a “fast camera”, because the motion of the air changes the distortion of light rays many times each second. Typical “fast cameras” run at video rate—30 times per second—or even faster; some take over 1000 frames per second!

What sort of thing is an **active optical element**? There are several different types, but most are basically flexible mirrors with lots of little actuators to bend them. Some are small:

Liquid Mirror

Figure 4.16.

and others are large:

Mirror on the MMT

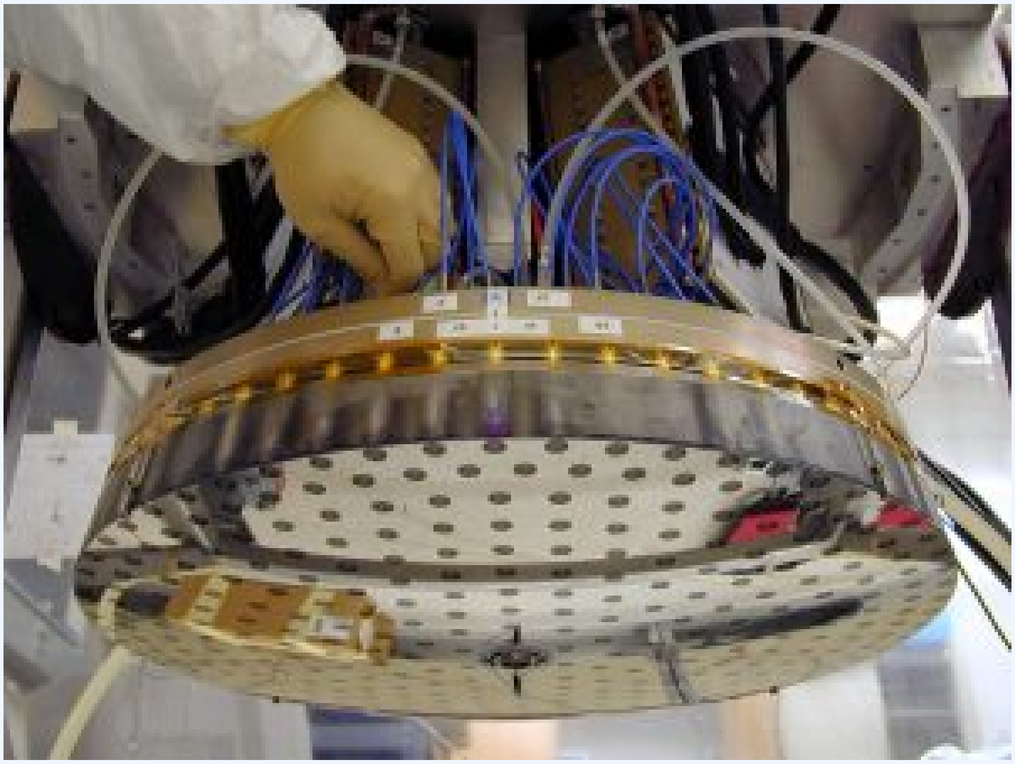


Figure 4.17. Secondary mirror on the MMT, from a talk by Laird Close

How well can these systems correct the air's distortions? Pretty darn well! It turns out to be easier for light of long wavelengths, in the near-infrared portion of the spectrum. Let's look at one example, taken with the MMT, a telescope of diameter roughly $D = 6.5$ m, at a wavelength of $\lambda = 1650$ nm.

Movie of Theta Ori 1

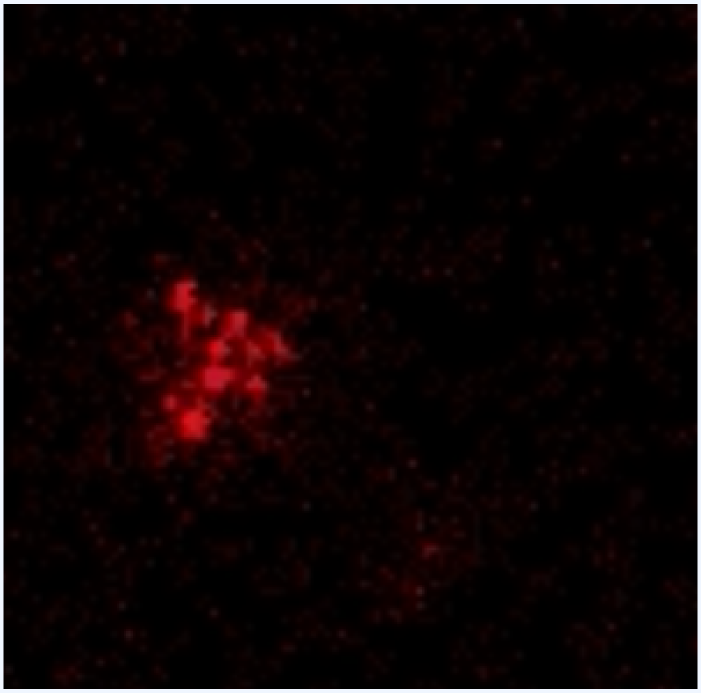


Figure 4.18. Movie of Theta Ori 1 observed by the MMT courtesy of a talk by Laird Close. Clicking on image will download AVI video file(2.13 MB, 40s duration).

Okay, let's compare that to the actual performance:

Performance of the MMT

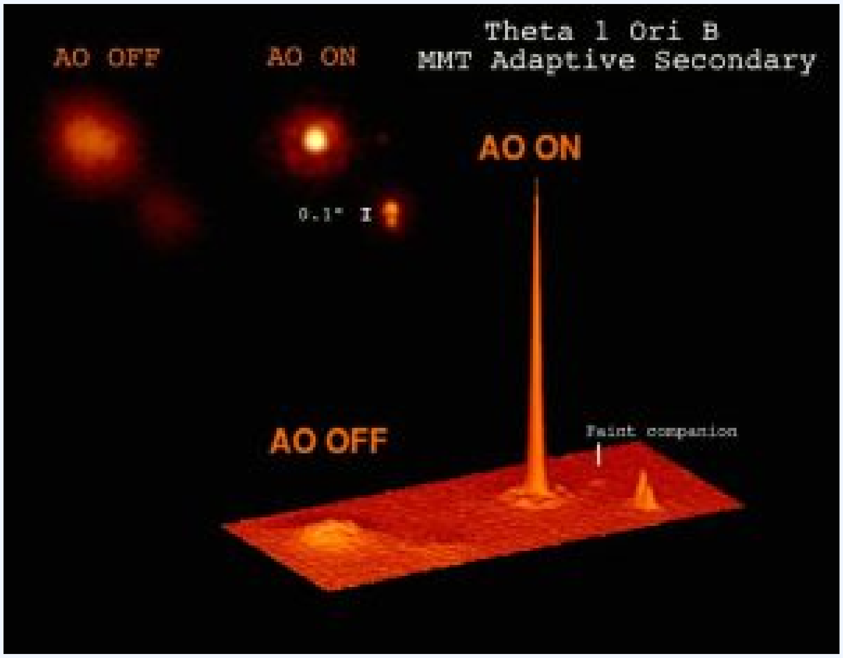


Figure 4.19.

Recently, astronomers have managed to use adaptive optics to improve images in the optical regime, at wavelengths we can see with our own eyes.

Attribution

"6.3 Visible-Light Detectors and Instruments" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

"Removing the effect of the atmosphere" from Adpative Optics by Michael Richmond is licensed under a Creative Commons Attribution Non-Commercial ShareAlike 4.0 International License, except where otherwise noted.

4.9 RADIO TELESCOPES AND INTERFEROMETRY

In addition to visible and infrared radiation, radio waves from astronomical objects can also be detected from the surface of Earth. In the early 1930s, Karl G. Jansky, an engineer at Bell Telephone Laboratories, was experimenting with antennas for long-range radio communication when he encountered some mysterious static—radio radiation coming from an unknown source as shown in Figure 4.13. He discovered that this radiation came in strongest about four minutes earlier on each successive day and correctly concluded that since Earth's sidereal rotation period (how long it takes us to rotate relative to the stars) is four minutes shorter than a solar day, the radiation must be originating from some region fixed on the celestial sphere. Subsequent investigation showed that the source of this radiation was part of the Milky Way Galaxy; Jansky had discovered the first source of cosmic radio waves.

First Radio Telescope

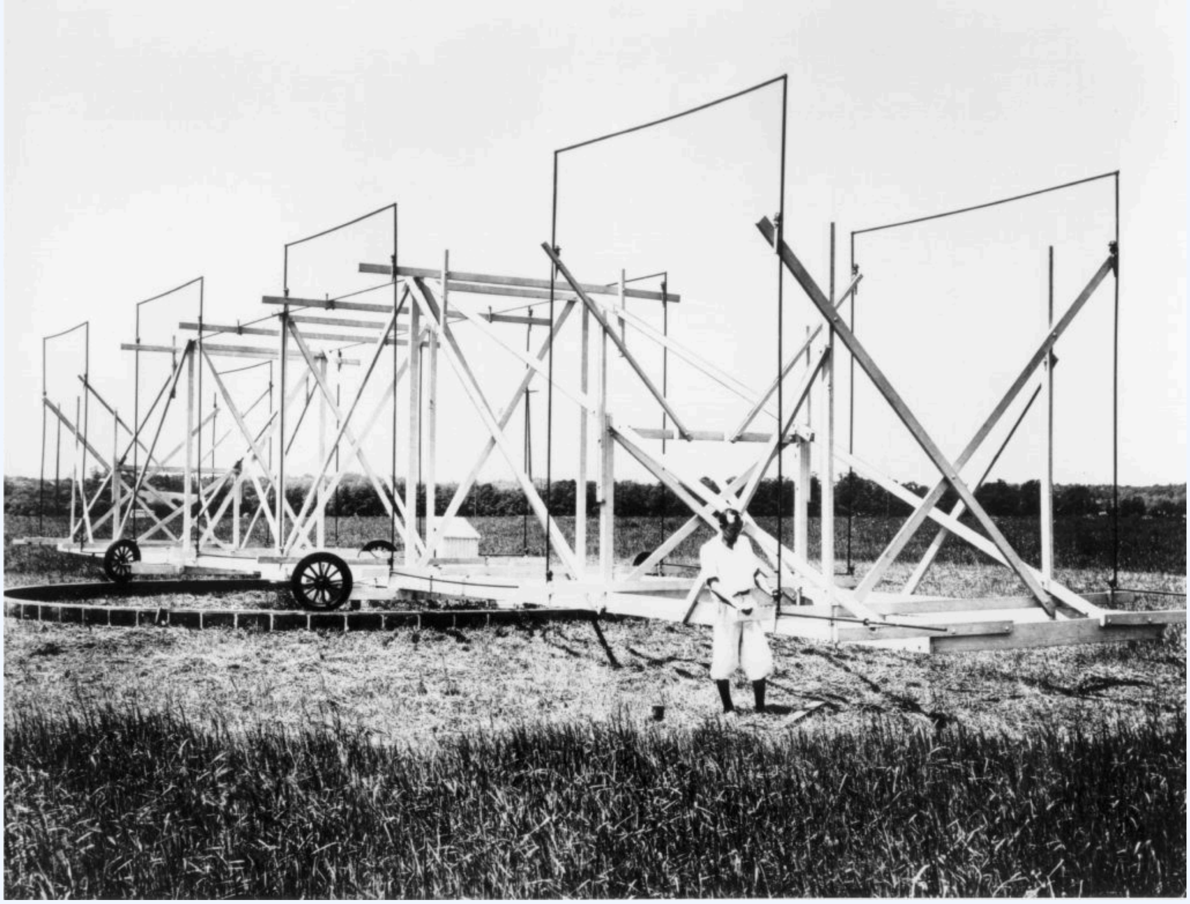


Figure 4.20. This rotating radio antenna was used by Jansky in his serendipitous discovery of radio radiation from the Milky Way.
Karl Jansky and his Merry-go-Round by NRAO/AUI/NSF, NRAO Image Use Policy.

In 1936, Grote Reber, who was an amateur astronomer interested in radio communications, used galvanized iron and wood to build the first antenna specifically designed to receive cosmic radio waves. Over the years, Reber built several such antennas and used them to carry out pioneering surveys of the sky for celestial radio sources; he remained active in radio astronomy for more than 30 years. During the first decade, he worked practically alone because professional astronomers had not yet recognized the vast potential of radio astronomy.

Detection of Radio Energy from Space

It is important to understand that radio waves cannot be "heard": they are not the sound waves you hear coming out of the radio receiver in your home or car. Like light, radio waves are a form of electromagnetic radiation, but unlike light, we cannot detect them with our senses—we must rely on electronic equipment to pick them up. In commercial radio broadcasting, we encode sound information (music or a newscaster's voice) into radio waves. These must be decoded at the other end and then turned back into sound by speakers or headphones.

The radio waves we receive from space do not, of course, have music or other program information encoded in them. If cosmic radio signals were translated into sound, they would sound like the static you hear when scanning between stations. Nevertheless, there is information in the radio waves we receive—information that can tell us about the chemistry and physical conditions of the sources of the waves.

Just as vibrating charged particles can produce electromagnetic waves, electromagnetic waves can make charged particles move back and forth. Radio waves can produce a current in conductors of electricity such as metals. An antenna is such a conductor: it intercepts radio waves, which create a feeble current in it. The current is then amplified in a radio receiver until it is strong enough to measure or record. Like your television or radio, receivers can be tuned to select a single frequency (channel). In astronomy, however, it is more common to use sophisticated data-processing techniques that allow thousands of separate frequency bands to be detected simultaneously. Thus, the astronomical radio receiver operates much like a spectrometer on a visible-light or infrared telescope, providing information about how much radiation we receive at each wavelength or frequency. After computer processing, the radio signals are recorded on magnetic disks for further analysis.

Radio waves are reflected by conducting surfaces, just as light is reflected from a shiny metallic surface, and according to the same laws of optics. A radio-reflecting telescope consists of a concave metal reflector (called a **dish**), analogous to a telescope mirror. The radio waves collected by the dish are reflected to a focus, where they can then be directed to a receiver and analyzed. Because humans are such visual creatures, radio astronomers often construct a pictorial representation of the radio sources they observe. Figure 4.14 shows such a radio image of a distant galaxy, where radio telescopes reveal vast jets and complicated regions of radio emissions that are completely invisible in photographs taken with light.

Radio Image of a Galaxy in Cygnus A

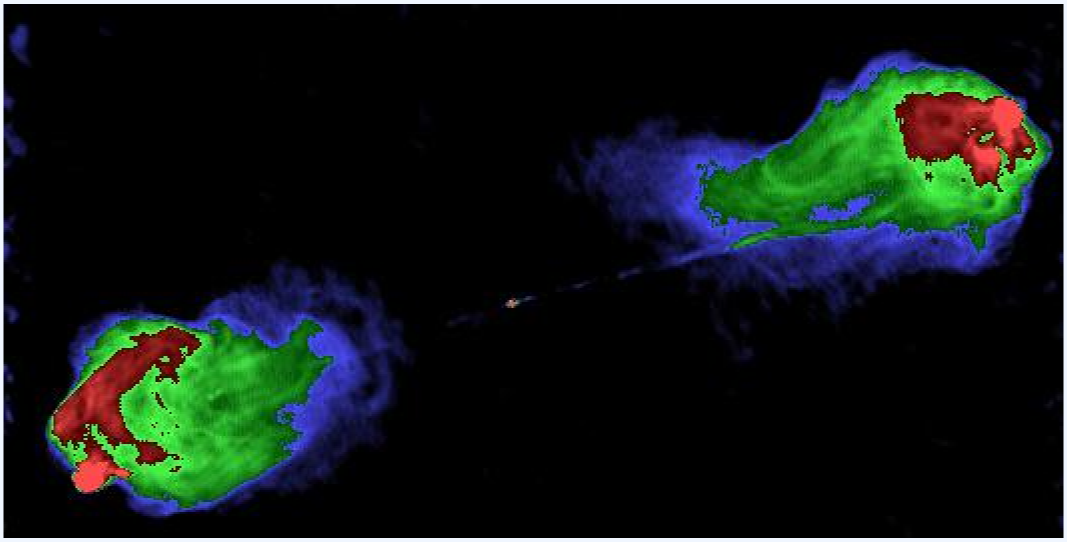


Figure 4.21. This image has been constructed of radio observations at the Very Large Array of a galaxy called Cygnus A. Colours have been added to help the eye sort out regions of different radio intensities. Red regions are the most intense, blue the least. The visible galaxy would be a small dot in the centre of the image. The radio image reveals jets of expelled material (more than 160,000 light-years long) on either side of the galaxy.
Cygnus A (3C405) by NRAO/AUI, adapted by Mhardcastle, CC BY-SA 3.0.

Radio astronomy is a young field compared with visible-light astronomy, but it has experienced tremendous growth in recent decades. The world's largest radio reflectors that can be pointed to any direction in the sky have apertures of 100 metres. One of these has been built at the US National Radio Astronomy Observatory in West Virginia as shown in Figure 4.15. Table 4.2 lists some of the major radio telescopes of the world.

Robert C. Byrd Green Bank Telescope

Figure 4.22. This fully steerable radio telescope in West Virginia went into operation in August 2000. Its dish is about 100 metres across.
Green Bank 100m diameter Radio Telescope by NRAO/AUI, CC BY 3.0.

Table 4.2. Major Radio Observatories of the World

Observatory	Location	Description	Website
Individual Radio Dishes			
Arecibo Observatory	Arecibo, Puerto Rico	305-m fixed dish	www.naic.edu
Green Bank Telescope (GBT)	Green Bank, WV	110 × 100-m steerable dish	www.science.nrao.edu/facilities/gbt
Effelsberg 100-m Telescope	Bonn, Germany	100-m steerable dish	www.mpifr-bonn.mpg.de/en/effelsberg
Lovell Telescope	Manchester, England	76-m steerable dish	www.jb.man.ac.uk/aboutus/lovell
Canberra Deep Space Communication Complex (CDSCC)	Tidbinbilla, Australia	70-m steerable dish	www.cdscc.nasa.gov
Goldstone Deep Space Communications Complex (GDSCC)	Barstow, CA	70-m steerable dish	www.gdscc.nasa.gov
Parkes Observatory	Parkes, Australia	64-m steerable dish	www.parkes.atnf.csiro.au
Arrays of Radio Dishes			
Square Kilometre Array (SKA)	South Africa and Western Australia	Thousands of dishes, km^2 collecting area, partial array in 2020	www.skatelescope.org
Atacama Large Millimeter/submillimeter Array (ALMA)	Atacama desert, Northern Chile	66 7-m and 12-m dishes	www.almaobservatory.org
Very Large Array (VLA)	Socorro, New Mexico	27-element array of 25-m dishes (36-km baseline)	www.science.nrao.edu/facilities/vla
Westerbork Synthesis Radio Telescope (WSRT)	Westerbork, the Netherlands	12-element array of 25-m dishes (1.6-km baseline)	www.astron.nl/radio-observatory/public/public-0
Very Long Baseline Array (VLBA)	Ten US sites, HI to the Virgin Islands	10-element array of 25-m dishes (9000 km baseline)	www.science.nrao.edu/facilities/vlba
Australia Telescope Compact Array (ATCA)	Several sites in Australia	8-element array (seven 22-m dishes plus Parkes 64 m)	www.narrabri.atnf.csiro.au
Multi-Element Radio Linked Interferometer Network (MERLIN)	Cambridge, England, and other British sites	Network of seven dishes (the largest is 32 m)	www.e-merlin.ac.uk
Millimeter-wave Telescopes			

Observatory	**Location**	**Description**	**Website**
IRAM	Granada, Spain	30-m steerable mm-wave dish	www.iram-institute.org
James Clerk Maxwell Telescope (JCMT)	Mauna Kea, HI	15-m steerable mm-wave dish	www.eaobservatory.org/jcmt
Nobeyama Radio Observatory (NRO)	Minamimaki, Japan	6-element array of 10-m wave dishes	www.nro.nao.ac.jp/en
Hat Creek Radio Observatory (HCRO)	Cassel, CA	6-element array of 5-m wave dishes	www.sri.com/research-development/specialized-facilities/hat-creek-radio-observatory

Radio Interferometry

As we discussed earlier, a telescope's ability to show us fine detail (its resolution) depends upon its aperture, but it also depends upon the wavelength of the radiation that the telescope is gathering. The longer the waves, the harder it is to resolve fine detail in the images or maps we make. Because radio waves have such long wavelengths, they present tremendous challenges for astronomers who need good resolution. In fact, even the largest radio dishes on Earth, operating alone, cannot make out as much detail as the typical small visible-light telescope used in a college astronomy lab. To overcome this difficulty, radio astronomers have learned to sharpen their images by linking two or more radio telescopes together electronically. Two or more telescopes linked together in this way are called an **interferometer**.

"Interferometer" may seem like a strange term because the telescopes in an interferometer work cooperatively; they don't "interfere" with each other. **Interference**, however, is a technical term for the way that multiple waves interact with each other when they arrive in our instruments, and this interaction allows us to coax more detail out of our observations. The resolution of an interferometer depends upon the separation of the telescopes, not upon their individual apertures. Two telescopes separated by 1 kilometre provide the same resolution as would a single dish 1 kilometre across (although they are not, of course, able to collect as much radiation as a radio-wave bucket that is 1 kilometre across).

To get even better resolution, astronomers combine a large number of radio dishes into an **interferometer array**. In effect, such an array works like a large number of two-dish interferometres, all observing the same part of the sky together. Computer processing of the results permits the reconstruction of a high-resolution radio image. The most extensive such instrument in the United States is the National Radio Astronomy Observatory's Very Large Array (VLA) near Socorro, New Mexico. It consists of 27 movable radio telescopes (on railroad tracks), each having an aperture of 25 metres, spread over a total span of about 36 kilometres. By electronically combining the signals from all of its individual telescopes, this array permits the radio astronomer to make pictures of the sky at radio wavelengths comparable to those obtained with a visible-light telescope, with a resolution of about 1 arcsecond.

The Atacama Large Millimetre/submillimetre array (ALMA) in the Atacama Desert of Northern Chile as shown in Figure 4.16, at an altitude of 16,400 feet, consists of 12 7-meter and 54 12-meter telescopes, and can achieve baselines up to 16 kilometres. Since it became operational in 2013, it has made observations at resolutions down to 6 milliarcseconds (0.006 arcseconds), a remarkable achievement for radio astronomy.

Atacama Large Millimeter/Submillimeter Array (ALMA)

Figure 4.23. Located in the Atacama Desert of Northern Chile, ALMA currently provides the highest resolution for radio observations.
ALMA by ESO/S. Guisard, CC BY 4.0.

Watch this documentary from the European Space Agency that explains the work that went into designing and building ALMA, discusses some of its first images, and explores its future. The URL is: https://youtu.be/_Ryctl1Gij4

Initially, the size of interferometer arrays was limited by the requirement that all of the dishes be physically wired together. The maximum dimensions of the array were thus only a few tens of kilometres. However, larger interferometer separations can be achieved if the telescopes do not require a physical connection. Astronomers, with the use of current technology and computing power, have learned to time the arrival of electromagnetic waves coming from space very precisely at each telescope and combine the data later. If the telescopes are as far apart as California and Australia, or as West Virginia and Crimea in Ukraine, the resulting resolution far surpasses that of visible-light telescopes.

The United States operates the Very Long Baseline Array (VLBA), made up of 10 individual telescopes stretching from the Virgin Islands to Hawaii as shown in Figure 4.17. The VLBA, completed in 1993, can form astronomical images with a resolution of 0.0001 arcseconds, permitting features as small as 10 astronomical units (AU) to be distinguished at the center of our Galaxy.

Very Long Baseline Array

Figure 4.24. This map shows the distribution of 10 antennas that constitute an array of radio telescopes stretching across the United States and its territories.

Recent advances in technology have also made it possible to do interferometry at visible-light and infrared wavelengths. At the beginning of the twenty-first century, three observatories with multiple telescopes each began using their dishes as interferometres, combining their light to obtain a much greater resolution. In addition, a dedicated interferometric array was built on Mt. Wilson in California. Just as in radio arrays, these observations allow astronomers to make out more detail than a single telescope could provide.

Table 4.3. Visible-Light Interferometres

Longest Baseline (m)	Telescope Name	Location	Mirrors	Status
400	CHARA Array (Center for High Angular Resolution Astronomy)	Mount Wilson, CA	Six 1-m telescopes	Operational since 2004
200	Very Large Telescope	Cerro Paranal, Chile	Four 8.2-m telescopes	Completed 2000
85	Keck I and II telescopes	Mauna Kea, HI	Two 10-m telescopes	Operated from 2001 to 2012
22.8	Large Binocular Telescope	Mount Graham, AZ	Two 8.4-m telescopes	First light 2004

Radar Astronomy

Radar is the technique of transmitting radio waves to an object in our solar system and then detecting the radio radiation that the object reflects back. The time required for the round trip can be measured electronically with great precision. Because we know the speed at which radio waves travel (the speed of light), we can determine the distance to the object or a particular feature on its surface (such as a mountain).

Radar observations have been used to determine the distances to planets and how fast things are moving in the solar system (using the Doppler effect). Radar waves have played important roles in navigating spacecraft throughout the solar system. In addition, as will be discussed in later chapters, radar observations have determined the rotation periods of Venus and Mercury, probed tiny Earth-approaching asteroids, and allowed us to investigate the mountains and valleys on the surfaces of Mercury, Venus, Mars, and the large moons of Jupiter.

Any radio dish can be used as a radar telescope if it is equipped with a powerful transmitter as well as a receiver. The most spectacular facility in the world for radar astronomy is the 1000-foot (305-meter) telescope at Arecibo in Puerto Rico ([link]). The Arecibo telescope is too large to be pointed directly at different parts of the sky. Instead, it is constructed in a huge natural "bowl" (more than a mere dish) formed by several hills, and it is lined with reflecting metal panels. A limited ability to track astronomical sources is achieved by moving the receiver system, which is suspended on cables 100 metres above the surface of the bowl. An even larger (500-meter) radar telescope is currently under construction. It is the Five-hundred-meter Aperture Spherical Telescope (FAST) in China and is expected to be completed in 2016.

Largest Radio and Radar Dish

Figure 4.25. The Arecibo Observatory, with its 1000-foot radio dish-filling valley in Puerto Rico, is part of the National Astronomy and Ionosphere Center, operated by SRI International, USRA, and UMET under a cooperative agreement with the National Science Foundation.
NASA Universe Arecibo Observatory Aerial View by H. Schweiker/WIYN and NOAO/AURA/NSF, CC BY 2.0.

Attribution

"6.4 Radio Telescopes" from Douglas College Astronomy 1105 by Douglas College Department of Physics

4.10 SPACE TELESCOPES

Earth's atmosphere blocks most radiation at wavelengths shorter than visible light, so we can only make direct ultraviolet, X-ray, and gamma ray observations from space (though indirect gamma ray observations can be made from Earth). Getting above the distorting effects of the atmosphere is also an advantage at visible and infrared wavelengths. The stars don't "twinkle" in space, so the amount of detail you can observe is limited only by the size of your instrument. On the other hand, it is expensive to place telescopes into space, and repairs can present a major challenge. This is why astronomers continue to build telescopes for use on the ground as well as for launching into space.

Airborne and Space Infrared Telescopes

Water vapor, the main source of atmospheric interference for making infrared observations, is concentrated in the lower part of Earth's atmosphere. For this reason, a gain of even a few hundred metres in elevation can make an important difference in the quality of an infrared observatory site. Given the limitations of high mountains, most of which attract clouds and violent storms, and the fact that the ability of humans to perform complex tasks degrades at high altitudes, it was natural for astronomers to investigate the possibility of observing infrared waves from airplanes and ultimately from space.

Infrared observations from airplanes have been made since the 1960s, starting with a 15-centimetre telescope on board a Learjet. From 1974 through 1995, NASA operated a 0.9-meter airborne telescope flying regularly out of the Ames Research Center south of San Francisco. Observing from an altitude of 12 kilometres, the telescope was above 99% of the atmospheric water vapour. More recently, NASA (in partnership with the German Aerospace Center) has constructed a much larger 2.5-meter telescope, called the Stratospheric Observatory for Infrared Astronomy (SOFIA), which flies in a modified Boeing 747SP as shown in Figure 4.19.

Stratospheric Observatory for Infrared Astronomy (SOFIA)

Figure 4.26. SOFIA allows observations to be made above most of Earth's atmospheric water vapour. SOFIA by NASA, NASA Media Licence.

Getting even higher and making observations from space itself have important advantages for infrared astronomy. First is the elimination of all interference from the atmosphere. Equally important is the opportunity to cool the entire optical system of the instrument in order to nearly eliminate infrared radiation from the telescope itself. If we tried to cool a telescope within the atmosphere, it would quickly become coated with condensing water vapor and other gases, making it useless. Only in the vacuum of space can optical elements be cooled to hundreds of degrees below freezing and still remain operational.

The first orbiting infrared observatory, launched in 1983, was the **Infrared Astronomical Satellite (IRAS)**, built as a joint project by the United States, the Netherlands, and Britain. IRAS was equipped with a 0.6-meter telescope cooled to a temperature of less than 10 K. For the first time, the infrared sky could be seen as if it were night, rather than through a bright foreground of atmospheric and telescope emissions. IRAS carried out a rapid but comprehensive survey of the entire infrared sky over a 10-month period, cataloguing about 350,000 sources of infrared radiation. Since then, several other infrared telescopes have operated in space with much better sensitivity and resolution due to improvements in infrared detectors. The most powerful of these infrared telescopes is the 0.85-meter Spitzer Space Telescope, which launched in 2003. A few of its observations are shown in Figure 4.20. With infrared observations, astronomers can detect cooler parts of cosmic objects, such as the dust clouds around star nurseries and the remnants of dying stars, that visible-light images don't reveal.

Observations from the Spitzer Space Telescope (SST)

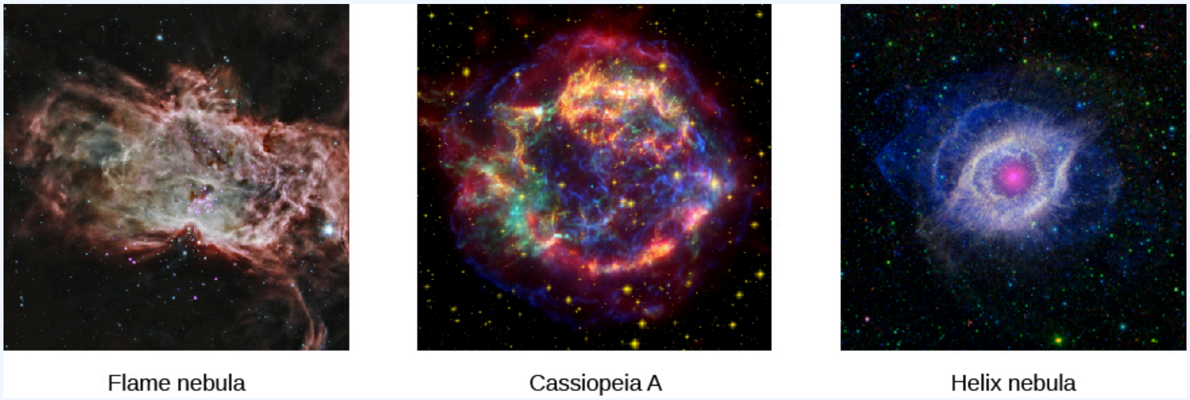


Figure 4.27. These infrared images—a region of star formation, the remnant of an exploded star, and a region where an old star is losing its outer shell—show just a few of the observations made and transmitted back to Earth from the SST. Since our eyes are not sensitive to infrared rays, we don't perceive colours from them. The colours in these images have been selected by astronomers to highlight details like the composition or temperature in these regions.
(a): PIA18249: Inside the Flame Nebula by X-ray: NASA/CXC/PSU/K.Getman, E.Feigelson, M.Kuhn and the MYStIX team; Infrared: NASA/JPL-Caltech, NASA Media Licence.
(b): PIA03519: Cassiopeia A: Death Becomes Her by NASA/JPL-Caltech/STScI/CXC/SAO, NASA Media Licence.
(c): PIA15817: The Helix Nebula: Unraveling at the Seams by NASA/JPL-Caltech, NASA Media Licence.

Hubble Space Telescope

In April 1990, a great leap forward in astronomy was made with the launch of the **Hubble Space Telescope (HST)**. With an aperture of 2.4 metres, this is the largest telescope put into space so far. (Its aperture was limited by the size of the payload bay in the Space Shuttle that served as its launch vehicle.) It was named for Edwin Hubble, the astronomer who discovered the expansion of the universe in the 1920s (whose work we will discuss in the chapters on Galaxies).

HST is operated jointly by NASA's Goddard Space Flight Centre and the Space Telescope Science Institute in Baltimore. It was the first orbiting observatory designed to be serviced by Shuttle astronauts and, over the years since it was launched, they made several visits to improve or replace its initial instruments and to repair some of the systems that operate the spacecraft — though this repair program has now been discontinued, and no more visits or improvements will be made. You can get more information about the history of the repairs by visiting https://www.nasa.gov/mission_pages/hubble/servicing/index.html .

With the Hubble, astronomers have obtained some of the most detailed images of astronomical objects from the solar system outward to the most distant galaxies. Among its many great achievements is the Hubble Ultra-Deep Field, an image of a small region of the sky observed for almost 100 hours. It contains views of about 10,000 galaxies, some of which formed when the universe was just a few percent of its current age as shown in Figure 4.21.

Hubble Ultra-Deep Field (HUDF)

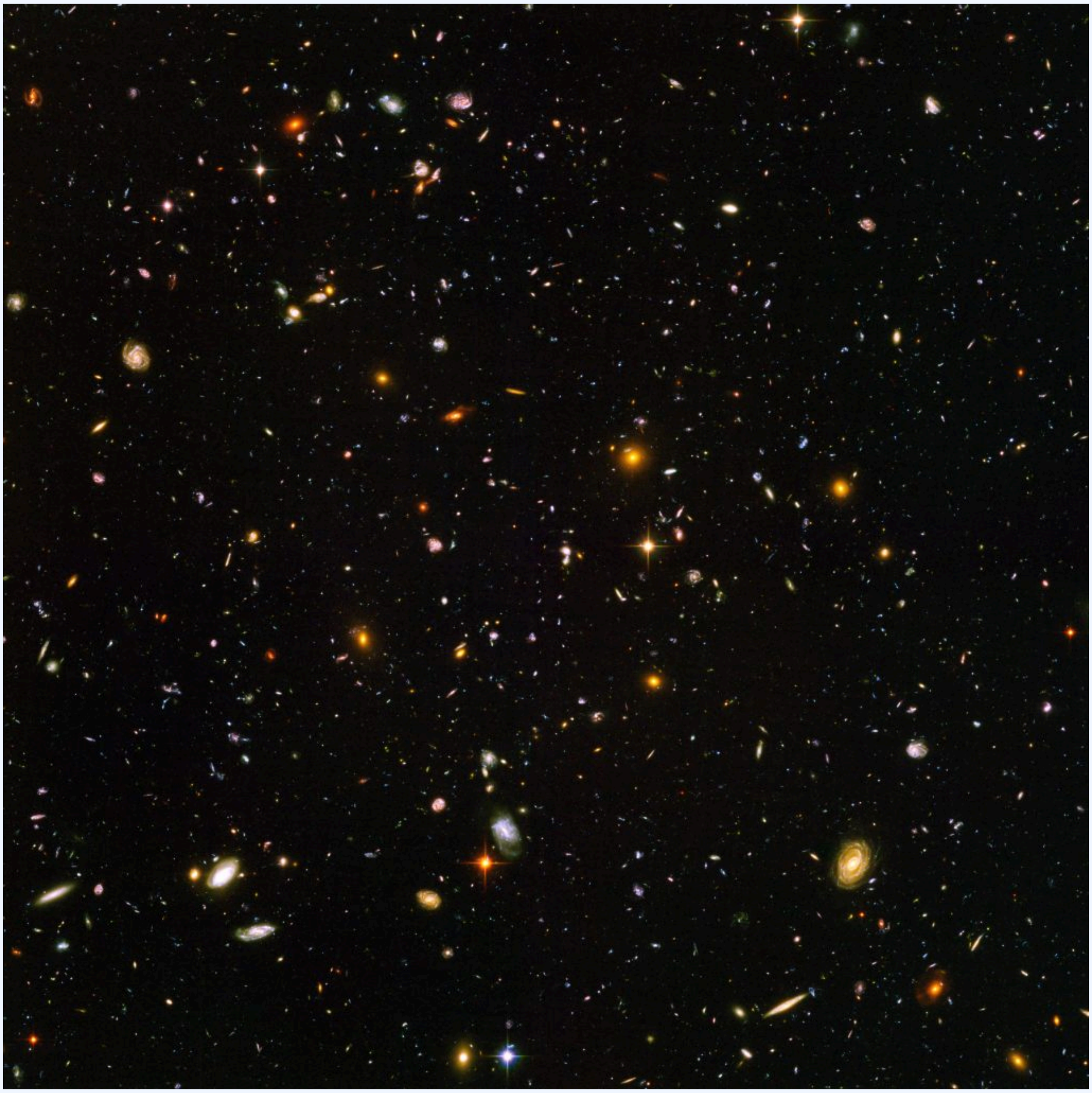


Figure 4.28. The Hubble Space Telescope has provided an image of a specific region of space built from data collected between September 24, 2003, and January 16, 2004. These data allow us to search for galaxies that existed approximately 13 billion years ago.
Hubble Ultra Deep Field by NASA, ESA, and S. Beckwith (STScI) and the HUDF Team, NASA Media Licence.

The HST's mirror was ground and polished to a remarkable degree of accuracy. If we were to scale up its

2.4-meter mirror to the size of the entire continental United States, there would be no hill or valley larger than about 6 centimetres in its smooth surface. Unfortunately, after it was launched, scientists discovered that the primary mirror had a slight error in its *shape,* equal to roughly 1/50 the width of a human hair. Small as that sounds, it was enough to ensure that much of the light entering the telescope did not come to a clear focus and that all the images were blurry. (In a misplaced effort to save money, a complete test of the optical system had not been carried out before launch, so the error was not discovered until HST was in orbit.)

The solution was to do something very similar to what we do for astronomy students with blurry vision: put corrective optics in front of their eyes. In December 1993, in one of the most exciting and difficult space missions ever flown, astronauts captured the orbiting telescope and brought it back into the shuttle payload bay. There they installed a package containing compensating optics as well as a new, improved camera before releasing HST back into orbit. The telescope now works as it was intended to, and further missions to it were able to install even more advanced instruments to take advantage of its capabilities.

High-Energy Observatories

Ultraviolet, X-ray, and direct gamma-ray (high-energy electromagnetic wave) observations can be made only from space. Such observations first became possible in 1946, with V2 rockets captured from Germany after World War II. The US Naval Research Laboratory put instruments on these rockets for a series of pioneering flights, used initially to detect ultraviolet radiation from the Sun. Since then, many other rockets have been launched to make X-ray and ultraviolet observations of the Sun, and later of other celestial objects.

Beginning in the 1960s, a steady stream of high-energy observatories has been launched into orbit to reveal and explore the universe at short wavelengths. Among recent X-ray telescopes is the Chandra X-ray Observatory, which was launched in 1999. It is shown in Figure 4.22. It is producing X-ray images with unprecedented resolution and sensitivity. Designing instruments that can collect and focus energetic radiation like X-rays and gamma rays is an enormous technological challenge. The 2002 Nobel Prize in physics was awarded to Riccardo Giacconi, a pioneer in the field of building and launching sophisticated X-ray instruments. In 2008, NASA launched the Fermi Gamma-ray Space Telescope, designed to measure cosmic gamma rays at energies greater than any previous telescope, and thus able to collect radiation from some of the most energetic events in the universe.

Chandra X-Ray Satellite

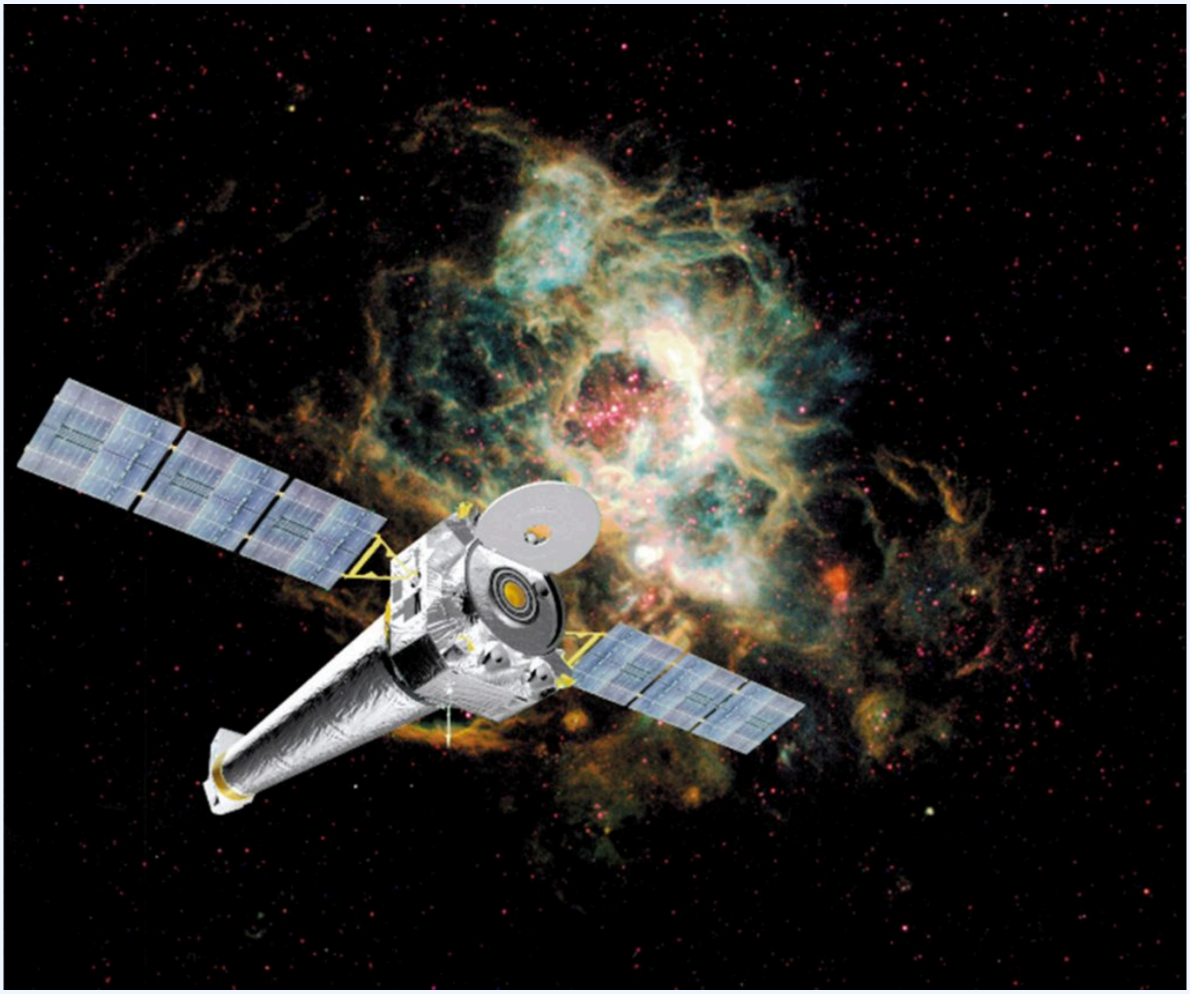


Figure 4.29. Chandra, the world's most powerful X-ray telescope, was developed by NASA and launched in July 1999.
Chandra X-Ray Observatory (Chandra XRO) by NASA, NASA Media Licence.

One major challenge is to design "mirrors" to reflect such penetrating radiation as X-rays and gamma rays, which normally pass straight through matter. However, although the technical details of design are more complicated, the three basic components of an observing system, as we explained earlier in this chapter, are the same at all wavelengths: a telescope to gather up the radiation, filters or instruments to sort the radiation according to wavelength, and some method of detecting and making a permanent record of the observations. Table 4.3 lists some of the most important active space observatories that humanity has launched.

Gamma-ray detections can also be made from Earth's surface by using the atmosphere as the primary detector. When a gamma ray hits our atmosphere, it accelerates charged particles (mostly electrons) in the atmosphere. Those energetic particles hit other particles in the atmosphere and give off their own radiation. The effect is a cascade of light and energy that can be detected on the ground. The VERITAS array in Arizona and the H.E.S.S. array in Namibia are two such ground-based gamma-ray observatories.

Table 4.3. Recent Observatories in Space

Observatory	**Date Operation Began**	**Bands of the Spectrum**	**Notes**	**Website**
Hubble Space Telescope (HST)	1990	visible, UV, IR	2.4-m mirror; images and spectra	www.hubblesite.org
Chandra X-Ray Observatory	1999	X-rays	X-ray images and spectra	www.chandra.si.edu
XMM-Newton	1999	X-rays	X-ray spectroscopy	http://www.cosmos.esa.int/web/xmm-newton
International Gamma-Ray Astrophysics Laboratory (INTEGRAL)	2002	X- and gamma-rays	higher resolution gamma-ray images	http://sci.esa.int/integral/
Spitzer Space Telescope	2003	IR	0.85-m telescope	www.spitzer.caltech.edu
Fermi Gamma-ray Space Telescope	2008	gamma-rays	first high-energy gamma-ray observations	fermi.gsfc.nasa.gov
Kepler	2009	visible-light	planet finder	http://kepler.nasa.gov
Wide-field Infrared Survey Explorer (WISE)	2009	IR	whole-sky map, asteroid searches	www.nasa.gov/mission_pages/WISE/main
Gaia	2013	visible-light	Precise map of the Milky Way	http://sci.esa.int/gaia/

Attribution

"6.5 Observations outside Earth's Atmosphere" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

4.11 THE FUTURE OF LARGE TELESCOPES

If you've ever gone on a hike, you have probably been eager to see what lies just around the next bend in the path. Researchers are no different, and astronomers and engineers are working on the technologies that will allow us to explore even more distant parts of the universe and to see them more clearly.

The premier space facility planned for the next decade is the James Webb Space Telescope as shown in Figure 4.23, which (in a departure from tradition) is named after one of the early administrators of NASA instead of a scientist. This telescope will have a mirror 6 metres in diameter, made up, like the Keck telescopes, of 36 small hexagons. These will have to unfold into place once the telescope reaches its stable orbit point, some 1.5 million kilometres from Earth (where no astronauts can currently travel if it needs repair.) The telescope is scheduled for launch in 2018 and should have the sensitivity needed to detect the very first generation of stars, formed when the universe was only a few hundred million years old. With the ability to measure both visible and infrared wavelengths, it will serve as the successor to both HST and the Spitzer Space Telescope.

James Webb Space Telescope (JWST)

Figure 4.30. This image shows some of the mirrors of the JWST as they underwent cryogenic testing. The mirrors were exposed to extreme temperatures in order to gather accurate measurements on changes in their shape as they heated and cooled.
James Webb Space Telescope Cryogenic Mirror testing by NASA/MSFC/David Higginbotham/Emmett Given, CC BY 2.0.

On the ground, astronomers have started building the Large Synoptic Survey Telescope (LSST), an 8.4-meter telescope with a significantly larger field of view than any existing telescopes. It will rapidly scan the sky to find **transients**, phenomena that change quickly, such as exploding stars and chunks of rock that orbit near Earth. The LSST is expected to see first light in 2021.

The international gamma-ray community is planning the Cherenkov Telescope Array (CTA), two arrays of telescopes, one in each hemisphere, which will indirectly measure gamma rays from the ground. The CTA will measure gamma-ray energies a thousand times as great as the Fermi telescope can detect.

Several groups of astronomers around the globe interested in studying visible light and infrared are exploring the feasibility of building ground-based telescopes with mirrors larger than 30 metres across. Stop and think what this means: 30 metres is one-third the length of a football field. It is technically impossible to build and transport a single astronomical mirror that is 30 metres or larger in diameter. The primary mirror of these giant telescopes will consist of smaller mirrors, all aligned so that they act as a very large mirror in combination. These include the Thirty-Meter Telescope for which construction has begun at the top of Mauna Kea in Hawaii.

The most ambitious of these projects is the European Extremely Large Telescope (E-ELT) as shown in Figure 4.24. (Astronomers try to outdo each other not only with the size of these telescopes, but also their names!) The design of the E-ELT calls for a 39.3-meter primary mirror, which will follow the Keck design and be made up of 798 hexagonal mirrors, each 1.4 metres in diameter and all held precisely in position so that they form a continuous surface.

Artist's Conception of the European Extremely Large Telescope

Figure 4.31. The primary mirror in The Extremely Large Telescope (ELT) is 39.3 metres across. The telescope is under construction in the Atacama Desert in Northern Chile.
How the ELT will look like at Cerro Armazones by ESO, CC BY 4.0.

Construction on the site in the Atacama Desert in Northern Chile started in 2014. The E-ELT, along with the Thirty Meter Telescope https://www.tmt.org and the Giant Magellan Telescope, which are being built by international consortia led by US astronomers, will combine light-gathering power with high-resolution imaging. These powerful new instruments will enable astronomers to tackle many important astronomical problems. For example, they should be able to tell us when, where, and how often planets form around other stars. They should even be able to provide us images and spectra of such planets and thus, perhaps, give us the first real evidence (from the chemistry of these planets' atmospheres) that life exists elsewhere.

Sizes of Planned and Existing Telescopes

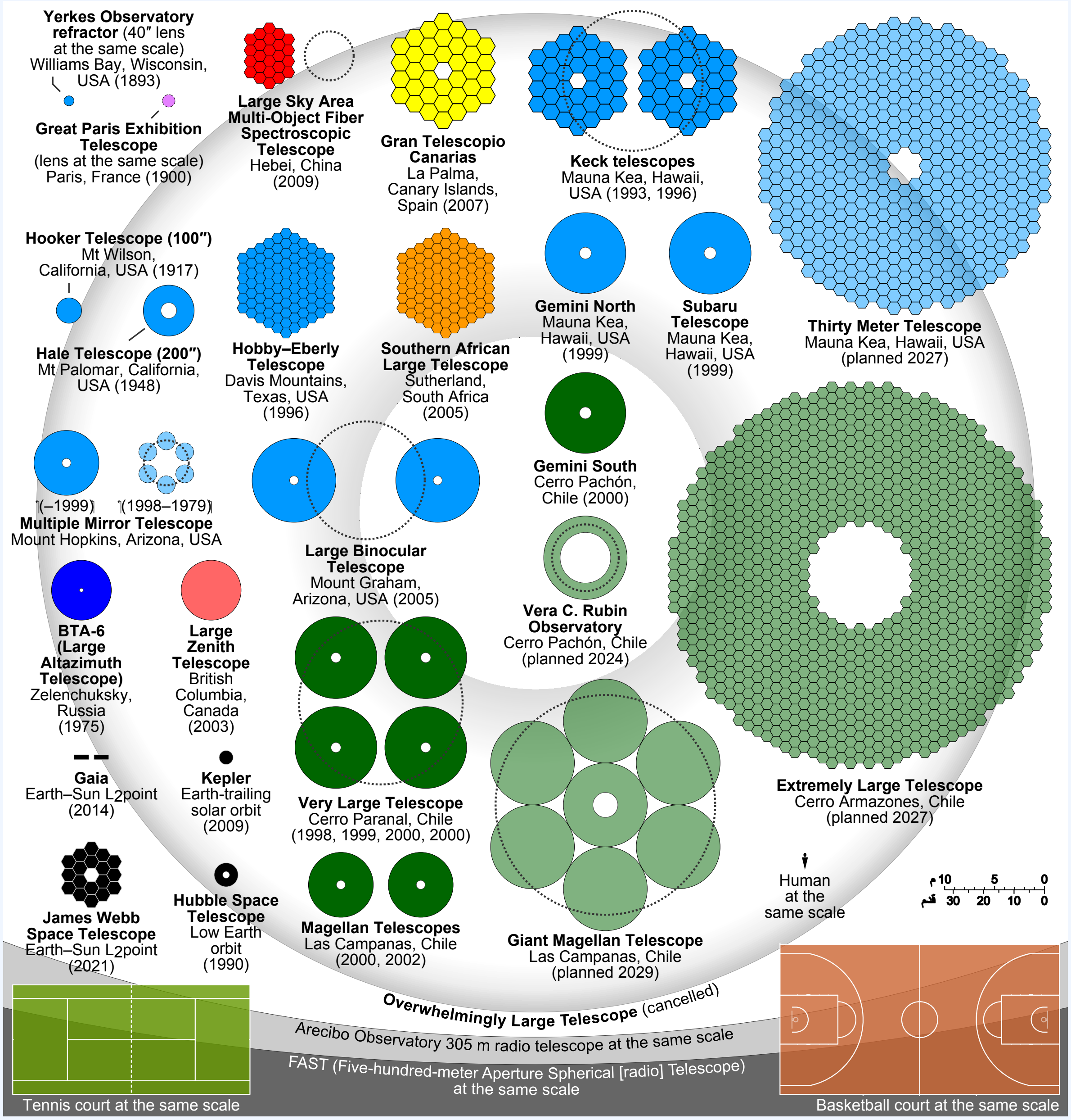


Figure 4.32. Sizes of planned and existing telescopes, compared to a basketball court and tennis court. Source: Wikimedia Commons.

Attribution

4.12 KEY TERMS

Active control: a process that measures a sag (like that of a Gemini mirror) many times each second and apply forces at 120 different locations to the back of the mirror to correct the sag. 4.5

Active optical element: flexible mirrors with lots of little actuators to bend them. 4.8

Adaptive optics: systems used with telescopes that can compensate for distortions in an image introduced by the atmosphere, thus resulting in sharper images. 4.8

Ancient *observatories*: special sites for observing the sky built by many ancient cultures that often had religious and ritual functions as well. 4.2

Aperture: diameter of the primary lens or mirror of a telescope. 4.3

Charge-coupled devices (CCDs): electronic detectors used to record astronomical images which are similar to the detectors used in video camcorders or in digital cameras. 4.8

Dish: a concave metal reflector. 4.9

Eyepiece: magnifying lens used to view the image produced by the objective lens or primary mirror of a telescope. 4.4

Focal length (of a lens): the distance from the lens to the location where the light rays focus, or image, behind the lens. 4.4

Focus (of a lens): the point at which all parallel rays of light converge after refracting due to the curvatures of a lens' surfaces. 4.4

Glass plates: the prime astronomical detectors used throughout most of the twentieth century, whether for photographing spectra or direct images of celestial objects. 4.8

Hubble Space Telescope (HST): launched in April 1990, this is the largest telescope put into space so far with an aperture of 2.4 metres. 4.10

Infrared Astronomical Satellite (IRAS): the first orbiting infrared observatory, launched in 1983 and built as a joint project by the United States, the Netherlands, and Britain. 4.10

Interference: a technical term for the way that multiple waves interact with each other when they arrive in our instruments, and this interaction allows us to coax more detail out of our observations. 4.9

Interferometer: two or more radio telescopes linked together electronically to sharpen their images. 4.9

Interferometer array: combination of multiple radio dishes to, in effect, work like a large number of two-dish interferometres. 4.9

Light pollution: when the air scatters the glare from lights, producing an illumination that hides the faintest stars and limits the distances that can be probed by telescopes. 4.5

Magnification: the size of the image formed by the lens in a telescope. 4.4

Primary mirror: a concave mirror that serves as the main optical element in reflecting telescopes and that is curved like the inner surface of a sphere, reflecting light in order to form an image. 4.4

Prime focus: location near the front end of the telescope where the mirror reflects the light to form an image. 4.4

Radar is the technique of transmitting radio waves to an object in our solar system and then detecting the radio radiation that the object reflects back. 4.9

Refracting telescope: a telescope based on a refractor design. 4.4

Refractor: a long tube with a large glass lens at one end that is used as the main optical element to form an image. 4.4

Resolution: the precision of detail present in an image: that is, the smallest features that can be distinguished. 4.7

Secondary mirror: used in some reflecting telescopes to allow more light to get through the system. 4.4

Seeing: unsteadiness of Earth's atmosphere, which blurs telescopic images; good seeing means the atmosphere is steady. 4.5

Transients: phenomena that change quickly, such as exploding stars and chunks of rock that orbit near Earth. 4.11

CHAPTER 5: AN OVERVIEW OF THE SOLAR SYSTEM

Chapter Overview

5.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify the main components of the solar system, including the Sun, planets, moons, asteroids, comets, and dust.
- Explain how the solar system formed according to the nebular hypothesis, and describe the key processes involved, such as gravitational collapse, nuclear fusion in the Sun, and planetesimal formation.
- Compare and contrast the terrestrial planets (Mercury, Venus, Earth, and Mars) with the Jovian planets (Jupiter, Saturn, Uranus, and Neptune) based on their physical properties, orbits, and atmospheric composition.

5.1 AN INVENTORY OF THE SOLAR SYSTEM

The solar system consists of the Sun and many smaller objects: the planets, their moons and rings, and such "debris" as asteroids, comets, and dust. Decades of observation and spacecraft exploration have revealed that most of these objects formed together with the Sun about 4.5 billion years ago. They represent clumps of material that condensed from an enormous cloud of gas and dust. The central part of this cloud became the Sun, and a small fraction of the material in the outer parts eventually formed the other objects.

During the past 50 years, we have learned more about the solar system than anyone imagined before the space age. In addition to gathering information with powerful new telescopes, we have sent spacecraft directly to many members of the planetary system. (Planetary astronomy is the only branch of our science in which we can, at least vicariously, travel to the objects we want to study.) With evocative names such as Voyager, Pioneer, *Curiosity*, and Pathfinder, our robot explorers have flown past, orbited, or landed on every planet, returning images and data that have dazzled both astronomers and the public. In the process, we have also investigated two dwarf planets, hundreds of fascinating moons, four ring systems, a dozen asteroids, and several comets (smaller members of our solar system that we will discuss later).

Our probes have penetrated the atmosphere of Jupiter and landed on the surfaces of Venus, Mars, our Moon, Saturn's moon Titan, the asteroids Eros and Itokawa, and the Comet Churyumov-Gerasimenko (usually referred to as 67P). Humans have set foot on the Moon and returned samples of its surface soil for laboratory analysis as shown in Figure 5.1. We have even discovered other places in our solar system that might be able to support some kind of life.

Astronauts on the Moon

Figure 5.1. The lunar lander and surface rover from the Apollo 15 mission are seen in this view of the one place beyond Earth that has been explored directly by humans.
Apollo 15 flag, rover, LM, Irwin by NASA/David R. Scott, modified by Bammesk, NASA Media Licence.

View this gallery of NASA images that trace the history of the Apollo mission.

An Inventory

The Sun, a star that is brighter than about 80% of the stars in the Galaxy, is by far the most massive member of the solar system, as shown in Table 5.1. It is an enormous ball about 1.4 million kilometres in diameter, with surface layers of incandescent gas and an interior temperature of millions of degrees. The Sun will be discussed in later chapters as our first, and best-studied, example of a star.

Table 5.1. Mass of Members of the Solar System

Object	Percentage of Total Mass of Solar System
Sun	99.80
Jupiter	0.10
Comets	0.0005–0.03 (estimate)
All other planets and dwarf planets	0.04
Moons and rings	0.00005
Asteroids	0.000002 (estimate)
Cosmic dust	0.0000001 (estimate)

Table 5.1 also shows that most of the material of the planets is actually concentrated in the largest one, Jupiter, which is more massive than all the rest of the planets combined. Astronomers were able to determine the masses of the planets centuries ago using Kepler's laws of planetary motion and Newton's law of gravity to measure the planets' gravitational effects on one another or on moons that orbit them. Today, we make even more precise measurements of their masses by tracking their gravitational effects on the motion of spacecraft that pass near them.

Beside Earth, five other planets were known to the ancients—Mercury, Venus, Mars, Jupiter, and Saturn—and two were discovered after the invention of the telescope: Uranus and Neptune. The eight planets all revolve in the same direction around the Sun. They orbit in approximately the same plane, like cars travelling on concentric tracks on a giant, flat racecourse. Each planet stays in its own "traffic lane," following a nearly circular orbit about the Sun and obeying the "traffic" laws discovered by Galileo, Kepler, and Newton. Besides these planets, we have also been discovering smaller worlds beyond Neptune that are called trans-Neptunian objects or TNOs as shown in Figure 5.2. The first to be found, in 1930, was Pluto, but others have been discovered during the twenty-first century. One of them, Eris, is about the same size as Pluto and has at least one moon (Pluto has five known moons.) The largest TNOs are also classed as *dwarf planets,* as is the largest asteroid, Ceres. To date, more than 1750 of these TNOs have been discovered.

Orbits of the Planets

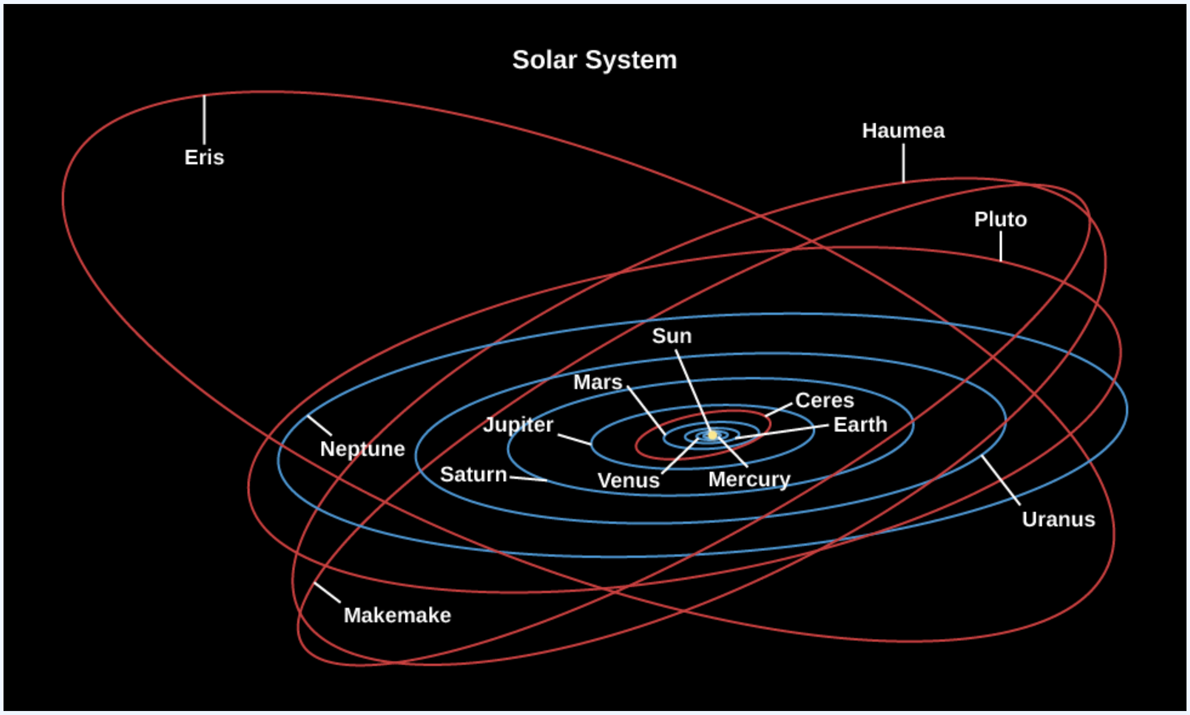


Figure 5.2. All eight major planets orbit the Sun in roughly the same plane. The five currently known dwarf planets are also shown: Eris, Haumea, Pluto, Ceres, and Makemake. Note that Pluto's orbit is not in the plane of the planets. Image by Open Stax CC BY 4.0

Each of the planets and dwarf planets also rotates (spins) about an axis running through it, and in most cases the direction of rotation is the same as the direction of revolution about the Sun. The exceptions are Venus, which rotates backward very slowly (that is, in a retrograde direction), and Uranus and Pluto, which also have strange rotations, each spinning about an axis tipped nearly on its side. We do not yet know the spin orientations of Eris, Haumea, and Makemake.

The four planets closest to the Sun (Mercury through Mars) are called the inner or **terrestrial planets**. Often, the Moon is also discussed as a part of this group, bringing the total of terrestrial objects to five. (We generally call Earth's satellite "the Moon," with a capital M, and the other satellites "moons," with lowercase m's.) The terrestrial planets are relatively small worlds, composed primarily of rock and metal. All of them have solid surfaces that bear the records of their geological history in the forms of craters, mountains, and volcanoes as shown in Figure 5.3.

Surface of Mercury

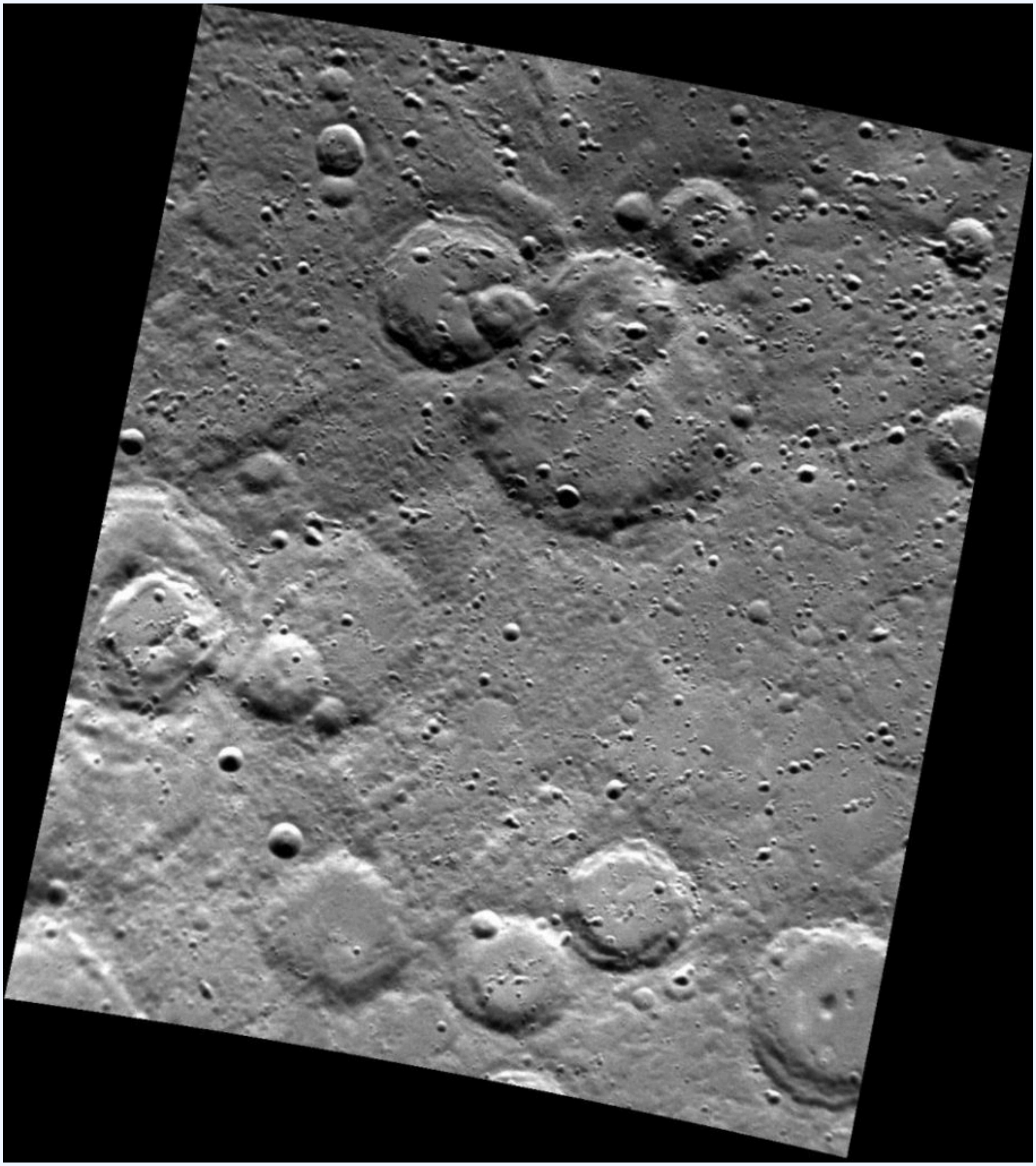


Figure 5.3. The pockmarked face of the terrestrial world of Mercury is more typical of the inner planets than the watery surface of Earth. This black-and-white image, taken with the Mariner 10 spacecraft, shows a region more than 400 kilometres wide.
PIA16757 by NASA/John Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington, JPL Media Licence.

The next four planets (Jupiter through Neptune) are much larger and are composed primarily of lighter ices, liquids, and gases. We call these four the jovian planets (after "Jove," another name for Jupiter in mythology) or giant planets—a name they richly deserve. They are shown in Figure 5.4. More than 1400 Earths could fit inside Jupiter, for example. These planets do not have solid surfaces on which future explorers might land. They are more like vast, spherical oceans with much smaller, dense cores.

The Four Giant Planets

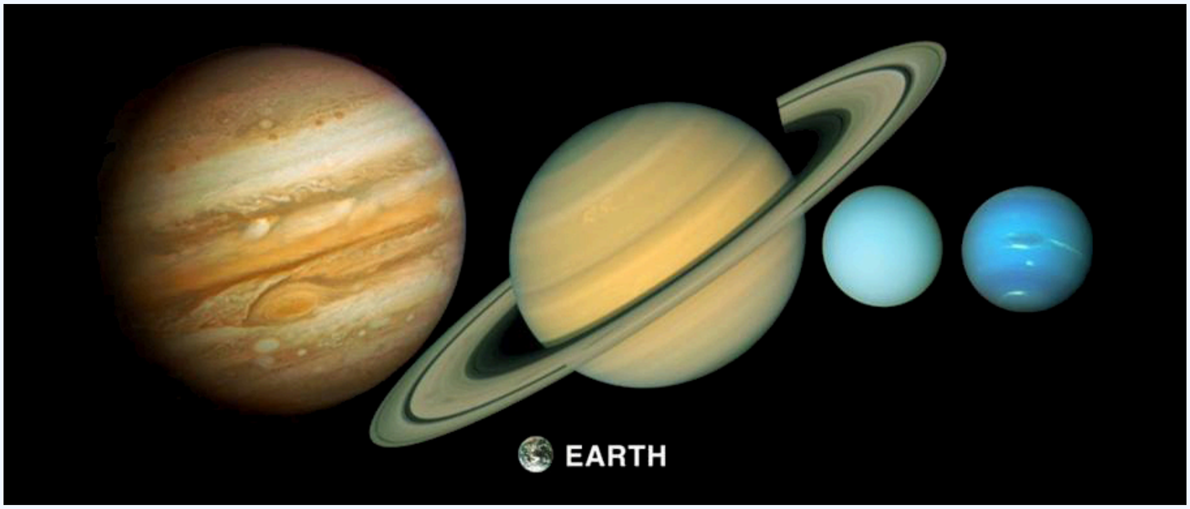


Figure 5.4. This montage shows the four giant planets: Jupiter, Saturn, Uranus, and Neptune. Below them, Earth is shown to scale.
First and Farthest by NASA, Solar System Exploration, NASA Media Licence.

Near the outer edge of the system lies Pluto, which was the first of the distant icy worlds to be discovered beyond Neptune (Pluto was visited by a spacecraft, the NASA New Horizons mission, in 2015. Figure 5.5 is the image it captured during its visit). Table 5.2 summarizes some of the main facts about the planets.

Pluto Close-up

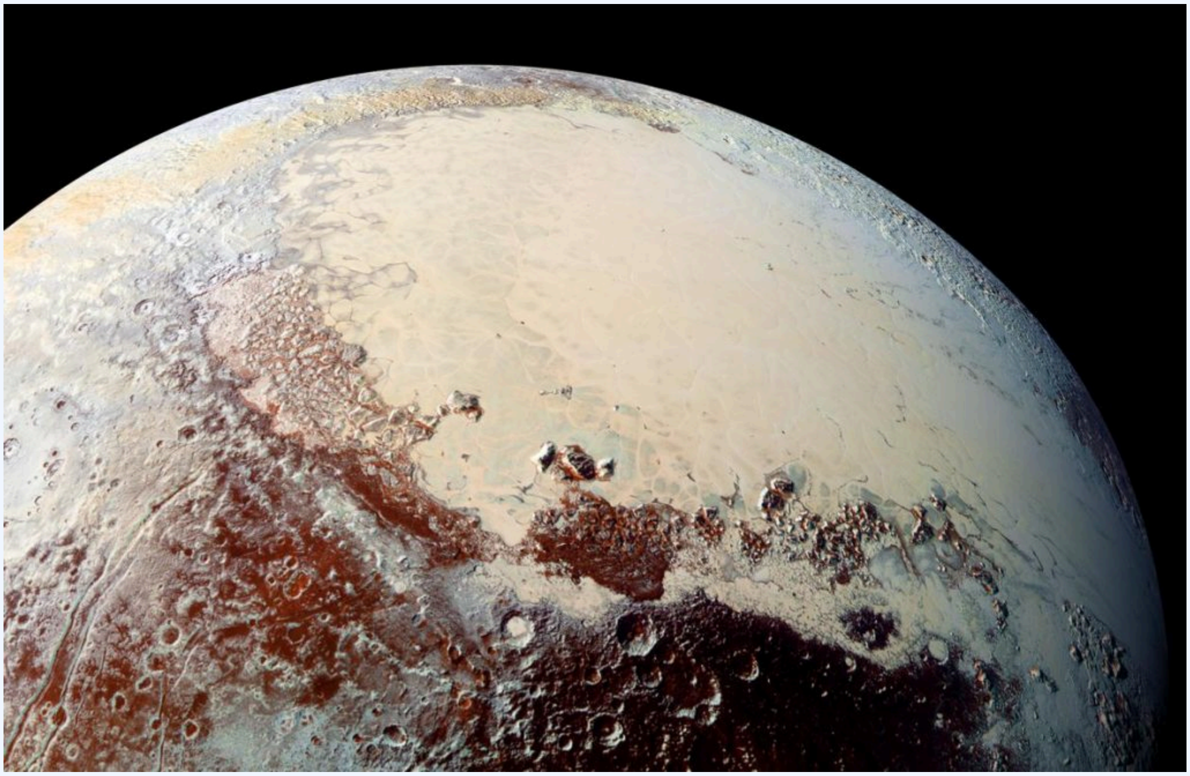


Figure 5.5. This intriguing image from the New Horizons spacecraft, taken when it flew by the dwarf planet in July 2015, shows some of its complex surface features. The rounded white area is temporarily being called the Sputnik Plain, after humanity's first spacecraft.
Sputnik Planum, in Color by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, JPL Media Licence.

Table 5.2. The Planets

Name	Distance from Sun	Revolution Period	Diameter	Mass	Density
	(AU)	(y)	(km)	(10^{23} kg)	(g/cm^3)
Mercury	0.39	0.24	4,878	3.3	5.4
Venus	0.72	0.62	12,120	48.7	5.2
Earth	1.00	1.00	12,756	59.8	5.5
Mars	1.52	1.88	6,787	6.4	3.9
Jupiter	5.20	11.86	142,984	18,991	1.3
Saturn	9.54	29.46	120,536	5686	0.7
Uranus	19.18	84.07	51,118	866	1.3
Neptune	30.06	164.82	49,660	1030	1.6

Example 5.1

Comparing Densities

Let's compare the densities of several members of the solar system. The density of an object equals its mass divided by its volume. The volume (V) of a sphere (like a planet) is calculated using the equation

$$V = \frac{4}{3}\pi R^3$$

where π (the Greek letter pi) has a value of approximately 3.14. Although planets are not perfect spheres, this equation works well enough. The masses and diametres of the planets are given in Table 5.2. Let's use Saturn's moon Mimas as our example, with a mass of $4 \times 10^1 9$ kg and a diameter of approximately 400 km (radius, $200 \text{ km} = 2 \times 10^5 \text{ m}$).

Solution

The volume of Mimas is

$$\frac{4}{3} \times 3.14 \times \left(2 \times 10^5 \text{ m}\right)^3 = 3.3 \times 10^{16} \text{ m}^3$$

Density is mass divided by volume:

$$\frac{4 \times 10^{19} \text{ kg}}{3.3 \times 10^{16} \text{ m}^3} = 1.2 \times 10^3 \text{ kg/m}^3$$

Note that the density of water in these units is 1000 kg/m^3, so Mimas must be made mainly of ice, not rock.

Exercise 5.1

Calculate the average density of our own planet, Earth. Show your work. How does it compare to the density of an ice moon like Mimas? See Table 5.2 for data.

Solution

For a sphere,

$$\text{density} = \frac{\text{mass}}{\left(\frac{4}{3}\pi R^3\right)} \text{kg/m}^3$$

For Earth, then,

$$\text{density} = \frac{6 \times 10^{24} \text{ kg}}{4.2 \times 2.6 \times 10^{20} \text{ m}^3} = 5.5 \times 10^3 \text{ kg/m}^3$$

This density is four to five times greater than Mimas'. In fact, Earth is the densest of the planets.

Learn more about NASA's mission to Pluto and see high-resolution images of Pluto's moon Charon.

Attribution

5.2 THE SCALE OF THE SOLAR SYSTEM

Astronomy often deals with dimensions and distances that far exceed our ordinary experience. What does 1.4 billion kilometres—the distance from the Sun to Saturn—really mean to anyone? It can be helpful to visualize such large systems in terms of a scale model.

In our imaginations, let us build a scale model of the solar system, adopting a scale factor of 1 billion (10^9)—that is, reducing the actual solar system by dividing every dimension by a factor of 10^9. Earth, then, has a diameter of 1.3 centimetres, about the size of a grape. The Moon is a pea orbiting this at a distance of 40 centimetres, or a little more than a foot away. The Earth-Moon system fits into a standard backpack.

In this model, the Sun is nearly 1.5 metres in diameter, about the average height of an adult, and our Earth is at a distance of 150 metres—about one city block—from the Sun. Jupiter is five blocks away from the Sun, and its diameter is 15 centimetres, about the size of a very large grapefruit. Saturn is 10 blocks from the Sun; Uranus, 20 blocks; and Neptune, 30 blocks. Pluto, with a distance that varies quite a bit during its 249-year orbit, is currently just beyond 30 blocks and getting farther with time. Most of the moons of the outer solar system are the sizes of various kinds of seeds orbiting the grapefruit, oranges, and lemons that represent the outer planets.

In our scale model, a human is reduced to the dimensions of a single atom, and cars and spacecraft to the size of molecules. Sending the Voyager spacecraft to Neptune involves navigating a single molecule from the Earth–grape toward a lemon 5 kilometres away with an accuracy equivalent to the width of a thread in a spider's web.

If that model represents the solar system, where would the nearest stars be? If we keep the same scale, the closest stars would be tens of thousands of kilometres away. If you built this scale model in the city where you live, you would have to place the representations of these stars on the other side of Earth or beyond.

Size Comparison of the Sun and Planets

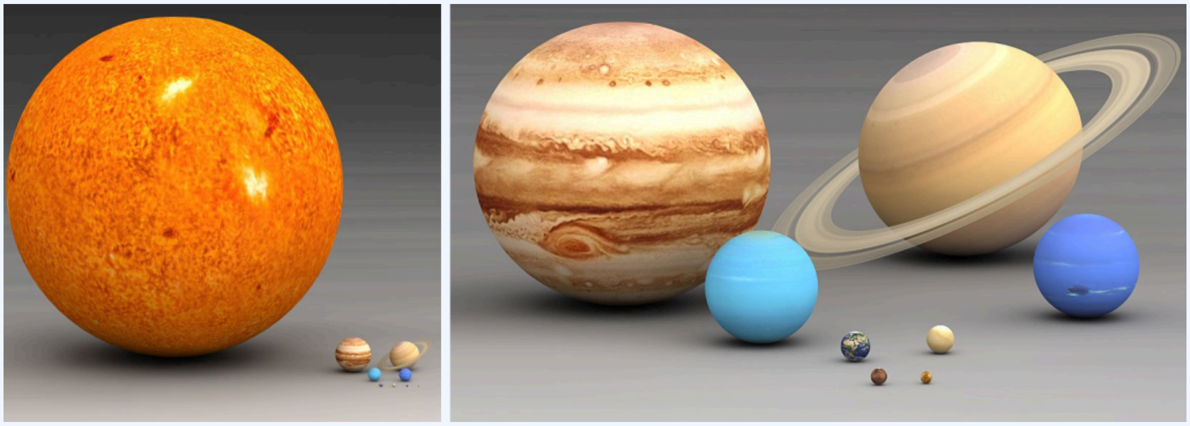


Figure 5.6. The left panel shows the sun as compared to the planets, while the right panel shows the planets only.
Planets and Sun size comparison by Lsmpascal, CC BY-SA 3.0.

By the way, model solar systems like the one we just presented have been built in cities throughout the world. In Sweden, for example, Stockholm's huge Globe Arena has become a model for the Sun, and Pluto is represented by a 12-centimetere sculpture in the small town of Delsbo, 300 kilometres away. Another model solar system is in Washington on the Mall between the White House and Congress (perhaps proving they are worlds apart?).

Names in the Solar System

We humans just don't feel comfortable until something has a name. Types of butterflies, new elements, and the mountains of Venus all need names for us to feel we are acquainted with them. How do we give names to objects and features in the solar system?

Planets and moons are named after gods and heroes in Greek and Roman mythology (with a few exceptions among the moons of Uranus, which have names drawn from English literature). When William Herschel, a German immigrant to England, first discovered the planet we now call Uranus, he wanted to name it Georgium Sidus (George's star) after King George III of his adopted country. This caused such an outcry among astronomers in other nations, however, that the classic tradition was upheld—and has been maintained ever since. Luckily, there were a lot of minor gods in the ancient pantheon, so plenty of names are left for the many small moons we are discovering around the giant planets.

Comets are often named after their discoverers (offering an extra incentive to comet hunters). Asteroids are named by their discoverers after just about anyone or anything they want. Recently, asteroid names have been used to recognize people who have made significant contributions to astronomy, including the three original authors of this book.

That was pretty much all the naming that was needed while our study of the solar system was confined to Earth. But now, our spacecraft have surveyed and photographed many worlds in great detail, and each world has a host of features that also need names. To make sure that naming things in space remains multinational, rational, and somewhat dignified, astronomers have given the responsibility of approving names to a special committee of the International Astronomical Union (IAU), the body that includes scientists from every country that does astronomy.

This IAU committee has developed a set of rules for naming features on other worlds. For example, craters on Venus are named for women who have made significant contributions to human knowledge and welfare. Volcanic features on Jupiter's moon Io, which is in a constant state of volcanic activity, are named after gods of fire and thunder from the mythologies of many cultures. Craters on Mercury commemorate famous novelists, playwrights, artists, and composers. On Saturn's moon Tethys, all the features are named after characters and places in Homer's great epic poem, *The Odyssey*. As we explore further, it may well turn out that more places in the solar system need names than Earth history can provide. Perhaps by then, explorers and settlers on these worlds will be ready to develop their own names for the places they may (if but for a while) call home.

You may be surprised to know that the meaning of the word *planet* has recently become controversial because we have discovered many other planetary systems that don't look very much like our own. Even within our solar system, the planets differ greatly in size and chemical properties. The biggest dispute concerns Pluto, which is much smaller than the other eight major planets. The category of dwarf planet was invented to include Pluto and similar icy objects beyond Neptune. But is a dwarf planet also a planet? Logically, it should be, but even this simple issue of grammar has been the subject of heated debate among both astronomers and the general public.

Attribution

5.3 FORMATION OF THE SOLAR SYSTEM

The most widely accepted explanation of how the solar system formed is called the **nebular hypothesis**. According to this hypothesis, the Sun and the planets of our solar system formed about 4.6 billion years ago from the collapse of a giant cloud of gas and dust, called a **nebula**.

The nebula was drawn together by gravity, which released gravitational potential energy. As small particles of dust and gas smashed together to create larger ones, they released kinetic energy. As the nebula collapsed, the gravity at the center increased, and the cloud started to spin because of its angular momentum. As it collapsed further, the spinning got faster, much as an ice skater spins faster when he pulls his arms to his sides during a spin.

Much of the cloud's mass migrated to its center, but the rest of the material flattened out in an enormous disk. The disk contained hydrogen and helium, along with heavier elements and even simple organic molecules.

As gravity pulled matter into the center of the disk, the density and pressure at the center became intense. When the pressure in the center of the disk was high enough, nuclear fusion within our star began, and the blazing star stopped the disk from collapsing further.

Meanwhile, the outer parts of the disk were cooling off. Matter condensed from the cloud, and small pieces of dust started clumping together to create ever bigger clumps of matter. Larger clumps, called **planetesimals**, attracted smaller clumps with their gravity. Gravity at the center of the disk attracted more massive particles, such as rock and metal, and lighter particles remained further out in the disk. Eventually, the planetesimals formed **protoplanets**, which grew to become the planets and moons that we find in our solar system today.

The gravitational sorting of material with the inner planets, Mercury, Venus, Earth, and Mars, dense rock and metal formed. The outer planets, Jupiter, Saturn, Uranus, and Neptune, condensed farther from the Sun from lighter materials such as hydrogen, helium, water, ammonia, and methane. Out by Jupiter and beyond, where it is frigid, these materials formed solid particles.

The nebular hypothesis was designed to explain some of the essential features of the solar system:

- Orbits of the planets lie in nearly the same plane with the Sun at the center
- Planets revolve in the same direction
- Planets mostly rotate in the same direction
- Axes of rotation of the planets are mostly nearly perpendicular to the orbital plane
- Oldest moon rocks are 4.5 billion years

Attribution

5.4 KEY TERMS

Nebula: a giant cloud of gas and dust. 5.3

Nebular hypothesis: the Sun and the planets of our solar system formed about 4.6 billion years ago from the collapse of a giant cloud of gas and dust, called a nebula. 5.3

Planetesimals: large clumps of matter that attracted smaller clumps with their gravity. 5.3

Protoplanets: former planetesimals formed through the gravitational attraction of more massive particles, such as rock and metal. 5.3

Terrestrial planet: any of the planets Mercury, Venus, Earth, or Mars; sometimes the Moon is included in the list. 5.1

CHAPTER 6: THE TERRESTRIAL PLANETS

Chapter Overview

6.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Define the characteristics that distinguish terrestrial planets from giant planets based on size, composition, and the abundance of elements in their structure.
- Compare and contrast the composition and features of the terrestrial planets.
- Discuss the giant impact hypothesis as the most plausible explanation for the origin of the Moon.

6.1 INTRODUCTION

The terrestrial planets are quite different from the giants. In addition to being much smaller, they are composed primarily of rocks and metals. These, in turn, are made of elements that are less common in the universe as a whole. The most abundant rocks, called silicates, are made of silicon and oxygen, and the most common metal is iron. We can tell from their densities that Mercury has the greatest proportion of metals (which are denser) and the Moon has the lowest. Earth, Venus, and Mars all have roughly similar bulk compositions: about one third of their mass consists of iron-nickel or iron-sulfur combinations; two thirds is made of silicates. Because these planets are largely composed of oxygen compounds (such as the silicate minerals of their crusts), their chemistry is said to be *oxidized*.

When we look at the internal structure of each of the terrestrial planets, we find that the densest metals are in a central core, with the lighter silicates near the surface. If these planets were liquid, like the giant planets, we could understand this effect as the result the sinking of heavier elements due to the pull of gravity. This leads us to conclude that, although the terrestrial planets are solid today, at one time they must have been hot enough to melt.

Differentiation is the process by which gravity helps separate a planet's interior into layers of different compositions and densities. The heavier metals sink to form a core, while the lightest minerals float to the surface to form a crust. Later, when the planet cools, this layered structure is preserved. In order for a rocky planet to differentiate, it must be heated to the melting point of rocks, which is typically more than 1300 K.

Attribution

6.2 EARTH

Earth has vast oceans of liquid water, large masses of exposed land, and a dynamic atmosphere with clouds of water vapour. Earth also has ice covering its polar regions. Earth's average surface temperature is 14 degrees Celsius (57 degrees Fahrenheit). Water is a liquid at this temperature, but the planet also has water in its other two states, solid and gas. The oceans and the atmosphere help keep Earth's surface temperatures reasonably steady.

Cross-Section of Earth

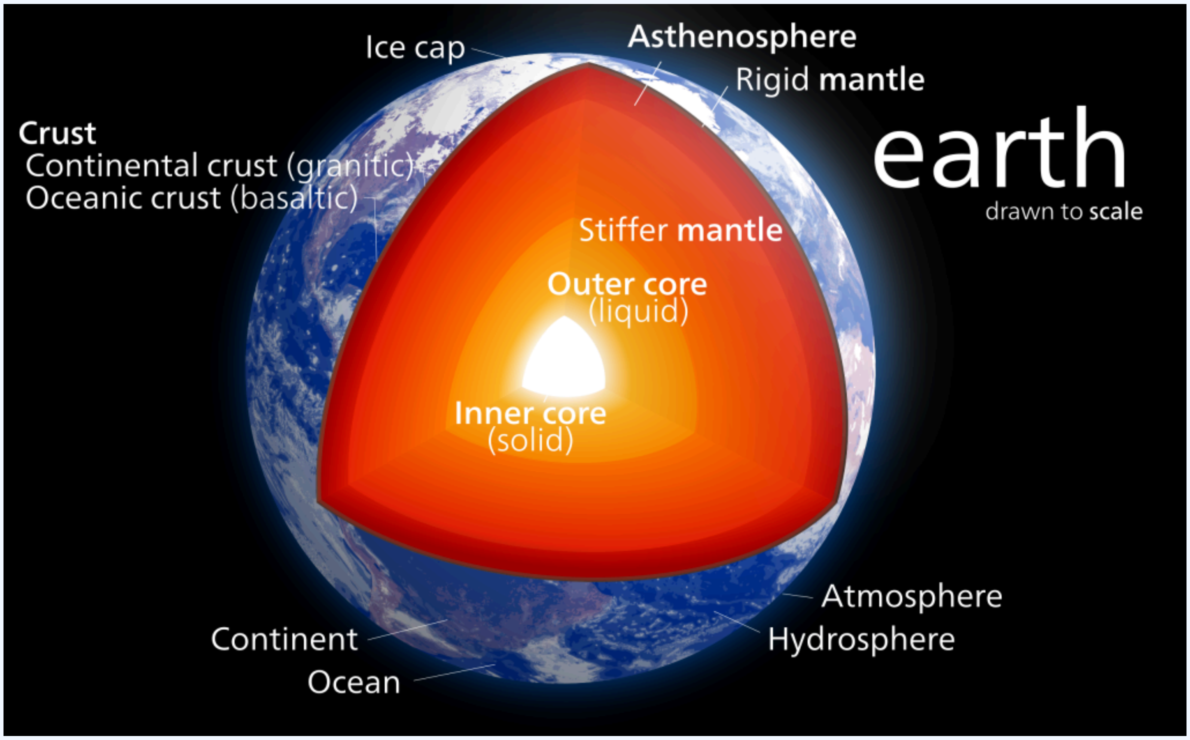


Figure 6.1. Internal structure of the Earth.
Earth by Kelvinsong, CC BY-SA 3.0.

Earth is the only planet known to have known life. The presence of liquid water, the atmosphere's ability to filter out harmful radiation, and many other features make the planet uniquely suited to harbour life. Life and

Earth now affect each other; for example, the evolution of plants allowed oxygen to enter the atmosphere in large enough quantities for animals to evolve. Although life has not been found elsewhere in the solar system, other planets or satellites may harbour primitive life forms. Life may also be found elsewhere in the universe.

The heat that remained from the planet's accretion, gravitational compression, and radioactive decay allowed the Earth to melt, probably more than once. As it subsequently cooled, gravity pulled metal into the centre to create the core. Heavier rocks formed the mantle, and lighter rocks formed the crust.

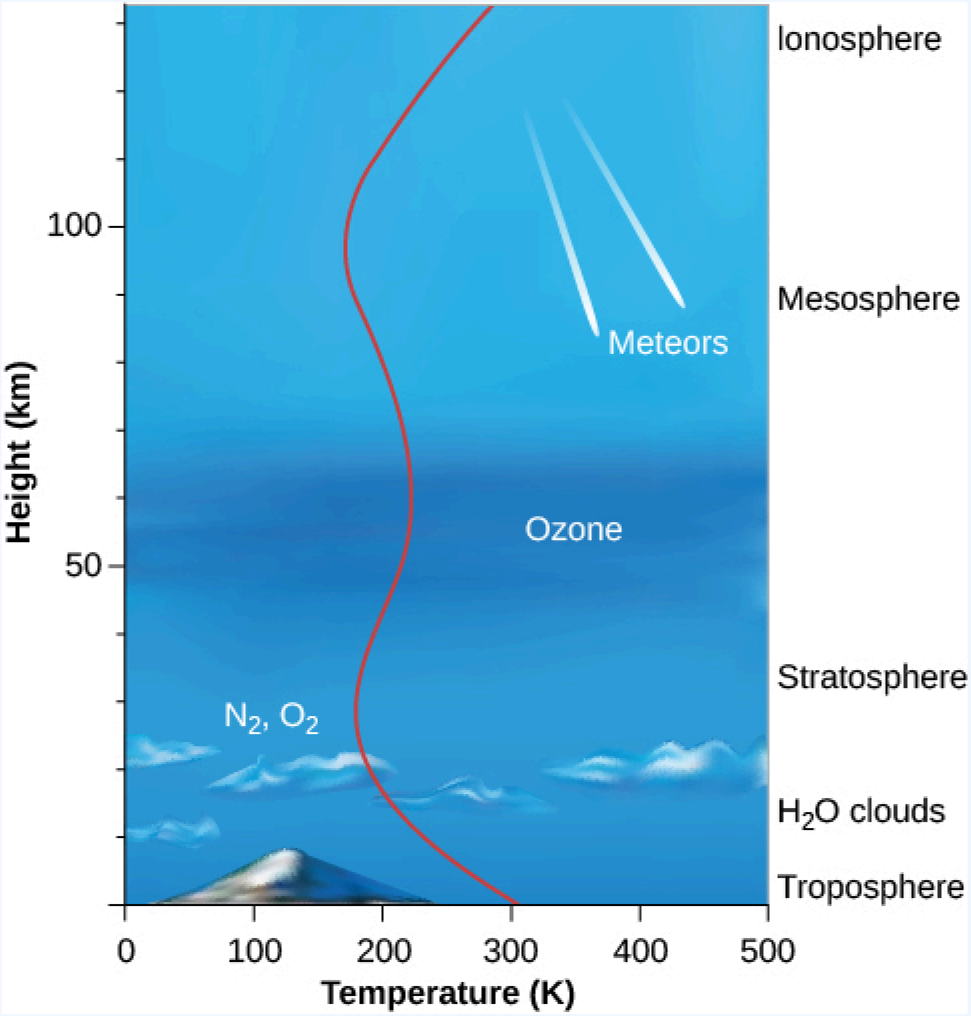
Earth's crust is divided into tectonic plates, which move around on the surface because of the convecting mantle below. The plates' movement causes other geological activity, such as earthquakes, volcanoes, and mountains. The locations of these features are mostly related to current or former plate boundaries. Earth is the only planet known to have plate tectonics.

Earth rotates on its axis once per day, by definition. Earth orbits the Sun once every 365.24 days, which is defined as a year. Earth has one large moon, which orbits Earth once every 29.5 days, a period known as a month.

We live at the bottom of the ocean of air that envelops our planet. The atmosphere, weighing down upon Earth's surface under the force of gravity, exerts a pressure at sea level that scientists define as 1 bar (a term that comes from the same root as *barometer*, an instrument used to measure atmospheric pressure). A **bar** of pressure means that each square centimetre of Earth's surface has a weight equivalent to 1.03 kilograms pressing down on it. Humans have evolved to live at this pressure; make the pressure a lot lower or higher and we do not function well.

The total mass of Earth's atmosphere is about 5×10^{18} kilograms. This sounds like a large number, but it is only about a millionth of the total mass of Earth. The atmosphere represents a smaller fraction of Earth than the fraction of your mass represented by the hair on your head.

The structure of the atmosphere is illustrated in the Figure 6.2. Most of the atmosphere is concentrated near the surface of Earth, within about the bottom 10 kilometres where clouds form and airplanes fly. Within this region—called the troposphere—warm air, heated by the surface, rises and is replaced by descending currents of cooler air; this is an example of convection. This circulation generates clouds and wind. Within the troposphere, temperature decreases rapidly with increasing elevation to values near 50 °C below freezing at its upper boundary, where the stratosphere begins. Most of the stratosphere, which extends to about 50 kilometres above the surface, is cold and free of clouds.



Structure of Earth's Atmosphere

Figure 6.2. Height increases up the left side of the diagram, and the names of the different atmospheric layers are shown at the right. In the upper ionosphere, ultraviolet radiation from the Sun can strip electrons from their atoms, leaving the atmosphere ionized. The curving red line shows the temperature (see the scale on the x-axis).

Near the top of the **stratosphere** is a layer of **ozone** (O_3), a heavy form of oxygen with three atoms per molecule instead of the usual two. Because ozone is a good absorber of ultraviolet light, it protects the surface from some of the Sun's dangerous ultraviolet radiation, making it possible for life to exist on Earth. The breakup of ozone adds heat to the stratosphere, reversing the decreasing temperature trend in the troposphere.

Because ozone is essential to our survival, we reacted with justifiable concern to evidence that became clear in the 1980s that atmospheric ozone was being destroyed by human activities. By international agreement, the production of industrial chemicals that cause ozone depletion, called chlorofluorocarbons, or CFCs, has been phased out. As a result, ozone loss has stopped and the "ozone hole" over the Antarctic is shrinking gradually. This is an example of how concerted international action can help maintain the habitability of Earth.

At heights above 100 kilometres, the atmosphere is so thin that orbiting satellites can pass through it with very little friction. Many of the atoms are ionized by the loss of an electron, and this region is often called the ionosphere. At these elevations, individual atoms can occasionally escape completely from the gravitational field of Earth. There is a continuous, slow leaking of atmosphere—especially of lightweight atoms, which move faster than heavy ones. Earth's atmosphere cannot, for example, hold on for long to hydrogen or helium, which escape into space. Earth is not the only planet to experience atmosphere leakage. Atmospheric leakage also created Mars' thin atmosphere. Venus' dry atmosphere evolved because its proximity to the Sun vaporized and dissociated any water, with the component gases lost to space.

Attribution

6.3 THE MOON: SURFACE, INTERIOR, AND ORIGIN

The Moon has only one-eightieth the mass of Earth and about one-sixth Earth's surface gravity—too low to retain an atmosphere. The Moon may be seen in Figure 6.3. Moving molecules of a gas can escape from a planet just the way a rocket does, and the lower the gravity, the easier it is for the gas to leak away into space. While the Moon can acquire a temporary atmosphere from impacting comets, this atmosphere is quickly lost by freezing onto the surface or by escape to surrounding space. The Moon today is dramatically deficient in a wide range of **volatiles**, those elements and compounds that evaporate at relatively low temperatures. Some of the Moon's properties are summarized in Table 6.1, along with comparative values for Mercury.

Two Sides of the Moon

Figure 6.3. The left image shows part of the hemisphere that faces Earth; several dark maria are visible. The right image shows part of the hemisphere that faces away from Earth; it is dominated by highlands. The resolution of this image is several kilometres, similar to that of high-powered binoculars or a small telescope.
The Moon's near side and far side. by NASA LRO / Jatan Mehta , NASA Media Licence.

Table 6.1. Properties of the Moon and Mercury

Property	Moon	Mercury
Mass (Earth = 1)	0.0123	0.055
Diameter (km)	3476	4878
Density (g/cm^3)	3.3	5.4
Surface gravity (Earth = 1)	0.17	0.38
Escape velocity (km/s)	2.4	4.3
Rotation period (days)	27.3	58.65
Surface area (Earth = 1)	0.27	0.38

If you look at the Moon through a telescope, you can see that it is covered by impact craters of all sizes. The most conspicuous of the Moon's surface features—those that can be seen with the unaided eye and that make up the feature often called "the man in the Moon"—are vast splotches of darker lava flows.

Centuries ago, early lunar observers thought that the Moon had continents and oceans and that it was a possible abode of life. They called the dark areas "seas" (*maria* in Latin, or *mare* in the singular, pronounced "mah ray"). Their names, Mare Nubium (Sea of Clouds), Mare Tranquillitatis (Sea of Tranquility), and so on, are still in use today. In contrast, the "land" areas between the seas are not named. Thousands of individual craters have been named, however, mostly for great scientists and philosophers such as the one in Figure 6.4. Among the most prominent craters are those named for Plato, Copernicus, Tycho, and Kepler. Galileo only has a small crater, however, reflecting his low standing among the Vatican scientists who made some of the first lunar maps.

We know today that the resemblance of lunar features to terrestrial ones is superficial. Even when they look somewhat similar, the origins of lunar features such as craters and mountains are very different from their terrestrial counterparts. The Moon's relative lack of internal activity, together with the absence of air and water, make most of its geological history unlike anything we know on Earth.

Sunrise on the Central Mountain Peaks of Tycho Crater, as Imaged by the NASA Lunar Reconnaissance Orbiter.

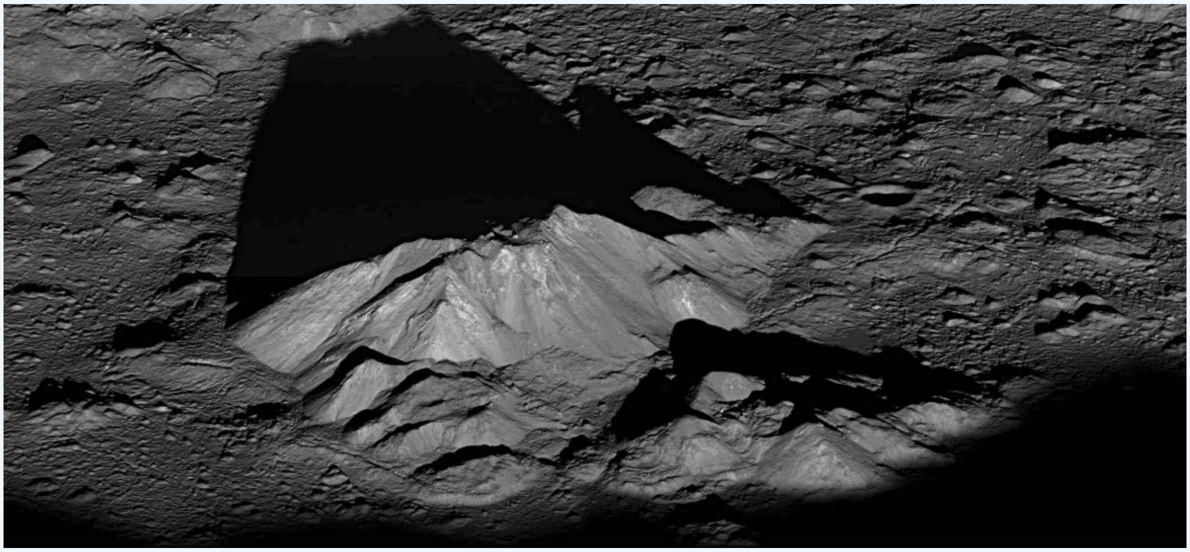


Figure 6.4. Tycho, about 82 kilometres in diameter, is one of the youngest of the very large lunar craters. The central mountain rises 12 kilometres above the crater floor.
Tycho Crater's Central Peak on the Moon by NASA Goddard/Arizona State University, NASA Media Licence.

Lunar History

To trace the detailed history of the Moon or of any planet, we must be able to estimate the ages of individual rocks. Once lunar samples were brought back by the Apollo astronauts, the radioactive dating techniques that had been developed for Earth were applied to them. The solidification ages of the samples ranged from about 3.3 to 4.4 billion years old, substantially older than most of the rocks on Earth.

Most of the crust of the Moon (83%) consists of silicate rocks called **anorthosites**; these regions are known as the **lunar highlands**. They are made of relatively low-density rock that solidified on the cooling Moon like slag floating on the top of a smelter. Because they formed so early in lunar history (between 4.1 and 4.4 billion years ago), the highlands are also extremely heavily cratered, bearing the scars of all those billions of years of impacts by interplanetary debris as shown in Figure 6.5.

Lunar Highlands

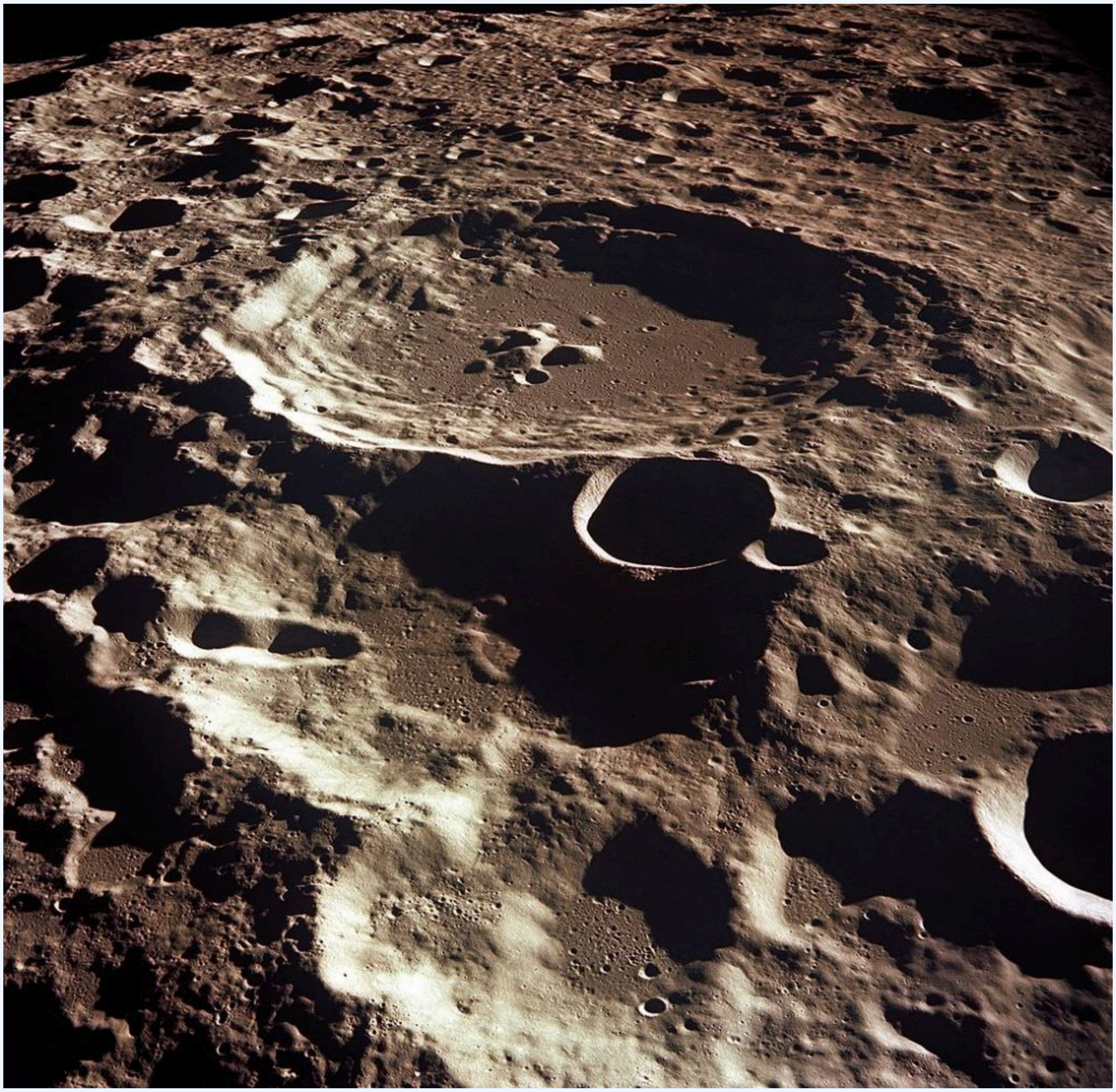


Figure 6.5. The old, heavily cratered lunar highlands make up 83% of the Moon's surface.
Crater Daedalus by NASA Goddard/Arizona State University, NASA Media Licence.

Unlike the mountains on Earth, the Moon's highlands do not have any sharp folds in their ranges. The highlands have low, rounded profiles that resemble the oldest, most eroded mountains on Earth as shown in Figure 6.6. Because there is no atmosphere or water on the Moon, there has been no wind, water, or ice to carve them into cliffs and sharp peaks, the way we have seen them shaped on Earth. Their smooth features are attributed to gradual erosion, mostly due to impact cratering from meteorites.

Lunar Mountain

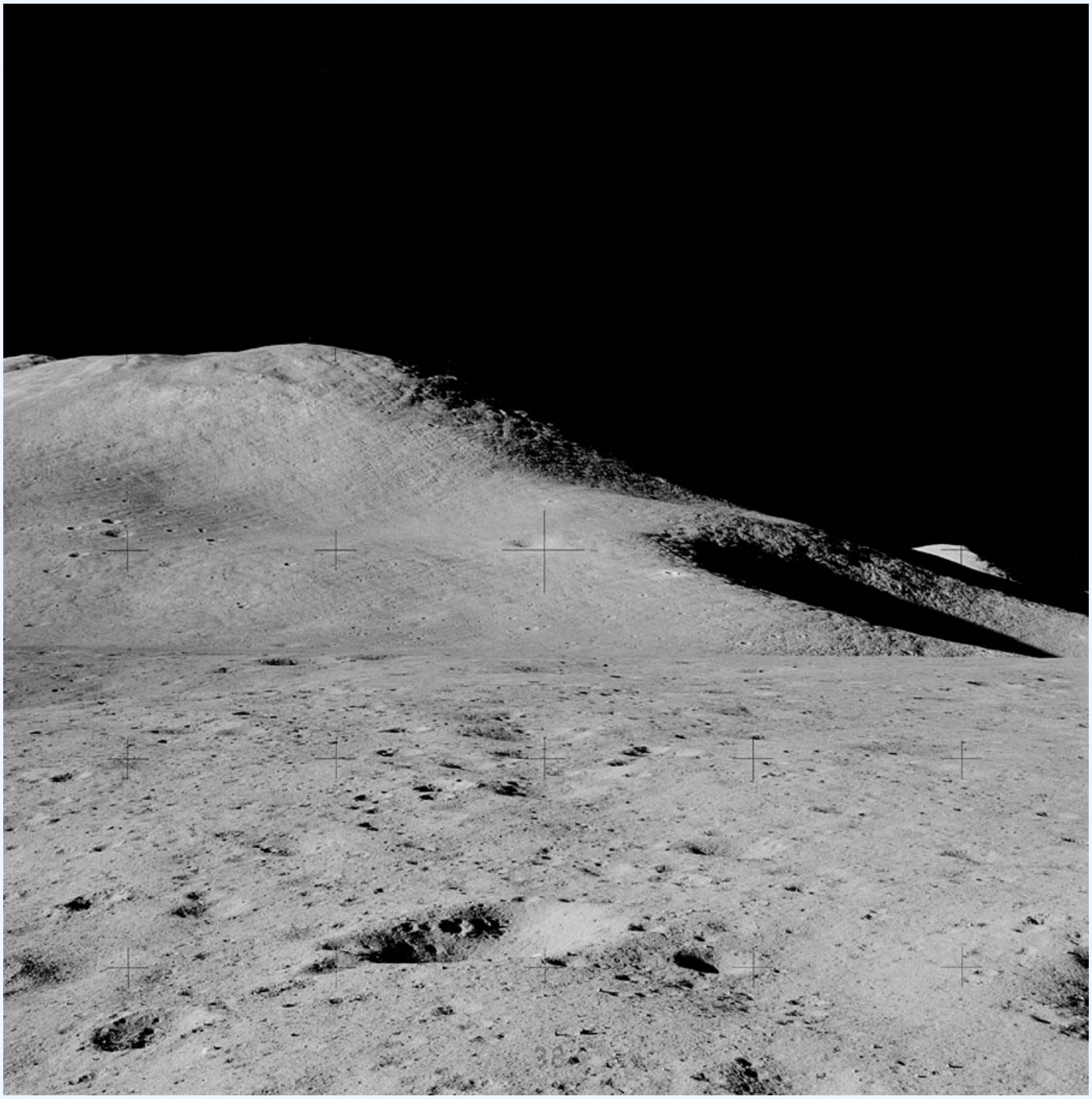


Figure 6.6. This photo of Mt. Hadley on the edge of Mare Imbrium was taken by Dave Scott, one of the Apollo 15 astronauts. Note the smooth contours of the lunar mountains, which have not been sculpted by water or ice.
Apollo 15 Mt. Hadley Delta by NASA/Apollo 15, NASA Media Licence.

The **maria** are much less cratered than the highlands, and cover just 17% of the lunar surface, mostly on the side of the Moon that faces Earth. The maria are pictured in Figure 6.7.

Lunar Maria

Figure 6.7. About 17% of the Moon's surface consists of the maria—flat plains of basaltic lava. This view of Mare Imbrium also shows numerous secondary craters and evidence of material ejected from the large crater Copernicus on the upper horizon. Copernicus is an impact crater almost 100 kilometres in diameter that was formed long after the lava in Imbrium had already been deposited.
Mare Imbrium of Copernicus crater on the Moon by NASA/Apollo 17, NASA Media Licence.

Today, we know that the maria consist mostly of dark-coloured basalt (volcanic lava) laid down in volcanic eruptions billions of years ago. Eventually, these lava flows partly filled the huge depressions called **impact basins**, which had been produced by collisions of large chunks of material with the Moon relatively early in its history. The basalt on the Moon, as shown in Figure 6.8, is very similar in composition to the crust under the oceans of Earth or to the lavas erupted by many terrestrial volcanoes. The youngest of the lunar impact basins is Mare Orientale, shown in Figure 6.9.

Rock from a Lunar Mare

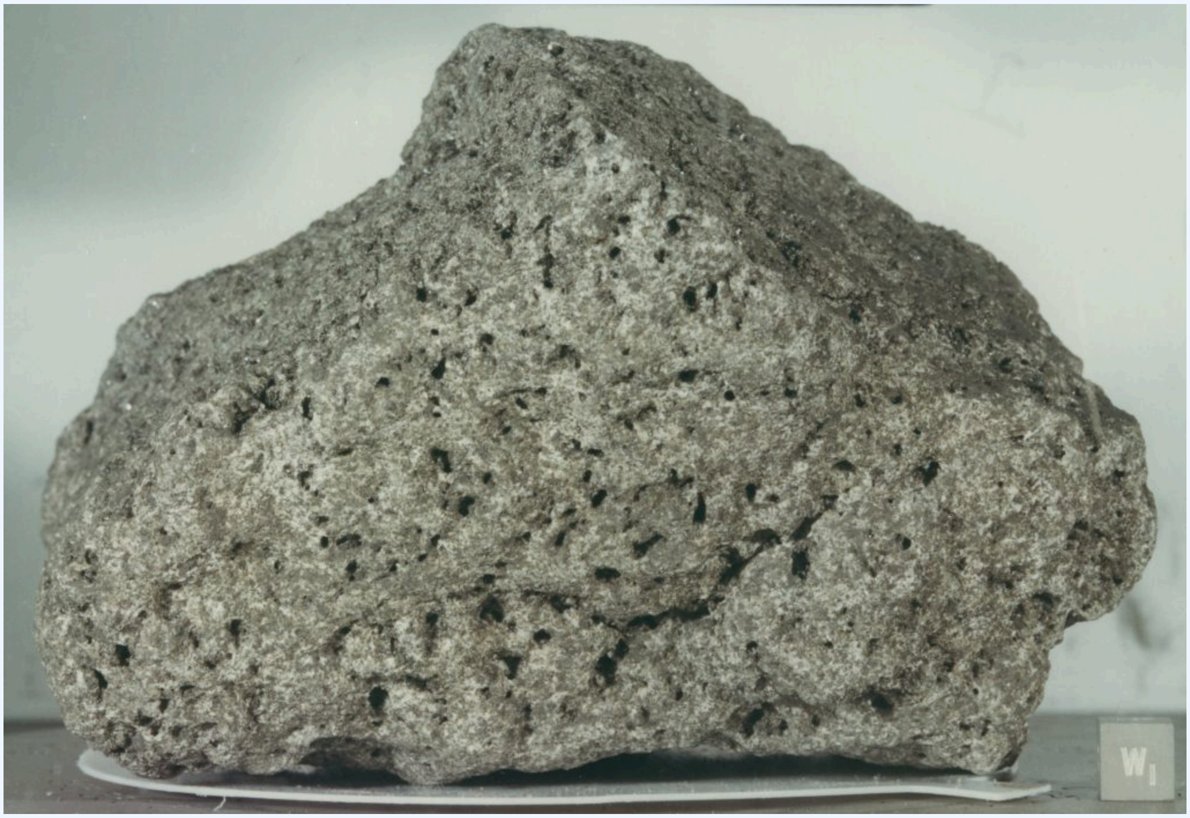


Figure 6.8. In this sample of basalt from the mare surface, you can see the holes left by gas bubbles, which are characteristic of rock formed from lava. All lunar rocks are chemically distinct from terrestrial rocks, a fact that has allowed scientists to identify a few lunar samples among the thousands of meteorites that reach Earth. Sample is about 10 cm across. Cube is 1 cm.
Lunar basalt 70017 by NASA/Apollo 17, NASA Media Licence.

Mare Orientale

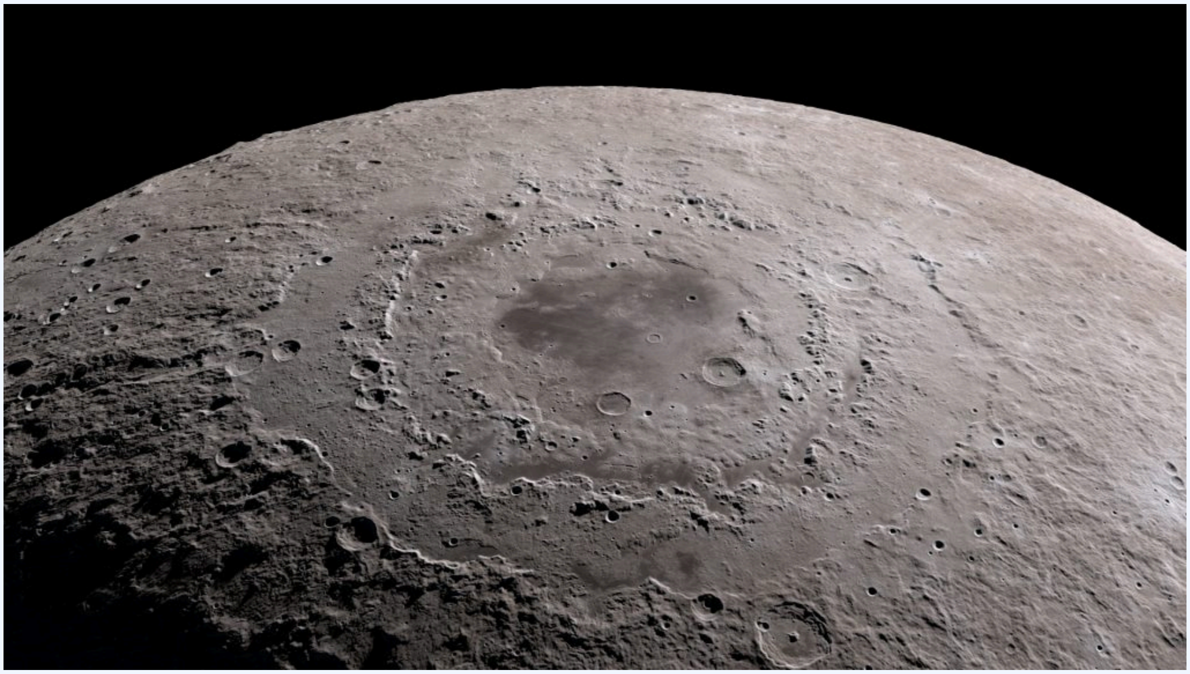


Figure 6.9. The youngest of the large lunar impact basins is Orientale, formed 3.8 billion years ago. Its outer ring is about 1000 kilometres in diameter, roughly the distance between New York City and Detroit, Michigan. Unlike most of the other basins, Orientale has not been completely filled in with lava flows, so it retains its striking "bull's-eye" appearance. It is located on the edge of the Moon as seen from Earth. Orientale Basin by NASA Goddard Space Flight Center, CC BY 2.0.

Volcanic activity may have begun very early in the Moon's history, although most evidence of the first half billion years is lost. What we do know is that the major mare volcanism, which involved the release of lava from hundreds of kilometres below the surface, ended about 3.3 billion years ago. After that, the Moon's interior cooled, and volcanic activity was limited to a very few small areas. The primary forces altering the surface come from the outside, not the interior.

Composition and Structure of the Moon

The composition of the Moon is not the same as that of Earth. With an average density of only 3.3 g/cm^3, the Moon must be made almost entirely of silicate rock. Compared to Earth, it is depleted in iron and other metals. It is as if the Moon were composed of the same silicates as Earth's mantle and crust, with the metals and the volatiles selectively removed. These differences in composition between Earth and Moon provide important clues about the origin of the Moon, a topic we will cover in detail later in this chapter.

Studies of the Moon's interior carried out with seismometres taken to the Moon as part of the Apollo program confirm the absence of a large metal core. The twin GRAIL spacecraft launched into lunar orbit in 2011 provided even more precise tracking of the interior structure. We also know from the study of lunar samples that water and other volatiles have been depleted from the lunar crust. The tiny amounts of water detected in these samples were originally attributed to small leaks in the container seal that admitted water vapor from Earth's atmosphere. However, scientists have now concluded that some chemically bound water is present in the lunar rocks.

Most dramatically, water ice has been detected in permanently shadowed craters near the lunar poles. In 2009, NASA crashed a small spacecraft called the Lunar Crater Observation and Sensing Satellite (LCROSS) into the crater Cabeus near the Moon's south pole. The impact at 9,000 kilometres per hour released energy equivalent to 2 tons of dynamite, blasting a plume of water vapor and other chemicals high above the surface. This plume was visible to telescopes in orbit around the Moon, and the LCROSS spacecraft itself made measurements as it flew through the plume. A NASA spacecraft called the Lunar Reconnaissance Orbiter (LRO) also measured the very low temperatures inside several lunar craters, and its sensitive cameras were even able to image crater interiors by starlight.

The total quantity of water ice in the Moon's polar craters is estimated to be hundreds of billions of tons. As liquid, this would only be enough water to fill a lake 100 miles across, but compared with the rest of the dry lunar crust, so much water is remarkable. Presumably, this polar water was carried to the Moon by comets and asteroids that hit its surface. Some small fraction of the water froze in a few extremely cold regions (cold traps) where the Sun never shines, such as the bottom of deep craters at the Moon's poles. One reason this discovery could be important is that it raises the possibility of future human habitation near the lunar poles, or even of a lunar base as a way-station on routes to Mars and the rest of the solar system. If the ice could be mined, it would yield both water and oxygen for human support, and it could be broken down into hydrogen and oxygen, a potent rocket fuel.

It is characteristic of modern science to ask how things originated. Understanding the origin of the Moon has proven to be challenging for planetary scientists, however. Part of the difficulty is simply that we know so much about the Moon (quite the opposite of our usual problem in astronomy). As we will see, one key problem is that the Moon is both tantalizingly similar to Earth and frustratingly different.

Early Ideas for the Origin of the Moon

Most of the earlier hypotheses for the Moon's origin followed one of three general ideas:

1. The fission hypothesis—the Moon was once part of Earth, but somehow separated from it early in their history.

2. The sister hypothesis—the Moon formed together with (but independent of) Earth, as we believe many moons of the outer planets formed.
3. The capture hypothesis—the Moon formed elsewhere in the solar system and was captured by Earth.

Unfortunately, there seem to be fundamental problems with each of these ideas. Perhaps the easiest hypothesis to reject is the capture hypothesis. Its primary drawback is that no one knows of any way that early Earth could have captured such a large moon from elsewhere. One body approaching another cannot go into orbit around it without a substantial loss of energy; this is the reason that spacecraft destined to orbit other planets are equipped with retro-rockets. Furthermore, if such a capture did take place, the captured object would go into a very eccentric orbit rather than the nearly circular orbit our Moon occupies today. Finally, there are too many compositional similarities between Earth and the Moon, particularly an identical fraction of the major isotopes[1] of oxygen, to justify seeking a completely independent origin.

The fission hypothesis, which states that the Moon separated from Earth, was suggested in the late nineteenth century. Modern calculations have shown that this sort of spontaneous fission or splitting is impossible. Furthermore, it is difficult to understand how a Moon made out of terrestrial material in this way could have developed the many distinctive chemical differences now known to characterize our neighbor.

Scientists were therefore left with the sister hypothesis—that the Moon formed alongside Earth—or with some modification of the fission hypothesis that can find a more acceptable way for the lunar material to have separated from Earth. But the more we learned about our Moon, the less these old ideas seem to fit the bill.

The Giant Impact Hypothesis – Currently accepted

In an effort to resolve these apparent contradictions, scientists developed a fourth hypothesis for the origin of the Moon, one that involves a giant impact early in Earth's history. There is increasing evidence that large chunks of material—objects of essentially planetary mass—were orbiting in the inner solar system at the time that the terrestrial planets formed. The giant impact hypothesis envisions Earth being struck obliquely by an object approximately one-tenth Earth's mass—a "bullet" about the size of Mars. This is very nearly the largest impact Earth could experience without being shattered.

Such an impact would disrupt much of Earth and eject a vast amount of material into space, releasing almost enough energy to break the planet apart. Computer simulations indicate that material totaling several percent of Earth's mass could be ejected in such an impact. Most of this material would be from the stony mantles of Earth and the impacting body, not from their metal cores. This ejected rock vapor then cooled and formed a ring of material orbiting Earth. It was this ring that ultimately condensed into the Moon.

While we do not have any current way of showing that the giant impact hypothesis is the correct model of the Moon's origin, it does offer potential solutions to most of the major problems raised by the chemistry of the Moon. First, since the Moon's raw material is derived from the mantles of Earth and the projectile, the absence of metals is easily understood. Second, most of the volatile elements would have been lost during the high-temperature phase following the impact, explaining the lack of these materials on the Moon. Yet, by making the Moon primarily of terrestrial mantle material, it is also possible to understand similarities such as identical abundances of various oxygen isotopes.

This is a link to the current research on this topic. https://www.psi.edu/epo/moon/moon.html.

Attribution

"9.1 General Properties of the Moon", "9.2 The Lunar Surface", and "9.4 The Origin of the Moon" are from Astronomy by Department of Physics and Astronomy at Douglas College is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

6.4 MARS

Mars

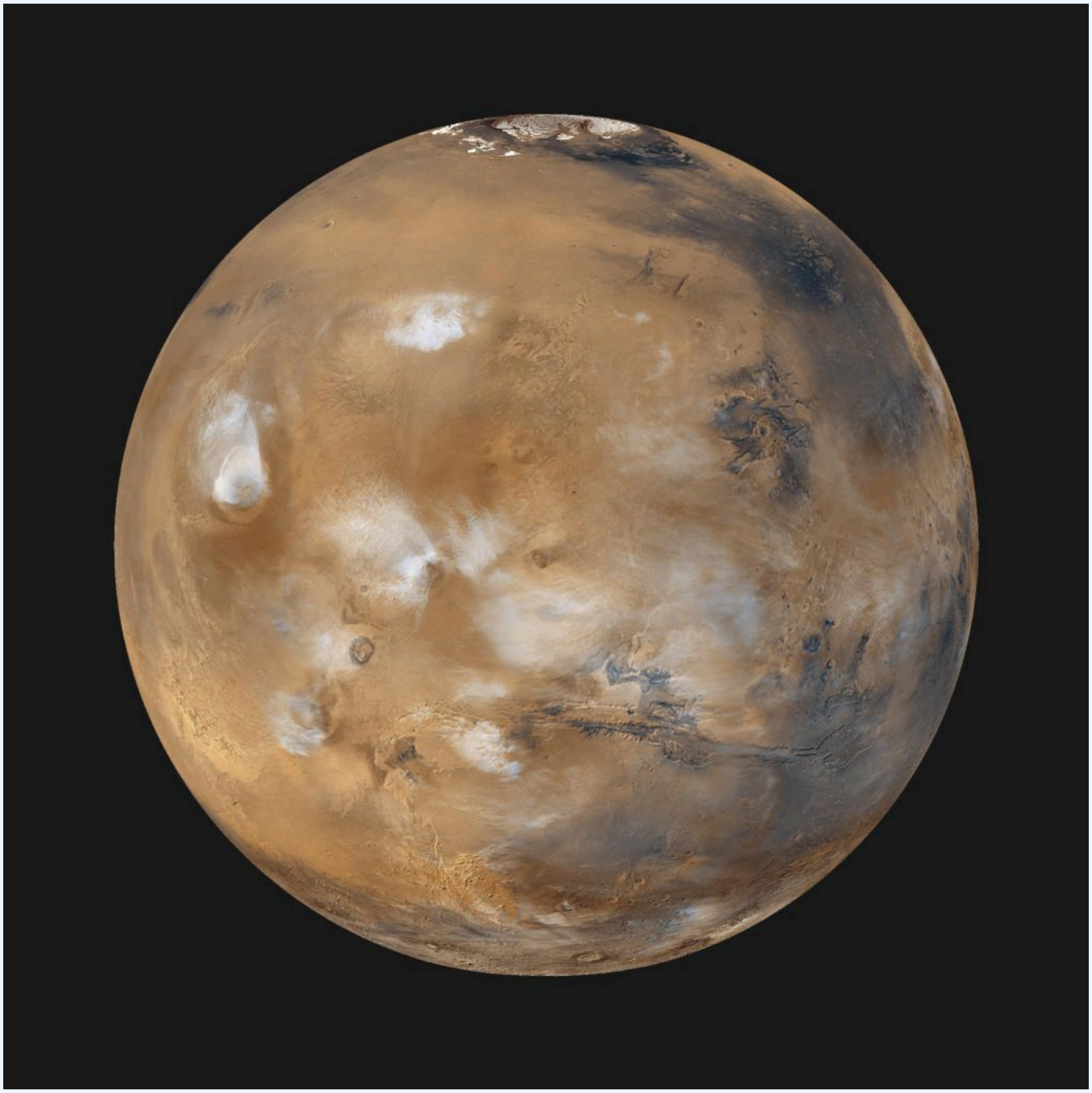


Figure 6.10. Mars – note the polar cap (top), clouds.
Mars by NASA/JPL/MSSS, NASA JPL Media Licence.

Mars is the fourth planet from the Sun. It is historically called the *Red Planet* because of its reddish appearance in Earth's night sky. Mars has some Earth-type characteristics, such as Polar Caps, clouds, and water in at least the ice or solid phase. It was very active volcanically at one time. Numerous spacecraft have successfully explored Mars from orbit and the surface. It is by far the most-visited planet (by spacecraft) in our Solar System, besides Earth.

Astronomers have determined the rotation period of Mars with great accuracy by watching the motion of permanent surface markings; its sidereal day is 24 hours 37 minutes 23 seconds, just a little longer than the rotation period of Earth. This high precision is not obtained by watching Mars for a single rotation, but by noting how many turns it makes over a long period of time. Good observations of Mars date back more than 200 years, a period during which tens of thousands of martian days have passed. As a result, the rotation period can be calculated to within a few hundredths of a second.

The rotational axis of Mars has a tilt of about 25°, similar to the tilt of Earth's axis. Thus, Mars experiences seasons very much like those on Earth. Because of the longer martian year (almost two Earth years), however, each season there lasts about six of our months.

Mars has a diameter of 6790 kilometres, just over half the diameter of Earth, giving it a total surface area very nearly equal to the continental (land) area of our planet. Its overall density of 3.9 g/cm^3 suggests a composition consisting primarily of silicates but with a small metal core. The planet has no global magnetic field, although there are areas of strong surface magnetization that indicate that there was a global field billions of years ago. Apparently, the red planet has no liquid material in its core today that would conduct electricity.

Thanks to the *Mars Global Surveyor*, we have mapped the entire planet, as shown in Figure 6.11. A laser altimeter on board made millions of separate measurements of the surface topography to a precision of a few metres—good enough to show even the annual deposition and evaporation of the polar caps. Like Earth, the Moon, and Venus, the surface of Mars has continental or highland areas as well as widespread volcanic plains. The total range in elevation from the top of the highest mountain (Olympus Mons) to the bottom of the deepest basin (Hellas) is 31 kilometres.

Mars Map from Laser Ranging

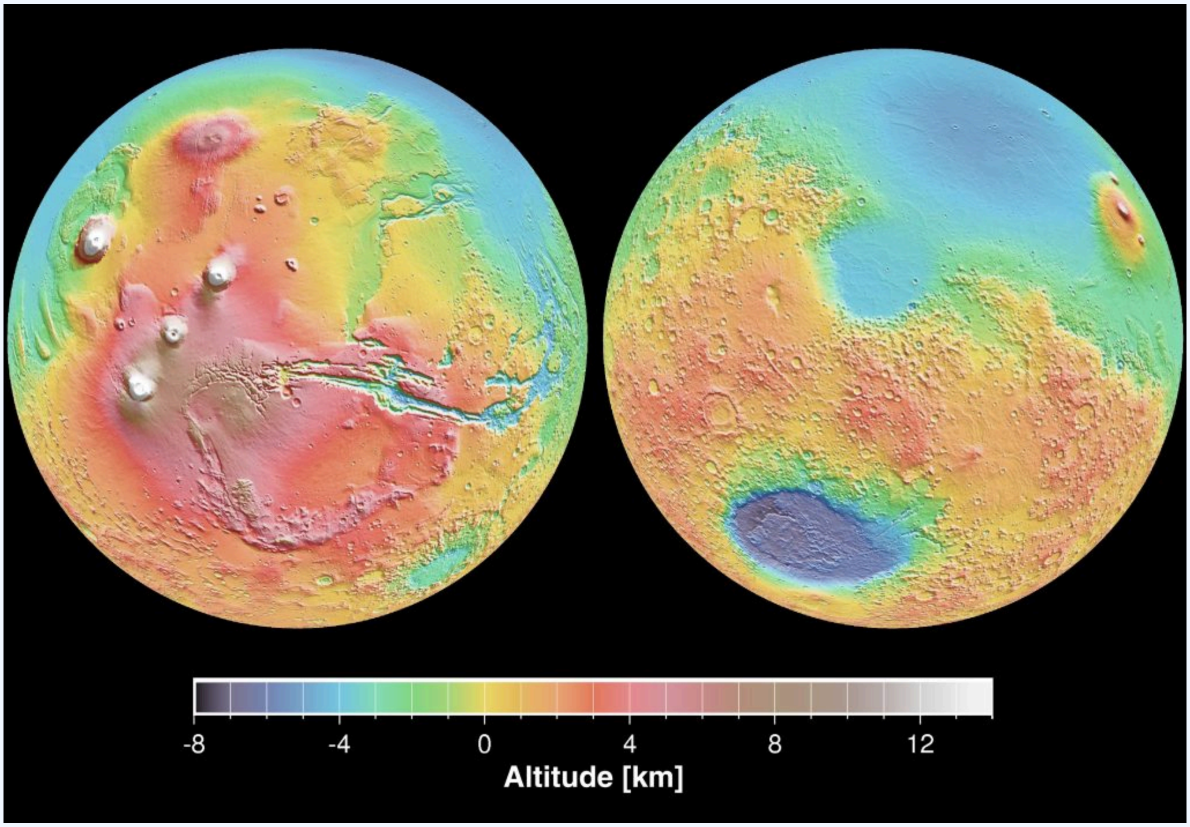


Fig 6.11. These globes are highly precise topographic maps, reconstructed from millions of individual elevation measurements made with the *Mars Global Surveyor*. Colour is used to indicate elevation. The hemisphere on the left includes the Tharsis bulge and Olympus Mons, the highest mountain on Mars; the hemisphere on the right includes the Hellas basin, which has the lowest elevation on Mars. Mars Topography by NASA/JPL, NASA JPL Media Licence.

Approximately half the planet consists of heavily cratered highland terrain, found primarily in the southern hemisphere. The other half, which is mostly in the north, contains younger, lightly cratered volcanic plains at an average elevation about 5 kilometres lower than the highlands. Remember that we saw a similar pattern on Earth, the Moon, and Venus. A geological division into older highlands and younger lowland plains seems to be characteristic of all the terrestrial planets except Mercury.

Lying across the north-south division of Mars is an uplifted continent the size of North America. This is the 10-kilometer-high Tharsis bulge, a volcanic region crowned by four great volcanoes that rise still higher into the martian sky.

The lowland plains of Mars look very much like the lunar maria, and they have about the same density of impact craters. Like the lunar maria, they probably formed between 3 and 4 billion years ago. Apparently, Mars experienced extensive volcanic activity at about the same time the Moon did, producing similar basaltic lavas.

The largest volcanic mountains of Mars are found in the Tharsis area (you can see them in Figure 6.11), although smaller volcanoes dot much of the surface. The most dramatic volcano on Mars is Olympus Mons (Mount Olympus), with a diameter larger than 500 kilometres and a summit that towers more than 20 kilometres above the surrounding plains—three times higher than the tallest mountain on Earth (Figure 6.12). The volume of this immense volcano is nearly 100 times greater than that of Mauna Loa in Hawaii. Placed on Earth's surface, Olympus would more than cover the entire state of Missouri.

Olympus Mons

Figure 6.12. The largest volcano on Mars, and probably the largest in the solar system, is Olympus Mons, illustrated in this computer-generated rendering based on data from the Mars Global Surveyor's laser altimeter. Placed on Earth, the base of Olympus Mons would completely cover the state of Missouri; the caldera, the circular opening at the top, is 65 kilometres across, about the size of Los Angeles. Olympus Mons – ESA Mars Express by ESA/DLR/FUBerlin/AndreaLuck, CC BY 2.0.

Images taken from orbit allow scientists to search for impact craters on the slopes of these volcanoes in order to estimate their age. Many of the volcanoes show a fair number of such craters, suggesting that they ceased activity a billion years or more ago. However, Olympus Mons has very, very few impact craters. Its present surface cannot be more than about 100 million years old; it may even be much younger. Some of the fresh-looking lava flows might have been formed a hundred years ago, or a thousand, or a million, but geologically speaking, they are quite young. This leads geologists to the conclusion that Olympus Mons possibly remains intermittently active today—something future Mars land developers may want to keep in mind.

The Tharsis bulge has many interesting geological features in addition to its huge volcanoes. In this part of the planet, the surface itself has bulged upward, forced by great pressures from below, resulting in extensive tectonic cracking of the crust. Among the most spectacular tectonic features on Mars are the canyons called the Valles Marineris (or Mariner Valleys, named after Mariner 9, which first revealed them to us), which are shown in Figure 6.13. They extend for about 5000 kilometres (nearly a quarter of the way around Mars) along the slopes of the Tharsis bulge. If it were on Earth, this canyon system would stretch all the way from Los Angeles to Washington, DC. The main canyon is about 7 kilometres deep and up to 100 kilometres wide, large enough for the Grand Canyon of the Colorado River to fit comfortably into one of its side canyons. Viewers of the movie "The Martian" can see a recreation of these canyonlands, as the film's hero takes a long trip through a spectacular (and somewhat exaggerated) presentation of this part of Mars.

Heavily Eroded Canyonlands on Mars

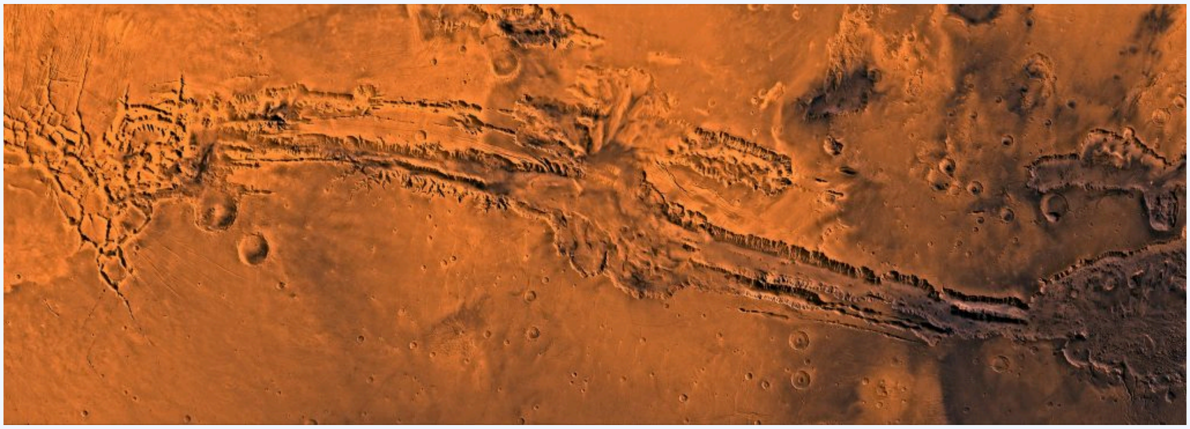


Figure 6.13. This image shows the Valles Marineris canyon complex, which is 3000 kilometres long and 8 kilometres deep.
PIA00422: Valles Marineris by NASA/JPL/USGS, NASA JPL Media Licence.

The atmosphere of Mars today has an average surface pressure of only 0.007 bar, less than 1% that of Earth. (This is how thin the air is about 30 kilometres above Earth's surface.) Martian air is composed primarily of carbon dioxide (95%), with about 3% nitrogen and 2% argon. The proportions of different gases are similar to those in the atmosphere of Venus, but a lot less of each gas is found in the thin air on Mars.

While winds on Mars can reach high speeds, they exert much less force than wind of the same velocity would on Earth because the atmosphere is so thin. The wind is able, however, to loft very fine dust particles, which can sometimes develop planet-wide dust storms. It is this fine dust that coats almost all the surface, giving Mars its distinctive red colour. In the absence of surface water, wind erosion plays a major role in sculpting the martian surface (Figure 6.14).

Wind Erosion on Mars

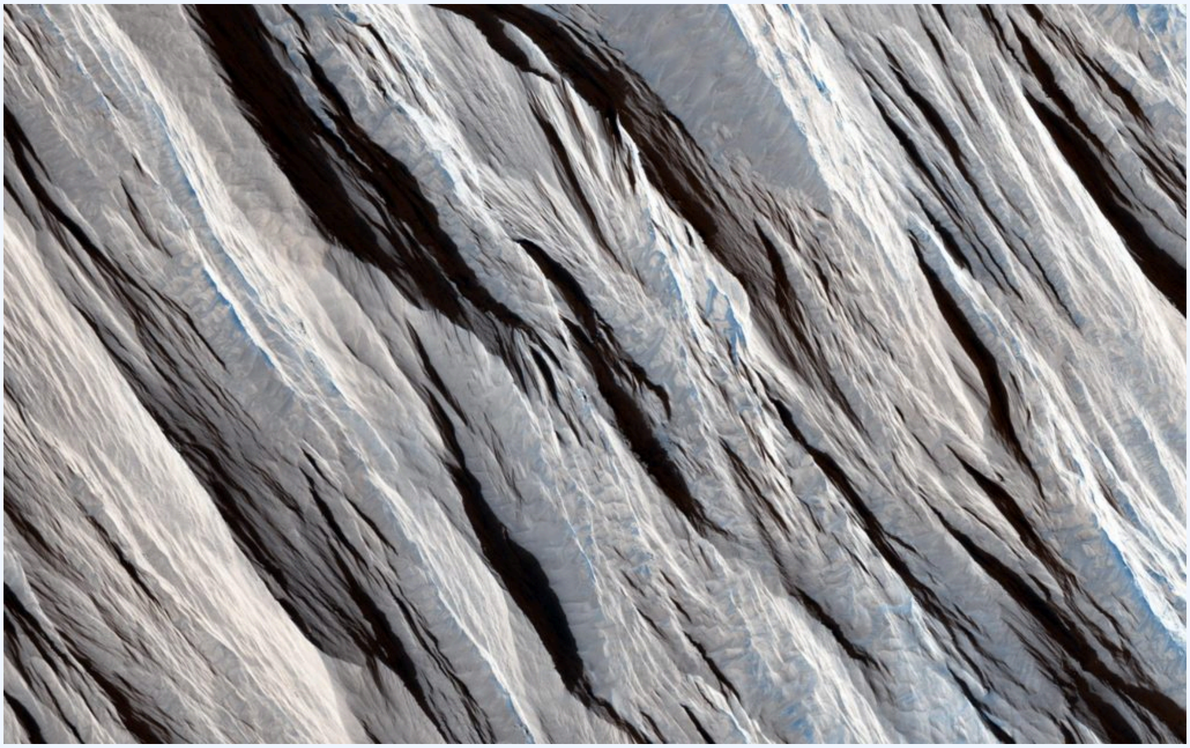


Figure 6.14. These long straight ridges, called yardangs, are aligned with the dominant wind direction. This is a high-resolution image from the Mars Reconnaissance Orbiter and is about 1 kilometre wide. On the Beauty of Yardangs by NASA/JPL-Caltech/Univ. of Arizona, NASA JPL Media Licence.

The issue of how strong the winds on Mars can be plays a big role in the 2015 hit movie *The Martian* in which the main character is stranded on Mars after being buried in the sand in a windstorm so great that his fellow astronauts have to leave the planet so their ship is not damaged. Astronomers have noted that the martian winds could not possibly be as forceful as depicted in the film because the air pressure is so low. In most ways, however, the depiction of Mars in this movie is remarkably accurate.

Although the atmosphere contains small amounts of water vapor and occasional clouds of water ice, liquid water is not stable under present conditions on Mars. Part of the problem is the low temperatures on the planet. But even if the temperature on a sunny summer day rises above the freezing point, the low pressure means that liquid water still cannot exist on the surface, except at the lowest elevations. At a pressure of less than 0.006 bar, the boiling point is as low or lower than the freezing point, and water changes directly from solid to vapor without an intermediate liquid state (as does "dry ice," carbon dioxide, on Earth). However, salts dissolved in water lower its freezing point, as we know from the way salt is used to thaw roads after snow and ice forms during winter on Earth. Salty water is therefore sometimes able to exist in liquid form on the martian surface, under the right conditions.

Several types of clouds can form in the martian atmosphere. First there are dust clouds, discussed above. Second are water-ice clouds similar to those on Earth. These often form around mountains, just as happens on our planet. Finally, the CO_2 of the atmosphere can itself condense at high altitudes to form hazes of dry ice crystals. The CO_2 clouds have no counterpart on Earth, since on our planet temperatures never drop low enough (down to about 150 K or about –125 °C) for this gas to condense.

Through a telescope, the most prominent surface features on Mars are the bright polar caps, which change with the seasons, similar to the seasonal snow cover on Earth. We do not usually think of the winter snow in northern latitudes as a part of our polar caps, but seen from space, the thin winter snow merges with Earth's thick, permanent ice caps to create an impression much like that seen on Mars (Figure 6.15).

Martian North Polar Cap

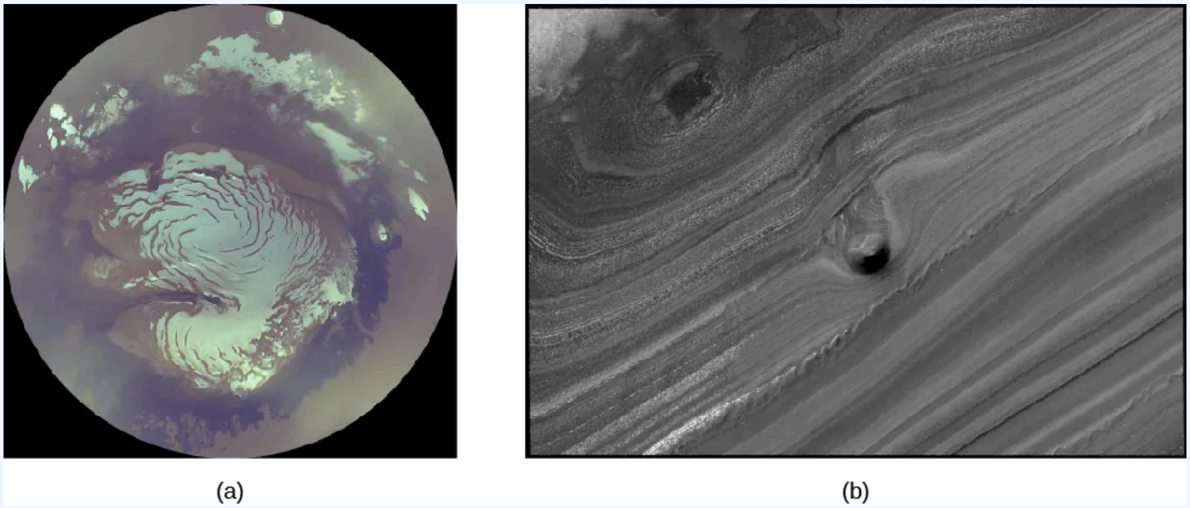


Figure 6.15. (a) This is a composite image of the north pole in summer, obtained in October 2006 by the Mars Reconnaissance Orbiter. It shows the mostly water-ice residual cap sitting atop light, tan-coloured, layered sediments. Note that although the border of this photo is circular, it shows only a small part of the planet. (b) Here we see a small section of the layered terrain near the martian north pole. There is a mound about 40 metres high that is sticking out of a trough in the centre of the picture.
(a): PIA01928: Mars Polar Cap During Transition Phase Instrument Checkout by NASA/JPL/MSSS, NASA Media Licence.
(b): PIA11231: Unusual Mound in North Polar Layered Deposits by NASA/JPL-Caltech/University of Arizona, NASA Media Licence.

The **seasonal caps** on Mars are composed not of ordinary snow but of frozen CO_2 (dry ice). These deposits condense directly from the atmosphere when the surface temperature drops below about 150 K. The caps develop during the cold martian winters and extend down to about 50° latitude by the start of spring.

Quite distinct from these thin seasonal caps of CO_2 are the **permanent** or **residual caps** that are always present near the poles. The southern permanent cap has a diameter of 350 kilometres and is composed of frozen CO_2 deposits together with a great deal of water ice. Throughout the southern summer, it remains at the freezing point of CO_2, 150 K, and this cold reservoir is thick enough to survive the summer heat intact.

The northern permanent cap is different. It is much larger, never shrinking to a diameter less than 1000 kilometres, and is composed of water ice. Summer temperatures in the north are too high for the frozen CO_2 to be retained. Measurements from the *Mars Global Surveyor* have established the exact elevations in the north polar region of Mars, showing that it is a large basin about the size of our own Arctic Ocean basin. The ice cap itself is about 3 kilometres thick, with a total volume of about 10 million km^3 (similar to that of Earth's

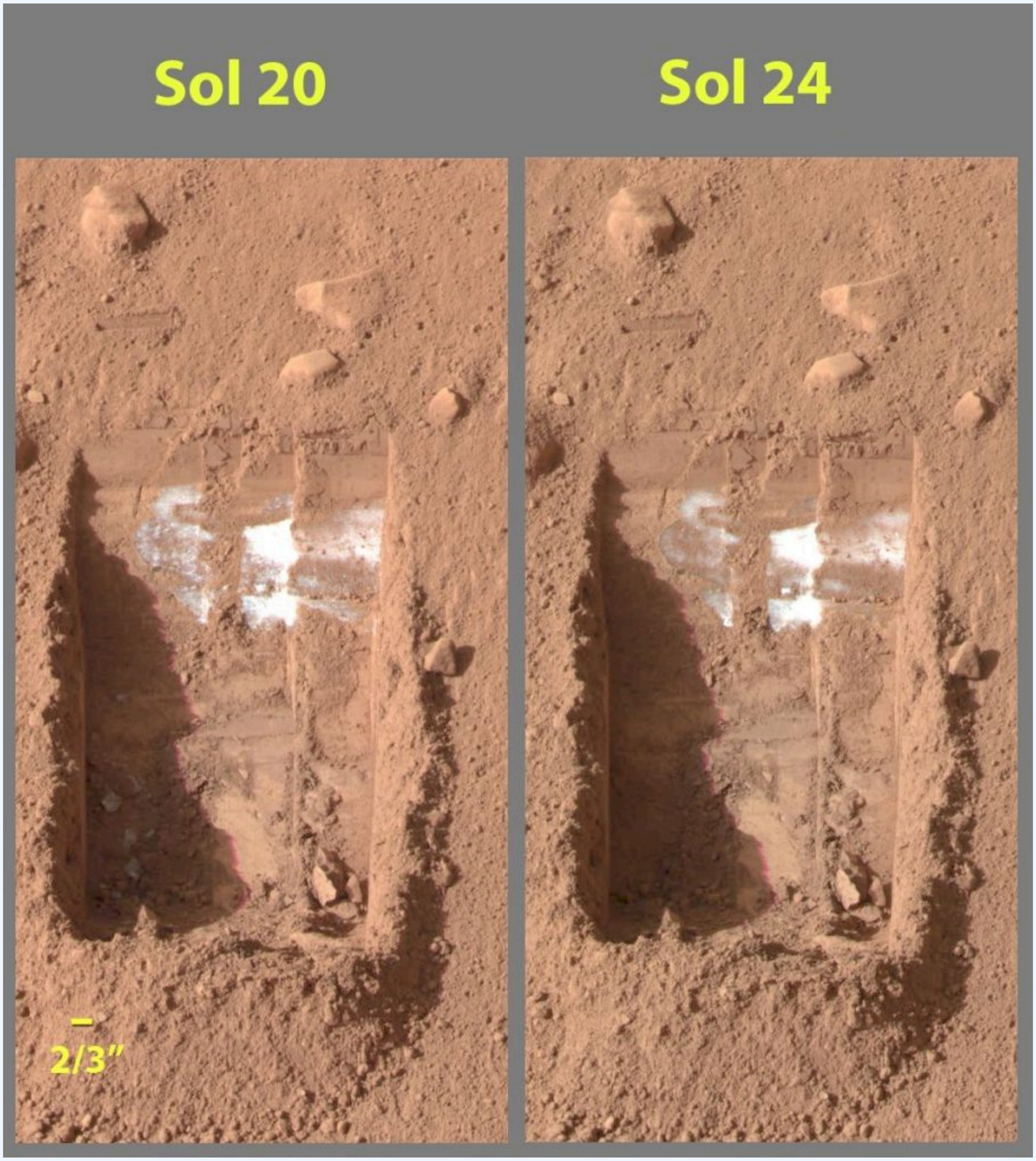
Mediterranean Sea). If Mars ever had extensive liquid water, this north polar basin would have contained a shallow sea. There is some indication of ancient shorelines visible, but better images will be required to verify this suggestion.

Images taken from orbit also show a distinctive type of terrain surrounding the permanent polar caps, as shown in Figure 6.15. At latitudes above 80° in both hemispheres, the surface consists of recent layered deposits that cover the older cratered ground below. Individual layers are typically ten to a few tens of metres thick, marked by alternating light and dark bands of sediment. Probably the material in the polar deposits includes dust carried by wind from the equatorial regions of Mars.

What do these terraced layers tell us about Mars? Some cyclic process is depositing dust and ice over periods of time. The time scales represented by the polar layers are tens of thousands of years. Apparently the martian climate experiences periodic changes at intervals similar to those between ice ages on Earth. Calculations indicate that the causes are probably also similar: the gravitational pull of the other planets produces variations in Mars' orbit and tilt as the great clockwork of the solar system goes through its paces.

The *Phoenix* spacecraft landed near the north polar cap in summer (Figure 6.16). Controllers knew that it would not be able to survive a polar winter, but directly measuring the characteristics of the polar region was deemed important enough to send a dedicated mission. The most exciting discovery came when the spacecraft tried to dig a shallow trench under the spacecraft. When the overlying dust was stripped off, they saw bright white material, apparently some kind of ice. From the way this ice sublimated over the next few days, it was clear that it was frozen water.

Evaporating Ice on Mars



Fiigure 6.16. We see a trench dug by the Phoenix lander in the north polar region four martian days apart in June 2008. If you look at the shadowed region in the bottom left of the trench, you can see three spots of ice in the left image which have sublimated away in the right image.
Disappearing Ice In Colour by NASA/JPL-Caltech/University of Arizona/Texas A&M University, NASA Media Licence.

Example 6.1

Comparing the Amount of Water on Mars and Earth

It is interesting to estimate the amount of water (in the form of ice) on Mars and to compare this with the amount of water on Earth. In each case, we can find the total volume of a layer on a sphere by multiplying the area of the sphere ($4\pi R^2$) by the thickness of the layer. For Earth, the ocean water is equivalent to a layer 3 km thick spread over the entire planet, and the radius of Earth is 6.378×10^6 m.

For Mars, most of the water we are sure of is in the form of ice near the poles. We can calculate the amount of ice in one of the residual polar caps if it is (for example) 2 km thick and has a radius of 400 km (the area of a circle is πR^2).

Solution

The volume of Earth's water is therefore the area

$$\begin{aligned} 4\pi R^2 &= 4\pi(6.378 \times 10^6)^2 \\ &= 5.1 \times 10^1 4m^2 \end{aligned}$$

multiplied by the thickness of $3000\ \mathrm{m}$:

$$5.1 \times 10^1 4\ \mathrm{m}^2 \times 3000\ \mathrm{m} = 1.5 \times 10^1 8\ \mathrm{m}^3$$

This gives $1.5 \times 10^1 8m^3$ of water. Since water has a density of 1 ton per cubic meter ($1000\ \mathrm{kg/m}^3$), we can calculate the mass:

$$1.5 \times 10^1 8m^3 \times 1\ \mathrm{ton/m}^3 = 1.5 \times 10^1 8\ \mathrm{tons}$$

For Mars, the ice doesn't cover the whole planet, only the caps; the polar cap area is

$$\pi R^2 = \pi(4 \times 10^5\ \mathrm{m})^2 = 5 \times 10^1 1\ \mathrm{m}^2$$

(Note that we converted kilometres to metres.)

The volume = area × height, so we have:

$$(2 \times 10^3\ \mathrm{m})(5 \times 10^1 1\ \mathrm{m}^2) = 1 \times 10^1 5\ \mathrm{m}^3 = 10^1 5\ \mathrm{m}^3$$

Therefore, the mass is:

$$10^1 5 \text{ m}^3 \times 1 \text{ ton/m}^3 = 10^1 5 \text{ tons}$$

This is about 0.1% that of Earth's oceans.

Exercise 6.1

A better comparison might be to compare the amount of ice in the Mars polar ice caps to the amount of ice in the Greenland ice sheet on Earth, which has been estimated as $2.85 \times 10^1 5 \text{ m}^3$. How does this compare with the ice on Mars?

Solution

The Greenland ice sheet has about 2.85 times as much ice as in the polar ice caps on Mars. They are about the same to the nearest power of 10.

Although no bodies of liquid water exist on Mars today, evidence has accumulated that rivers flowed on the red planet long ago. Two kinds of geological features appear to be remnants of ancient watercourses, while a third class—smaller gullies—suggests intermittent outbreaks of liquid water even today. We will examine each of these features in turn.

In the highland equatorial plains, there are multitudes of small, sinuous (twisting) channels—typically a few metres deep, some tens of metres wide, and perhaps 10 or 20 kilometres long (Figure 6.17). They are called **runoff channels** because they look like what geologists would expect from the surface runoff of ancient rain storms. These runoff channels seem to be telling us that the planet had a very different climate long ago. To estimate the age of these channels, we look at the cratering record. Crater counts show that this part of the planet is more cratered than the lunar maria but less cratered than the lunar highlands. Thus, the runoff channels are probably older than the lunar maria, presumably about 4 billion years old.

The second set of water-related features we see are **outflow channels** (Figure 6.17) are much larger than the runoff channels. The largest of these, which drain into the Chryse basin where Pathfinder landed, are 10 kilometres or more wide and hundreds of kilometres long. Many features of these outflow channels have convinced geologists that they were carved by huge volumes of running water, far too great to be produced by ordinary rainfall. Where could such floodwater have come from on Mars?

Runoff and Outflow Channels

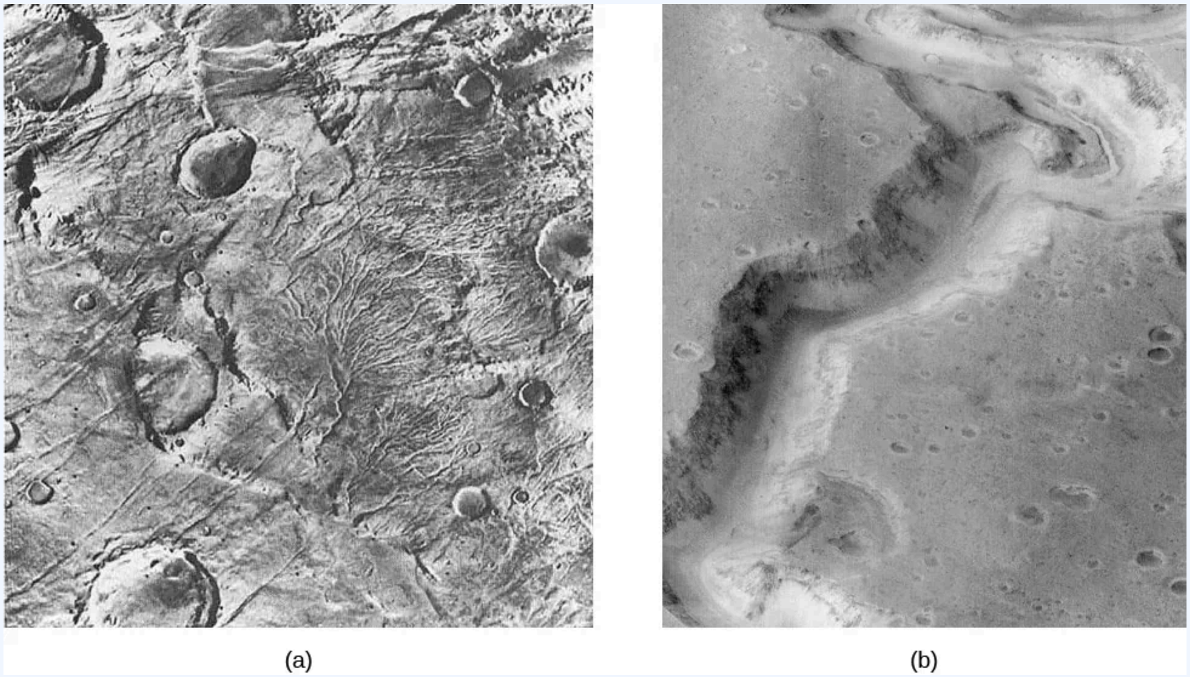


Figure 6.17. (a) These runoff channels in the old martian highlands are interpreted as the valleys of ancient rivers fed by either rain or underground springs. The width of this image is about 200 kilometres. (b) This intriguing channel, called Nanedi Valles, resembles Earth riverbeds in some (but not all) ways. The tight curves and terraces seen in the channel certainly suggest the sustained flow of a fluid like water. The channel is about 2.5 kilometres across.
(a): Photo No. A85-0760-3 by NASA, NASA Media Licence.
(b): PIA02094: Water: Sustained Flow by NASA/JPL/MSSS, NASA Media Licence.

As far we can tell, the regions where the outflow channels originate contained abundant water frozen in the soil as permafrost. Some local source of heating must have released this water, leading to a period of rapid and catastrophic flooding. Perhaps this heating was associated with the formation of the volcanic plains on Mars, which date back to roughly the same time as the outflow channels.

Note that neither the runoff channels nor the outflow channels are wide enough to be visible from Earth, nor do they follow straight lines. They could not have been the “canals” Percival Lowell imagined seeing on the red planet.

The third type of water feature, the **smaller gullies**, was discovered by the *Mars Global Surveyor* (Figure 6.18). The *Mars Global Surveyor’s* camera images achieved a resolution of a few metres, good enough to see something as small as a truck or bus on the surface. On the steep walls of valleys and craters at high latitudes, there are many erosional features that look like gullies carved by flowing water. These gullies are very young: not only are there no superimposed impact craters, but in some instances, the gullies seem to cut across recent wind-deposited dunes. Perhaps there is liquid water underground that can occasionally break out to produce short-lived surface flows before the water can freeze or evaporate.

Gullies on the Wall of Garni Crater

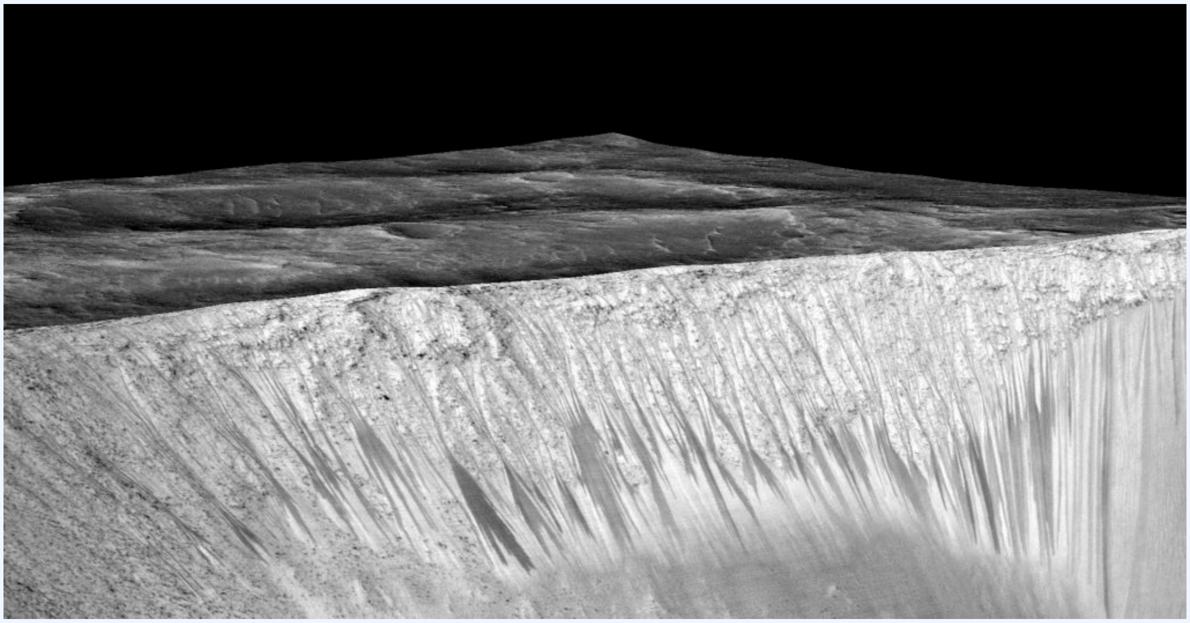


Figure 6.18. This high-resolution image is from the Mars Reconnaissance Orbiter. The dark streaks, which are each several hundred metres long, change in a seasonal pattern that suggests they are caused by the temporary flow of surface water.
Dark, Recurring Streaks on Walls of Garni Crater on Mars by NASA/JPL-Caltech/Univ. of Arizona, NASA Media Licence.

The gullies also have the remarkable property of changing regularly with the martian seasons. Many of the dark streaks (visible in Figure 6.18) elongate within a period of a few days, indicating that something is flowing downhill—either water or dark sediment. If it is water, it requires a continuing source, either from the atmosphere or from springs that tap underground water layers (aquifers.) Underground water would be the most exciting possibility, but this explanation seems inconsistent with the fact that many of the dark streaks start at high elevations on the walls of craters.

Additional evidence that the dark streaks (called by the scientists **recurring slope lineae**) are caused by water was found in 2015 when spectra were obtained of the dark streaks (Figure 6.19). These showed the presence of hydrated salts produced by the evaporation of salty water. If the water is salty, it could remain liquid long enough to flow downstream for distances of a hundred metres or more, before it either evaporates or soaks into the ground. However, this discovery still does not identify the ultimate source of the water.

Evidence for Liquid Water on Mars

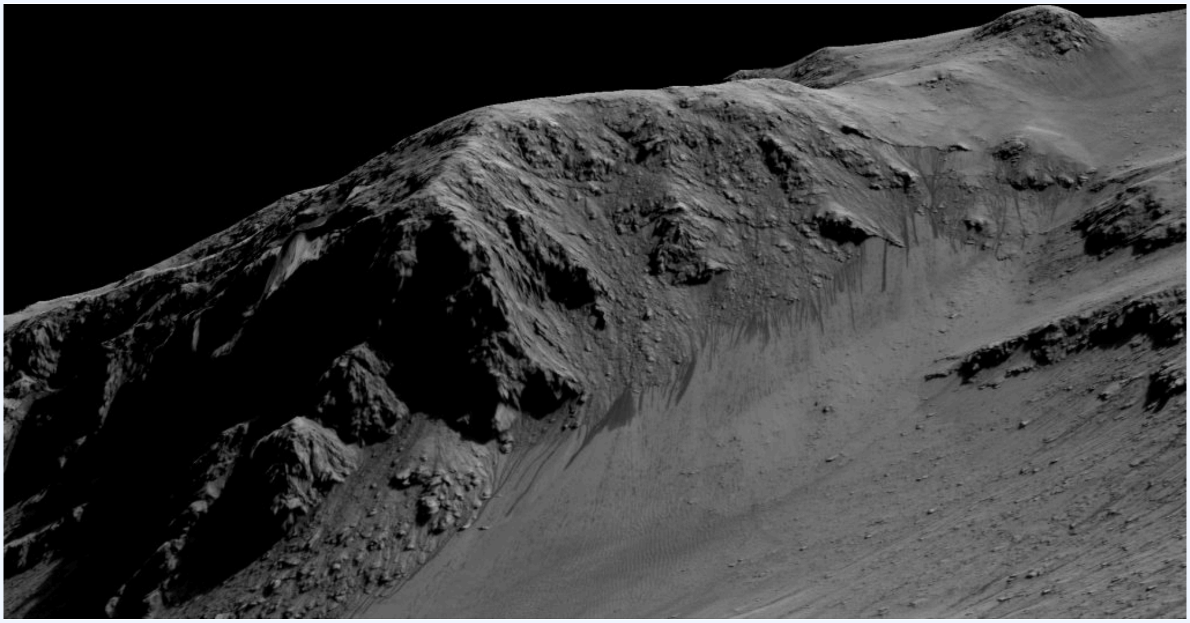


Figure 6.19. The dark streaks in Horowitz crater, which move downslope, have been called recurring slope lineae. The streaks in the centre of the image go down the wall of the crater for about a distance of 100 metres. Spectra taken of this region indicate that these are locations where salty liquid water flows on or just below the surface of Mars. (The vertical dimension is exaggerated by a factor of 1.5 compared to horizontal dimensions.)
Recurring "Lineae" on Slopes at Horowitz Crater by NASA/JPL-Caltech/Univ. of Arizona, NASA Media Licence.

Mars has two small moons, **Deimos** and **Phobos** . The names are taken from the Greek for Dread (Deimos) and Fear (Phobos). They were discovered by Asaph Hall in 1877. Phobos is the larger of the two moons (barely larger than 16 miles at its biggest point) and orbits Mars once every 11 hours and 6 minutes. Both moons are odd-shaped and very small. Both Phobos and Deimos were most-likely captured asteroids or the result of a collision with Mars. As the images show, both moons have craters. Neither moon has an atmosphere.

Deimos

Figure 6.20. Photo of Deimos.
Colour-enhanced views of Deimos by NASA/JPL-Caltech/Univ. of Arizona, NASA Media Licence.

Phobos

Figure 6.21. Photo of Phobos taken by NASA's Mars Reconnaissance Orbiter in 2008. Colour-enhanced views of Phobos by NASA/JPL-Caltech/Univ. of Arizona, NASA Media Licence.

Attribution

6.5 VENUS

Venus appears very bright, and even a small telescope reveals that it goes through phases like the Moon. Galileo discovered that Venus displays a full range of phases, and he used this as an argument to show that Venus must circle the Sun and not Earth. The planet's actual surface is not visible because it is shrouded by dense clouds that reflect about 70% of the sunlight that falls on them, frustrating efforts to study the underlying surface, even with cameras in orbit around the planet (Figure 6.22).

Venus as Photographed by the Pioneer Venus Orbiter

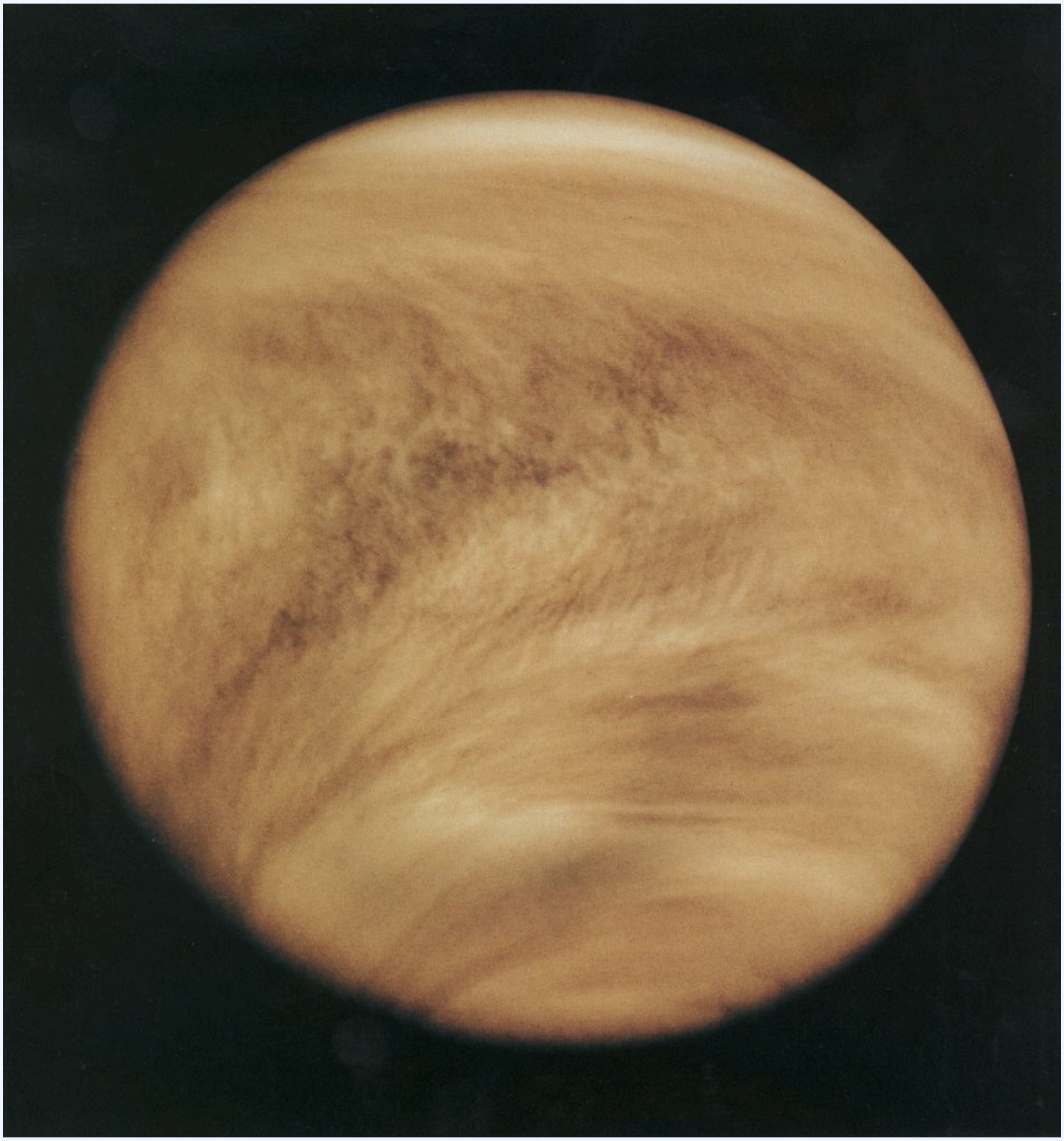


Figure 6.22. This ultraviolet image shows an upper-atmosphere cloud structure that would be invisible at visible wavelengths. Note that there is not even a glimpse of the planet's surface.
Ultraviolet image of Venus' clouds as seen by the Pioneer Venus Orbiter (Feb. 26, 1979) by NASA/GSFC, NASA Media Licence.

Since no surface detail can be seen through Venus' clouds, its rotation period can be found only by bouncing radar signals off the planet. The first radar observations of Venus' rotation were made in the early 1960s. Later, topographical surface features were identified on the planet that showed up in the reflected radar signals. The rotation period of Venus, precisely determined from the motion of such "radar features" across its disk, is 243 days. Even more surprising than how *long* Venus takes to rotate is the fact that it spins in a backward or retrograde direction (east to west).

Stop for a moment and think about how odd this slow rotation makes the calendar on Venus. The planet takes 225 Earth days to orbit the Sun and 243 Earth days to spin on its axis. So the day on Venus (as defined by its spinning once) is longer than the year! As a result, the time the Sun takes to return to the same place in Venus' sky—another way we might define the meaning of a day—turns out to be 117 Earth days. (If you say "See you tomorrow" on Venus, you'll have a long time to wait.) Although we do not know the reason for Venus' slow backward rotation, we can guess that it may have suffered one or more extremely powerful collisions during the formation process of the solar system.

Since Venus has about the same size and composition as Earth, we might expect its geology to be similar. This is partly true, but Venus does not exhibit the same kind of *plate tectonics* as Earth, and we will see that its lack of erosion results in a very different surface appearance.

Nearly 50 spacecraft have been launched to Venus, but only about half were successful. Although the 1962 US Mariner 2 flyby was the first, the Soviet Union launched most of the subsequent missions to Venus. In 1970, Venera 7 became the first probe to land and broadcast data from the surface of Venus. It operated for 23 minutes before succumbing to the high surface temperature. Additional Venera probes and landers followed, photographing the surface and analyzing the atmosphere and soil.

To understand the geology of Venus, however, we needed to make a global study of its surface, a task made very difficult by the perpetual cloud layers surrounding the planet. The problem resembles the challenge facing air traffic controllers at an airport, when the weather is so cloudy or smoggy that they can't locate the incoming planes visually. The solution is similar in both cases: use a radar instrument to probe through the obscuring layer.

The first global radar map was made by the US Pioneer Venus orbiter in the late 1970s, followed by better maps from the twin Soviet Venera 15 and 16 radar orbiters in the early 1980s. However, most of our information on the geology of Venus is derived from the US *Magellan* spacecraft, which mapped Venus with a powerful *imaging radar*. *Magellan* produced images with a resolution of 100 metres, much better than that of previous missions, yielding our first detailed look at the surface of our sister planet (Figure 6.23). (The *Magellan* spacecraft returned more data to Earth than all previous planetary missions combined; each 100 minutes of data transmission from the spacecraft provided enough information, if translated into characters, to fill two 30-volume encyclopedias.)

Radar Map of Venus

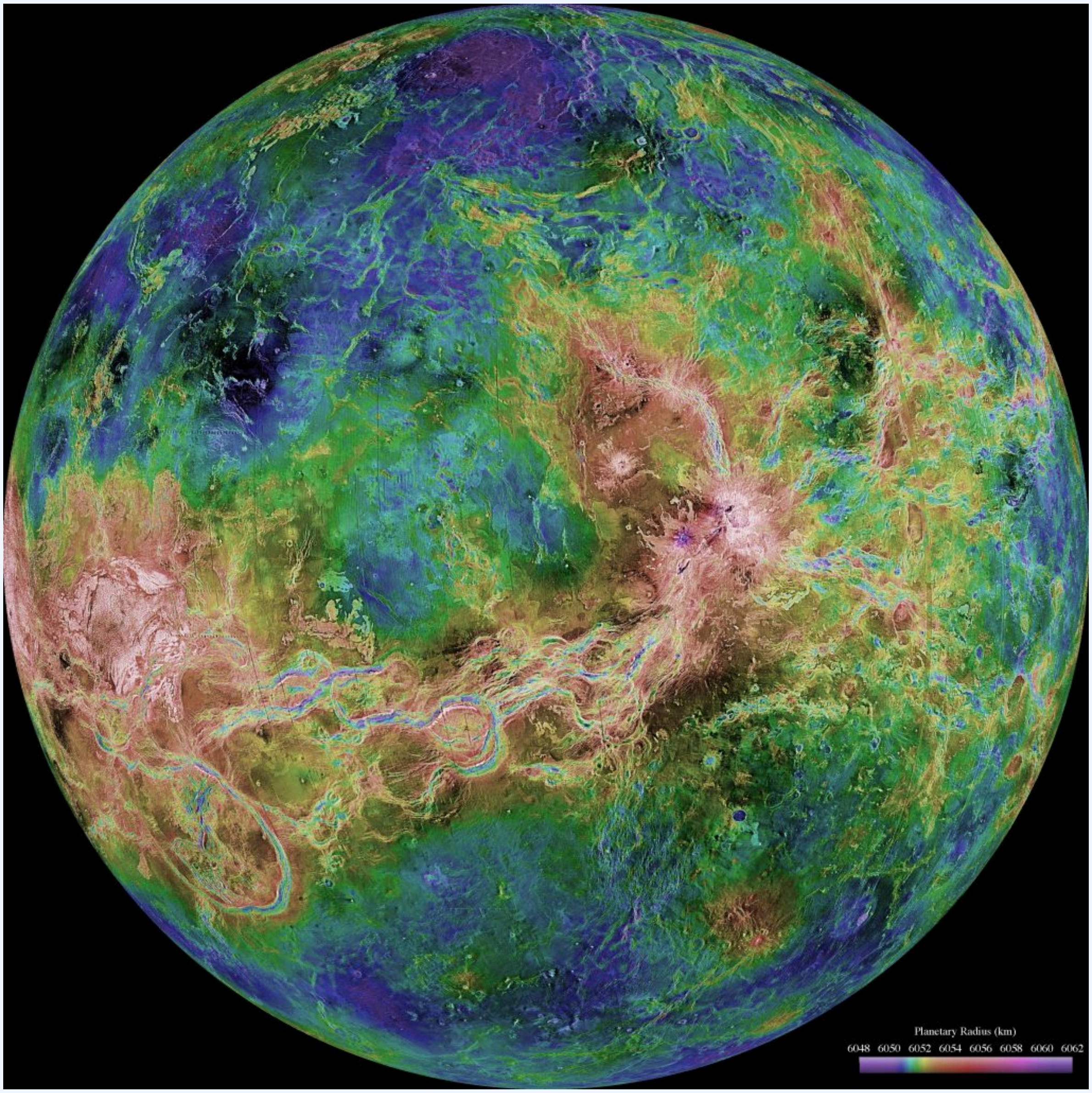


Figure 6.23. This composite image has a resolution of about 3 kilometres. Colours have been added to indicate elevation, with blue meaning low and brown and white high. The large continent Aphrodite stretches around the equator, where the bright (therefore rough) surface has been deformed by tectonic forces in the crust of Venus.
Hemispheric View of Venus by NASA/JPL/USGS, NASA Media Licence.

Consider for a moment how good *Magellan*'s resolution of 100 metres really is. It means the radar images from

Venus can show anything on the surface larger than a football field. Suddenly, a whole host of topographic features on Venus became accessible to our view. As you look at the radar images throughout this chapter, bear in mind that these are constructed from radar reflections, not from visible-light photographs. For example, bright features on these radar images are an indication of rough terrain, whereas darker regions are smoother.

The radar maps of Venus reveal a planet that looks much the way Earth might look if our planet's surface were not constantly being changed by erosion and deposition of sediment. Because there is no water or ice on Venus and the surface wind speeds are low, almost nothing obscures or erases the complex geological features produced by the movements of Venus' crust, by volcanic eruptions, and by impact craters. Having finally penetrated below the clouds of Venus, we find its surface to be naked, revealing the history of hundreds of millions of years of geological activity.

About 75% of the surface of Venus consists of lowland lava plains. Superficially, these plains resemble the basaltic ocean basins of Earth, but they were not produced in quite the same way. There is no evidence of subduction zones on Venus, indicating that, unlike Earth, this planet never experienced plate tectonics. Although **convection** (the rising of hot materials) in its mantle generated great stresses in the crust of Venus, they did not start large continental plates moving. The formation of the lava plains of Venus more nearly resembles that of the lunar maria. Both were the result of widespread lava eruptions without the crustal spreading associated with plate tectonics.

Rising above the lowland lava plains are two full-scale continents of mountainous terrain. The largest continent on Venus, called Aphrodite, is about the size of Africa (you can see it stand out in Figure 6.23). Aphrodite stretches along the equator for about one-third of the way around the planet. Next in size is the northern highland region Ishtar, which is about the size of Australia. Ishtar contains the highest region on the planet, the Maxwell Mountains, which rise 11 kilometres above the surrounding lowlands. (The Maxwell Mountains are the only feature on Venus named after a man. They commemorate James Clerk Maxwell, whose theory of electromagnetism led to the invention of radar. All other features are named for women, either from history or mythology.)

One of the first questions astronomers addressed with the high-resolution *Magellan* images was the age of the surface of Venus. Remember that the age of a planetary surface is rarely the age of the world it is on. A young age merely implies an active geology in that location. Such ages can be derived from counting impact craters. Figure 6.24 is an example of what these craters look like on the Venus radar images. The more densely cratered the surface, the greater its age. The largest crater on Venus (called Mead) is 275 kilometres in diameter, slightly larger than the largest known terrestrial crater (Chicxulub), but much smaller than the lunar impact basins.

Impact Craters on Venus

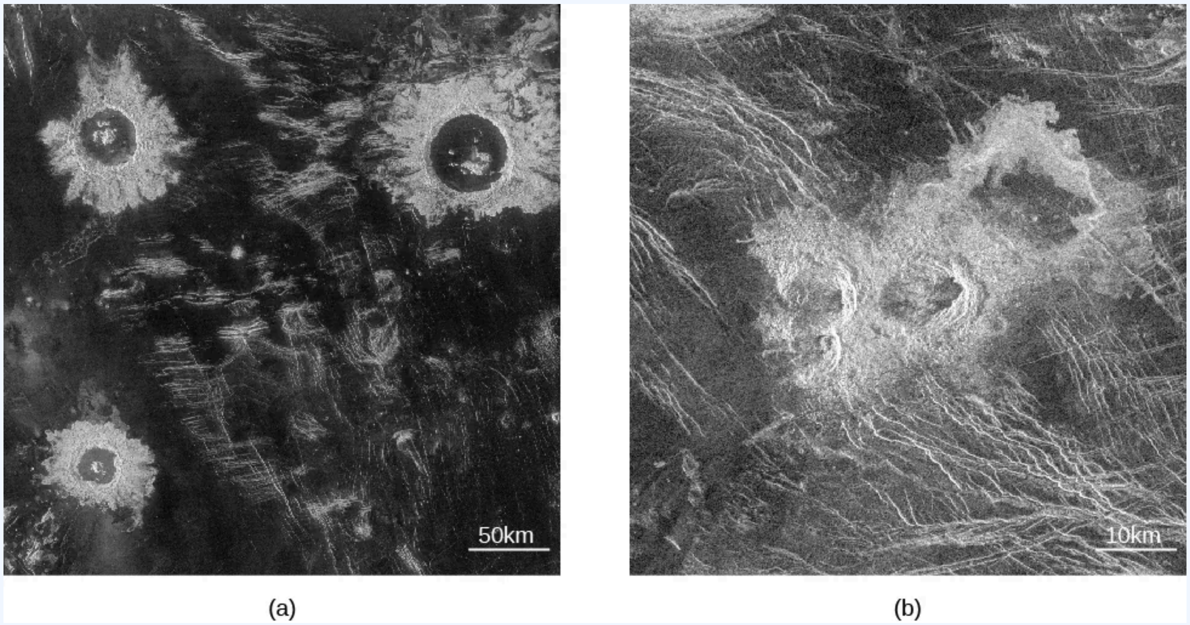


Figure 6.24. (a) These large impact craters are in the Lavinia region of Venus. Because they are rough, the crater rims and ejecta appear brighter in these radar images than do the smoother surrounding lava plains. The largest of these craters has a diameter of 50 kilometres. (b) This small, complex crater is named after writer Gertrude Stein. The triple impact was caused by the breaking apart of the incoming asteroid during its passage through the thick atmosphere of Venus. The projectile had an initial diameter of between 1 and 2 kilometres.
(a): PIA00086: Mosaic of Large Impact Craters by NASA/JPL, NASA Media Licence.
(b): PIA00088: Venus – Stein Triplet Crater by NASA/JPL, NASA Media Licence.

The large craters in the venusian plains indicate an average surface age that is only between 300 and 600 million years. These results indicate that Venus is indeed a planet with persistent geological activity, intermediate between that of Earth's ocean basins (which are younger and more active) and that of its continents (which are older and less active).

Like Earth, Venus is a planet that has experienced widespread volcanism. In the lowland plains, volcanic eruptions are the principal way the surface is renewed, with large flows of highly fluid lava destroying old craters and generating a fresh surface. In addition, numerous younger volcanic mountains and other structures are associated with surface hot spots—places where convection in the planet's mantle transports the interior heat to the surface.

The largest individual volcano on Venus, called Sif Mons, is about 500 kilometres across and 3 kilometres high—broader but lower than the Hawaiian volcano Mauna Loa. At its top is a volcanic crater, or **caldera**, about 40 kilometres across, and its slopes show individual lava flows up to 500 kilometres long. Thousands of smaller volcanoes dot the surface, down to the limit of visibility of the *Magellan* images, which correspond to cones or domes about the size of a shopping mall parking lot. Most of these seem similar to terrestrial volcanoes. Other volcanoes have unusual shapes, such as the “pancake domes” illustrated in Figure 6.25.

Pancake-Shaped Volcanoes on Venus

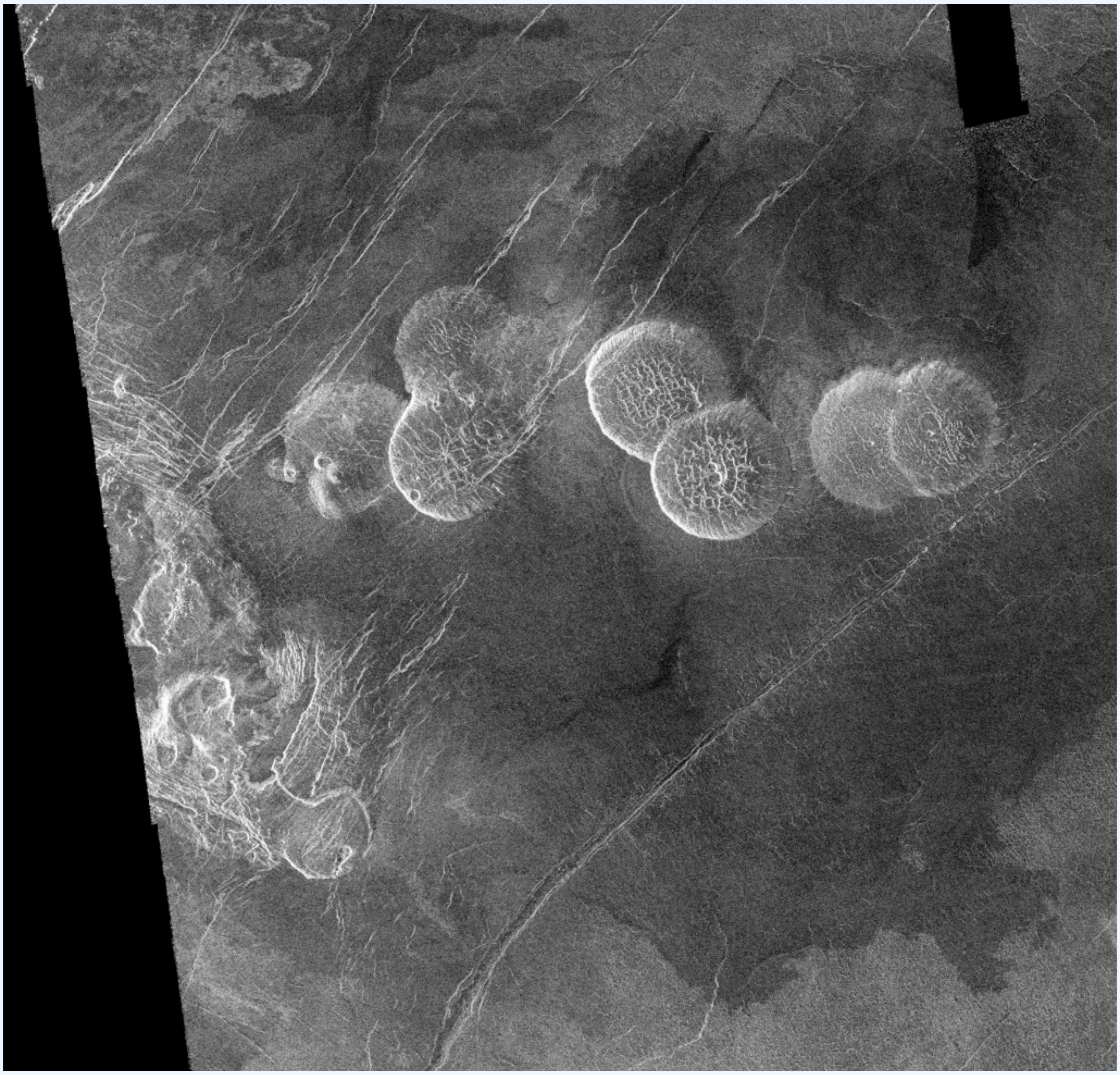


Figure 6.25. These remarkable circular domes, each about 25 kilometres across and about 2 kilometres tall, are the result of eruptions of highly viscous (sludgy) lava that spreads out evenly in all directions. Venus – Alpha Regio by NASA/JPL, NASA JPL Media Licence.

All of the volcanism is the result of eruption of lava onto the surface of the planet. But the hot lava rising from the interior of a planet does not always make it to the surface. On both Earth and Venus, this upwelling lava can collect to produce bulges in the crust. Many of the granite mountain ranges on Earth, such as the Sierra Nevada in California, involve such subsurface volcanism. These bulges are common on Venus, where they produce large circular or oval features called ***coronae*** (singular: corona) (Figure 6.26).

The "Miss Piggy" Corona

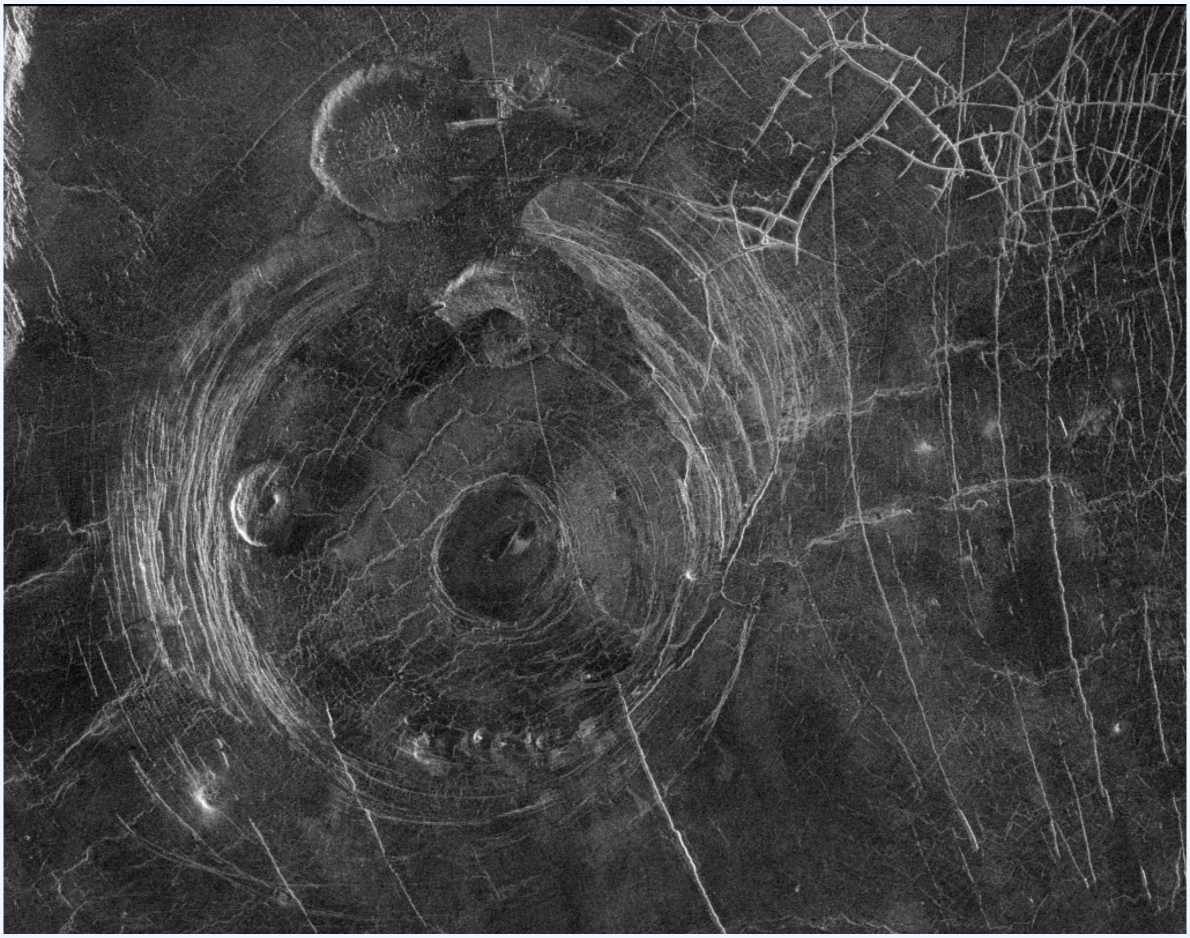


Figure 6.26. Fotla Corona is located in the plains to the south of Aphrodite Terra. Curved fracture patterns show where the material beneath has put stress on the surface. A number of pancake and dome volcanoes are also visible. Fotla was a Celtic fertility goddess. Some students see a resemblance between this corona and Miss Piggy of the Muppets (her left ear, at the top of the picture, is the pancake volcano in the upper center of the image).
Venus – Aine Corona by NASA/JPL, NASA JPL Media Licence.

The successful Venera landers of the 1970s found themselves on an extraordinarily inhospitable planet, with a surface pressure of 90 bars and a temperature hot enough to melt lead and zinc. Despite these unpleasant conditions, the spacecraft were able to photograph their surroundings and collect surface samples for chemical analysis before their instruments gave out. The diffuse sunlight striking the surface was tinted red by the clouds, and the illumination level was equivalent to a heavy overcast on Earth.

The probes found that the rock in the landing areas is igneous, primarily basalts. Examples of the Venera photographs are shown in Figure 6.27. Each picture shows a flat, desolate landscape with a variety of rocks, some of which may be ejecta from impacts. Other areas show flat, layered lava flows. There have been no further landings on Venus since the 1970s. However, NASA has plans for a radar orbiter with much higher resolution than Magellan, that will focus on the complex structures in the Aphrodite region. Some geologists think this area shows evidence of geologic plates, with past activity analogous to the plate tectonics on Earth.

Surface of Venus

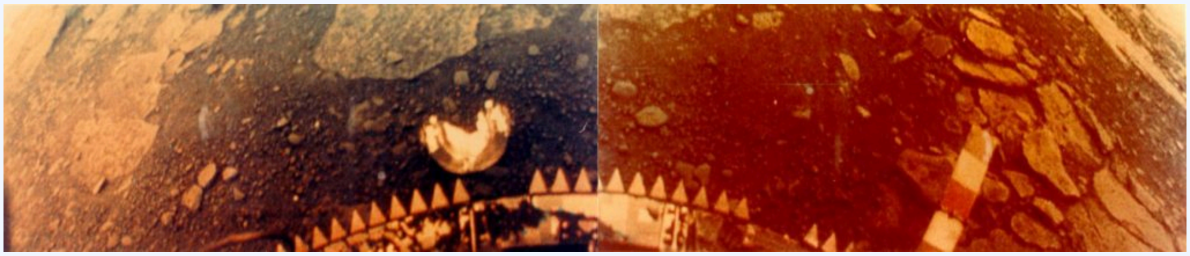


Figure 6.27. These views of the surface of Venus are from the Venera 13 spacecraft. Everything is orange because the thick atmosphere of Venus absorbs the bluer colours of light. The horizon is visible in the upper corner of each image.
Colour version of the left & right halves of the Venera 13 image by NASA/Russian Space Agency, NASA Media Licence.

The thick atmosphere of Venus produces the high surface temperature and shrouds the surface in a perpetual red twilight. Sunlight does not penetrate directly through the heavy clouds, but the surface is fairly well lit by diffuse light (about the same as the light on Earth under a heavy overcast). The weather at the bottom of this deep atmosphere remains perpetually hot and dry, with calm winds. Because of the heavy blanket of clouds and atmosphere, one spot on the surface of Venus is similar to any other as far as weather is concerned. The most abundant gas on Venus is carbon dioxide (CO_2), which accounts for 96% of the atmosphere. The second most common gas is nitrogen. The predominance of carbon dioxide over nitrogen is not surprising when you recall that Earth's atmosphere would also be mostly carbon dioxide if this gas were not locked up in marine sediments.

Table 6.3 compares the compositions of the atmospheres of Venus, Mars, and Earth. Expressed in this way, as percentages, the proportions of the major gases are very similar for Venus and Mars, but in total quantity, their atmospheres are dramatically different. With its surface pressure of 90 bars, the venusian atmosphere is more than 10,000 times more massive than its martian counterpart. Overall, the atmosphere of Venus is very dry; the absence of water is one of the important ways that Venus differs from Earth.

Table 6.3. Atmospheric Composition of Earth, Venus, and Mars

Gas	Earth	Venus	Mars
Carbon dioxide (CO_2)	0.03%	96%	95.3%
Nitrogen (N_2)	78.1%	3.5%	2.7%
Argon (Ar)	0.93%	0.006%	1.6%
Oxygen (O_2)	21.0%	0.003%	0.15%
Neon (Ne)	0.002%	0.001%	0.0003%

The atmosphere of Venus has a huge **troposphere** (region of convection) that extends up to at least 50 kilometres above the surface (Figure 6.28). Within the troposphere, the gas is heated from below and circulates slowly, rising near the equator and descending over the poles. Being at the base of the atmosphere of Venus is something like being a kilometer or more below the ocean surface on Earth. There, the mass of water evens out temperature variations and results in a uniform environment—the same effect the thick atmosphere has on Venus.

Venus' Atmosphere

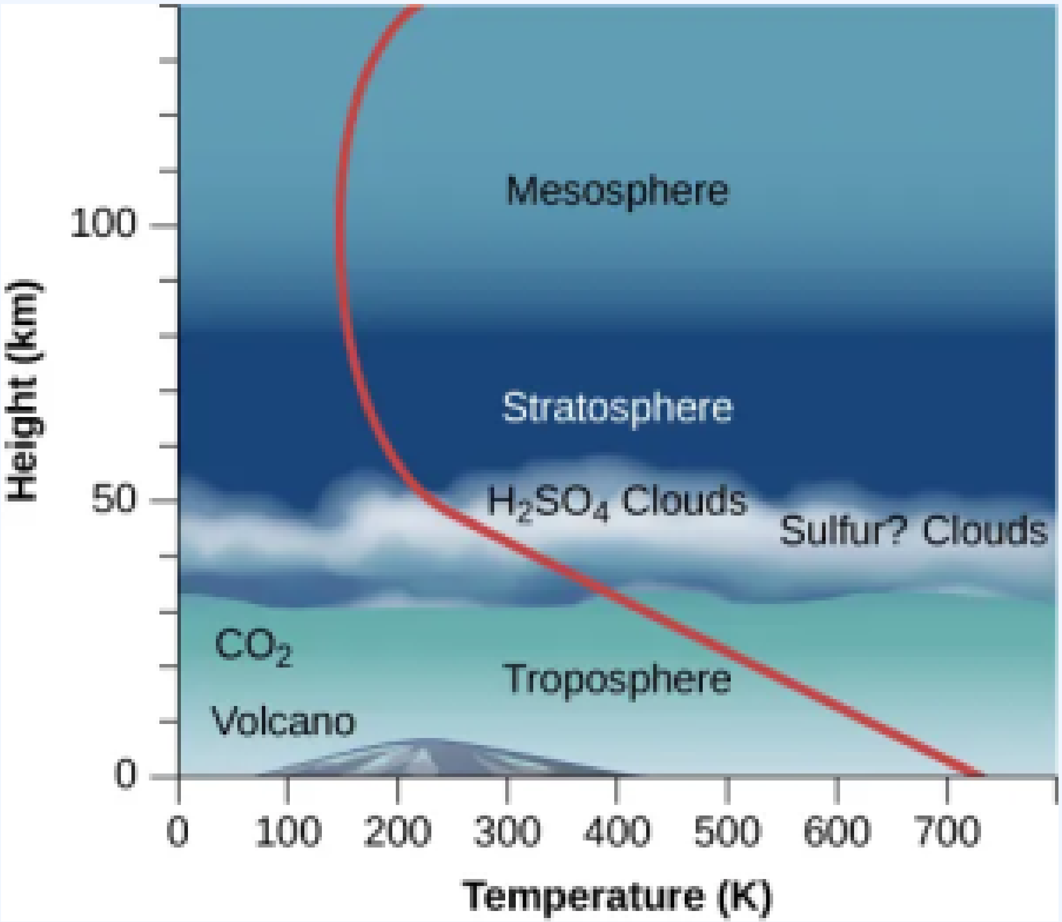


Figure 6.28. The layers of the massive atmosphere of Venus shown here are based on data from the Pioneer and Venera entry probes. Height is measured along the left axis, the bottom scale shows temperature, and the red line allows you to read off the temperature at each height. Notice how steeply the temperature rises below the clouds, thanks to the planet's huge greenhouse effect.

In the upper troposphere, between 30 and 60 kilometres above the surface, a thick cloud layer is composed primarily of sulfuric acid droplets. Sulfuric acid (H_2SO_4) is formed from the chemical combination of sulfur dioxide (SO_2) and water (H_2O). In the atmosphere of Earth, sulfur dioxide is one of the primary gases emitted by volcanoes, but it is quickly diluted and washed out by rainfall. In the dry atmosphere of Venus, this unpleasant substance is apparently stable. Below 30 kilometres, the Venus atmosphere is clear of clouds.

The high surface temperature of Venus was discovered by radio astronomers in the late 1950s and confirmed by the Mariner and Venera probes. How can our neighbor planet be so hot? Although Venus is somewhat closer to the Sun than is Earth, its surface is hundreds of degrees hotter than you would expect from the extra sunlight it receives. Scientists wondered what could be heating the surface of Venus to a temperature above 700 K. The answer turned out to be the *greenhouse effect*.

The greenhouse effect works on Venus just as it does on Earth, but since Venus has so much more CO_2—almost a million times more—the effect is much stronger. The thick CO_2 acts as a blanket, making it very difficult for the infrared (heat) radiation from the ground to get back into space. As a result, the surface heats up. The energy balance is only restored when the planet is radiating as much energy as it receives from the Sun, but this can happen only when the temperature of the lower atmosphere is very high. One way of thinking of greenhouse heating is that it must raise the surface temperature of Venus until this energy balance is achieved.

Let us try to reconstruct the possible evolution of Venus from an earthlike beginning to its present state. Venus may once have had a climate similar to that of Earth, with moderate temperatures, water oceans, and much of its CO_2 dissolved in the ocean or chemically combined with the surface rocks. Then we allow for modest additional heating—by gradual increase in the energy output of the Sun, for example. When we calculate how Venus' atmosphere would respond to such effects, it turns out that even a small amount of extra heat can lead to increased evaporation of water from the oceans and the release of gas from surface rocks.

This in turn means a further increase in the atmospheric CO_2 and H_2O, gases that would amplify the greenhouse effect in Venus' atmosphere. That would lead to still more heat near Venus' surface and the release of further CO_2 and H_2O. Unless some other processes intervene, the temperature thus continues to rise. Such a situation is called the runaway greenhouse effect.

Attribution

"10.1 The Nearest Planets: An Overview", "10.2 The Geology of Venus", and "10.3 The Massive Atmosphere of Venus" from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

6.6 MERCURY

The planet Mercury is similar to the Moon in many ways. Like the Moon, it has no atmosphere, and its surface is heavily cratered. As described later in this chapter, it also shares with the Moon the likelihood of a violent birth.

Mercury is the nearest planet to the Sun, and, in accordance with Kepler's third law, it has the shortest period of revolution about the Sun (88 of our days) and the highest average orbital speed (48 kilometres per second). It is appropriately named for the fleet-footed messenger god of the Romans. Because Mercury remains close to the Sun, it can be difficult to pick out in the sky. As you might expect, it's best seen when its eccentric orbit takes it as far from the Sun as possible.

The **semimajor axis** of Mercury's orbit—that is, the planet's average distance from the Sun—is 58 million kilometres, or 0.39 AU. However, because its orbit has the high eccentricity of 0.206, Mercury's actual distance from the Sun varies from 46 million kilometres at perihelion to 70 million kilometres at aphelion.

Mercury's mass is one-eighteenth that of Earth, making it the smallest terrestrial planet. Mercury is the smallest planet (except for the dwarf planets), having a diameter of 4878 kilometres, less than half that of Earth. Mercury's density is 5.4 g/cm^3, much greater than the density of the Moon, indicating that the composition of those two objects differs substantially.

Mercury's composition is one of the most interesting things about it and makes it unique among the planets. Mercury's high density tells us that it must be composed largely of heavier materials such as metals. The most likely models for Mercury's interior suggest a metallic iron-nickel core amounting to 60% of the total mass, with the rest of the planet made up primarily of silicates. The core has a diameter of 3500 kilometres and extends out to within 700 kilometres of the surface. We could think of Mercury as a metal ball the size of the Moon surrounded by a rocky crust 700 kilometres thick (Figure 6.29). Unlike the Moon, Mercury does have a weak magnetic field. The existence of this field is consistent with the presence of a large metal core, and it suggests that at least part of the core must be liquid in order to generate the observed magnetic field.

Mercury's Internal Structure

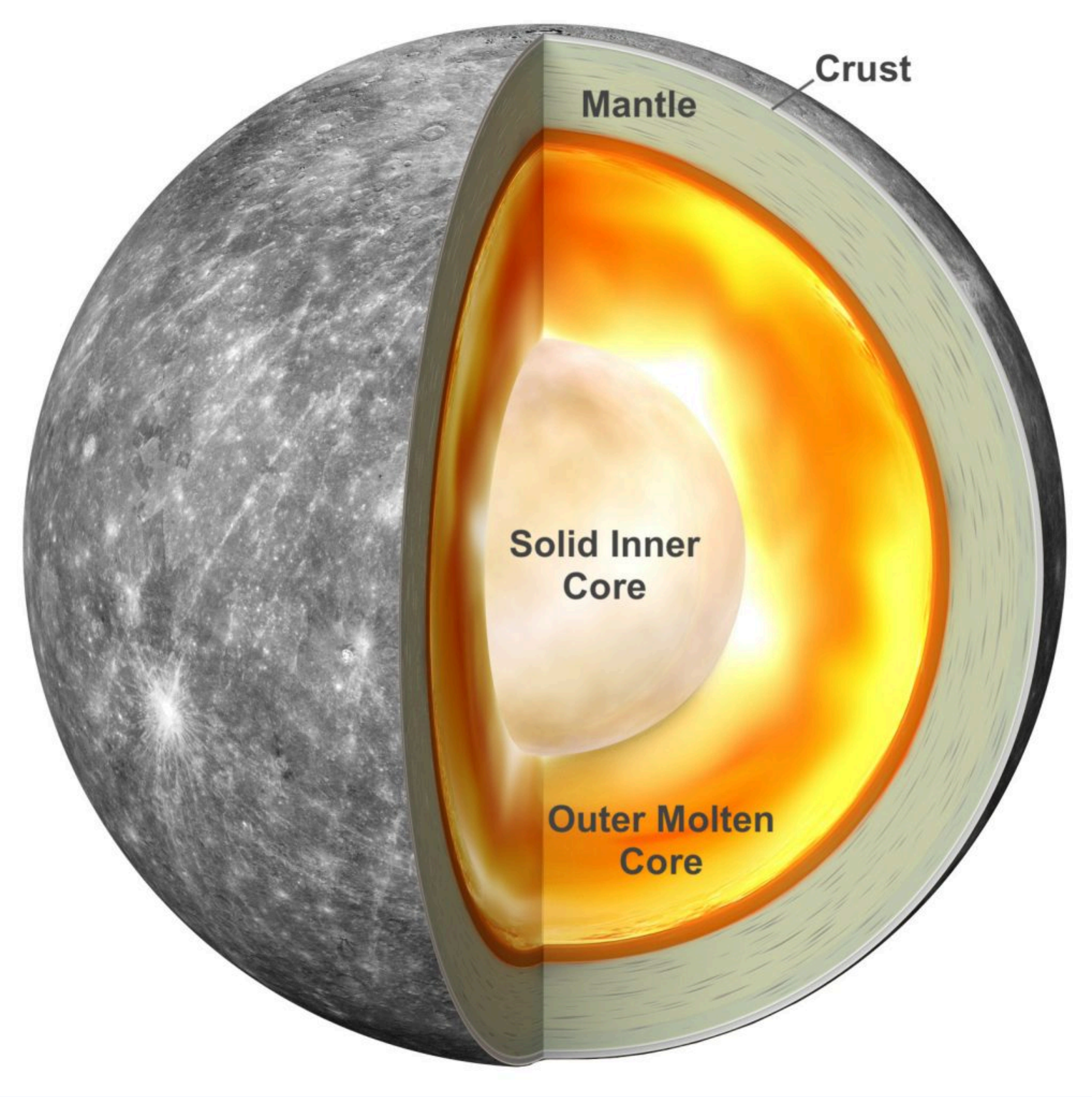


Figure 6.29. The interior of Mercury is dominated by a metallic core about the same size as our Moon. Mercury's internal structure by NASA's Goddard Space Flight Center, NASA Media Licence.

Example 6.2

Densities of Worlds

The average density of a body equals its mass divided by its volume. For a sphere, density is:

$$\text{density} = \frac{\text{mass}}{\frac{4}{3}\pi R^3}$$

Astronomers can measure both mass and radius accurately when a spacecraft flies by a body. Using the information in this chapter, we can calculate the approximate average density of the Moon.

Solution

For a sphere,

$$\text{density} = \frac{\text{mass}}{\frac{4}{3}\pi R^3} = \frac{7.35 \times 10^{2}2 \text{ kg}}{4.2 \times 5.2 \times 10^{1}8 \text{ m}^3} = 3.4 \times 10^3 \text{ kg/m}^3$$

Table 6.1 gives a value of 3.3 g/cm^3, which is $3.3 \times 10^3 \text{ kg/m}^3$.

Exercise 6.2

Using the information in this chapter, calculate the average density of Mercury. Show your work. Does your calculation agree with the figure we give in this chapter?

Solution

$$\text{density} = \frac{\text{mass}}{\frac{4}{3}\pi R^3} = \frac{3.3 \times 10^{2}3 \text{ kg}}{4.2 \times 1.45 \times 10^{1}9 \text{ m}^3} = 5.4 \times 10^3 \text{ kg/m}^3$$

That matches the value given in Table 6.1 when g/cm^3 is converted into kg/m^3.

Mercury's Strange Rotation

Visual studies of Mercury's indistinct surface markings were once thought to indicate that the planet kept one face to the Sun (as the Moon does to Earth). Thus, for many years, it was widely believed that Mercury's rotation period was equal to its revolution period of 88 days, making one side perpetually hot while the other was always cold.

Radar observations of Mercury in the mid-1960s, however, showed conclusively that Mercury does not keep one side fixed toward the Sun. If a planet is turning, one side seems to be approaching Earth while the other is moving away from it. The resulting Doppler shift spreads or broadens the precise transmitted radar-wave frequency into a range of frequencies in the reflected signal (Figure 6.30). The degree of broadening provides an exact measurement of the rotation rate of the planet.

Doppler Radar Measures Rotation

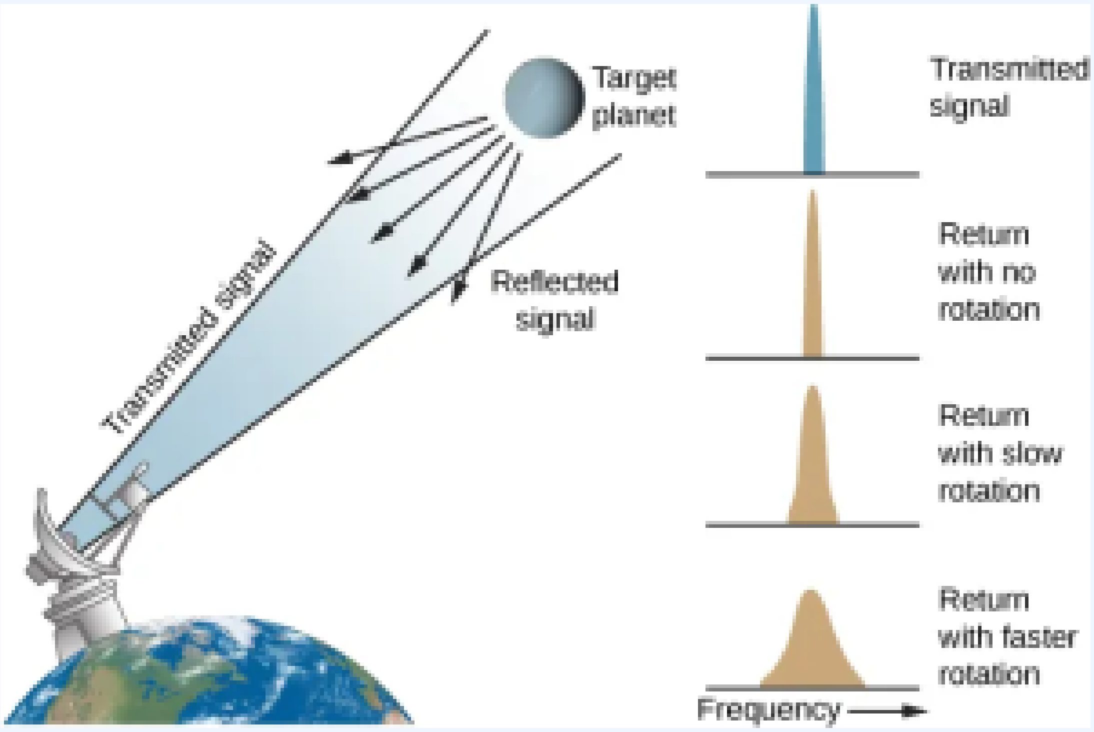


Figure 6.30. When a radar beam is reflected from a rotating planet, the motion of one side of the planet's disk toward us and the other side away from us causes Doppler shifts in the reflected signal. The effect is to cause both a redshift and a blueshift, widening the spread of frequencies in the radio beam.

Mercury's period of rotation (how long it takes to turn with respect to the distant stars) is 59 days, which is just two-thirds of the planet's period of revolution. Subsequently, astronomers found that a situation where the spin and the orbit of a planet (its year) are in a 2:3 ratio turns out to be stable.

Mercury, being close to the Sun, is very hot on its daylight side; but because it has no appreciable atmosphere, it gets surprisingly cold during the long nights. The temperature on the surface climbs to 700 K (430 °C) at noontime. After sunset, however, the temperature drops, reaching 100 K (–170 °C) just before dawn. (It is even colder in craters near the poles that receive no sunlight at all.) The range in temperature on Mercury is thus 600 K (or 600 °C), a greater difference than on any other planet.

Mercury rotates three times for each two orbits around the Sun. It is the only planet that exhibits this relationship between its spin and its orbit, and there are some interesting consequences for any observers who might someday be stationed on the surface of Mercury.

Here on Earth, we take for granted that days are much shorter than years. Therefore, the two astronomical ways of defining the local "day"—how long the planet takes to rotate and how long the Sun takes to return to the same position in the sky—are the same on Earth for most practical purposes. But this is not the case on Mercury. While Mercury rotates (spins once) in 59 Earth days, the time for the Sun to return to the same place in Mercury's sky turns out to be two Mercury years, or 176 Earth days. (Note that this result is not intuitively obvious, so don't be upset if you didn't come up with it.) Thus, if one day at noon a Mercury explorer suggests to her companion that they should meet at noon the next day, this could mean a very long time apart!

To make things even more interesting, recall that Mercury has an eccentric orbit, meaning that its distance from the Sun varies significantly during each mercurian year. By Kepler's law, the planet moves fastest in its orbit when closest to the Sun. Let's examine how this affects the way we would see the Sun in the sky during one 176-Earth-day cycle. We'll look at the situation as if we were standing on the surface of Mercury in the center of a giant basin that astronomers call Caloris (Figure 6.31).

Caloris Basin

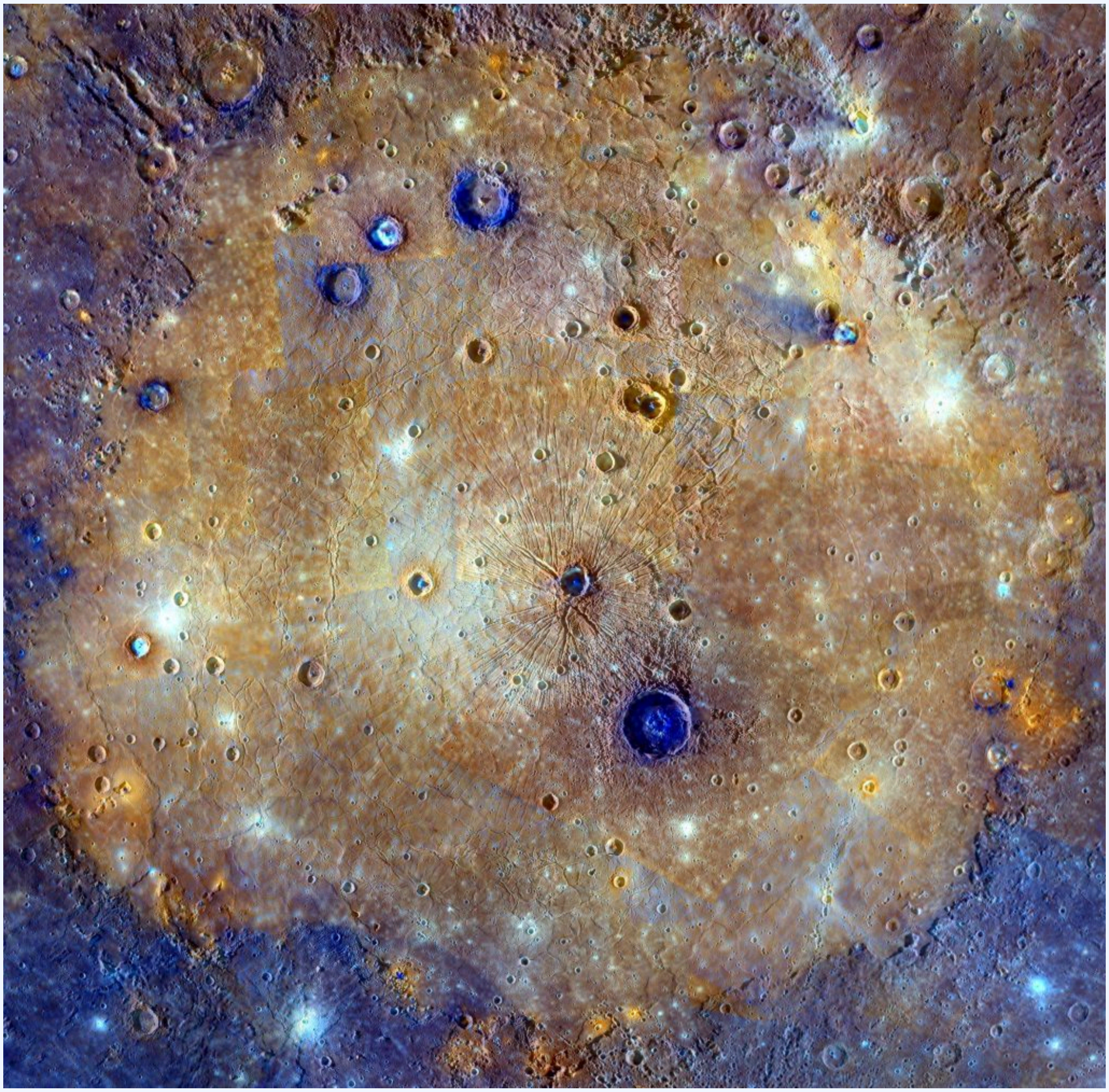


Figure 6.31. This partially flooded impact basin is the largest known structural feature on Mercury. The smooth plains in the interior of the basin have an area of almost two million square kilometres. Mercury's Caloris Basin by NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington, NASA Media Licence.

At the location of Caloris, Mercury is most distant from the Sun at sunrise; this means the rising Sun looks smaller in the sky (although still more than twice the size it appears from Earth). As the Sun rises higher and higher, it looks bigger and bigger; Mercury is now getting closer to the Sun in its eccentric orbit. At the same time, the apparent motion of the Sun slows down as Mercury's faster motion in orbit begins to catch up with its rotation.

At noon, the Sun is now three times larger than it looks from Earth and hangs almost motionless in the sky. As the afternoon wears on, the Sun appears smaller and smaller, and moves faster and faster in the sky. At sunset, a full Mercury year (or 88 Earth days after sunrise), the Sun is back to its smallest apparent size as it dips out of sight. Then it takes another Mercury year before the Sun rises again. (By the way, sunrises and sunsets are much more sudden on Mercury, since there is no atmosphere to bend or scatter the rays of sunlight.)

Astronomers call locations like the Caloris Basin the "hot longitudes" on Mercury because the Sun is closest to the planet at noon, just when it is lingering overhead for many Earth days. This makes these areas the hottest places on Mercury.

We bring all this up not because the exact details of this scenario are so important but to illustrate how many of the things we take for granted on Earth are not the same on other worlds. As we've mentioned before, one of the best things about taking an astronomy class should be ridding you forever of any "Earth chauvinism" you might have. The way things are on our planet is just one of the many ways nature can arrange reality.

This Day on Mercury visualization shows how Mercury's slow rotation leads to very long days and how this combines with its highly elliptical orbit to cause some unusual motions of the sun in the sky. Note the stick figure on the surface of Mercury; the Day/Night box shows what that stick figure would see in the sky.

The first close-up look at Mercury came in 1974, when the US spacecraft Mariner 10 passed 9500 kilometres from the surface of the planet and transmitted more than 2000 photographs to Earth, revealing details with a resolution down to 150 metres. Subsequently, the planet was mapped in great detail by the MESSENGER spacecraft, which was launched in 2004 and made multiple flybys of Earth, Venus, and Mercury before settling into orbit around Mercury in 2011. It ended its life in 2015, when it was commanded to crash into the surface of the planet.

Mercury's surface strongly resembles the Moon in appearance (Figure 6.32 and Figure 6.33). It is covered with thousands of craters and larger basins up to 1300 kilometres in diameter. Some of the brighter craters are rayed, like Tycho and Copernicus on the Moon, and many have central peaks. There are also **scarps** (cliffs) more than a kilometer high and hundreds of kilometres long, as well as ridges and plains.

MESSENGER instruments measured the surface composition and mapped past volcanic activity. One of its most important discoveries was the verification of water ice (first detected by radar) in craters near the poles, similar to the situation on the Moon, and the unexpected discovery of organic (carbon-rich) compounds mixed with the water ice.

Scientists working with data from the MESSENGER mission put together a rotating globe of Mercury, in false colour, showing some of the variations in the composition of the planet's surface. You can watch it spin.

Mercury's Topography

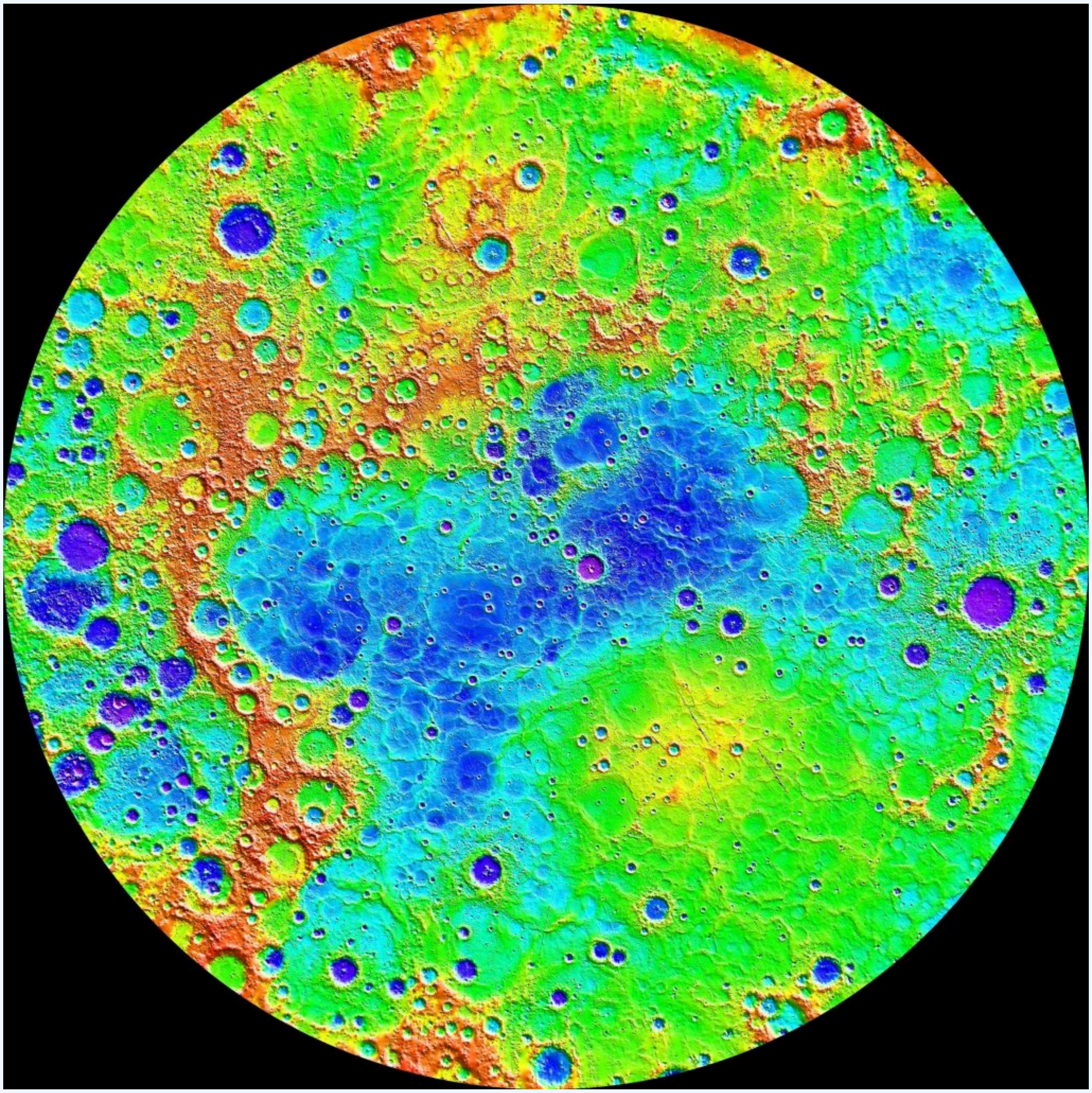


Figure 6.32. The topography of Mercury's northern hemisphere is mapped in great detail from MESSENGER data. The lowest regions are shown in purple and blue, and the highest regions are shown in red. The difference in elevation between the lowest and highest regions shown here is roughly 10 kilometres. The permanently shadowed low-lying craters near the north pole contain radar-bright water ice.
PIA19420: The Ups and Downs of Mercury's Topography by NASA/Johns Hopkins University Applied Physics Laboratory/Carnegie Institution of Washington, NASA Media Licence.

Most of the mercurian features have been named in honour of artists, writers, composers, and other contributors to the arts and humanities, in contrast with the scientists commemorated on the Moon. Among the named craters are Bach, Shakespeare, Tolstoy, Van Gogh, and Scott Joplin.

There is no evidence of plate tectonics on Mercury. However, the planet's distinctive long scarps can sometimes be seen cutting across craters; this means the scarps must have formed later than the craters (Figure 6.33). These long, curved cliffs appear to have their origin in the slight compression of Mercury's crust. Apparently, at some point in its history, the planet shrank, wrinkling the crust, and it must have done so after most of the craters on its surface had already formed.

If the standard cratering chronology applies to Mercury, this shrinkage must have taken place during the last 4 billion years and not during the solar system's early period of heavy bombardment.

Discovery Scarp on Mercury

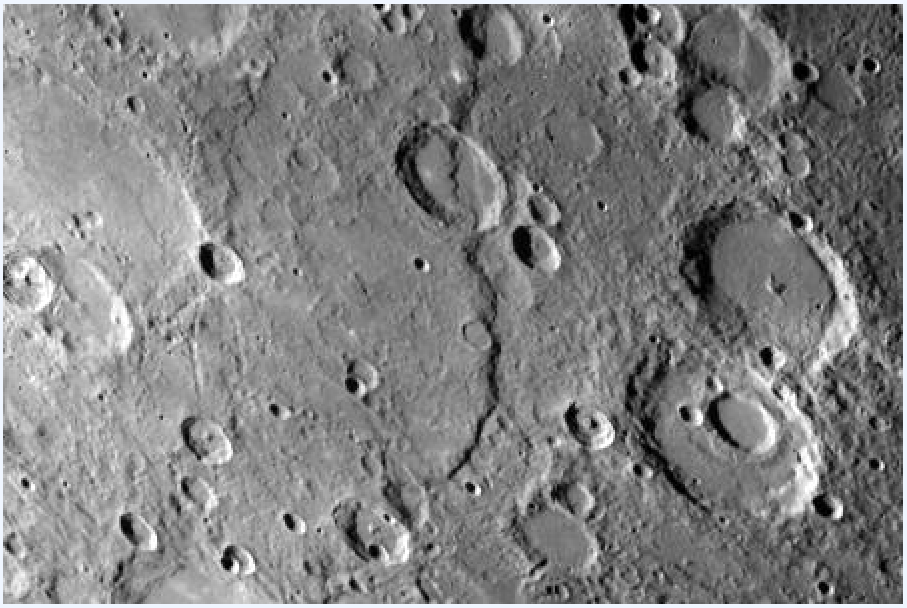


Figure 6.33. This long cliff, nearly 1 kilometre high and more than 100 kilometres long, cuts across several craters. Astronomers conclude that the compression that made "wrinkles" like this in the planet's surface must have taken place after the craters were formed.
Discovery Scarp by NASA/JPL/Northwestern University, NASA JPL Media Licence.

Attribution

6.7 KEY TERMS

Anorthosites: silicate rocks that make up most of the Moon's crust and are made of relatively low-density rock that solidified on the cooling Moon like slag floating on the top of a smelter. 6.3

Bar: a force of 100,000 Newtons acting on a surface area of 1 square meter; the average pressure of Earth's atmosphere at sea level is 1.013 bars. 6.2

Caldera: volcanic crater on Venus. 6.5

Convection: the rising of hot materials. 6.5

Coronae: large circular or oval features on that are produced by bulges that form from the upwelling lava below the surface of a planet. 6.5

Deimos: one of Mars' two small and odd-shaped moons that were most-likely captured asteroids or the result of a collision with Mars. 6.4

Differentiation is the process by which gravity helps separate a planet's interior into layers of different compositions and densities.

Impact basins: huge depressions produced by collisions of large chunks of material with the Moon relatively early in its history. 6.3

Lunar highlands: the lighter, heavily cratered regions of the Moon, which are generally several kilometres higher than the maria. 6.3

Maria: areas of the Moon consisting mostly of flat plains with dark-coloured basalt (volcanic lava). 6.3

Outflow channels: a set of water-related features on Mars that were carved by huge volumes of running water, far too great to be produced by ordinary rainfall. 6.4

Ozone (O_3): a heavy form of oxygen with three atoms per molecule instead of the usual two. 6.2

Permanent (residual) polar caps: polar caps that are always present near the poles on Mars. 6.4

Phobos: the larger of Mars' two small and odd-shaped moons that were most-likely captured asteroids or the result of a collision with Mars. 6.4

Recurring slope lineae: dark streaks on Mars that elongate within a period of a few days, indicating that something is flowing downhill—either water or dark sediment. 6.4

Runoff channels: multitudes of small, sinuous (twisting) channels on Mars that look like what geologists would expect from the surface runoff of ancient rain storms. 6.4

Scarps: cliffs on Mercury that are more than a kilometer high and hundreds of kilometres long. 6.6

Seasonal polar caps: deposits of frozen CO_2 (dry ice) on Mars that condense directly from the atmosphere when the surface temperature drops below about 150 K and develop during the cold martian winters. 6.4

Semimajor axis: a planet's average distance from the Sun (or from its star) on its orbit. 6.6

Smaller gullies: a type of water feature on Mars that suggests intermittent outbreaks of liquid water even today as they have the remarkable property of changing regularly with the martian seasons. 6.4

Stratoshphere: the layer of Earth's atmosphere above the troposphere and below the ionosphere. 6.2

Troposphere: a region of convection in a planet's atmosphere. 6.5

Volatiles: elements and compounds that evaporate at relatively low temperatures. 6.3

CHAPTER 7: THE GAS AND ICE GIANTS

Chapter Overview

7.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Describe the basic characteristics of the giant planets in our solar system, including their distances from the Sun, orbital periods, and comparative sizes in relation to Earth and the Sun.
- Compare and contrast the cloud compositions and weather patterns of Jupiter, Saturn, Uranus, and Neptune.
- Discuss the diverse magnetic tilts observed in the giant planets and explore the implications of these magnetic fields on their respective magnetospheres and charged particle interactions.
- Evaluate the significance of the Voyager missions' trajectory known as the "Grand Tour" and describe how gravity-assisted flybys were used to visit multiple outer planets with a single spacecraft.

7.1 EXPLORING THE OUTER PLANETS

The giant planets hold most of the mass in our planetary system. Jupiter alone exceeds the mass of all the other planets combined (Figure 7.1). The material available to build these planets can be divided into three classes by what they are made of: “gases,” “ices,” and “rocks” (see Table 7.1). The **“gases”** are primarily hydrogen and helium, the most abundant elements in the universe. The way it is used here, the term **“ices”** refers to composition only and not whether a substance is actually in a solid state. “Ices” means compounds that form from the next most abundant elements: oxygen, carbon, and nitrogen. Common ices are water, methane, and ammonia, but ices may also include carbon monoxide, carbon dioxide, and others. **“Rocks”** are even less abundant than ices, and include everything else: magnesium, silicon, iron, and so on.

Jupiter

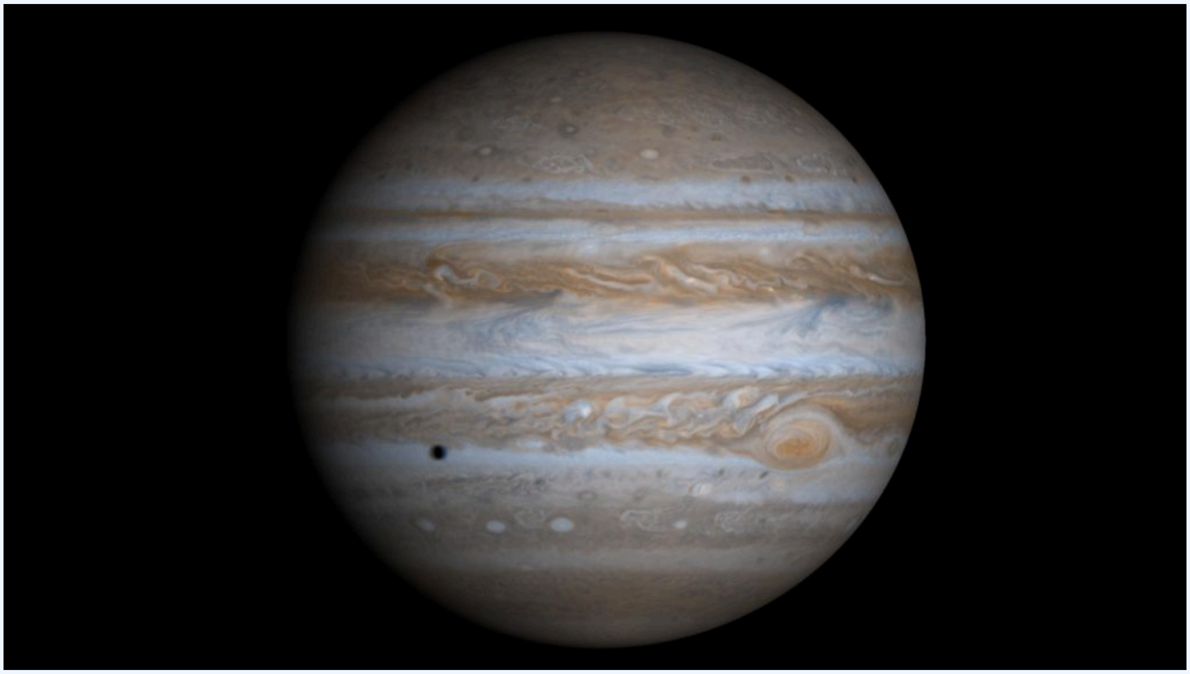


Figure 7.1. The Cassini spacecraft imaged Jupiter on its way to Saturn in 2012. The giant storm system called the Great Red Spot is visible to the lower right. The dark spot to the lower left is the shadow of Jupiter's moon Europa.
PIA02873: High Resolution Globe of Jupiter by NASA/JPL/University of Arizona, NASA Media Licence.

Table 7.1 Abundances in the Outer Solar System

Type of Material	Name	Approximate % (by Mass)
Gas	Hydrogen (H_2)	75
Gas	Helium (He)	24
Ice	Water (H_2O)	0.6
Ice	Methane (CH_4)	0.4
Ice	Ammonia (NH_3)	0.1
Rock	Magnesium (Mg), iron (Fe), silicon (Si)	0.3

In the outer solar system, gases dominate the two largest planets, Jupiter and Saturn, hence their nickname "gas giants.” Uranus and Neptune are sometimes called “ice giants” because their interiors contain far more of the “ice” component than their larger cousins. The chemistry for all four giant planet atmospheres is dominated by hydrogen. This hydrogen caused the chemistry of the outer solar system to become **reducing**, meaning that other elements tend to combine with hydrogen first. In the early solar system, most of the oxygen combined with hydrogen to make H_2O and was thus unavailable to form the kinds of oxidized compounds with other elements that are more familiar to us in the inner solar system (such as CO_2). As a result, the compounds detected in the atmosphere of the giant planets are mostly hydrogen-based gases such as methane (CH_4) and ammonia (NH_3), or more complex hydrocarbons (combinations of hydrogen and carbon) such as ethane (C_2H_6) and acetylene (C_2H_2).

Exploration of the Outer Solar System So Far

Eight spacecraft, seven from the United States and one from Europe, have penetrated beyond the asteroid belt into the realm of the giants. Table 7.2 summarizes the spacecraft missions to the outer solar system.

Table 7.2 Missions to the Giant Planets

Planet	Spacecraft[1]	Encounter Date	Type
Jupiter	Pioneer 10	December 1973	Flyby
	Pioneer 11	December 1974	Flyby
	Voyager 1	March 1979	Flyby
	Voyager 2	July 1979	Flyby
	Ulysses	February 1992	Flyby during gravity assist
	Galileo	December 1995	Orbiter and probe
	Cassini	December 2002	Flyby
	New Horizons	February 2007	Flyby during gravity assist
	Juno	July 2016	Orbiter
Saturn	Pioneer 11	September 1979	Flyby
	Voyager 1	November 1980	Flyby
	Voyager 2	August 1981	Flyby
	Cassini	July 2004	Orbiter
Uranus	Voyager 2	January 1986	Flyby
Neptune	Voyager 2	August 1989	Flyby

The challenges of exploring so far away from Earth are considerable. Flight times to the giant planets are measured in years to decades, rather than the months required to reach Venus or Mars. Even at the speed of light, messages take hours to pass between Earth and the spacecraft. If a problem develops near Saturn, for example, a wait of hours for the alarm to reach Earth and for instructions to be routed back to the spacecraft could spell disaster. Spacecraft to the outer solar system must therefore be highly reliable and capable of a greater degree of independence and autonomy. Outer solar system missions also must carry their own power sources since the Sun is too far away to provide enough energy, or else they must have very large arrays of solar cells. Heaters are required to keep instruments at proper operating temperatures, and spacecraft must have radio transmitters powerful enough to send their data to receivers on distant Earth.

The first spacecraft to investigate the regions past Mars were the NASA Pioneers 10 and 11, launched in 1972 and 1973 as pathfinders to Jupiter. One of their main objectives was simply to determine whether a spacecraft could actually navigate through the belt of asteroids that lies beyond Mars without getting destroyed by collisions with asteroidal dust. Another objective was to measure the radiation hazards in the magnetosphere (or zone of magnetic influence) of Jupiter. Both spacecraft passed through the asteroid belt without incident, but the energetic particles in Jupiter's magnetic field nearly wiped out their electronics, providing information necessary for the safe design of subsequent missions.

Pioneer 10 flew past Jupiter in 1973, after which it sped outward toward the limits of the solar system. Pioneer 11 undertook a more ambitious program, using the gravity of Jupiter to aim for Saturn, which it reached in 1979. The twin Voyager spacecraft launched the next wave of outer planet exploration in 1977. Voyagers 1 and 2 each carried 11 scientific instruments, including cameras and spectrometres, as well as devices to measure the characteristics of planetary magnetospheres. Since they kept going outward after their planetary encounters, these are now the most distant spacecraft ever launched by humanity.

Voyager 1 reached Jupiter in 1979 and used a gravity assist from that planet to take it on to Saturn in 1980. Voyager 2 arrived at Jupiter four months later, but then followed a different path to visit all the outer planets, reaching Saturn in 1981, Uranus in 1986, and Neptune in 1989. This trajectory was made possible by the approximate alignment of the four giant planets on the same side of the Sun. About once every 175 years, these planets are in such a position, and it allows a single spacecraft to visit them all by using gravity-assisted flybys to adjust course for each subsequent encounter; such a maneuver has been nicknamed a "Grand Tour" by astronomers.

The Jet Propulsion Laboratory has a nice video called Voyager: The Grand Tour that describes the Voyager mission and what it found.

Engineering and Space Science: Teaching an Old Spacecraft New Tricks

By the time Voyager 2 arrived at Neptune in 1989, 12 years after its launch, the spacecraft was beginning to show signs of old age. The arm on which the camera and other instruments were located was "arthritic": it could no longer move easily in all directions. The communications system was "hard of hearing": part of its radio receiver had stopped working. The "brains" had significant "memory loss": some of the onboard computer memory had failed. And the whole spacecraft was beginning to run out of energy: its generators had begun showing serious signs of wear.

To make things even more of a challenge, Voyager's mission at Neptune was in many ways the most difficult of all four flybys. For example, since sunlight at Neptune is 900 times weaker than at Earth, the onboard camera had to take much longer exposures in this light-starved environment. This was a nontrivial requirement, given that the spacecraft was hurtling by Neptune at ten times the speed of a rifle bullet.

The solution was to swivel the camera backward at exactly the rate that would compensate for the forward motion of the spacecraft. Engineers had to preprogram the ship's computer to execute an incredibly complex series of maneuvers for each image. The beautiful Voyager images of Neptune are a testament to the ingenuity of spacecraft engineers.

The sheer distance of the craft from its controllers on Earth was yet another challenge. Voyager 2 received instructions and sent back its data via on-board radio transmitter. The distance from Earth to Neptune is about 4.8 billion kilometres. Over this vast distance, the power that reached us from Voyager 2 at Neptune was approximately 10^{-16} watts, or 20 billion times less power than it takes to operate a digital watch. Thirty-eight different antennas on four continents were used by NASA to collect the faint signals from the spacecraft and decode the precious information about Neptune that they contained.

Enter the Orbiters: Galileo, Cassini, and Juno

The Pioneer and Voyager missions were flybys of the giant planets: they each produced only quick looks before the spacecraft sped onward. For more detailed studies of these worlds, we require spacecraft that can go into orbit around a planet. For Jupiter and Saturn, these orbiters were the Galileo, Cassini, and Juno spacecraft. To date, no orbiter missions have been started for Uranus and Neptune, although planetary scientists have expressed keen interest.

The Galileo spacecraft was launched toward Jupiter in 1989 and arrived in 1995. Galileo began its investigations by deploying an entry probe into Jupiter, for the first direct studies of the planet's outer atmospheric layers.

The probe plunged at a shallow angle into Jupiter's atmosphere, travelling at a speed of 50 kilometres *per second*—that's fast enough to fly from New York to San Francisco in 100 seconds! This was the highest speed at which any probe has so far entered the atmosphere of a planet, and it put great demands on the heat shield protecting it. The high entry speed was a result of acceleration by the strong gravitational attraction of Jupiter.

Atmospheric friction slowed the probe within 2 minutes, producing temperatures at the front of its heat shield as high as 15,000 °C. As the probe's speed dropped to 2500 kilometres per hour, the remains of the glowing heat shield were jettisoned, and a parachute was deployed to lower the instrumented probe spacecraft more gently into the atmosphere (Figure 7.2). The data from the probe instruments were relayed to Earth via the main Galileo spacecraft.

Galileo Probe Falling into Jupiter and Juno Image of Jupiter's South Pole

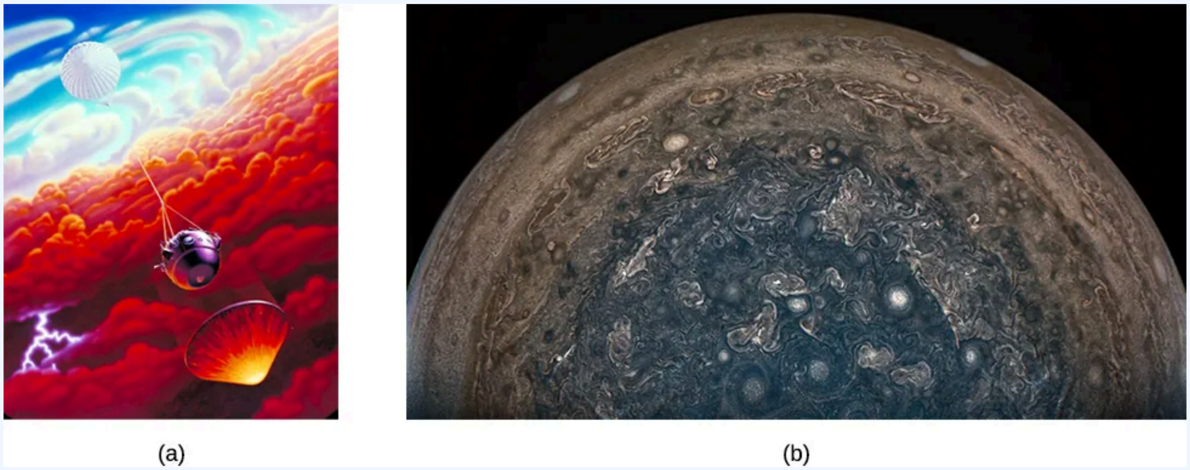


Figure 7.2. (a) This artist's depiction shows the Galileo probe descending into the clouds via parachute just after the protective heat shield separated. The probe made its measurements of Jupiter's atmosphere on December 7, 1995. (b) This Juno image, taken in 2017 from about 100,000 kilometres above the cloudtops, shows the south polar region of Jupiter with its dramatic complex of storms and clouds. The enhanced-colour image was processed for NASA/JPL by citizen scientist John Landino.
(a): Galileo probe – artistic impression by NASA/Ames Research Center/C. Kallas, NASA Media Licence.
(b): PIA21382: Jovian Stormy Weather by Enhanced image by John Landino (CC-BY) based on images provided courtesy of NASA/JPL-Caltech/SwRI/MSSS., NASA Media Licence.

The probe continued to operate for an hour, descending 200 kilometres into the atmosphere. A few minutes later the polyester parachute melted, and within a few hours the main aluminum and titanium structure of the probe vaporized to become a part of Jupiter itself. About 2 hours after receipt of the final probe data, the main spacecraft fired its retro-rockets so it could be captured into orbit around the planet, where its primary objectives were to study Jupiter's large and often puzzling moons.

The Cassini mission to Saturn, a cooperative venture between NASA and the European Space Agency, was similar to Galileo in its two-fold approach. Launched in 1997, Cassini arrived in 2004 and went into orbit around Saturn, beginning extensive studies of its rings and moons, as well as the planet itself. In January 2005, Cassini deployed an entry probe into the atmosphere of Saturn's large moon, Titan, where it successfully landed on the surface.

The Voyager and Galileo missions to Jupiter were primarily designed to study the moons and the atmosphere of the planet. The next NASA mission, an orbiter called Juno, arrived at Jupiter in July 2016. In order to meet its objectives of studying the jovian magnetosphere, it has a very elongated (eccentric) 55-day orbit, that takes it from 4 thousand kilometres above the cloud tops out to 76 thousand kilometres. The orbit takes the craft over Jupiter's poles, giving us remarkable close-ups of the polar regions (previous spacecraft viewed the planet from lower latitudes).

Juno was originally designed without a camera, but fortunately scientists rectified this omission, adding a simple downward-looking colour camera to use during close passes by Jupiter. Recognizing the value of such images, both scientific and artistic, it was decided to post the raw images and encourage "citizen scientists" to process them. The product has been many dramatic, brightly coloured views of Jupiter, such as Figure 7.3.

Earth as Seen from Saturn

Figure 7.3. This popular Cassini image shows Earth as a tiny dot (marked with an arrow) seen below Saturn's rings. It was taken in July 2013, when Saturn was 1.4 billion kilometres from Earth. The Day the Earth Smiled by NASA/JPL-Caltech/Space Science Institute, NASA JPL Media Licence.

Footnotes

[1] Both the Ulysses and the New Horizons spacecraft (designed to study the Sun and Pluto, respectively) flew past Jupiter for a gravity boost (gaining energy by “stealing” a little bit from the giant planet’s rotation)

Attribution

“11.1 Exploring the Outer Planets” from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff,

7.2 THE GIANT PLANETS

Let us now examine the four giant (or **jovian**) planets in more detail. Our approach is not just to catalogue their characteristics, but to compare them with each other, noting their similarities and differences and attempting to relate their properties to their differing masses and distances from the Sun.

Basic Characteristics

The giant planets are very far from the Sun. Jupiter is more than five times farther from the Sun than Earth's distance (5 AU), and takes just under 12 years to circle the Sun. Saturn is about twice as far away as Jupiter (almost 10 AU) and takes nearly 30 years to complete one orbit. Uranus orbits at 19 AU with a period of 84 years, while Neptune, at 30 AU, requires 165 years for each circuit of the Sun. These long timescales make it difficult for us short-lived humans to study seasonal change on the outer planets.

Jupiter and Saturn have many similarities in composition and internal structure, although Jupiter is nearly four times more massive. Uranus and Neptune are smaller and differ in composition and internal structure from their large siblings. Some of the main properties of these four planets are summarized in Table 7.3.

Table 7.3. Basic Properties of the Jovian Planets

Planet	Distance (AU)	Period (years)	Diameter (km)	Mass (Earth = 1)	Density (g/cm^3)	Rotation (hours)
Jupiter	5.2	11.9	142,800	318	1.3	9.9
Saturn	9.5	29.5	120,540	95	0.7	10.7
Uranus	19.2	84.1	51,200	14	1.3	17.2
Neptune	30.0	164.8	49,500	17	1.6	16.1

Jupiter, the giant among giants, has enough mass to make 318 Earths. Its diameter is about 11 times that of Earth (and about one tenth that of the Sun). Jupiter's average density is 1.3 g/cm^3, much lower than that of any of the terrestrial planets. (Recall that water has a density of 1 g/cm^3.) Jupiter's material is spread out over a volume so large that about 1,300 Earths could fit within it.

Saturn's mass is 95 times that of Earth, and its average density is only 0.7 g/cm^3—the lowest of any planet. Since this is less than the density of water, Saturn would be light enough to float.

Uranus and Neptune each have a mass about 15 times that of Earth and, hence, are only 5% as massive as Jupiter. Their densities of 1.3 g/cm^3 and 1.6 g/cm^3, respectively, are much higher than that of Saturn. This is one piece of evidence that tells us that their composition must differ fundamentally from the gas giants. When astronomers began to discover other planetary systems (exoplanets), we found that planets the size of Uranus and Neptune are common, and that there are even more exoplanets intermediate in size between Earth and these ice giants, a type of planet not found in our solar system.

Appearance and Rotation

When we look at the planets, we see only their atmospheres, composed primarily of hydrogen and helium gas. The uppermost clouds of Jupiter and Saturn, the part we see when looking down at these planets from above, are composed of ammonia crystals. On Neptune, the upper clouds are made of methane. On Uranus, we see no obvious cloud layer at all, but only a deep and featureless haze.

Seen through a telescope, Jupiter is a colourful and dynamic planet. Distinct details in its cloud patterns allow us to determine the rotation rate of its atmosphere at the cloud level, although such atmosphere rotation may have little to do with the spin of the underlying planet. Much more fundamental is the rotation of the mantle and core; these can be determined by periodic variations in radio waves coming from Jupiter, which are controlled by its magnetic field. Since the magnetic field (which we will discuss below) originates deep inside the planet, it shares the rotation of the interior. The rotation period we measure in this way is 9 hours 56 minutes, which gives Jupiter the shortest "day" of any planet. In the same way, we can measure that the underlying rotation period of Saturn is 10 hours 40 minutes. Uranus and Neptune have slightly longer rotation periods of about 17 hours, also determined from the rotation of their magnetic fields.

A brief video made from Hubble Space Telescope photos shows the rotation of Jupiter with its many atmospheric features.

Remember that Earth and Mars have seasons because their spin axes, instead of "standing up straight," are tilted relative to the orbital plane of the solar system. This means that as Earth revolves around the Sun, sometimes one hemisphere and sometimes the other "leans into" the Sun.

What are the seasons like for the giant planets? The spin axis of Jupiter is tilted by only 3°, so there are no seasons to speak of. Saturn, however, does have seasons, since its spin axis is inclined at 27° to the perpendicular to its orbit. Neptune has about the same tilt as Saturn (29°); therefore, it experiences similar seasons (only more slowly). The strangest seasons of all are on Uranus, which has a spin axis tilted by 98° with respect to the north direction. Practically speaking, we can say that Uranus orbits on its side, and its ring and moon system follow along, orbiting about Uranus' equator (Figure 7.4).

Infrared Image of Uranus

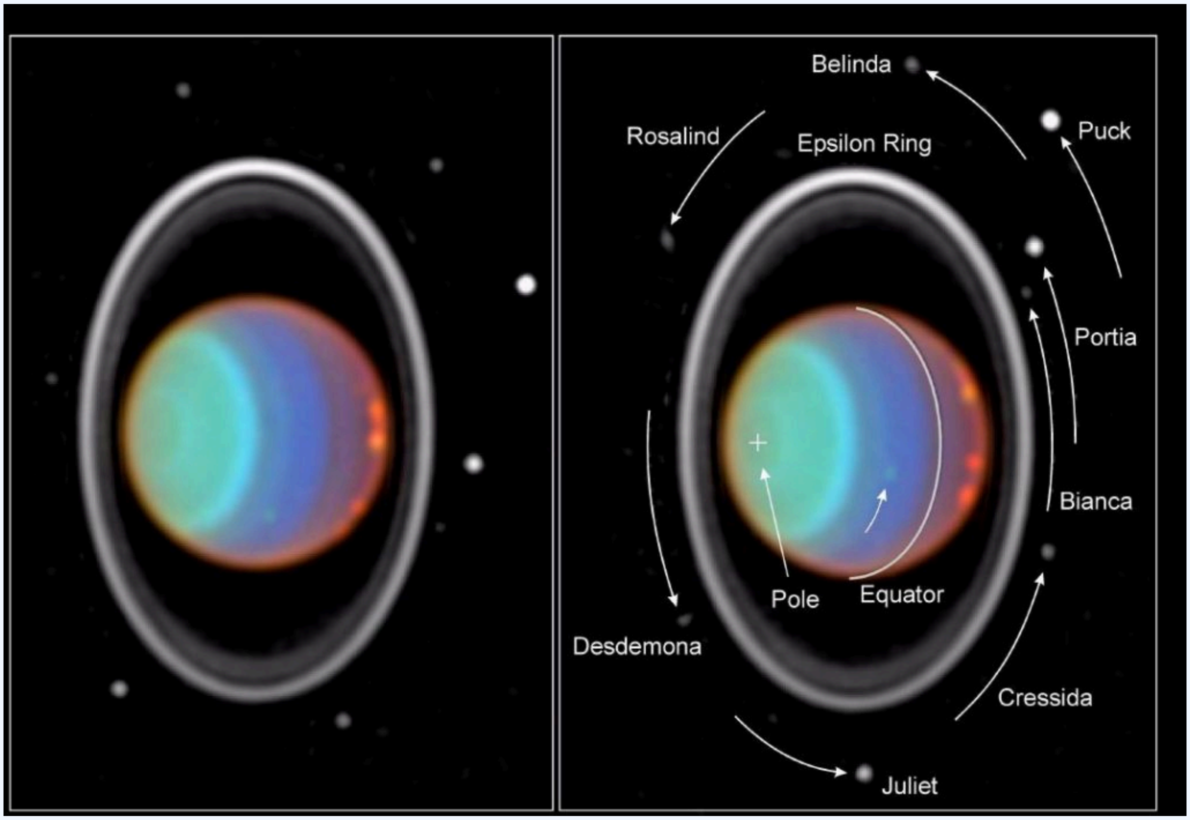


Figure 7.4. The infrared camera on the Hubble Space Telescope took these false-colour images of the planet Uranus, its ring system, and moons in 1997. The south pole of the planet (marked with a "+" on the right image) faces the Sun; its green colour shows a strong local haze. The two images were taken 90 minutes apart, and during that time the five reddish clouds can be seen to rotate around the parallel to the equator. The rings (which are very faint in the visible light, but prominent in infrared) and eight moons can be seen around the equator. This was the "bull's eye" arrangement that Voyager saw as it approached Uranus in 1986.
Hubble Tracks Clouds on Uranus by NASA/JPL/STScI, NASA Media Licence.

We don't know what caused Uranus to be tipped over like this, but one possibility is a collision with a large planetary body when our system was first forming. Whatever the cause, this unusual tilt creates dramatic seasons. When Voyager 2 arrived at Uranus, its south pole was facing directly into the Sun. The southern hemisphere was experiencing a 21-year sunlit summer, while during that same period the northern hemisphere was plunged into darkness. For the next 21-year season, the Sun shines on Uranus' equator, and both hemispheres go through cycles of light and dark as the planet rotates (Figure 7.5). Then there are 21 years of an illuminated northern hemisphere and a dark southern hemisphere. After that the pattern of alternating day and night repeats.

Just as on Earth, the seasons are even more extreme at the poles. If you were to install a floating platform at the south pole of Uranus, for example, it would experience 42 years of light and 42 years of darkness. Any future astronauts crazy enough to set up camp there could spend most of their lives without ever seeing the Sun.

Strange Seasons on Uranus

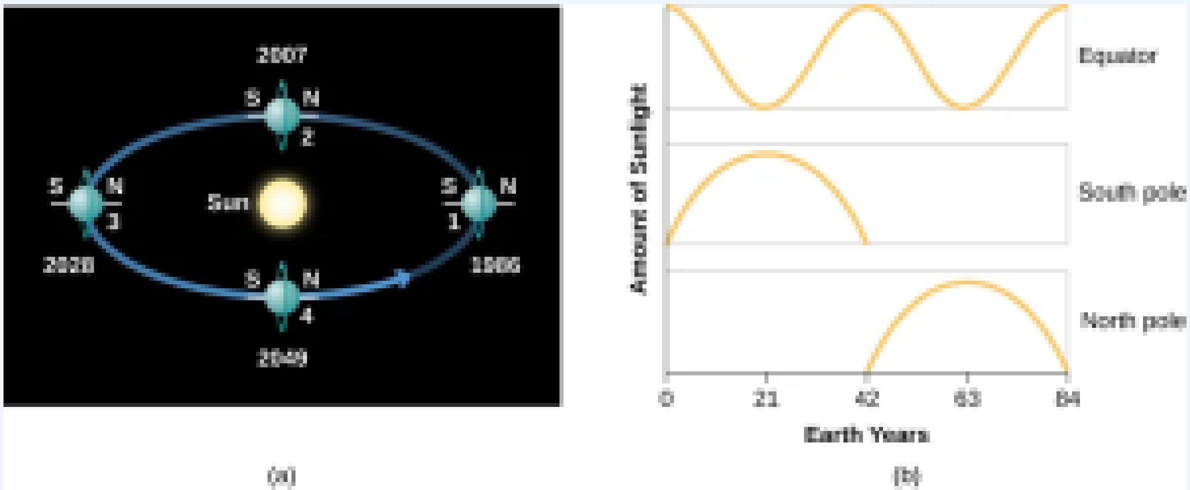


Figure 7.5. (a) This diagram shows the orbit of Uranus as seen from above. At the time Voyager 2 arrived (position 1), the South Pole was facing the Sun. As we move counterclockwise in the diagram, we see the planet 21 years later at each step. (b) This graph compares the amount of sunlight seen at the poles and the equator of Uranus over the course of its 84-year revolution around the Sun.

Composition and Structure

Although we cannot see into these planets, astronomers are confident that the interiors of Jupiter and Saturn are composed primarily of hydrogen and helium. Of course, these gases have been measured only in their atmosphere, but calculations first carried out more than 50 years ago showed that these two light gases are the only possible materials out of which a planet with the observed masses and densities of Jupiter and Saturn could be constructed.

The deep internal structures of these two planets are difficult to predict. This is mainly because these planets are so big that the hydrogen and helium in their centers become tremendously compressed and behave in ways that these gases can never behave on Earth. The best theoretical models we have of Jupiter's structure predict a central pressure greater than 100 million bars and a central density of about 31 g/cm^3. (By contrast, Earth's core has a central pressure of 4 million bars and a central density of 17 g/cm^3.)

At the pressures inside the giant planets, familiar materials can take on strange forms. A few thousand kilometres below the visible clouds of Jupiter and Saturn, pressures become so great that hydrogen changes from a gaseous to a liquid state. Still deeper, this liquid hydrogen is further compressed and begins to act like a metal, something it never does on Earth. (In a metal, electrons are not firmly attached to their parent nuclei but can wander around. This is why metals are such good conductors of electricity.) On Jupiter, the greater part of the interior is liquid metallic hydrogen.

Because Saturn is less massive, it has only a small volume of metallic hydrogen, but most of its interior is liquid. Uranus and Neptune are too small to reach internal pressures sufficient to liquefy hydrogen. We will return to the discussion of the metallic hydrogen layers when we examine the magnetic fields of the giant planets.

Each of these planets has a core composed of heavier materials, as demonstrated by detailed analyses of their gravitational fields. Presumably these cores are the original rock-and-ice bodies that formed before the capture of gas from the surrounding nebula. The cores exist at pressures of tens of millions of bars. While scientists speak of the giant planet cores being composed of rock and ice, we can be sure that neither rock nor ice assumes any familiar forms at such pressures and temperatures. Remember that what is really meant by "rock" is any material made up primarily of iron, silicon, and oxygen, while the term "ice" in this chapter denotes materials composed primarily of the elements carbon, nitrogen, and oxygen in combination with hydrogen.

Figure 7.6 illustrates the likely interior structures of the four jovian planets. It appears that all four have similar cores of rock and ice. On Jupiter and Saturn, the cores constitute only a few percent of the total mass, consistent with the initial composition of raw materials shown in Table 7.1. However, most of the mass of Uranus and Neptune resides in these cores, demonstrating that the two outer planets were unable to attract massive quantities of hydrogen and helium when they were first forming.

Internal Structures of the Jovian Planets

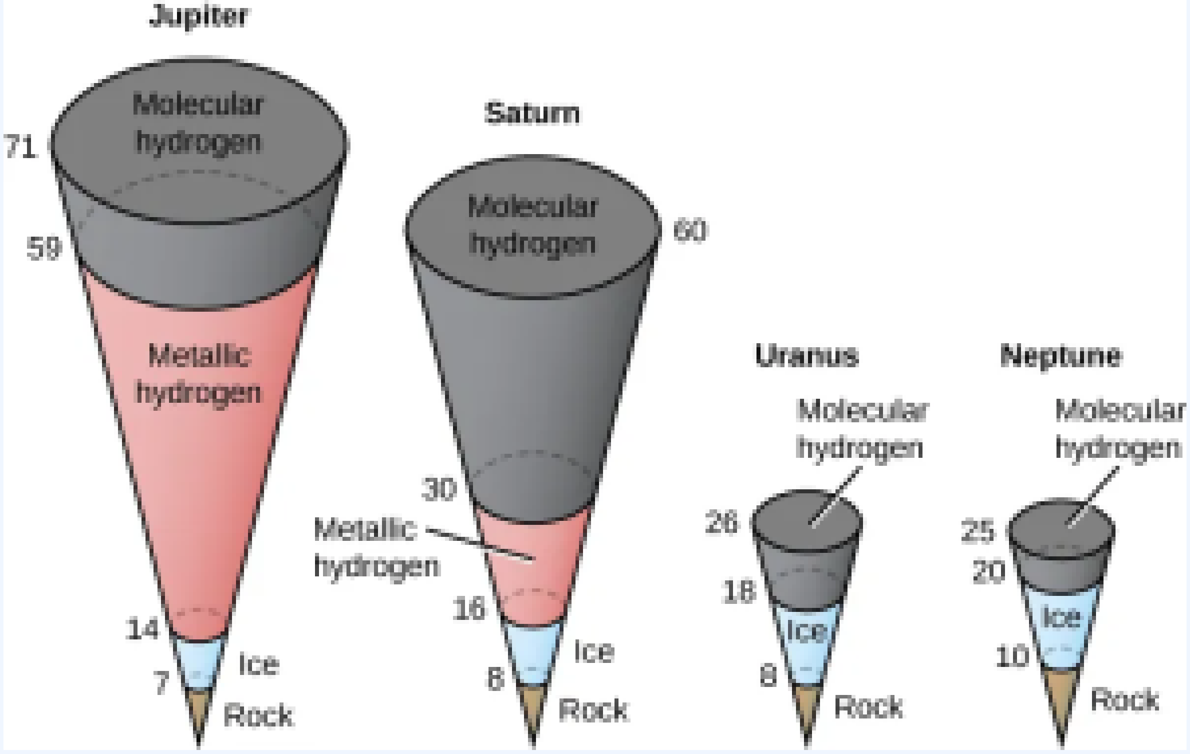


Figure 7.6. Jupiter and Saturn are composed primarily of hydrogen and helium (but hydrogen dominates), but Uranus and Neptune consist in large part of compounds of carbon, nitrogen, and oxygen. (The diagrams are drawn to scale; numbers show radii in thousands of kilometres.)

Internal Heat Sources

Because of their large sizes, all the giant planets were strongly heated during their formation by the collapse of surrounding material onto their cores. Jupiter, being the largest, was the hottest. Some of this primordial heat can still remain inside such large planets. In addition, it is possible for giant, largely gaseous planets to generate heat after formation by slowly contracting. (With so large a mass, even a minuscule amount of shrinking can generate significant heat.) The effect of these internal energy sources is to raise the temperatures in the interiors and atmospheres of the planets higher than we would expect from the heating effect of the Sun alone.

Jupiter has the largest internal energy source, amounting to 4×10^{17} watts; that is, it is heated from inside with energy equivalent to 4 million billion 100-watt lightbulbs. This energy is about the same as the total solar energy absorbed by Jupiter. The atmosphere of Jupiter is therefore something of a cross between a normal planetary atmosphere (like Earth's), which obtains most of its energy from the Sun, and the atmosphere of a star, which is entirely heated by an internal energy source. Most of the internal energy of Jupiter is primordial heat, left over from the formation of the planet 4.5 billion years ago.

Saturn has an internal energy source about half as large as that of Jupiter, which means (since its mass is only about one quarter as great) that it is producing twice as much energy per kilogram of material as does Jupiter. Since Saturn is expected to have much less primordial heat, there must be another source at work generating most of this 2×10^{17} watts of power. This source is the separation of helium from hydrogen in Saturn's interior. In the liquid hydrogen mantle, the heavier helium forms droplets that sink toward the core, releasing gravitational energy. In effect, Saturn is still differentiating—letting lighter material rise and heavier material fall.

Uranus and Neptune are different. Neptune has a small internal energy source, while Uranus does not emit a measurable amount of internal heat. As a result, these two planets have almost the same atmospheric temperature, in spite of Neptune's greater distance from the Sun. No one knows why these two planets differ in their internal heat, but all this shows how nature can contrive to make each world a little bit different from its neighbors.

Magnetic Fields

Each of the giant planets has a strong magnetic field, generated by electric currents in its rapidly spinning interior. Associated with the magnetic fields are the planets' **magnetospheres**, which are regions around the planet within which the planet's own magnetic field dominates over the general interplanetary magnetic field. The magnetospheres of these planets are their largest features, extending millions of kilometres into space.

In the late 1950s, astronomers discovered that Jupiter was a source of radio waves that got more intense at longer rather than at shorter wavelengths—just the reverse of what is expected from thermal radiation (radiation caused by the normal vibrations of particles within all matter). Such behavior is typical, however, of the radiation emitted when high-speed electrons are accelerated by a magnetic field. We call this synchrotron radiation because it was first observed on Earth in particle accelerators, called synchrotrons. This was our first hint that Jupiter must have a strong magnetic field.

Later observations showed that the radio waves are coming from a region surrounding Jupiter with a diameter several times that of the planet itself (Figure 7.7). The evidence suggested that a vast number of charged atomic particles must be circulating around Jupiter, spiraling around the lines of force of a magnetic field associated with the planet. This is just what we observe happening, but on a smaller scale, in the Van Allen belt around Earth. The magnetic fields of Saturn, Uranus, and Neptune, discovered by the spacecraft that first passed close to these planets, work in a similar way, but are not as strong.

Jupiter in Radio Waves

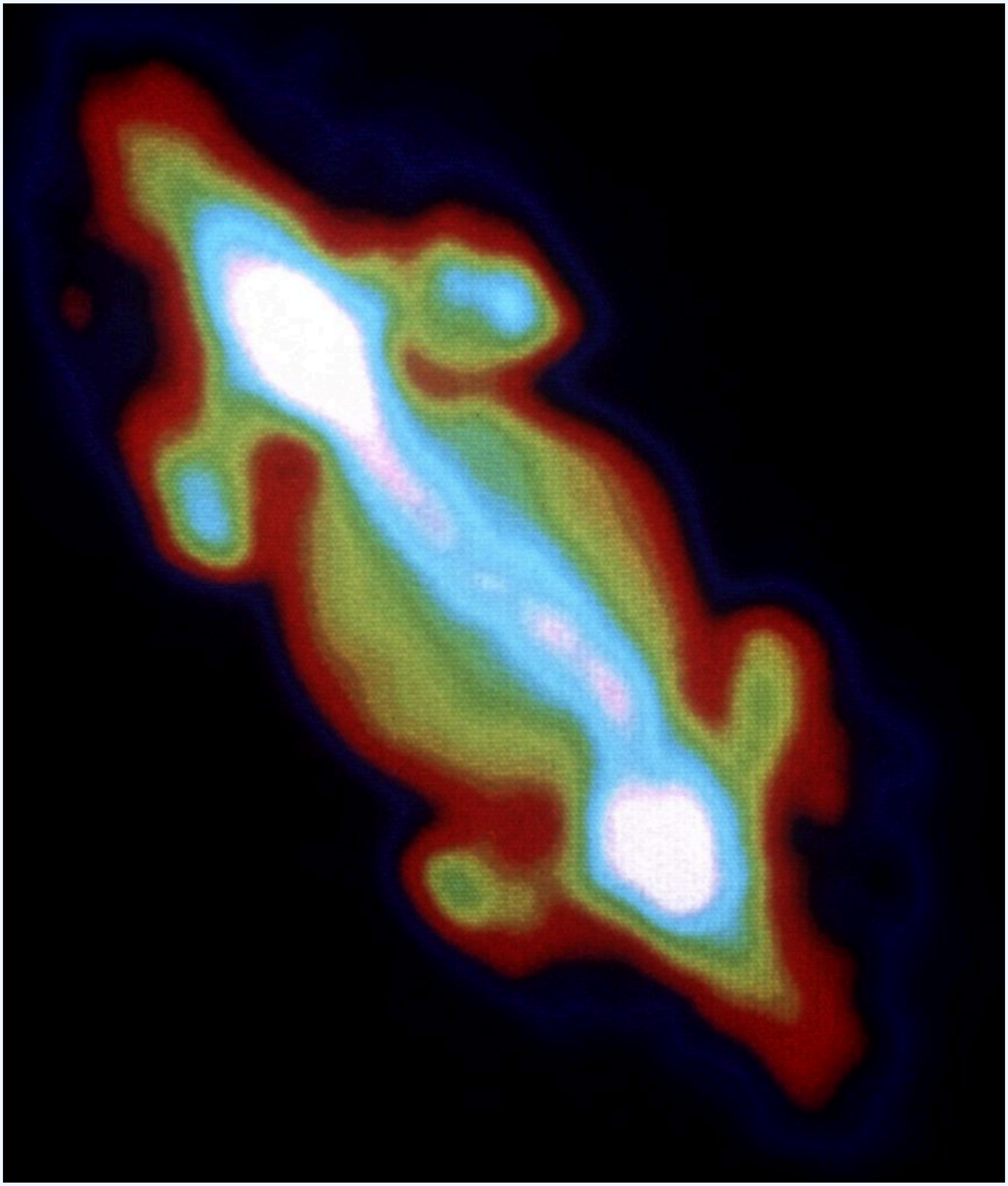


Figure 7.7. This false-colour image of Jupiter was made with the Very Large Array (of radio telescopes) in New Mexico. We see part of the magnetosphere, brightest in the middle because the largest number of charged particles are in the equatorial zone of Jupiter. The planet itself is slightly smaller than the green oval in the center. Different colours are used to indicate different intensities of synchrotron radiation. Jupiter in Radio Waves by NRAO/AUI/NSF, NRAO Media Licence.

Learn more about the magnetosphere of Jupiter and why we continue to be interested in it from this brief NASA video.

Inside each magnetosphere, charged particles spiral around in the magnetic field; as a result, they can be accelerated to high energies. These charged particles can come from the Sun or from the neighborhood of the planet itself. In Jupiter's case, Io, one of its moons, turns out to have volcanic eruptions that blast charged particles into space and right into the jovian magnetosphere.

The axis of Jupiter's magnetic field (the line that connects the magnetic north pole with the magnetic south pole) is not aligned exactly with the axis of rotation of the planet; rather, it is tipped by about 10°. Uranus and Neptune have even greater magnetic tilts, of 60° and 55°, respectively. Saturn's field, on the other hand, is perfectly aligned with its rotation axis. Why different planets have such different magnetic tilts is not well understood.

The physical processes around the jovian planets turn out to be milder versions of what astronomers find in many distant objects, from the remnants of dead stars to the puzzling distant powerhouses we call quasars. One reason to study the magnetospheres of the giant planets and Earth is that they provide nearby accessible analogues of more energetic and challenging cosmic processes.

Attribution

7.3 ATMOSPHERES OF THE GIANT PLANETS

The atmospheres of the jovian planets are the parts we can observe or measure directly. Since these planets have no solid surfaces, their atmospheres are more representative of their general compositions than is the case with the terrestrial planets. These atmospheres also present us with some of the most dramatic examples of weather patterns in the solar system. As we will see, storms on these planets can grow bigger than the entire planet Earth.

Atmospheric Composition

When sunlight reflects from the atmospheres of the giant planets, the atmospheric gases leave their "fingerprints" in the spectrum of light. Spectroscopic observations of the jovian planets began in the nineteenth century, but for a long time, astronomers were not able to interpret the spectra they observed. As late as the 1930s, the most prominent features photographed in these spectra remained unidentified. Then better spectra revealed the presence of molecules of methane (CH_4) and ammonia (NH_3) in the atmospheres of Jupiter and Saturn.

At first astronomers thought that methane and ammonia might be the main constituents of these atmospheres, but now we know that hydrogen and helium are actually the dominant gases. The confusion arose because neither hydrogen nor helium possesses easily detected spectral features in the visible spectrum. It was not until the Voyager spacecraft measured the far-infrared spectra of Jupiter and Saturn that a reliable abundance for the elusive helium could be found.

The compositions of the two atmospheres are generally similar, except that on Saturn there is less helium as the result of the precipitation of helium that contributes to Saturn's internal energy source. The most precise measurements of composition were made on Jupiter by the Galileo entry probe in 1995; as a result, we know the abundances of some elements in the jovian atmosphere even better than we know those in the Sun.

Voyagers in Astronomy

James Van Allen: Several Planets under His Belt

The career of physicist James Van Allen spanned the birth and growth of the space age, and he played a major role in its development. Born in Iowa in 1914, Van Allen received his PhD from the University of Iowa. He then worked for several research institutions and served in the Navy during World War II.

After the war, Van Allen (Figure 7.8) was appointed Professor of Physics at the University of Iowa. He and his collaborators began using rockets to explore cosmic radiation in Earth's outer atmosphere. To reach extremely high altitudes, Van Allen designed a technique in which a balloon lifts and then launches a small rocket (the rocket is nicknamed "the rockoon").

James Van Allen (1914–2006)

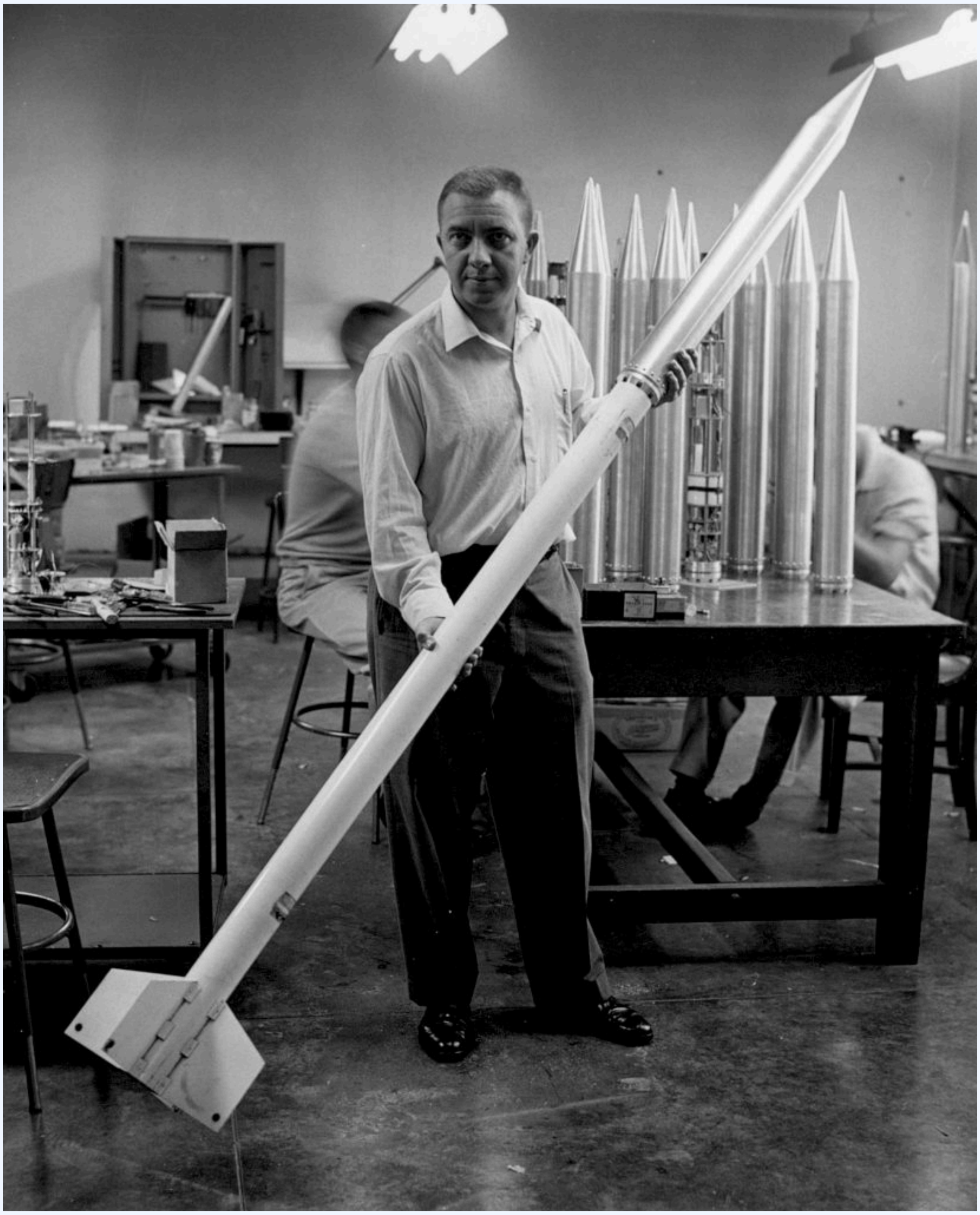


Figure 7.8. In this 1950s photograph, Van Allen holds a "rockoon."
1955. James Van Allen holding a Loki rockoon in an Iowa fabrication lab courtesy of University of Iowa

James A. Van Allen Collection

Over dinner one night in 1950, Van Allen and several colleagues came up with the idea of the International Geophysical Year (IGY), an opportunity for scientists around the world to coordinate their investigations of the physics of Earth, especially research done at high altitudes. In 1955, the United States and the Soviet Union each committed themselves to launching an Earth-orbiting satellite during IGY, a competition that began what came to be known as the space race. The IGY (stretched to 18 months) took place between July 1957 and December 1958.

The Soviet Union won the first lap of the race by launching Sputnik 1 in October 1957. The US government spurred its scientists and engineers to even greater efforts to get something into space to maintain the country's prestige. However, the primary US satellite program, Vanguard, ran into difficulties: each of its early launches crashed or exploded. Simultaneously, a second team of rocket engineers and scientists had quietly been working on a military launch vehicle called Jupiter-C. Van Allen spearheaded the design of the instruments aboard a small satellite that this vehicle would carry. On January 31, 1958, Van Allen's Explorer 1 became the first US satellite in space.

Unlike Sputnik, Explorer 1 was equipped to make scientific measurements of high-energy charged particles above the atmosphere. Van Allen and his team discovered a belt of highly charged particles surrounding Earth, and these belts now bear his name. This first scientific discovery of the space program made Van Allen's name known around the world.

Van Allen and his colleagues continued to measure the magnetic and particle environment around planets with increasingly sophisticated spacecraft, including Pioneers 10 and 11, which made exploratory surveys of the environments of Jupiter and Saturn. Some scientists refer to the charged-particle zones around those planets as Van Allen belts as well. (Once, when Van Allen was giving a lecture at the University of Arizona, the graduate students in planetary science asked him if he would leave his belt at the school. It is now proudly displayed as the university's "Van Allen belt.")

Van Allen was a strong supporter of space science and an eloquent senior spokesperson for the

American scientific community, warning NASA not to put all its efforts into human spaceflight, but to also use robotic spacecraft as productive tools for space exploration.

Clouds and Atmospheric Structure

The clouds of Jupiter (Figure 7.9) are among the most spectacular sights in the solar system, much beloved by makers of science-fiction films. They range in colour from white to orange to red to brown, swirling and twisting in a constantly changing kaleidoscope of patterns. Saturn shows similar but much more subdued cloud activity; instead of vivid colours, its clouds have a nearly uniform butterscotch hue (Figure 7.12).

Jupiter's Colourful Clouds

Figure 7.9. The vibrant colours of the clouds on Jupiter present a puzzle to astronomers: given the cool temperatures and the composition of nearly 90% hydrogen, the atmosphere should be colourless. One hypothesis suggests that perhaps colourful hydrogen compounds rise from warm areas. The actual colours are a bit more muted.
PIA00014: Jupiter Great Red Spot by NASA/JPL, NASA Media Licence.

Different gases freeze at different temperatures. At the temperatures and pressures of the upper atmospheres of Jupiter and Saturn, methane remains a gas, but ammonia can condense and freeze. (Similarly, water vapor condenses high in Earth's atmosphere to produce clouds of ice crystals.) The primary clouds that we see around these planets, whether from a spacecraft or through a telescope, are composed of frozen ammonia crystals. The ammonia clouds mark the upper edge of the planets' tropospheres; above that is the stratosphere, the coldest part of the atmosphere.

Saturn over Five Years

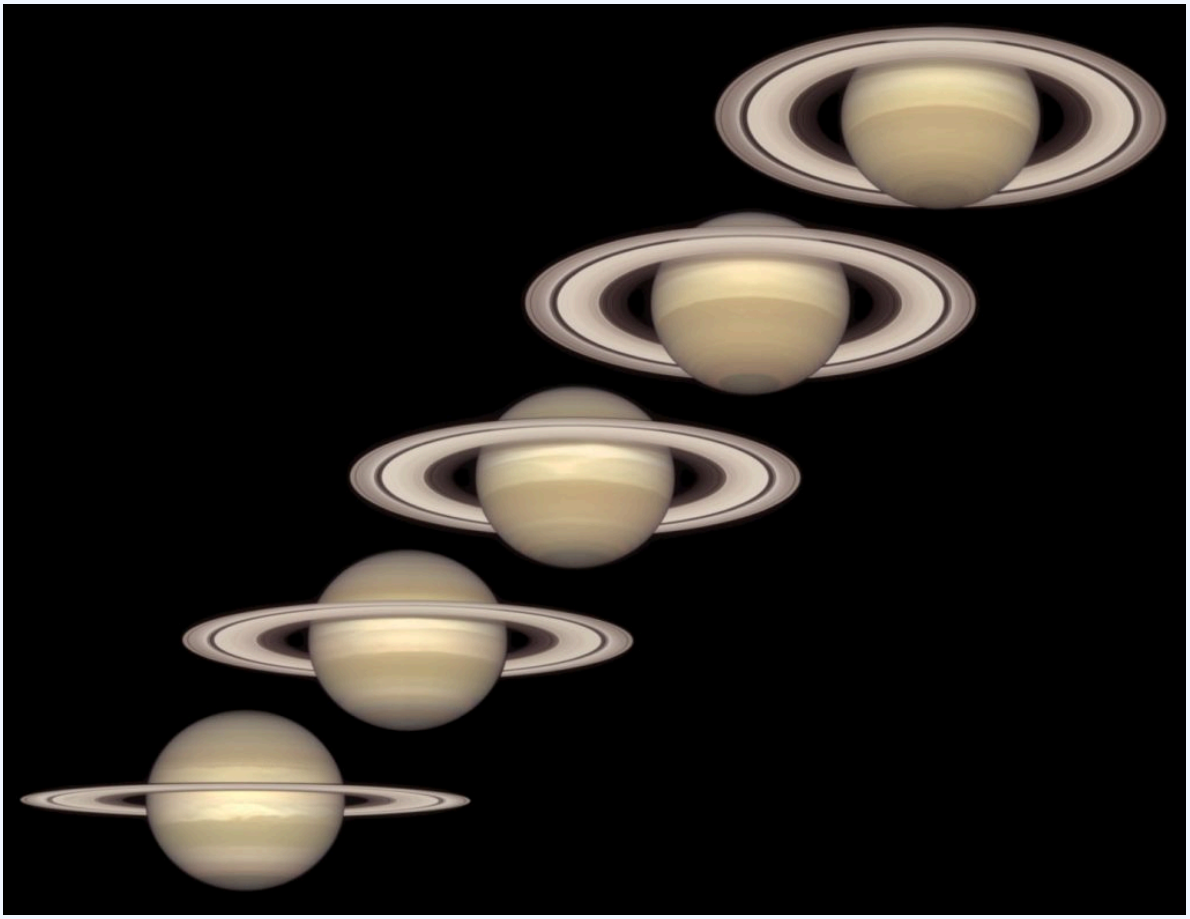


Figure 7.10. These beautiful images of Saturn were recorded by the Hubble Space Telescope between 1996 and 2000. Since Saturn is tilted by 27°, we see the orientation of Saturn's rings around its equator change as the planet moves along its orbit. Note the horizontal bands in the atmosphere.
Saturn from 1996 to 2000 by NASA/Hubble, NASA Hubble Media Licence.

The diagrams in Figure 7.11 show the structure and clouds in the atmospheres of all four jovian planets. On both Jupiter and Saturn, the temperature near the cloud tops is about 140 K (only a little cooler than the polar caps of Mars). On Jupiter, this cloud level is at a pressure of about 0.1 bar (one tenth the atmospheric pressure at the surface of Earth), but on Saturn it occurs lower in the atmosphere, at about 1 bar. Because the ammonia clouds lie so much deeper on Saturn, they are more difficult to see, and the overall appearance of the planet is much blander than is Jupiter's appearance.

Atmospheric Structure of the Jovian Planets

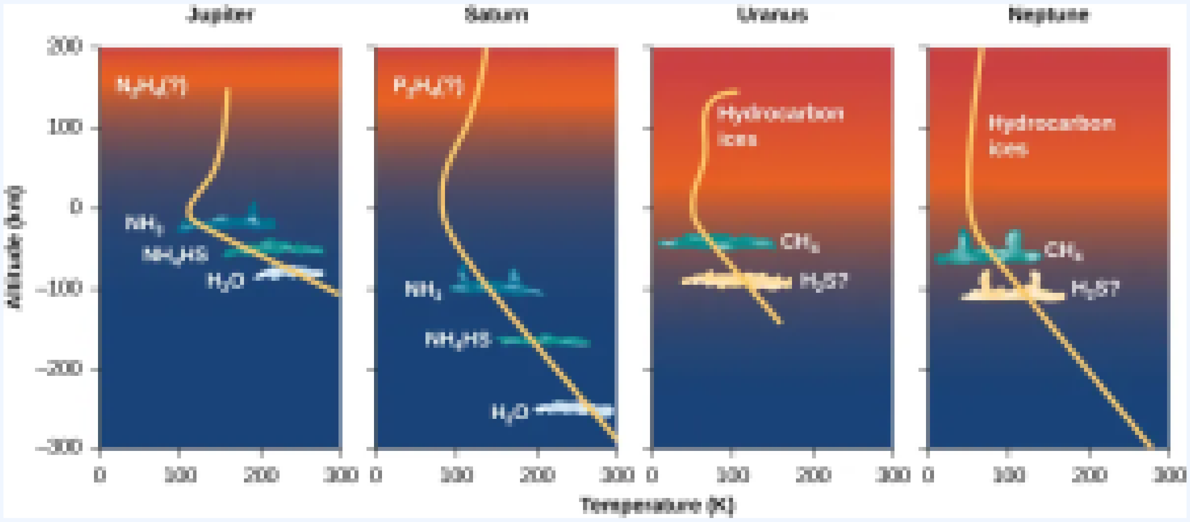


Figure 7.11. In each diagram, the yellow line shows how the temperature (see the scale on the bottom) changes with altitude (see the scale at the left). The location of the main layers on each planet is also shown.

Within the tropospheres of these planets, the temperature and pressure both increase with depth. Through breaks in the ammonia clouds, we can see tantalizing glimpses of other cloud layers that can form in these deeper regions of the atmosphere—regions that were sampled directly for Jupiter by the Galileo probe that fell into the planet.

As it descended to a pressure of 5 bars, the probe should have passed into a region of frozen water clouds, then below that into clouds of liquid water droplets, perhaps similar to the common clouds of the terrestrial troposphere. At least this is what scientists expected. But the probe saw no water clouds, and it measured a surprisingly low abundance of water vapor in the atmosphere. It soon became clear to the Galileo scientists that the probe happened to descend through an unusually dry, cloud-free region of the atmosphere—a giant downdraft of cool, dry gas. Andrew Ingersoll of Caltech, a member of the Galileo team, called this entry site the "desert" of Jupiter. It's a pity that the probe did not enter a more representative region, but that's the luck of the cosmic draw. The probe continued to make measurements to a pressure of 22 bars but found no other cloud layers before its instruments stopped working. It also detected lightning storms, but only at great distances, further suggesting that the probe itself was in a region of clear weather.

Above the visible ammonia clouds in Jupiter's atmosphere, we find the clear stratosphere, which reaches a minimum temperature near 120 K. At still higher altitudes, temperatures rise again, just as they do in the upper atmosphere of Earth, because here the molecules absorb ultraviolet light from the Sun. The cloud colours are due to impurities, the product of chemical reactions among the atmospheric gases in a process we call photochemistry. In Jupiter's upper atmosphere, photochemical reactions create a variety of fairly complex compounds of hydrogen and carbon that form a thin layer of smog far above the visible clouds. We show this smog as a fuzzy orange region in Figure 7.11; however, this thin layer does not block our view of the clouds beneath it.

The visible atmosphere of Saturn is composed of approximately 75% hydrogen and 25% helium, with trace amounts of methane, ethane, propane, and other hydrocarbons. The overall structure is similar to that of Jupiter. Temperatures are somewhat colder, however, and the atmosphere is more extended due to Saturn's lower surface gravity. Thus, the layers are stretched out over a longer distance, as you can see in Figure 7.11. Overall, though, the same atmospheric regions, condensation cloud, and photochemical reactions that we see on Jupiter should be present on Saturn (Figure 7.12).

Cloud Structure on Saturn

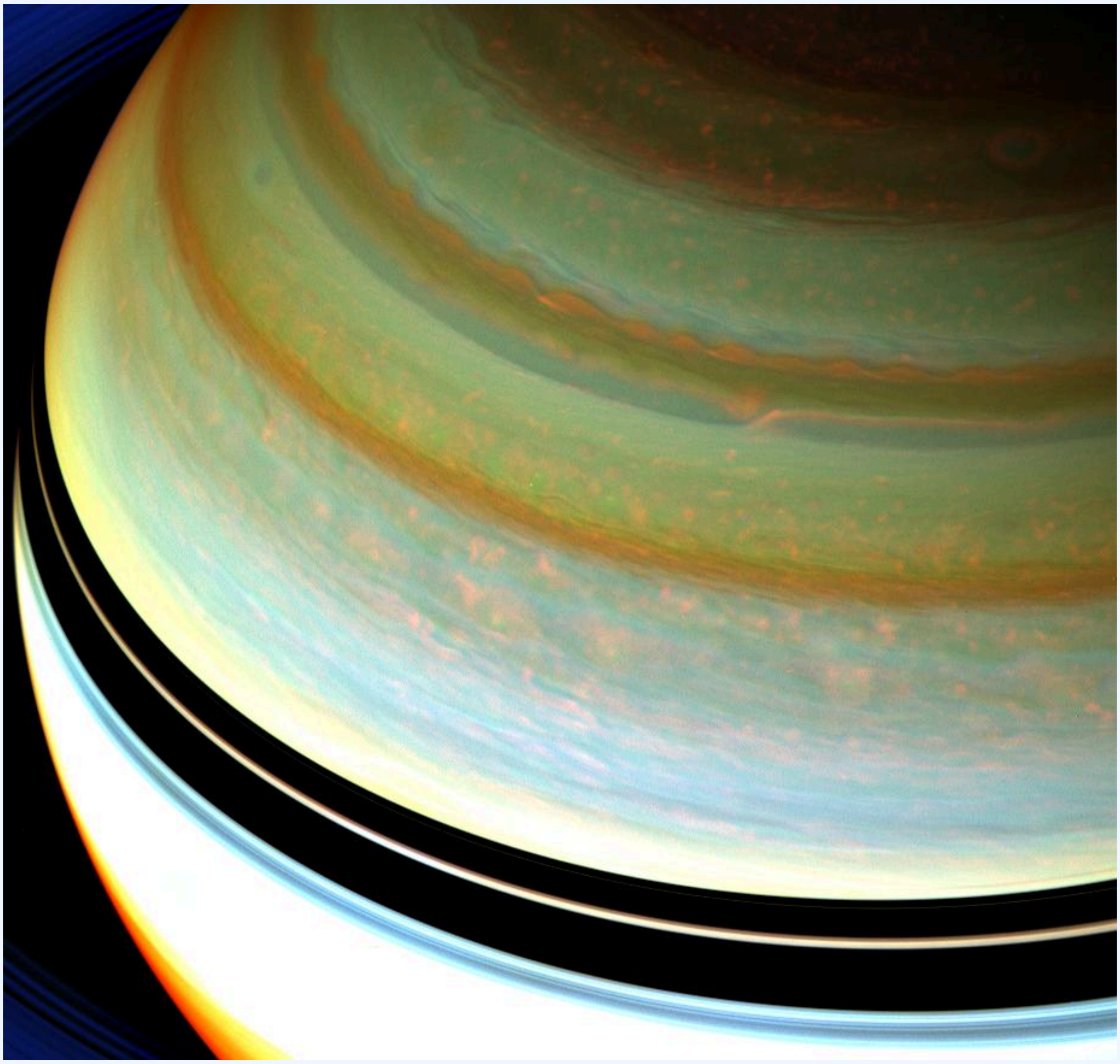


Figure 7.12.I n this Cassini image, colours have been intensified, so we can see the bands and zones and storms in the atmosphere. The dark band is the shadow of the rings on the planet.
Strong Jet in False Colours by NASA/JPL-Caltech/Space Science Institute, NASA JPL Media Licence.

Saturn has one anomalous cloud structure that has mystified scientists: a hexagonal wave pattern around the north pole, shown in Figure 7.13. The six sides of the hexagon are each longer than the diameter of Earth.

Winds are also extremely high on Saturn, with speeds of up to 1800 kilometres per hour measured near the equator.

Hexagon Pattern on Saturn's North Pole

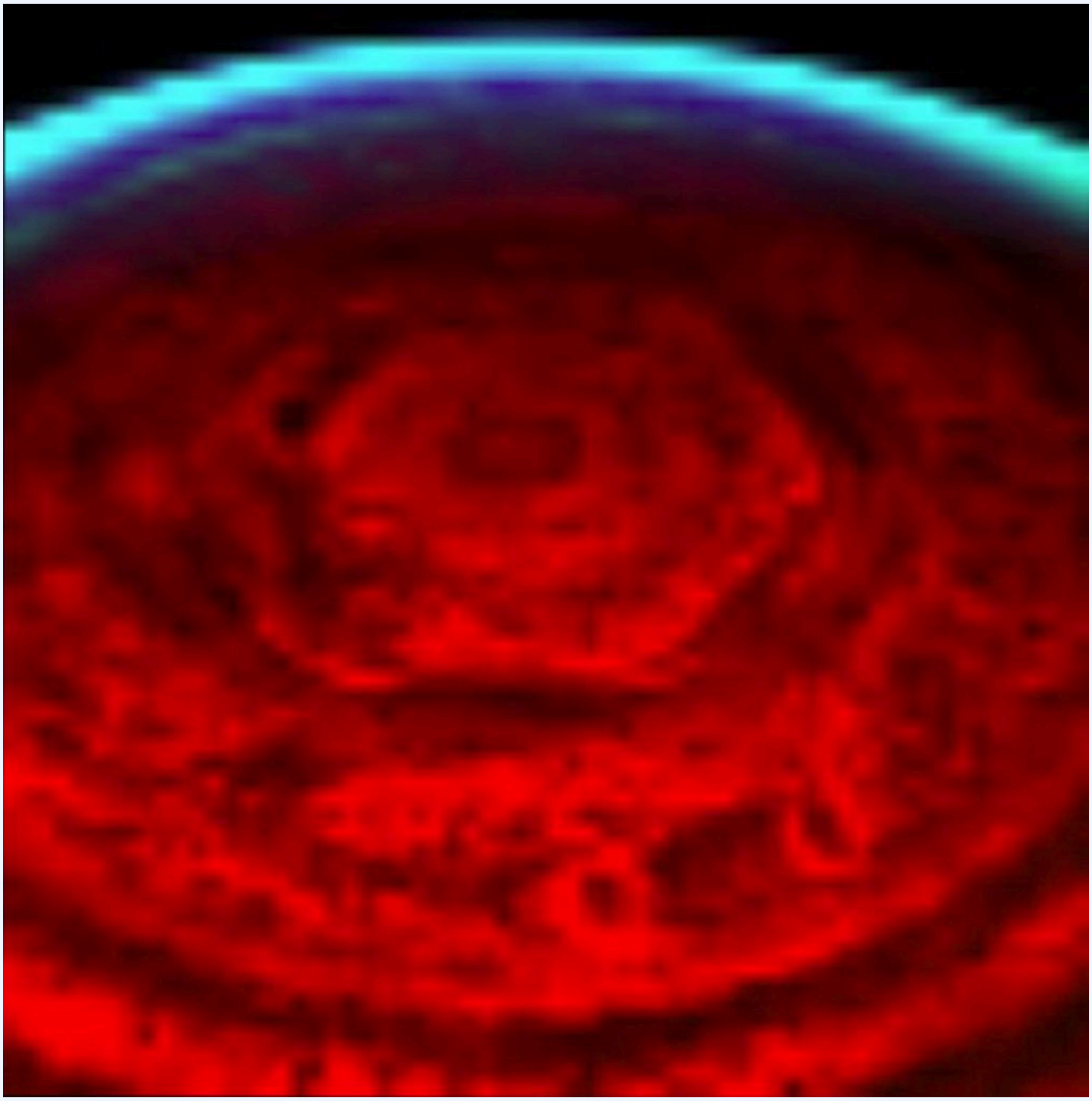


Figure 7.13. In this infrared nighttime image from the Cassini mission, the path of Saturn's hexagonal jet stream is visible as the planet's north pole emerges from the darkness of winter.
PIA09186: Saturn's Strange Hexagon by NASA/JPL/University of Arizona, NASA Media Licence.

See images of Saturn's hexagon with exaggerated colour in this brief NASA video.

Unlike Jupiter and Saturn, Uranus is almost entirely featureless as seen at wavelengths that range from the ultraviolet to the infrared. Calculations indicate that the basic atmospheric structure of Uranus should resemble that of Jupiter and Saturn, although its upper clouds (at the 1-bar pressure level) are composed of methane rather than ammonia. However, the absence of an internal heat source suppresses up-and-down movement and leads to a very stable atmosphere with little visible structure.

Neptune differs from Uranus in its appearance, although their basic atmospheric temperatures are similar. The upper clouds are composed of methane, which forms a thin cloud layer near the top of the troposphere at a temperature of 70 K and a pressure of 1.5 bars. Most the atmosphere above this level is clear and transparent, with less haze than is found on Uranus. The scattering of sunlight by gas molecules lends Neptune a pale blue colour similar to that of Earth's atmosphere (Figure 7.14). Another cloud layer, perhaps composed of hydrogen sulfide ice particles, exists below the methane clouds at a pressure of 3 bars.

Neptune

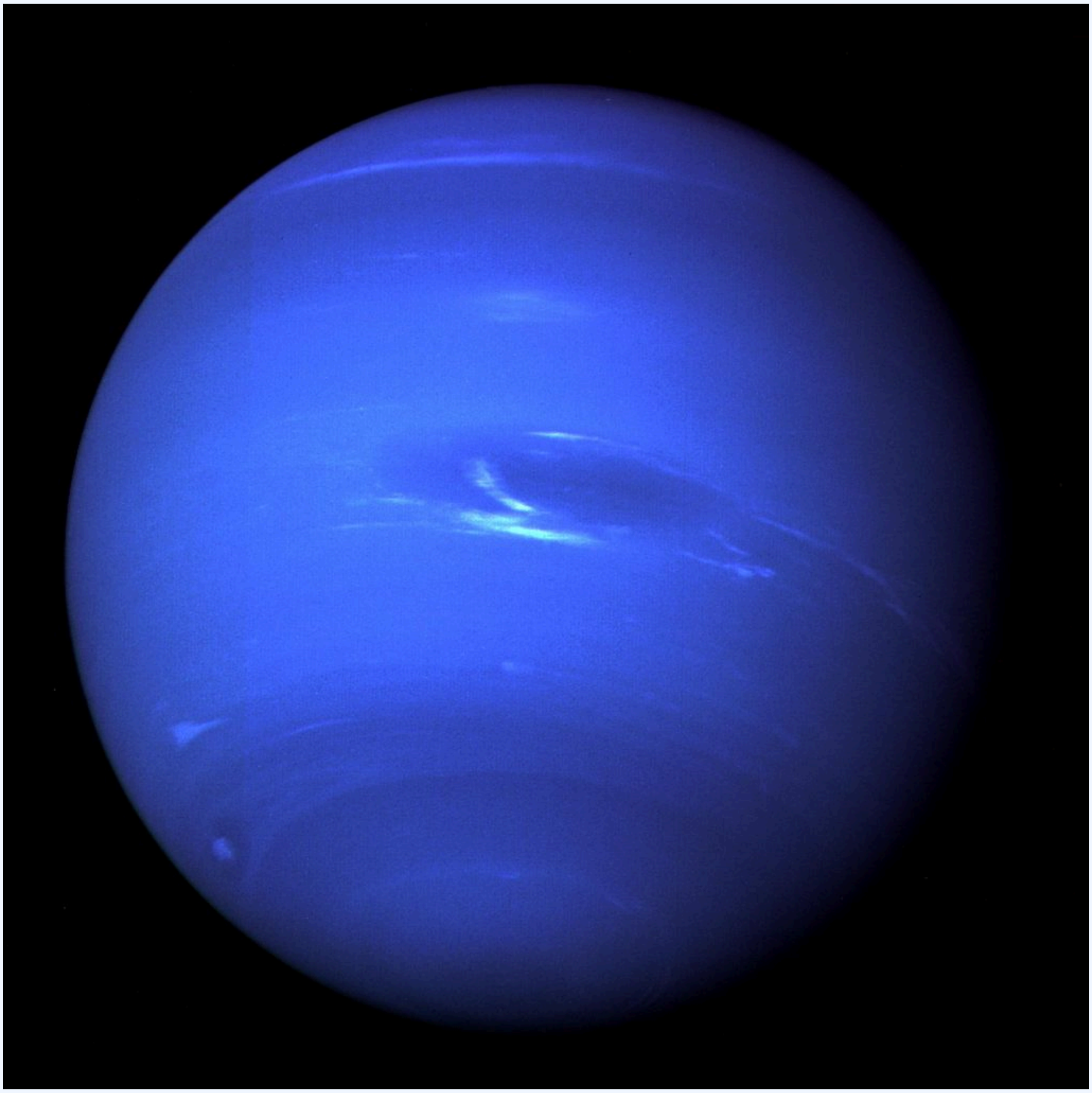


Figure 7.14. The planet Neptune is seen here as photographed by Voyager in 1989. The blue colour, exaggerated with computer processing, is caused by the scattering of sunlight in the planet's upper atmosphere.
Neptune Full Disk View by NASA/JPL, NASA JPL Media Licence.

Unlike Uranus, Neptune has an atmosphere in which convection currents—vertical drafts of gas—emanate from the interior, powered by the planet's internal heat source. These currents carry warm gas above the 1.5-bar cloud level, forming additional clouds at elevations about 75 kilometres higher. These high-altitude clouds form bright white patterns against the blue planet beneath. Voyager photographed distinct shadows on the methane cloud tops, permitting the altitudes of the high clouds to be calculated. Figure 7.15 is a remarkable close-up of Neptune's outer layers that could never have been obtained from Earth.

High Clouds in the Atmosphere of Neptune

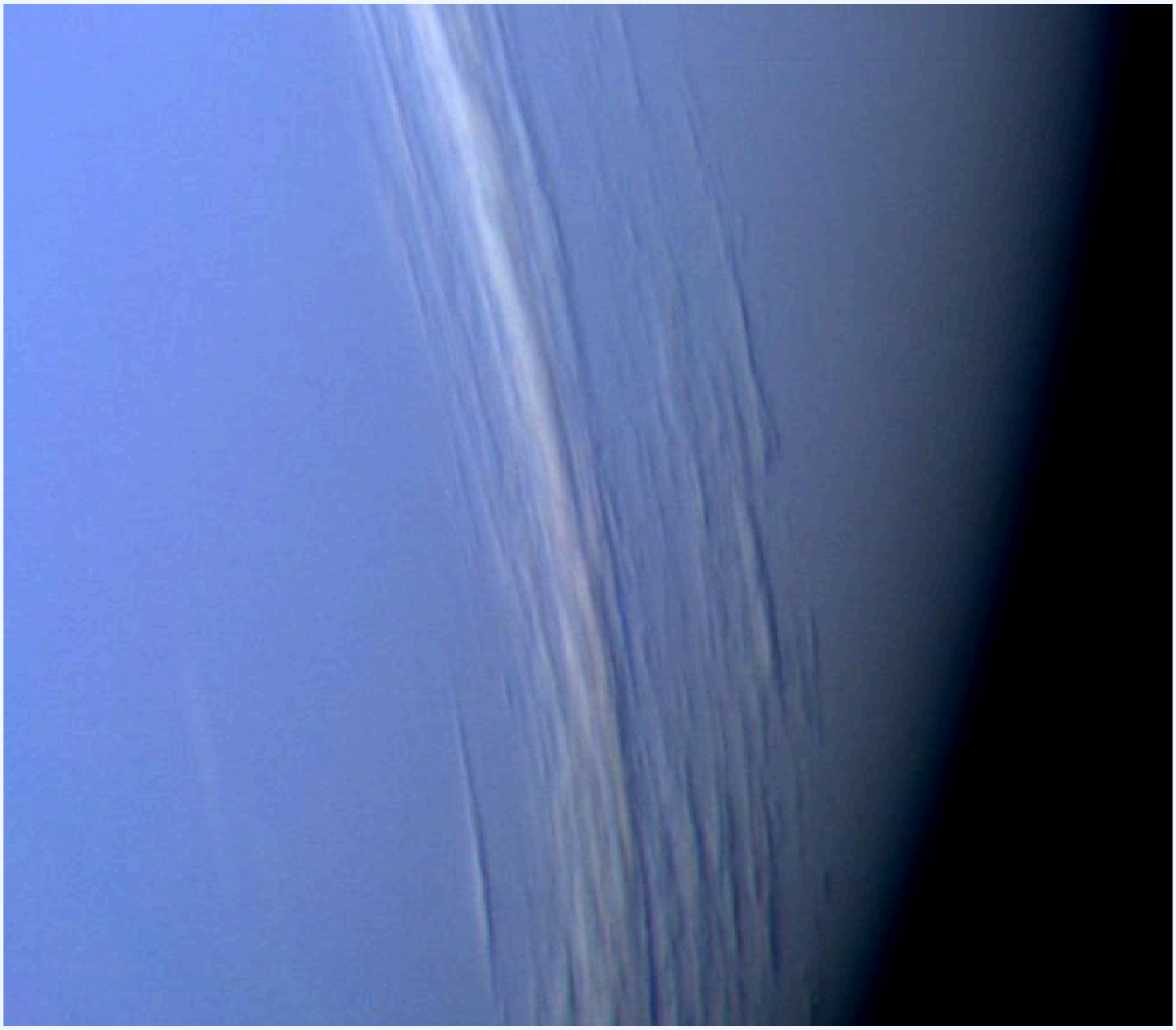


Figure 7.15. These bright, narrow cirrus clouds are made of methane ice crystals. From the shadows they cast on the thicker cloud layer below, we can measure that they are about 75 kilometres higher than the main clouds.
PIA00058: Neptune Clouds Showing Vertical Relief by NASA/JPL, NASA Media Licence.

Giant Storms on Giant Planets

The largest and most famous of Jupiter's storms is the Great Red Spot, a reddish oval in the southern hemisphere that changes slowly; it was 25,000 kilometres long when Voyager arrived in 1979, but it had shrunk to 20,000 kilometres by the end of the Galileo mission in 2000 (Figure 7.17). The giant storm has persisted in Jupiter's atmosphere ever since astronomers were first able to observe it after the invention of the telescope, more than 300 years ago. However, it has continued to shrink and has become more nearly circular, raising speculation that we may see its end within a few decades. Measurements from the Juno spacecraft of variations in Jupiter's gravity field indicate that the depth of the Red Spot storm system is only a few hundred kilometres—a finding that challenges some of the models astronomers have made of this long-lasting weather system.

Storms on Jupiter

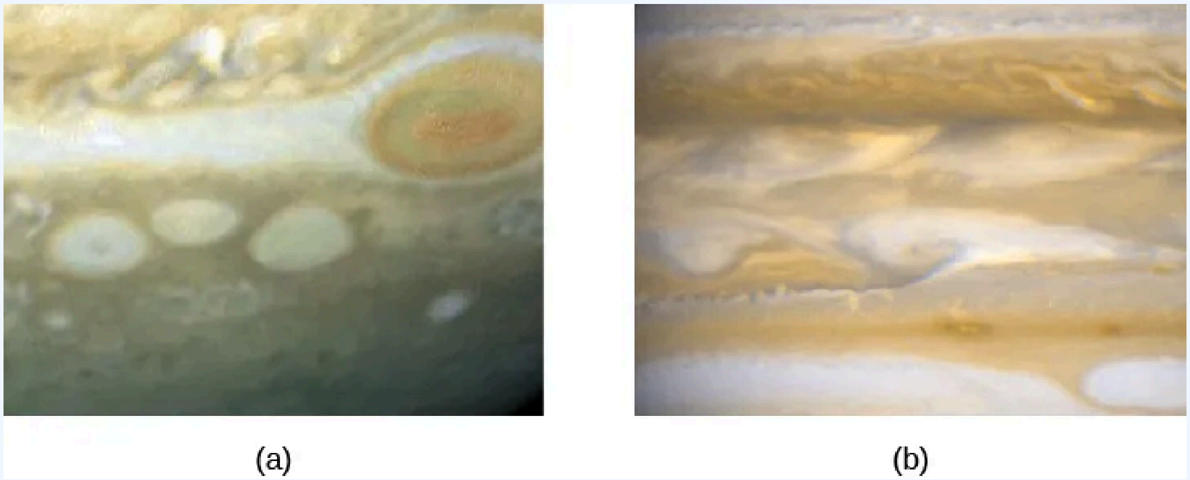


Figure 7.16. Two examples of storms on Jupiter illustrate the use of enhanced colour and contrast to bring out faint features. (a) The three oval-shaped white storms below and to the left of Jupiter's Great Red Spot are highly active, and moved closer together over the course of seven months between 1994 and 1995. (b) The clouds of Jupiter are turbulent and ever-changing, as shown in this Hubble Space Telescope image from 2007.
(a): PIA01262: Hubble Tracks Jupiter Storms by NASA/ New Mexico State Univ., NASA Media Licence.
(b): Jupiter – March 25, 2007 (Full Field) by NASA, ESA and A. Simon-Miller (NASA Goddard Space Flight Center), NASA Hubble Media Licence.

Jupiter's Great Red Spot

Figure 7.17. This is the largest storm system on Jupiter, as seen during the Voyager spacecraft flyby. Below and to the right of the Red Spot is one of the white ovals, which are similar but smaller high-pressure features. The white oval is roughly the size of planet Earth, to give you a sense of the huge scale of the weather patterns we are seeing. The colours on the Jupiter image have been somewhat exaggerated here so astronomers (and astronomy students) can study their differences more effectively. PIA01384: Jupiter's Great Red Spot by NASA/JPL, NASA Media Licence.

In addition to its longevity, the Red Spot differs from terrestrial storms in being a high-pressure region; on our planet, such storms are regions of lower pressure. The Red Spot's counterclockwise rotation has a period of six days. Three similar but smaller disturbances (about as big as Earth) formed on Jupiter in the 1930s. They look like white ovals, and one can be seen clearly below and to the right of the Great Red Spot in Figure 7.17. In 1998, the Galileo spacecraft watched as two of these ovals collided and merged into one.

We don't know what causes the Great Red Spot or the white ovals, but we do have an idea how they can last so long once they form. On Earth, the lifetime of a large oceanic hurricane or typhoon is typically a few weeks, or even less when it moves over the continents and encounters friction with the land. Jupiter has no solid surface to slow down an atmospheric disturbance; furthermore, the sheer size of the disturbances lends them stability. We can calculate that on a planet with no solid surface, the lifetime of anything as large as the Red Spot should be measured in centuries, while lifetimes for the white ovals should be measured in decades, which is pretty much what we have observed.

Despite Neptune's smaller size and different cloud composition, Voyager showed that it had an atmospheric feature surprisingly similar to Jupiter's Great Red Spot. Neptune's Great Dark Spot was nearly 10,000 kilometres long. On both planets, the giant storms formed at latitude 20° S, had the same shape, and took up about the same fraction of the planet's diameter. The Great Dark Spot rotated with a period of 17 days, versus about 6 days for the Great Red Spot. When the Hubble Space Telescope examined Neptune in the mid-1990s, however, astronomers could find no trace of the Great Dark Spot on their images.

Although many of the details of the weather on the jovian planets are not yet understood, it is clear that if you are a fan of dramatic weather, these worlds are the place to look. We study the features in these atmospheres not only for what they have to teach us about conditions in the jovian planets, but also because we hope they can help us understand the weather on Earth just a bit better.

Example 7.1

Storms and Winds

The wind speeds in circular storm systems can be formidable on both Earth and the giant planets. Think about our big terrestrial hurricanes. If you watch their behavior in satellite images shown on weather outlets, you will see that they require about one day to rotate. If a storm has a diameter of $400\ \text{km}$ and rotates once in $24\ \text{h}$, what is the wind speed?

Solution

Speed equals distance divided by time. The distance in this case is the circumference ($2\pi R$ or πd), or approximately $1250\ \text{km}$, and the time is $24\ \text{h}$, so the speed at the edge of the storm would be about $52\ \text{km/h}$. Toward the center of the storm, the wind speeds can be much higher.

Exercise 7.1

Jupiter's Great Red Spot rotates in 6 d and has a circumference equivalent to a circle with radius $10,000 \text{ km}$. Calculate the wind speed at the outer edge of the spot.

Solution

For the Great Red Spot of Jupiter, the circumference $(2\pi R)$ is about $63,000 \text{ km}$. Six d equals 144 h, suggesting a speed of about 436 km/h. This is much faster than wind speeds on Earth.

Footnotes

[2] Recall from earlier chapters that convection is a process in which liquids, heated from underneath, have regions where hot material rises and cooler material descends. You can see convection at work if you heat oatmeal on a stovetop or watch miso soup boil.

Attribution

7.4 KEY TERMS

Gases: a class of material available to build the giant planets that consists primarily of hydrogen and helium, the most abundant elements in the universe. 7.1

Ices: a class of material available to build the giant planets which consist of compounds that form from the next most abundant elements: oxygen, carbon, and nitrogen. 7.1

Jovian planets: the four giant planets: Jupiter, Saturn, Uranus, and Neptune. 7.2

Reducing (chemistry): the chemistry present in the outer solar system where other elements tend to combine with hydrogen first. 7.1

Rocks: a class of material available to build the giant planets that consist of even less abundant elements than those of ices, and include everything else: magnesium, silicon, iron, and so on. 7.1

Magnetospheres: regions around the planet within which the planet's own magnetic field dominates over the general interplanetary magnetic field. 7.2

CHAPTER 8: OTHER OBJECTS IN THE SOLAR SYSTEM

Chapter Overview

8.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Describe the key characteristics of the Galilean moons of Jupiter, including their distances from Jupiter, sizes, compositions, surface features, and geological activities.
- Explain the concept of tidal heating and its role in shaping geological activity.
- Evaluate the potential for life on certain moons in the solar system based on the evidence of subsurface oceans and other factors.
- Compare and contrast the ring systems of Saturn, Uranus, and Neptune, highlighting their differences in mass, structure, and composition.
- Outline the differences between asteroids and comets and their roles in the solar system.
- Analyze the impact of asteroid collisions on planetary formation and the history of Earth.

8.1 THE GALILEAN MOONS OF JUPITER

From 1996 to 1999, the Galileo spacecraft careered through the jovian system on a complex but carefully planned trajectory that provided repeated close encounters with the large Galilean moons. More recently, the Juno spacecraft, also a Jupiter orbiter, has provided a few close looks at Ganymede and Europa. (Beginning in 2004, we received an even greater bonanza of information about Titan, obtained from the Cassini spacecraft and its Huygens probe, which landed on its surface. We include Titan, Saturn's one big moon, here for comparison.) Table 8.1 summarizes some basic facts about these large moons (plus our own Moon for comparison).

Table 8.1. The Largest Moons

Name	Diameter (km)	Mass (Earth's Moon = 1)	Density (g/cm^3)	Reflectivity (%)
Moon	3476	1.0	3.3	12
Callisto	4820	1.5	1.8	20
Ganymede	5270	2.0	1.9	40
Europa	3130	0.7	3.0	70
Io	3640	1.2	3.5	60
Titan	5150	1.9	1.9	20

Callisto: An Ancient, Primitive World

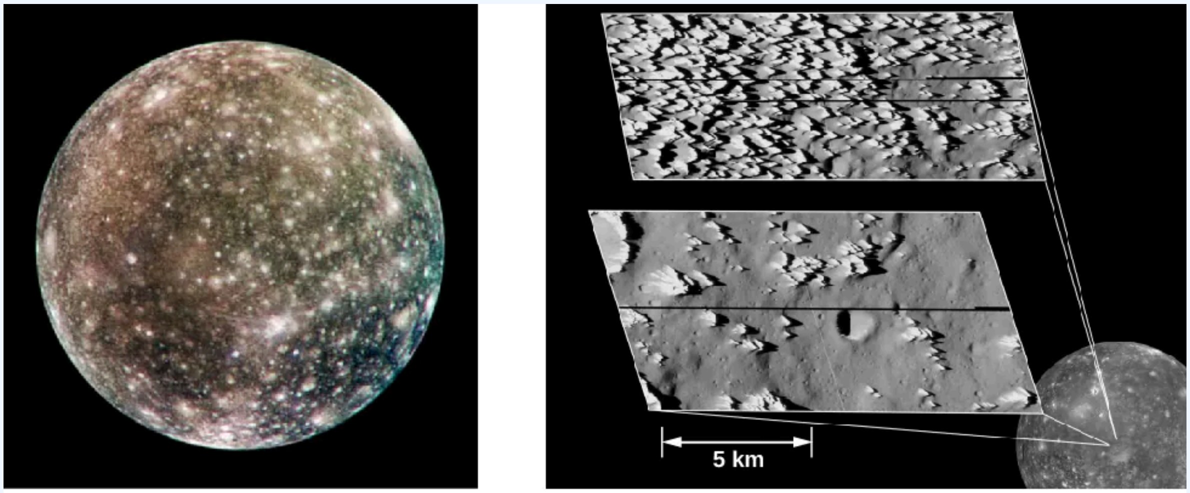
We begin our discussion of the Galilean moons with the outermost one, Callisto, not because it is remarkable but because it is not. This makes it a convenient object with which other, more active, worlds can be compared. Its distance from Jupiter is about 2 million kilometres, and it orbits the planet in 17 days. Like our own Moon, Callisto rotates in the same period as it revolves, so it always keeps the same face toward Jupiter. Callisto's day thus equals its month: 17 days. Its noontime surface temperature is only 130 K (about 140 °C below freezing), so that water ice is stable (it never evaporates) on its surface year round.

Callisto has a diameter of 4820 kilometres, almost the same as the planet Mercury (Figure 8.1). Yet its mass is only one-third as great, which means its density (the mass divided by the volume) must be only one-third as great as well. This tells us that Callisto has far less of the rocky and metallic materials found in the inner planets and must instead be an icy body through much of its interior. Callisto can show us how the geology of an icy object compares with those made primarily of rock.

Unlike the worlds we have studied so far, Callisto has not fully **differentiated** (separated into layers of different density materials). We can tell that it lacks a dense core from the details of its gravitational pull on the Galileo spacecraft. This surprised scientists, who expected that all the big icy moons would be differentiated. It should be easier for an icy body to differentiate than for a rocky one because the melting temperature of ice is so low. Only a little heating will soften the ice and get the process started, allowing the rock and metal to sink to the center while the slushy ice floats to the surface. Yet Callisto seems to have frozen solid before the process of differentiation was complete.

The surface of Callisto is covered with impact craters, like the lunar highlands. The survival of these craters tells us that an icy object can retain impact craters on its surface. Callisto is unique among the planet-sized objects of the solar system in the apparent absence of interior forces to drive geological change. You might say that this moon was stillborn, and it has remained geologically dead for more than 4 billion years (Figure 8.1).

Callisto



(a) (b)

Figure 8.1. (a) Jupiter's outermost large moon shows a heavily cratered surface. Astronomers believe that the bright areas are mostly ice, while the darker areas are more eroded, ice-poor material. (b) These high-resolution images, taken by NASA's Galileo spacecraft in May 2001, show the icy spires (top) on Callisto's surface, with darker dust that has slid down as the ice erodes, collecting in the low-lying areas. The spires are about 80 to 100 metres tall. As the surface erodes even further, the icy spires eventually disappear, leaving impact craters exposed, as shown in the lower image.
(a): PIA03456: Global Callisto in Color by NASA/JPL/DLR, NASA Media Licence.
(b): PIA03455: Callisto Close-up with Jagged Hills by NASA/JPL/Arizona State University/Academic Research Lab, NASA Media Licence.

In thinking about ice so far from the Sun, we must take care not to judge its behavior from the much warmer ice we know and love on Earth. At the temperatures of the outer solar system, ice on the surface is nearly as hard as rock, and it behaves similarly. Ice on Callisto does not deform or flow like ice in glaciers on Earth.

Ganymede, the Largest Moon

Ganymede, the largest moon in the solar system, also shows a great deal of cratering (Figure 8.2). We can use crater counts on solid worlds to estimate the age of the surface. The more craters, the longer the surface has been exposed to battering from space, and the older it must therefore be. About one-quarter of Ganymede's surface seems to be as old and heavily cratered as that of Callisto; the rest formed more recently, as we can tell by the sparse covering of impact craters as well as the relative freshness of those craters. If we judge from crater counts, this fresher terrain on Ganymede is somewhat younger than the lunar maria or the martian volcanic plains, perhaps 2 to 3 billion years old.

The differences between Ganymede and Callisto are more than skin deep. Ganymede is a differentiated world, like the terrestrial planets. Measurements of its gravity field tell us that the rock sank to form a core about the size of our Moon, with a mantle and crust of ice "floating" above it. In addition, the Galileo spacecraft discovered that Ganymede has a magnetic field, the sure signature of a partially molten interior. There is very likely liquid water trapped within the interior. Thus, Ganymede is not a dead world but rather a place of intermittent geological activity powered by an internal heat source. Some surface features could be as young as the surface of Venus (a few hundred million years).

The younger terrain was formed by tectonic and volcanic forces (Figure 8.2). In some places, the crust apparently cracked, flooding many of the craters with water from the interior. Extensive mountain ranges were formed from compression of the crust, forming long ridges with parallel valleys spaced a few kilometres apart. In some areas, older impact craters were split and pulled apart. There are even indications of large-scale crustal movements that are similar to the plate tectonics of Earth.

Ganymede

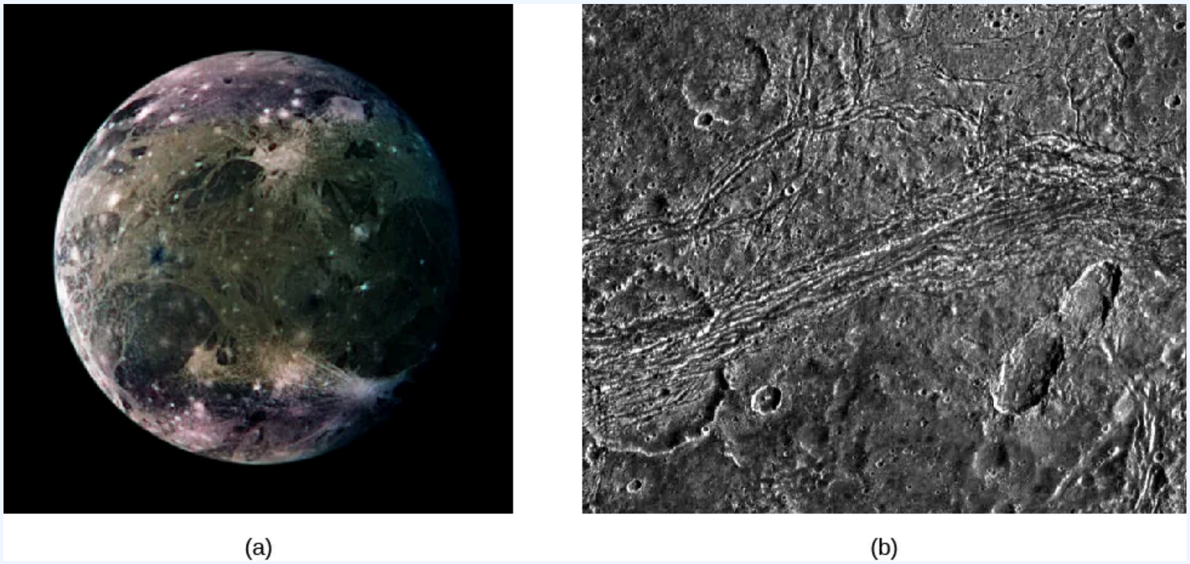


Figure 8.2. (a) This global view of Ganymede, the largest moon in the solar system, was taken by Voyager 2. The colours are enhanced to make spotting differences easier. Darker places are older, more heavily cratered regions; the lighter areas are younger (the reverse of our Moon). The brightest spots are sites of geologically recent impacts. (b) This close-up of Nicholson Regio on Ganymede shows an old impact crater (on the lower left-hand side) that has been split and pulled apart by tectonic forces. Against Ganymede's dark terrain, a line of grooves and ridges appears to cut through the crater, deforming its circular shape.
(a): PIA01666: Ganymede's Trailing Hemisphere by NASA/JPL/DLR, NASA Media Licence.
(b): PIA01612: A Tumultuous Past for Ganymede's Dark Terrain by NASA/JPL/Brown University, NASA Media Licence.

Why is Ganymede so different from Callisto? Possibly the small difference in size and internal heating between the two led to this divergence in their evolution. But more likely the gravity of Jupiter is to blame for Ganymede's continuing geological activity. Ganymede is close enough to Jupiter that **tidal forces** from the giant planet may have episodically heated its interior and triggered major convulsions on its crust.

A tidal force results from the unequal gravitational pull on two sides of a body. In a complex kind of modern dance, the large moons of Jupiter are caught in the varying gravity grip of both the giant planet and each other. This leads to gravitational flexing or kneading in their centers, which can heat them—an effect called tidal heating. (A fuller explanation is given in the section on Io.) We will see as we move inward to Europa and Io that the role of jovian tides becomes more important for moons close to the planet.

Europa, a Moon with an Ocean

Europa and Io, the inner two Galilean moons, are not icy worlds like most of the moons of the outer planets. With densities and sizes similar to our Moon, they appear to be predominantly rocky objects. How did they fail to acquire a majority share of the ice that must have been plentiful in the outer solar system at the time of their formation?

The most probable cause is Jupiter itself, which was hot enough to radiate a great deal of infrared energy during the first few million years after its formation. This infrared radiation would have heated the disk of material near the planet that would eventually coalesce into the closer moons. Thus, any ice near Jupiter was vaporized, leaving Europa and Io with compositions similar to planets in the inner solar system.

Despite its mainly rocky composition, Europa has an ice-covered surface, as astronomers have long known from examining spectra of sunlight reflected from it. In this it resembles Earth, which has a layer of water on its surface, but in Europa's case the water is capped by a thick crust of ice. There are very few impact craters in this ice, indicating that the surface of Europa is in a continual state of geological self-renewal. Judging from crater counts, the surface must be no more than a few million years old, and perhaps substantially less. In terms of its ability to erase impact craters, Europa is more geologically active than Earth.

When we look at close-up photos of Europa, we see a strange, complicated surface (Figure 8.3). For the most part, the icy crust is extremely smooth, but it is crisscrossed with cracks and low ridges that often stretch for thousands of kilometres. Some of these long lines are single, but most are double or multiple, looking rather like the remnants of a colossal freeway system.

Evidence for an Ocean on Europa

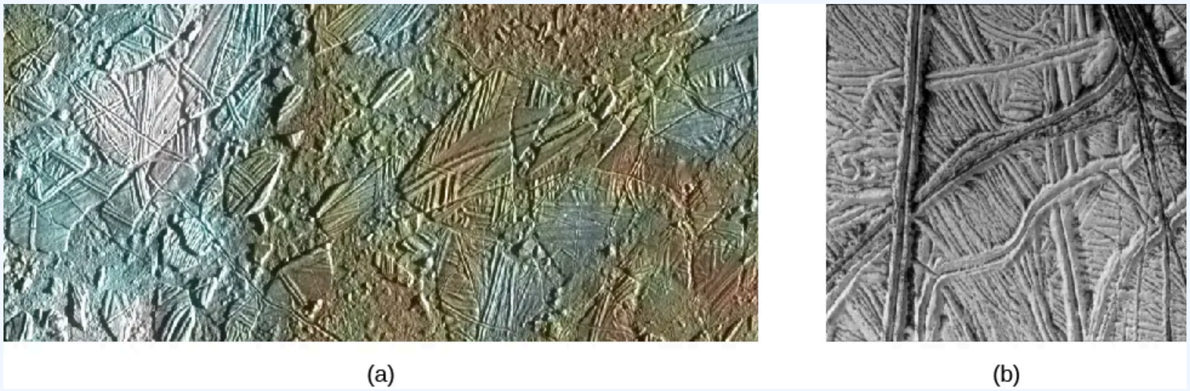


Figure 8.3. (a) A close-up of an area called Conamara Chaos is shown here with enhanced colour. This view is 70 kilometres wide in its long dimension. It appears that Conamara is a region where Europa's icy crust is (or recently was) relatively thin and there is easier access to the possible liquid or slushy ocean beneath. Not anchored to solid crust underneath, many of the ice blocks here seem to have slid or rotated from their original positions. In fact, the formations seen here look similar to views of floating sea-ice and icebergs in Earth's Arctic Ocean. (b) In this high-resolution view, the ice is wrinkled and crisscrossed by long ridges. Where these ridges intersect, we can see which ones are older and which younger; the younger ones cross over the older ones. While superficially this system of ridges resembles a giant freeway system on Europa, the ridges are much wider than our freeways and are a natural result of the flexing of the moon.
(a): PIA01127: Europa – Ice Rafting View by NASA/JPL/University of Arizona, NASA Media Licence.
(b): High-Resolution Image of Europa's Ridged Plains by NASA/JPL, NASA Media Licence.

It is very difficult to make straight lines on a planetary surface. In discussing Mars, we explained that when Percival Lowell saw what appeared to him to be straight lines (the so-called martian "canals"), he attributed them to the engineering efforts of intelligent beings. We now know the lines on Mars were optical illusions, but the lines on Europa are real. These long cracks can form in the icy crust if it is floating without much friction on an ocean of liquid water (Figure 8.4).

Young Double Ridge on Europa

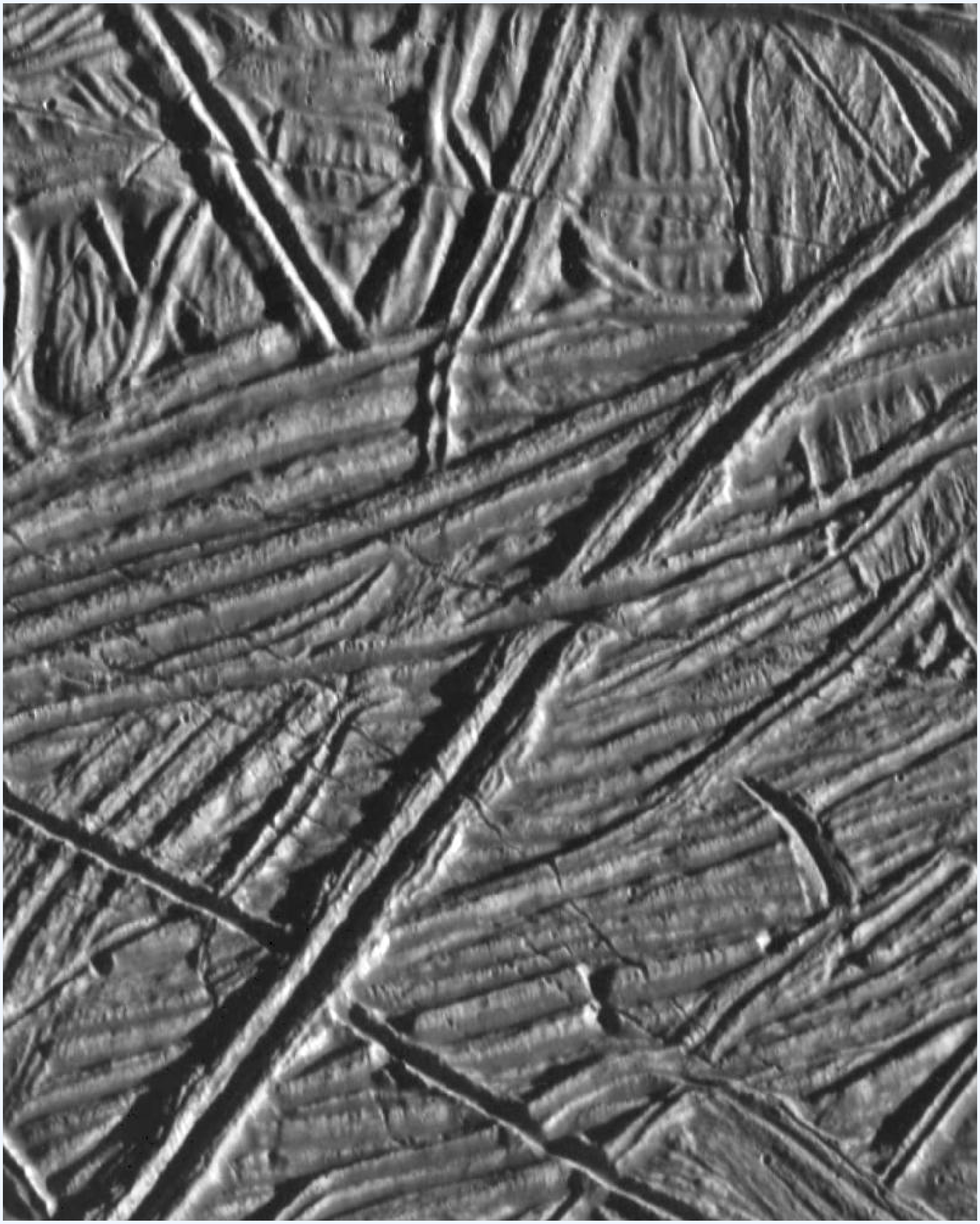


Figure 8.4. Very High-Resolution Galileo Image of One Young Double Ridge on Europa. The area in this picture is only 15 kilometres across. It appears to have formed when viscous icy material was forced up

through a long, straight crack in the crust. Note how the young ridge going from top left toward bottom right lies on top of older features, which are themselves on top of even older ones.
Europa Under Stress by NASA/JPL, NASA JPL Media Licence.

The close-up Galileo images appear to confirm the existence of a global ocean. In many places, the surface of Europa looks just as we would expect for a thick layer of ice that was broken up into giant icebergs and ice floes and then refrozen in place. When the ice breaks, water or slush from below may be able to seep up through the cracks and make the ridges and multiple-line features we observe. Many episodes of ice cracking, shifting, rotating, and refreezing are required to explain the complexity we see. The icy crust might vary in thickness from a kilometer or so up to 20 kilometres. Further confirmation that a liquid ocean exists below the ice comes from measurements of the small magnetic field induced by Europa's interactions with the magnetosphere of Jupiter. The "magnetic signature" of Europa is that of a liquid water ocean, not one of ice or rock.

If Europa really has a large ocean of liquid water under its ice, then it may be the only place in the solar system, other than Earth, with really large amounts of liquid water.[1] To remain liquid, this ocean must be warmed by heat escaping from the interior of Europa. Hot (or at least warm) springs might be active there, analogous to those we have discovered in the deep oceans of Earth. The necessary internal heat is generated by tidal heating (see the discussion later in this chapter).

What makes the idea of an ocean with warm springs exciting is the discovery in Earth's oceans of large ecosystems clustered around deep ocean hot springs. Such life derives all its energy from the mineral-laden water and thrives independent of the sunlight shining on Earth's surface. Is it possible that similar ecosystems could exist today under the ice of Europa?

Many scientists now think that Europa is the most likely place beyond Earth to find life in the solar system. In response, NASA is designing a Europa mission to characterize its liquid ocean and its ice crust, and to identify locations where material from inside has risen to the surface. Such interior material might reveal direct evidence for microbial life. This "Europa Clipper" mission will not orbit Europa, where the intense energetic particles from Jupiter's magnetosphere would quickly destroy its electronics, but instead will swoop in for brief close flybys. It is planned for launch in 2024.

Io, a Volcanic Moon

Io, the innermost of Jupiter's Galilean moons, is in many ways a close twin of our Moon, with nearly the same size and density. We might therefore expect it to have experienced a similar history. Its appearance, as photographed from space, tells us another story, however (Figure 8.5). Instead of being a dead cratered world, Io turns out to have the highest level of volcanism in the solar system, greatly exceeding that of Earth.

Two Sides of Io

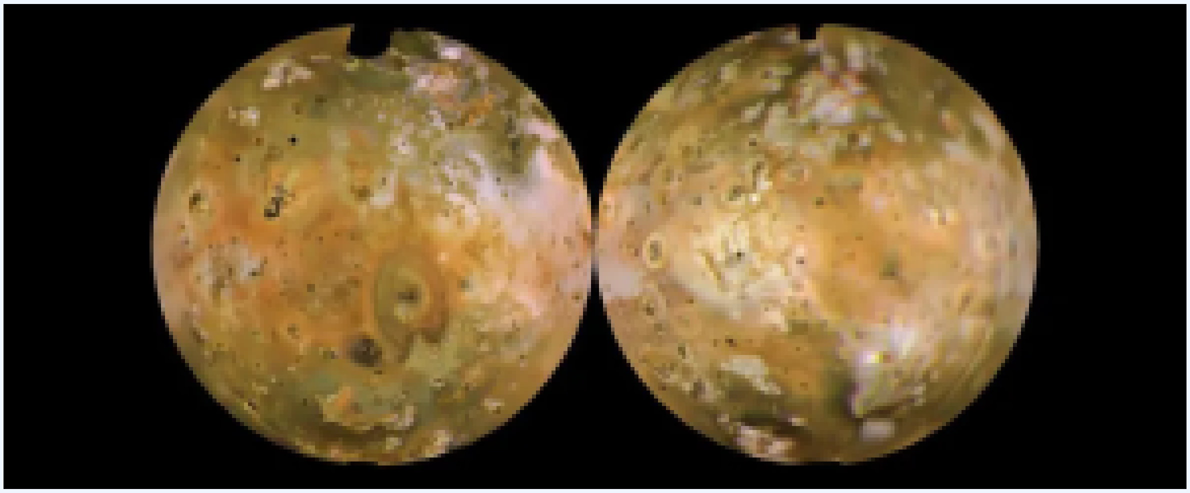


Figure 8.5. This composite image shows both sides of the volcanically active moon Io. The orange deposits are sulfur snow; the white is sulfur dioxide. (Carl Sagan once quipped that Io looks as if it desperately needs a shot of penicillin.) (Photo by NASA/JPL/USGS JPL Image Use Policy)

Io's active volcanism was discovered by the Voyager spacecraft. Eight volcanoes were seen erupting when Voyager 1 passed in March 1979, and six of these were still active four months later when Voyager 2 passed. With the improved instruments carried by the Galileo spacecraft, more than 50 eruptions were found during 1997 alone. Many of the eruptions produce graceful plumes that extend hundreds of kilometres out into space (Figure 8.6).

Volcanic Eruptions on Io

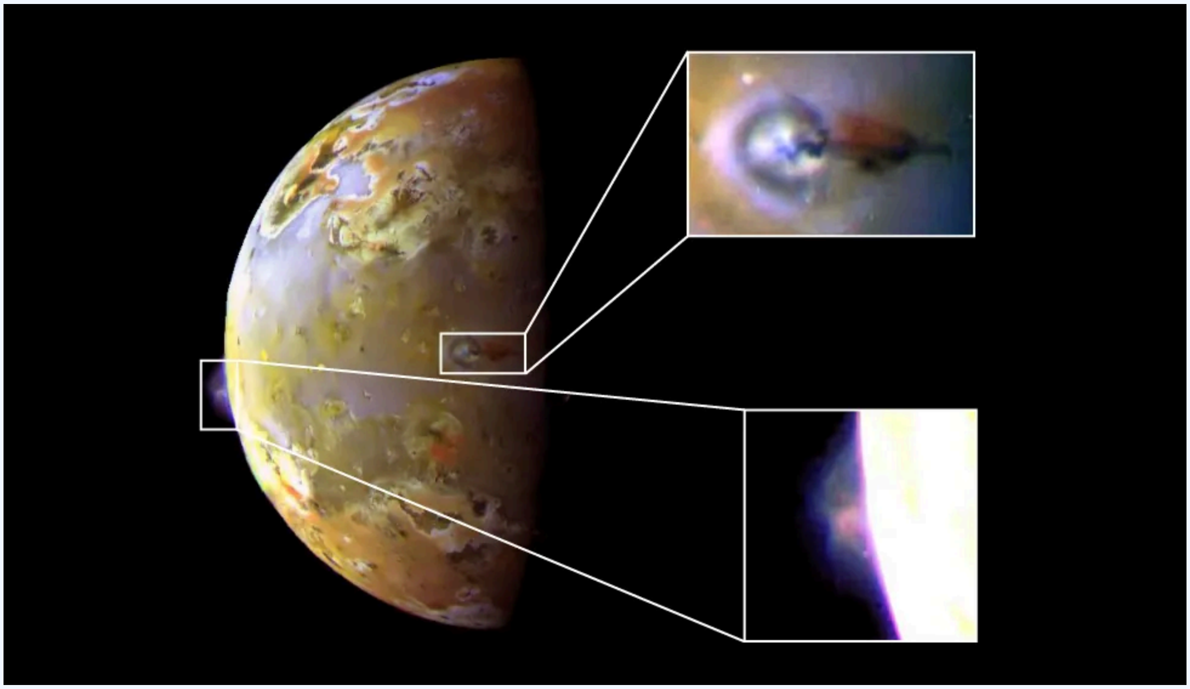


Figure 8.6. This composite image from NASA's Galileo spacecraft shows close-ups (the two inset photos) of two separate volcanic eruptions on Jupiter's volcanic moon, Io. In the upper inset image, you can see a close up of a bluish plume rising about 140 kilometres above the surface of the volcano. In the lower inset image is the Prometheus plume, rising about 75 kilometres from Io's surface. The Prometheus plume is named for the Greek god of fire.
PIA00703: Active Volcanic Plumes on Io by NASA/JPL, NASA Media Licence.

The Galileo data show that most of the volcanism on Io consists of hot silicate lava, like the volcanoes on Earth. Sometimes the hot lava encounters frozen deposits of sulfur and sulfur dioxide. When these icy deposits are suddenly heated, the result is great eruptive plumes far larger than any ejected from terrestrial volcanoes. As the rising plumes cool, the sulfur and sulfur dioxide recondense as solid particles that fall back to the surface in colourful "snowfalls" that extend as much as a thousand kilometres from the vent. Major new surface features were even seen to appear between Galileo orbits, as shown in Figure 8.7.

Volcanic Changes on Io

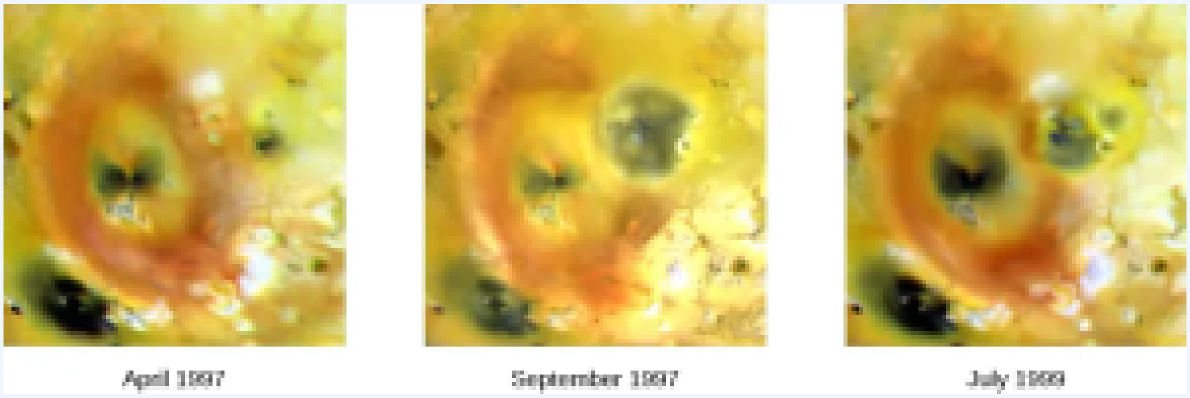


Figure 8.7. These three images were taken of the same 1700-kilometer-square region of Io in April 1997, September 1997, and July 1999. The dark volcanic center called Pillan Patera experienced a huge eruption, producing a dark deposit some 400 kilometres across (seen as the grey area in the upper center of the middle image). In the right image, however, some of the new dark deposit is already being covered by reddish material from the volcano Pele. Also, a small unnamed volcano to the right of Pillan has erupted since 1997, and some of its dark deposit and a yellow ring around it are visible on the right image (to the right of the grey spot). The colour range is exaggerated in these images. (Photo by NASA/JPL/University of Arizona JPL Image Use Policy modified to add dates)

As the Galileo mission drew to a close, controllers were willing to take risks in getting close to Io. Approaching this moon is a dangerous maneuver because the belts of atomic particles trapped in Jupiter's magnetic environment are at their most intense near Io's orbit. Indeed, in its very first pass by Io, the spacecraft absorbed damaging radiation beyond its design levels. To keep the system working at all, controllers had to modify or disable various fault-protection software routines in the onboard computers. In spite of these difficulties, the spacecraft achieved four successful Io flybys, obtaining photos and spectra of the surface with unprecedented resolution.

Maps of Io reveal more than 100 recently active volcanoes. Huge flows spread out from many of these vents, covering about 25% of the moon's total surface with still-warm lava. From these measurements, it seems clear that the bright surface colours that first attracted attention to Io are the result of a thin veneer of sulfur compounds. The underlying volcanism is driven by eruptions of molten silicates, just like on Earth (Figure 8.8).

Lava Fountains on Io

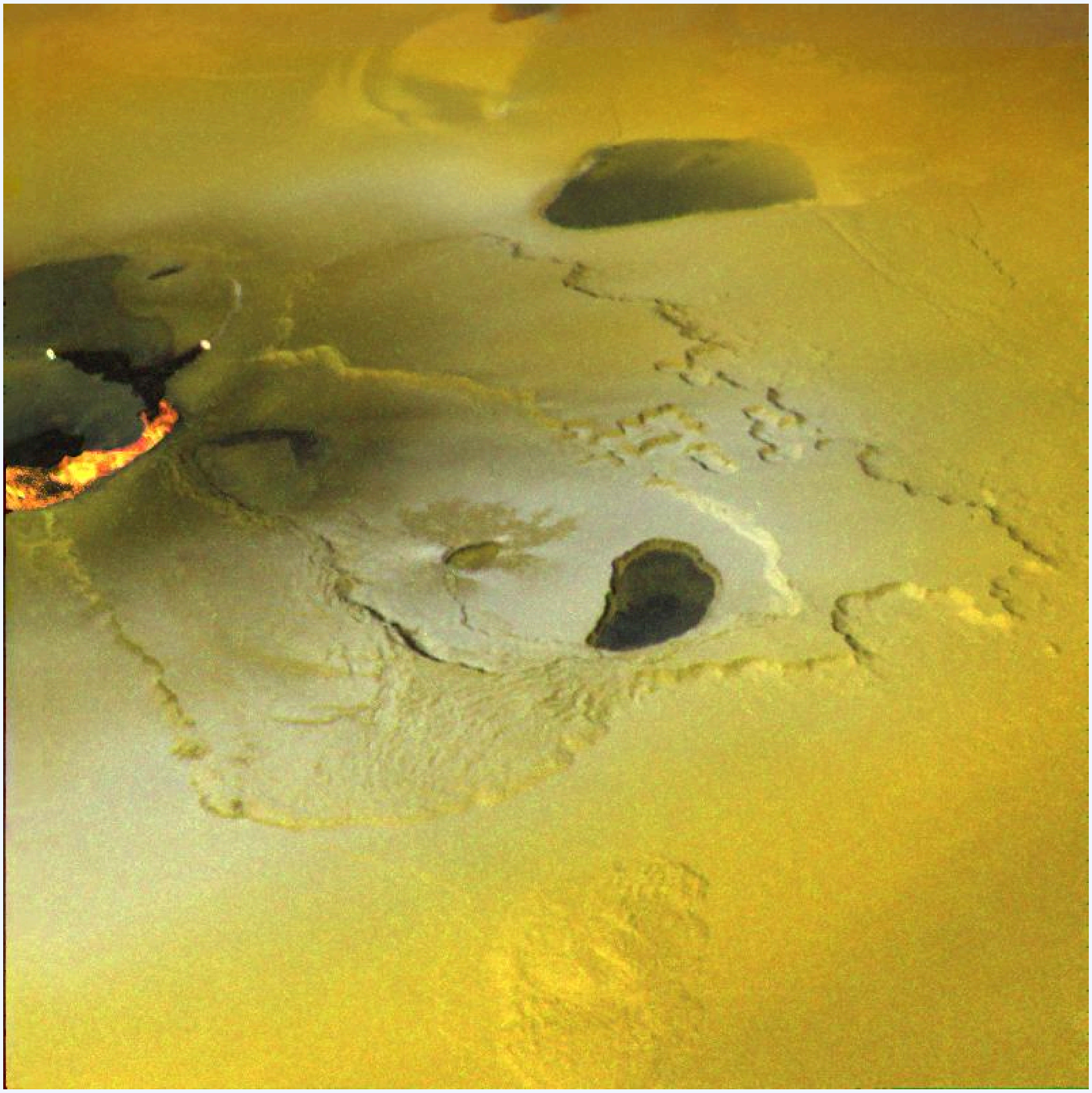


Figure 8.8. Galileo captured a number of eruptions along the chain of huge volcanic calderas (or pits) on Io called Tvashtar Catena in this false-colour image combining infrared and visible light. The bright orange-yellow areas at left are places where fresh, hot lava is erupting from below ground. PIA02550: Ongoing Volcanic Eruption at Tvashtar Catena, Io by NASA/JPL/University of Arizona, NASA Media Licence.

Tidal Heating

How can Io remain volcanically active in spite of its small size? The answer, as we hinted earlier, lies in the effect of gravity, through tidal heating. Io is about the same distance from Jupiter as our Moon is from Earth. Yet Jupiter is more than 300 times more massive than Earth, causing forces that pull Io into an elongated shape, with a several-kilometer-high bulge extending toward Jupiter.

If Io always kept exactly the same face turned toward Jupiter, this bulge would not generate heat. However, Io's orbit is not exactly circular due to gravitational perturbations (tugs) from Europa and Ganymede. In its slightly eccentric orbit, Io twists back and forth with respect to Jupiter, at the same time moving nearer and farther from the planet on each revolution. The twisting and flexing heat Io, much as repeated flexing of a wire coat hanger heats the wire.

After billions of years, this constant flexing and heating have taken their toll on Io, driving away water and carbon dioxide and other gases, so that now sulfur and sulfur compounds are the most volatile materials remaining. Its interior is entirely melted, and the crust itself is constantly recycled by volcanic activity.

In moving inward toward Jupiter from Callisto to Io, we have encountered more and more evidence of geological activity and internal heating, culminating in the violent volcanism on Io. Three of these surfaces are compared in Figure 8.9. Just as the character of the planets in our solar system depends in large measure on their distance from the Sun (and on the amount of heat they receive), so it appears that distance from a giant planet like Jupiter can play a large role in the composition and evolution of its moons (at least partly due to differences in internal heating of each moon by Jupiter's unrelenting tidal forces).

Three Icy Moons

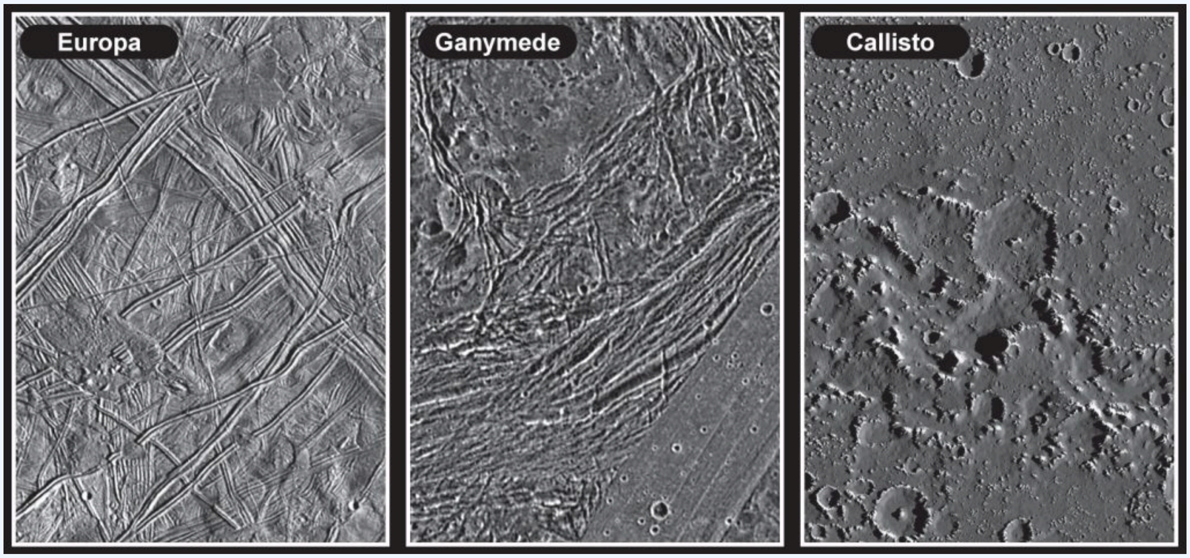


Figure 8.9. These Galileo images compare the surfaces of Europa, Ganymede, and Callisto at the same resolution. Note that the number of craters (and thus the age of the surface we see) increases as we go from Europa to Ganymede to Callisto. The Europa image is one of those where the system of cracks and ridges resembles a freeway system.
PIA01656: Europa, Ganymede, and Callisto: Surface Comparison at High Spatial Resolution by NASA/JPL/DLR, NASA Media Licence.

Footnotes

- [1] Ganymede and Saturn's moon Enceladus may have smaller amounts of liquid water under their surfaces.

Attribution

"12.2 The Galilean Moons of Jupiter" from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C.

8.2 THE LARGE MOONS OF SATURN AND NEPTUNE

We shift our attention now to small worlds in the more distant parts of the solar system. Saturn's large moon, Titan, turns out to be a weird cousin of Earth, with many similarities in spite of frigid temperatures. The Cassini observations of Titan have provided some of the most exciting recent discoveries in planetary science. Neptune's moon Triton also has unusual characteristics and resembles Pluto, which we will discuss in the following section.

Titan, a Moon with Atmosphere and Hydrocarbon Lakes

Titan, first seen in 1655 by the Dutch astronomer Christiaan Huygens, was the first moon discovered after Galileo saw the four large moons of Jupiter. Titan has roughly the same diameter, mass, and density as Callisto or Ganymede. Presumably it also has a similar composition—about half ice and half rock. However, Titan is unique among moons, with a thick atmosphere and lakes and rivers and falling rain (although these are not composed of water but of hydrocarbons such as ethane and methane, which can stay liquid at the frigid temperatures on Titan). Titan is the only moon or planet other than Earth where we have found evidence of bodies of surface liquids.

The 1980 Voyager flyby of Titan determined that the surface density of its atmosphere is four times greater than that on Earth. The atmospheric pressure on this moon is 1.6 bars, higher than that on any other moon and, remarkably, even higher than that of the terrestrial planets Mars and Earth. The atmospheric composition is primarily nitrogen, an important way in which Titan's atmosphere resembles Earth's.

Also detected in Titan's atmosphere were carbon monoxide (CO), hydrocarbons (compounds of hydrogen and carbon) such as methane (CH_4), ethane (C_2H_6), and propane (C_3H_8), and nitrogen compounds such as hydrogen cyanide (HCN), cyanogen (C_2N_2), and cyanoacetylene (HC_3N). Their presence indicates an active chemistry in which sunlight interacts with atmospheric nitrogen and methane to create a rich mix of organic molecules. There are also multiple layers of hydrocarbon haze and clouds in the atmosphere, as illustrated in Figure 8.10.

Structure of Titan's Atmosphere

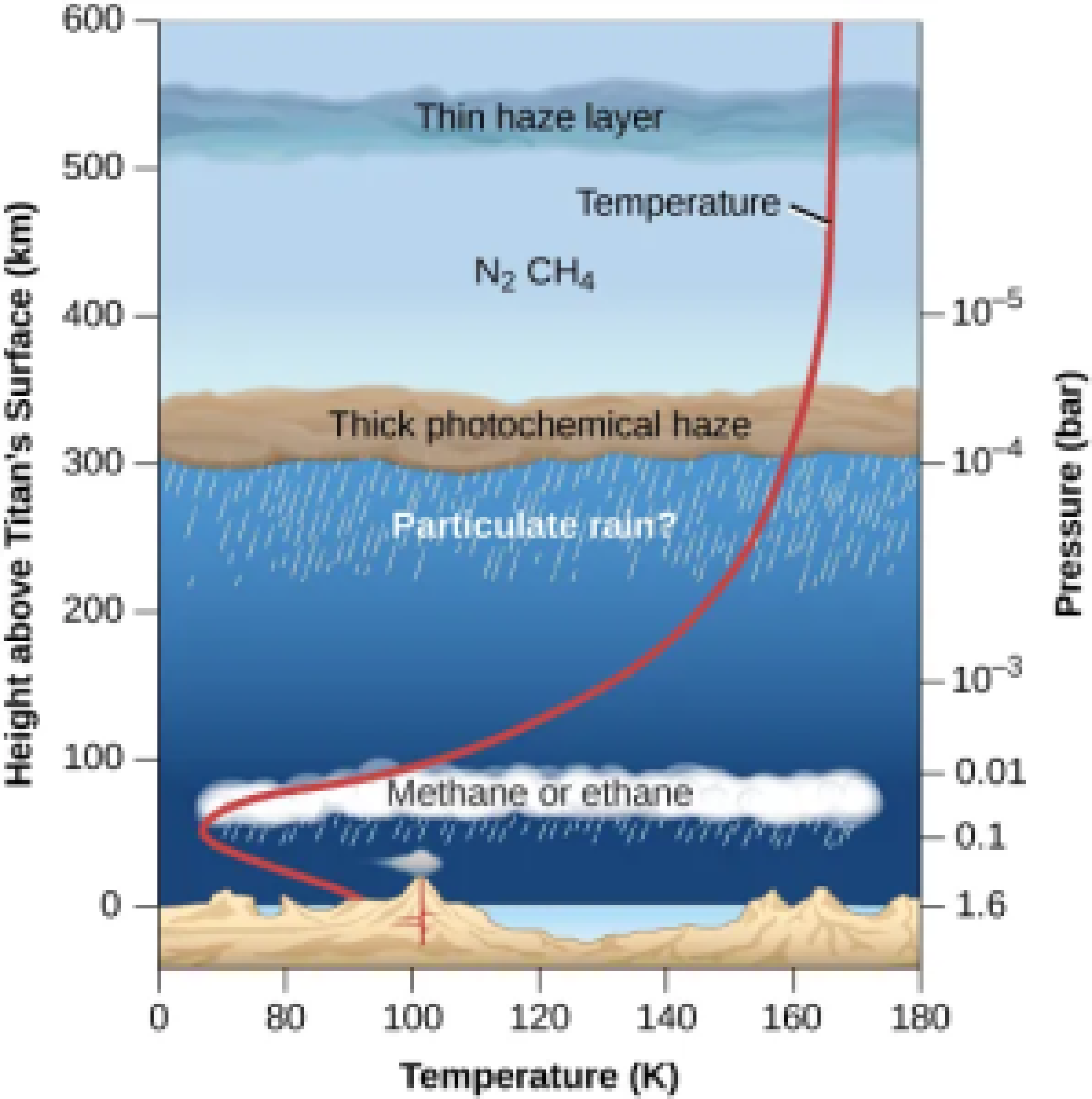


Figure 8.10. Some characteristics of Titan's atmosphere resemble those of Earth's atmosphere, although it is much colder than our planet. The red line indicates the temperature of Titan's atmosphere at different altitudes.

These Voyager discoveries motivated a much more ambitious exploration program using the NASA Cassini Saturn orbiter and a probe to land on Titan called Huygens, built by the European Space Agency. The orbiter, which included several cameras, spectrometres, and a radar imaging system, made dozens of close flybys of Titan between 2004 and 2015, each yielding radar and infrared images of portions of the surface. The Huygens probe successfully descended by parachute through the atmosphere, photographing the surface from below the clouds, and landing on January 14, 2005. This was the first (and so far the only) spacecraft landing on a moon in the outer solar system.

At the end of its parachute descent, the 319-kilogram Huygens probe safely touched down, slid a short distance, and began sending data back to Earth, including photos and analyses of the atmosphere. It appeared to have landed on a flat, boulder-strewn plain, but both the surface and the boulders were composed of water ice, which is as hard as rock at the temperature of Titan (see Figure 8.11).

The photos taken during descent showed a variety of features, including drainage channels, suggesting that Huygens had landed on the shore of an ancient hydrocarbon lake. The sky was deep orange, and the brightness of the Sun was a thousand times less than sunlight on Earth (but still more than a hundred times brighter than under the full moon on Earth). Titan's surface temperature was 94 K (−179 °C). The warmer spacecraft heated enough of the ice where it landed for its instruments to measure released hydrocarbon gas. Measurements on the surface continued for more than an hour before the probe succumbed to the frigid temperature.

Views of the Surface of Titan

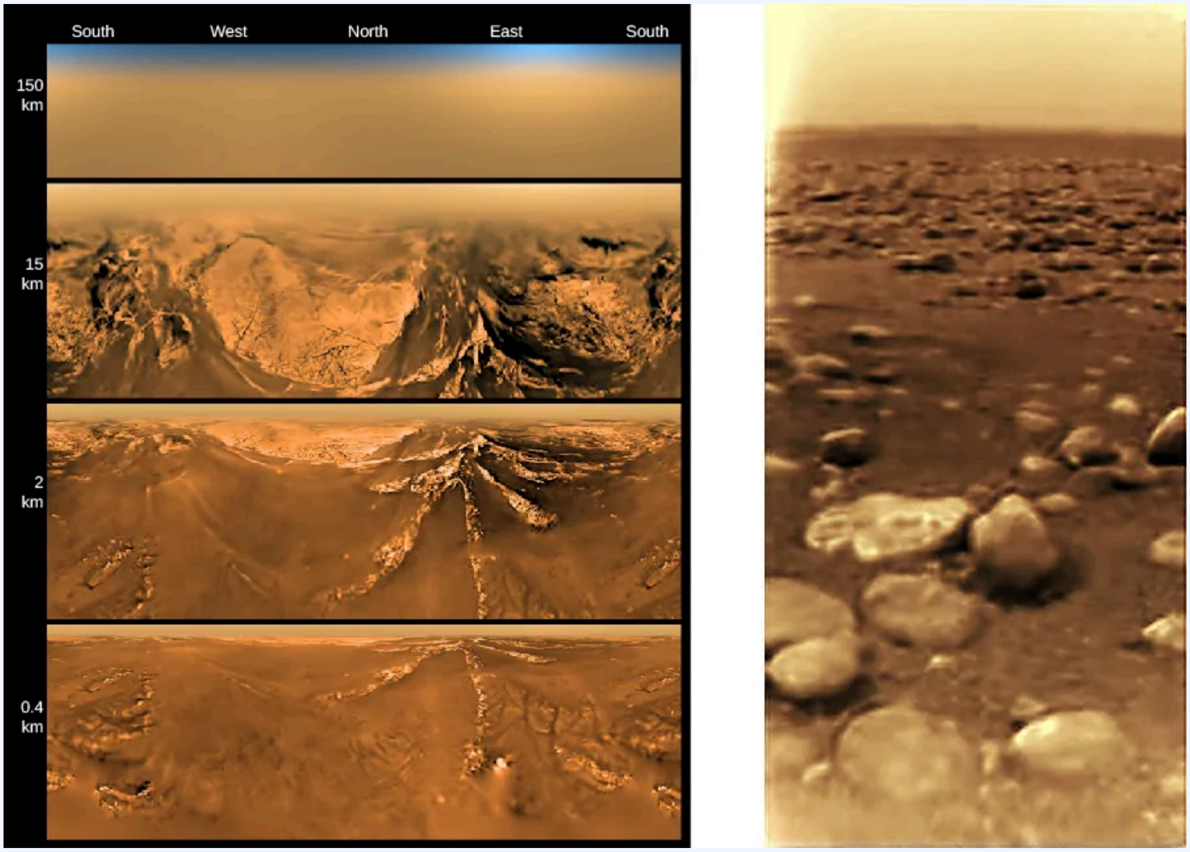


Figure 8.11. The left image shows the views of Titan from the descent camera, in a flattened projection, at different altitudes. The right image, taken after landing, shows a boulder-strewn surface illuminated by faint reddish sunlight. The boulders are composed of water ice.
(a): PIA08427: Mercator Projection of Huygens's View at Different Altitudes by ESA/NASA/JPL/University of Arizona, NASA Media Licence.
(b): PIA07232: First Color View of Titan's Surface by NASA/JPL/ESA/University of Arizona, NASA Media Licence.

Radar and infrared imaging of Titan from the Cassini orbiter gradually built up a picture of a remarkably active surface on this moon, complex and geologically young (Figure 8.12). There are large methane lakes near the polar regions that interact with the methane in the atmosphere, much as Earth's water oceans interact with the water vapor in our atmosphere. The presence of many erosional features indicates that atmospheric methane can condense and fall as rain, then flow down valleys to the big lakes. Thus, Titan has a low-temperature equivalent of the water cycle on Earth, with liquid on the surface that evaporates, forms clouds, and then condenses to fall as rain—but on Titan the liquid is a combination of methane, ethane, and a trace of other hydrocarbons. It is a weirdly familiar and yet utterly alien landscape.

Titan's Lakes

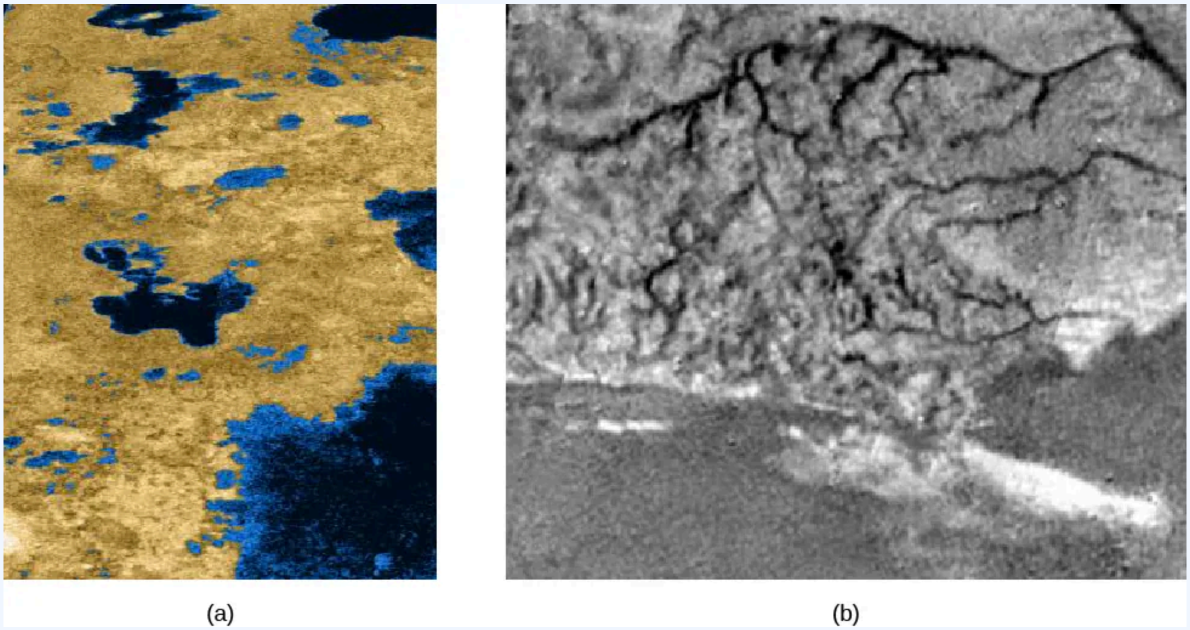


Figure 8.12. (a) This Cassini image from a September 2006 flyby shows the liquid lakes on Titan. Their composition is most likely a combination of methane and ethane. (Since this is a radar image, the colours are artificially added. The dark blue areas are the smooth surfaces of the liquid lakes, and yellow is the rougher solid terrain around them.) (b) This mosaic of Titan's surface from the Cassini-Huygens mission shows in detail a high ridge area and many narrow, sinuous erosion channels that appear to be part of a widespread network of "rivers" carved by flowing hydrocarbons.
(a): PIA09102: Liquid Lakes on Titan by NASA/JPL-Caltech/USGS, NASA Media Licence.
(b): PIA07236: Mosaic of River Channel and Ridge Area on Titan by NASA/JPL/ESA/University of Arizona, NASA Media Licence.

These discoveries raise the question of whether there could be life on Titan. Hydrocarbons are fundamental for the formation of the large carbon molecules that are essential to life on our planet. However, the temperature on Titan is far too low for liquid water or for many of the chemical processes that are essential to life as we know it. There remains, though, an intriguing possibility that Titan might have developed a different form of low-temperature carbon-based life that could operate with liquid hydrocarbons playing the role of water. The discovery of such "life as we don't know it" could be even more exciting than finding life like ours on Mars. If such a truly alien life is present on Titan, its existence would greatly expand our understanding of the nature of life and of habitable environments.

NASA has selected a new mission to Titan for launch in 2027. Called *Dragonfly*, this mission is a drone that will fly in Titan's atmosphere, with emphasis on study of pre-biotic chemistry. Other future proposed missions include a balloon operating in the atmosphere and a "boat" floating in one of the Titan lakes.

Enceladus

Saturn has a very faint, tenuous ring, called the E Ring, associated with its small icy moon Enceladus. The particles in the E Ring are very small and composed of water ice. Since such a tenuous cloud of ice crystals will tend to dissipate, the ongoing existence of the E Ring strongly suggests that it is being continually replenished by a source at Enceladus. This icy moon is very small—only 500 kilometres in diameter—but the Voyager images showed that the craters on about half of its surface have been erased, indicating geological activity sometime in the past few million years. It was with great anticipation that the Cassini scientists maneuvered the spacecraft orbit to allow multiple close flybys of Enceladus starting in 2005.

Those awaiting the Cassini flyby results were not disappointed. High-resolution images showed long, dark stripes of smooth ground near its south pole, which were soon nicknamed "tiger stripes" (Figure 8.13). Infrared measurements revealed that these tiger stripes are warmer than their surroundings. Best of all, dozens of cryovolcanic vents on the tiger stripes were seen to be erupting geysers of salty water and ice (Figure 8.14). Estimates suggested that 200 kilograms of material were shooting into space each second—not a lot, but enough for the spacecraft to sample.

Enceladus

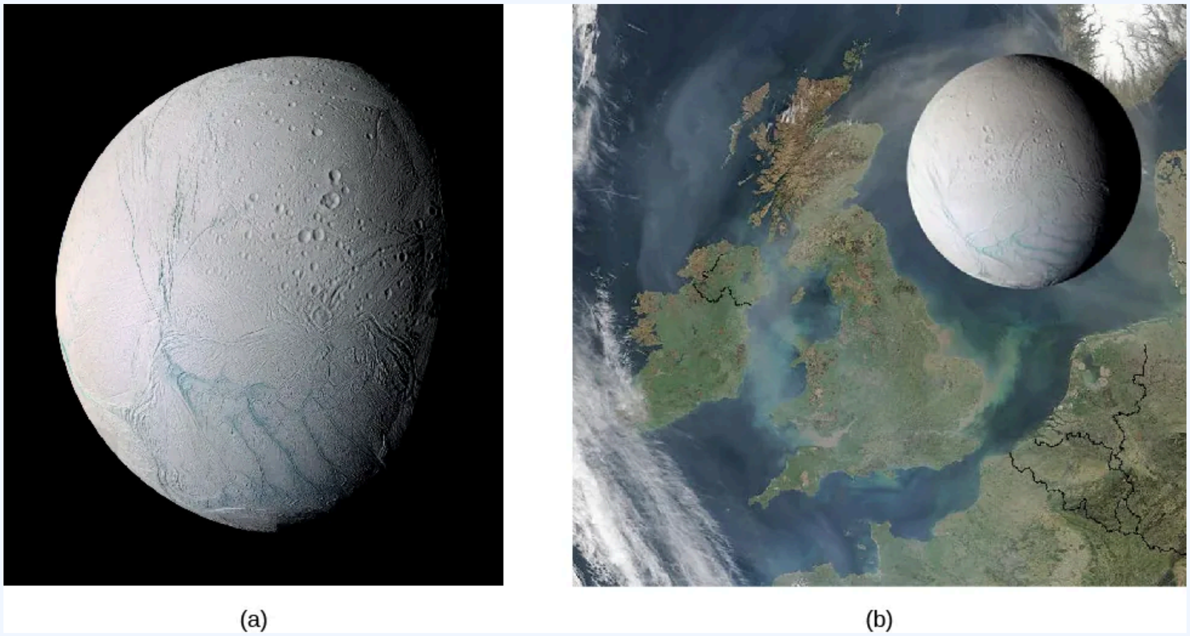


Figure 8.13. (a) This image shows both smooth and cratered terrain on Saturn's moon, and also "tiger stripes" in the south polar region (lower part of image). These dark stripes (shown here in exaggerated colour) have elevated temperatures and are the source of the many geysers discovered on Enceladus. They are about 130 kilometres long and 40 kilometres apart. (b) Here Enceladus is shown to scale with Great Britain and the coast of Western Europe, to emphasize that it is a small moon, only about 500 kilometres in diameter.
(a): PIA06254: Zooming In On Enceladus (Mosaic) by NASA/JPL/Space Science Institute, NASA Media Licence.
(b): PIA07724: Enceladus to Scale by NASA/JPL/Space Science Institute, NASA Media Licence.

When Cassini was directed to fly into the plumes, it measured their composition and found them to be similar to material we see liberated from comets. The vapour and ice plumes consisted mostly of water, but with trace amounts of nitrogen, ammonia, methane, and other hydrocarbons. Minerals found in the geysers in trace amounts included ordinary salt, meaning that the geyser plumes were high-pressure sprays of salt water.

Based on the continuing study of Enceladus' bulk properties and the ongoing geysers, in 2015 the Cassini mission scientists tentatively identified a subsurface ocean of water feeding the geysers. These discoveries suggested that in spite of its small size, Enceladus should be added to the list of worlds that we would like to explore for possible life. Since its subsurface ocean is conveniently escaping into space, it might be much easier to sample than the ocean of Europa, which is deeply buried below its thick crust of ice.

Geysers on Enceladus

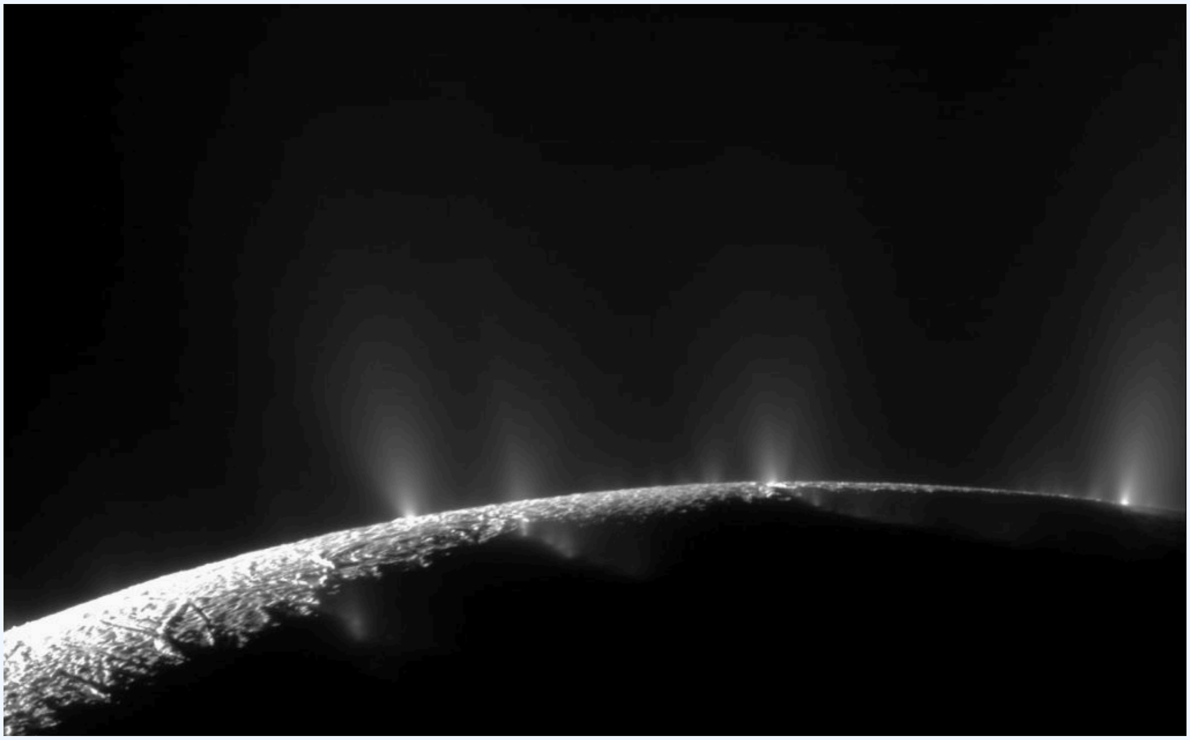


Figure 8.14. This Cassini image shows a number of water geysers on Saturn's small moon Enceladus, apparently salty water from a subsurface source escaping through cracks in the surface. You can see curved lines of geysers along the four "tiger stripes" on the surface.
PIA11688: Bursting at the Seams: the Geyser Basin of Enceladus by NASA/JPL/Space Science Institute, NASA Media Licence.

Triton and Its Volcanoes

Neptune's largest moon Triton (don't get its name confused with Titan) has a diameter of 2720 kilometres and a density of 2.1 g/cm^3, indicating that it's probably composed of about 75% rock mixed with 25% water ice. Measurements indicate that Triton's surface has the coldest temperature of any of the worlds our robot representatives have visited. Because its reflectivity is so high (about 80%), Triton reflects most of the solar energy that falls on it, resulting in a surface temperature between 35 and 40 K.

The surface material of Triton is made of frozen water, nitrogen, methane, and carbon monoxide. Methane and nitrogen exist as gas in most of the solar system, but they are frozen at Triton's temperatures. Only a small quantity of nitrogen vapor persists to form an atmosphere. Although the surface pressure of this atmosphere is only 16 millionths of a bar, this is sufficient to support thin haze or cloud layers.

Triton's surface, like that of many other moons in the outer solar system, reveals a long history of geological evolution (Figure 8.15). Although some impact craters are found, many regions have been flooded fairly recently by the local version of "lava" (perhaps water or water-ammonia mixtures). There are also mysterious regions of jumbled or mountainous terrain.

Neptune's Moon Triton

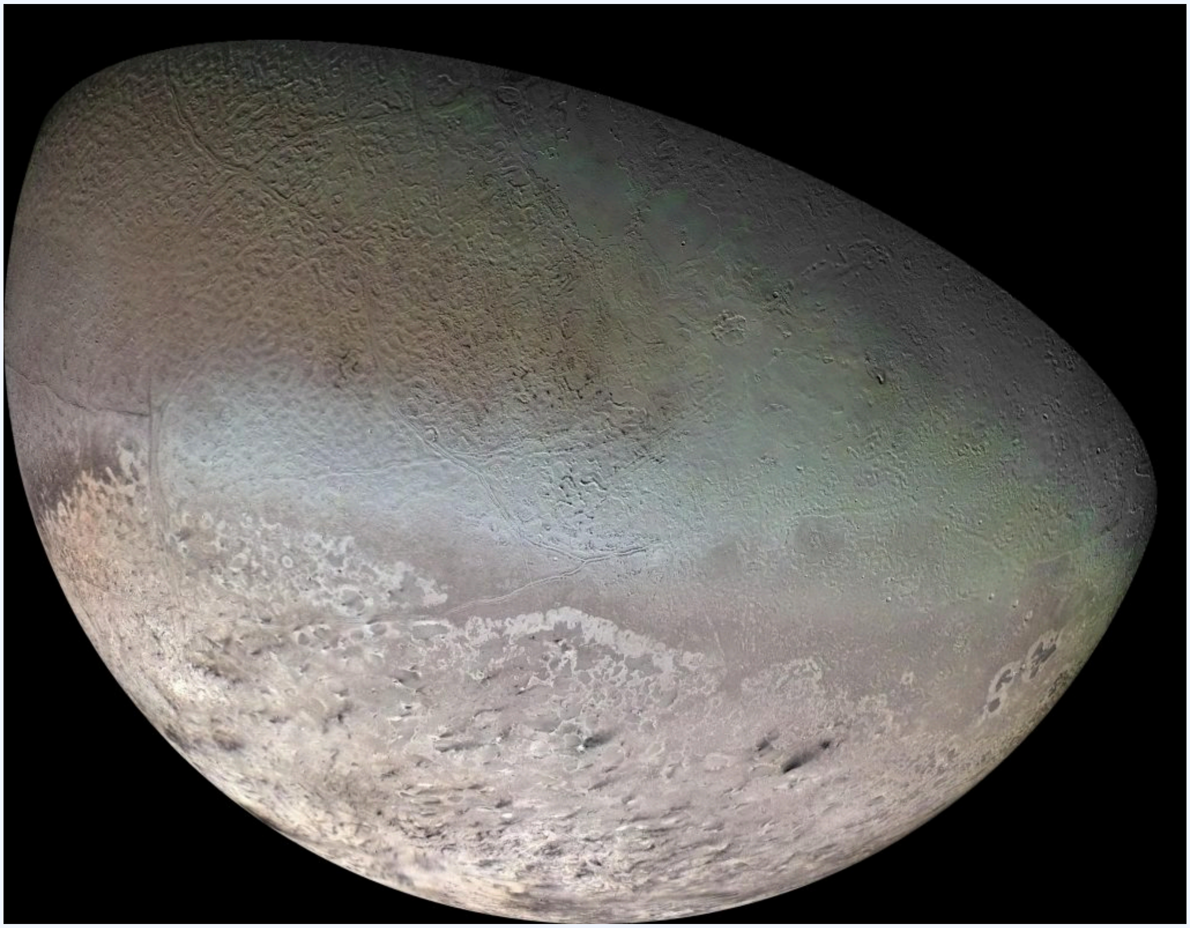


Figure 8.15. This mosaic of Voyager 2 images of Triton shows a wide range of surface features. The pinkish area at the bottom is Triton's large southern polar cap. The south pole of Triton faces the Sun here, and the slight heating effect is driving some of the material northward, where it is colder.
PIA00317: Global Color Mosaic of Triton by NASA/JPL/USGS, NASA Media Licence.

The Voyager flyby of Triton took place at a time when the moon's southern pole was tipped toward the Sun, allowing this part of the surface to enjoy a period of relative warmth. (Remember that "warm" on Triton is still outrageously colder than anything we experience on Earth.) A polar cap covers much of Triton's southern hemisphere, apparently evaporating along the northern edge. This polar cap may consist of frozen nitrogen that was deposited during the previous winter.

Remarkably, the Voyager images showed that the evaporation of Triton's polar cap generates geysers or volcanic plumes of nitrogen gas (see Figure 8.16). (Fountains of such gas rose about 10 kilometres high, visible in the thin atmosphere because dust from the surface rose with them and coloured them dark.) These plumes differ from the volcanic plumes of Io in their composition and also in that they derive their energy from sunlight warming the surface rather than from internal heat.

Triton's Geysers

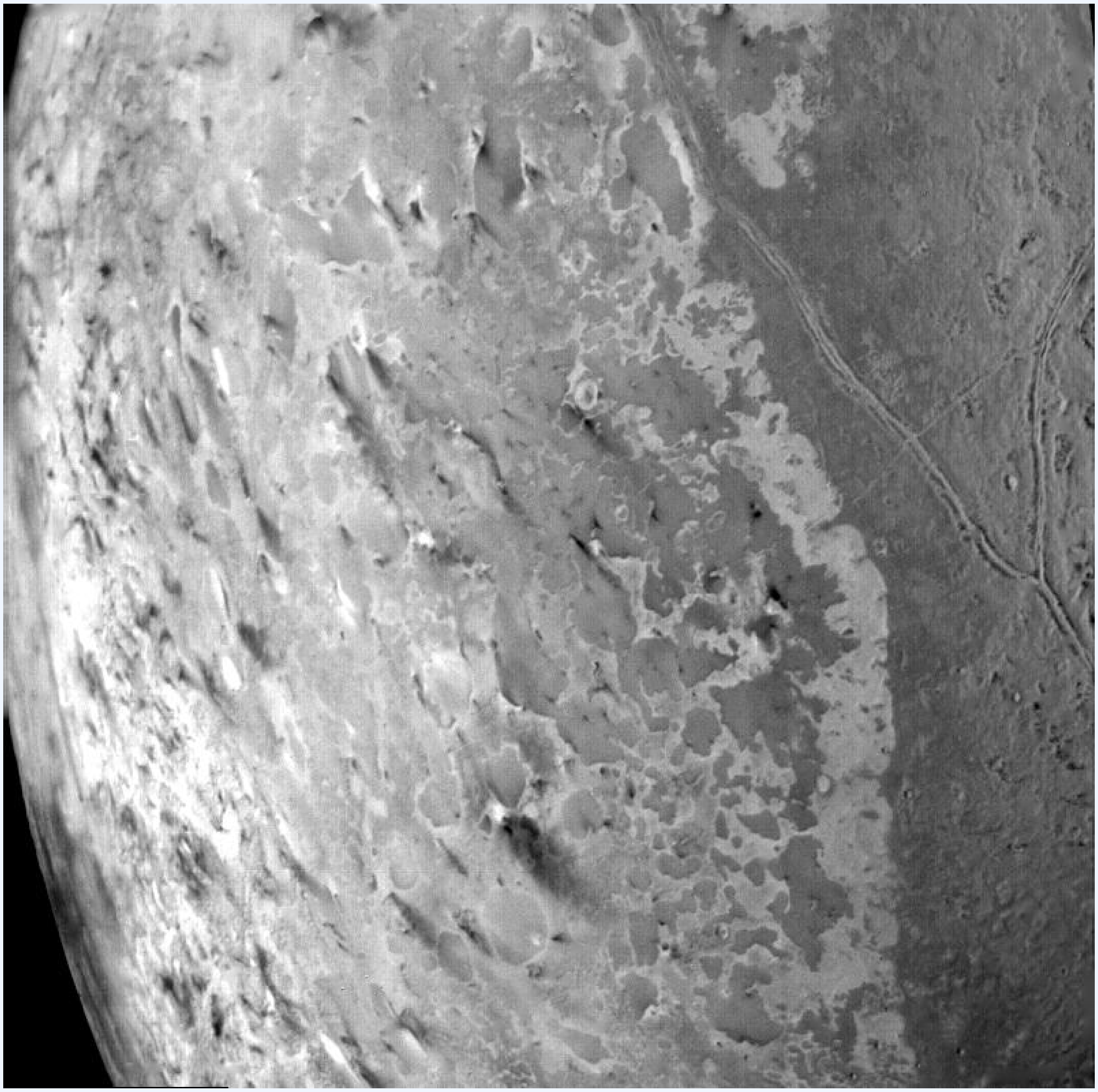


Figure 8.16. This close-up view shows some of the geysers on Neptune's moon Triton, with the long trains of dust pointing to the lower right in this picture.
PIA00059: Triton South Polar Terrain by NASA/JPL, NASA Media Licence.

Attribution

8.3 PLANETARY RINGS

In addition to their moons, all four of the giant planets have rings, with each ring system consisting of billions of small particles or "moonlets" orbiting close to their planet. Each of these rings displays a complicated structure that is related to interactions between the ring particles and the larger moons. However, the four ring systems are very different from each other in mass, structure, and composition, as outlined in Table 8.2.

Table 8.2. Properties of the Ring Systems

Planet	Outer Radius (km)	Outer Radius (R_{planet})	Mass (kg)	Reflectivity (%)
Jupiter	128,000	1.8	10^{10}(?)	?
Saturn	140,000	2.3	10^{19}	60
Uranus	51,000	2.2	10^{14}	5
Neptune	63,000	2.5	10^{12}	5

Saturn's large ring system is made up of icy particles spread out into several vast, flat rings containing a great deal of fine structure. The Uranus and Neptune ring systems, on the other hand, are nearly the reverse of Saturn's: they consist of dark particles confined to a few narrow rings with broad empty gaps in between. Jupiter's ring and at least one of Saturn's are merely transient dust bands, constantly renewed by dust grains eroded from small moons. In this section, we focus on the two most massive ring systems, those of Saturn and Uranus.

Rings of Saturn

Saturn's rings are one of the most beautiful sights in the solar system (Figure 8.17). From outer to inner, the three brightest rings are labelled with the extremely unromantic names of A, B, and C Rings. Table 8.3 gives the dimensions of the rings in both kilometres and units of the radius of Saturn, R_{Saturn}. The B Ring is the brightest and has the most closely packed particles, whereas the A and C Rings are translucent.

The total mass of the B Ring, which is probably close to the mass of the entire ring system, is about equal to that of an icy moon 250 kilometres in diameter (suggesting that the ring could have originated in the breakup of such a moon). Between the A and B Rings is a wide gap named the Cassini Division after Gian Domenico Cassini, who first glimpsed it through a telescope in 1675 and whose name planetary scientists also gave to the Cassini spacecraft that explored the Saturn system.

Saturn's Rings as Seen from Above and Below

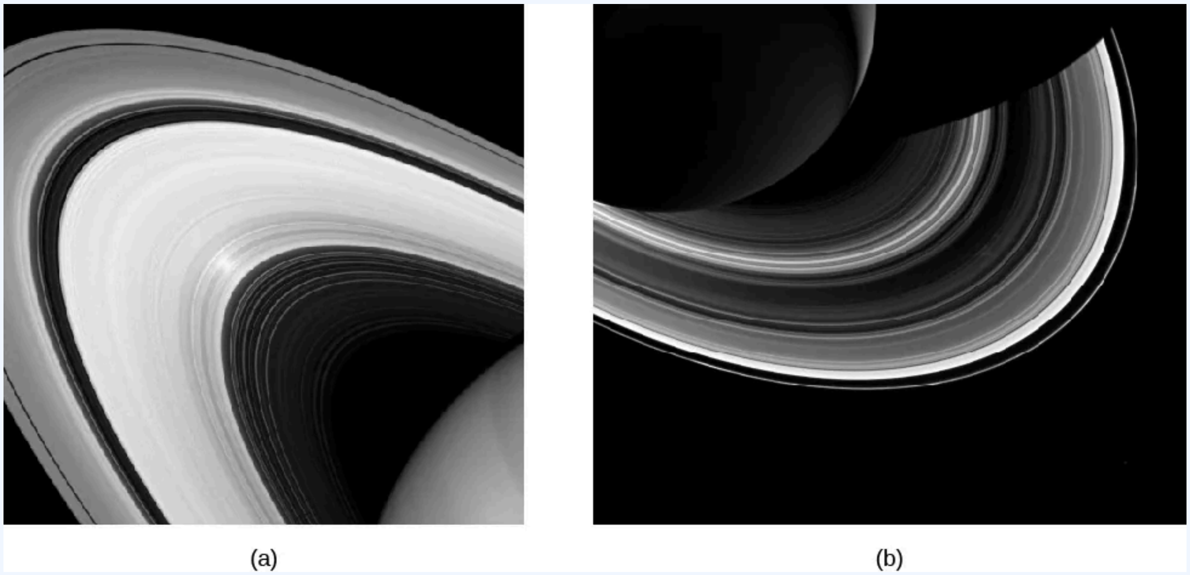


Figure 8.17. (a) The view from above is illuminated by direct sunlight. (b) The illumination seen from below is sunlight that has diffused through gaps in the rings.
(a): PIA17152: Polarized Surge by NASA/JPL-Caltech/Space Science Institute, NASA Media Licence.
(b): PIA17154: Shadows and Rings by NASA/JPL-Caltech/Space Science Institute, NASA Media Licence.

Table 8.3. Selected Features in the Rings of Saturn

Ring Name[3]	Outer Edge (R_{Saturn})	Outer Edge (km)	Width (km)
F	2.324	140,180	90
A	2.267	136,780	14,600
Cassini Division	2.025	122,170	4590
B	1.949	117,580	25,580
C	1.525	92,000	17,490

Saturn's rings are very broad and very thin. The width of the main rings is 70,000 kilometres, yet their average thickness is only 20 metres. If we made a scale model of the rings out of paper, we would have to make them 1 kilometre across. On this scale, Saturn itself would loom as high as an 80-story building. The ring particles are composed primarily of water ice, and they range from grains the size of sand up to house-sized boulders. An insider's view of the rings would probably resemble a bright cloud of floating snowflakes and hailstones, with a few snowballs and larger objects, many of them loose aggregates of smaller particles (Figure 8.18).

Rings of Saturn as Seen from the Inside

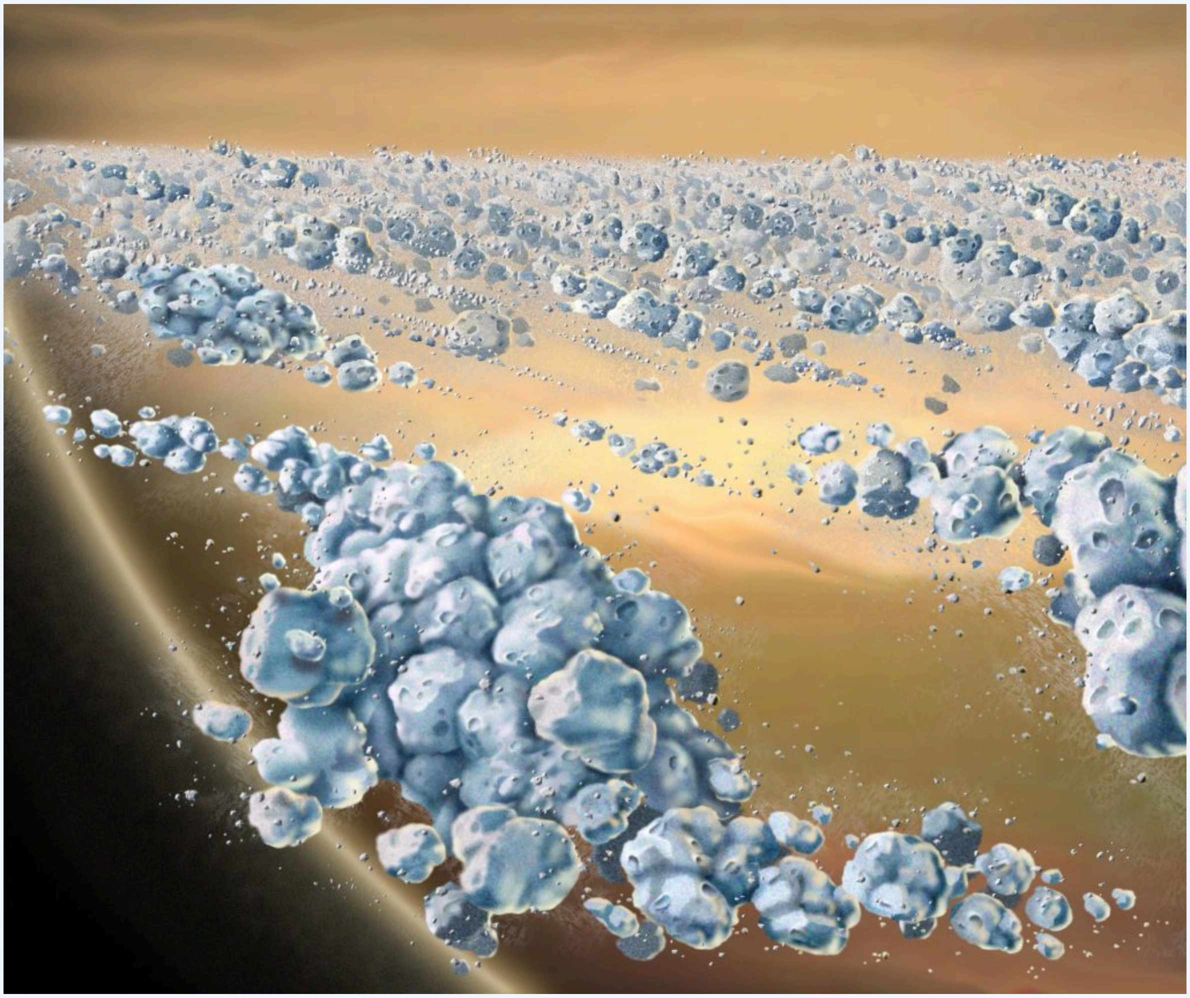


Figure 8.18. Artist's idealized impression of the rings of Saturn as seen from the inside. Note that the rings are mostly made of pieces of water ice of different sizes. Toward the end of its mission, the Cassini spacecraft got nearer to the rings of Saturn, but it never got this close.
PIA10081: Saturn's Recycling Rings by NASA/JPL/University of Colorado, NASA Media Licence.

Rings of Uranus and Neptune

Uranus' rings are narrow and black, making them almost invisible from Earth. The nine main rings were discovered in 1977 from observations made of a star as Uranus passed in front of it. We call such a passage of one astronomical object in front of another an **occultation**. During the 1977 occultation, astronomers expected the star's light to disappear as the planet moved across it. But in addition, the star dimmed briefly several times before Uranus reached it, as each narrow ring passed between the star and the telescope. Thus, the rings were mapped out in detail even though they could not be seen or photographed directly, like counting the number of cars in a train at night by watching the blinking of a light as the cars successively pass in front of it. When Voyager approached Uranus in 1986, it was able to study the rings at close range; the spacecraft also photographed two new rings (Figure 8.19).

Rings of Uranus

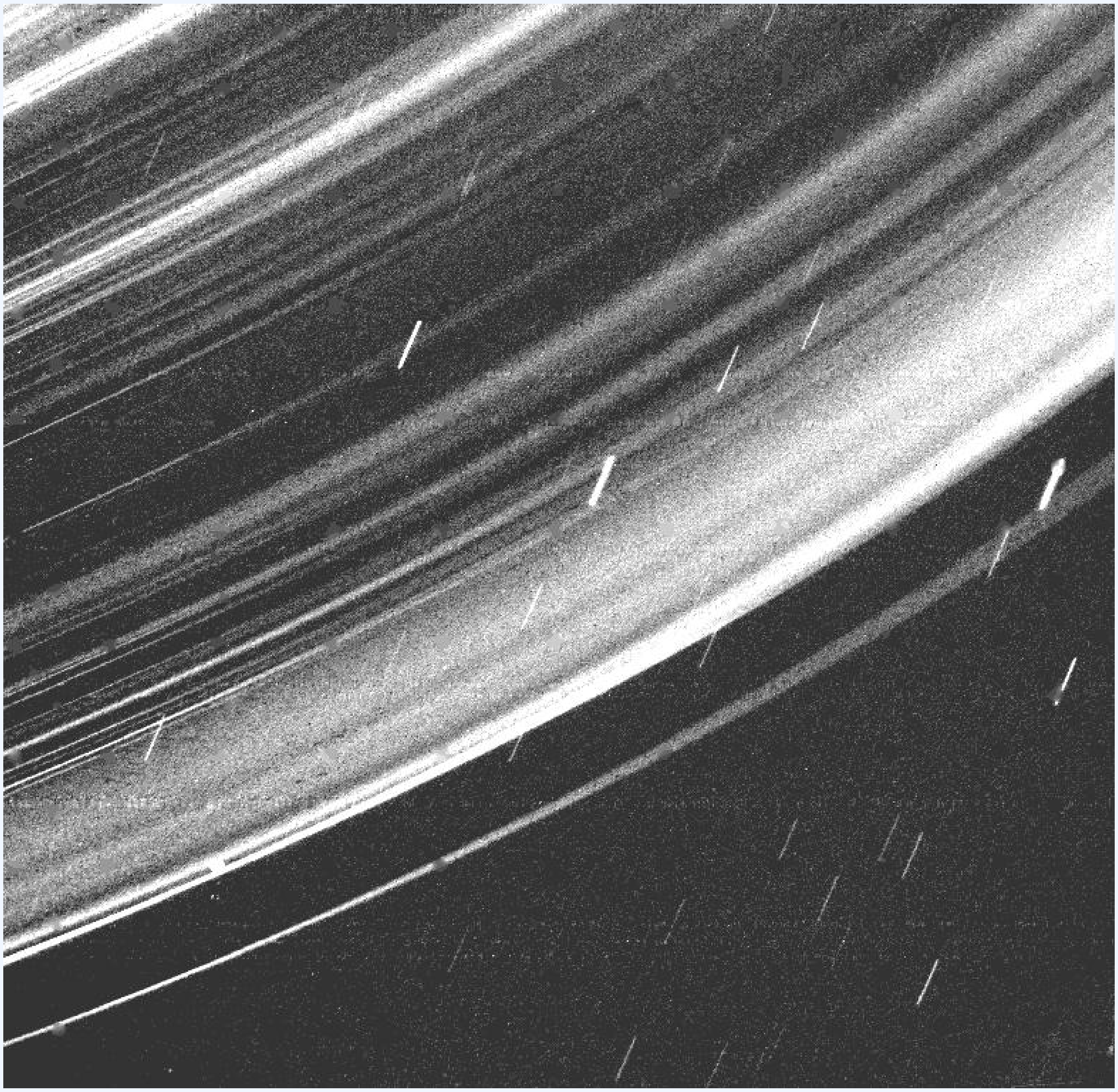


Figure 8.19. The Voyager team had to expose this image for a long time to get a glimpse of Uranus' narrow dark rings. You can see the grainy structure of "noise" in the electronics of the camera in the picture background.
PIA00142: Uranus Ring System by NASA/JPL, NASA Media Licence.

The outermost and most massive of Uranus' rings is called the Epsilon Ring. It is only about 100 kilometres wide and probably no more than 100 metres thick (similar to the F Ring of Saturn). The Epsilon Ring encircles

Uranus at a distance of 51,000 kilometres, about twice the radius of Uranus. This ring probably contains as much mass as all of Uranus' other ten rings combined; most of them are narrow ribbons less than 10 kilometres wide, just the reverse of the broad rings of Saturn.

The individual particles in the uranian rings are nearly as black as lumps of coal. While astronomers do not understand the composition of this material in detail, it seems to consist in large part of carbon and hydrocarbon compounds. Organic material of this sort is rather common in the outer solar system. Many of the asteroids and comets are also composed of dark, tarlike materials. In the case of Uranus, its ten small inner moons have a similar composition, suggesting that one or more moons might have broken up to make the rings.

Neptune's rings are generally similar to those of Uranus but even more tenuous (Figure 8.20). There are only four of them, and the particles are not uniformly distributed along their lengths. Because these rings are so difficult to investigate from Earth, it will probably be a long time before we understand them very well.

Rings of Neptune

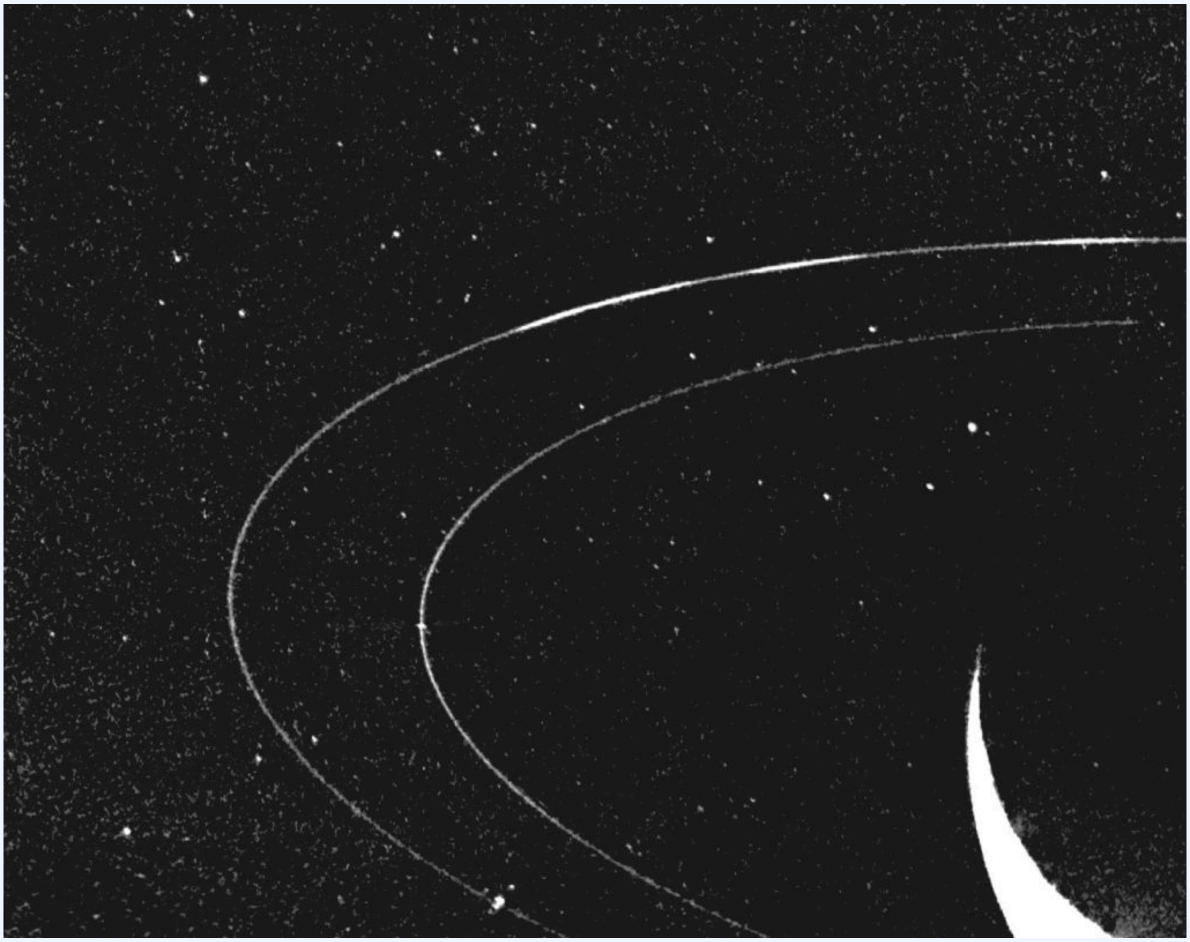


Figure 8.20. This long exposure of Neptune's rings was photographed by Voyager 2. Note the two denser regions of the outer ring.
PIA01493: Neptune's Rings by NASA/JPL, NASA Media Licence.

Interactions between Rings and Moons

Much of our fascination with planetary rings is a result of their intricate structures, most of which owe their existence to the gravitational effect of moons, without which the rings would be flat and featureless. Indeed, it is becoming clear that without moons there would probably be no rings at all because, left to themselves, thin disks of small particles gradually spread and dissipate.

Most of the gaps in Saturn's rings, and also the location of the outer edge of the A Ring, result from gravitational resonances with small inner moons. A resonance takes place when two objects have orbital periods that are exact ratios of each other, such as 1:2 or 2:3. For example, any particle in the gap at the inner side of the Cassini Division of Saturn's rings would have a period equal to one-half that of Saturn's moon Mimas. Such a particle would be nearest Mimas in the same part of its orbit every second revolution. The repeated gravitational tugs of Mimas, acting always in the same direction, would perturb it, forcing it into a new orbit outside the gap. In this way, the Cassini Division became depleted of ring material over long periods of time.

The Cassini mission revealed a great deal of fine structure in Saturn's rings. Unlike the earlier Voyager flybys, Cassini was able to observe the rings for more than a decade, revealing a remarkable range of changes, on time scales from a few minutes to several years. Many of the features newly seen in Cassini data indicated the presence of condensations or small moons only a few tens of metres across imbedded in the rings. As each small moon moves, it produces waves in the surrounding ring material like the wake left by a moving ship. Even when the moon is too small to be resolved, its characteristic waves could be photographed by Cassini.

One of the most interesting rings of Saturn is the narrow F Ring, which contains several apparent ringlets within its 90-kilometer width. In places, the F Ring breaks up into two or three parallel strands that sometimes show bends or kinks. Most of the rings of Uranus and Neptune are also narrow ribbons like the F Ring of Saturn. Clearly, the gravity of some objects must be keeping the particles in these thin rings from spreading out.

As we have seen, the largest features in the rings of Saturn are produced by gravitational resonances with the inner moons, while much of the fine structure is caused by smaller embedded moons. In the case of Saturn's F Ring, close-up images revealed that it is bounded by the orbits of two moons, called Pandora and Prometheus (Figure 8.21). These two small moons (each about 100 kilometres in diameter) are referred to as **shepherd moons**, since their gravitation serves to "shepherd" the ring particles and keep them confined to a narrow ribbon. A similar situation applies to the Epsilon Ring of Uranus, which is shepherded by the moons Cordelia and Ophelia. These two shepherds, each about 50 kilometres in diameter, orbit about 2000 kilometres inside and outside the ring.

Saturn's F Ring and Its Shepherd Moons

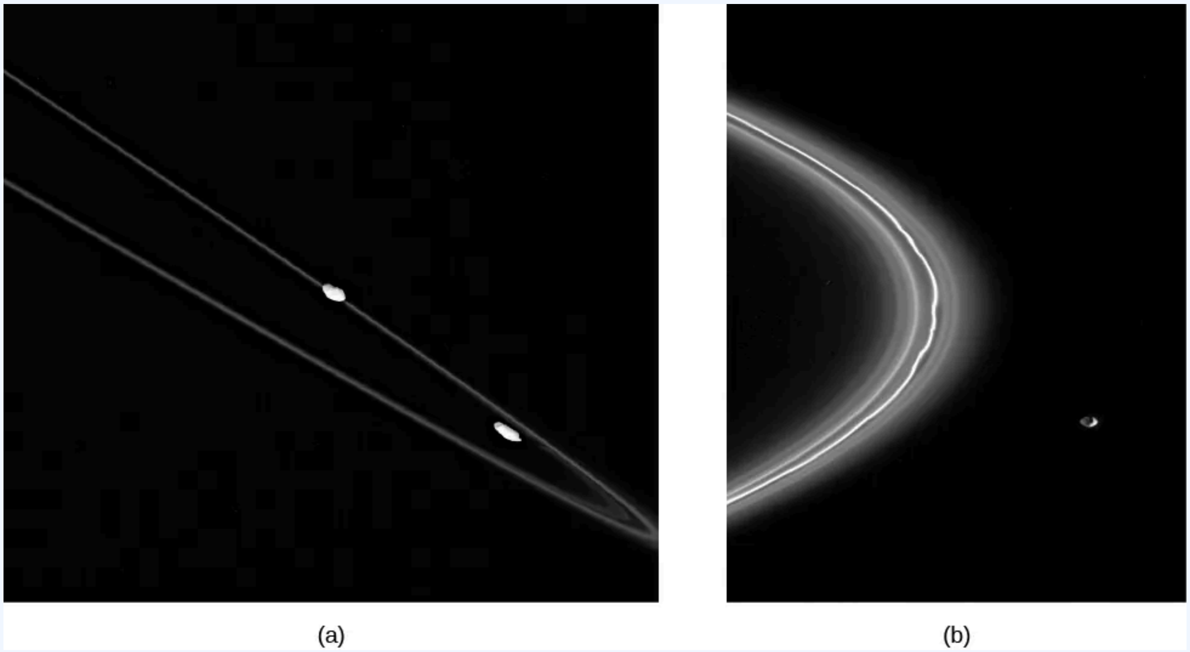


Figure 8.21. (a) This Cassini image shows the narrow, complex F Ring of Saturn, with its two small shepherd moons Pandora (left) and Prometheus (right). (b) In this closer view, the shepherd moon Pandora (84 kilometres across) is seen next to the F ring, in which the moon is perturbing the main (brightest) strand of ring particles as it passes. You can see the dark side of Pandora on this image because it is being illuminated by the light reflected from Saturn.
(a): PIA07653: Close to the Shepherd Moons by NASA/JPL/Space Science Institute, NASA Media Licence.
(b): PIA07523: Pandora's Flocks by NASA/JPL/Space Science Institute, NASA Media Licence.

Attribution

8.4 PLUTO AND CHARON

Almost everything we know about Pluto and its moons comes from the New Horizons spacecraft, which flew by in 2015 before travelling on into the outermost parts of the planetary system. Using data from the New Horizons probe, astronomers have measured the diameter of Pluto as 2370 kilometres, only 60 percent as large as our Moon. From the diameter and mass, we find a density of 1.9 g/cm^3, suggesting that Pluto is a mixture of rocky materials and water ice in about the same proportions as many outer-planet moons.

Parts of Pluto's surface are highly reflective, and its spectrum demonstrates the presence on its surface of frozen methane, carbon monoxide, and nitrogen. The maximum surface temperature ranges from about 50 K when Pluto is farthest from the Sun to 60 K when it is closest. Even this small difference is enough to cause a partial sublimation (going from solid to gas) of the methane and nitrogen ice. This generates an atmosphere when Pluto is close to the Sun, and it freezes out when Pluto is farther away. Observations of distant stars seen through this thin atmosphere indicate that the surface pressure is about a ten-thousandth of Earth's. Because Pluto is a few degrees warmer than Triton, its atmospheric pressure is about ten times greater. This atmosphere contains several distinct haze layers, presumably caused by photochemical reactions, like those in Titan's atmosphere (Figure 8.22).

Haze Layers in the Atmosphere of Pluto

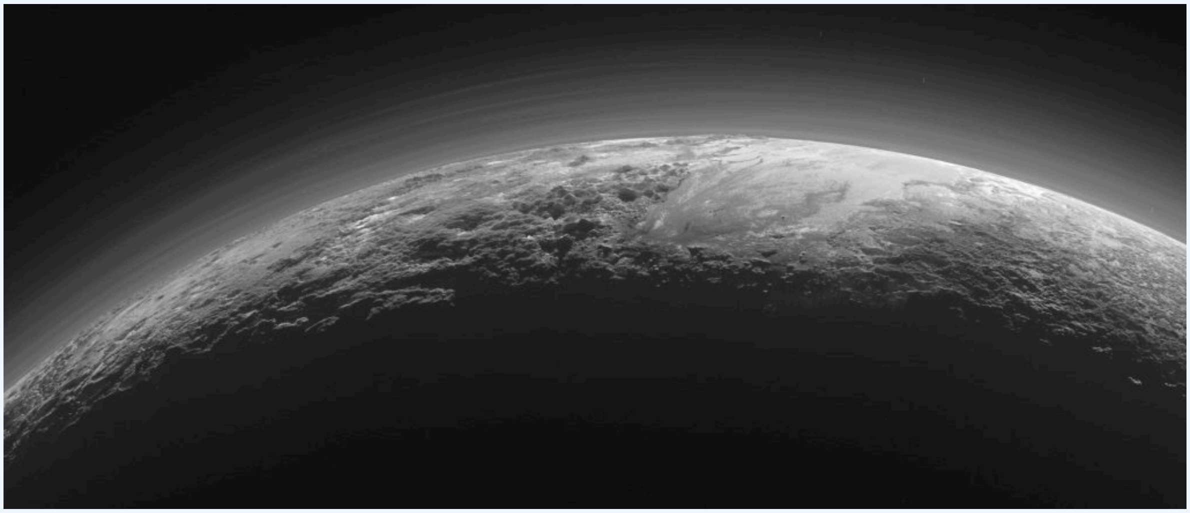


Figure 8.22. This is one of the highest-resolution photos of Pluto, taken by the New Horizons spacecraft 15 minutes after its closest approach. It shows 12 layers of haze. Note also the range of mountains with heights up to 3500 metres.
PIA19948: Pluto's Majestic Mountains, Frozen Plains and Foggy Hazes by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, NASA Media Licence.

Reaching Pluto with a spacecraft was a major challenge, especially in an era when reduced NASA budgets could not support large, expensive missions like Galileo and Cassini. Yet like Galileo and Cassini, a Pluto mission would require a nuclear electric system that used the heat from plutonium to generate the energy to power the instruments and keep them operating far from the warmth of the Sun. NASA made available one of the last of its nuclear generators for such a mission. Assuming an affordable but highly capable spacecraft could be built, there was still the problem of getting to Pluto, nearly 5 billion kilometres from Earth, without waiting decades. The answer was to use Jupiter's gravity to slingshot the spacecraft toward Pluto.

The 2006 launch of New Horizons started the mission with a high speed, and the Jupiter flyby just a year later gave it the required additional boost. The New Horizons spacecraft arrived at Pluto in July 2015, travelling at a relative speed of 14 kilometres per second (or about 50,000 kilometres per hour). With this high speed, the entire flyby sequence was compressed into just one day. Most of the data recorded near closest approach could not be transmitted to Earth until many months later, but when it finally arrived, astronomers were rewarded with a treasure trove of images and data.

Pluto is not the geologically dead world that many anticipated for such a small object—far from it. The division of the surface into areas with different composition and surface texture is apparent in the global colour photo shown in Figure 8.23. The reddish colour is enhanced in this image to bring out differences in colour more clearly. The darker parts of the surface are cratered, but adjacent to them is a nearly featureless light area in the lower right quadrant of this image. This is a huge ice-filled depression called the Sputnik Plains, named for the first artificial Earth satellite. The absence of impact craters suggests a surface no more than 10 million years old. The dark areas show the colours of photochemical haze or smog similar to that in the atmosphere of Titan. The dark material that is staining these old surfaces could come from Pluto's atmospheric haze or from chemical reactions taking place at the surface due to the action of sunlight.

Global Colour Image of Pluto

Figure 8.23. This New Horizons image clearly shows the variety of terrains on Pluto. The dark area in the lower left is covered with impact craters, while the large light area in the centre and lower right is a flat basin devoid of craters. The colours you see are somewhat enhanced to bring out subtle differences. PIA19952: The Rich Colour Variations of Pluto by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, NASA Media Licence.

Figure 8.24 shows some of the remarkable variety of surface features New Horizons revealed. we are finding on

Pluto. Toward the right on this image, we see the shoreline of the Sputnik Plains, showing evidence of relatively recent geological activity. The nitrogen and carbon monoxide ice that fills the Sputnik depression shows cells or polygons with an average width of more than 30 kilometres, caused by slow convection in the ice.

Surface of Pluto

Figure 8.24. This enhanced colour view of a strip of Pluto's surface about 80 kilometres long shows a variety of different surface features. From left to right, we first cross a region of "badlands" with some craters showing, and then move across a wide range of mountains made of water ice and coated with the redder material we saw in the previous image. Then, at right, we arrive at the "shoreline" of the great sea of frozen nitrogen that the mission scientists have named the "Sputnik Plains." This nitrogen sea is divided into mysterious cells or segments that are many kilometres across.
PIA20213: Pluto's Close-up, Now in Colour by NASA/Johns Hopkins University Applied Physics Laboratory/ Southwest Research Institute, NASA Media Licence.

Figure 8.25 shows another view of the boundary between different types of geology. The width of this image is 250 kilometres, and it shows dark, ancient, heavily cratered terrain; dark, uncratered terrain with a hilly surface; smooth, geologically young terrain; and a small cluster of mountains more than 3000 metres high. In the best images, the light areas of nitrogen ice seem to have flowed much like glaciers on Earth, covering some of the older terrain underneath them.

The isolated mountains in the midst of the smooth nitrogen plains are probably also made of water ice, which is very hard at the temperatures on Pluto and can float on frozen nitrogen. Additional mountains, and some hilly terrain that reminded the mission scientists of snakeskin, are visible in part (b) of Figure 8.25.

Diversity of Terrains on Pluto

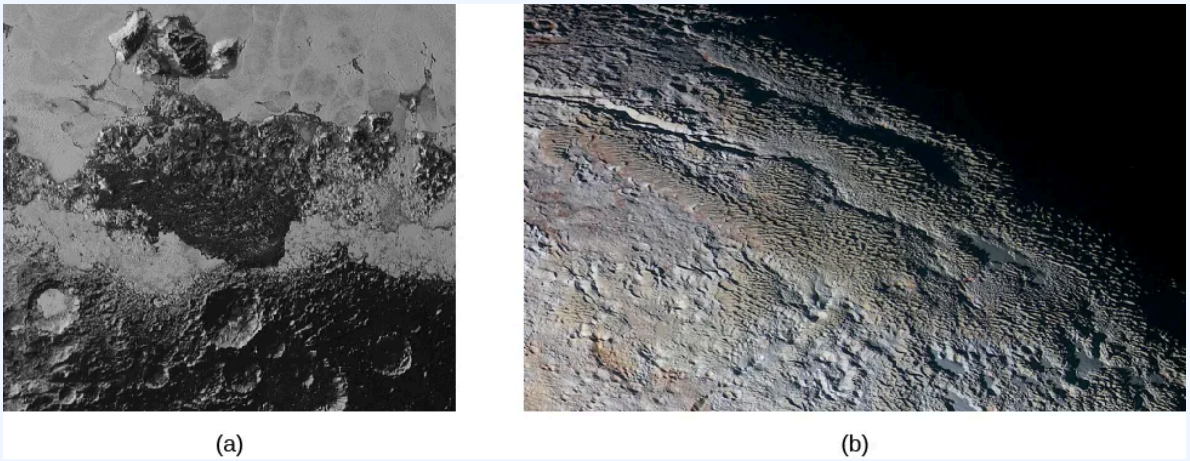


Figure 8.25. (a) In this photo, about 250 kilometres across, we can see many different kinds of terrain. At the bottom are older, cratered highlands; a V-shaped region of hills without cratering points toward the bottom of the image. Surrounding the V-shaped dark region is the smooth, brighter frozen nitrogen plain, acting as glaciers on Earth do. Some isolated mountains, made of frozen water ice, are floating in the nitrogen near the top of the picture. (b) This scene is about 390 kilometres across. The rounded mountains, quite different from those we know on Earth, are named Tartarus Dorsa. The patterns, made of repeating ridges with the more reddish terrain between them, are not yet understood.
(a): PIA19933: Dark Areas by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, NASA Media Licence.
(b): PIA19957: 'Snakeskin' Terrain by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, NASA Media Licence.

To add to the mysteries of Pluto, we show in Figure 8.26 one of the best New Horizons images of Pluto's large moon Charon. Charon is roughly half Pluto's size (its diameter is about the size of Texas) and keeps the same side toward Pluto, just as our Moon keeps the same side toward Earth. What is unique about the Pluto-Charon system, however, is that Pluto also keeps its same face toward Charon. Like two dancers embracing, these two constantly face each other as they spin across the celestial dance floor. Astronomers call this a double tidal lock.

Pluto's Large Moon Charon

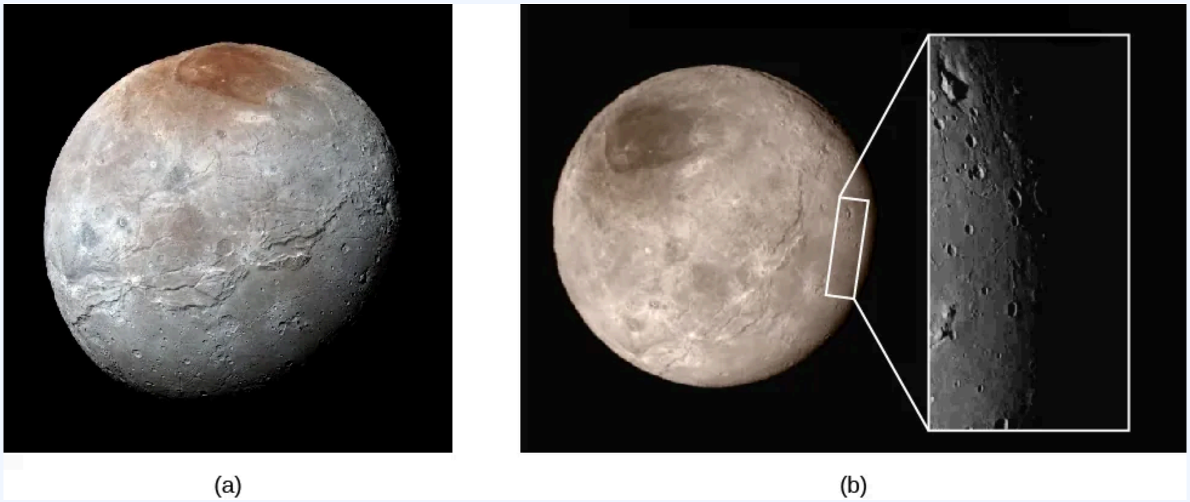


Figure 8.26. (a) In this New Horizons image, the colour has been enhanced to bring out the colour of the moon's strange red polar cap. Charon has a diameter of 1214 kilometres, and the resolution of this image is 3 kilometres. (b) Here we see the moon from a slightly different angle, in true colour. The inset shows an area about 390 kilometres from top to bottom. Near the top left is an intriguing feature—what appears to be a mountain in the middle of a depression or moat.
(a): PIA19968: Charon in Enhanced Colour by NASA/Johns Hopkins University Applied Physics Laboratory/ Southwest Research Institute, NASA Media Licence.
(b): PIA19713: Close-Up of Charon's 'Mountain in a Moat' by NASA/Johns Hopkins University Applied Physics Laboratory/Southwest Research Institute, NASA Media Licence.

What New Horizons showed was another complex world. There are scattered craters in the lower part of the image, but much of the rest of the surface appears smooth. Crossing the center of the image is a belt of rough terrain, including what appear to be tectonic valleys, as if some forces had tried to split Charon apart. Topping off this strange image is a distinctly red polar cap, of unknown composition. Many features on Charon are not yet understood, including what appears to be a mountain in the midst of a low-elevation region.

Attribution

"12.4 Pluto and Charon" from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff, ©

8.5 TRANS-NEPTUNIAN OBJECTS, THE KUIPER BELT, AND THE OORT CLOUD

Trans-Neptunian Objects (TNO) are any solar system minor planet that orbits the sun at a greater average distance than Neptune. Pluto is now considered a TNO, as is Eris. As of July 2014, over 1,500 trans-Neptunian Objects have been cataloged and of these, some 200 have been designated as dwarf planets. From Earth, astronomers study TNO heat emissions, colours, and spectra.The **Kuiper Belt** is a region beyond the orbit of Neptune at 30 Astronomical Units (AU) to about 50 AU from the sun. It is sometimes called the **Edgeworth–Kuiper Belt**. The Kuiper Belt is much larger than the Asteroid Belt. It's about twenty times as wide and twenty to two hundred times as massive. Kuiper Belt Objects, called **KBOs** , are composed of rock and metal, like the asteroids, but also frozen ices like ammonia, methane, and water. Trans-Neptunian Objects are Kuiper Belt Objects (KBOs), but KBOs are not TNOs because the distance range of KBOs from the sun is much farther out in the solar system. At least two moons – Saturn's Phoebe and Neptune's Triton – are believed to be captured KBOs. To date, over 1,000 KBOs have been discovered and around 100,000 KBOs are suspected to exist.

KBOs are fragments from the original solar system's protoplanetary disc that did not combine to form the more-massive and larger bodies. Astronomers believe that belts of fragments like the solar system's Kuiper Belt are present around other stellar systems being studied.

Beyond Neptune

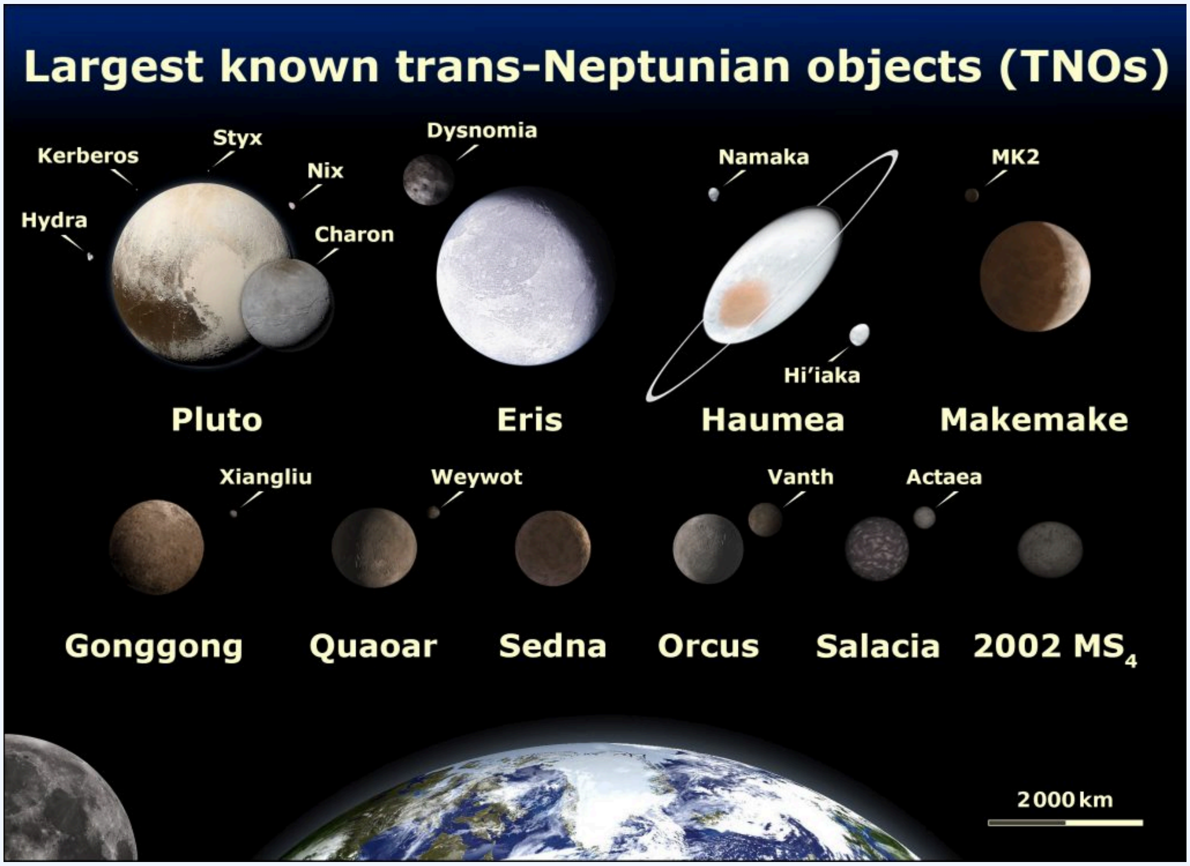


Figure 8.27. Comparison of the largest TNOs: Pluto, Eris, Haumea, Makemake, Gonggong, Quaoar, Sedna, Orcus, Salacia, and 2002 MS4. All but two of these TNOs (Sedna and 2002 MS4) are known to have at least one moon. The top four are IAU-accepted dwarf planets while the bottom six are dwarf-planet candidates that are accepted as dwarf planets by several astronomers.
Largest known Trans Neptunian Objects by Lexicon, CC BY-SA 3.0.

The **Oort Cloud** is a hypothesized spherical cloud of icy objects up to 50,000 AU from the sun. This spherical cloud is the primary source of long-period comets. No Oort Cloud Objects have been found, to date. Any object within the Oort Cloud would be called an **Oort Cloud Object, (OCO)**. The Oort Cloud would represent the physical boundary of the solar system.

Attribution

8.6 DWARF PLANETS

Ceres

Ceres is the most massive object in the asteroid belt. Before 2006, Ceres was considered the largest of the asteroids, with only about 1.3 percent of the mass of the Earth's Moon. However, unlike the asteroids, Ceres has enough mass that its gravity causes it to be shaped like a sphere. Like Pluto, Ceres is rocky. Ceres orbits the Sun, is round, and is not a moon. As part of the asteroid belt, its orbit is full of other smaller bodies, so Ceres fails the fourth criterion for being a planet.

Ceres' Internal Structure

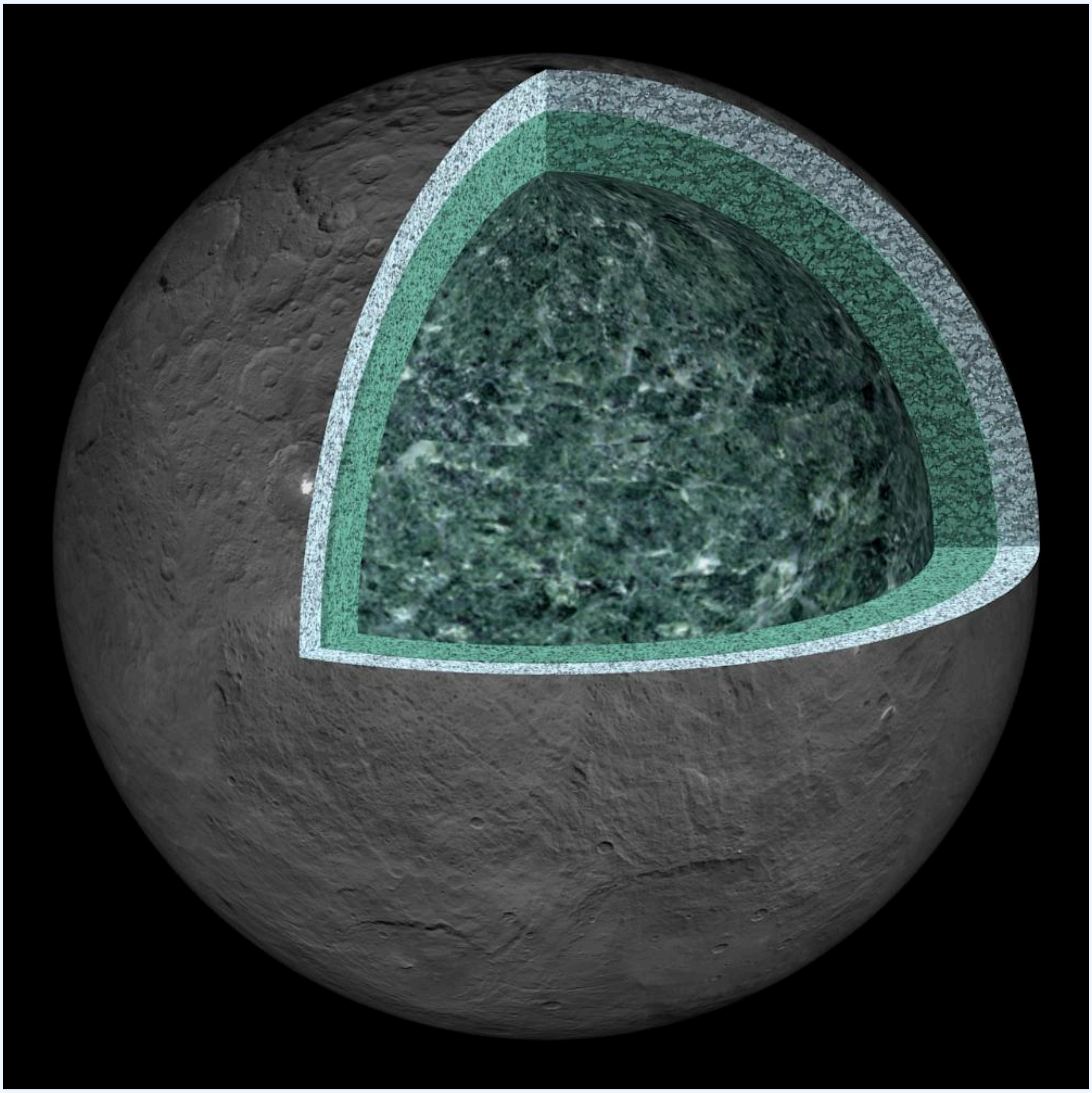


Figure 8.28. Informed by NASA's Dawn mission data, researchers inferred that Ceres possesses discrete layers. The inner "mantle" consists of hydrated rocks, whereas the outer crust, approximately 40 kilometres thick, contains ice, salts, and minerals. Interposed is a layer with potential brine, extending down 100 kilometres. Details further below this depth remain undisclosed.
PIA22660: Ceres' Internal Structure (Artist's Concept) by NASA/JPL-Caltech/UCLA/MPS/DLR/IDA, NASA Media Licence.

Makemake

Makemake is the third largest and second brightest dwarf planet we have discovered so far. With a diameter estimated to be between 1,300 and 1,900 km, it is about three-quarters the size of Pluto. Makemake orbits the Sun in 310 years at a distance between 38.5 to 53 AU. It is thought to be made of methane, ethane, and nitrogen ices.

Eris

Eris is the most widely-known dwarf planet in the solar system, which is roughly 27 percent more massive than Pluto. The object was not discovered until 2003 because it is about three times farther from the Sun than Pluto, and almost 100 times farther from the Sun than Earth is. For a short time, Eris was considered the "tenth planet" in the solar system, but its discovery helped to prompt astronomers to better define planets and dwarf planets in 2006. Eris also has a small moon, Dysnomia, that orbits it once about every 16 days.

Astronomers know there may be other dwarf planets in the outer reaches of the solar system. Haumea was made a dwarf planet in 2008, and so now the total is five. Quaoar, Varuna, and Orcus may be added to the list of dwarf planets in the future. We still have a lot to discover and explore.

Attribution

8.7 ASTEROIDS AND METEOROIDS

Asteroids

Asteroids are tiny, rocky bodies that orbit the Sun. "Asteroid" means "star-like," and in a telescope, asteroids look like points of light, just like stars. Asteroids are irregularly shaped because they do not have enough gravity to become round. They are also too small to maintain an atmosphere, and without internal heat, they are not geologically active. Collisions with other bodies may break up the asteroid or create craters on its surface.

Asteroid impacts have had dramatic impacts on the shaping of the planets, including Earth. Early impacts caused the planets to grow as they cleared their portions of space. An impact with an asteroid about the size of Mars caused Earth's fragments to fly into space and ultimately create the moon. Asteroid impacts are linked to mass extinctions throughout Earth history.

The Asteroid Belt

Hundreds of thousands of asteroids have been discovered in our solar system. They are still being discovered at a rate of about 5,000 new asteroids per month. The majority of the asteroids are found in between the orbits of Mars and Jupiter in a region called the **Asteroid Belt**. Although there are many thousands of asteroids in the asteroid belt, their total mass adds up to only about 4 percent of Earth's moon.

Scientists think that the bodies in the asteroid belt formed during the formation of the solar system. The asteroids might have come together to make a single planet, but they were pulled apart by Jupiter's intense gravity.

Near-Earth Asteroids

More than 4,500 asteroids cross Earth's orbit; they are near-Earth asteroids. Between 500 and 1,000 of these are over 1 km in diameter. Any object whose orbit crosses Earth's can collide with Earth, and many asteroids do. On average, each year, a rock about 5–10 m in diameter hits Earth. Since past asteroid impacts have been implicated in mass extinctions, astronomers are always on the lookout for new asteroids and follow the known near-Earth asteroids closely to predict a possible collision as early as possible.

Scientists are interested in asteroids because they are representatives of the earliest solar system. Eventually, asteroids could be mined for rare minerals or construction projects in space. A few missions have studied asteroids directly. NASA's DAWN mission orbited asteroid Vesta from July 2011 to September 2012 and is on its way to meet dwarf planet Ceres in 2015.

Thousands of objects, including comets and asteroids, are zooming around our solar system; some could be on a collision course with Earth. A meteor is a streak of light across the sky. People call them shooting stars, but small pieces of matter are burning up as they enter Earth's atmosphere from space.

Meteors are called meteoroids before they reach Earth's atmosphere. Meteoroids are smaller than asteroids and range from the size of boulders down to the size of tiny sand grains. Still, smaller objects are called interplanetary dust. When Earth passes through a cluster of meteoroids, there is a meteor shower. These clusters are often remnants left behind by comet tails.

Meteorites

Although most meteors burn up in the atmosphere, larger meteoroids may strike the Earth's surface to create a meteorite. **Meteorites** are valuable to scientists because they provide clues about our solar system. Many meteorites are from asteroids that formed when the solar system formed. A few meteorites are made of rocky material that is thought to have come from Mars when an asteroid impact shot material off the Martian surface and into space.

Attribution

"2.7 Other Objects in the Solar System" from Physical Geography and Natural Disasters by R. Adam Dastrup, MA, GISP is licensed under a Creative Commons Attribution Non-Commerical Share-Alike 4.0 International License, except where otherwise noted.

8.8 COMETS

Comets differ from asteroids primarily in their icy composition, a difference that causes them to brighten dramatically as they approach the Sun, forming a temporary atmosphere. In some early cultures, these so-called "hairy stars" were considered omens of disaster. Today, we no longer fear comets, but eagerly anticipate those that come close enough to us to put on a good sky show.

Appearance of Comets

A **comet** is a relatively small chunk of icy material (typically a few kilometres across) that develops an atmosphere as it approaches the Sun. Later, there may be a very faint, nebulous tail, extending several million kilometres away from the main body of the comet. Comets have been observed from the earliest times: accounts of comets are found in the histories of virtually all ancient civilizations. The typical comet, however, is not spectacular in our skies, instead having the appearance of a rather faint, diffuse spot of light somewhat smaller than the Moon and many times less brilliant. (Comets seemed more spectacular to people before the invention of artificial lighting, which compromises our view of the night sky.)

Like the Moon and planets, comets appear to wander among the stars, slowly shifting their positions in the sky from night to night. Unlike the planets, however, most comets appear at unpredictable times, which perhaps explain why they frequently inspired fear and superstition in earlier times. Comets typically remain visible for periods that vary from a couple of weeks to several months. We'll say more about what they are made of and how they become visible after we discuss their motions.

Note that still images of comets give the impression that they are moving rapidly across the sky, like a bright meteor or shooting star. Looking only at such images, it is easy to confuse comets and meteors. But seen in the real sky, they are very different: the meteor burns up in our atmosphere and is gone in a few seconds, whereas the comet may be visible for weeks in nearly the same part of the sky.

Comet Orbits

The study of comets as members of the solar system dates from the time of Isaac Newton, who first suggested that they orbited the Sun on extremely elongated ellipses. Newton's colleague Edmund Halley developed these ideas, and in 1705, he published calculations of 24 comet orbits. In particular, he noted that the orbits of the bright comets that had appeared in the years 1531, 1607, and 1682 were so similar that the three could well be the same comet, returning to perihelion (closest approach to the Sun) at average intervals of 76 years. If so, he predicted that the object should next return about 1758. Although Halley had died by the time the comet appeared as he predicted, it was given the name Comet Halley (rhymes with "valley") in honour of the astronomer who first recognized it as a permanent member of our solar system, orbiting around the Sun. Its aphelion (furthest point from the Sun) is beyond the orbit of Neptune.

We now know from historical records that Comet Halley has actually been observed and recorded on every passage near the Sun since 239 BCE at intervals ranging from 74 to 79 years. The period of its return varies somewhat because of orbital changes produced by the pull of the giant planets. In 1910, Earth was brushed by the comet's tail, causing much needless public concern. Comet Halley last appeared in our skies in 1986 (Figure 8.29), when it was met by several spacecraft that gave us a wealth of information about its makeup; it will return in 2061.

Comet Halley

Figure 8.29. This composite of three images (one in red, one in green, one in blue) shows Comet Halley as seen with a large telescope in Chile in 1986. During the time the three images were taken in sequence, the comet moved among the stars. The telescope was moved to keep the image of the comet steady, causing the stars to appear in triplicate (once in each colour) in the background.
Comet Halley from La Silla in 1986 by ESO, ESO Media Licence.

Edmund Halley: Astronomy's Renaissance Man

Edmund Halley (Figure 8.30), a brilliant astronomer who made contributions in many fields of science and statistics, was by all accounts a generous, warm, and outgoing person. In this, he was quite the opposite of his good friend Isaac Newton, whose great work, the *Principia*, Halley encouraged, edited, and helped pay to publish. Halley himself published his first scientific paper at age 20, while still in college. As a result, he was given a royal commission to go to Saint Helena (a remote island off the coast of Africa where Napoleon would later be exiled) to make the first telescopic survey of the southern sky. After returning, he received the equivalent of a master's degree and was elected to the prestigious Royal Society in England, all at the age of 22.

In addition to his work on comets, Halley was the first astronomer to recognize that the so-called "fixed" stars move relative to each other, by noting that several bright stars had changed their positions since Ptolemy's publication of the ancient Greek catalogs. He wrote a paper on the possibility of an infinite universe, proposed that some stars may be variable, and discussed the nature and size of *nebulae* (glowing cloudlike structures visible in telescopes). While in Saint Helena, Halley observed the planet Mercury going across the face of the Sun and developed the mathematics of how such transits could be used to establish the size of the solar system.

In other fields, Halley published the first table of human life expectancies (the precursor of life-insurance statistics); wrote papers on monsoons, trade winds, and tides (charting the tides in the English Channel for the first time); laid the foundations for the systematic study of Earth's magnetic field; studied evaporation and how inland waters become salty; and even designed an underwater diving bell. He served as a British diplomat, advising the emperor of Austria and squiring the future czar of Russia around England (avidly discussing, we are told, both the importance of science and the quality of local brandy).

In 1703, Halley became a professor of geometry at Oxford, and in 1720, he was appointed Astronomer Royal of England. He continued observing Earth and the sky and publishing his ideas for another 20 years, until death claimed him at age 85.

Edmund Halley (1656–1742)

Figure 8.30. Halley was a prolific contributor to the sciences. His study of comets at the turn of the eighteenth century helped predict the orbit of the comet that now bears his name.
Portrait of Edmond Halley (1656–1742) by Thomas Murray, CC0 1.0.

Only a few comets return in a time measureable in human terms (shorter than a century or two), like Comet Halley does; these are called **short-periodcomets**. Many short-period comets have had their orbits changed by coming too close to one of the giant planets—most often Jupiter (and they are thus sometimes called Jupiter-family comets). Most comets have long periods and will take thousands of years to return, if they return at all. As we will see later in this chapter, most Jupiter-family comets come from a different source than the **long-periodcomets** (those with orbital periods longer than about a century).

Observational records exist for thousands of comets. We were visited by two bright comets in recent decades. First, in March 1996, came Comet Hyakutake, with a very long tail. A year later, Comet Hale-Bopp appeared; it was as bright as the brightest stars and remained visible for several weeks, even in urban areas. Since then, there have been few comets visible to the naked eye, and astronomers (both professional and amateur) eagerly await such a visitor.

Table 8.4 lists some well-known comets whose history or appearance is of special interest.

Table 8.4. Some Interesting Comets

Name	Period	Significance
Great Comet of 1577	Long	Tycho Brahe showed it was beyond the Moon (a big step in our understanding)
Great Comet of 1843	Long	Brightest recorded comet; visible in daytime
Daylight Comet of 1910	Long	Brightest comet of the twentieth century
West	Long	Nucleus broke into pieces (1976)
Hyakutake	Long	Passed within 15 million km of Earth (1996)
Hale–Bopp	Long	Brightest recent comet (1997)
Swift-Tuttle	133 years	Parent comet of Perseid meteor shower
Halley	76 years	First comet found to be periodic; explored by spacecraft in 1986
Borrelly	6.8 years	Flyby by Deep Space 1 spacecraft (2000)
Biela	6.7 years	Broke up in 1846 and not seen again
Churyumov-Gerasimenko	6.5 years	Target of Rosetta mission (2014–16)
Wild 2	6.4 years	Target of Stardust sample return mission (2004)
Tempel 1	5.7 years	Target of Deep Impact mission (2005)
Encke	3.3 years	Shortest known period

The Comet's Nucleus

When we look at an active comet, all we normally see is its temporary atmosphere of gas and dust illuminated by sunlight. This atmosphere is called the comet’s head or **coma**. Since the gravity of such small bodies is very weak, the atmosphere is rapidly escaping all the time; it must be replenished by new material, which has to come from somewhere. The source is the small, solid nucleus inside, just a few kilometres across, usually hidden by the glow from the much-larger atmosphere surrounding it. The **nucleus** is the real comet, the fragment of ancient icy material responsible for the atmosphere and the tail (Figure 8.31).

Parts of a Comet

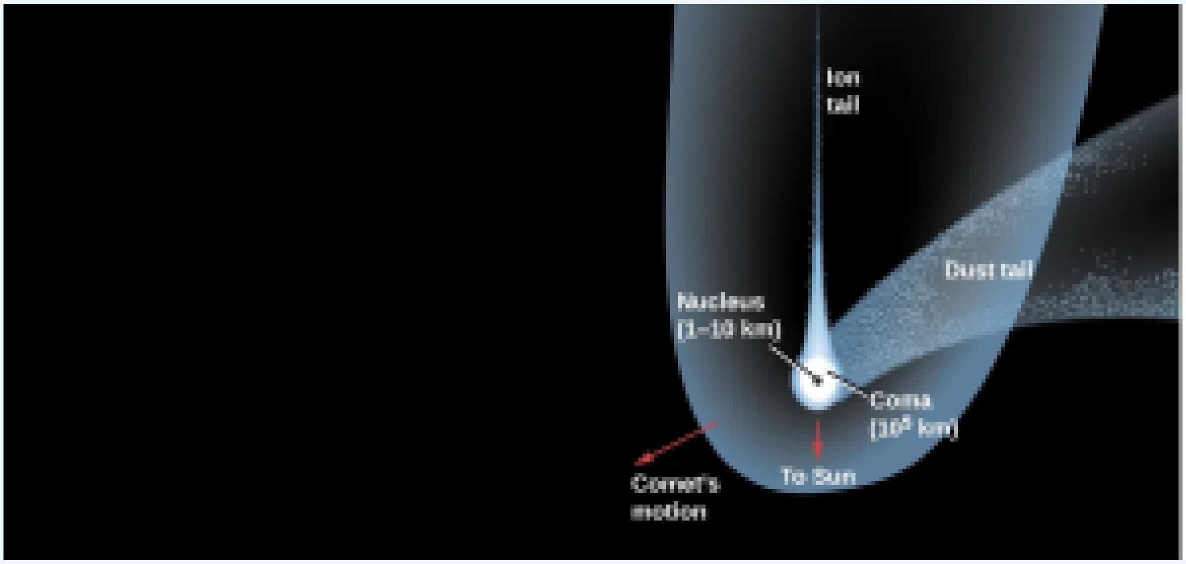


Figure 8.31. This schematic illustration shows the main parts of a comet. Note that the different structures are not to scale.

The modern theory of the physical and chemical nature of comets was first proposed by Harvard astronomer Fred Whipple in 1950. Before Whipple’s work, many astronomers thought that a comet’s nucleus might be a loose aggregation of solids, sort of an orbiting “gravel bank,” Whipple proposed instead that the nucleus is a solid object a few kilometres across, composed in substantial part of water ice (but with other ices as well) mixed with silicate grains and dust. This proposal became known as the “dirty snowball” model.

The water vapor and other volatiles that escape from the nucleus when it is heated can be detected in the comet’s head and tail, and therefore, we can use spectra to analyze what atoms and molecules the nucleus ice consists of. However, we are somewhat less certain of the non-icy component. We have never identified a fragment of solid matter from a comet that has survived passage through Earth’s atmosphere. However, spacecraft that have approached comets have carried dust detectors, and some comet dust has even been returned to Earth (see Figure 8.32). It seems that much of the “dirt” in the dirty snowball is dark, primitive hydrocarbons and silicates, rather like the material thought to be present on the dark, primitive asteroids.

Captured Comet Dust

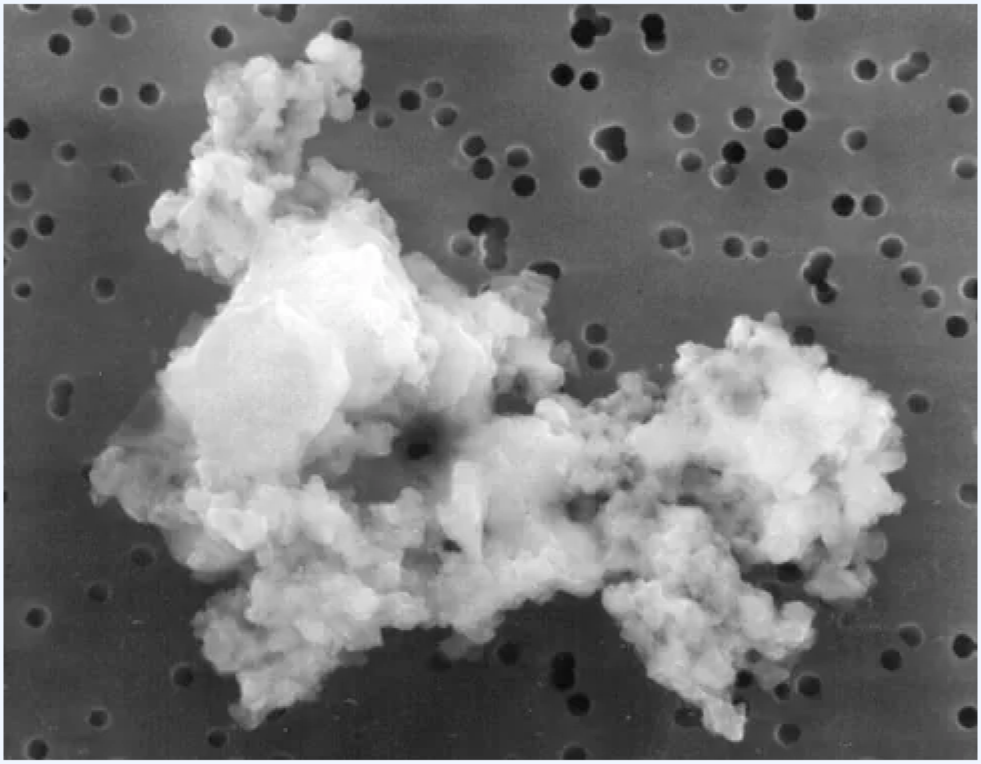


Figure 8.32. This particle (seen through a microscope) is believed to be a tiny fragment of cometary dust, collected in the upper atmosphere of Earth. It measures about 10 microns, or 1/100 of a millimetre, across.
Interplanetary Dust Particle by NASA/JPL, NASA JPL Media Licence.

Since the nuclei of comets are small and dark, they are difficult to study from Earth. Spacecraft did obtain direct measurements of a comet nucleus, however, in 1986, when three spacecraft swept past Comet Halley at close range (see Figure 8.33). Subsequently, other spacecraft have flown close to other comets. In 2005, the NASA *Deep Impact* spacecraft even carried a probe for a high-speed impact with the nucleus of Comet Tempel 1. But by far, the most productive study of a comet has been by the 2015 Rosetta mission, which we will discuss shortly.

Close-up of Comet Halley

Figure 8.33. This historic photograph of the black, irregularly shaped nucleus of Comet Halley was obtained by the ESA Giotto spacecraft from a distance of about 1000 kilometres. The bright areas are jets of material escaping from the surface. The length of the nucleus is 10 kilometres, and details as small as 1 kilometre can be made out.
PIA17485: Comet Halley by NASA/ESA/Giotto Project, NASA Media Licence.

Attribution

"13.3 The "Long-Haired" Comets" from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

8.9 KEY TERMS

Asteroids are tiny, rocky bodies that orbit the Sun. 8.7

Asteroid Belt: a region between the orbits of Mars and Jupiter where the majority of asteroids are found. 8.7

Coma: a comet's temporary atmosphere of gas and dust illuminated by sunlight. 8.8

Comet: a relatively small chunk of icy material (typically a few kilometres across) that develops an atmosphere as it approaches the Sun. 8.8

Ceres is the most massive object in the asteroid belt. 8.6

Differentiated: to be separated into layers of different density materials. 8.1

Eris is the most widely-known dwarf planet in the solar system, which is roughly 27 percent more massive than Pluto. 8.6

Kuiper Belt is a region beyond the orbit of Neptune at 30 Astronomical Units (AU) to about 50 AU from the sun. 8.5

Kuiper Belt Objects (KBOs): objects composed of rock and metal, like the asteroids, but also frozen ices like ammonia, methane, and water. 8.5

Long-periodcomets: comets with orbital periods longer than about a century. 8.8

Makemake is the third largest and second brightest dwarf planet we have discovered so far. 8.6

Meteorite: a large meteoroid formed from an asteroid that strikes the Earth's surface. 8.7

Nucleus (of a comet): the real source of a comet, which is the fragment of ancient icy material responsible for the atmosphere and the tail. 8.8

Occultation: the passage of one astronomical object in front of another. 8.3

Oort Cloud: a hypothesized spherical cloud of icy objects up to 50,000 AU from the sun. 8.5

Oort Cloud Object (OCO): name that would be given to any object found in the Oort Cloud. 8.5

Shepherd moons: moons whose gravitation serves to "shepherd" the ring particles and keep them confined to a narrow ribbon; examples include two of Saturn's moons: Pandora and Prometheus. 8.3

Short-period comets: comets with orbital periods shorter than about a century. 8.8

Tidal force: the result from the unequal gravitational pull on two sides of a body. 8.1

Trans-Neptunian Objects (TNO) are any solar system minor planet that orbits the sun at a greater average distance than Neptune. 8.5

CHAPTER 9: STARS: AN INTRODUCTORY SURVEY

Chapter Overview

9.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Describe the role of stars as fundamental building blocks of galaxies and explain how their age, composition, and distribution provide insights into galactic history and evolution.
- Apply the principles of parallax and stellar distances to calculate the approximate distance of a star from Earth when provided with its measured parallax angle, illustrating an understanding of the inverse relationship between parallax and distance.
- Define luminosity and explain its significance in relation to stars' energy output.
- Explain the inverse square law and how it determines the relationship between distance and perceived brightness of stars.
- Outline the challenges astronomers face in directly imaging exoplanets and describe the techniques used to overcome these challenges.

9.1 INTRODUCTION

"Stars are the most widely recognized astronomical objects, and represent the most fundamental building blocks of galaxies. The age, distribution, and composition of the stars in a galaxy trace the history, dynamics, and evolution of that galaxy. Moreover, stars are responsible for the manufacture and distribution of heavy elements such as carbon, nitrogen, and oxygen, and their characteristics are intimately tied to the characteristics of the planetary systems that may coalesce about them. Consequently, the study of the birth, life, and death of stars is central to the field of astronomy (Science Mission Directorate, n.d.)."

Although constellations have stars that usually only appear to be close together, stars may be found in the same portion of space. Stars that are grouped tightly together are called **star systems**. Larger groups of hundreds or thousands of stars are called **star clusters**. The image shown here is a famous star cluster known as Pleiades, which can be seen with the naked autumn sky.

Although the star humans know best is a single star, many stars – in fact, more than half of the bright stars in our galaxy – are star systems. A system of two stars orbiting each other is a binary star. A system with more than two stars orbiting each other is a multiple star system. The stars in a binary or multiple star system are often so close together that they appear as only through a telescope can the pair be distinguished.

Star clusters are divided into two main types, open clusters and globular clusters. **Open clusters** are groups of up to a few thousand stars that are loosely held together by gravity. Pleiades is an open cluster that is also called the Seven Sisters. Open clusters tend to be blue and often contain glowing gas and dust and are made of young stars formed from the same nebula. The stars may eventually be pulled apart by gravitational attraction to other objects.

Globular clusters are groups of tens to hundreds of thousands of stars held tightly together by gravity. Globular clusters have a definite, spherical shape and contain mostly reddish stars. The stars are closer together, closer to the center of the cluster. Globular clusters do not have much dust in them — the dust has already formed into stars.

Attribution

"2.1 The Universe" from Physical Geography and Natural Disasters by R. Adam Dastrup, MA, GISP is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, except where otherwise noted.

Science Mission Directorate. (n.d.). “Stars”. Retrieved April 26, 2020, from https://science.nasa.gov/astrophysics/focus-areas/how-do-stars-form-and-evolve

9.2 PARALLAX AND THE PARSEC

Distance within the Solar System

The work of Copernicus and Kepler established the ***relative* distances** of the planets—that is, how far from the Sun one planet is compared to another. But their work could not establish the ***absolute* distances** (in light-seconds or metres or other standard units of length). This is like knowing the height of all the students in your class only as compared to the height of your astronomy instructor, but not in inches or centimetres. Somebody's height has to be measured directly.

Similarly, to establish absolute distances, astronomers had to measure one distance in the solar system directly. Generally, the closer to us the object is, the easier such a measurement would be. Estimates of the distance to Venus were made as Venus crossed the face of the Sun in 1761 and 1769, and an international campaign was organized to estimate the distance to the asteroid Eros in the early 1930s, when its orbit brought it close to Earth. More recently, Venus crossed (or *transited*) the surface of the Sun in 2004 and 2012, and allowed us to make a modern distance estimate, although, as we will see below, by then it wasn't needed. This transit is pictured in Figure 9.1.

Venus Transits the Sun, 2012

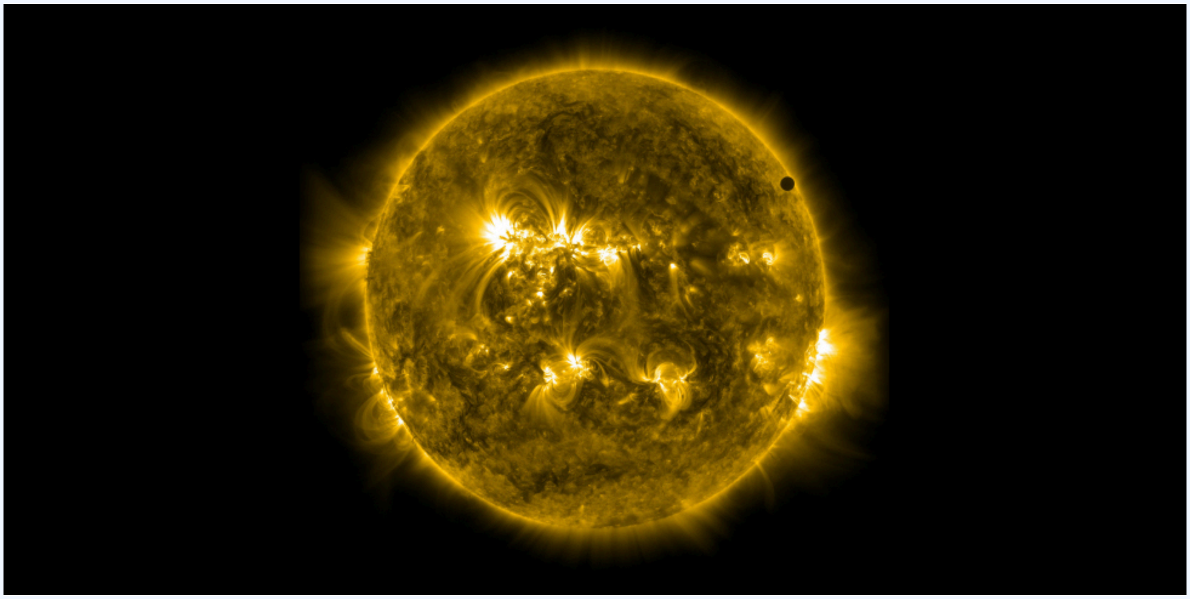


Figure 9.1. This striking "picture" of Venus crossing the face of the Sun (it's the black dot at about 2 o'clock) is more than just an impressive image. Taken with the Solar Dynamics Observatory spacecraft and special filters, it shows a modern transit of Venus. Such events allowed astronomers in the 1800s to estimate the distance to Venus. They measured the time it took Venus to cross the face of the Sun from different latitudes on Earth. The differences in times can be used to estimate the distance to the planet. Today, radar is used for much more precise distance estimates.
SDO's Ultra-high Definition View of 2012 Venus Transit – 171 Angstrom by NASA/SDO, AIA, NASA Media License.

The key to our modern determination of solar system dimensions is radar, a type of radio wave that can bounce off solid objects. A radar is pictured in Figure 9.2. As discussed in several earlier chapters, by timing how long a radar beam (travelling at the speed of light) takes to reach another world and return, we can measure the distance involved very accurately. In 1961, radar signals were bounced off Venus for the first time, providing a direct measurement of the distance from Earth to Venus in terms of light-seconds (from the roundtrip travel time of the radar signal).

Subsequently, radar has been used to determine the distances to Mercury, Mars, the satellites of Jupiter, the rings of Saturn, and several asteroids. Note, by the way, that it is not possible to use radar to measure the distance to the Sun directly because the Sun does not reflect radar very efficiently. But we can measure the distance to many other solar system objects and use Kepler's laws to give us the distance to the Sun.

Radar Telescope

Figure 9.2. This dish-shaped antenna, part of the NASA Deep Space Network in California's Mojave Desert, is 70 metres wide. Nicknamed the "Mars antenna," this radar telescope can send and receive radar waves, and thus measure the distances to planets, satellites, and asteroids. Image by NASA/JPL-Caltech, NASA Media License.

From the various (related) solar system distances, astronomers selected the average distance from Earth to the Sun as our standard "measuring stick" within the solar system. When Earth and the Sun are closest, they are about 147.1 million kilometres apart; when Earth and the Sun are farthest, they are about 152.1 million kilometres apart. The average of these two distances is called the astronomical unit (AU). We then express all the other distances in the solar system in terms of the AU. Years of painstaking analyses of radar measurements have led to a determination of the length of the AU to a precision of about one part in a billion. The length of

1 AU can be expressed in light travel time as 499.004854 light-seconds, or about 8.3 light-minutes. If we use the definition of the meter given previously, this is equivalent to 1 AU = 149,597,870,700 metres.

These distances are, of course, given here to a much higher level of precision than is normally needed. In this text, we are usually content to express numbers to a couple of significant places and leave it at that. For our purposes, it will be sufficient to round off these numbers:

$$\begin{aligned}&\text{speed of light: } c = 3\times10^8\,\text{m/s} = 3\times10^5\ \text{km/s}\\&\text{length of light-second: ls} = 3\times10^8\,\text{m} = 3\times10^5\,\text{km}\\&\text{astronomical unit: AU} = 1.50\times10^{11}\,\text{m} = 1.50\times10^8\,\text{km} = 500\ \text{light-seconds}\end{aligned}$$

We now know the absolute distance scale within our own solar system with fantastic accuracy. This is the first link in the chain of cosmic distances.

The distances between the celestial bodies in our solar system are sometimes difficult to grasp or put into perspective. This interactive website provides a "map" that shows the distances by using a scale at the bottom of the screen and allows you to scroll (using your arrow keys) through screens of "empty space" to get to the next planet—all while your current distance from the Sun is visible on the scale. Direct link: http://joshworth.com/dev/pixelspace/pixelspace_solarsystem.html

It is an enormous step to go from the planets to the stars. For example, our Voyager 1 probe, which was launched in 1977, has now traveled farther from Earth than any other spacecraft. As of July 2023, Voyager 1 is 160 AU from the Sun. The nearest star, however, is hundreds of thousands of AU from Earth. Even so, we can, in principle, survey distances to the stars using the same technique that a civil engineer employs to survey the distance to an inaccessible mountain or tree—the method of **triangulation**.

Triangulation in Space

A practical example of triangulation is your own depth perception. As you are pleased to discover every morning when you look in the mirror, your two eyes are located some distance apart. You therefore view the world from two different vantage points, and it is this dual perspective that allows you to get a general sense of how far away objects are.

To see what we mean, take a pen and hold it a few inches in front of your face. Look at it first with one eye (closing the other) and then switch eyes. Note how the pen seems to shift relative to objects across the room.

Now hold the pen at arm's length: the shift is less. If you play with moving the pen for a while, you will notice that the farther away you hold it, the less it seems to shift. Your brain automatically performs such comparisons and gives you a pretty good sense of how far away things in your immediate neighbourhood are.

If your arms were made of rubber, you could stretch the pen far enough away from your eyes that the shift would become imperceptible. This is because our depth perception fails for objects more than a few tens of metres away. In order to see the shift of an object a city block or more from you, your eyes would need to be spread apart a lot farther.

Let's see how surveyors take advantage of the same idea. Suppose you are trying to measure the distance to a tree across a deep river as shown in Figure 9.3. You set up two observing stations some distance apart. That distance (line AB) is called the **baseline**. Now the direction to the tree (C in the figure) in relation to the baseline is observed from each station. Note that C appears in different directions from the two stations. This apparent change in direction of the remote object due to a change in vantage point of the observer is called **parallax**.

Triangulation

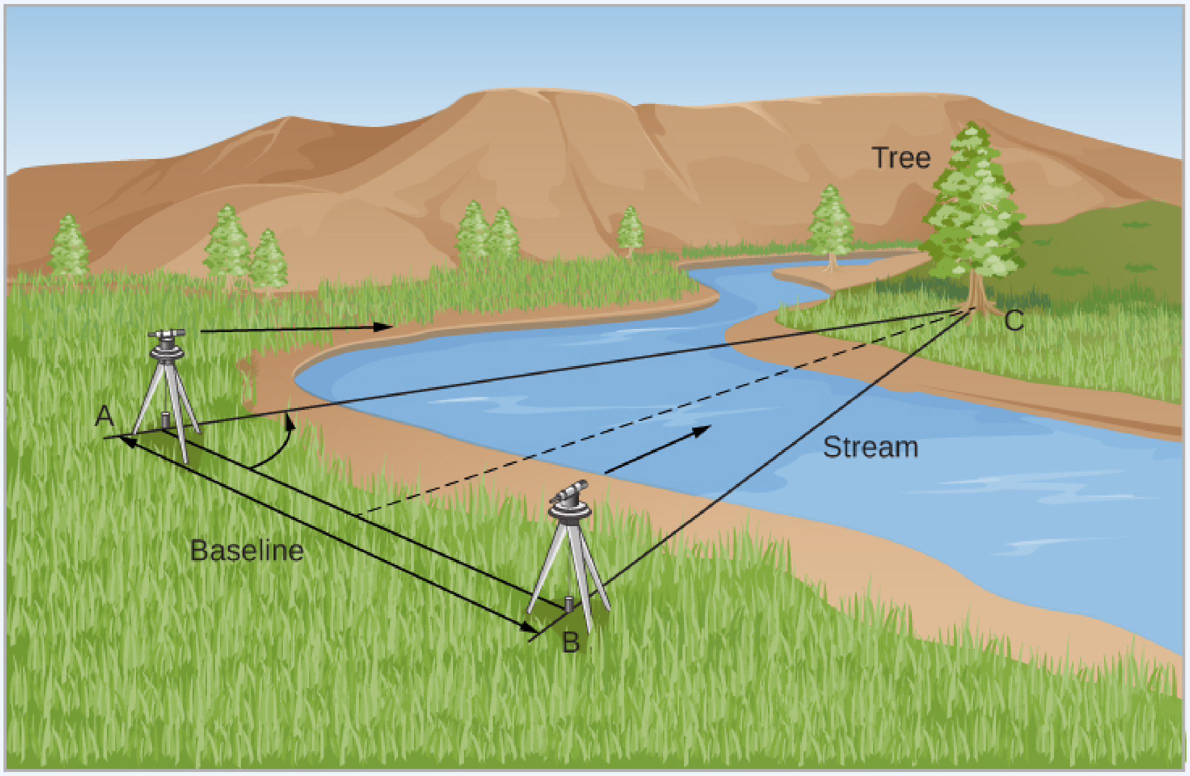


Figure 9.3. Triangulation allows us to measure distances to inaccessible objects. By getting the angle to a tree from two different vantage points, we can calculate the properties of the triangle they make and thus the distance to the tree.

The parallax is also the angle that lines AC and BC make—in mathematical terms, the angle subtended by the baseline. A knowledge of the angles at A and B and the length of the baseline, AB, allows the triangle ABC to be solved for any of its dimensions—say, the distance AC or BC. The solution could be reached by constructing a scale drawing or by using trigonometry to make a numerical calculation. If the tree were farther away, the whole triangle would be longer and skinnier, and the parallax angle would be smaller. Thus, we have the general rule that the smaller the parallax, the more distant the object we are measuring must be.

In practice, the kinds of baselines surveyors use for measuring distances on Earth are completely useless when we try to gauge distances in space. The farther away an astronomical object lies, the longer the baseline has to be to give us a reasonable chance of making a measurement. Unfortunately, nearly all astronomical objects are very far away. To measure their distances requires a very large baseline and highly precise angular measurements. The Moon is the only object near enough that its distance can be found fairly accurately with measurements made without a telescope. Ptolemy determined the distance to the Moon correctly to within a few percent. He used the turning Earth itself as a baseline, measuring the position of the Moon relative to the stars at two different times of night.

With the aid of telescopes, later astronomers were able to measure the distances to the nearer planets and asteroids using Earth's diameter as a baseline. This is how the AU was first established. To reach for the stars, however, requires a much longer baseline for triangulation and extremely sensitive measurements. Such a baseline is provided by Earth's annual trip around the Sun.

As Earth travels from one side of its orbit to the other, it graciously provides us with a baseline of 2 AU, or about 300 million kilometres. Although this is a much bigger baseline than the diameter of Earth, the stars are *so far away* that the resulting parallax shift is *still* not visible to the naked eye—not even for the closest stars.

The first successful detections of stellar parallax were in the year 1838, when Friedrich Bessel in Germany, pictured in Figure 9.4, Thomas Henderson, a Scottish astronomer working at the Cape of Good Hope, and Friedrich Struve in Russia independently measured the parallaxes of the stars 61 Cygni, Alpha Centauri, and Vega, respectively. Even the closest star, Alpha Centauri, showed a total displacement of only about 1.5 arcseconds during the course of a year.

Friedrich Wilhelm Bessel (1784–1846), Thomas J. Henderson (1798–1844), and Friedrich Struve (1793–1864).

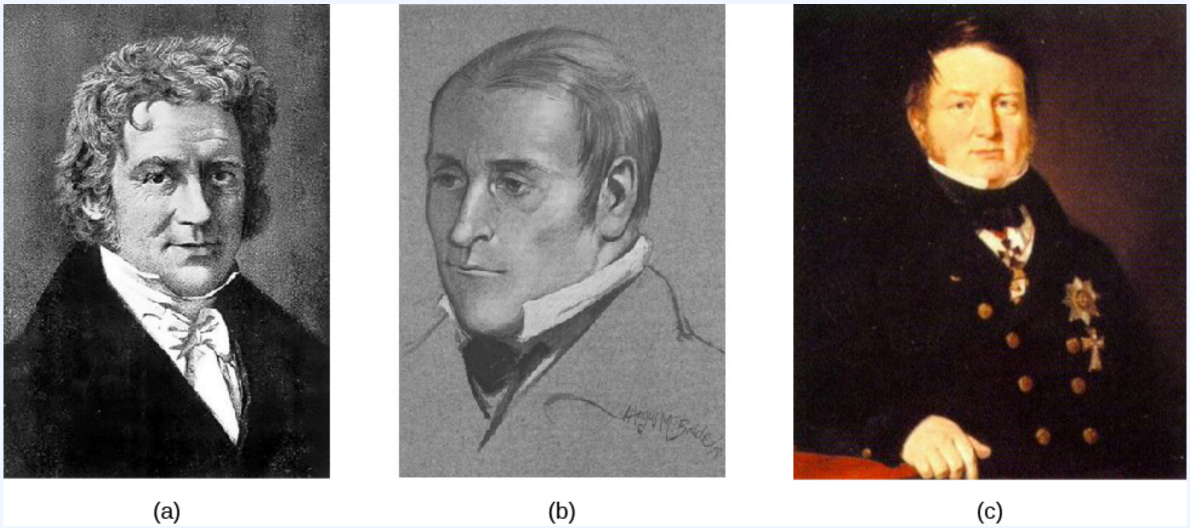


Figure 9.4**.** (a) Bessel made the first authenticated measurement of the distance to a star (61 Cygni) in 1838, a feat that had eluded many dedicated astronomers for almost a century. But two others, (b) Scottish astronomer Thomas J. Henderson and (c) Friedrich Struve, in Russia, were close on his heels.

Figure 9.5 shows how such measurements work. Seen from opposite sides of Earth's orbit, a nearby star shifts position when compared to a pattern of more distant stars. Astronomers actually define parallax to be *one-half* the angle that a star shifts when seen from opposite sides of Earth's orbit (the angle labelled *P* in Figure 9.5). The reason for this definition is just that they prefer to deal with a baseline of 1 AU instead of 2 AU.

Parallax

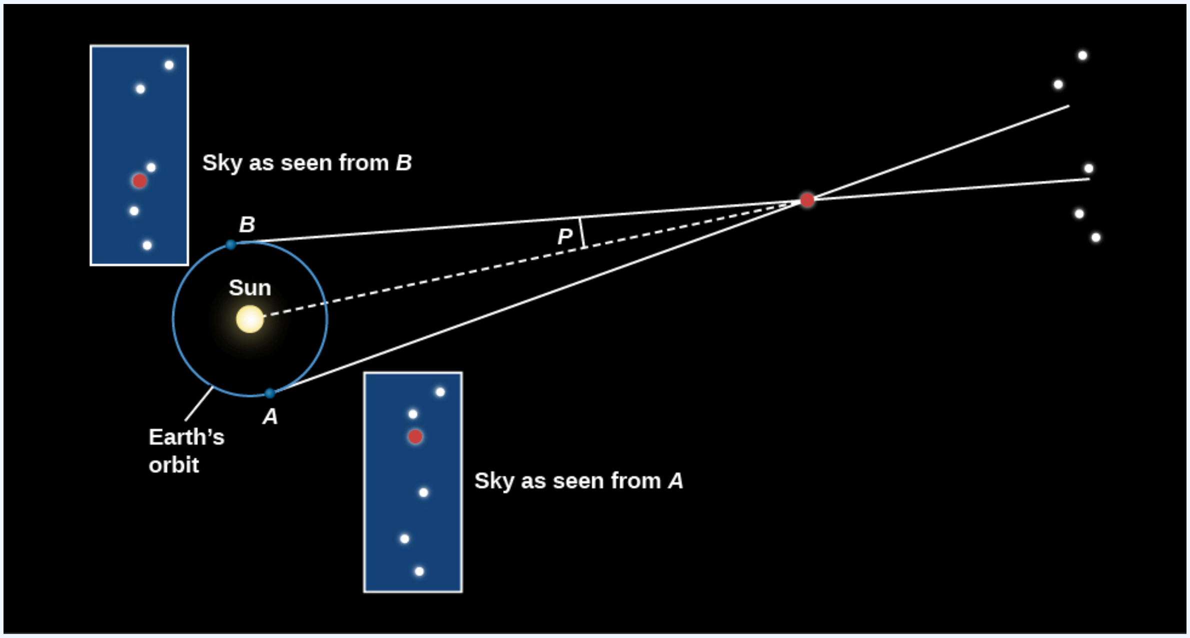


Figure 9.5. As Earth revolves around the Sun, the direction in which we see a nearby star varies with respect to distant stars. We define the parallax of the nearby star to be one half of the total change in direction, and we usually measure it in arcseconds.

With a baseline of one AU, how far away would a star have to be to have a parallax of 1 arcsecond? The answer turns out to be 206,265 AU, or 3.26 light-years. This is equal to 3.1×10^{13} kilometres (in other words, 31 trillion kilometres). We give this unit a special name, the **parsec (pc)**—derived from "the distance at which we have a *par*allax of one *sec*ond." The distance (*D*) of a star in parsecs is just the reciprocal of its parallax (*p*) in arcseconds; that is,

$$D = \frac{1}{p}$$

Thus, a star with a parallax of 0.1 arcsecond would be found at a distance of 10 parsecs, and one with a parallax of 0.05 arcsecond would be 20 parsecs away.

Back in the days when most of our distances came from parallax measurements, a parsec was a useful unit of distance, but it is not as intuitive as the light-year. One advantage of the light-year as a unit is that it emphasizes the fact that, as we look out into space, we are also looking back into time. The light that we see from a star 100 light-years away left that star 100 years ago. What we study is not the star as it is now, but rather as it was in the past. The light that reaches our telescopes today from distant galaxies left them before Earth even existed.

In this text, we will use light-years as our unit of distance, but many astronomers still use parsecs when they write technical papers or talk with each other at meetings. To convert between the two distance units, just bear in mind: 1 parsec = 3.26 light-year, and 1 light-year = 0.31 parsec.

Example 9.1

How Far Is a Light-Year?

A light-year is the distance light travels in 1 year. Given that light travels at a speed of 300,000 km/s, how many kilometres are there in a light-year?

Solution

We learned earlier that speed = distance/time. We can rearrange this equation so that distance = velocity × time. Now, we need to determine the number of seconds in a year.

There are approximately 365 days in 1 year. To determine the number of seconds, we must estimate the number of seconds in 1 day.

We can change units as follows (notice how the units of time cancel out):

$$1\text{ day } \times \ 24\text{ hr/day } \times \ 60\text{ min/hr } \times \ 60\text{ s/min} = 86,400\text{ s/day}$$

Next, to get the number of seconds per year:

$$365\text{ days/year } \times \ 86,400\text{ s/day} = 31,536,000\text{ s/year}$$

Now we can multiply the speed of light by the number of seconds per year to get the distance traveled by light in 1 year:

$$\begin{aligned}\text{distance} &= \text{velocity } \times \text{ time}\\ &= 300,000\text{ km/s } \times \ 31,536,000\text{ s}\\ &= 9.46 \ \times \ 10^{12}\text{ km}\end{aligned}$$

That's almost 10,000,000,000,000 km that light covers in a year. To help you imagine how long this distance is, we'll mention that a string 1 light-year long could fit around the circumference of Earth 236 million times.

Exercise 9.1

The solution for Example 9.1 is a really large number. What happens if we put it in terms that might be a little more understandable, like the diameter of Earth? Earth's diameter is about 12,700 km.

Solution

$$\begin{aligned} \text{1 light-year} &= 9.46 \times 10^{12}\ \text{km} \\ &= 9.46 \times 10^{12}\ \text{km} \times \frac{\text{1 Earth diameter}}{\text{12,700 km}} \\ &= 7.45 \times 10^{8}\ \text{Earth diametres} \end{aligned}$$

That means that 1 light-year is about 745 million times the diameter of Earth.

Attribution

"19.1 Fundamental Units of Distance" and "19.2 Surveying the Stars" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

9.3 THE SOLAR NEIGHBOURHOOD

No known star (other than the Sun) is within 1 light-year or even 1 parsec of Earth. The stellar neighbours nearest the Sun are three stars in the constellation of Centaurus. To the unaided eye, the brightest of these three stars is Alpha Centauri, which is only 30° from the south celestial pole and hence not visible from the mainland United States. Alpha Centauri itself is a binary star—two stars in mutual revolution—too close together to be distinguished without a telescope. These two stars are 4.4 light-years from us. Nearby is a third faint star, known as Proxima Centauri. Proxima, with a distance of 4.3 light-years, is slightly closer to us than the other two stars. If Proxima Centauri is part of a triple star system with the binary Alpha Centauri, as seems likely, then its orbital period may be longer than 500,000 years.

Table 9.1. The Nearest Stars, Brown Dwarfs, and White Dwarfs

Star	System	Discovery Name	Distance (light-year)	Spectral Type	Location: RA[1][1]	Location: Dec[2][2]	Luminosity (Sun = 1)
		Sun	—	G2 V	—	—	1
1	1	Proxima Centauri	4.2	M5.5 V	14 29	−62 40	5×10^{-5}
2	2	Alpha Centauri A	4.4	G2 V	14 39	−60 50	1.5
3		Alpha Centauri B	4.4	K2 IV	14 39	−60 50	0.5
4	3	Barnard's Star	6.0	M4 V	17 57	+04 42	4.4×10^{-4}
5	4	Wolf 359	7.8	M6 V	10 56	+07 00	2×10^{-5}
6	5	Lalande 21 185	8.3	M2 V	11 03	+35 58	5.7×10^{-3}
7	6	Sirius A	8.6	A1 V	06 45	−16 42	23.1
8		Sirius B	8.6	DA2[3]	06 45	−16 43	2.5×10^{-3}
9	7	Luyten 726-8 A	8.7	M5.5 V	01 39	−17 57	6×10^{-5}
10		Luyten 726-8 B (UV Ceti)	8.7	M6 V	01 39	−17 57	4×10^{-5}
11	8	Ross 154	9.7	M.05 V	18 49	−23 50	5×10^{-4}
12	9	Ross 248 (HH Andromedae)	10.3	M5.5 V	23 41	+44 10	1.0×10^{-4}
13	10	Epsilon Eridani	10.5	K2 V	03 32	−09 27	0.29
14	11	Lacaille 9352	10.7	M0.5 V	23 05	−35 51	0.011
15	12	Ross 128 (FI Virginis)	10.9	M4 V	11 47	+00 48	3.4×10^{-4}
16	13	Luyten 789-6 A (EZ Aquarii A)	11.3	M5 V	22 38	−15 17	5×10^{-5}
17		Luyten 789-6 B (EZ Aquarii B)	11.3	M5.5 V	22 38	−15 15	5×10^{-5}
18		Luyten 789-6 C (EZ Aquarii C)	11.3	M6.5 V	22 38	−15 17	2×10^{-5}
19	14	61 Cygni A	11.4	K5 V	21 06	+38 44	0.086
20		61 Cygni B	11.4	K7 V	21 06	+38 44	0.041
21	15	Procyon A	11.4	F51V	07 39	+05 13	7.38
22		Procyon B	11.4	wd[4]	07 39	+05 13	5.5×10^{-4}
23	16	Sigma 2398 A	11.5	M3 V	18 42	+59 37	0.003

Star	**System**	**Discovery Name**	**Distance (light-year)**	**Spectral Type**	**Location: RA**[1]	**Location: Dec**[2]	**Luminosity (Sun = 1)**
24		Sigma 2398 B	11.5	M3.5 V	18 42	+59 37	1.4×10^{-3}
25	17	Groombridge 34 A (GX Andromedae)	11.6	M1.5 V	00 18	+44 01	6.4×10^{-3}

Attribution

"19.2 Surveying the Stars" and "A.9 The Nearest Stars, Brown Dwarfs, and White Dwarfs" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

---

1. Location (right ascension) given for Epoch 2000.0
2. Location (declination) given for Epoch 2000.0
3. White dwarf stellar remnant
4. White dwarf stellar remnant

9.4 LUMINOSITY AND APPARENT BRIGHTNESS

Luminosity

Perhaps the most important characteristic of a star is its **luminosity**—the total amount of energy at all wavelengths that it emits per second. Earlier, we saw that the Sun puts out a tremendous amount of energy every second. (And there are stars far more luminous than the Sun out there.) To make the comparison among stars easy, astronomers express the luminosity of other stars in terms of the Sun's luminosity. For example, the luminosity of Sirius is about 25 times that of the Sun. We use the symbol L_{Sun} to denote the Sun's luminosity; hence, that of Sirius can be written as 25 L_{Sun}. In a later chapter, we will see that if we can measure how much energy a star emits and we also know its mass, then we can calculate how long it can continue to shine before it exhausts its nuclear energy and begins to die.

Apparent Brightness

Astronomers are careful to distinguish between the luminosity of the star (the total energy output) and the amount of energy that happens to reach our eyes or a telescope on Earth. Stars are democratic in how they produce radiation; they emit the same amount of energy in every direction in space. Consequently, only a minuscule fraction of the energy given off by a star actually reaches an observer on Earth. We call the amount of a star's energy that reaches a given area (say, one square meter) each second here on Earth its **apparent brightness**. If you look at the night sky, you see a wide range of apparent brightnesses among the stars. Most stars, in fact, are so dim that you need a telescope to detect them.

If all stars were the same luminosity—if they were like standard bulbs with the same light output—we could use the difference in their apparent brightnesses to tell us something we very much want to know: how far away they are. Imagine you are in a big concert hall or ballroom that is dark except for a few dozen 25-watt bulbs placed in fixtures around the walls. Since they are all 25-watt bulbs, their luminosity (energy output) is the same. But from where you are standing in one corner, they do *not* have the same apparent brightness. Those close to you appear brighter (more of their light reaches your eye), whereas those far away appear dimmer (their light has spread out more before reaching you). In this way, you can tell which bulbs are closest to you. In the same way, if all the stars had the same luminosity, we could immediately infer that the brightest-appearing stars were close by and the dimmest-appearing ones were far away.

We know exactly how light fades with increasing distance. The energy we receive is inversely proportional to the square of the distance. If, for example, we have two stars of the same luminosity and one is twice as far away as the other, it will look four times dimmer than the closer one. If it is three times farther away, it will look nine (three squared) times dimmer, and so forth.

Alas, the stars do not all have the same luminosity. (Actually, we are pretty glad about that because having many different types of stars makes the universe a much more interesting place.) But this means that if a star looks dim in the sky, we cannot tell whether it appears dim because it has a low luminosity but is relatively nearby, or because it has a high luminosity but is very far away. To measure the luminosities of stars, we must first compensate for the dimming effects of distance on light, and to do that, we must know how far away they are. Distance is among the most difficult of all astronomical measurements. We will return to how it is determined after we have learned more about the stars. For now, we will describe how astronomers specify the apparent brightness of stars.

Attribution

9.5 PLANETS OF OTHER SUNS

You might think that with the advanced telescopes and detectors astronomers have today, they could directly image planets around nearby stars (which we call **exoplanets**). This has proved extremely difficult, however, not only because the exoplanets are faint, but also because they are generally lost in the brilliant glare of the star they orbit. The detection techniques that work best are indirect: they observe the effects of the planet on the star it orbits, rather than seeing the planet itself.

The first technique that yielded many planet detections is very high-resolution stellar spectroscopy. The **Doppler effect** lets astronomers measure the star's **radial velocity**: that is, the speed of the star, toward us or away from us, relative to the observer. If there is a massive planet in orbit around the star, the gravity of the planet causes the star to wobble, changing its radial velocity by a small but detectable amount. The distance of the star does not matter, as long as it is bright enough for us to take very high quality spectra.

Measurements of the variation in the star's radial velocity as the planet goes around the star can tell us the mass and orbital period of the planet. If there are several planets present, their effects on the radial velocity can be disentangled, so the entire planetary system can be deciphered—as long as the planets are massive enough to produce a measureable Doppler effect. This detection technique is most sensitive to large planets orbiting close to the star, since these produce the greatest wobble in their stars. It has been used on large ground-based telescopes to detect hundreds of planets, including one around Proxima Centauri, the nearest star to the Sun.

The second indirect technique is based on the slight dimming of a star when one of its planets **transits**, or crosses over the face of the star, as seen from Earth. Astronomers do not see the planet, but only detect its presence from careful measurements of a change in the brightness of the star over long periods of time. If the slight dips in brightness repeat at regular intervals, we can determine the orbital period of the planet. From the amount of starlight obscured, we can measure the planet's size.

While some transits have been measured from Earth, large-scale application of this transit technique requires a telescope in space, above the atmosphere and its distortions of the star images. It has been most successfully applied from the NASA Kepler space observatory, which was built for the sole purpose of "staring" for 5 years at a single part of the sky, continuously monitoring the light from more than 150,000 stars. The primary goal of Kepler was to determine the frequency of occurrence of exoplanets of different sizes around different classes of stars. Like the Doppler technique, the transit observations favour discovery of large planets and short-period orbits.

Recent detection of exoplanets using both the Doppler and transit techniques has been incredibly successful. Within two decades, we went from no knowledge of other planetary systems to a catalog of *thousands* of exoplanets. Most of the exoplanets found so far are more massive than or larger in size than Earth.

It is not that Earth analogs do not exist. Rather, the shortage of small rocky planets is an observational bias: smaller planets are more difficult to detect.

Analyses of the data to correct for such biases or selection effects indicate that small planets (like the terrestrial planets in our system) are actually much more common than giant planets. Also relatively common are "**super Earths**," planets with two to ten times the mass of our planet as shown in Figure 9.6. We don't have any of these in our solar system, but nature seems to have no trouble making them elsewhere. Overall, the Kepler data suggest that approximately one quarter of stars have exoplanet systems, implying the existence of at least 50 billion planets in our Galaxy alone. For up to date information you can visit https://exoplanets.nasa.gov/keplerscience/ .

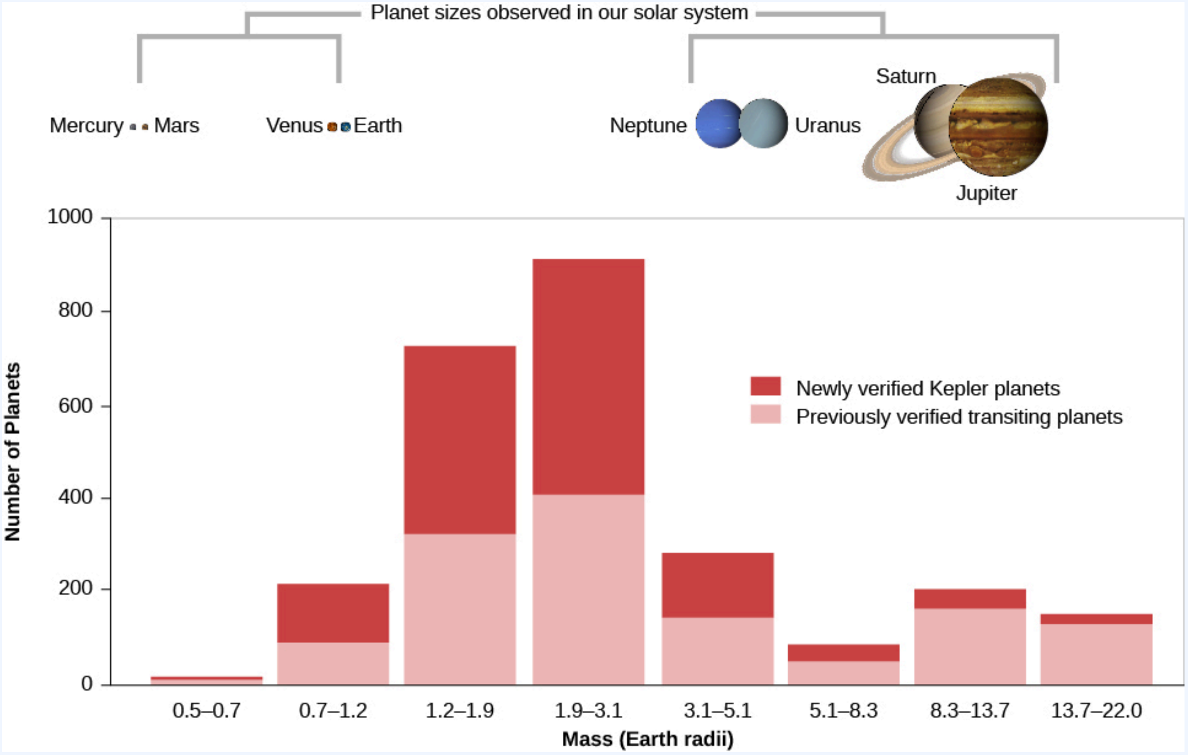

Transiting Planets by Size

Figure 9.6. This bar graph shows the planets found so far using the transit method (the vast majority found by the Kepler mission). The orange parts of each bar indicate the planets announced by the Kepler team in May 2016. Note that the largest number of planets found so far are in two categories that we don't have in our own solar system—planets whose size is between Earth's and Neptune's. Modification of image by NASA, NASA Media License.

Let's look more closely at the progress in the detection of exoplanets. Figure 9.7 shows the planets that were discovered each year by the two techniques we discussed. In the early years of exoplanet discovery, most of the planets were similar in mass to Jupiter. This is because, as mentioned above, the most massive planets were easiest to detect. In more recent years, planets smaller than Neptune and even close to the size of Earth have been detected.

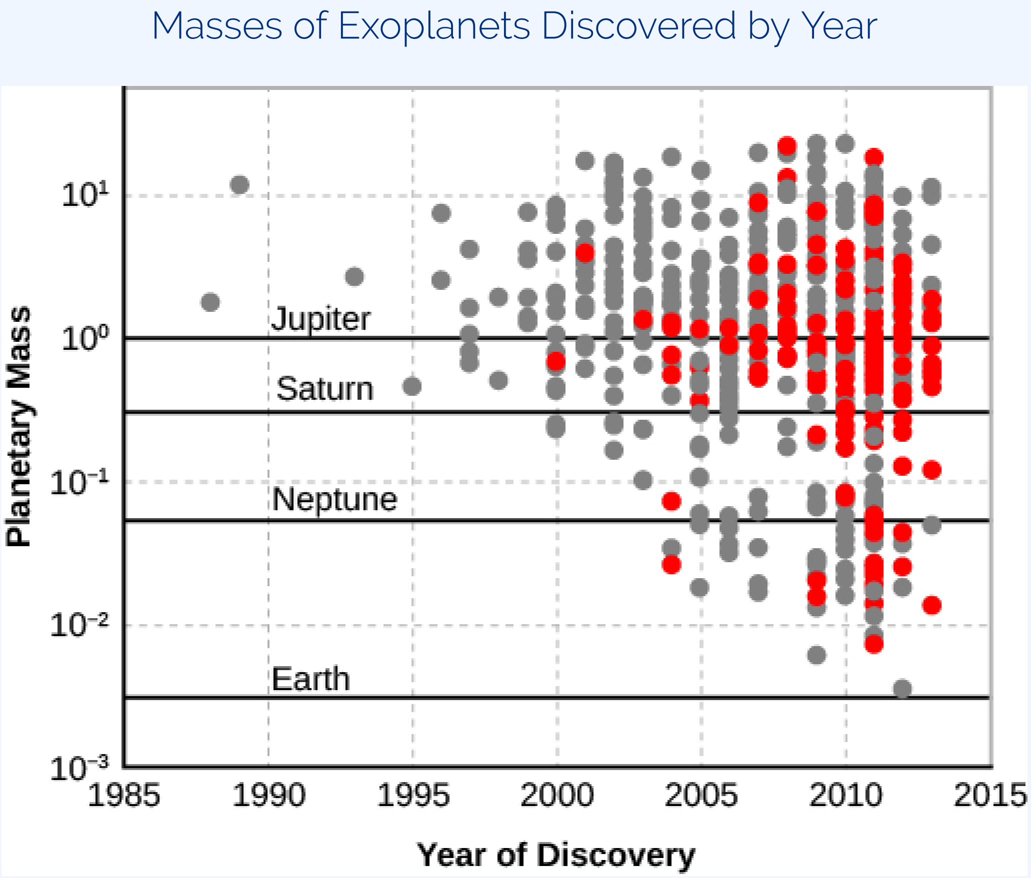


Figure 9.7. Horizontal lines are drawn to reference the masses of Jupiter, Saturn, Neptune, and Earth. The gray dots indicate planets discovered by measuring the radial velocity of the star, and the red dots are for planets that transit their stars. In the early years, the only planets that could be detected were similar in mass to Jupiter. Improvements in technology and observing strategies enabled the detection of lower mass planets as time went on, and now even smaller worlds are being found. (Note that this tally ends in 2014.)

We also know that many exoplanets are in multiplanet systems. This is one characteristic that our solar system shares with exosystems. Since large disks can give rise to more than one centre of condensation, it is not too surprising that multi-planet systems are a typical outcome of planet formation. Astronomers have tried to measure whether multiple planet systems all lie in the same plane using astrometry. This is a difficult measurement to make with current technology, but it is an important measurement that could help us understand the origin and evolution of planetary systems.

Many of the planetary systems discovered so far do not resemble our own solar system. Consequently, we have had to reassess some aspects of the "standard models" for the formation of planetary systems. Science sometimes works in this way, with new data contradicting our expectations. The press often talks about a scientist making experiments to "confirm" a theory. Indeed, it is comforting when new data support a hypothesis or theory and increase our confidence in an earlier result. But the most exciting and productive moments in science often come when new data *don't* support existing theories, forcing scientists to rethink their position and develop new and deeper insights into the way nature works.

Nothing about the new planetary systems contradicts the basic idea that planets form from the aggregation (clumping) of material within circumstellar disks. However, the existence of "hot Jupiters"—planets of jovian mass that are closer to their stars than the orbit of Mercury—poses the biggest problem. As far as we know, a giant planet cannot be formed without the condensation of water ice, and water ice is not stable so close to the heat of a star. It seems likely that all the giant planets, "hot" or "normal," formed at a distance of several astronomical units from the star, but we now see that they did not necessarily stay there. This discovery has led to a revision in our understanding of planet formation that now includes "planet migrations" within the protoplanetary disk, or later gravitational encounters between sibling planets that scatter one of the planets inward.

Many exoplanets have large orbital eccentricity (recall this means the orbits are not circular). High eccentricities were not expected for planets that form in a disk. This discovery provides further support for the scattering of planets when they interact gravitationally. When planets change each other's motions, their orbits could become much more eccentric than the ones with which they began.

There are several suggestions for ways migration might have occurred. Most involve interactions between the giant planets and the remnant material in the circumstellar disk from which they formed. These interactions would have taken place when the system was very young, while material still remained in the disk. In such cases, the planet travels at a faster velocity than the gas and dust and feels a kind of "headwind" (or friction) that causes it to lose energy and spiral inward. It is still unclear how the spiraling planet stops before it plunges into the star. Our best guess is that this plunge into the star is the fate for many protoplanets; however, clearly some migrating planets can stop their inward motions and escape this destruction, since we find hot Jupiters in many mature planetary systems.

The best possible evidence for an earthlike planet elsewhere would be an image. After all, "seeing is believing" is a very human prejudice. But imaging a distant planet is a formidable challenge indeed. Suppose, for example, you were a great distance away and wished to detect reflected light from Earth. Earth intercepts and reflects less than one billionth of the Sun's radiation, so its apparent brightness in visible light is less than one billionth that of the Sun. Compounding the challenge of detecting such a faint speck of light, the planet is swamped by the blaze of radiation from its parent star.

Even today, the best telescope mirrors' optics have slight imperfections that prevent the star's light from coming into focus in a completely sharp point.

Direct imaging works best for young gas giant planets that emit infrared light and reside at large separations from their host stars. Young giant planets emit more infrared light because they have more internal energy, stored from the process of planet formation. Even then, clever techniques must be employed to subtract out the light from the host star. In 2008, three such young planets were discovered orbiting HR 8799, a star in the constellation of Pegasus, shown in Figure 9.8. Two years later, a fourth planet was detected closer to the star. Additional planets may reside even closer to HR 8799, but if they exist, they are currently lost in the glare of the star.

Since then, a number of planets around other stars have been found using direct imaging. However, one challenge is to tell whether the objects we are seeing are indeed planets or if they are brown dwarfs (failed stars) in orbit around a star.

Exoplanets around HR 8799

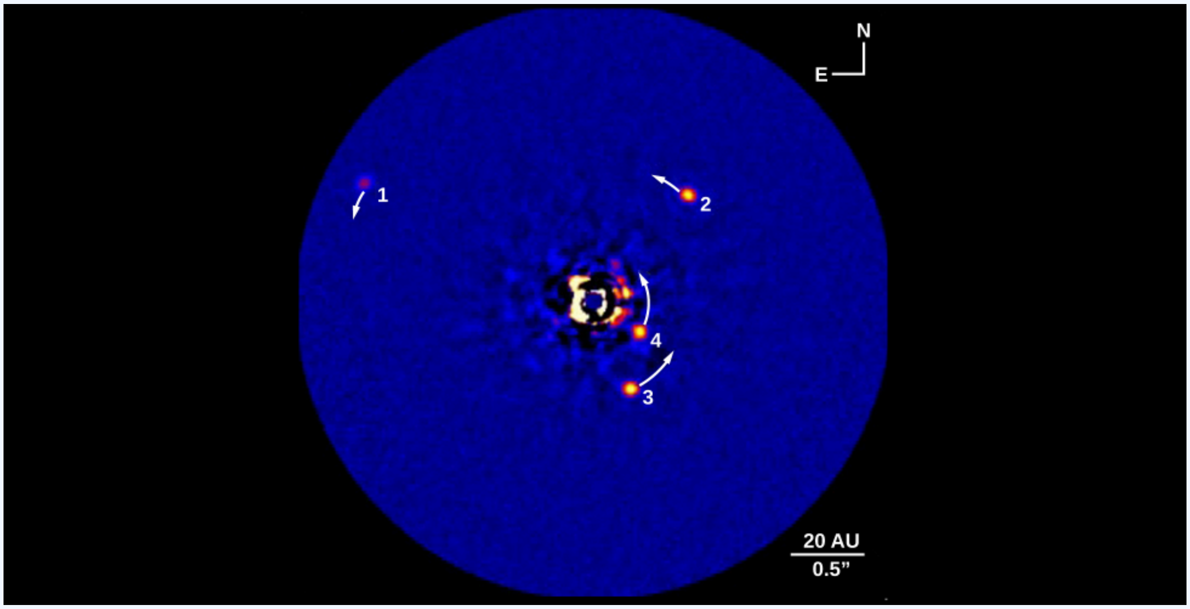


Figure 9.8. This image shows Keck telescope observations of four directly imaged planets orbiting HR 8799. A size scale for the system gives the distance in AU (remember that one astronomical unit is the distance between Earth and the Sun.) HR 8799 planetary system by NRC-HIA/C. MAROIS/W. M. KECK OBSERVATORY, used under fair dealing.

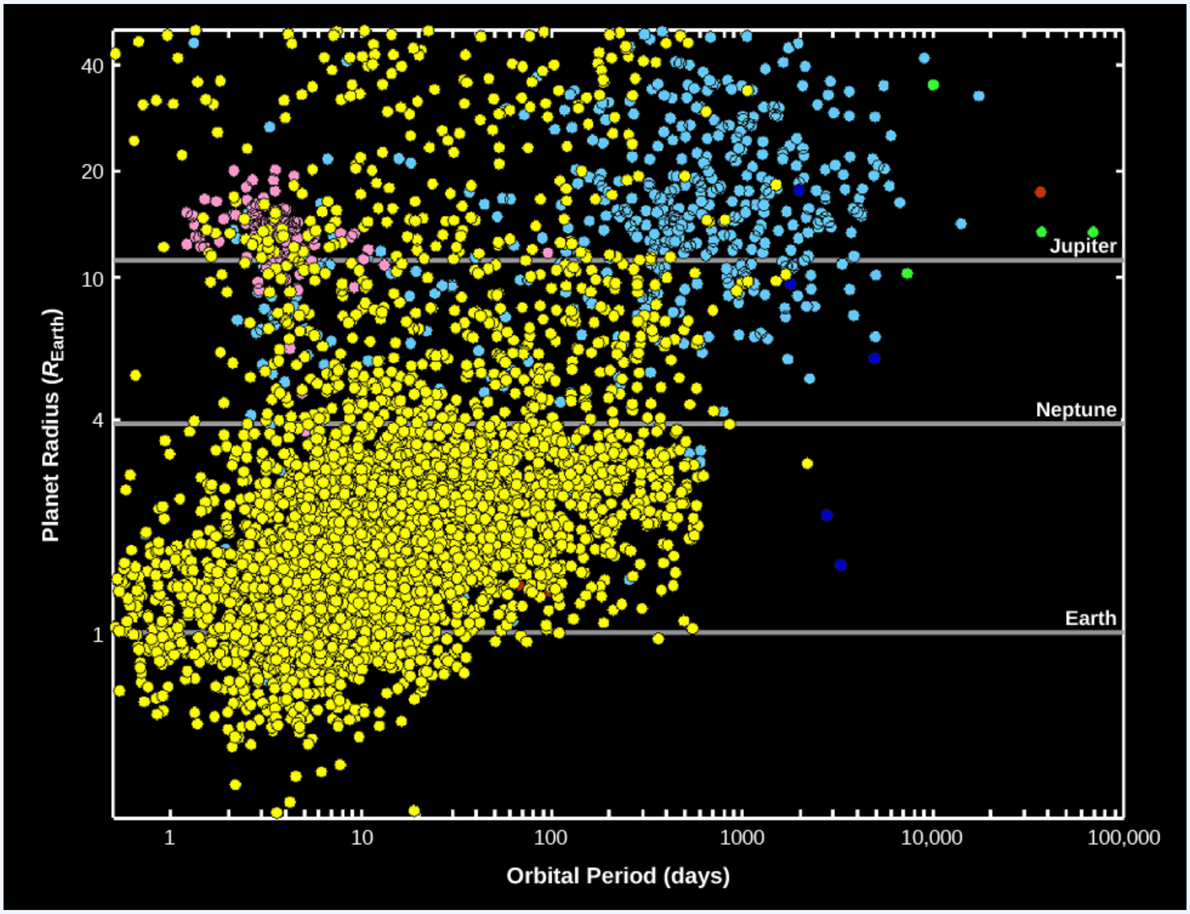
Direct imaging is an important technique for characterizing an exoplanet. The brightness of the planet can be measured at different wavelengths. These observations provide an estimate for the temperature of the planet's atmosphere; in the case of HR 8799 planet 1, the colour suggests the presence of thick clouds. Spectra can also be obtained from the faint light to analyze the atmospheric constituents. A spectrum of HR 8799 planet 1 indicates a hydrogen-rich atmosphere, while the closer planet 4 shows evidence for methane in the atmosphere.

Another way to overcome the blurring effect of Earth's atmosphere is to observe from space. Infrared may be the optimal wavelength range in which to observe because planets get brighter in the infrared while stars like our Sun get fainter, thereby making it easier to detect a planet against the glare of its star. Special optical techniques can be used to suppress the light from the central star and make it easier to see the planet itself. However, even if we go into space, it will be difficult to obtain images of Earth-size planets.

Before the discovery of exoplanets, most astronomers expected that other planetary systems would be much like our own—planets following roughly circular orbits, with the most massive planets several AU from their parent star. Such systems do exist in large numbers, but many exoplanets and planetary systems are very different from those in our solar system. Another surprise is the existence of whole classes of exoplanets that we simply don't have in our solar system: planets with masses between the mass of Earth and Neptune, and planets that are several times more massive than Jupiter.

Kepler Results

The Kepler telescope has been responsible for the discovery of most exoplanets, especially at smaller sizes, as illustrated in Figure 9.9, where the Kepler discoveries are plotted in yellow. You can see the wide range of sizes, including planets substantially larger than Jupiter and smaller than Earth. The absence of Kepler-discovered exoplanets with orbital periods longer than a few hundred days is a consequence of the 4-year lifetime of the mission. (Remember that three evenly spaced transits must be observed to register a discovery.) At the smaller sizes, the absence of planets much smaller than one earth radius is due to the difficulty of detecting transits by very small planets. In effect, the "discovery space" for Kepler was limited to planets with orbital periods less than 400 days and sizes larger than Mars.



Exoplanet Discoveries through 2015

Figure 9.9. The vertical axis shows the radius of each planet compared to Earth. Horizontal lines show the size of Earth, Neptune, and Jupiter. The horizontal axis shows the time each planet takes to make one orbit (and is given in Earth days). Recall that Mercury takes 88 days and Earth takes a little more than 365 days to orbit the Sun. The yellow and red dots show planets discovered by transits, and the blue dots are the discoveries by the radial velocity (Doppler) technique. Exoplanet Discoveries by NASA/Kepler mission, NASA Media License.

One of the primary objectives of the Kepler mission was to find out how many stars hosted planets and especially to estimate the frequency of earthlike planets. Although Kepler looked at only a very tiny fraction of the stars in the Galaxy, the sample size was large enough to draw some interesting conclusions. While the observations apply only to the stars observed by Kepler, those stars are reasonably representative, and so astronomers can extrapolate to the entire Galaxy.

Figure 9.10 shows that the Kepler discoveries include many rocky, Earth-size planets, far more than Jupiter-size gas planets. This immediately tells us that the initial Doppler discovery of many hot Jupiters was a biased sample, in effect, finding the odd planetary systems because they were the easiest to detect. However, there is one huge difference between this observed size distribution and that of planets in our solar system. The most common planets have radii between 1.4 and 2.8 that of Earth, sizes for which we have no examples in the solar system. These have been nicknamed super-Earths, while the other large group with sizes between 2.8 and 4 that of Earth are often called **mini-Neptunes**.

Kepler Discoveries

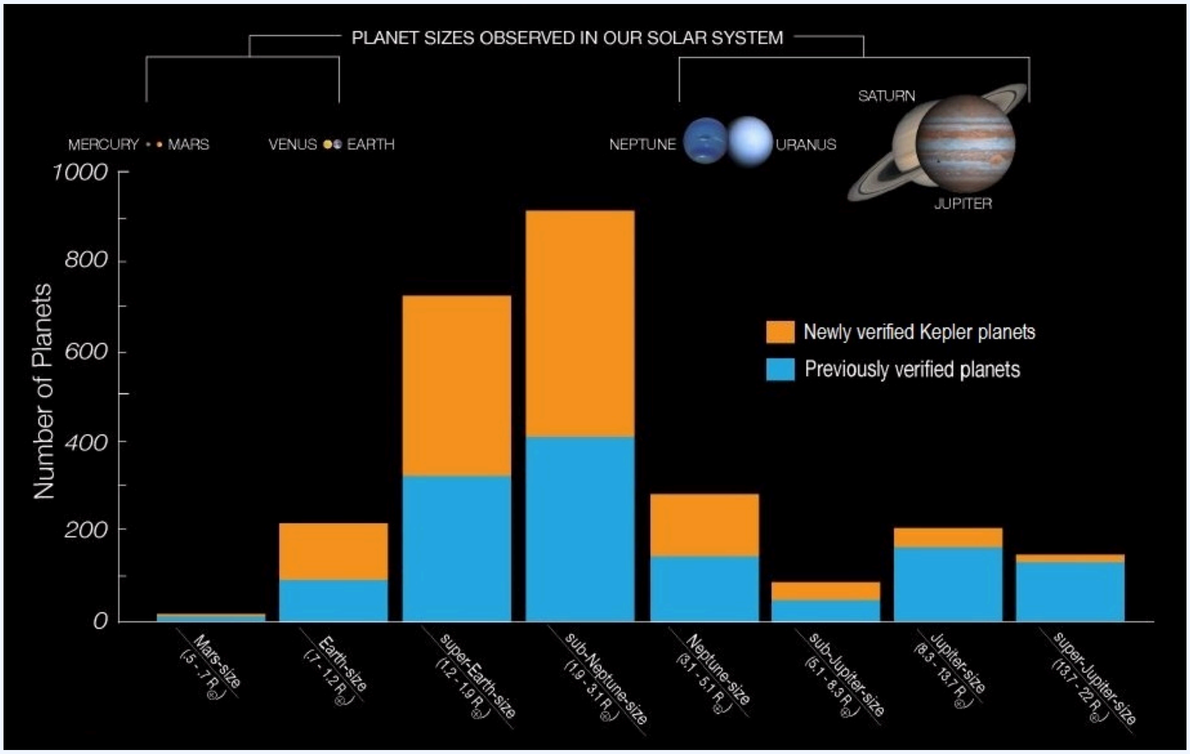


Figure 9.10. This bar graph shows the number of planets of each size range found among the first 2213 Kepler planet discoveries. Sizes range from half the size of Earth to 20 times that of Earth. On the vertical axis, you can see the fraction that each size range makes up of the total. Note that planets that are between 1.4 and 4 times the size of Earth make up the largest fractions, yet this size range is not represented among the planets in our solar system. Known Planets by Size by NASA/Kepler mission, NASA Media License.

What a remarkable discovery it is that the most common types of planets in the Galaxy are completely absent from our solar system and were unknown until Kepler's survey. However, recall that really small planets were difficult for the Kepler instruments to find. So, to estimate the frequency of Earth-size exoplanets, we need to correct for this sampling bias. The result is the corrected size distribution shown in Figure 9.11. Notice that in this graph, we have also taken the step of showing not the number of Kepler detections but the average number of planets per star for solar-type stars (spectral types F, G, and K).

Size Distribution of Planets for Stars Similar to the Sun

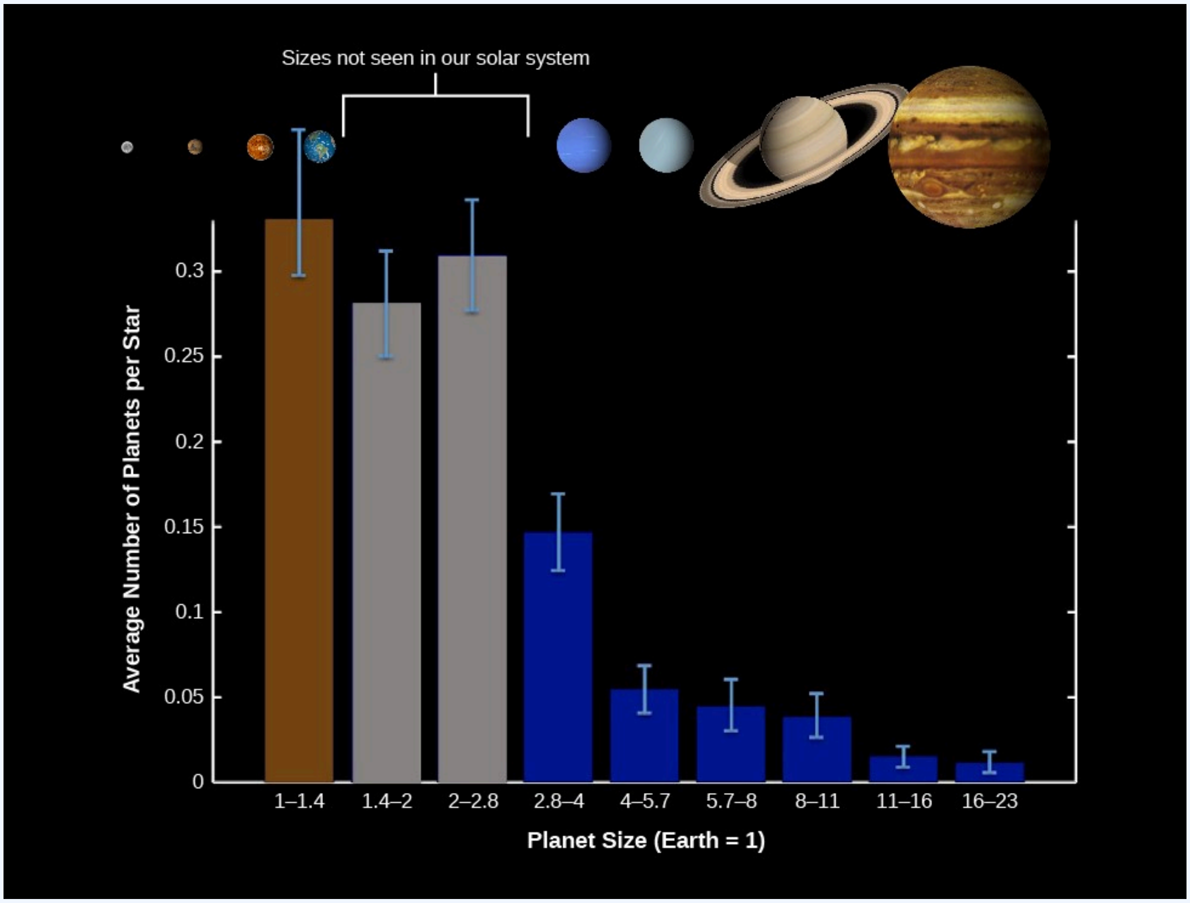


Figure 9.11. We show the average number of planets per star in each planet size range. (The average is less than one because some stars will have zero planets of that size range.) This distribution, corrected for biases in the Kepler data, shows that Earth-size planets may actually be the most common type of exoplanets.

We see that the most common planet sizes of are those with radii from 1 to 3 times that of Earth—what we have called "Earths" and "super-Earths." Each group occurs in about one-third to one-quarter of stars. In other words, if we group these sizes together, we can conclude there is nearly one such planet per star! And remember, this census includes primarily planets with orbital periods less than 2 years. We do not yet know how many undiscovered planets might exist at larger distances from their star.

To estimate the number of Earth-size planets in our Galaxy, we need to remember that there are approximately 100 billion stars of spectral types F, G, and K. Therefore, we estimate that there are about 30 billion Earth-size planets in our Galaxy. If we include the super-Earths too, then there could be one hundred billion in the whole Galaxy. This idea—that planets of roughly Earth's size are so numerous—is surely one of the most important discoveries of modern astronomy.

Attribution

"14.4 Comparison with Other Planetary Systems – Exoplanets", "21.4 Planets beyond the Solar System: Search and Discovery", and "21.5 Exoplanets Everywhere: What We Are Learning" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

9.6 BROWN DWARFS

A **star** is defined as an object that during some part of its lifetime derives 100% of its energy from the same process that makes the Sun shine — the fusion of hydrogen nuclei (protons) into helium. Objects with masses less than about 7.5% of the mass of our Sun (about 0.075 M_{Sun}) do not become hot enough for hydrogen fusion to take place. Even before the first such "failed star" was found, this class of objects, with masses intermediate between stars and planets, was given the name **brown dwarfs**.

Brown dwarfs are very difficult to observe because they are extremely faint and cool, and they put out most of their light in the infrared part of the spectrum. It was only after the construction of very large telescopes, like the Keck telescopes in Hawaii, and the development of very sensitive infrared detectors, that the search for brown dwarfs succeeded. The first brown dwarf was discovered in 1988, and, as of the summer of 2015, there are more than 2200 known brown dwarfs.

Most brown dwarfs start out with atmospheric temperatures and spectra like those of true stars, even though the brown dwarfs are not hot and dense enough in their interiors to fuse hydrogen. In fact, the spectra of brown dwarfs and true stars are so similar over a specific range of spectral types that it is not possible to distinguish the two types of objects based on spectra alone. An independent measure of mass is required to determine whether a specific object is a brown dwarf or a very low mass star. Since brown dwarfs cool steadily throughout their lifetimes, the spectral type of a given brown dwarf changes with time over a billion years or more.

An interesting property of brown dwarfs is that they are all about the same radius as Jupiter, regardless of their masses. Amazingly, this covers a range of masses from about 13 to 80 times the mass of Jupiter (M_{J}). This can make distinguishing a low-mass brown dwarf from a high-mass planet very difficult.

So, what is the difference between a low-mass brown dwarf and a high-mass planet? The International Astronomical Union considers the distinctive feature to be **deuterium fusion**. Although brown dwarfs do not sustain regular (proton-proton) hydrogen fusion, they are capable of fusing deuterium (a rare form of hydrogen with one proton and one neutron in its nucleus). The fusion of deuterium can happen at a lower temperature than the fusion of hydrogen. If an object has enough mass to fuse deuterium (about 13 M_{J} or 0.012 M_{Sun}), it is a brown dwarf. Objects with less than 13 M_{J} do not fuse deuterium and are usually considered planets.

Attribution

9.7 KEY TERMS

Absolute distances (of the planets): how far from the Sun one planet is compared to another. 9.2

Apparent brightness: the amount of a star's energy that reaches a given area (say, one square meter) each second here on Earth. 9.4

Baseline: the distance between the two different vantage points in a triangulation setup. 9.2

Brown dwarfs: objects with masses intermediate between stars and planets that do not become hot enough for hydrogen fusion to take place. 9.6

Deuterium fusion: a process that brown dwarfs are capable of completing which involves fusing deuterium (a rare form of hydrogen with one proton and one neutron in its nucleus) and which happens at a lower temperature than the fusion of hydrogen. 9.6

Doppler effect: the very high-resolution stellar spectroscopy technique that lets astronomers measure a star's radial velocity. 9.5

Exoplanets: planets of other nearby stars. 9.5

Globular clusters are groups of tens to hundreds of thousands of stars held tightly together by gravity. 9.1

Luminosity: the total amount of energy at all wavelengths that a star emits per second. 9.4

Mini-Neptune: a planet that is intermediate between the largest terrestrial planet in our solar system (Earth) and the smallest jovian planet (Neptune); generally, mini-Neptunes have sizes between 2.8 and 4 times Earth's size. 9.5

Open clusters are groups of up to a few thousand stars that are loosely held together by gravity. 9.1

Parallax: the apparent change in direction of the remote object due to a change in vantage point of the observer. 9.2

Parsec (pc): a unit of distance in astronomy, equal to 3.26 light-years; at a distance of 1 parsec, a star has a parallax of 1 arcsecond. 9.2

Radial velocity: the speed of a star, toward us or away from us, relative to the observer. 9.5

Relative distances: distances in light-seconds or metres or other standard units of length. 9.2

Star clusters: larger groups of hundreds or thousands of stars. 9.1

Star: an object that during some part of its lifetime derives 100% of its energy from the same process that makes the Sun shine — the fusion of hydrogen nuclei (protons) into helium. 9.6

Star systems: stars that are grouped tightly together. 9.1

Super Earth: a planet larger than Earth, generally between 1.4 and 2.8 times the size of our planet. 9.5

Transit: when one astronomical object moves in front of another. 9.5

Triangulation: a technique that allows us to measure distances to inaccessible objects by getting the angle to the object from two different vantage points where we can calculate the properties of the triangle they make and thus the distance to the object. 9.2

CHAPTER 10: STARS: ALL ABOUT SOL

Chapter Overview

10.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify the different types of star clusters.
- Explain the concept of parallax and how it is used to determine the distances to nearby stars.
- Define luminosity in the context of stars and explain how it is measured.
- Summarize the challenges astronomers face when measuring apparent brightness and determining distances to stars.
- Calculate the luminosity of a star given its apparent brightness and distance from Earth.
- Evaluate the impact of different types of stars on the detectability of exoplanets using direct imaging and transit methods.
- Identify the defining characteristics of a brown dwarf, including its energy source and temperature range, and explain why brown dwarfs are difficult to observe and why their detection required the development of specific observational tools and techniques.

10.1 INTRODUCTION

The Sun, like all stars, is an enormous ball of extremely hot, largely ionized gas, shining under its own power. And we do mean enormous. The Sun could fit 109 Earths side-by-side across its diameter, and it has enough volume (takes up enough space) to hold about 1.3 million Earths.

The Sun does not have a solid surface or continents like Earth, nor does it have a solid core as shown in Figure 10.1. However, it does have a lot of structure and can be discussed as a series of layers, not unlike an onion. In this section, we describe the huge changes that occur in the Sun's extensive interior and atmosphere, and the dynamic and violent eruptions that occur daily in its outer layers.

Earth and the Sun

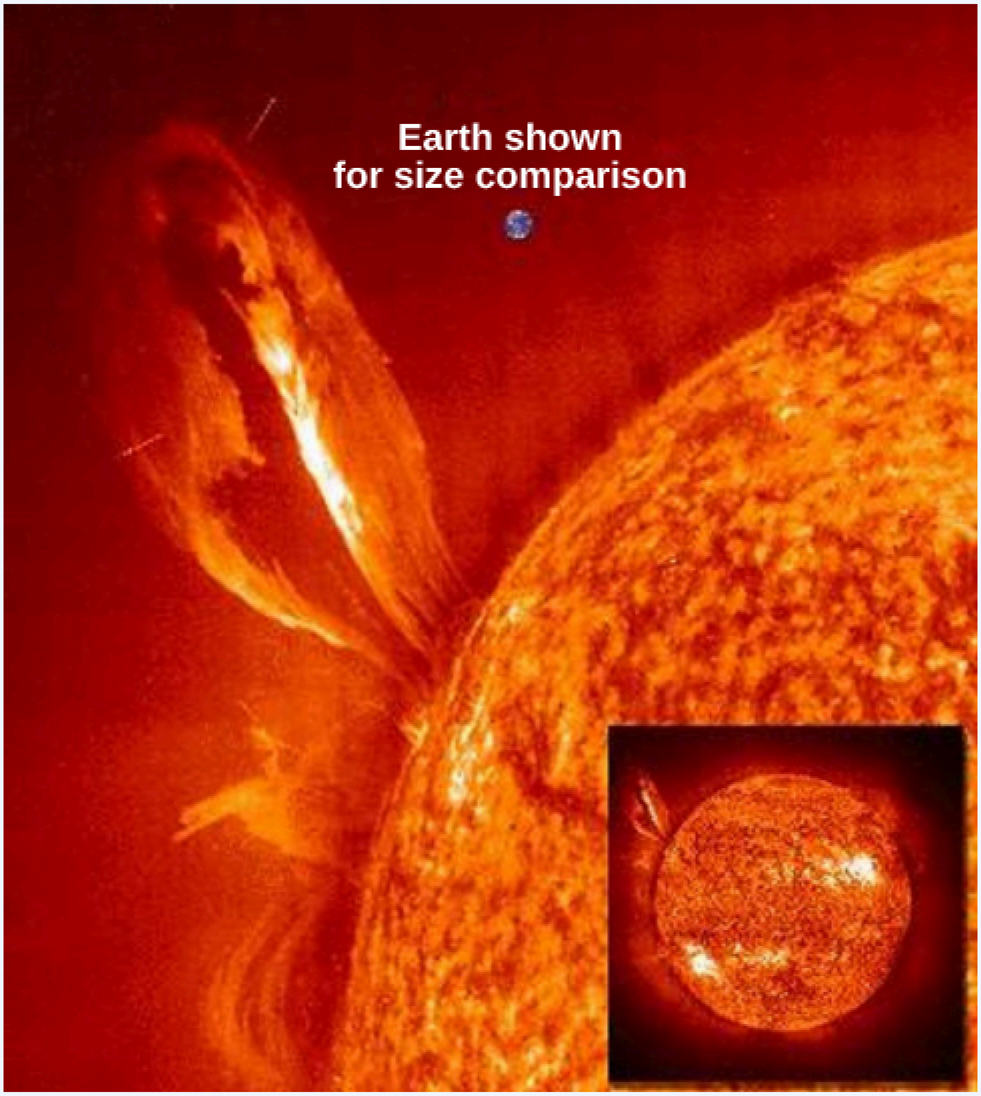


Figure 10.1. Here, Earth is shown to scale with part of the Sun and a giant loop of hot gas erupting from its surface. The inset shows the entire Sun, smaller.
Solar eruption larger than Earth by SOHO/EIT/ESA,ESA Standard License)

Some of the basic characteristics of the Sun are listed in Table 10.1. Although some of the terms in that table may be unfamiliar to you right now, you will get to know them as you read further.

Table 10.1. Characteristics of the Sun

Characteristic	**How Found**	**Value**
Mean distance	Radar reflection from planets	1 AU (149,597,892 km)
Maximum distance from Earth		1.521×10^8 km
Minimum distance from Earth		1.471×10^8 km
Mass	Orbit of Earth	333,400 Earth masses (1.99×10^{30} kg)
Diameter of photosphere	Angular size and distance	109.3 × Earth diameter (1.39×10^6 km)
Mean density	Mass/volume	1.41 g/cm^3 (1400 kg/m^3)
Gravitational acceleration at photosphere (surface gravity)	GM/R^2	27.9 × Earth surface gravity = 273 m/s^2
Solar constant	Instrument sensitive to radiation at all wavelengths	1370 W/m^2
Luminosity	Solar constant × area of spherical surface 1 AU in radius	3.8×10^{26} W
Spectral class	Spectrum	G2V
Effective temperature	Derived from luminosity and radius of the Sun	5800 K
Rotation period at equator	Sunspots and Doppler shift in spectra taken at the edge of the Sun	24 days 16 hours

---

Attribution

"15.1 The Structure and Composition of the Sun – and the connection to the Auroras" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

10.2 WHAT POWERS THE SUN

The Sun

Figure 10.2. It takes an incredible amount of energy for the Sun to shine, as it has and will continue to do for billions of years.
Sun and Fog by Ed Dunens, CC-BY-4.0

The Sun puts out an incomprehensible amount of energy—so much that its ultraviolet radiation can cause sunburns from 93 million miles away. It is also very old. As you learned earlier, evidence shows that the Sun formed about 4.5 billion years ago and has been shining ever since. How can the Sun produce so much energy for so long?

The Sun's energy output is about 4×10^{26} watts. This is unimaginably bright: brighter than a trillion cities together each with a trillion 100-watt light bulbs. Most known methods of generating energy fall far short of the capacity of the Sun. The total amount of energy produced over the entire life of the Sun is staggering, since the Sun has been shining for billions of years. Scientists were unable to explain the seemingly unlimited energy of stars like the Sun prior to the twentieth century.

When striving to understand how the Sun can put out so much energy for so long, scientists considered many different types of energy. Nineteenth-century scientists knew of two possible sources for the Sun's energy: chemical and gravitational energy. The source of chemical energy most familiar to them was the burning (the chemical term is *oxidation*) of wood, coal, gasoline, or other fuel. We know exactly how much energy the burning of these materials can produce. We can thus calculate that even if the immense mass of the Sun consisted of a burnable material like coal or wood, our star could not produce energy at its present rate for more than few thousand years. However, we know from geologic evidence that water was present on Earth's surface nearly 4 billion years ago, so the Sun must have been shining brightly (and making Earth warm) at least as long as that. Today, we also know that at the temperatures found in the Sun, nothing like solid wood or coal could survive.

It was only in the twentieth century that the true source of the Sun's energy was identified. The two key pieces of information required to solve the puzzle were the structure of the nucleus of the atom and the fact that mass can be converted into energy.

The Sun, then, taps the energy contained in the nuclei of atoms through nuclear fusion. Let's look at what happens in more detail. Deep inside the Sun, a three-step process takes four hydrogen nuclei and fuses them together to form a single helium nucleus. The helium nucleus is slightly less massive than the four hydrogen nuclei that combine to form it, and that mass is converted into energy.

The initial step required to form one helium nucleus from four hydrogen nuclei is shown in Figure 10.3. At the high temperatures inside the Sun's core, two protons combine to make a ***deuterium* nucleus**, which is an isotope (or version) of hydrogen that contains one proton and one neutron. In effect, one of the original protons has been converted into a neutron in the fusion reaction. Electric charge has to be conserved in nuclear reactions, and it is conserved in this one. A **positron** (antimatter electron) emerges from the reaction and carries away the positive charge originally associated with one of the protons.

Proton-Proton Chain, Step 1

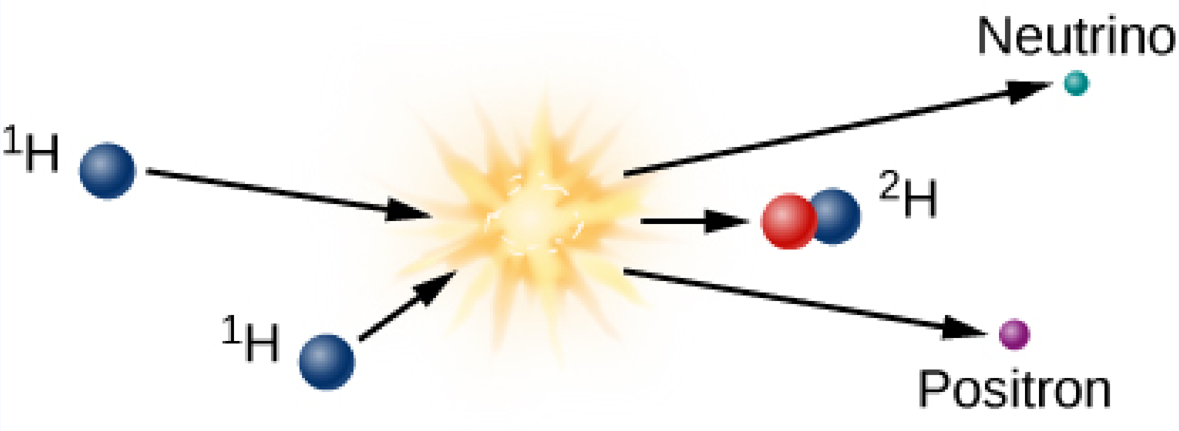


Figure 10.3. This is the first step in the process of fusing hydrogen into helium in the Sun. High temperatures are required because this reaction starts with two hydrogen nuclei, which are protons (shown in blue at left) that must overcome electrical repulsion to combine, forming a hydrogen nucleus with a proton and a neutron (shown in red). Note that hydrogen containing one proton and one neutron is given its own name: deuterium. Also produced in this reaction are a positron, which is an antielectron, and an elusive particle named the neutrino.

Since it is antimatter, this positron will instantly collide with a nearby electron, and both will be annihilated, producing electromagnetic energy in the form of gamma-ray photons. This gamma ray, which has been created in the centre of the Sun, finds itself in a world crammed full of fast-moving nuclei and electrons. The gamma ray collides with particles of matter and transfers its energy to one of them. The particle later emits another gamma-ray photon, but often the emitted photon has a bit less energy than the one that was absorbed.

Such interactions happen to gamma rays again and again and again as they make their way slowly toward the outer layers of the Sun, until their energy becomes so reduced that they are no longer gamma rays but X-rays. Later, as the photons lose still more energy through collisions in the crowded centre of the Sun, they become ultraviolet photons.

By the time they reach the Sun's surface, most of the photons have given up enough energy to be ordinary light—and they are the sunlight we see coming from our star. (To be precise, each gamma-ray photon is ultimately converted into many separate lower-energy photons of sunlight.) So, the sunlight given off by the Sun today had its origin as a gamma ray produced by nuclear reactions deep in the Sun's core. The length of time that photons require to reach the surface depends on how far a photon on average travels between collisions, and the travel time depends on what model of the complicated solar interior we accept. Estimates are somewhat uncertain but indicate that the emission of energy from the surface of the Sun can lag its production in the interior by 100,000 years to as much as 1,000,000 years.

In addition to the positron, the fusion of two hydrogen atoms to form deuterium results in the emission of a neutrino. Because neutrinos interact so little with ordinary matter, those produced by fusion reactions near the centre of the Sun travel directly to the Sun's surface and then out into space, in all directions. Neutrinos move at nearly the speed of light, and they escape the Sun about two seconds after they are created.

Proton-Proton Chain, Step 2

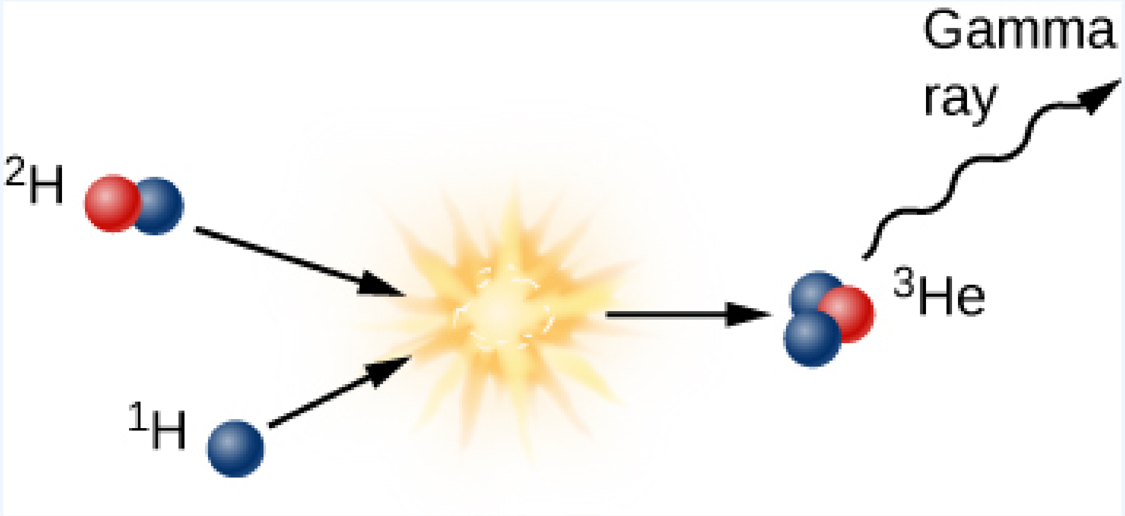


Figure 10.4. This is the second step of the proton-proton chain, the fusion reaction that converts hydrogen into helium in the Sun. This step combines one hydrogen nucleus, which is a proton (shown in blue), with the deuterium nucleus from the previous step (shown as a red and blue particle). The product of this is an isotope of helium with two protons (blue) and one neutron (red) and energy in the form of gamma-ray radiation.

The second step in forming helium from hydrogen is to add another proton to the deuterium nucleus to create a helium nucleus that contains two protons and one neutron as illustrated in Figure 10.4. In the process, some mass is again lost and more gamma radiation is emitted. Such a nucleus is helium because an element is defined by its number of protons; any nucleus with two protons is called helium. But this form of helium, which we call helium-3 (and write in shorthand as ^{3}He) is not the isotope we see in the Sun's atmosphere or on Earth. That helium has two neutrons and two protons and hence is called helium-4 (^{4}He).

To get to helium-4 in the Sun, helium-3 must combine with another helium-3 in the third step of fusion as illustrated in Figure 10.5. Note that two energetic protons are left over from this step; each of them comes out of the reaction ready to collide with other protons and to start step 1 in the chain of reactions all over again.

Proton-Proton Chain, Step 3

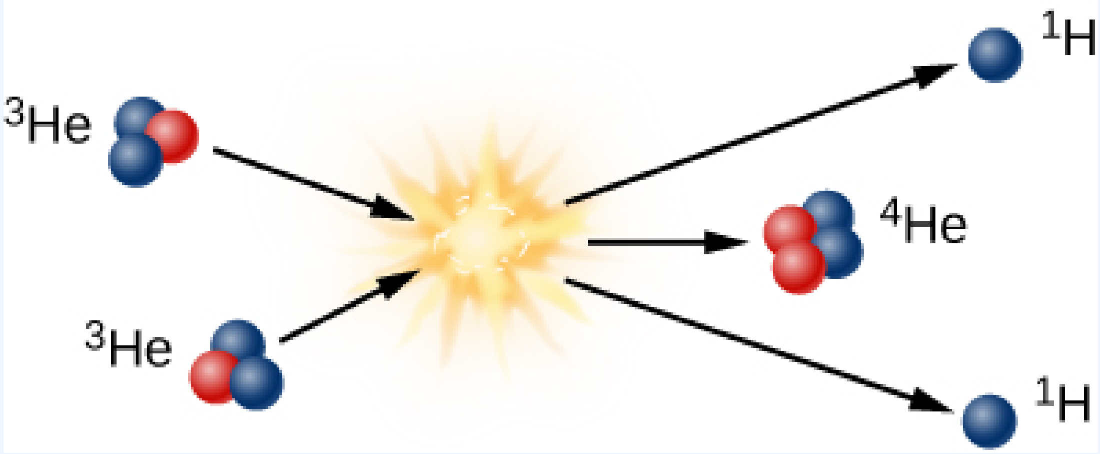


Figure 10.5. This is the third step in the fusion of hydrogen into helium in the Sun. Note that the two helium-3 nuclei from the second step must combine before the third step becomes possible. The two protons that come out of this step have the energy to collide with other protons in the Sun and start step one again.

Attribution

10.3 ANATOMY OF THE SUN

Let's begin by asking what the solar atmosphere is made of. We can use a star's *absorption line spectrum* to determine what elements are present. It turns out that the Sun contains the same elements as Earth but *not* in the same proportions. About 73% of the Sun's mass is hydrogen, and another 25% is helium. All the other chemical elements (including those we know and love in our own bodies, such as carbon, oxygen, and nitrogen) make up only 2% of our star. The 10 most abundant gases in the Sun's visible surface layer are listed in Table 10.2. Examine that table and notice that the composition of the Sun's outer layer is very different from Earth's crust, where we live. (In our planet's crust, the three most abundant elements are oxygen, silicon, and aluminum.) Although not like our planet's, the makeup of the Sun is quite typical of stars in general.

Table 10.2. The Abundance of Elements in the Sun

Element	Percentage by Number of Atoms	Percentage By Mass
Hydrogen	92.0	73.4
Helium	7.8	25.0
Carbon	0.02	0.20
Nitrogen	0.008	0.09
Oxygen	0.06	0.80
Neon	0.01	0.16
Magnesium	0.003	0.06
Silicon	0.004	0.09
Sulfur	0.002	0.05
Iron	0.003	0.14

The fact that our Sun and the stars all have similar compositions and are made up of mostly hydrogen and helium was first shown in a brilliant thesis in 1925 by Cecilia Payne-Gaposchkin, the first woman to get a PhD in astronomy in the United States, pictured in Figure 10.6. However, the idea that the simplest light gases—hydrogen and helium—were the most abundant elements in stars was so unexpected and so shocking that she assumed her analysis of the data must be wrong. At the time, she wrote, "The enormous abundance derived for these elements in the stellar atmosphere is almost certainly not real." Even scientists sometimes find it hard to accept new ideas that do not agree with what everyone "knows" to be right.

Cecilia Payne-Gaposchkin (1900–1979)

Figure 10.6. Her 1925 doctoral thesis laid the foundations for understanding the composition of the Sun and the stars. Yet, being a woman, she was not given a formal appointment at Harvard, where she worked, until 1938 and was not appointed a professor until 1956.
Cecilia Helena Payne Gaposchkin (1900-1979) by Smithsonian Institution Archives, Public Domain.

Before Payne-Gaposchkin's work, everyone assumed that the composition of the Sun and stars would be much like that of Earth. It was 3 years after her thesis that other studies proved beyond a doubt that the enormous abundance of hydrogen and helium in the Sun is indeed real. (And, as we will see, the composition of the Sun and the stars is much more typical of the makeup of the universe than the odd concentration of heavier elements that characterizes our planet.)

Most of the elements found in the Sun are in the form of atoms, with a small number of molecules, all in the form of gases: the Sun is so hot that no matter can survive as a liquid or a solid. In fact, the Sun is so hot that many of the atoms in it are **ionized**, that is, stripped of one or more of their electrons. This removal of electrons from their atoms means that there is a large quantity of free electrons and positively charged ions in the Sun, making it an electrically charged environment—quite different from the neutral one in which you are reading this text. (Scientists call such a hot ionized gas a plasma.)

In the nineteenth century, scientists observed a spectral line at 530.3 nanometres in the Sun's outer atmosphere, called the corona (a layer we will discuss in a minute.) This line had never been seen before, and so it was assumed that this line was the result of a new element found in the corona, quickly named coronium. It was not until 60 years later that astronomers discovered that this emission was in fact due to highly ionized iron—iron with 13 of its electrons stripped off. This is how we first discovered that the Sun's atmosphere had a temperature of more than a million degrees.

Figure 10.7 shows what the Sun would look like if we could see all parts of it from the centre to its outer atmosphere; the terms in the figure will become familiar to you as you read on.

Parts of the Sun

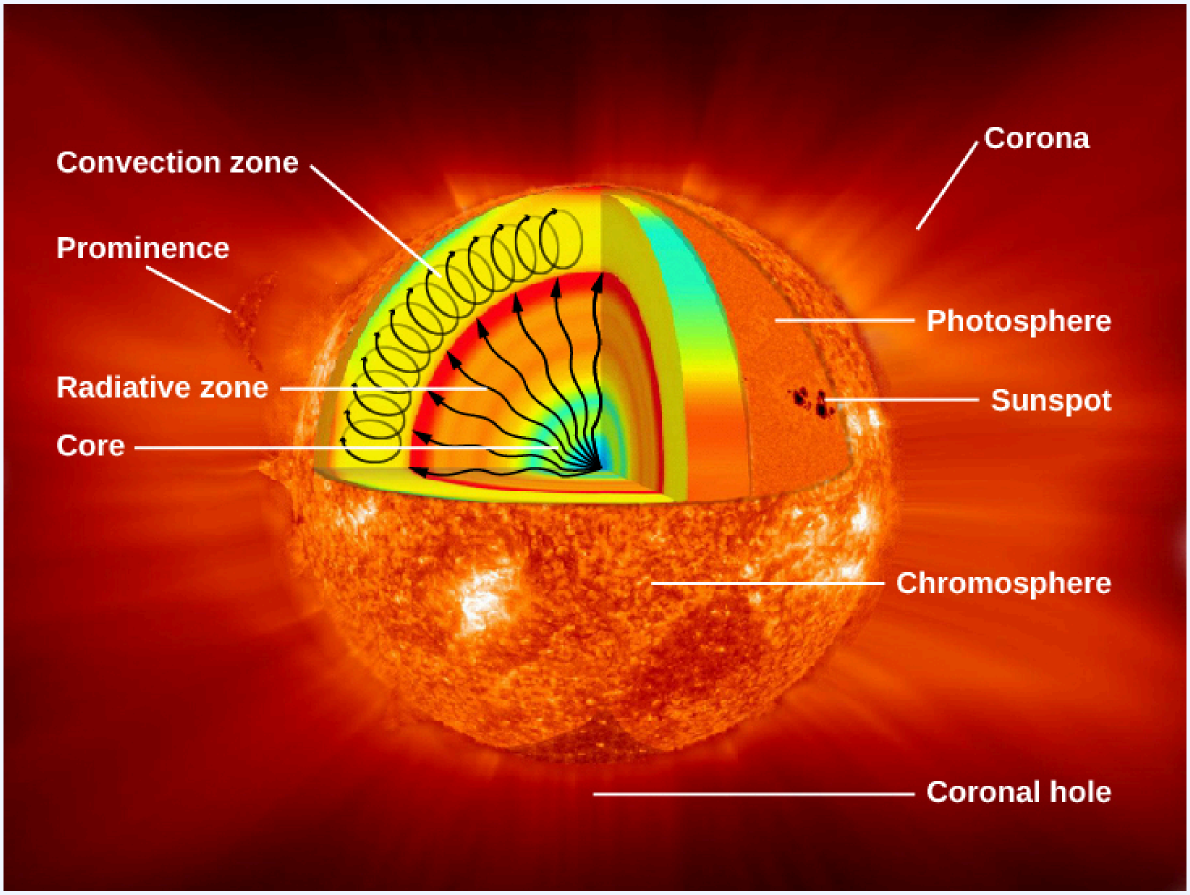


Figure 10.7. This illustration shows the different parts of the Sun, from the hot core where the energy is generated through regions where energy is transported outward, first by radiation, then by convection, and then out through the solar atmosphere. The parts of the atmosphere are also labelled the photosphere, chromosphere, and corona. Some typical features in the atmosphere are shown, such as coronal holes and prominences.
The Sun by NASA/Goddard, NASA Media License.

The Sun's layers are different from each other, and each plays a part in producing the energy that the Sun ultimately emits. We will begin with the core and work our way out through the layers. The Sun's **core** is extremely dense and is the source of all of its energy. Inside the core, nuclear energy is being released. The core is approximately 20% of the size of the solar interior and is thought to have a temperature of approximately 15 million K, making it the hottest part of the Sun.

Above the core is a region known as the **radiative zone**—named for the primary mode of transporting energy across it. This region starts at about 25% of the distance to the solar surface and extends up to about 70% of the way to the surface. The light generated in the core is transported through the radiative zone very slowly, since the high density of matter in this region means a photon cannot travel too far without encountering a particle, causing it to change direction and lose some energy.

The **convective zone** is the outermost layer of the solar interior. It is a thick layer approximately 200,000 kilometres deep that transports energy from the edge of the radiative zone to the surface through giant convection cells, similar to a pot of boiling oatmeal. The plasma at the bottom of the convective zone is extremely hot, and it bubbles to the surface where it loses its heat to space. Once the plasma cools, it sinks back to the bottom of the convective zone.

Now that we have given a quick overview of the structure of the whole Sun, in this section, we will embark on a journey through the visible layers of the Sun, beginning with the **photosphere**—the visible surface.

Earth's air is generally transparent. But on a smoggy day in many cities, it can become opaque, which prevents us from seeing through it past a certain point. Something similar happens in the Sun. Its outer atmosphere is transparent, allowing us to look a short distance through it. But when we try to look through the atmosphere deeper into the Sun, our view is blocked. The photosphere is the layer where the Sun becomes opaque and marks the boundary past which we cannot see.

Solar Photosphere plus Sunspots

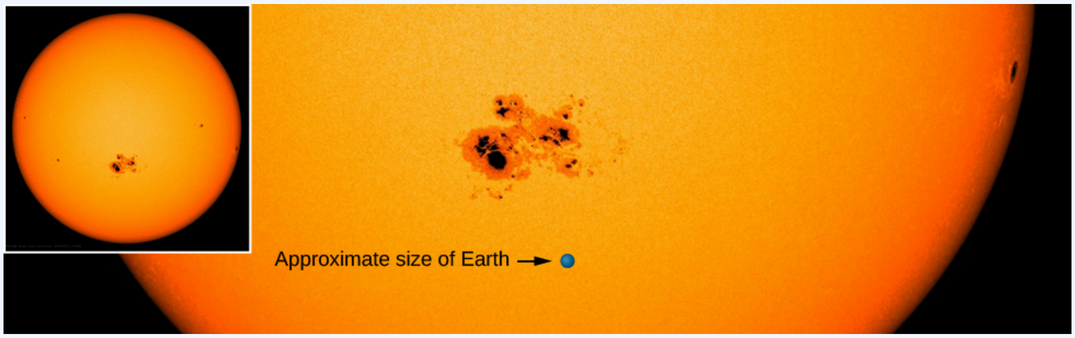


Figure 10.8. This photograph shows the photosphere—the visible surface of the Sun. Also shown is an enlarged image of a group of sunspots; the size of Earth is shown for comparison. Sunspots appear darker because they are cooler than their surroundings. The typical temperature at the centre of a large sunspot is about 3800 K, whereas the photosphere has a temperature of about 5800 K.
Modification of work by NASA/SDO, NASA Media License

NASA put together a static image and a video of the Sun's atmosphere in different wavelengths. Static image can be found at https://www.nasa.gov/content/goddard/sdo/wheel-of-colourful-sun-sdo5/

As we saw, the energy that emerges from the photosphere was originally generated deep inside the Sun. This energy is in the form of photons, which make their way slowly toward the solar surface. Outside the Sun, we can observe *only* those photons that are emitted into the solar photosphere, where the density of atoms is sufficiently low and the photons can finally escape from the Sun without colliding with another atom or ion.

As an analogy, imagine that you are attending a big campus rally and have found a prime spot near the centre of the action. Your friend arrives late and calls you on your cell phone to ask you to join her at the edge of the crowd. You decide that friendship is worth more than a prime spot, and so you work your way out through the dense crowd to meet her. You can move only a short distance before bumping into someone, changing direction, and trying again, making your way slowly to the outside edge of the crowd. All this while, your efforts are not visible to your waiting friend at the edge. Your friend can't see you until you get very close to the edge because of all the bodies in the way. So too photons making their way through the Sun are constantly bumping into atoms, changing direction, working their way slowly outward, and becoming visible only when they reach the atmosphere of the Sun where the density of atoms is too low to block their outward progress.

Astronomers have found that the solar atmosphere changes from almost perfectly transparent to almost completely opaque in a distance of just over 400 kilometres; it is this thin region that we call the *photosphere*, a word that comes from the Greek for "light sphere." When astronomers speak of the "diameter" of the Sun, they mean the size of the region surrounded by the photosphere.

The photosphere looks sharp only from a distance. If you were falling into the Sun, you would not feel any surface but would just sense a gradual increase in the density of the gas surrounding you. It is much the same as falling through a cloud while skydiving. From far away, the cloud looks as if it has a sharp surface, but you do not feel a surface as you fall into it. (One big difference between these two scenarios, however, is temperature. The Sun is so hot that you would be vaporized long before you reached the photosphere. Skydiving in Earth's atmosphere is much safer.)

We might note that the atmosphere of the Sun is not a very dense layer compared to the air in the room where you are reading this text. At a typical point in the photosphere, the pressure is less than 10% of Earth's pressure at sea level, and the density is about one ten-thousandth of Earth's atmospheric density at sea level.

Observations with telescopes show that the photosphere has a mottled appearance, resembling grains of rice spilled on a dark tablecloth or a pot of boiling oatmeal. This structure of the photosphere is called **granulation**, it is shown in Figure 10.9. Granules, which are typically 700 to 1000 kilometres in diameter (about the width of Texas), appear as bright areas surrounded by narrow, darker (cooler) regions. The lifetime of an individual granule is only 5 to 10 minutes. Even larger are supergranules, which are about 35,000 kilometres across (about the size of two Earths) and last about 24 hours.

Granulation Pattern

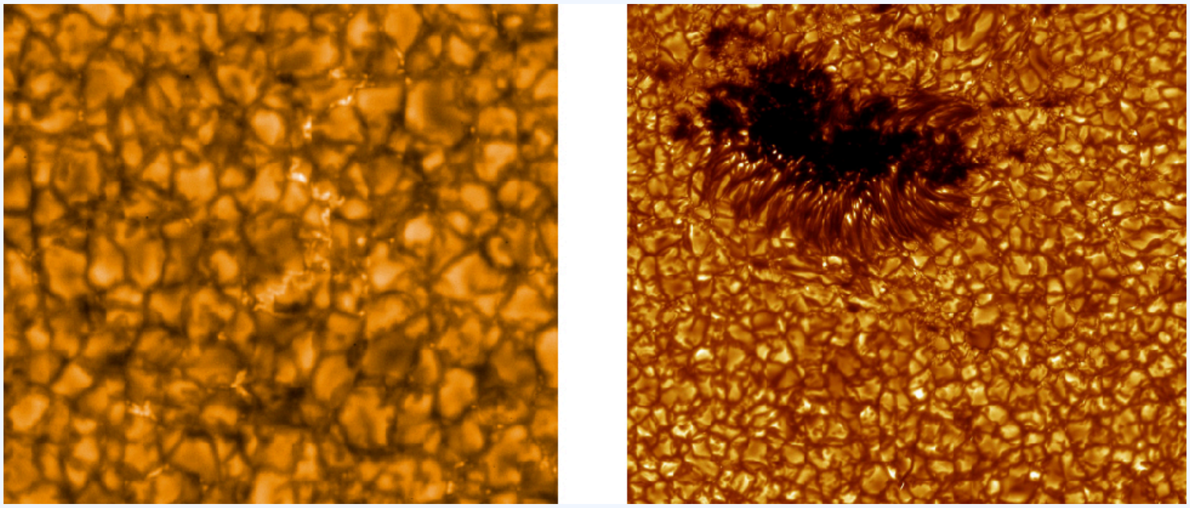


Figure 10.9. The surface markings of the convection cells create a granulation pattern on this dramatic image (left) taken from the Japanese Hinode spacecraft. You can see the same pattern when you heat up miso soup. The right image shows an irregular-shaped sunspot and granules on the Sun's surface, seen with the Swedish Solar Telescope on August 22, 2003.
Credit left: Hinode Views the Sun's Surface by Hinode JAXA/NASA/PPARC , NASA Media License
Credit right: ISP/SST/Oddbjorn Engvold, Jun Elin Wiik, Luc Rouppe van der Voort, Public Domain (Replace with alternative we can credit properly)

The motions of the granules can be studied by examining the Doppler shifts in the spectra of gases just above them. The bright granules are columns of hotter gases rising at speeds of 2 to 3 kilometres per second from below the photosphere. As this rising gas reaches the photosphere, it spreads out, cools, and sinks down again into the darker regions between the granules. Measurements show that the centres of the granules are hotter than the intergranular regions by 50 to 100 K.

The Sun's outer gases extend far beyond the photosphere as shown in Figure 10.10. Because they are transparent to most visible radiation and emit only a small amount of light, these outer layers are difficult to observe. The region of the Sun's atmosphere that lies immediately above the photosphere is called the **chromosphere**. Until this century, the chromosphere was visible only when the photosphere was concealed by the Moon during a total solar eclipse. In the seventeenth century, several observers described what appeared to them as a narrow red "streak" or "fringe" around the edge of the Moon during a brief instant after the Sun's photosphere had been covered. The name *chromosphere*, from the Greek for "coloured sphere," was given to this red streak.

The Sun's Atmosphere

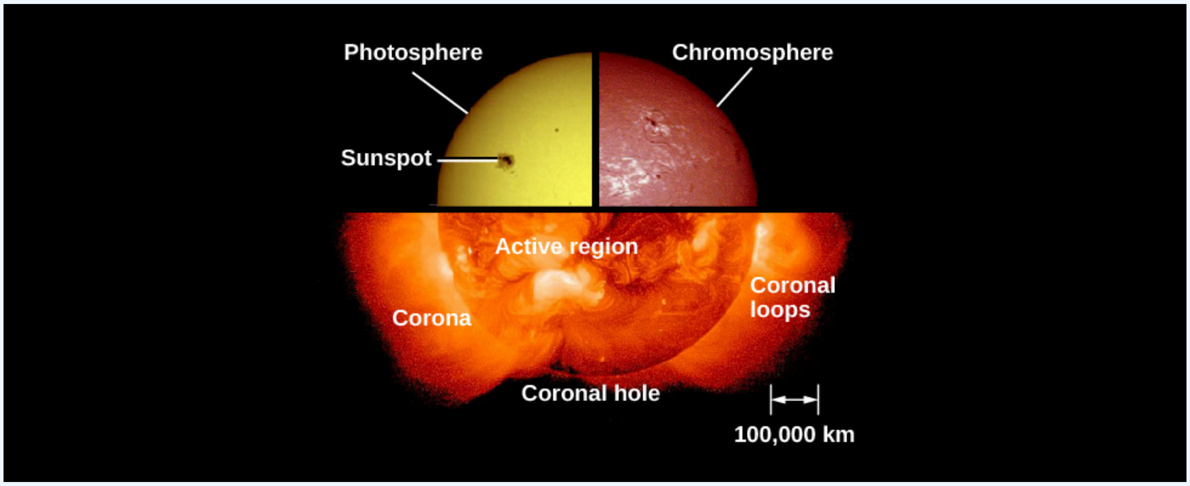


Figure 10.10. Composite image showing the three components of the solar atmosphere: the photosphere or surface of the Sun taken in ordinary light; the chromosphere, imaged in the light of the strong red spectral line of hydrogen (H-alpha); and the corona as seen with X-rays.

Observations made during eclipses show that the chromosphere is about 2000 to 3000 kilometres thick, and its spectrum consists of bright emission lines, indicating that this layer is composed of hot gases emitting light at discrete wavelengths. The reddish colour of the chromosphere arises from one of the strongest emission lines in the visible part of its spectrum—the bright red line caused by hydrogen, the element that, as we have already seen, dominates the composition of the Sun.

In 1868, observations of the chromospheric spectrum revealed a yellow emission line that did not correspond to any previously known element on Earth. Scientists quickly realized they had found a new element and named it *helium* (after *helios*, the Greek word for "Sun"). It took until 1895 for helium to be discovered on our planet. Today, students are probably most familiar with it as the light gas used to inflate balloons, although it turns out to be the second-most abundant element in the universe.

The temperature of the chromosphere is about 10,000 K. This means that the chromosphere is hotter than the photosphere, which should seem surprising. In all the situations we are familiar with, temperatures fall as one moves away from the source of heat, and the chromosphere is farther from the centre of the Sun than the photosphere is.

The outermost part of the Sun's atmosphere is called the **corona**. Like the chromosphere, the corona was first observed during total eclipses. Unlike the chromosphere, the corona has been known for many centuries: it was referred to by the Roman historian Plutarch and was discussed in some detail by Kepler.

The corona extends millions of kilometres above the photosphere and emits about half as much light as the full moon. The reason we don't see this light until an eclipse occurs is the overpowering brilliance of the photosphere. Just as bright city lights make it difficult to see faint starlight, so too does the intense light from the photosphere hide the faint light from the corona. While the best time to see the corona from Earth is during a total solar eclipse, it can be observed easily from orbiting spacecraft. Its brighter parts can now be photographed with a special instrument—a coronagraph—that removes the Sun's glare from the image with an occulting disk (a circular piece of material held so it is just in front of the Sun).

Coronagraph

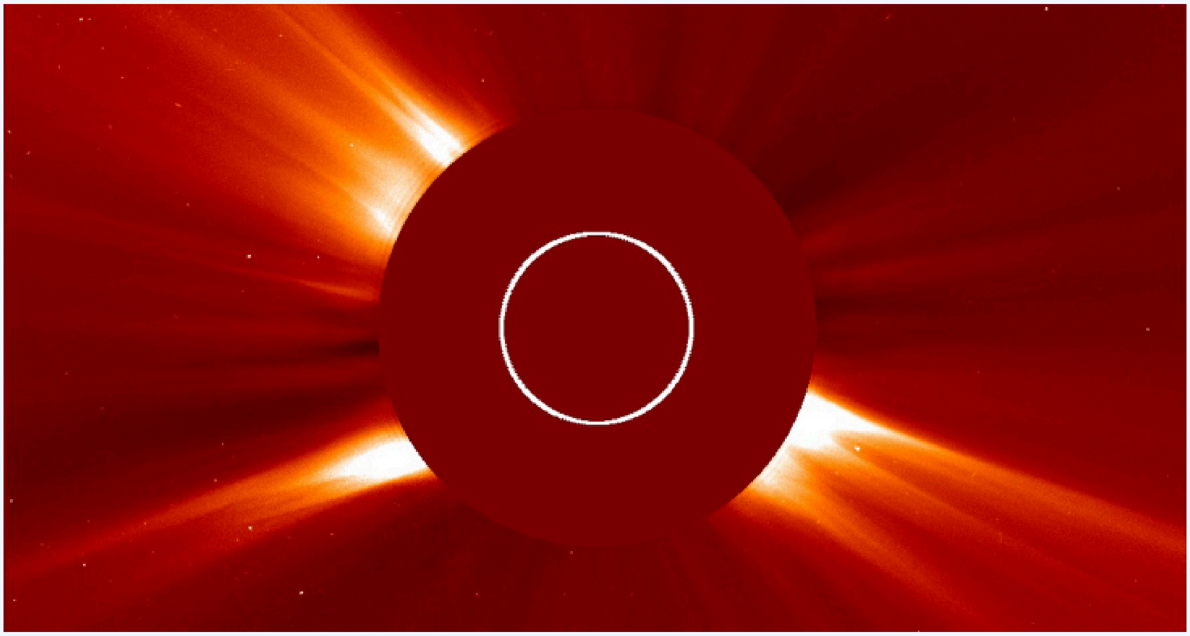


Figure 10.11. This image of the Sun was taken March 2, 2016. The larger dark circle in the center is the disk the blocks the Sun's glare, allowing us to see the corona. The smaller inner circle is where the Sun would be if it were visible in this image.
A moment in the sun's atmosphere by NASA/SOHO, NASA Media License

Studies of its spectrum show the corona to be very low in density. At the bottom of the corona, there are only about 10^9 atoms per cubic centimetre, compared with about 10^{16} atoms per cubic centimetre in the upper photosphere and 10^{19} molecules per cubic centimetre at sea level in Earth's atmosphere. The corona thins out very rapidly at greater heights, where it corresponds to a high vacuum by Earth laboratory standards. The corona extends so far into space—far past Earth—that here on our planet, we are technically living in the Sun's atmosphere.

One of the most remarkable discoveries about the Sun's atmosphere is that it produces a stream of charged particles (mainly protons and electrons) that we call the **solar wind**. These particles flow outward from the Sun into the solar system at a speed of about 400 kilometres per second (almost 1 million miles per hour)! The solar wind exists because the gases in the corona are so hot and moving so rapidly that they cannot be held back by solar gravity. (This wind was actually discovered by its effects on the charged tails of comets; in a sense, we can see the comet tails blow in the solar breeze the way wind socks at an airport or curtains in an open window flutter on Earth.)

Although the solar wind material is very, very rarified (i.e., *extremely* low density), the Sun has an enormous surface area. Astronomers estimate that the Sun is losing about 10 million tons of material each year through this wind. While this amount of lost mass seems large by Earth standards, it is completely insignificant for the Sun.

From where in the Sun does the solar wind emerge? In visible photographs, the solar corona appears fairly uniform and smooth. X-ray and extreme ultraviolet pictures, however, show that the corona has loops, plumes, and both bright and dark regions. Large dark regions of the corona that are relatively cool and quiet are called **coronal holes**, shown in Figure 10.12. In these regions, magnetic field lines stretch far out into space away from the Sun, rather than looping back to the surface. The solar wind comes predominantly from coronal holes, where gas can stream away from the Sun into space unhindered by magnetic fields. Hot coronal gas, on the other hand, is present mainly where magnetic fields have trapped and concentrated it.

Coronal Hole

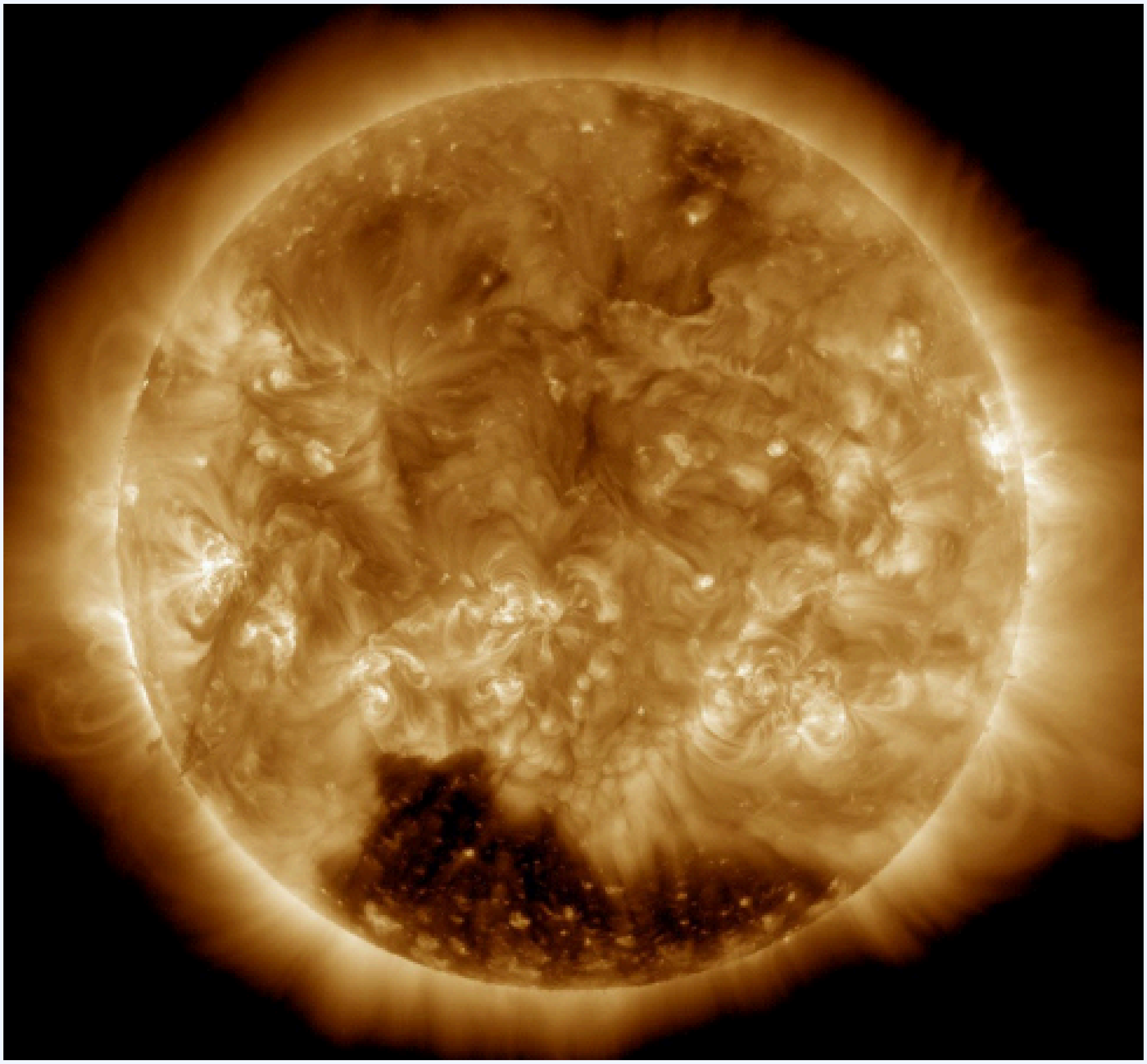


Figure 10.12. The dark area visible near the Sun's south pole on this Solar Dynamics Observer spacecraft image is a coronal hole.
Solar Dynamics Observatory Welcomes the New Year by NASA/SDO, NASA Media License

At the surface of Earth, we are protected to some degree from the solar wind by our atmosphere and Earth's magnetic field. However, the magnetic field lines come into Earth at the north and south magnetic poles. Here, charged particles accelerated by the solar wind can follow the field down into our atmosphere. As the particles strike molecules of air, they cause them to glow, producing beautiful curtains of light called the **auroras**, or the northern and southern lights as shown in Figure 10.13.

Aurora

Figure 10.13. The colourful glow in the sky results from charged particles in a solar wind interacting with Earth's magnetic fields. The stunning display captured here occurred over Jokulsarlon Lake in Iceland in 2013.
Aurora Borealis by Moyan Brenn, CC BY 4.0

Attribution

"15.1 The Structure and Composition of the Sun " from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

10.4 THE ACTIVE SUN

Before the invention of the telescope, the Sun was thought to be an unchanging and perfect sphere. We now know that the Sun is in a perpetual state of change: its surface is a seething, bubbling cauldron of hot gas. Areas that are darker and cooler than the rest of the surface come and go. Vast plumes of gas erupt into the chromosphere and corona. Occasionally, there are even giant explosions on the Sun that send enormous streamers of charged particles and energy hurtling toward Earth. When they arrive, these can cause power outages and other serious effects on our planet.

Sunspots

The first evidence that the Sun changes came from studies of **sunspots**, which are large, dark features seen on the surface of the Sun caused by increased magnetic activity. They look darker because the spots are typically at a temperature of about 3800 K, whereas the bright regions that surround them are at about 5800 K, pictured in Figure 10.14. Occasionally, these spots are large enough to be visible to the unaided eye, and we have records going back over a thousand years from observers who noticed them when haze or mist reduced the Sun's intensity. (We emphasize what your parents have surely told you: looking at the Sun for even a brief time can cause permanent eye damage. This is the one area of astronomy where we don't encourage you to do your own observing without getting careful instructions or filters from your instructor.)

Sunspots

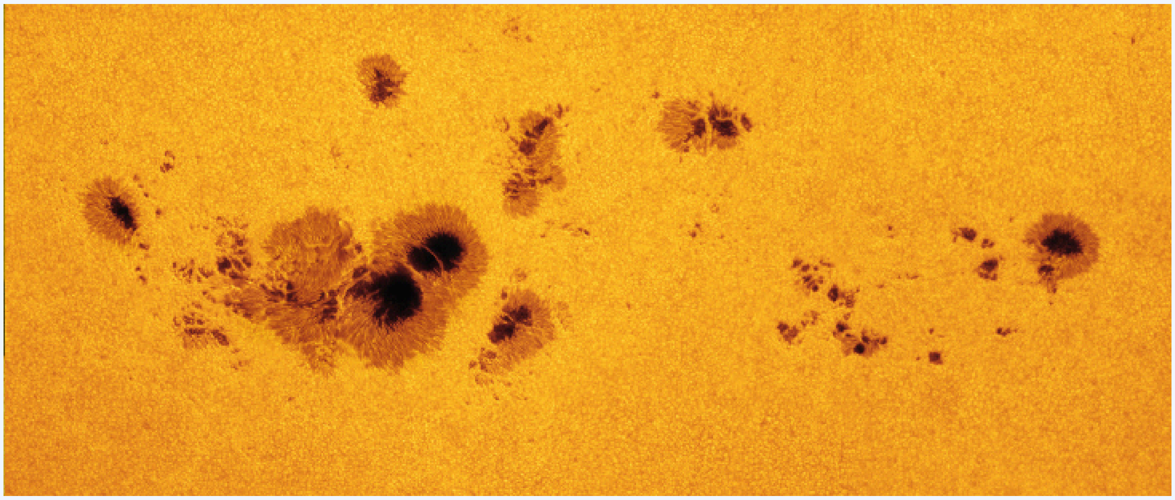


Figure 10.14. This image of sunspots, cooler and thus darker regions on the Sun, was taken in July 2012. You can see the dark, central region of each sunspot (called the umbra) surrounded by a less dark region (the penumbra). The largest spot shown here is about 11 Earths wide. Although sunspots appear dark when seen next to the hotter gases of the photosphere, an average sunspot, cut out of the solar surface and left standing in the night sky, would be about as bright as the full moon. The mottled appearance of the Sun's surface is granulation.
Solar Archipelago by NASA Goddard Space Flight Center, Alan Friedman, CC-BY-4.0

While we understand that sunspots look darker because they are cooler, they are nevertheless hotter than the surfaces of many stars. If they could be removed from the Sun, they would shine brightly. They appear dark only in contrast with the hotter, brighter photosphere around them.

Individual sunspots come and go, with lifetimes that range from a few hours to a few months. If a spot lasts and develops, it usually consists of two parts: an inner darker core, the **umbra**, and a surrounding less dark region, the **penumbra**. Many spots become much larger than Earth, and a few, like the largest one shown in Figure 10.14, have reached diametres over 140,000 kilometres. Frequently, spots occur in groups of 2 to 20 or more. The largest groups are very complex and may have over 100 spots. Like storms on Earth, sunspots are not fixed in position, but they drift slowly compared with the Sun's rotation.

By recording the apparent motions of the sunspots as the turning Sun carried them across its disk as shown Figure 10.15, Galileo, in 1612, demonstrated that the Sun rotates on its axis with a rotation period of approximately 1 month. Our star turns in a west-to-east direction, like the orbital motions of the planets. The Sun, however, is a gas and does not have to rotate rigidly, the way a solid body like Earth does. Modern observations show that the speed of rotation of the Sun varies according to latitude, that is, it's different as you go north or south of the Sun's equator. The rotation period is about 25 days at the equator, 28 days at latitude 40°, and 36 days at latitude 80°. We call this **behaviour differential rotation.**

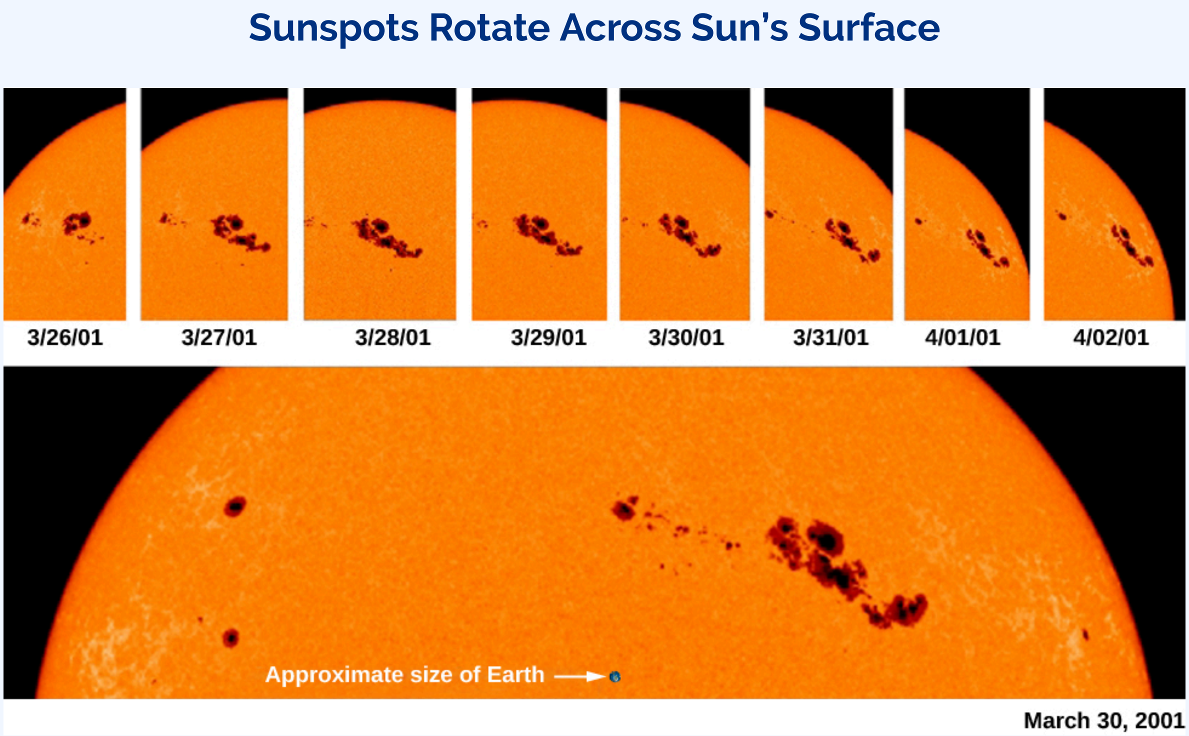

Sunspots Rotate Across Sun's Surface

Figure 10.15. This sequence of photographs of the Sun's surface tracks the movement of sunspots across the visible hemisphere of the Sun. On March 30, 2001, this group of sunspots extended across an area about 13 times the diameter of Earth. This region produced many flares and coronal mass ejections. Sunspot series by SOHO/NASA/ESA, NASA Media License.

The Sunspot Cycle

Between 1826 and 1850, Heinrich Schwabe, a German pharmacist and amateur astronomer, kept daily records of the number of sunspots. What he was really looking for was a planet inside the orbit of Mercury, which he hoped to find by observing its dark silhouette as it passed between the Sun and Earth. He failed to find the hoped-for planet, but his diligence paid off with an even-more important discovery: the sunspot cycle. He found that the number of sunspots varied systematically, in cycles about a decade long.

What Schwabe observed was that, although individual spots are short lived, the total number visible on the Sun at any one time was likely to be very much greater at certain times—the periods of **sunspot**

maximum—than at other times—the periods of **sunspot minimum**. We now know that sunspot maxima occur at an *average* interval of 11 years, but the intervals between successive maxima have ranged from as short as 9 years to as long as 14 years. During sunspot maxima, more than 100 spots can often be seen at once. Even then, less than one-half of one percent of the Sun's surface is covered by spots. During sunspot minima, sometimes no spots are visible. The Sun's activity reached its most recent maximum in 2014.

Sunspots are not the only features that vary during a solar cycle. There are dramatic changes in the chromosphere and corona as well. To see what happens in the chromosphere, we must observe the emission lines from elements such as hydrogen and calcium, which emit useful spectral lines at the temperatures in that layer. The hot corona, on the other hand, can be studied by observations of X-rays and of extreme ultraviolet and other wavelengths at high energies.

Plages and Prominences

As we saw, emission lines of hydrogen and calcium are produced in the hot gases of the chromosphere. Astronomers routinely photograph the Sun through filters that transmit light only at the wavelengths that correspond to these emission lines. Pictures taken through these special filters show bright "clouds" in the chromosphere around sunspots; these bright regions are known as **plages**, pictured in Figure 10.16. These are regions within the chromosphere that have higher temperature and density than their surroundings. The plages actually contain all of the elements in the Sun, not just hydrogen and calcium. It just happens that the spectral lines of hydrogen and calcium produced by these clouds are bright and easy to observe.

Plages on the Sun

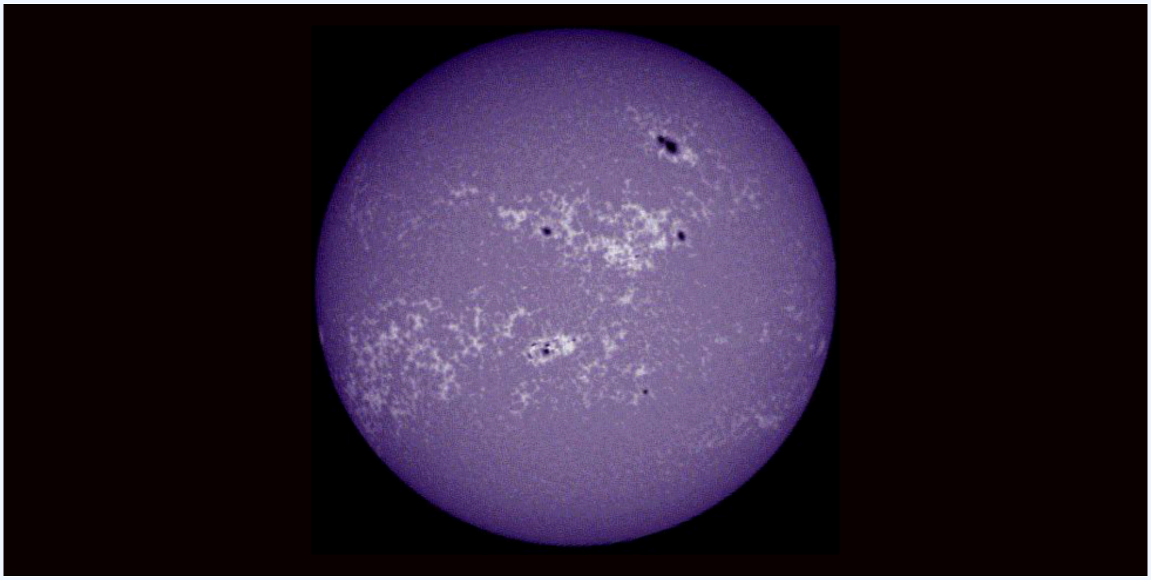


Figure 10.16. This image of the Sun was taken with a filter that transmits only the light of the spectral line produced by singly ionized calcium. The bright cloud-like regions are the plages.
The Chromospheric Network by SOHO/NASA/ESA, NASA Media License.

Moving higher into the Sun's atmosphere, we come to the spectacular phenomena called **prominences**, shown in Figure 10.17, which usually originate near sunspots. Eclipse observers often see prominences as red features rising above the eclipsed Sun and reaching high into the corona. Some, the ***quiescent*** **prominences**, are graceful loops of plasma (ionized gas) that can remain nearly stable for many hours or even days. The relatively rare ***eruptive*** **prominences** appear to send matter upward into the corona at high speeds, and the most active ***surge*** **prominences** may move as fast as 1300 kilometres per second (almost 3 million miles per hour). Some eruptive prominences have reached heights of more than 1 million kilometres above the photosphere; Earth would be completely lost inside one of those awesome displays shown in Figure 10.17.

Prominences

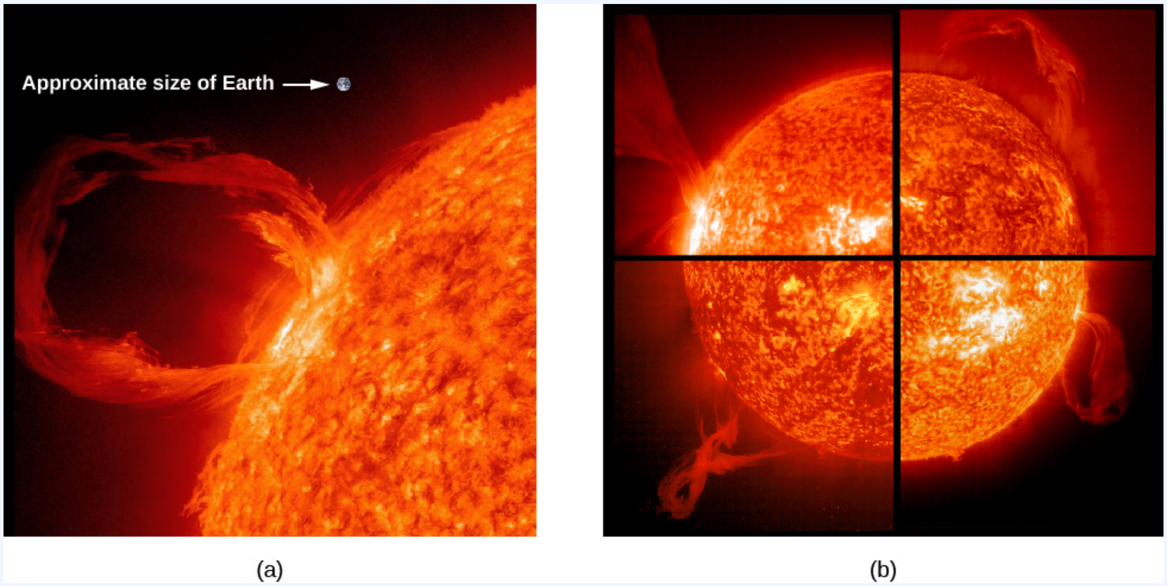


Figure 10.17. (a) This image of an eruptive prominence was taken in the light of singly ionized helium in the extreme ultraviolet part of the spectrum. The prominence is a particularly large one. An image of Earth is shown at the same scale for comparison. (b) A prominence is a huge cloud of relatively cool (about 60,000 K in this case), fairly dense gas suspended in the much hotter corona. These pictures, taken in ultraviolet, are colour coded so that white corresponds to the hottest temperatures and dark red to cooler ones. The four images were taken, moving clockwise from the upper left, on May 15, 2001; March 28, 2000; January 18, 2000; and February 2, 2001.
Credit a: What is a solar prominence? by SOHO/NASA, NASA Media License
Credit b: Image of a pair of similarly shaped prominences from 11 January 1998 by NASA/SDO, NASA Media License.

Flares and Coronal Mass Ejections

The most violent event on the surface of the Sun is a rapid eruption called a **solar flare** as pictured in Figure 10.18. A typical flare lasts for 5 to 10 minutes and releases a total amount of energy equivalent to that of perhaps a million hydrogen bombs. The largest flares last for several hours and emit enough energy to power the entire United States at its current rate of electrical consumption for 100,000 years. Near sunspot maximum, small flares occur several times per day, and major ones may occur every few weeks.

Solar Flare

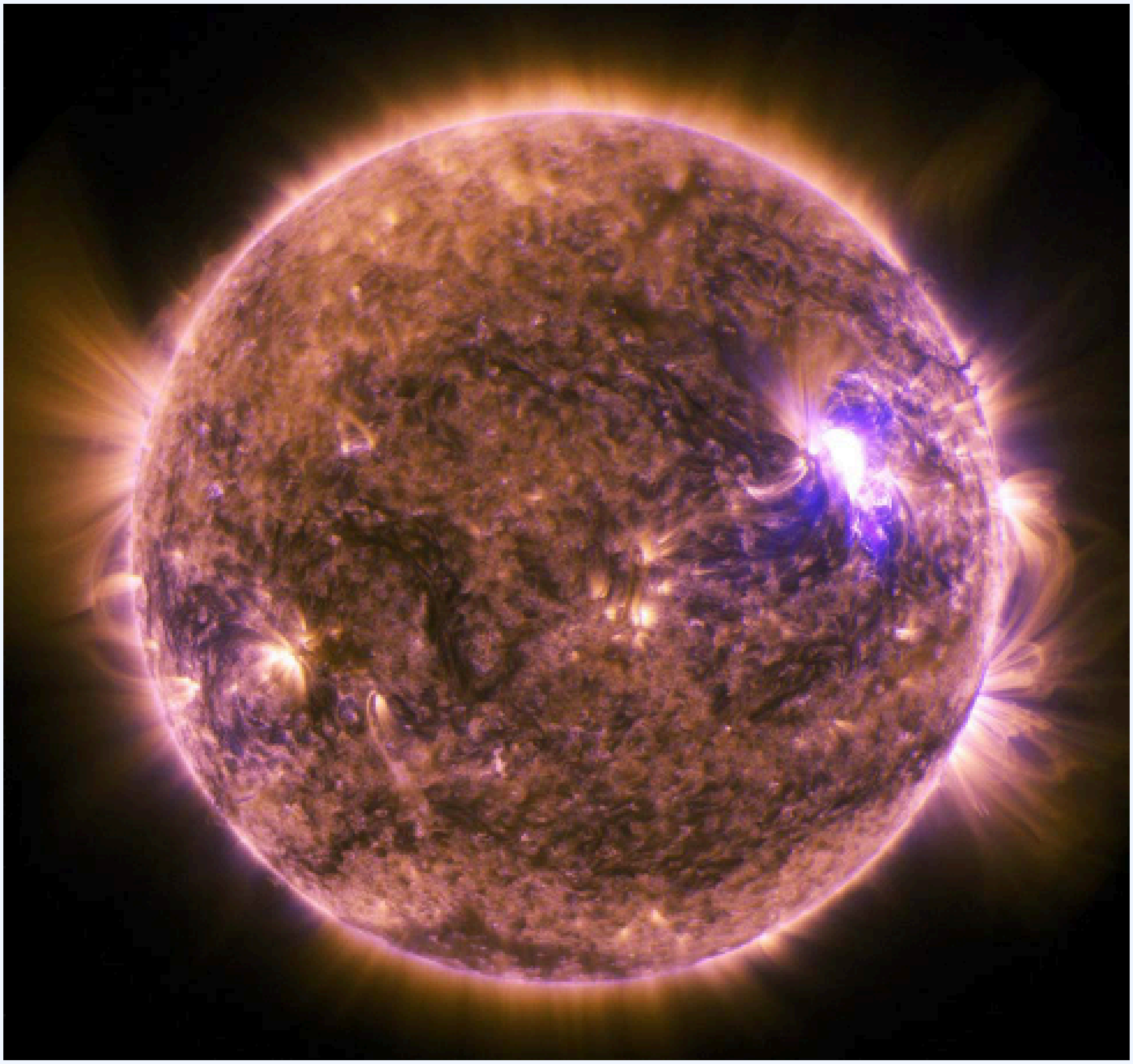


Figure 10.18. The bright white area seen on the right side of the Sun in this image from the Solar Dynamics Observer spacecraft is a solar flare that was observed on June 25, 2015.
Holiday Lights on the Sun: Imagery of a Solar Flare by NASA/SDO, NASA Media License.

Flares, like the one shown in Figure 10.18, are often observed in the red light of hydrogen, but the visible emission is only a tiny fraction of the energy released when a solar flare explodes. At the moment of the explosion, the matter associated with the flare is heated to temperatures as high as 10 million K. At such high temperatures, a flood of X-ray and ultraviolet radiation is emitted.

Flares seem to occur when magnetic fields pointing in opposite directions release energy by interacting with and destroying each other—much as a stretched rubber band releases energy when it breaks.

What is different about flares is that their magnetic interactions cover a large volume in the solar corona and release a tremendous amount of electromagnetic radiation. In some cases, immense quantities of coronal material—mainly protons and electrons—may also be ejected at high speeds (500–1000 kilometres per second) into interplanetary space. Such a **coronal mass ejection (CME)** can affect Earth in a number of ways (which we will discuss in the section on space weather).

Flare and Coronal Mass Ejection

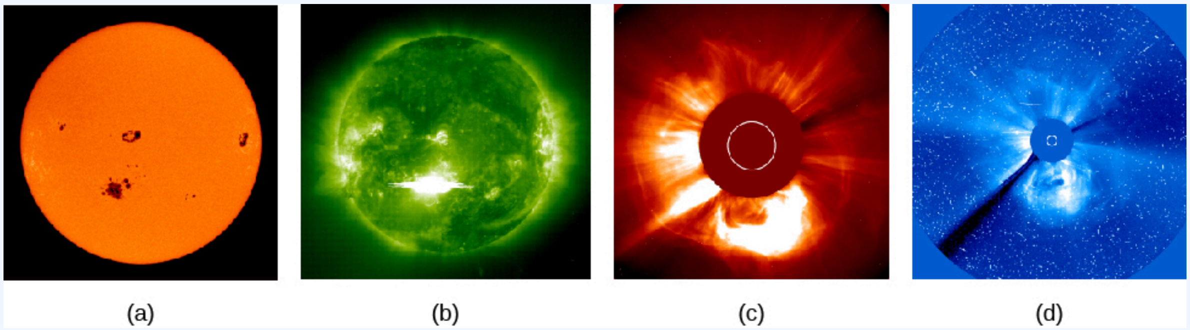


Figure 10.19. This sequence of four images shows the evolution over time of a giant eruption on the Sun. (a) The event began at the location of a sunspot group, and (b) a flare is seen in far-ultraviolet light. (c) Fourteen hours later, a CME is seen blasting out into space. (d) Three hours later, this CME has expanded to form a giant cloud of particles escaping from the Sun and is beginning the journey out into the solar system. The white circle in (c) and (d) shows the diameter of the solar photosphere. The larger dark area shows where light from the Sun has been blocked out by a specially designed instrument to make it possible to see the faint emission from the corona.
MDI, EIT 195, LASCO C2, C3 CMEs by SOHO/EIT/LASCO/MDI (NASA & ESA), NASA Media License.

Active Regions

To bring the discussion of the last two sections together, astronomers now realize that sunspots, flares, and bright regions in the chromosphere and corona tend to occur together on the Sun in time and space. That is, they all tend to have similar longitudes and latitudes, but they are located at different heights in the atmosphere. Because they all occur together, they vary with the sunspot cycle.

Solar Cycle

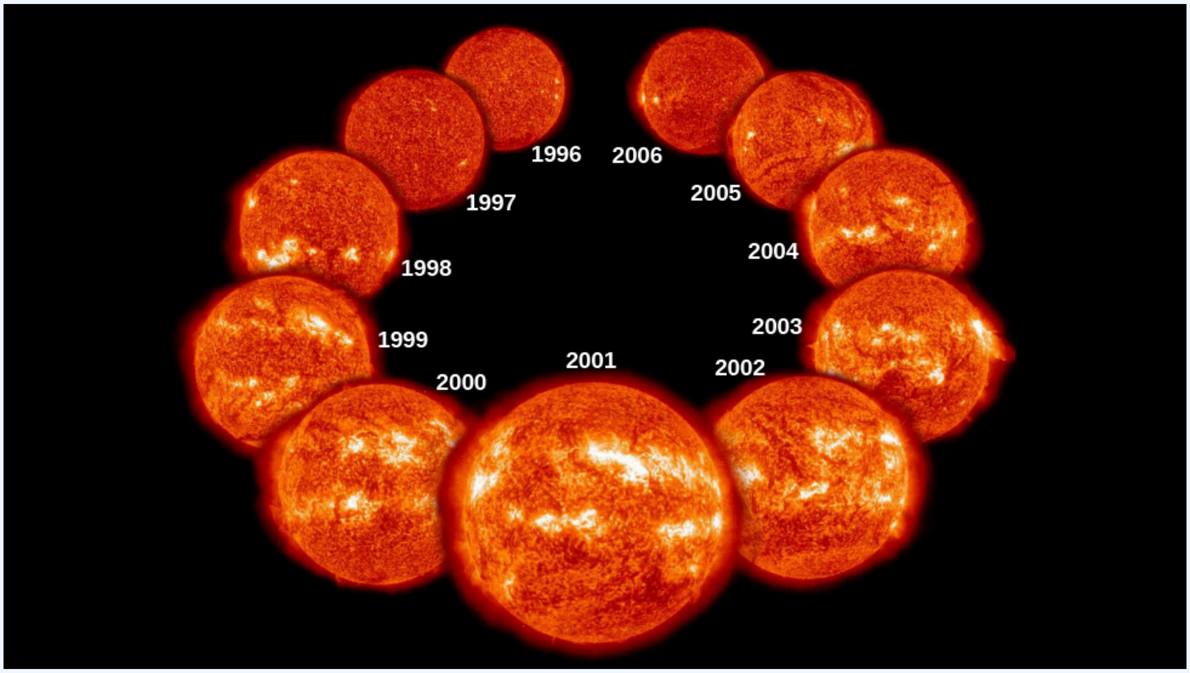


Figure 10.20. This dramatic sequence of images taken from the SOHO satellite over a period of 11 years shows how active regions change during the solar cycle. The images were taken in the ultraviolet region of the spectrum and show that active regions on the Sun increase and decrease during the cycle. Sunspots are located in the cooler photosphere, beneath the hot gases shown in this image, and vary in phase with the emission from these hot gases—more sunspots and more emission from hot gases occur together. Collage of Solar Cycle 23 by ESA/NASA/SOHO, NASA Media License.

For example, flares are more likely to occur near sunspot maximum, and the corona is much more conspicuous at that time as shown in Figure 10.20. A place on the Sun where a number of these phenomena are seen is called an **active region**, pictured in Figure 10.21. As you might deduce from our earlier discussion, active regions are always associated with strong magnetic fields.

Solar Active Region Observed at Different Heights in the Sun's Atmosphere

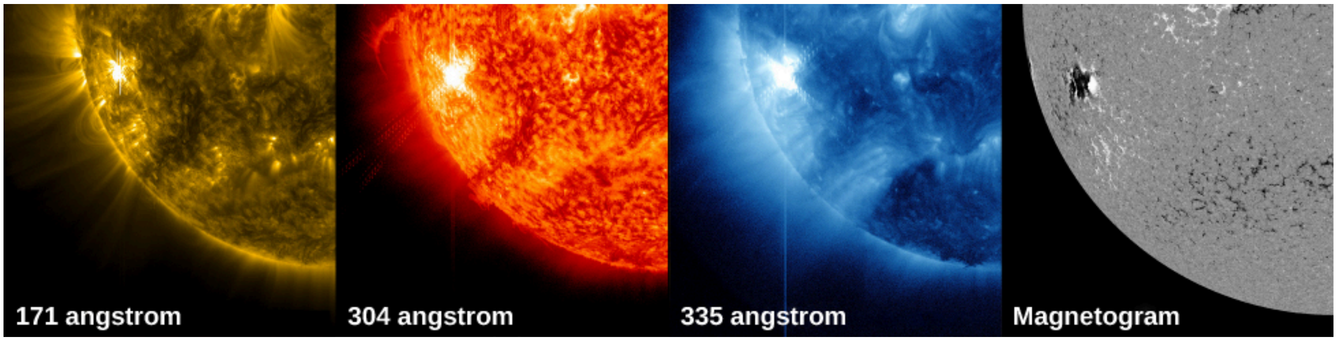


Figure 10.21. These four images of a solar flare on October 22, 2012, show from the left: light from the Sun at a wavelength of 171 angstroms, which shows the structure of loops of solar material in the corona; ultraviolet at 304 angstroms, which shows light from the region of the Sun's atmosphere where flares originate; light at 335 angstroms, which highlights radiation from active regions in the corona; a magnetogram, which shows magnetically active regions on the Sun. Note how these different types of activity all occur above a sunspot region with a strong magnetic field.
Modification of image by NASA/SDO/Goddard, NASA Media License.

Attribution

10.5 THE SOLAR INTERIOR

Fusion of protons can occur in the centre of the Sun only if the temperature exceeds 12 million K. How do we know that the Sun is actually this hot? To determine what the interior of the Sun might be like, it is necessary to resort to complex calculations. Since we can't see the interior of the Sun, we have to use our understanding of physics, combined with what we see at the surface, to construct a mathematical model of what must be happening in the interior. Astronomers use observations to build a computer program containing everything they think they know about the physical processes going on in the Sun's interior. The computer then calculates the temperature and pressure at every point inside the Sun and determines what nuclear reactions, if any, are taking place. For some calculations, we can use observations to determine whether the computer program is producing results that match what we see. In this way, the program evolves with ever-improving observations.

The computer program can also calculate how the Sun will change with time. After all, the Sun must change. In its center, the Sun is slowly depleting its supply of hydrogen and creating helium instead. Will the Sun get hotter? Cooler? Larger? Smaller? Brighter? Fainter? Ultimately, the changes in the center could be catastrophic, since eventually all the hydrogen fuel hot enough for fusion will be exhausted. Either a new source of energy must be found, or the Sun will cease to shine. We will describe the ultimate fate of the Sun in later chapters. For now, let's look at some of the things we must teach the computer about the Sun in order to carry out such calculations.

The Sun Is a Plasma

The Sun is so hot that all of the material in it is in the form of an ionized gas, called a **plasma**. Plasma acts much like a hot gas, which is easier to describe mathematically than either liquids or solids. The particles that constitute a gas are in rapid motion, frequently colliding with one another. This constant bombardment is the **pressure** of the gas, it is illustrated in Figure 10.22.

Gas Pressure

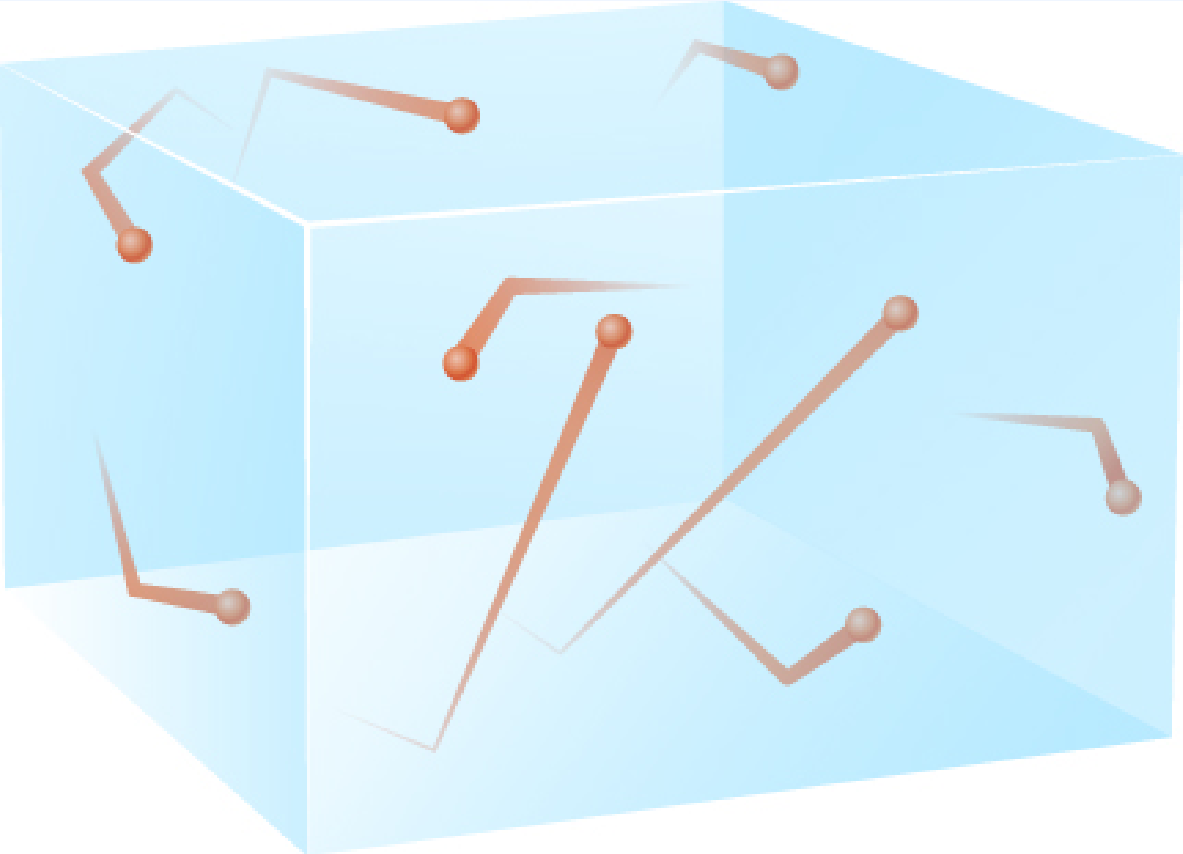


Figure 10.22. The particles in a gas are in rapid motion and produce pressure through collisions with the surrounding material. Here, particles are shown bombarding the sides of an imaginary container.

More particles within a given volume of gas produce more pressure because the combined impact of the moving particles increases with their number. The pressure is also greater when the molecules or atoms are moving faster. Since the molecules move faster when the temperature is hotter, higher temperatures produce higher pressure.

The Sun Is Stable

The Sun, like the majority of other stars, is stable; it is neither expanding nor contracting. Such a star is said to be in a condition of **equilibrium**. All the forces within it are balanced, so that at each point within the star, the temperature, pressure, density, and so on are maintained at constant values. We will see in later chapters that even these stable stars, including the Sun, are changing as they evolve, but such evolutionary changes are so gradual that, for all intents and purposes, the stars are still in a state of equilibrium at any given time.

The mutual gravitational attraction between the masses of various regions within the Sun produces tremendous forces that tend to collapse the Sun toward its centre. Yet we know from the history of Earth that the Sun has been emitting roughly the same amount of energy for billions of years, so clearly it has managed to resist collapse for a very long time. The gravitational forces must therefore be counterbalanced by some other force. That force is due to the pressure of gases within the Sun as shown in Figure 10.23. Calculations show that, in order to exert enough pressure to prevent the Sun from collapsing due to the force of gravity, the gases at its center must be maintained at a temperature of 15 million K. Think about what this tells us. Just from the fact that the Sun is not contracting, we can conclude that its temperature must indeed be high enough at the centre for protons to undergo fusion.

Hydrostatic Equilibrium

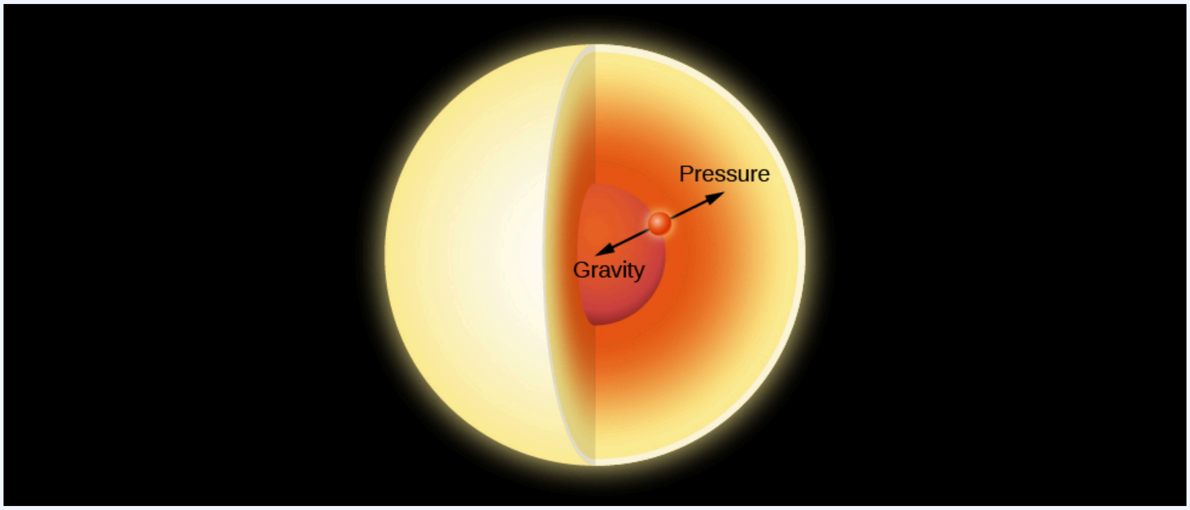


Figure 10.23. In the interior of a star, the inward force of gravity is exactly balanced at each point by the outward force of gas pressure.

The Sun maintains its stability in the following way. If the internal pressure in such a star were not great enough to balance the weight of its outer parts, the star would collapse somewhat, contracting and building up the pressure inside. On the other hand, if the pressure were greater than the weight of the overlying layers, the star would expand, thus decreasing the internal pressure. Expansion would stop, and equilibrium would again be reached when the pressure at every internal point equaled the weight of the stellar layers above that point. An analogy is an inflated balloon, which will expand or contract until an equilibrium is reached between the pressure of the air inside and outside. The technical term for this condition is **hydrostatic equilibrium**. Stable stars are all in hydrostatic equilibrium; so are the oceans of Earth as well as Earth's atmosphere. The air's own pressure keeps it from falling to the ground.

The Sun Is Not Cooling Down

As everyone who has ever left a window open on a cold winter night knows, heat always flows from hotter to cooler regions. As energy filters outward toward the surface of a star, it must be flowing from inner, hotter regions. The temperature cannot ordinarily get cooler as we go inward in a star, or energy would flow in and heat up those regions until they were at least as hot as the outer ones. Scientists conclude that the temperature is highest at the centre of a star, dropping to lower and lower values toward the stellar surface. (The high temperature of the Sun's chromosphere and corona may therefore appear to be a paradox. But these high temperatures are maintained by magnetic effects, which occur in the Sun's atmosphere.)

The outward flow of energy through a star robs it of its internal heat, and the star would cool down if that energy were not replaced. Similarly, a hot iron begins to cool as soon as it is unplugged from its source of electric energy. Therefore, a source of fresh energy must exist within each star. In the Sun's case, we have seen that this energy source is the ongoing fusion of hydrogen to form helium.

Heat Transfer in a Star

Since the nuclear reactions that generate the Sun's energy occur deep within it, the energy must be transported from the centre of the Sun to its surface—where we see it in the form of both heat and light. There are three ways in which energy can be transferred from one place to another. In **conduction**, atoms or molecules pass on their energy by colliding with others nearby. This happens, for example, when the handle of a metal spoon heats up as you stir a cup of hot coffee. In convection, currents of warm material rise, carrying their energy with them to cooler layers. A good example is hot air rising from a fireplace. In **radiation**, energetic photons move away from hot material and are absorbed by some material to which they convey some or all of their energy. You can feel this when you put your hand close to the coils of an electric heater, allowing infrared photons to heat up your hand. Conduction and convection are both important in the interiors of planets. In stars, which are much more transparent, radiation and convection are important, whereas conduction can usually be ignored.

Stellar *convection* occurs as currents of hot gas flow up and down through the star as shown in Figure 10.24. Such currents travel at moderate speeds and do not upset the overall stability of the star. They don't even result in a net transfer of mass either inward or outward because, as hot material rises, cool material falls and replaces it. This results in a convective circulation of rising and falling cells as seen in Figure 10.24. In much the same way, heat from a fireplace can stir up air currents in a room, some rising and some falling, without driving any air into or out the room. Convection currents carry heat very efficiently outward through a star. In the Sun, convection turns out to be important in the central regions and near the surface.

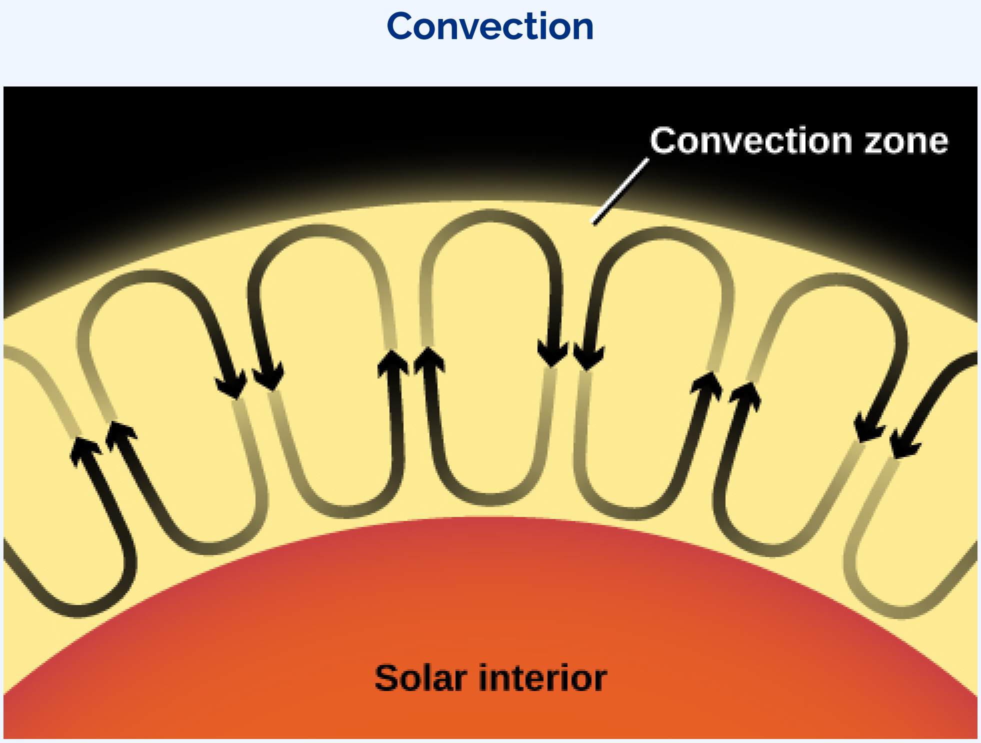


Figure 10.24. Rising convection currents carry heat from the Sun's interior to its surface, whereas cooler material sinks downward. Of course, nothing in a real star is as simple as diagrams in textbooks suggest.

Unless convection occurs, the only significant mode of energy transport through a star is by electromagnetic radiation. Radiation is not an efficient means of energy transport in stars because gases in stellar interiors are very opaque, that is, a photon does not go far (in the Sun, typically about 0.01 meter) before it is absorbed. The absorbed energy is always reemitted, but it can be reemitted in any direction. A photon absorbed when travelling outward in a star has almost as good a chance of being radiated back toward the centre of the star as toward its surface.

A particular quantity of energy, therefore, zigzags around in an almost random manner and takes a long time to work its way from the centre of a star to its surface as shown in Figure 10.25. Estimates are somewhat uncertain, but in the Sun, as we saw, the time required is probably between 100,000 and 1,000,000 years. If the photons were not absorbed and reemitted along the way, they would travel at the speed of light and could reach the surface in a little over 2 seconds, just as neutrinos do. Pictured in Figure 10.26.

Photons Deep in the Sun

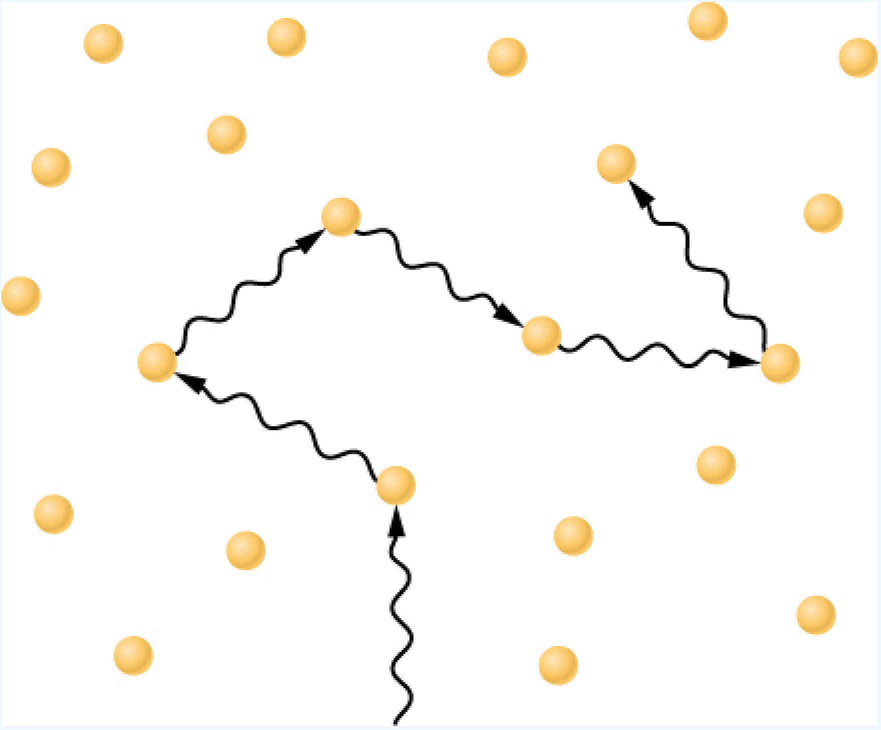


Figure 10.25. A photon moving through the dense gases in the solar interior travels only a short distance before it interacts with one of the surrounding atoms. The photon usually has a lower energy after each interaction and may then travel in any random direction.

Photon and Neutrino Paths in the Sun

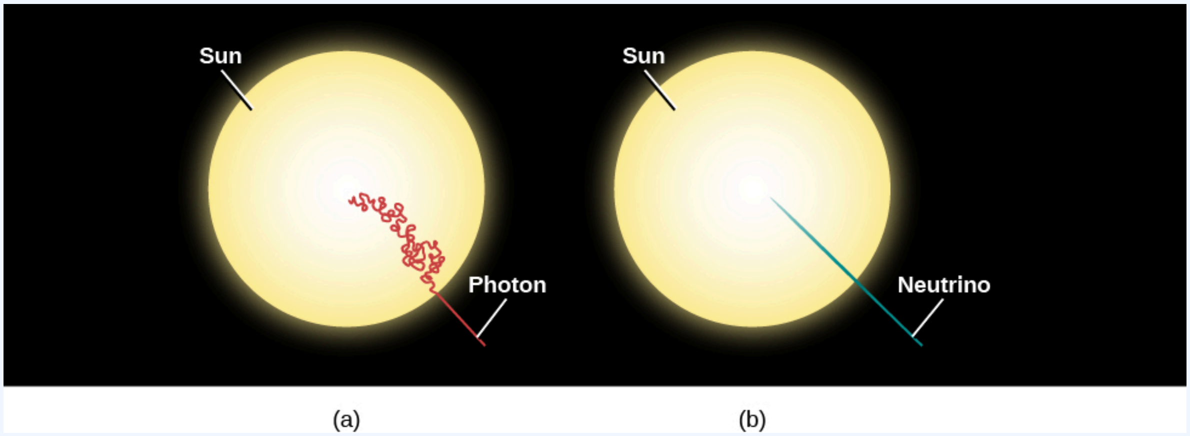


Figure 10.26. (a) Because photons generated by fusion reactions in the solar interior travel only a short distance before being absorbed or scattered by atoms and sent off in random directions, estimates are that it takes between 100,000 and 1,000,000 years for energy to make its way from the centre of the Sun to its surface. (b) In contrast, neutrinos do not interact with matter but traverse straight through the Sun at the speed of light, reaching the surface in only a little more than 2 seconds.

Model Stars

Scientists use the principles we have just described to calculate what the Sun's interior is like. These physical ideas are expressed as mathematical equations that are solved to determine the values of temperature, pressure, density, the efficiency with which photons are absorbed, and other physical quantities throughout the Sun. The solutions obtained, based on a specific set of physical assumptions, provide a theoretical model for the interior of the Sun.

Figure 10.27 schematically illustrates the predictions of a theoretical model for the Sun's interior. Energy is generated through fusion in the core of the Sun, which extends only about one-quarter of the way to the surface but contains about one-third of the total mass of the Sun. At the centre, the temperature reaches a maximum of approximately 15 million K, and the density is nearly 150 times that of water. The energy generated in the core is transported toward the surface by radiation until it reaches a point about 70% of the distance from the centre to the surface. At this point, convection begins, and energy is transported the rest of the way, primarily by rising columns of hot gas.

Interior Structure of the Sun

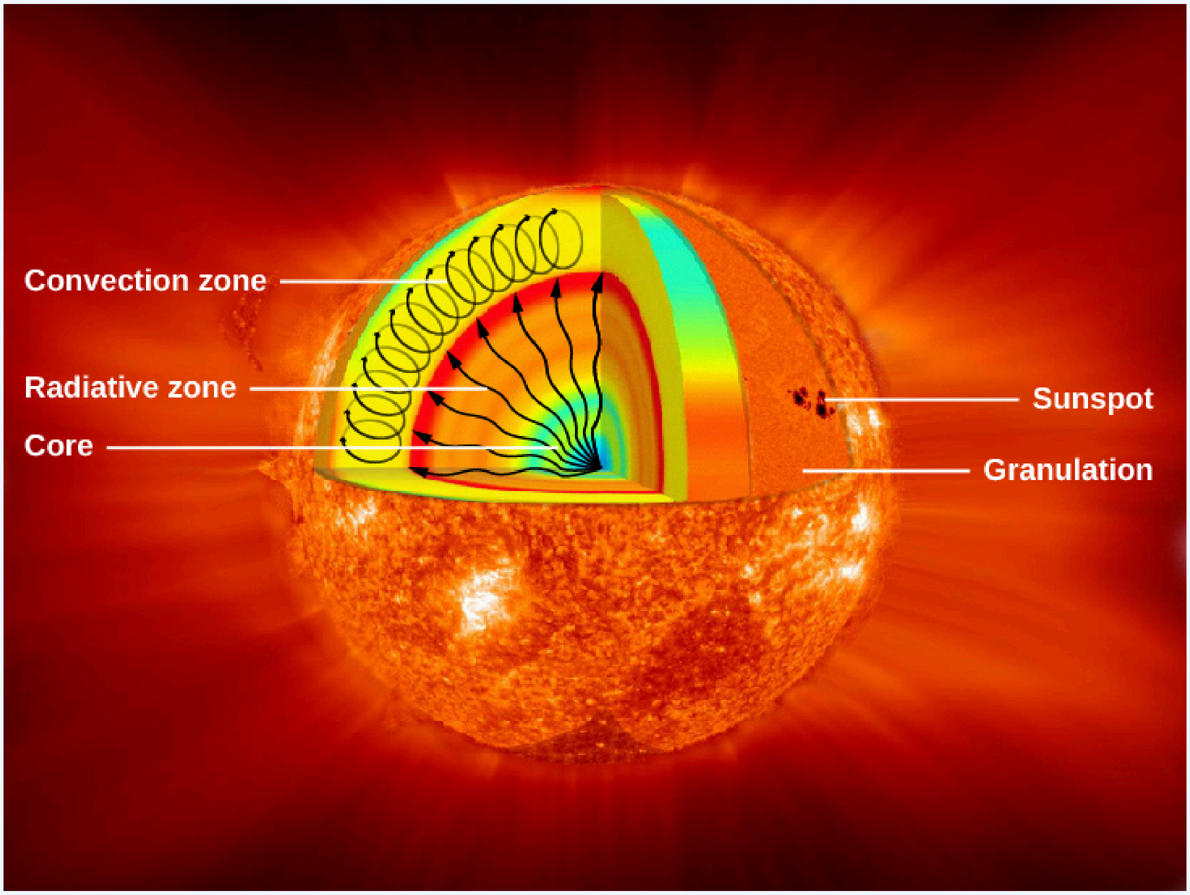


Figure 10.27. Energy is generated in the core by the fusion of hydrogen to form helium. This energy is transmitted outward by radiation—that is, by the absorption and reemission of photons. In the outermost layers, energy is transported mainly by convection.
Modification of The Sun by NASA/Goddard, NASA Media License.

Figure 10.28 shows how the temperature, density, rate of energy generation, and composition vary from the centre of the Sun to its surface.

Interior of the Sun

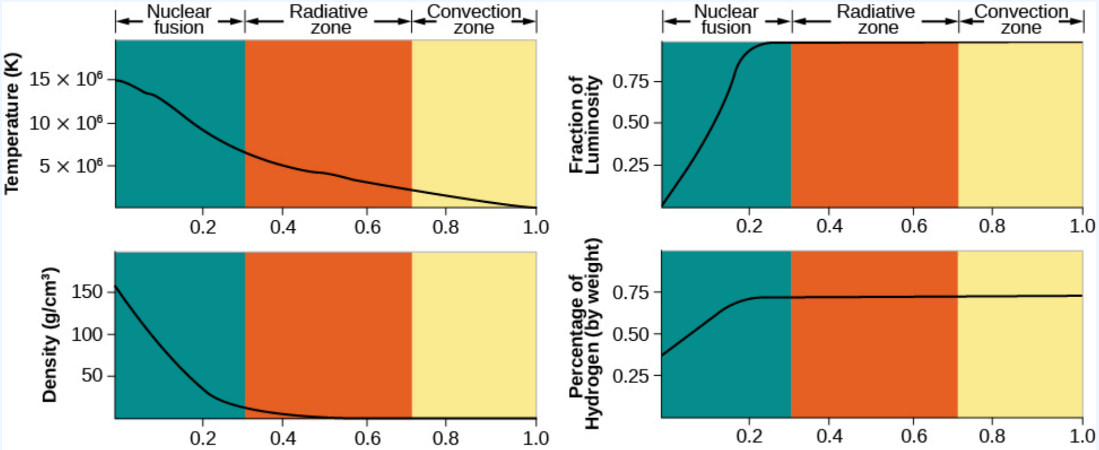


Figure 10.28. Diagrams showing how temperature, density, rate of energy generation, and the percentage (by mass) abundance of hydrogen vary inside the Sun. The horizontal scale shows the fraction of the Sun's radius: the left edge is the very center, and the right edge is the visible surface of the Sun, which is called the photosphere.

Attribution

"16.3 The Solar Interior: Theory" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

10.6 KEY TERMS

Active region: A place on the Sun where a sunspots, flares, and bright regions in the chromosphere and corona are seen. 10.4

Aurora: charged particles accelerated by the solar wind that can follow Earth's magnetic field down into our atmosphere; these particles strike molecules of air, causing them to glow and produce beautiful curtains of light. 10.3

Behaviour differential rotation: the type of rotation possessed by the Sun which varies according to latitude, that is, it's different as you go north or south of the Sun's equator. 10.4

Chromosphere: the region of the Sun's atmosphere that lies immediately above the photosphere, is about 2000 to 3000 kilometres thick, and whose spectrum consists of bright emission lines, indicating that this layer is composed of hot gases emitting light at discrete wavelengths. 10.3

Conduction: process by which heat is directly transmitted through a substance when there is a difference of temperature between adjoining regions caused by atomic or molecular collisions. 10.5

Convective zone: the outermost layer of the solar interior which is approximately 200,000 kilometres deep and transports energy from the edge of the radiative zone to the surface through giant convection cells. 10.3

Corona: the outermost part of the Sun's atmosphere which extends millions of kilometres above the photosphere and emits about half as much light as the full moon. 10.3

Coronal hole: a region in the Sun's outer atmosphere that appears darker because there is less hot gas there. 10.3

Coronal mass ejection (CME): event where immense quantities of coronal material—mainly protons and electrons—may also be ejected at high speeds (500–1000 kilometres per second) into interplanetary space. 10.4

Deuterium nucleus: an isotope (or version) of hydrogen that contains one proton and one neutron. 10.2

Equilibrium: a condition where a star is stable—neither expanding nor contracting—and where all the forces within it are balanced, so that at each point within the star, the temperature, pressure, density, and so on are maintained at constant values. 10.5

Eruptive prominences: relatively rare prominences that appear to send matter upward into the corona at high speeds. 10.4

Granulation: the rice-grain-like structure of the solar photosphere; granulation is produced by upwelling currents of gas that are slightly hotter, and therefore brighter, than the surrounding regions, which are flowing downward into the Sun. 10.3

Hydrostatic equilibrium: balance between the weights of various layers, as in a star or Earth's atmosphere, and the pressures that support them. 10.5

Ionized: atoms that are stripped of one or more of their electrons. 10.3

Penumbra: a region surrounding the umbra in a lasting sunspot that is less dark. 10.4

Photosphere: the layer where the Sun becomes opaque and marks the boundary past which we cannot see. 10.3

Plages: bright regions within the chromosphere that have higher temperature and density than their surroundings. 10.4

Plasma: the material in the Sun which is in the form of an ionized gas and acts much like a hot gas. 10.5

Positron: particle with the same mass as an electron, but positively charged. 10.2

Pressure (of a gas): the constant bombardment where the particles that constitute a gas are in rapid motion, frequently colliding with one another. 10.5

Prominences: phenomena higher in the Sun's atmosphere that usually originate near sunspots. 10.4

Proton-proton chain: series of thermonuclear reactions by which nuclei of hydrogen are built up into nuclei of helium. 10.2

Quiescent prominences: graceful loops of plasma (ionized gas) that can remain nearly stable for many hours or even days. 10.4

Radiation: emission of energy as electromagnetic waves or photons also the transmitted energy itself. 10.5

Radiative zone: the region above the Sun's core named for the primary mode of transporting energy across it; this region starts at about 25% of the distance to the solar surface and extends up to about 70% of the way to the surface. 10.3

Solar flare: a rapid eruption on the surface of the Sun that lasts for 5 to 10 minutes and releases a total amount of energy equivalent to that of perhaps a million hydrogen bombs. 10.4

Solar wind: a flow of hot, charged particles leaving the Sun. 10.3

Stellar convection: a process that occurs as currents of hot gas flow up and down through a star. 10.5

Sunspots: large, dark features seen on the surface of the Sun caused by increased magnetic activity. 10.4

Sunspot cycel: the semiregular 11-year period with which the frequency of sunspots fluctuates. 10.4

Sunspot maximum: a period of time when the the total number visible spots on the Sun is likely to be much greater. 10.4

Sunspot minimum: a period of time when the the total number visible spots on the Sun is likely to be much lower. 10.4

Sun's core: an area located at the center of the Sun that is extremely dense and is the source of all of its energy. 10.3

Surge prominences: the most active prominences may move as fast as 1300 kilometres per second (almost 3 million miles per hour). 10.4

Umbra: inner dark core of a sunspot that lasts and develops. 10.4

CHAPTER 11: STARS: CLASSIFICATION, EVOLUTION, AND DEATH

Chapter Overview

11.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify the two primary components used to classify stars on the Hertzsprung-Russell (H-R) diagram.
- Compare and contrast the spectral lines observed in the spectra of very hot, intermediate, and very cool stars, explaining the differences in line strengths and types of elements present.
- Define the stages of stellar evolution and explain the factors that determine the lifetime of a star and how the rate of fusion changes as the star evolves.
- Compare and contrast the evolution of low-mass stars and massive stars in terms of nuclear fusion reactions and element synthesis.
- Assess the impact of supernova explosions on the universe, including their role in synthesizing heavy elements and distributing them throughout space, and their influence on stellar and galactic evolution.
- Discuss the life cycle and evolution of neutron stars and pulsars.
- Summarize the main characteristics of a black hole, including its formation, size, and the behavior of matter falling into it.

11.1 CLASSIFICATION BY COLOUR

Look at the beautiful picture of the stars in the Sagittarius Star Cloud shown in Figure 11.1. The stars show a multitude of colours, including red, orange, yellow, white, and blue. As we have seen, stars are not all the same colour because they do not all have identical temperatures. To define *colour* precisely, astronomers have devised quantitative methods for characterizing the colour of a star and then using those colours to determine stellar temperatures. In the chapters that follow, we will provide the temperature of the stars we are describing, and this section tells you how those temperatures are determined from the colours of light the stars give off.

Sagittarius Star Cloud

Figure 11.1. This image, which was taken by the Hubble Space Telescope, shows stars in the direction toward the centre of the Milky Way Galaxy. The bright stars glitter like coloured jewels on a black velvet background. The colour of a star indicates its temperature. Blue-white stars are much hotter than the Sun, whereas red stars are cooler. On average, the stars in this field are at a distance of about 25,000 light-years (which means it takes light 25,000 years to traverse the distance from them to us) and the width of the field is about 13.3 light-years.
The Sagittarius Star Cloud: A Sky Full of Glittering Jewels by Hubble Heritage Team (AURA/STScI/NASA)), NASA Media License.

Blue colours dominate the visible light output of very hot stars (with much additional radiation in the ultraviolet). On the other hand, cool stars emit most of their visible light energy at red wavelengths (with more radiation coming off in the infrared) Table 11.1 shows star colours and their corresponding temperature. The colour of a star therefore provides a measure of its intrinsic or true surface temperature (apart from the effects of reddening by interstellar dust).

"Cosmic Latte" or the Average Colour of Stars

Figure 11.2. Cosmic latte or the average colour of the stars in the universe.
The Average Color of the Universe by Karl Glazebrook & Ivan Baldry (JHU), NASA Media License.

Colour does not depend on the distance to the object. This should be familiar to you from everyday experience. The colour of a traffic signal, for example, appears the same no matter how far away it is. If we could somehow take a star, observe it, and then move it much farther away, its apparent brightness (magnitude) would change. But this change in brightness is the same for all wavelengths, and so its colour would remain the same.

Table 11.1. Example Star Colours and Corresponding Approximate Temperatures

Star Colour	Approximate Temperature	Example
Blue	25,000 K	Spica
White	10,000 K	Vega
Yellow	6000 K	Sun
Orange	4000 K	Aldebaran
Red	3000 K	Betelgeuse

Go to this PhET interactive simulation from the University of Colorado to see the colour of a star changing as the temperature is changed. The link opens directly below this.

The hottest stars have temperatures of over 40,000 K, and the coolest stars have temperatures of about 2000 K. Our Sun's surface temperature is about 6000 K; its peak wavelength colour is a slightly greenish-yellow. In space, the Sun would look white, shining with about equal amounts of reddish and bluish wavelengths of light. It looks somewhat yellow as seen from Earth's surface because our planet's nitrogen molecules scatter some of the shorter (i.e., blue) wavelengths out of the beams of sunlight that reach us, leaving more long wavelength light behind. This also explains why the sky is blue: the blue sky is sunlight scattered by Earth's atmosphere.

Attribution

"17.2 Colours of Stars" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.2 CLASSIFICATION BY SPECTRA

When the spectra of different stars were first observed, astronomers found that they were not all identical. Since the dark lines are produced by the chemical elements present in the stars, astronomers first thought that the spectra differ from one another because stars are not all made of the same chemical elements. This hypothesis turned out to be wrong. *The primary reason that stellar spectra look different is because the stars have different temperatures*. Most stars have nearly the same composition as the Sun, with only a few exceptions.

Hydrogen, for example, is by far the most abundant element in most stars. However, lines of hydrogen are not seen in the spectra of the hottest and the coolest stars. In the atmospheres of the hottest stars, hydrogen atoms are completely ionized. Because the electron and the proton are separated, ionized hydrogen cannot produce absorption lines.

In the atmospheres of the coolest stars, hydrogen atoms have their electrons attached and can switch energy levels to produce lines. However, practically all of the hydrogen atoms are in the lowest energy state (unexcited) in these stars and thus can absorb only those photons able to lift an electron from that first energy level to a higher level. Photons with enough energy to do this lie in the ultraviolet part of the electromagnetic spectrum, and there are very few ultraviolet photons in the radiation from a cool star. What this means is that if you observe the spectrum of a very hot or very cool star with a typical telescope on the surface of Earth, the most common element in that star, hydrogen, will show very weak spectral lines or none at all.

The hydrogen lines in the visible part of the spectrum (called **Balmer lines**) are strongest in stars with intermediate temperatures—not too hot and not too cold. Calculations show that the optimum temperature for producing visible hydrogen lines is about 10,000 K. At this temperature, an appreciable number of hydrogen atoms are excited to the second energy level. They can then absorb additional photons, rise to still-higher levels of excitation, and produce a dark absorption line. Similarly, every other chemical element, in each of its possible stages of ionization, has a characteristic temperature at which it is most effective in producing absorption lines in any particular part of the spectrum.

Astronomers use the patterns of lines observed in stellar spectra to sort stars into a spectral class. Because a star's temperature determines which absorption lines are present in its spectrum, these spectral classes are a measure of its surface temperature. There are seven standard spectral classes. From hottest to coldest, these seven spectral classes are designated O, B, A, F, G, K, and M. Recently, astronomers have added three additional classes for even cooler objects—L, T, and Y.

At this point, you may be looking at these letters with wonder and asking yourself why astronomers didn't call the spectral types A, B, C, and so on. You will see, as we tell you the history, that it's an instance where tradition won out over common sense.

In the 1880s, Williamina Fleming devised a system to classify stars based on the strength of hydrogen absorption lines. Spectra with the strongest lines were classified as "A" stars, the next strongest "B," and so on down the alphabet to "O" stars, in which the hydrogen lines were very weak. But we saw above that hydrogen lines alone are not a good indicator for classifying stars, since their lines disappear from the visible light spectrum when the stars get too hot or too cold.

In the 1890s, Annie Jump Cannon revised this classification system, focusing on just a few letters from the original system: A, B, F, G, K, M, and O. Instead of starting over, Cannon also rearranged the existing classes—in order of decreasing temperature—into the sequence we have learned: O, B, A, F, G, K, M. She classified around 500,000 stars over her lifetime, classifying up to three stars per minute by looking at the stellar spectra.

For a deep dive into spectral types, explore the interactive project at the Sloan Digital Sky Survey in which you can practice classifying stars yourself. Direct link: http://cas.sdss.org/dr12/en/proj/advanced/spectraltypes/studentclasses.aspx

To help astronomers remember this crazy order of letters, Cannon created a mnemonic, "Oh Be A Fine Girl, Kiss Me." (If you prefer, you can easily substitute "Guy" for "Girl.") Other mnemonics, which we hope will not be relevant for you, include "Oh Brother, Astronomers Frequently Give Killer Midterms" and "Oh Boy, An F Grade Kills Me!" With the new L, T, and Y spectral classes, the mnemonic might be expanded to "Oh Be A Fine Girl (Guy), Kiss Me Like That, Yo!"

Each of these spectral classes, except possibly for the Y class which is still being defined, is further subdivided into 10 subclasses designated by the numbers 0 through 9. A B0 star is the hottest type of B star; a B9 star is the coolest type of B star and is only slightly hotter than an A0 star.

And just one more item of vocabulary: for historical reasons, astronomers call all the elements heavier than helium *metals*, even though most of them do not show metallic properties. (If you are getting annoyed at the peculiar jargon that astronomers use, just bear in mind that every field of human activity tends to develop its own specialized vocabulary. Just try reading a credit card or social media agreement form these days without training in law.)

Let's take a look at some of the details of how the spectra of the stars change with temperature. (It is these details that allowed Annie Cannon to identify the spectral types of stars as quickly as three per minute!) As Figure 11.3 shows, in the hottest O stars (those with temperatures over 28,000 K), only lines of ionized helium and highly ionized atoms of other elements are conspicuous. Hydrogen lines are strongest in A stars with atmospheric temperatures of about 10,000 K. Ionized metals provide the most conspicuous lines in stars with temperatures from 6000 to 7500 K (spectral type F). In the coolest M stars (below 3500 K), absorption bands of titanium oxide and other molecules are very strong. By the way, the spectral class assigned to the Sun is G2. The sequence of spectral classes is summarized in Table 11.2.

Absorption Lines in Stars of Different Temperatures

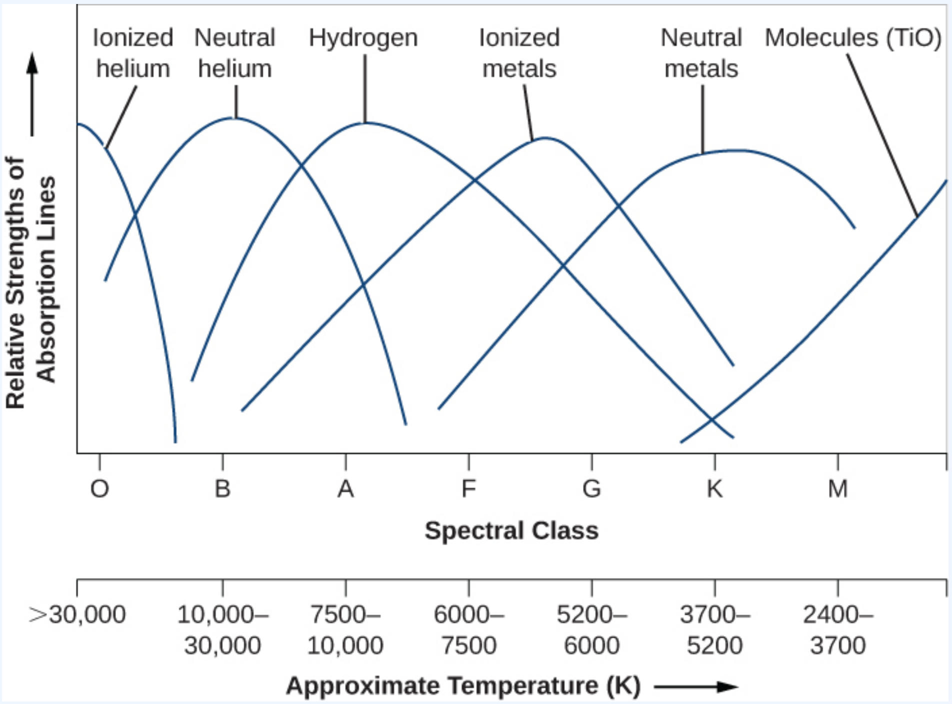


Figure 11.3. This graph details of how the spectra of the stars change with temperature.

Table 11.2. Spectral Classes for Stars

Spectral Class	Colour	Approximate Temperature (K)	Principal Features	Examples
O	Blue	> 30,000	Neutral and ionized helium lines, weak hydrogen lines	10 Lacertae
B	Blue-white	10,000–30,000	Neutral helium lines, strong hydrogen lines	Rigel, Spica
A	White	7500–10,000	Strongest hydrogen lines, weak ionized calcium lines, weak ionized metal (e.g., iron, magnesium) lines	Sirius, Vega
F	Yellow-white	6000–7500	Strong hydrogen lines, strong ionized calcium lines, weak sodium lines, many ionized metal lines	Canopus, Procyon
G	Yellow	5200–6000	Weaker hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of ionized and neutral metals	Sun, Capella
K	Orange	3700–5200	Very weak hydrogen lines, strong ionized calcium lines, strong sodium lines, many lines of neutral metals	Arcturus, Aldebaran
M	Red	2400–3700	Strong lines of neutral metals and molecular bands of titanium oxide dominate	Betelgeuse, Antares
L	Red	1300–2400	Metal hydride lines, alkali metal lines (e.g., sodium, potassium, rubidium)	Teide 1
T	Magenta	700–1300	Methane lines	Gliese 229B
Y	Infrared[1]	< 700	Ammonia lines	WISE 1828+2650

To see how spectral classification works, let's use Figure 11.3. Suppose you have a spectrum in which the hydrogen lines are about half as strong as those seen in an A star. Looking at the lines in our figure, you see that the star could be either a B star or a G star. But if the spectrum also contains helium lines, then it is a B star, whereas if it contains lines of ionized iron and other metals, it must be a G star.

Attribution

11.3 CLASSIFICATION BY MASS AND SIZE

How large can the mass of a star be? Stars more massive than the Sun are rare. Most stars have less mass than the Sun. The Sun, itself, is a **low-mass star**, as are all stars with mass less than 1.33 times the mass of the Sun.

None of the stars within 30 light-years of the Sun has a mass greater than four times that of the Sun. Those with mass between 1.33 and 4 times the mass of the Sun are **intermediate-mass stars**.

Searches at large distances from the Sun have led to the discovery of a few stars with masses up to about 100 times that of the Sun, and a handful of stars (a few out of several billion) may have masses as large as 250 solar masses. Any star with mass exceeding four times that of the Sun is categorized as **high-mass**.

Table 11.3 Spectral Class and Stellar Mass versus Stable Lifetime

Spectral Type	Mass (Mass of Sun = 1)	Stable Lifetime (years)
O5	40	1 million
B0	16	10 million
A0	3.3	500 million
F0	1.7	2.7 billion
G0	1.1	9 billion
K0	0.8	14 billion
M0	0.4	200 billion

The table above shows that the most massive stars only have a few million years of stability. A star of 1 solar mass remains stable for roughly 10 billion years, while a star of about 0.4 solar mass can be stable for some 200 billion years, which is longer than the current age of the universe. A star spends *most* of its total lifetime (an average of about 90% of the total) as a stable star.

As far as size (radius or diameter) is concerned:

- **Giant stars** have radii between 10 and 100 times that of the Sun
- **Dwarf stars** have radii equal to, or less than, that of the Sun
- **Supergiant stars** have radii more than 100 times than that of the Sun

The results of many stellar size measurements over the years have shown that most nearby stars are roughly the size of the Sun, with typical diametres of a million kilometres or so. Faint stars, as we might have expected, are generally smaller than more luminous stars. However, there are some dramatic exceptions to this simple generalization.

A few of the very luminous stars, those that are also red (indicating relatively low surface temperatures), turn out to be truly enormous. These stars are called, appropriately enough, giant stars or supergiant stars. An example is Betelgeuse, the second brightest star in the constellation of Orion and one of the dozen brightest stars in our sky. Its diameter, remarkably, is greater than 10 AU (1.5 *billion* kilometres), large enough to fill the entire inner solar system almost as far out as Jupiter.

Attribution

11.4 THE HERTZSPRUNG-RUSSELL (H-R) DIAGRAM

Table 11.4 summarizes some of the characteristics by which we might classify stars and how those are measured. We have looked at a few of these characteristics already in this chapter. When the characteristics of large numbers of stars were measured at the beginning of the twentieth century, astronomers were able to begin a deeper search for patterns and relationships in these data.

Table 11.4. Measuring the Characteristics of Stars

Characteristic	Technique
Surface temperature	1. Determine the colour (very rough). 2. Measure the spectrum and get the spectral type.
Chemical composition	Determine which lines are present in the spectrum.
Luminosity	Measure the apparent brightness and compensate for distance.
Radial velocity	Measure the Doppler shift in the spectrum.
Rotation	Measure the width of spectral lines.
Mass	Measure the period and radial velocity curves of spectroscopic binary stars.
Diameter	1. Measure the way a star's light is blocked by the Moon. 2. Measure the light curves and Doppler shifts for eclipsing binary stars.

To help understand what sorts of relationships might be found, let's look briefly at a range of data about human beings. If you want to understand humans by comparing and contrasting their characteristics—without assuming any previous knowledge of these strange creatures—you could try to determine which characteristics lead you in a fruitful direction. For example, you might plot the heights of a large sample of humans against their weights (which is a measure of their mass). Such a plot is shown in Figure 11.4 and it has some interesting features. In the way we have chosen to present our data, height increases upward, whereas weight increases to the left. Notice that humans are not randomly distributed in the graph. Most points fall along a sequence that goes from the upper left to the lower right.

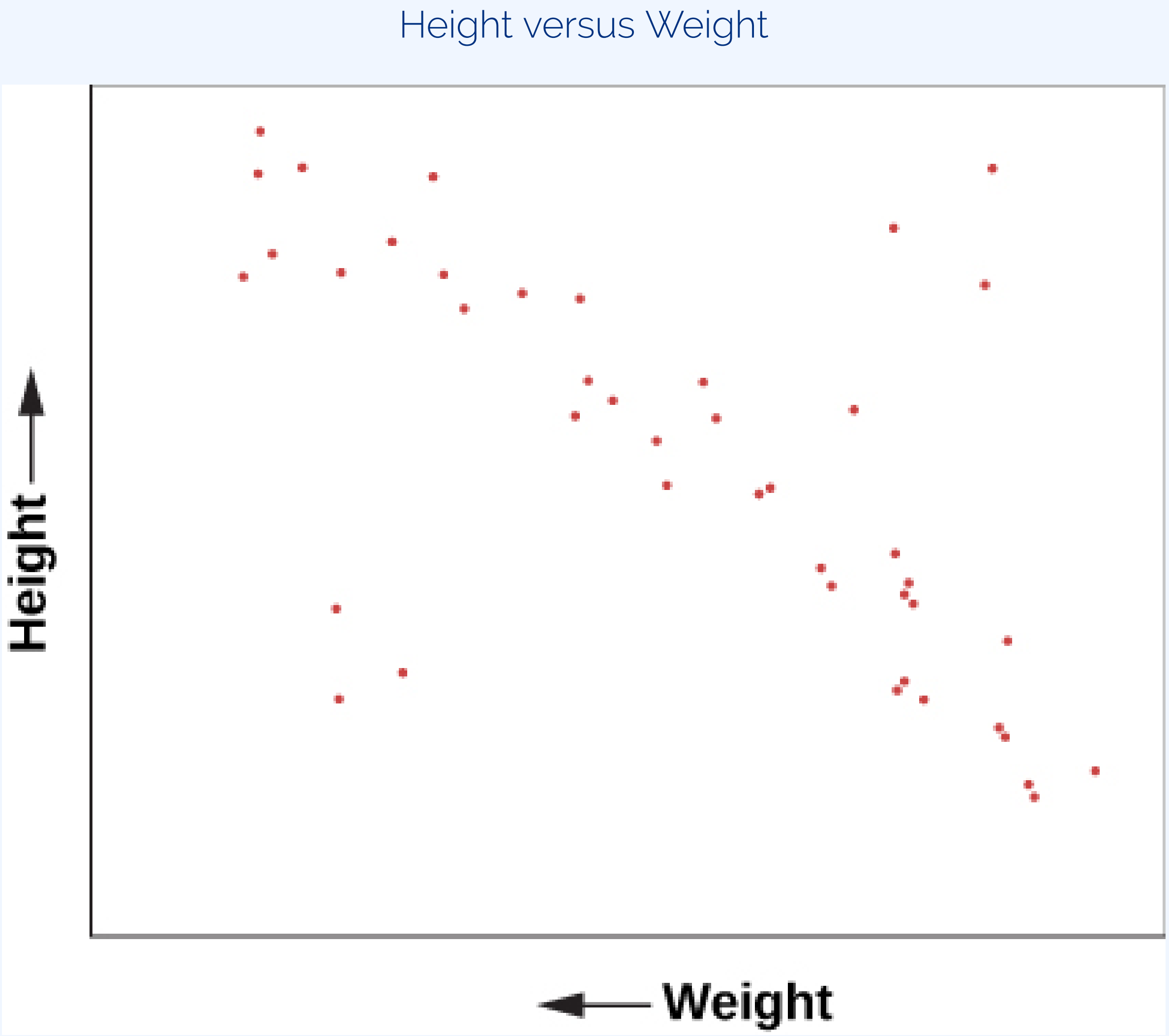


Figure 11.4. The plot of the heights and weights of a representative group of human beings. Most points lie along a “main sequence” representing most people, but there are a few exceptions.

We can conclude from this graph that human height and weight are related. Generally speaking, taller human beings weigh more, whereas shorter ones weigh less. This makes sense if you are familiar with the structure of human beings. Typically, if we have bigger bones, we have more flesh to fill out our larger frame. It’s not mathematically exact—there is a wide range of variation—but it’s not a bad overall rule. And, of course, there are some dramatic exceptions. You occasionally see a short human who is very overweight and would thus be more to the bottom left of our diagram than the average sequence of people. Or you might have a very tall, skinny fashion model with great height but relatively small weight, who would be found near the upper right.

A similar diagram has been found extremely useful for understanding the lives of stars. In 1913, American astronomer Henry Norris Russell plotted the luminosities of stars against their spectral classes (a way of denoting their surface temperatures). This investigation, and a similar independent study in 1911 by Danish astronomer Ejnar Hertzsprung, led to the extremely important discovery that the temperature and luminosity of stars are related.

Hertzsprung (1873–1967) and Russell (1877–1957)

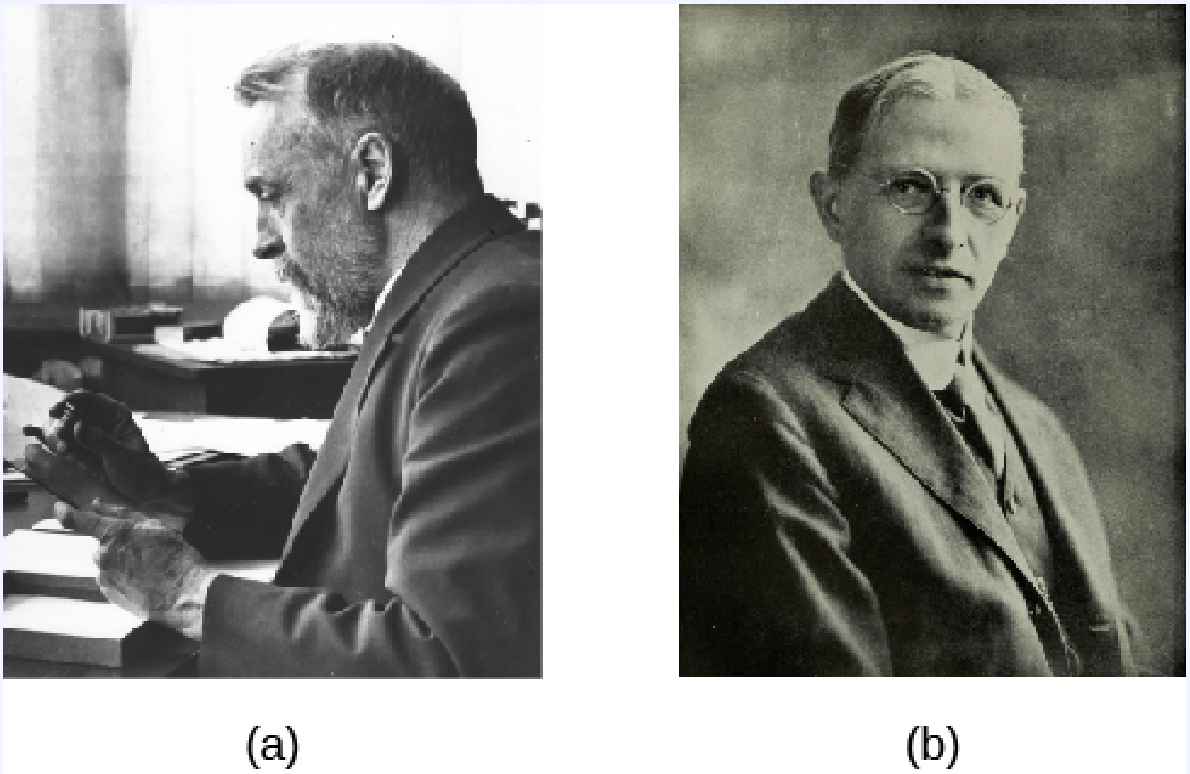


Figure 11.5. (a) Ejnar Hertzsprung and (b) Henry Norris Russell independently discovered the relationship between the luminosity and surface temperature of stars that is summarized in what is now called the H–R diagram.

Henry Norris Russell

When Henry Norris Russell graduated from Princeton University, his work had been so brilliant that the faculty decided to create a new level of honours degree beyond “summa cum laude” for him. His students later remembered him as a man whose thinking was three times faster than just about anybody else’s. His memory was so phenomenal, he could correctly quote an enormous number of poems and limericks, the entire Bible, tables of mathematical functions, and almost anything he had learned about astronomy. He was nervous, active, competitive, critical, and very articulate; he tended to dominate every meeting he attended. In outward appearance, he was an old-fashioned product of the nineteenth century who wore high-top black shoes and high starched collars, and carried an umbrella every day of his life. His 264 papers were enormously influential in many areas of astronomy.

Born in 1877, the son of a Presbyterian minister, Russell showed early promise. When he was 12, his family sent him to live with an aunt in Princeton so he could attend a top preparatory school. He lived in the same house in that town until his death in 1957 (interrupted only by a brief stay in Europe for graduate work). He was fond of recounting that both his mother and his maternal grandmother had won prizes in mathematics, and that he probably inherited his talents in that field from their side of the family.

Before Russell, American astronomers devoted themselves mainly to surveying the stars and making impressive catalogs of their properties, especially their spectra. Russell began to see that interpreting the spectra of stars required a much more sophisticated understanding of the physics of the atom, a subject that was being developed by European physicists in the 1910s and 1920s. Russell embarked on a lifelong quest to ascertain the physical conditions inside stars from the clues in their spectra; his work inspired, and was continued by, a generation of astronomers, many trained by Russell and his collaborators.

Russell also made important contributions in the study of binary stars and the measurement of star masses, the origin of the solar system, the atmospheres of planets, and the measurement of distances in astronomy, among other fields. He was an influential teacher and popularizer of astronomy, writing a column on astronomical topics for *Scientific American* magazine for more than 40 years. He and two colleagues wrote a textbook for college astronomy classes that helped train astronomers and astronomy enthusiasts over several decades. That book set the scene for the kind of textbook you are now reading, which not only lays out the facts of astronomy but also explains how they fit together. Russell gave lectures around the country, often emphasizing the importance of understanding modern physics in order to grasp what was happening in astronomy.

Harlow Shapley, director of the Harvard College Observatory, called Russell “the dean of American astronomers.” Russell was certainly regarded as the leader of the field for many years and was consulted on many astronomical problems by colleagues from around the world. Today, one of the highest recognitions that an astronomer can receive is an award from the American Astronomical Society called the Russell Prize, set up in his memory.

Following Hertzsprung and Russell, let us plot the temperature (or spectral class) of a selected group of nearby stars against their luminosity and see what we find. Such a plot is shown in Figure 11.6 and is frequently called the **Hertzsprung–Russell diagram**, abbreviated H–R diagram. It is one of the most important and widely used diagrams in astronomy, with applications that extend far beyond the purposes for which it was originally developed more than a century ago.

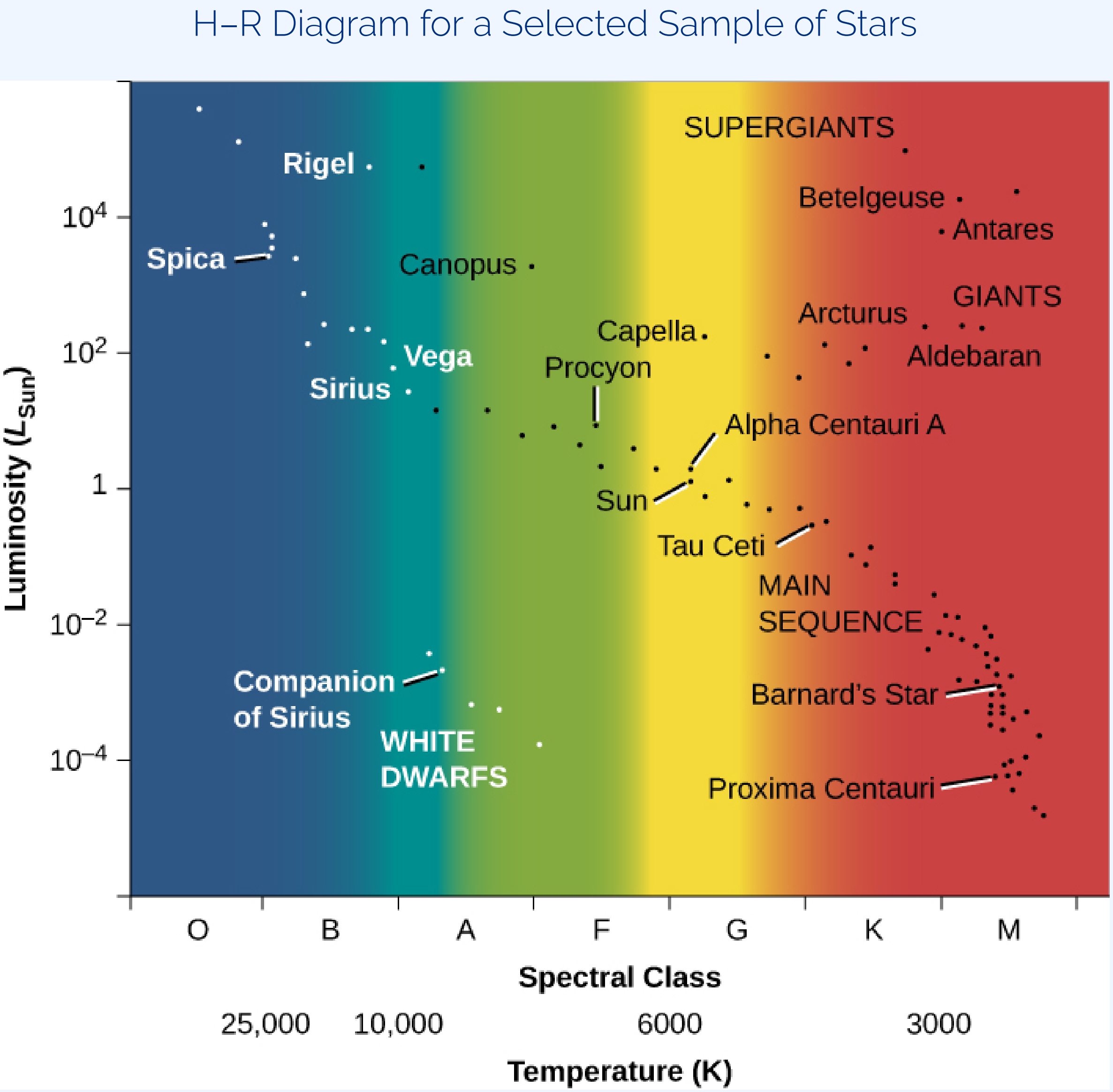


Figure 11.6. In such diagrams, luminosity is plotted along the vertical axis. Along the horizontal axis, we can plot either temperature or spectral type (also sometimes called spectral class). Several of the brightest stars are identified by name. Most stars fall on the main sequence.

It is customary to plot H–R diagrams in such a way that temperature increases toward the left and luminosity toward the top. Notice the similarity to our plot of height and weight for people that was shown in Figure 11.4. Stars, like people, are not distributed over the diagram at random, as they would be if they exhibited all combinations of luminosity and temperature. Instead, we see that the stars cluster into certain parts of the H–R diagram. The great majority are aligned along a narrow sequence running from the upper left (hot, highly luminous) to the lower right (cool, less luminous). This band of points is called the main sequence. It represents a relationship between *temperature* and *luminosity* that is followed by most stars. We can summarize this relationship by saying that hotter stars are more luminous than cooler ones.

A number of stars, however, lie above the main sequence on the H–R diagram, in the upper-right region, where stars have low temperature and high luminosity. How can a star be at once cool, meaning each square meter on the star does not put out all that much energy, and yet very luminous? The only way is for the star to be enormous—to have so many square metres on its surface that the *total* energy output is still large. These stars must be *giants* or *supergiants*, the stars of huge diameter we discussed earlier.

There are also some stars in the lower-left corner of the diagram, which have high temperature and low luminosity. If they have high surface temperatures, each square meter on that star puts out a lot of energy. How then can the overall star be dim? It must be that it has a very small total surface area; such stars are known as white dwarfs (white because, at these high temperatures, the colours of the electromagnetic radiation that they emit blend together to make them look bluish-white). We will say more about these puzzling objects in a moment. Figure 11.7. is a schematic H–R diagram for a large sample of stars, drawn to make the different types more apparent.

Schematic H–R Diagram for Many Stars

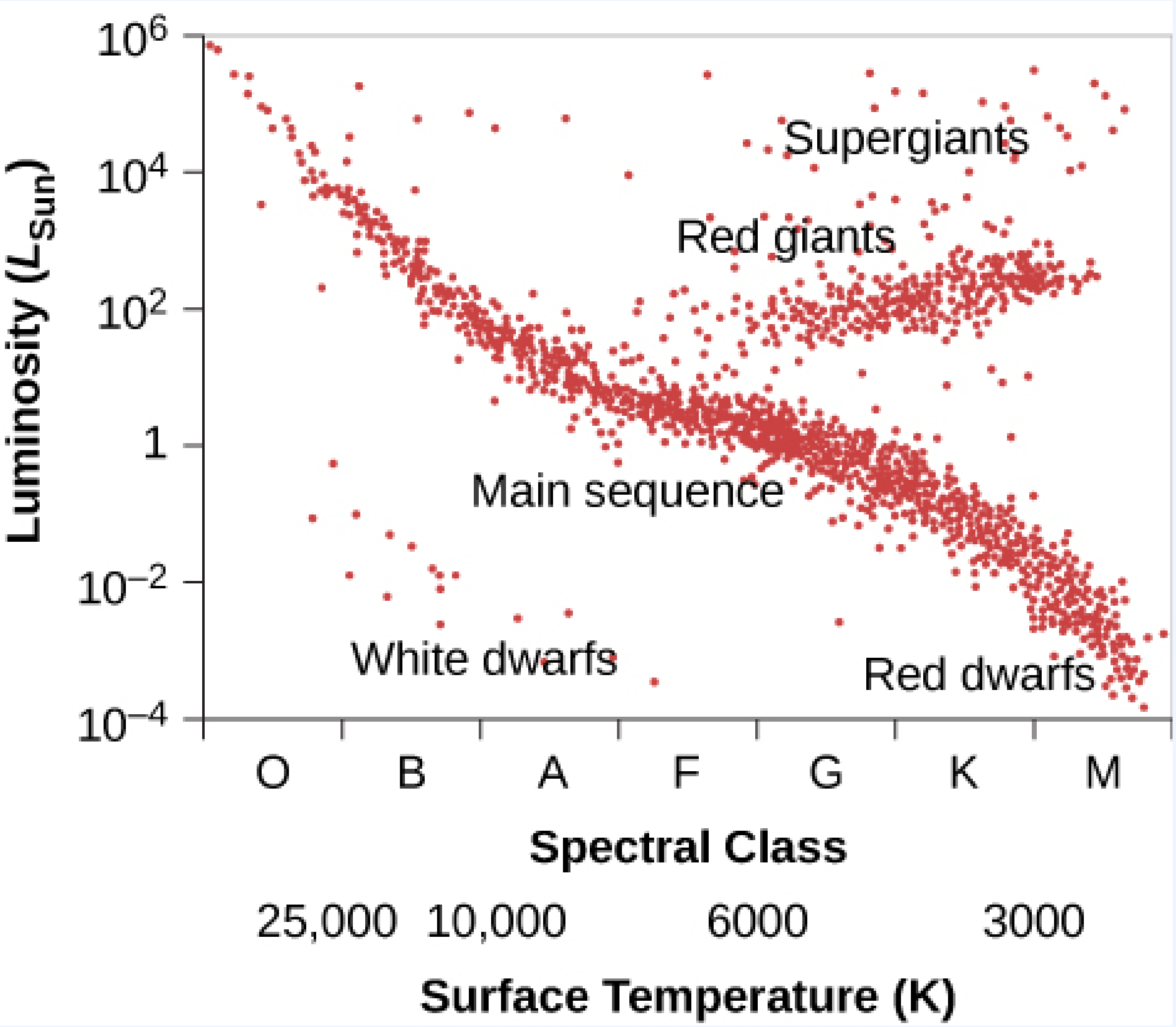


Figure 11.7. Ninety percent of all stars on such a diagram fall along a narrow band called the main sequence. A minority of stars are found in the upper right; they are both cool (and hence red) and bright, and must be giants. Some stars fall in the lower left of the diagram; they are both hot and dim, and must be white dwarfs.

Now, think back to our discussion of star surveys. It is difficult to plot an H–R diagram that is truly representative of all stars because most stars are so faint that we cannot see those outside our immediate neighbourhood. The stars plotted in Figure 11.6 were selected because their distances are known. This sample omits many intrinsically faint stars that are nearby but have not had their distances measured, so it shows fewer faint main-sequence stars than a “fair” diagram would. To be truly representative of the stellar population, an H–R diagram should be plotted for all stars within a certain distance. Unfortunately, our knowledge is reasonably complete only for stars within 10 to 20 light-years of the Sun, among which there are no giants or supergiants. Still, from many surveys (and more can now be done with new, more powerful telescopes), we estimate that about 90% of the true stars overall (excluding brown dwarfs) in our part of space are main-sequence stars, about 10% are white dwarfs, and fewer than 1% are giants or supergiants.

These estimates can be used directly to understand the lives of stars. Permit us another quick analogy with people. Suppose we survey people just like astronomers survey stars, but we want to focus our attention on the location of young people, ages 6 to 18 years. Survey teams fan out and take data about where such youngsters are found at all times during a 24-hour day. Some are found in the local pizza parlor, others are asleep at home, some are at the movies, and many are in school. After surveying a very large number of young people, one of the things that the teams determine is that, averaged over the course of the 24 hours, one-third of all youngsters are found in school.

How can they interpret this result? Does it mean that two-thirds of students are truants and the remaining one-third spend all their time in school? No, we must bear in mind that the survey teams counted youngsters throughout the full 24-hour day. Some survey teams worked at night, when most youngsters were at home asleep, and others worked in the late afternoon, when most youngsters were on their way home from school (and more likely to be enjoying a pizza). If the survey was truly representative, we *can* conclude, however, that if an average of one-third of all youngsters are found in school, then humans ages 6 to 18 years must spend about one-third of *their time* in school.

We can do something similar for stars. We find that, on average, 90% of all stars are located on the main sequence of the H–R diagram. If we can identify some activity or life stage with the main sequence, then it follows that stars must spend 90% of their lives in that activity or life stage.

Attribution

“17.3 The Spectra of Stars (and Brown Dwarfs)” from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.5 THE DEATH OF A SUN-LIKE STAR

Ant Nebula

Figure 11.8. During the later phases of stellar evolution, stars expel some of their mass, which returns to the interstellar medium to form new stars. This Hubble Space Telescope image shows a star losing mass. Known as Menzel 3, or the Ant Nebula, this beautiful region of expelled gas is about 3000 light-years away from the Sun. We see a central star that has ejected mass preferentially in two opposite directions. The object is about 1.6 light-years long. The image is colour coded—red corresponds to an emission line of sulfur, green to nitrogen, blue to hydrogen, and blue/violet to oxygen.
Ant Nebula by NASA, ESA and The Hubble Heritage Team (STScI/AURA)), ESA Standard License.

The Sun and other stars cannot last forever. Eventually they will exhaust their nuclear fuel and cease to shine. But how do they change during their long lifetimes? And what do these changes mean for the future of Earth?

We now turn from the birth of stars to the rest of their life stories. This is not an easy task since stars live much longer than astronomers. Thus, we cannot hope to see the life story of any single star unfold before our eyes or telescopes. To learn about their lives, we must survey as many of the stellar inhabitants of the Galaxy as possible. With thoroughness and a little luck, we can catch at least a few of them in each stage of their lives. As you've learned, stars have many different characteristics, with the differences sometimes resulting from their different masses, temperatures, and luminosities, and at other times derived from changes that occur as they age. Through a combination of observation and theory, we can use these differences to piece together the life story of a star.

One of the best ways to get a “snapshot” of a group of stars is by plotting their properties on an H–R diagram. We have already used the H–R diagram to follow the evolution of protostars up to the time they reach the main sequence. Now we’ll see what happens next.

Once a star has reached the main-sequence stage of its life, it derives its energy almost entirely from the conversion of hydrogen to helium via the process of nuclear fusion in its core . Since hydrogen is the most abundant element in stars, this process can maintain the star’s equilibrium for a long time. Thus, all stars remain on the main sequence for most of their lives. Some astronomers like to call the main-sequence phase the star’s “prolonged adolescence” or “adulthood” (continuing our analogy to the stages in a human life).

The left-hand edge of the main-sequence band in the H–R diagram is called the zero-age main sequence. We use the term ***zero-age*** to mark the time when a star stops contracting, settles onto the main sequence, and begins to fuse hydrogen in its core. The zero-age main sequence is a continuous line in the H–R diagram that shows where stars of different masses but similar chemical composition can be found when they begin to fuse hydrogen.

Since only 0.7% of the hydrogen used in fusion reactions is converted into energy, fusion does not change the *total* mass of the star appreciably during this long period. It does, however, change the chemical composition in its central regions where nuclear reactions occur: hydrogen is gradually depleted, and helium accumulates. This change of composition changes the luminosity, temperature, size, and interior structure of the star. When a star’s luminosity and temperature begin to change, the point that represents the star on the H–R diagram moves away from the zero-age main sequence.

Calculations show that the temperature and density in the inner region slowly increase as helium accumulates in the centre of a star. As the temperature gets hotter, each proton acquires more energy of motion on average; this means it is more likely to interact with other protons, and as a result, the rate of fusion also increases. For the proton-proton cycle, the rate of fusion goes up roughly as the temperature to the fourth power.

If the rate of fusion goes up, the rate at which energy is being generated also increases, and the luminosity of the star gradually rises. Initially, however, these changes are small, and stars remain within the main-sequence band on the H–R diagram for most of their lifetimes.

Example 11.1

Star Temperature and Rate of Fusion

If a star's temperature were to double, by what factor would its rate of fusion increase?

Solution
Since the rate of fusion (like temperature) goes up to the fourth power, it would increase by a factor of 2^4, or 16 times.

Exercise 11.1

If the rate of fusion of a star increased 256 times, by what factor would the temperature increase?

Solution
The temperature would increase by a factor of $256^{0.25}$ (that is, the 4th root of 256), or 4 times.

Lifetimes on the Main Sequence

How many years a star remains in the main-sequence band depends on its mass. You might think that a more massive star, having more fuel, would last longer, but it's not that simple. The lifetime of a star in a particular stage of evolution depends on how much nuclear fuel it has and on *how quickly* it uses up that fuel. (In the same way, how long people can keep spending money depends not only on how much money they have but also on how quickly they spend it. This is why many lottery winners who go on spending sprees quickly wind up poor again.) In the case of stars, more massive ones use up their fuel much more quickly than stars of low mass.

The reason massive stars are such spendthrifts is that, as we saw above, the rate of fusion depends *very* strongly on the star's core temperature. And what determines how hot a star's central regions get? It is the *mass* of the star—the weight of the overlying layers determines how high the pressure in the core must be: higher mass requires higher pressure to balance it. Higher pressure, in turn, is produced by higher temperature. The higher the temperature in the central regions, the faster the star races through its storehouse of central hydrogen. Although massive stars have more fuel, they burn it so prodigiously that their lifetimes are much shorter than those of their low-mass counterparts. You can also understand now why the most massive main-sequence stars are also the most luminous. Like new rock stars with their first platinum album, they spend their resources at an astounding rate.

Eventually, all the hydrogen in a star's core, where it is hot enough for fusion reactions, is used up. The core then contains only helium, "contaminated" by whatever small percentage of heavier elements the star had to begin with. The helium in the core can be thought of as the accumulated "ash" from the nuclear "burning" of hydrogen during the main-sequence stage.

Energy can no longer be generated by hydrogen fusion in the stellar core because the hydrogen is all gone and, as we will see, the fusion of helium requires much higher temperatures. Since the central temperature is not yet high enough to fuse helium, there is no nuclear energy source to supply heat to the central region of the star. The long period of stability now ends, gravity again takes over, and the core begins to contract. Once more, the star's energy is partially supplied by gravitational energy, in the way described by Kelvin and Helmholtz. As the star's core shrinks, the energy of the inward-falling material is converted to heat.

The heat generated in this way, like all heat, flows outward to where it is a bit cooler. In the process, the heat raises the temperature of a layer of hydrogen that spent the whole long main-sequence time just outside the core. Like an understudy waiting in the wings of a hit Broadway show for a chance at fame and glory, this hydrogen was almost (but not quite) hot enough to undergo fusion and take part in the main action that sustains the star. Now, the additional heat produced by the shrinking core puts this hydrogen "over the limit," and a shell of hydrogen nuclei just outside the core becomes hot enough for hydrogen fusion to begin.

New energy produced by fusion of this hydrogen now pours outward from this shell and begins to heat up layers of the star farther out, causing them to expand. Meanwhile, the helium core continues to contract, producing more heat right around it. This leads to more fusion in the shell of fresh hydrogen outside the core as shown in Figure 11.9. The additional fusion produces still more energy, which also flows out into the upper layer of the star.

Star Layers during and after the Main Sequence

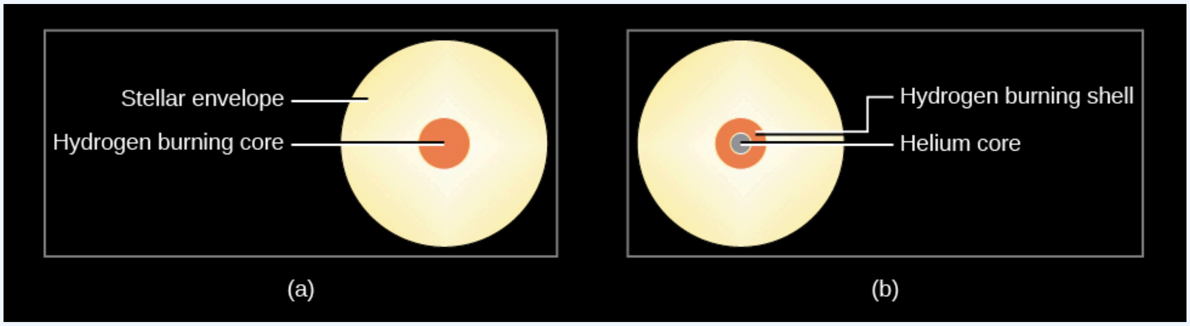


Figure 11.9. (a) During the main sequence, a star has a core where fusion takes place and a much larger envelope that is too cold for fusion. (b) When the hydrogen in the core is exhausted (made of helium, not hydrogen), the core is compressed by gravity and heats up. The additional heat starts hydrogen fusion in a layer just outside the core. Note that these parts of the Sun are not drawn to scale.

Most stars actually generate more energy each second when they are fusing hydrogen in the shell surrounding the helium core than they did when hydrogen fusion was confined to the central part of the star; thus, they increase in luminosity. With all the new energy pouring outward, the outer layers of the star begin to expand, and the star eventually grows and grows until it reaches enormous proportions as illustrated in Figure 11.10.

Relative Sizes of Stars

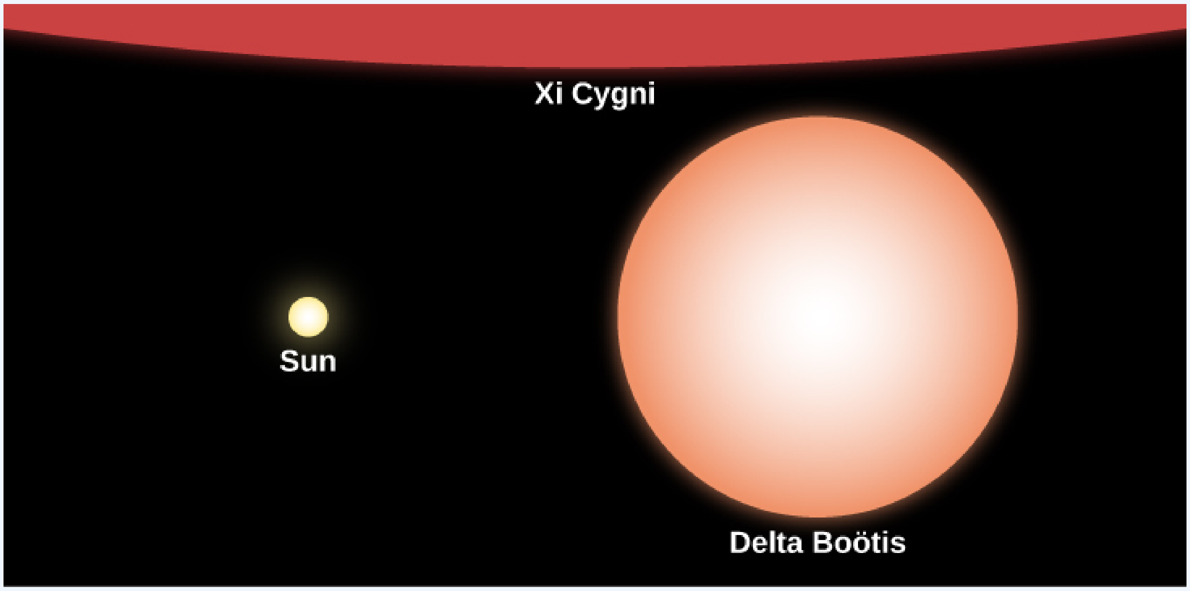


Figure 11.10. This image compares the size of the Sun to that of Delta Boötis, a giant star, and Xi Cygni, a supergiant. Note that Xi Cygni is so large in comparison to the other two stars that only a small portion of it is visible at the top of the frame.

When you take the lid off a pot of boiling water, the steam can expand and it cools down. In the same way, the expansion of a star's outer layers causes the temperature at the surface to decrease. As it cools, the star's overall colour becomes redder.

So the star becomes simultaneously more luminous and cooler. On the H–R diagram, the star therefore leaves the main-sequence band and moves upward (brighter) and to the right (cooler surface temperature). Over time, massive stars become red supergiants, and lower-mass stars like the Sun become red giants. You might also say that these stars have "split personalities": their cores are contracting while their outer layers are expanding. (Note that red giant stars do not actually look deep red; their colours are more like orange or orange-red.)

Just how different are these red giants and supergiants from a main-sequence star? Table 11.5 compares the Sun with the red supergiant Betelgeuse, which is visible above Orion's belt as the bright red star that marks the hunter's armpit. Relative to the Sun, this supergiant has a much larger radius, a much lower average density, a cooler surface, and a much hotter core.

Table 11.5 Comparing a Supergiant with the Sun

Property	Sun	Betelgeuse
Mass (2×10^{33} g)	1	16
Radius (km)	700,000	500,000,000
Surface temperature (K)	5,800	3,600
Core temperature (K)	15,000,000	160,000,000
Luminosity (4×10^{26} W)	1	46,000
Average density (g/cm^3)	1.4	1.3×10^{-7}
Age (millions of years)	4,500	10

Red giants can become so large that if we were to replace the Sun with one of them, its outer atmosphere would extend to the orbit of Mars or even beyond as shown in Figure 11.11. This is the next stage in the life of a star as it moves (to continue our analogy to human lives) from its long period of "youth" and "adulthood" to "old age." (After all, many human beings today also see their outer layers expand a bit as they get older.) By considering the relative ages of the Sun and Betelgeuse, we can also see that the idea that "bigger stars die faster" is indeed true here. Betelgeuse is a mere 10 million years old, which is relatively young compared with our Sun's 4.5 billion years, but it is already nearing its death throes as a red supergiant.

Betelgeuse

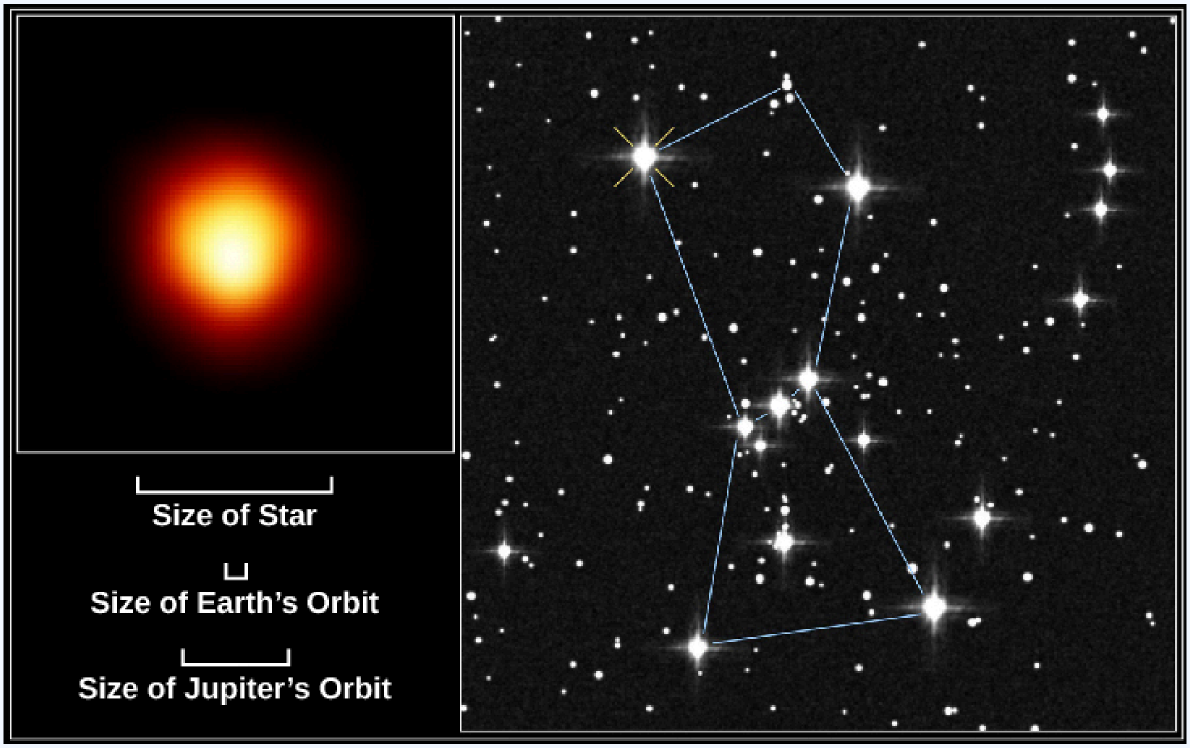


Figure 11.11. Betelgeuse is in the constellation Orion, the hunter; in the right image, it is marked with a yellow "X" near the top left. In the left image, we see it in ultraviolet with the Hubble Space Telescope, in the first direct image ever made of the surface of another star. As shown by the scale at the bottom, Betelgeuse has an extended atmosphere so large that, if it were at the center of our solar system, it would stretch past the orbit of Jupiter.
Atmosphere of Betelgeuse by Andrea Dupree (Harvard-Smithsonian CfA), Ronald Gilliland (STScI), NASA and ESA, ESA Standard License.

The "life story" we have related so far applies to almost all stars: each starts as a contracting protostar, then lives most of its life as a stable main-sequence star, and eventually moves off the main sequence toward the red-giant region.

As we have seen, the pace at which each star goes through these stages depends on its mass, with more massive stars evolving more quickly. But after this point, the life stories of stars of different masses diverge, with a wider range of possible behaviour according to their masses, their compositions, and the presence of any nearby companion stars.

Let's begin by considering stars with composition like that of the Sun and whose *initial* masses are comparatively low—no more than about twice the mass of our Sun. (Such mass may not seem too low, but stars with masses less than this all behave in a fairly similar fashion. We will see what happens to more massive stars in the next section.) Because there are much more low-mass stars than high-mass stars in the Milky Way, the vast majority of stars—including our Sun—follow the scenario we are about to relate. By the way, we carefully used the term **initial masses** of stars because, as we will see, stars can lose quite a bit of mass in the process of aging and dying.

Remember that red giants start out with a helium core where no energy generation is taking place, surrounded by a shell where hydrogen is undergoing fusion. The core, having no source of energy to oppose the inward pull of gravity, is shrinking and growing hotter. As time goes on, the temperature in the core can rise to much hotter values than it had in its main-sequence days. Once it reaches a temperature of 100 million K (but not before such point), three helium atoms can begin to fuse to form a single carbon nucleus. This process is called the **triple-alpha process**, so named because physicists call the nucleus of the helium atom an alpha particle.

When the triple-alpha process begins in low-mass (about 0.8 to 2.0 solar masses) stars, calculations show that the entire core is ignited in a quick burst of fusion called a helium flash. (More massive stars also ignite helium but more gradually and not with a flash.) As soon as the temperature at the centre of the star becomes high enough to start the triple-alpha process, the extra energy released is transmitted quickly through the entire helium core, producing very rapid heating. The heating speeds up the nuclear reactions, which provide more heating, and which accelerates the nuclear reactions even more. We have runaway generation of energy, which reignites the entire helium core in a flash.

You might wonder why the next major step in nuclear fusion in stars involves three helium nuclei and not just two. Although it is a lot easier to get two helium nuclei to collide, the product of this collision is not stable and falls apart very quickly. It takes three helium nuclei coming together *simultaneously* to make a stable nuclear structure. Given that each helium nucleus has two positive protons and that such protons repel one another, you can begin to see the problem. It takes a temperature of 100 million K to slam three helium nuclei (six protons) together and make them stick. But when that happens, the star produces a carbon nucleus.

Stars in Your Little Finger

Stop reading for a moment and look at your little finger. It's full of carbon atoms because carbon is a fundamental chemical building block for life on Earth. Each of those carbon atoms was once inside a red giant star and was fused from helium nuclei in the triple-alpha process. All the carbon on Earth—in you, in the charcoal you use for barbecuing, and in the diamonds you might exchange with a loved one—was "cooked up" by previous generations of stars. How the carbon atoms (and other elements) made their way from inside some of those stars to become part of

Earth is something we will discuss in the next chapter. For now, we want to emphasize that our description of stellar evolution is, in a very real sense, the story of our own cosmic “roots”—the history of how our own atoms originated among the stars. We are made of “star-stuff.”

After the helium flash, the star, having survived the “energy crisis” that followed the end of the main-sequence stage and the exhaustion of the hydrogen fuel at its centre, finds its balance again. As the star readjusts to the release of energy from the triple-alpha process in its core, its internal structure changes once more: its surface temperature increases and its overall luminosity decreases. The star then continues to fuse the helium in its core for a while, returning to the kind of equilibrium between pressure and gravity that characterized the main-sequence stage. During this time, a newly formed carbon nucleus at the centre of the star can sometimes be joined by another helium nucleus to produce a nucleus of oxygen — another building block of life.

However, at a temperature of 100 million K, the inner core is converting its helium fuel to carbon (and a bit of oxygen) at a rapid rate. Thus, the new period of stability cannot last very long: it is far shorter than the main-sequence stage. Soon, all the helium hot enough for fusion will be used up, just like the hot hydrogen that was used up earlier in the star’s evolution. Once again, the inner core will not be able to generate energy via fusion. Once more, gravity will take over, and the core will start to shrink again. We can think of stellar evolution as a story of a constant struggle against gravitational collapse. A star can avoid collapsing as long as it can tap energy sources, but once any particular fuel is used up, it starts to collapse again.

The star’s situation is analogous to the end of the main-sequence stage (when the central hydrogen got used up), but the star now has a somewhat more complicated structure. Again, the star’s core begins to collapse under its own weight. Heat released by the shrinking of the carbon and oxygen core flows into a shell of helium just above the core. This helium, which had not been hot enough for fusion into carbon earlier, is heated just enough for fusion to begin and to generate a new flow of energy.

Farther out in the star, there is also a shell where fresh hydrogen has been heated enough to fuse helium. The star now has a multi-layered structure like an onion: a carbon-oxygen core, surrounded by a shell of helium fusion, a layer of helium, a shell of hydrogen fusion, and finally, the extended outer layers of the star as shown in Figure 11.12. As energy flows outward from the two fusion shells, once again the outer regions of the star begin to expand. Its brief period of stability is over; the star moves back to the red-giant domain on the H–R diagram for a short time. But this is a brief and final burst of glory.

Layers inside a Low-Mass Star before Death

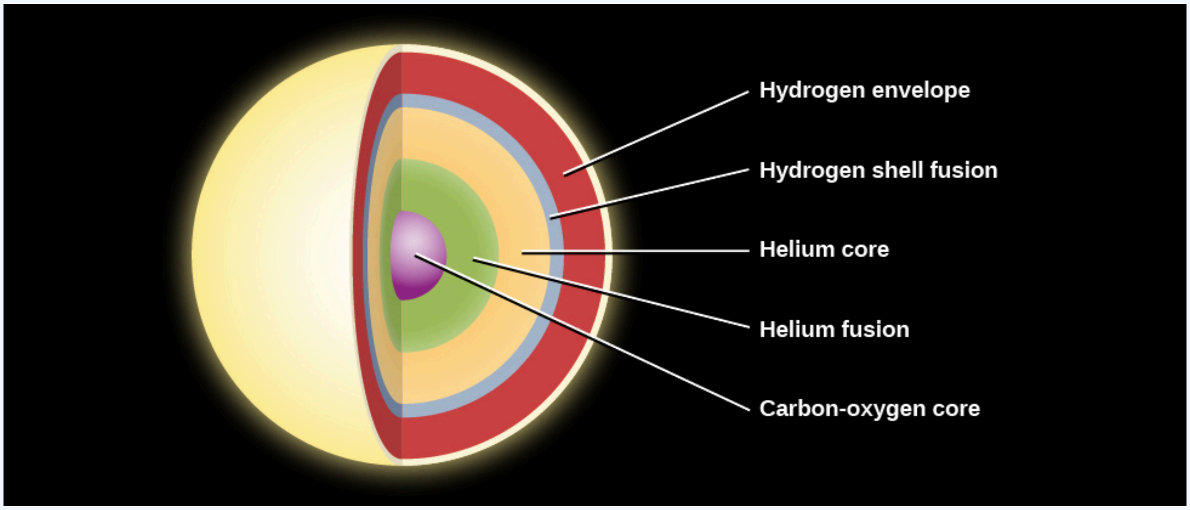


Figure 11.12. Here we see the layers inside a star with an initial mass that is less than twice the mass of the Sun. These include, from the centre outward, the carbon-oxygen core, a layer of helium hot enough to fuse, a layer of cooler helium, a layer of hydrogen hot enough to fuse, and then cooler hydrogen beyond.

Recall that the last time the star was in this predicament, helium fusion came to its rescue. The temperature at the star's centre eventually became hot enough for the *product* of the previous step of fusion (helium) to become the *fuel* for the next step (helium fusing into carbon). But the step after the fusion of helium nuclei requires a temperature so hot that the kinds of lower-mass stars (less than 2 solar masses) we are discussing simply cannot compress their cores to reach it. No further types of fusion are possible for such a star.

In a star with a mass similar to that of the Sun, the formation of a carbon-oxygen core thus marks the end of the generation of nuclear energy at the centre of the star. The star must now confront the fact that its death is near. Table 11.6 summarizes the stages discussed so far in the life of a star with the same mass as that of the Sun. One thing that gives us confidence in our calculations of stellar evolution is that when we make H–R diagrams of older clusters, we actually see stars in each of the stages that we have been discussing.

Table 11.6. The Evolution of a Star with the Sun's Mass

Stage	Time in This Stage (years)	Surface Temperature (K)	Luminosity (L_{Sun})	Diameter (Sun = 1)
Main sequence	11 billion	6000	1	1
Becomes red giant	1.3 billion	3100 at minimum	2300 at maximum	165
Helium fusion	100 million	4800	50	10
Giant again	20 million	3100	5200	180

When stars swell up to become red giants, they have very large radii and therefore a low escape velocity. Radiation pressure, stellar pulsations, and violent events like the helium flash can all drive atoms in the outer atmosphere away from the star, and cause it to lose a substantial fraction of its mass into space. Astronomers estimate that by the time a star like the Sun reaches the point of the helium flash, for example, it will have lost as much as 25% of its mass. And it can lose still more mass when it ascends the red-giant branch for the second time. As a result, aging stars are surrounded by one or more expanding shells of gas, each containing as much as 10–20% of the Sun's mass (or 0.1–0.2 M_{Sun}).

When nuclear energy generation in the carbon-oxygen core ceases, the star's core begins to shrink again and to heat up as it gets more and more compressed. (Remember that this compression will not be halted by another type of fusion in these low-mass stars.) The whole star follows along, shrinking and also becoming very hot—reaching surface temperatures as high as 100,000 K. Such hot stars are very strong sources of stellar winds and ultraviolet radiation, which sweep outward into the shells of material ejected when the star was a red giant. The winds and the ultraviolet radiation heat the shells, ionize them, and set them aglow (just as ultraviolet radiation from hot, young stars produces H II regions).

The result is the creation of some of the most beautiful objects in the cosmos as seen on the gallery of Figure 11.13. These objects were given an extremely misleading name when first found in the eighteenth century: planetary nebulae. The name is derived from the fact that a few planetary nebulae, when viewed through a small telescope, have a round shape bearing a superficial resemblance to planets. Actually, they have nothing to do with planets, but once names are put into regular use in astronomy, it is extremely difficult to change them. There are tens of thousands of planetary nebulae in our own Galaxy, although many are hidden from view because their light is absorbed by interstellar dust.

Gallery of Planetary Nebulae

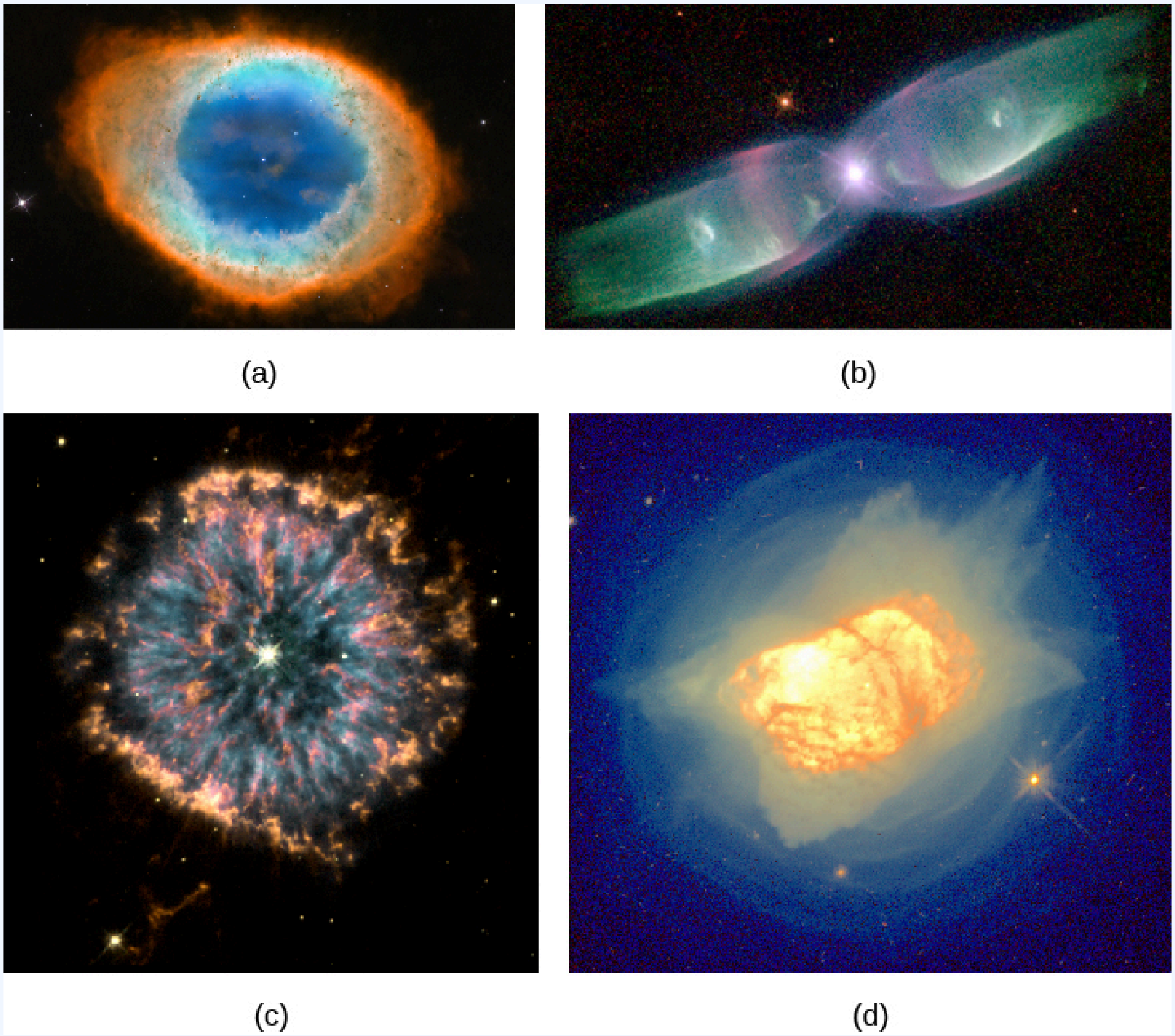


Figure 11.13. This series of beautiful images depicting some intriguing planetary nebulae highlights the capabilities of the Hubble Space Telescope. (a) Perhaps the best known planetary nebula is the Ring Nebula (M57), located about 2000 light-years away in the constellation of Lyra. The ring is about 1 light-year in diameter, and the central star has a temperature of about 120,000 °C. Careful study of this image has shown scientists that, instead of looking at a spherical shell around this dying star, we may be looking down the barrel of a tube or cone. The blue region shows emission from very hot helium, which is located very close to the star; the red region isolates emission from ionized nitrogen, which is radiated by the coolest gas farthest from the star; and the green region represents oxygen emission, which is produced at intermediate temperatures and is at an intermediate distance from the star. (b) This planetary nebula, M2-9, is an example of a butterfly nebula. The central star (which is part of a binary system) has ejected mass preferentially in two opposite directions. In other images, a disk, perpendicular to the two long streams of gas, can be seen around the two stars in the middle. The stellar outburst that resulted in the expulsion of matter occurred about 1200 years ago. Neutral oxygen is shown in red, once-ionized nitrogen in green, and twice-ionized oxygen in blue. The planetary nebula is about 2100 light-years away in the constellation of Ophiuchus. (c) In this image of the planetary nebula NGC 6751, the blue regions mark the hottest gas, which forms a ring around the central star. The orange and red regions show the locations

of cooler gas. The origin of these cool streamers is not known, but their shapes indicate that they are affected by radiation and stellar winds from the hot star at the centre. The temperature of the star is about 140,000 °C. The diameter of the nebula is about 600 times larger than the diameter of our solar system. The nebula is about 6500 light-years away in the constellation of Aquila. (d) This image of the planetary nebula NGC 7027 shows several stages of mass loss. The faint blue concentric shells surrounding the central region identify the mass that was shed slowly from the surface of the star when it became a red giant. Somewhat later, the remaining outer layers were ejected but not in a spherically symmetric way. The dense clouds formed by this late ejection produce the bright inner regions. The hot central star can be seen faintly near the centre of the nebulosity. NGC 7027 is about 3000 light-years away in the direction of the constellation of Cygnus.

Credit a: Messier 57 (The Ring Nebula) by NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, NASA Media License.
Credit b: Planetary Nebula M-29 by Bruce Balick (University of Washington), Vincent Icke (Leiden University, The Netherlands), Garrelt Mellema (Stockholm University), and NASA, NASA Media License.
Credit c: The Glowing Eye of Planetary Nebula NGC 6751 by NASA, The Hubble Heritage Team (STScI/AURA), NASA Media License.
Credit d: The Stellar Death Process: Planetary Nebula NGC 7027 by H. Bond (STScI) and NASA, NASA Media License.

As shows, sometimes a planetary nebula appears to be a simple ring. Others have faint shells surrounding the bright ring, which is evidence that there were multiple episodes of mass loss when the star was a red giant (see image (d)). In a few cases, we see two lobes of matter flowing in opposite directions. Many astronomers think that a considerable number of planetary nebulae basically consist of the same structure, but that the shape we see depends on the viewing angle as seen on Figure 11.14. According to this idea, the dying star is surrounded by a very dense, doughnut-shaped disk of gas. (Theorists do not yet have a definite explanation for why the dying star should produce this ring, but many believe that binary stars, which are common, are involved.)

Model to Explain the Different Shapes of Planetary Nebulae

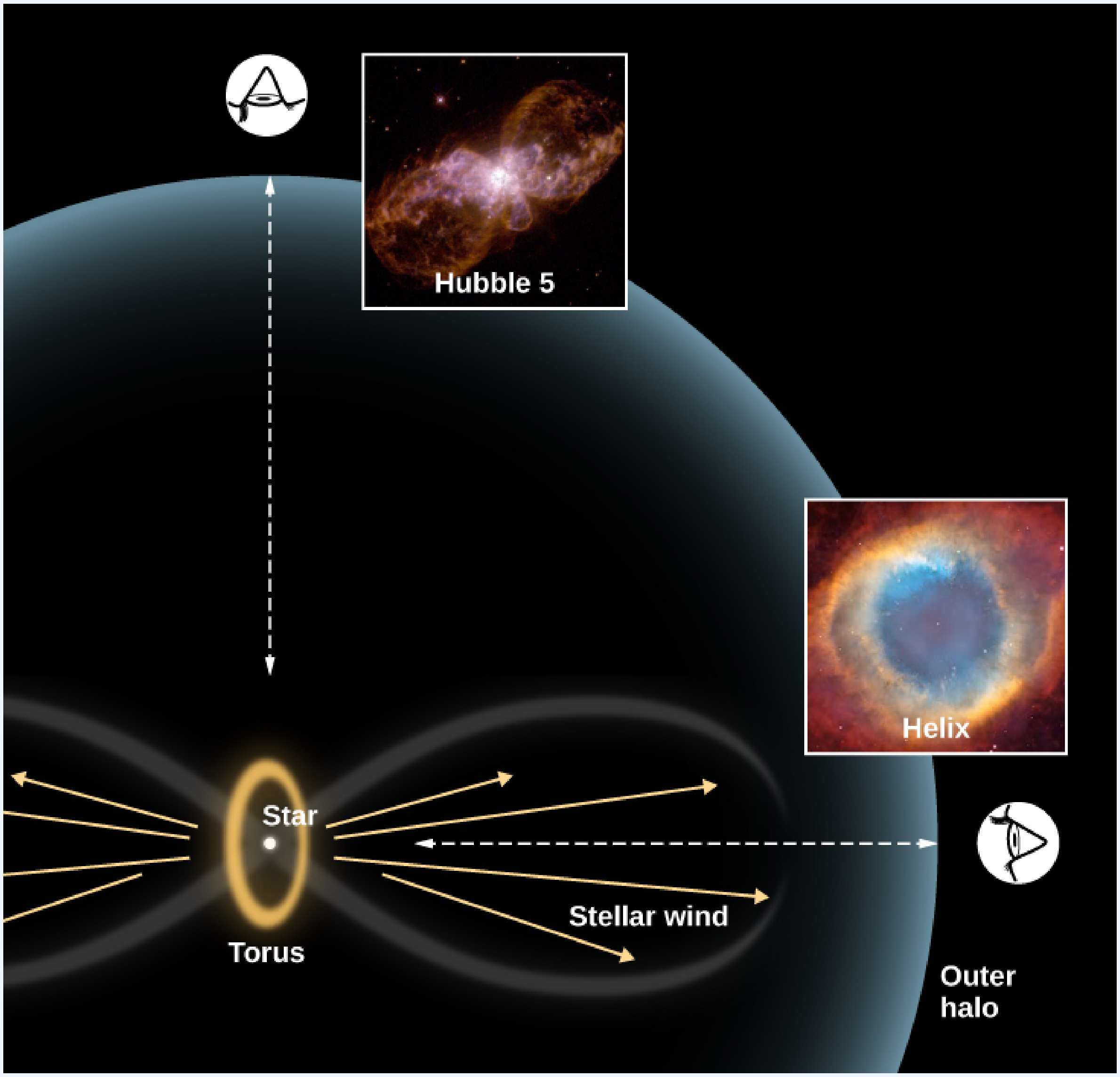


Figure 11.14. The range of different shapes that we see among planetary nebulae may, in many cases, arise from the same geometric shape, but seen from a variety of viewing directions. The basic shape is a hot central star surrounded by a thick torus (or doughnut-shaped disk) of gas. The star's wind cannot flow out into space very easily in the direction of the torus, but can escape more freely in the two directions perpendicular to it. If we view the nebula along the direction of the flow (Helix Nebula), it will appear nearly circular (like looking directly down into an empty ice-cream cone). If we look along the equator of the torus, we see both outflows and a very elongated shape (Hubble 5). Current research on planetary nebulae focuses on the reasons for having a torus around the star in the first place. Many astronomers suggest that the basic cause may be that many of the central stars are actually close binary stars, rather than single stars.

A view of Hubble 5 by Bruce Balick (University of Washington), Vincent Icke (Leiden University, The Netherlands), Garrelt Mellema (Stockholm University), and NASA/ESA,ESA Standard License.
Helix Nebula: HST/CTIO Image by NASA, ESA, C.R. O'Dell (Vanderbilt University), and M. Meixner, P.

McCullough), ESA Standard License.

As the star continues to lose mass, any less dense gas that leaves the star cannot penetrate the torus, but the gas *can* flow outward in directions perpendicular to the disk. If we look perpendicular to the direction of outflow, we see the disk and both of the outward flows. If we look "down the barrel" and into the flows, we see a ring. At intermediate angles, we may see wonderfully complex structures. Compare the viewpoints in the two images above.

Planetary nebula shells usually expand at speeds of 20–30 km/s, and a typical planetary nebula has a diameter of about 1 light-year. If we assume that the gas shell has expanded at a constant speed, we can calculate that the shells of all the planetary nebulae visible to us were ejected within the past 50,000 years at most. After this amount of time, the shells have expanded so much that they are too thin and tenuous to be seen. That's a pretty short time that each planetary nebula can be observed (when compared to the whole lifetime of the star). Given the number of such nebulae we nevertheless see, we must conclude that a large fraction of all stars evolve through the planetary nebula phase. Since we saw that low-mass stars are much more common than high-mass stars, this confirms our view of planetary nebulae as sort of "last gasp" of low-mass star evolution.

Attribution

"22.0 Thinking Ahead" and "22.1 Evolution from the Main Sequence to Red Giants" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.6 WHITE DWARFS

In the last section, we left the life story of a star with a mass like the Sun's just after it had climbed up to the red-giant region of the H–R diagram for a second time and had shed some of its outer layers to form a planetary nebula. Recall that during this time, the *core* of the star was undergoing an "energy crisis." Earlier in its life, during a brief stable period, helium in the core had gotten hot enough to fuse into carbon (and oxygen). But after this helium was exhausted, the star's core had once more found itself without a source of pressure to balance gravity and so had begun to contract.

This collapse is the final event in the life of the core. Because the star's mass is relatively low, it cannot push its core temperature high enough to begin another round of fusion (in the same way larger-mass stars can). The core continues to shrink until it reaches a density equal to nearly a million times the density of water! That is 200,000 times greater than the average density of Earth. At this extreme density, a new and different way for matter to behave kicks in and helps the star achieve a final state of equilibrium. In the process, what remains of the star becomes a **white dwarf**.

Because white dwarfs are far denser than any substance on Earth, the matter inside them behaves in a very unusual way—unlike anything we know from everyday experience. At this high density, gravity is incredibly strong and tries to shrink the star still further, but all the *electrons* resist being pushed closer together and set up a powerful pressure inside the core. This pressure is the result of the fundamental rules that govern the behaviour of electrons. According to these rules (known to physicists as the Pauli exclusion principle), which have been verified in studies of atoms in the laboratory, no two electrons can be in the same place at the same time doing the same thing. We specify the *place* of an electron by its position in space, and we specify what it is doing by its motion and the way it is spinning.

The temperature in the interior of a star is always so high that the atoms are stripped of virtually all their electrons. For most of a star's life, the density of matter is also relatively low, and the electrons in the star are moving rapidly. This means that no two of them will be in the same place moving in exactly the same way at the same time. But this all changes when a star exhausts its store of nuclear energy and begins its final collapse.

As the star's core contracts, electrons are squeezed closer and closer together. Eventually, a star like the Sun becomes so dense that further contraction would in fact require two or more electrons to violate the rule against occupying the same place and moving in the same way. Such a dense gas is said to be degenerate (a term coined by physicists and not related to the electron's moral character). The electrons in a **degenerate gas** resist further crowding with tremendous pressure. (It's as if the electrons said, "You can press inward all you want, but there is simply no room for any other electrons to squeeze in here without violating the rules of our existence.")

The degenerate electrons do not require an input of heat to maintain the pressure they exert, and so a star with this kind of structure, if nothing disturbs it, can last essentially forever. (Note that the repulsive force between degenerate electrons is different from, and much stronger than, the normal electrical repulsion between charges that have the same sign.)

The electrons in a degenerate gas do move about, as do particles in any gas, but not with a lot of freedom. A particular electron cannot change position or momentum until another electron in an adjacent stage gets out of the way. The situation is much like that in the parking lot after a big football game. Vehicles are closely packed, and a given car cannot move until the one in front of it moves, leaving an empty space to be filled.

Of course, the dying star also has atomic nuclei in it, not just electrons, but it turns out that the nuclei must be squeezed to much higher densities before their quantum nature becomes apparent. As a result, in white dwarfs, the nuclei do not exhibit degeneracy pressure. Hence, in the white dwarf stage of stellar evolution, it is the degeneracy pressure of the electrons, and not of the nuclei, that halts the collapse of the core.

White dwarfs, then, are stable, compact objects with electron-degenerate cores that cannot contract any further. Calculations showing that white dwarfs are the likely end state of low-mass stars were first carried out by the Indian-American astrophysicist Subrahmanyan Chandrasekhar. He was able to show how much a star will shrink before the degenerate electrons halt its further contraction and hence what its final diameter will be.

When Chandrasekhar made his calculation about white dwarfs, he found something very surprising: the radius of a white dwarf shrinks as the mass in the star increases (the larger the mass, the more tightly packed the electrons can become, resulting in a smaller radius). According to the best theoretical models, a white dwarf with a mass of about 1.4 M_{Sun} or larger would have a radius of zero. What the calculations are telling us is that even the force of degenerate electrons cannot stop the collapse of a star with more mass than this. The maximum mass that a star can end its life with and still become a white dwarf—1.4 M_{Sun}—is called the **Chandrasekhar limit**. Stars with end-of-life masses that exceed this limit have a different kind of end in store—one that we will explore in the next section.

Subrahmanyan Chandrasekhar

Born in 1910 in Lahore, India, Subrahmanyan Chandrasekhar (known as Chandra to his friends and colleagues) grew up in a home that encouraged scholarship and an interest in science. His uncle, C. V. Raman, was a physicist who won the 1930 Nobel Prize. A precocious student, Chandra tried to read as much as he could about the latest ideas in physics and astronomy, although obtaining technical books was not easy in India at the time. He finished college at age 19 and won a scholarship to study in England. It was during the long boat voyage to get to graduate school that he first began doing calculations about the structure of white dwarf stars.

Chandra developed his ideas during and after his studies as a graduate student, showing—as we have discussed—that white dwarfs with masses greater than 1.4 times the mass of the Sun cannot exist and that the theory predicts the existence of other kinds of stellar corpses. He wrote later that he felt very shy and lonely during this period, isolated from students, afraid to assert himself, and sometimes waiting for hours to speak with some of the famous professors he had read about in India. His calculations soon brought him into conflict with certain distinguished astronomers, including Sir Arthur Eddington, who publicly ridiculed Chandra's ideas. At a number of meetings of astronomers, such leaders in the field as Henry Norris Russell refused to give Chandra the opportunity to defend his ideas, while allowing his more senior critics lots of time to criticize them.

Yet Chandra persevered, writing books and articles elucidating his theories, which turned out not only to be correct, but to lay the foundation for much of our modern understanding of the death of stars. In 1983, he received the Nobel Prize in physics for this early work.

In 1937, Chandra came to the United States and joined the faculty at the University of Chicago, where he remained for the rest of his life. There he devoted himself to research and teaching, making major contributions to many fields of astronomy, from our understanding of the motions of stars through the Galaxy to the behaviour of the bizarre objects called black holes. In 1999, NASA named its sophisticated orbiting X-ray telescope (designed in part to explore such stellar corpses) the Chandra X-ray Observatory.

S. Chandrasekhar (1910–1995)

Figure 11.15 Chandra’s research provided the basis for much of what we now know about stellar corpses.
Portrait of a young Chandrasekhar by American Institute of Physics, Public Domain

Chandra spent a great deal of time with his graduate students, supervising the research of more than 50 PhDs during his life. He took his teaching responsibilities very seriously: during the 1940s, while based at the Yerkes Observatory, he willingly drove the more than 100-mile trip to the university each week to teach a class of only a few students.

Chandra also had a deep devotion to music, art, and philosophy, writing articles and books about the relationship between the humanities and science. He once wrote that “one can learn science the way one enjoys music or art. . . . Heisenberg had a marvelous phrase ‘shuddering before the beautiful’. . . that is the kind of feeling I have.”

Using the Hubble Space Telescope, astronomers were able to detect images of faint white dwarf stars and other “stellar corpses” in the M4 star cluster, located about 7200 light-years away. Direct link to the Hubble web site: https://hubblesite.org/news/news-releases . Direct link to images of white dwarf stars: https://sci.esa.int/web/hubble/-/55902-hubble-traces-the-migration-of-white-dwarfs-in-cluster-47-tucanae-heic1510 .

Since a stable white dwarf can no longer contract or produce energy through fusion, its only energy source is the heat represented by the motions of the atomic nuclei in its interior. The light it emits comes from this internal stored heat, which is substantial. Gradually, however, the white dwarf radiates away all its heat into space. After many billions of years, the nuclei will be moving much more slowly, and the white dwarf will no longer shine as shown in Figure 11.16. It will then be a **black dwarf**—a cold stellar corpse with the mass of a star and the size of a planet. It will be composed mostly of carbon, oxygen, and neon, the products of the most advanced fusion reactions of which the star was capable.

Visible Light and X-Ray Images of the Sirius Star System

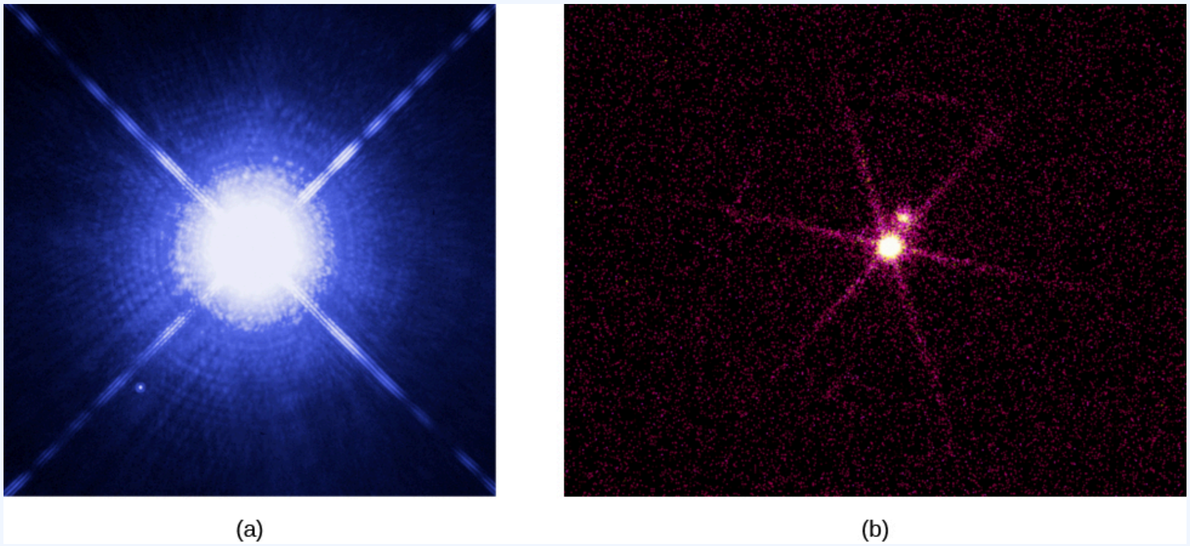


Figure 11.16. (a) This image taken by the Hubble Space Telescope shows Sirius A (the large bright star), and its companion star, the white dwarf known as Sirius B (the tiny, faint star at the lower left). Sirius A and B are 8.6 light-years from Earth and are our fifth-closest star system. Note that the image has intentionally been overexposed to allow us to see Sirius B. (b) The same system is shown in X-ray taken with the Chandra Space Telescope. Note that Sirius A is fainter in X-rays than the hot white dwarf that is Sirius B. The Dog Star, Sirius A, and its tiny companion by NASA, ESA, H. Bond, M. Barstow(University of Leicester), ESA Standard License.
Sirius A and B:A Double Star System In The Constellation Canis Major by NASA/SAO/CXC, NASA Media License.

We have one final surprise as we leave our low-mass star in the stellar graveyard. Calculations show that as a degenerate star cools, the atoms inside it in essence “solidify” into a giant, highly compact lattice (organized rows of atoms, just like in a crystal). When carbon is compressed and crystallized in this way, it becomes a giant ***diamond-like* star**. A white dwarf star might make the most impressive engagement present you could ever see, although any attempt to mine the diamond-like material inside would crush an ardent lover instantly!

Learn about a recent “diamond star” find, a cold, white dwarf star detected in 2014, which is considered the coldest and dimmest found to date, at the website of the National Radio Astronomy Observatory. Direct link: https://public.nrao.edu/news/cold-white-dwarf/ .

Please note that coldest star means about 3000 K. Our own star is about 6000 K.

Attribution

"23.1 The Death of Low-Mass Stars" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.7 VARIABLE STARS AND NOVAE

Sometimes a star will change luminosity not due to an eclipse, but due to some sort of physical characteristic. These are called **variable stars** or **variables**. These changes in brightness can range from 1/1000 of a magnitude to 20 magnitudes over a period of a fraction of a second to many years. Over 150,000 variables are known, and many others are suspected to be variables.

There are two major classes of variable stars: **Pulsating Variables** and **Cataclysmic Variables**. **Pulsating variable stars** swell and shrink, which affects the star's brightness. One important class of pulsating variable stars is the **Cepheid Variables**.

Cataclysmic variables are binary stars, which consist of a white dwarf primary and an orbiting secondary star. The secondary star is transferring matter to the primary star. This causes the primary star to irregularly outburst a significant increase in brightness. Cataclysmic variables were originally called **novae** (singular: nova) from the Latin term *new* , since the star seemingly appeared out of nowhere.

Cataclysmic Variable

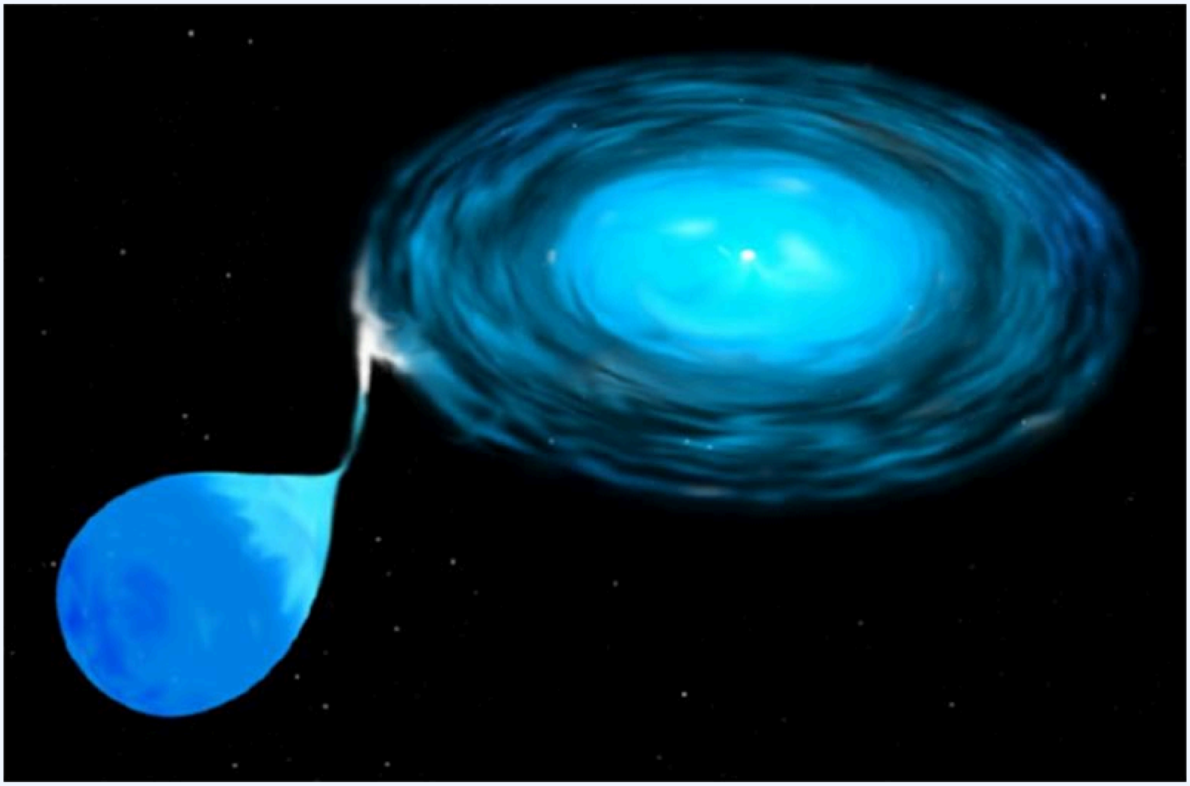


Figure 11.17. the cataclysmic variable will eventually dim back down significantly until more stellar fuel is transferred from the secondary star to the primary star. This process will go on until there is no more stellar fuel to transfer. The primary star (upper right) is pulling stellar material off the secondary star (lower left). Once a significant amount of mass builds up on the primary star, it will outburst and brighten significantly. Then the star will dim back down, repeating the process as long as fuel is available from its secondary star.
Cataclysmic Variable Star by STSci, Public Domain.

Attribution

"Variable Stars" and "Cataclysmic Variables" from Introduction to Astronomy by Florida State College at Jacksonville and Dr. Mike Reynolds is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

11.8 TYPE IA SUPERNOVAE

If a white dwarf accumulates matter from a companion star at a much faster rate, it can be pushed over the Chandrasekhar limit. The evolution of such a binary system is shown in Figure 11.18. When its mass approaches the Chandrasekhar mass limit (exceeds 1.4 M_{Sun}), such an object can no longer support itself as a white dwarf, and it begins to contract. As it does so, it heats up, and new nuclear reactions can begin in the degenerate core. The star "simmers" for the next century or so, building up internal temperature. This simmering phase ends in less than a second, when an enormous amount of fusion (especially of carbon) takes place all at once, resulting in an explosion. The fusion energy produced during the final explosion is so great that it completely destroys the white dwarf. Gases are blown out into space at velocities of about 10,000 kilometres per second, and afterward, no trace of the white dwarf remains.

Evolution of a Binary System

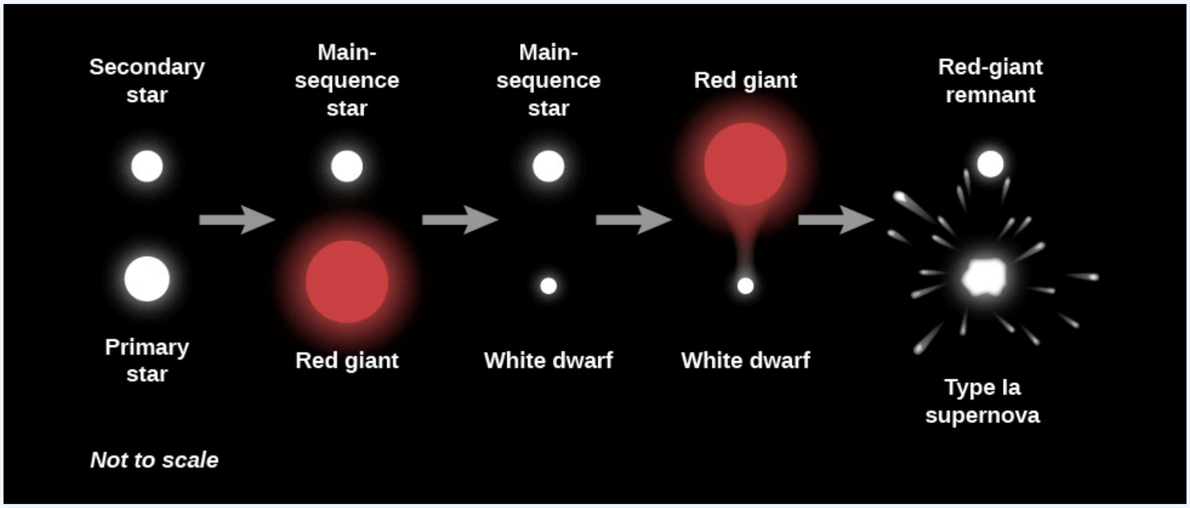


Figure 11.18. The more massive star evolves first to become a red giant and then a white dwarf. The white dwarf then begins to attract material from its companion, which in turn evolves to become a red giant. Eventually, the white dwarf acquires so much mass that it is pushed over the Chandrasekhar limit and becomes a type Ia supernova.

Such an explosion is also called a **supernova**, since, like the destruction of a high-mass star, it produces a huge amount of energy in a very short time. However, unlike the explosion of a high-mass star, which can leave behind a neutron star or black hole remnant, the white dwarf is completely destroyed in the process, leaving behind no remnant. We call these white dwarf explosions type **Ia supernovae**.

Attribution

"23.5 The Evolution of Binary Star Systems" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.9 THE DEATH OF A HIGH-MASS STAR

If what we have described so far were the whole story of the evolution of stars and elements, we would have a big problem on our hands. We will see in a later chapter that in our best models of the first few minutes of the universe, everything starts with the two simplest elements—hydrogen and helium (plus a tiny bit of lithium). All the predictions of the models imply that no heavier elements were produced at the beginning of the universe. Yet when we look around us on Earth, we see lots of other elements besides hydrogen and helium. These elements must have been made (fused) somewhere in the universe, *and the only place hot enough to make them is inside stars*. One of the fundamental discoveries of twentieth-century astronomy is that the stars are the source of most of the chemical richness that characterizes our world and our lives.

We have already seen that carbon and some oxygen are manufactured inside the lower-mass stars that become red giants. But where do the heavier elements we know and love (such as the silicon and iron inside Earth, and the gold and silver in our jewelry) come from? The kinds of stars we have been discussing so far never get hot enough at their centers to make these elements. It turns out that such heavier elements can be formed only late in the lives of *more massive* stars.

Massive stars evolve in much the same way that the Sun does (but always more quickly)—up to the formation of a carbon-oxygen core. One difference is that for stars with more than about twice the mass of the Sun, helium begins fusion more gradually, rather than with a sudden flash. Also, when more massive stars become red giants, they become so bright and large that we call them **supergiants**. Such stars can expand until their outer regions become as large as the orbit of Jupiter, which is precisely what the Hubble Space Telescope has shown for the star Betelgeuse. They also lose mass very effectively, producing dramatic winds and outbursts as they age. Figure 11.19. shows a wonderful image of the very massive star Eta Carinae, with a great deal of ejected material clearly visible.

Eta Carinae

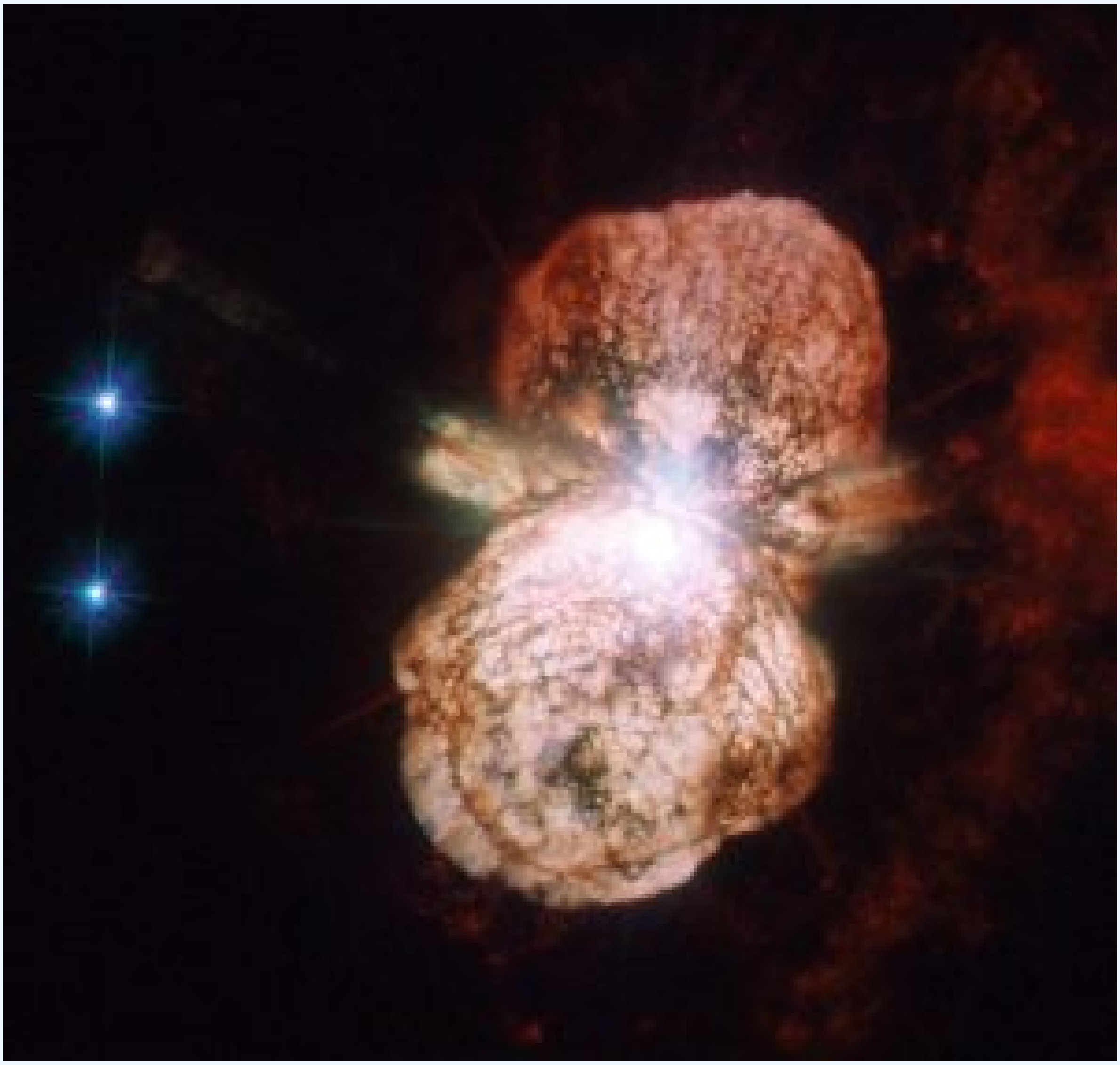


Figure 11.19. With a mass at least 100 times that of the Sun, the hot supergiant Eta Carinae is one of the most massive stars known. This Hubble Space Telescope image records the two giant lobes and equatorial disk of material it has ejected in the course of its evolution. The pink outer region is material ejected in an outburst seen in 1843, the largest of such mass loss event that any star is known to have survived. Moving away from the star at a speed of about 1000 km/s, the material is rich in nitrogen and other elements formed in the interior of the star. The inner blue-white region is the material ejected at lower speeds and is thus still closer to the star. It appears blue-white because it contains dust and reflects the light of Eta Carinae, whose luminosity is 4 million times that of our Sun.
Preview of a forthcoming supernova by ESA/NASA, ESA Standard License.

But the crucial way that massive stars diverge from the story we have outlined is that they can start additional kinds of fusion in their centres and in the shells surrounding their central regions. The outer layers of a star with a mass greater than about 8 solar masses have a weight that is enough to compress the carbon-oxygen core until it becomes hot enough to ignite fusion of carbon nuclei. Carbon can fuse into still more oxygen, and at still higher temperatures, oxygen and then neon, magnesium, and finally silicon can build even heavier elements. Iron is, however, the endpoint of this process. The fusion of iron atoms produces products that are *more* massive than the nuclei that are being fused and therefore the process *requires* energy, as opposed to releasing energy, which all fusion reactions up to this point have done. This required energy comes at the expense of the star itself, which is now on the brink of death as shown in Figure 11.20.

Interior Structure of a Massive Star Just before It Exhausts Its Nuclear Fuel

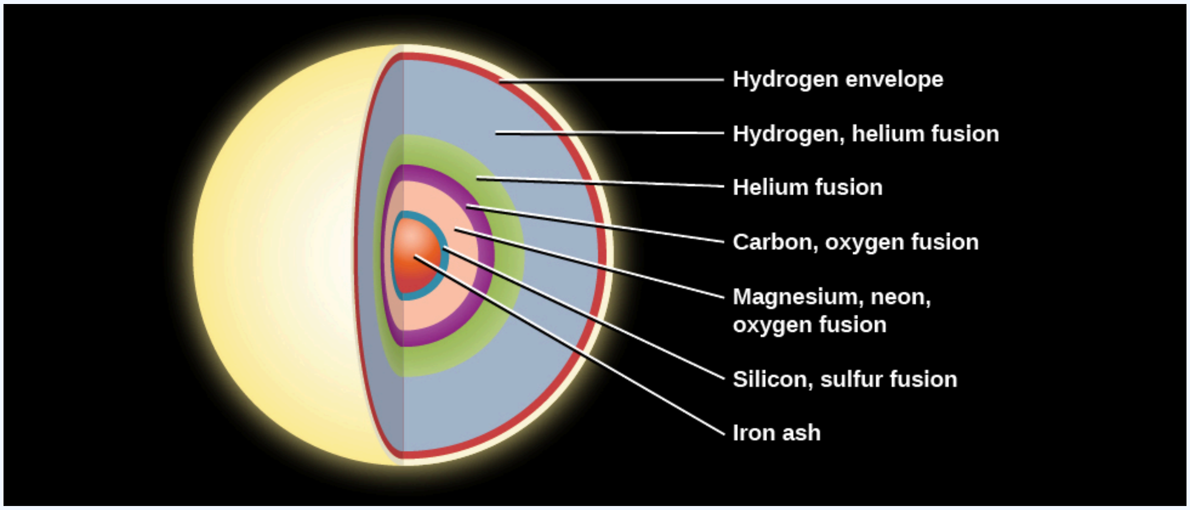


Figure 11.20. High-mass stars can fuse elements heavier than carbon. As a massive star nears the end of its evolution, its interior resembles an onion. Hydrogen fusion is taking place in an outer shell, and progressively heavier elements are undergoing fusion in the higher-temperature layers closer to the centre. All of these fusion reactions generate energy and enable the star to continue shining. Iron is different. The fusion of iron requires energy, and when iron is finally created in the core, the star has only minutes to live.

Physicists have now found nuclear pathways whereby virtually all chemical elements of atomic weights up to that of iron can be built up by this **nucleosynthesis** (the making of new atomic nuclei) in the centres of the more massive red giant stars. This still leaves the question of where elements *heavier* than iron come from. We will see in the next chapter that when massive stars finally exhaust their nuclear fuel, they most often die in a spectacular explosion—a supernova. Heavier elements can be synthesized in the stunning violence of such explosions.

Not only can we explain in this way where the elements that make up our world and others come from, but our theories of nucleosynthesis inside stars are even able to predict the *relative abundances* with which the elements occur in nature. The way stars build up elements during various nuclear reactions really can explain why some elements (oxygen, carbon, and iron) are common and others are quite rare (gold, silver, and uranium).

Compared with the main-sequence lifetimes of stars, the events that characterize the last stages of stellar evolution pass very quickly (especially for massive stars). As the star's luminosity increases, its rate of nuclear fuel consumption goes up rapidly—just at that point in its life when its fuel supply is beginning to run down.

After the prime fuel—hydrogen—is exhausted in a star's core, we saw that other sources of nuclear energy are available to the star in the fusion of, first, helium, and then of other more complex elements. But the energy yield of these reactions is much less than that of the fusion of hydrogen to helium. And to trigger these reactions, the central temperature must be higher than that required for the fusion of hydrogen to helium, leading to even more rapid consumption of fuel. Clearly this is a losing game, and very quickly the star reaches its end.

Attribution

11.10 TYPE II SUPERNOVAE

After the helium in its core is exhausted, the evolution of a massive star takes a significantly different course from that of lower-mass stars. In a massive star, the weight of the outer layers is sufficient to force the carbon core to contract until it becomes hot enough to fuse carbon into oxygen, neon, and magnesium. This cycle of contraction, heating, and the ignition of another nuclear fuel repeats several more times. After each of the possible nuclear fuels is exhausted, the core contracts again until it reaches a new temperature high enough to fuse still-heavier nuclei. The products of carbon fusion can be further converted into silicon, sulphur, calcium, and argon. And these elements, when heated to a still-higher temperature, can combine to produce iron. Massive stars go through these stages very, very quickly. In really massive stars, some fusion stages toward the very end can take only months or even days! This is a far cry from the millions of years they spend in the main-sequence stage.

At this stage of its evolution, a massive star resembles an onion with an iron core. As we get farther from the centre, we find shells of decreasing temperature in which nuclear reactions involve nuclei of progressively lower mass—silicon and sulphur, oxygen, neon, carbon, helium, and finally, hydrogen as shown in Figure 11.21.

Structure of an Old Massive Star

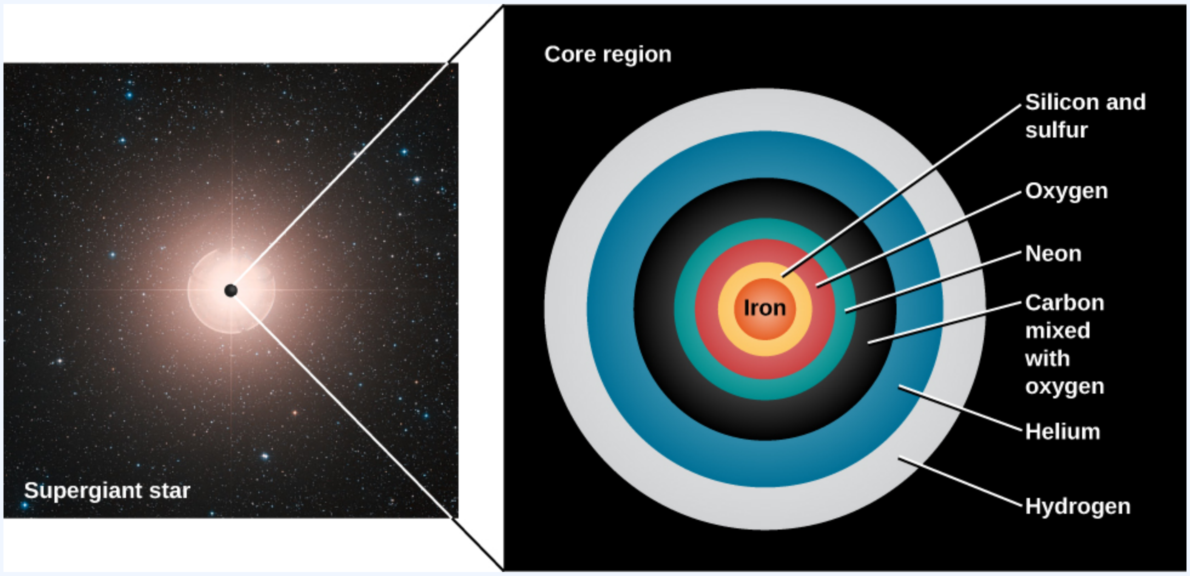


Figure 11.21. Just before its final gravitational collapse, the core of a massive star resembles an onion. The iron core is surrounded by layers of silicon and sulphur, oxygen, neon, carbon mixed with some oxygen, helium, and finally hydrogen. Outside the core, the composition is mainly hydrogen and helium. (Note that this diagram is not precisely to scale but is just meant to convey the general idea of what such a star would be like.)
Modification of Digitized Sky Survey image of Betelgeuse by ESO, Digitized Sky Survey, CC-BY-4.0.

But there is a limit to how long this process of building up elements by fusion can go on. The fusion of silicon into iron turns out to be the last step in the sequence of nonexplosive element production. Up to this point, each fusion reaction has *produced* energy because the nucleus of each fusion product has been a bit more stable than the nuclei that formed it. As discussed before, light nuclei give up some of their binding energy in the process of fusing into more tightly bound, heavier nuclei. It is this released energy that maintains the outward pressure in the core so that the star does not collapse. But of all the nuclei known, iron is the most tightly bound and thus the most stable.

You might think of the situation like this: all smaller nuclei want to "grow up" to be like iron, and they are willing to pay (*produce* energy) to move toward that goal. But iron is a mature nucleus with good self-esteem, perfectly content being iron; it requires payment (must *absorb* energy) to change its stable nuclear structure. This is the exact opposite of what has happened in each nuclear reaction so far: instead of *providing* energy to balance the inward pull of gravity, any nuclear reactions involving iron would *remove* some energy from the core of the star.

Unable to generate energy, the star now faces catastrophe.

When nuclear reactions stop, the core of a massive star is supported by degenerate electrons, just as a white dwarf is. For stars that begin their evolution with masses of at least 10 M_{Sun}, this core is likely made mainly of iron. (For stars with initial masses in the range 8 to 10 M_{Sun}, the core is likely made of oxygen, neon, and magnesium, because the star never gets hot enough to form elements as heavy as iron. The exact composition of the cores of stars in this mass range is very difficult to determine because of the complex physical characteristics in the cores, particularly at the very high densities and temperatures involved.) We will focus on the more massive iron cores in our discussion.

While no energy is being generated within the white dwarf core of the star, fusion still occurs in the shells that surround the core. As the shells finish their fusion reactions and stop producing energy, the ashes of the last reaction fall onto the white dwarf core, increasing its mass. As we saw earlier, a higher mass means a smaller core due to the larger gravity pressure. The core can contract because even a degenerate gas is still mostly empty space. Electrons and atomic nuclei are, after all, extremely small. The electrons and nuclei in a stellar core may be crowded compared to the air in your room, but there is still lots of space between them.

The electrons at first resist being crowded closer together, and so the core shrinks only a small amount. Ultimately, however, the iron core reaches a mass so large that even degenerate electrons can no longer support it. When the density reaches 4×10^{11} g/cm^3 (400 billion times the density of water), some electrons are actually squeezed into the atomic nuclei, where they combine with protons to form neutrons and neutrinos. This transformation is not something that is familiar from everyday life, but becomes very important as such a massive star core collapses.

Some of the electrons are now gone, so the core can no longer resist the crushing mass of the star's overlying layers. The core begins to shrink rapidly. More and more electrons are now pushed into the atomic nuclei, which ultimately become so saturated with neutrons that they cannot hold onto them.

At this point, the neutrons are squeezed out of the nuclei and can exert a new force. As is true for electrons, it turns out that the neutrons strongly resist being in the same place and moving in the same way. The force that can be exerted by such *degenerate neutrons* is much greater than that produced by degenerate electrons, so unless the core is too massive, they can ultimately stop the collapse.

This means the collapsing core can reach a stable state as a crushed ball made mainly of neutrons, which astronomers call a **neutron star**. We don't have an exact number (a "Chandrasekhar limit") for the maximum mass of a neutron star, but calculations tell us that the upper mass limit of a body made of neutrons might only be about 3 M_{Sun}. So if the mass of the core were greater than this, then even neutron degeneracy would not be able to stop the core from collapsing further. The dying star must end up as something even more extremely compressed, which until recently was believed to be only one possible type of object—the state of ultimate compaction known as a black hole. This is because no force was believed to exist that could stop a collapse beyond the neutron star stage.

When the collapse of a high-mass star's core is stopped by degenerate neutrons, the core is saved from further destruction, but it turns out that the rest of the star is literally blown apart. Here's how it happens.

The collapse that takes place when electrons are absorbed into the nuclei is very rapid. In less than a second, a core with a mass of about 1 M_{Sun}, which originally was approximately the size of Earth, collapses to a diameter of less than 20 kilometres. The speed with which material falls inward reaches one-fourth the speed of light. The collapse halts only when the density of the core exceeds the density of an atomic nucleus (which is the densest form of matter we know). A typical neutron star is so compressed that to duplicate its density, we would have to squeeze all the people in the world into a single sugar cube! This would give us one sugar cube's worth (one cubic centimetre's worth) of a neutron star.

The neutron degenerate core strongly resists further compression, abruptly halting the collapse. The shock of the sudden jolt initiates a shock wave that starts to propagate outward. However, this shock alone is not enough to create a star explosion. The energy produced by the outflowing matter is quickly absorbed by atomic nuclei in the dense, overlying layers of gas, where it breaks up the nuclei into individual neutrons and protons.

Our understanding of nuclear processes indicates (as we mentioned above) that each time an electron and a proton in the star's core merge to make a neutron, the merger releases a **neutrino**. These ghostly subatomic particles carry away some of the nuclear energy. It is their presence that launches the final disastrous explosion of the star. The total energy contained in the neutrinos is huge. In the initial second of the star's explosion, the power carried by the neutrinos (10^{46} watts) is greater than the power put out by all the stars in over a billion galaxies.

While neutrinos ordinarily do not interact very much with ordinary matter (we earlier accused them of being downright antisocial), matter near the center of a collapsing star is so dense that the neutrinos do interact with it to some degree. They deposit some of this energy in the layers of the star just outside the core. This huge, sudden input of energy reverses the infall of these layers and drives them explosively outward. Most of the mass of the star (apart from that which went into the neutron star in the core) is then ejected outward into space. As we saw earlier, such an explosion requires a star of at least 8 M_{Sun}, and the neutron star can have a mass of at most 3 $M_{\text{Sun.}}$ Consequently, at least five times the mass of our Sun is ejected into space in each such explosive event!

The resulting explosion is called a **supernova**, it is pictured in Figure 11.22 below. When these explosions happen close by, they can be among the most spectacular celestial events, as we will discuss in the next section.

Five Supernova Explosions in Other Galaxies

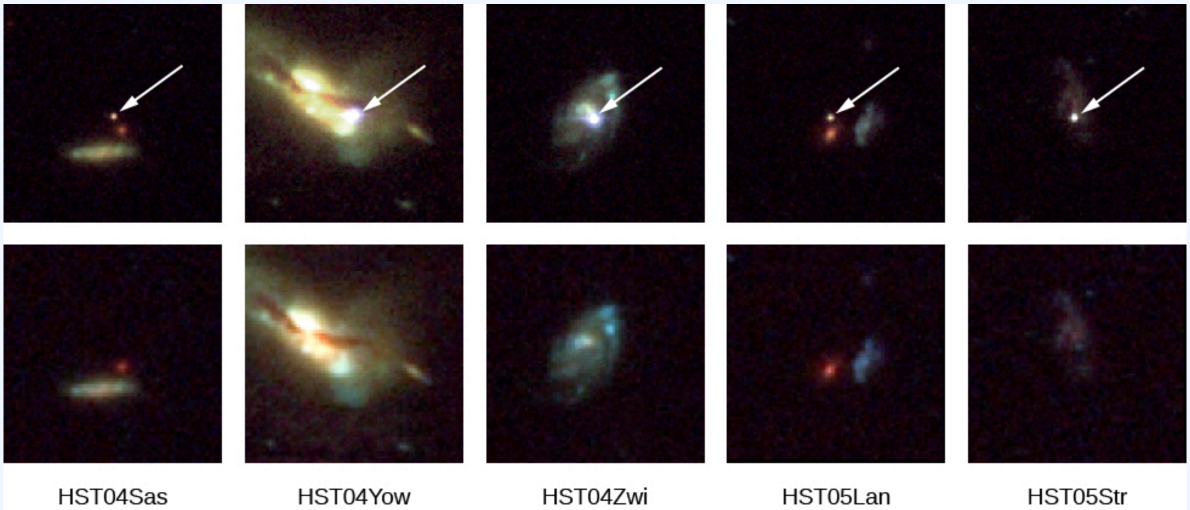


Figure 11.22. The arrows in the top row of images point to the supernovae. The bottom row shows the host galaxies before or after the stars exploded. Each of these supernovae exploded between 3.5 and 10 billion years ago. Note that the supernovae when they first explode can be as bright as an entire galaxy. Host Galaxies of Distant Supernovae by NASA, ESA, and A. Riess (STScI), NASA Media License.

Table 11.6 summarizes the discussion so far about what happens to stars and substellar objects of different initial masses at the ends of their lives. Like so much of our scientific understanding, this list represents a progress report: it is the best we can do with our present models and observations. The mass limits corresponding to various outcomes may change somewhat as models are improved. There is much we do not yet understand about the details of what happens when stars die.

Table 11.6. The Ultimate Fate of Stars and Substellar Objects with Different Masses

Initial Mass (Mass of Sun = 1)	Final State at the End of Its Life
<0.01	Planet
0.01 to 0.08	Brown dwarf
0.08 to 0.25	White dwarf made mostly of helium
0.25 to 8	White dwarf made mostly of carbon and oxygen
8 to 10	White dwarf made of oxygen, neon, and magnesium
10 to 40	Supernova explosion that leaves a neutron star
> 40	Supernova explosion that leaves a black hole

Attribution

11.11 SUPERNOVA OBSERVATIONS

Supernovae were discovered long before astronomers realized that these spectacular cataclysms mark the death of stars. The word *nova* means "new" in Latin; before telescopes, when a star too dim to be seen with the unaided eye suddenly flared up in a brilliant explosion, observers concluded it must be a brand-new star. Twentieth-century astronomers reclassified the explosions with the greatest luminosity as *super*novae.

From historical records of such explosions, from studies of the remnants of supernovae in our Galaxy, and from analyses of supernovae in other galaxies, we estimate that, on average, one supernova explosion occurs somewhere in the Milky Way Galaxy every 25 to 100 years. Unfortunately, however, no supernova explosion has been observable in our Galaxy since the invention of the telescope. Either we have been exceptionally unlucky or, more likely, recent explosions have taken place in parts of the Galaxy where interstellar dust blocks light from reaching us.

SUPERNOVAE IN HISTORY

Although many supernova explosions in our own Galaxy have gone unnoticed, a few were so spectacular that they were clearly seen and recorded by sky watchers and historians at the time. We can use these records, going back two millennia, to help us pinpoint where the exploding stars were and thus where to look for their remnants today.

The most dramatic supernova was observed in the year 1006. It appeared in May as a brilliant point of light visible during the daytime, perhaps 100 times brighter than the planet Venus. It was bright enough to cast shadows on the ground during the night and was recorded with awe and fear by observers all over Europe and Asia. No one had seen anything like it before; Chinese astronomers, noting that it was a temporary spectacle, called it a "guest star."

Astronomers David Clark and Richard Stephenson have scoured records from around the world to find more than 20 reports of the 1006 supernova (SN 1006) which is pictured in Figure 11.23.

This has allowed them to determine with some accuracy where in the sky the explosion occurred. They place it in the modern constellation of Lupus; at roughly the position they have determined, we find a supernova remnant, now quite faint. From the way its filaments are expanding, it indeed appears to be about 1000 years old.

Supernova 1006 Remnant

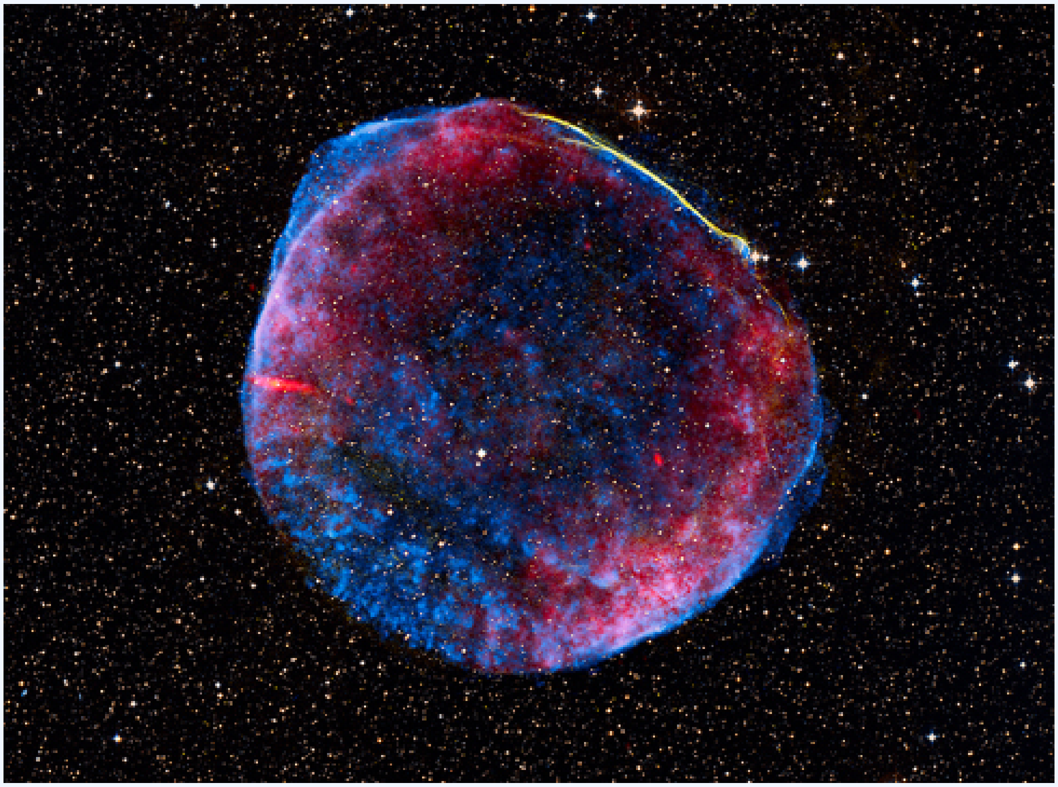


Figure 11.23. This composite view of SN 1006 from the Chandra X-Ray Observatory shows the X-rays coming from the remnant in blue, visible light in white-yellow, and radio emission in red.
SN 1006 Supernova Remnant by NASA, ESA, Zolt Levay(STScI), NASA Media License.

Another guest star, now known as SN 1054, was clearly recorded in Chinese records in July 1054. The remnant of that star is one of the most famous and best-studied objects in the sky, called the Crab Nebula. It is a marvellously complex object, which has been key to understanding the death of massive stars. When its explosion was first seen, we estimate that it was about as bright as the planet Jupiter: nowhere near as dazzling as the 1006 event but still quite dramatic to anyone who kept track of objects in the sky. Another fainter supernova was seen in 1181.

The next supernova became visible in November 1572 and, being brighter than the planet Venus, was quickly spotted by a number of observers, including the young Tycho Brahe. His careful measurements of the star over a year and a half showed that it was not a comet or something in Earth's atmosphere since it did not move relative to the stars. He correctly deduced that it must be a phenomenon belonging to the realm of the stars, not of the solar system. The remnant of Tycho's Supernova (as it is now called) can still be detected in many different bands of the electromagnetic spectrum.

Not to be outdone, Johannes Kepler, Tycho Brahe's scientific heir, found his own supernova in 1604, now known as Kepler's Supernova. Fainter than Tycho's, it nevertheless remained visible for about a year. Kepler wrote a book about his observations that was read by many with an interest in the heavens, including Galileo.

No supernova has been spotted in our Galaxy for the past 300 years. Since the explosion of a visible supernova is a chance event, there is no way to say when the next one might occur. Around the world, dozens of professional and amateur astronomers keep a sharp lookout for "new" stars that appear overnight, hoping to be the first to spot the next guest star in our sky and make a little history themselves.

At their maximum brightness, the most luminous supernovae have about 10 billion times the luminosity of the Sun. For a brief time, a supernova may outshine the entire galaxy in which it appears. After maximum brightness, the star's light fades and disappears from telescopic visibility within a few months or years. At the time of their outbursts, supernovae eject material at typical velocities of 10,000 kilometres per second (and speeds twice that have been observed). A speed of 20,000 kilometres per second corresponds to about 45 million miles per hour, truly an indication of great cosmic violence.

Supernovae are classified according to the appearance of their spectra, but in this chapter, we will focus on the two main causes of supernovae. Type Ia supernovae are ignited when a lot of material is dumped on degenerate white dwarfs as shown in Figure 11.24; these supernovae will be discussed later in this chapter. For now, we will continue our story about the death of massive stars and focus on type II supernovae, which are produced when the core of a massive star collapses.

Supernova 2014J

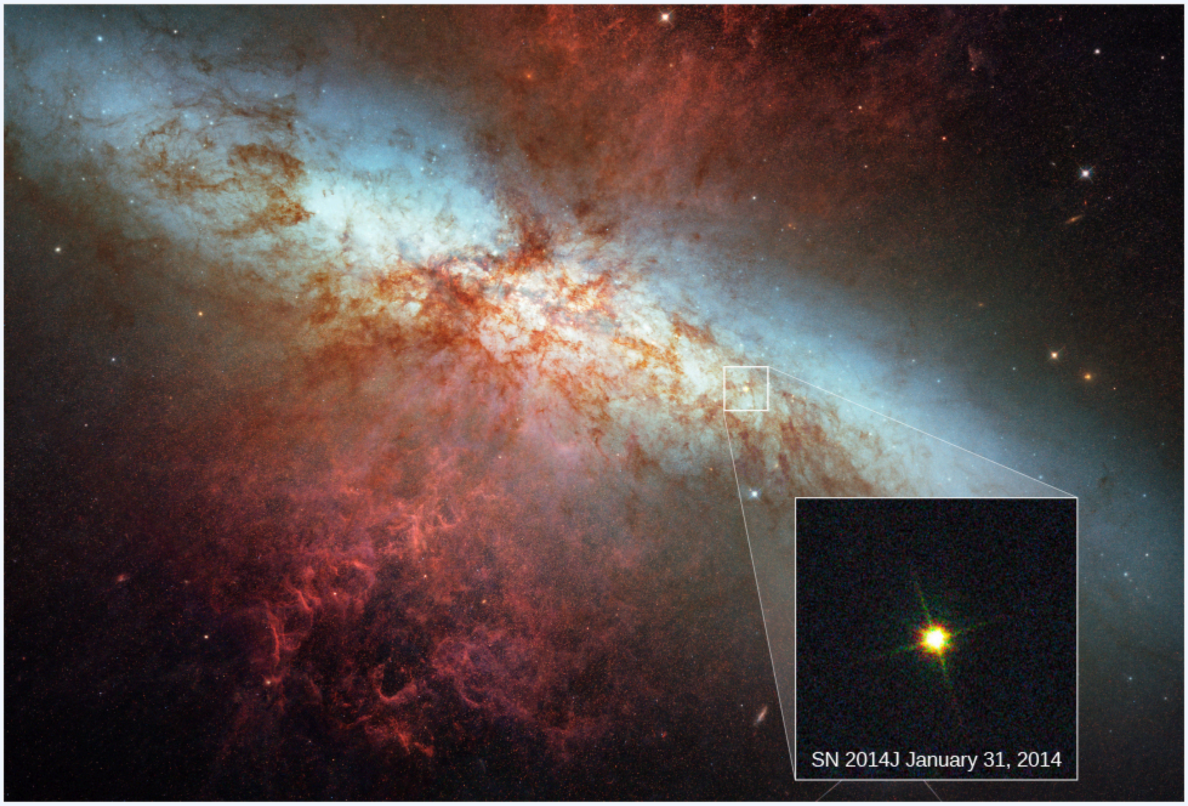


Figure 11.24. This image of supernova 2014J, located in Messier 82 (M82), which is also known as the Cigar galaxy, was taken by the Hubble Space Telescope and is superposed on a mosaic image of the galaxy also taken with Hubble. The supernova event is indicated by the box and the inset. This explosion was produced by a type Ia supernova, which is theorized to be triggered in binary systems consisting of a white dwarf and another star—and could be a second white dwarf, a star like our Sun, or a giant star. This type of supernova will be discussed later in this chapter. At a distance of approximately 11.5 million light-years from Earth, this is the closest supernova of type Ia discovered in the past few decades. In the image, you can see reddish plumes of hydrogen coming from the central region of the galaxy, where a considerable number of young stars are being born.
Hubble Monitors Supernova In Nearby Galaxy M82 by NASA, ESA, A. Goobar (Stockholm University), and the Hubble Heritage Team (STScI/AURA), NASA Media Licence.

Supernova 1987A

Our most detailed information about what happens when a type II supernova occurs comes from an event that was observed in 1987. Before dawn on February 24, Ian Shelton, a Canadian astronomer working at an observatory in Chile, pulled a photographic plate from the developer. Two nights earlier, he had begun a survey of the Large Magellanic Cloud, a small galaxy that is one of the Milky Way's nearest neighbours in space. Where he expected to see only faint stars, he saw a large bright spot. Concerned that his photograph was flawed, Shelton went outside to look at the Large Magellanic Cloud . . . and saw that a new object had indeed appeared in the sky as shown in Figure 11.25. He soon realized that he had discovered a supernova, one that could be seen with the unaided eye even though it was about 160,000 light-years away.

Hubble Space Telescope Image of SN 1987A

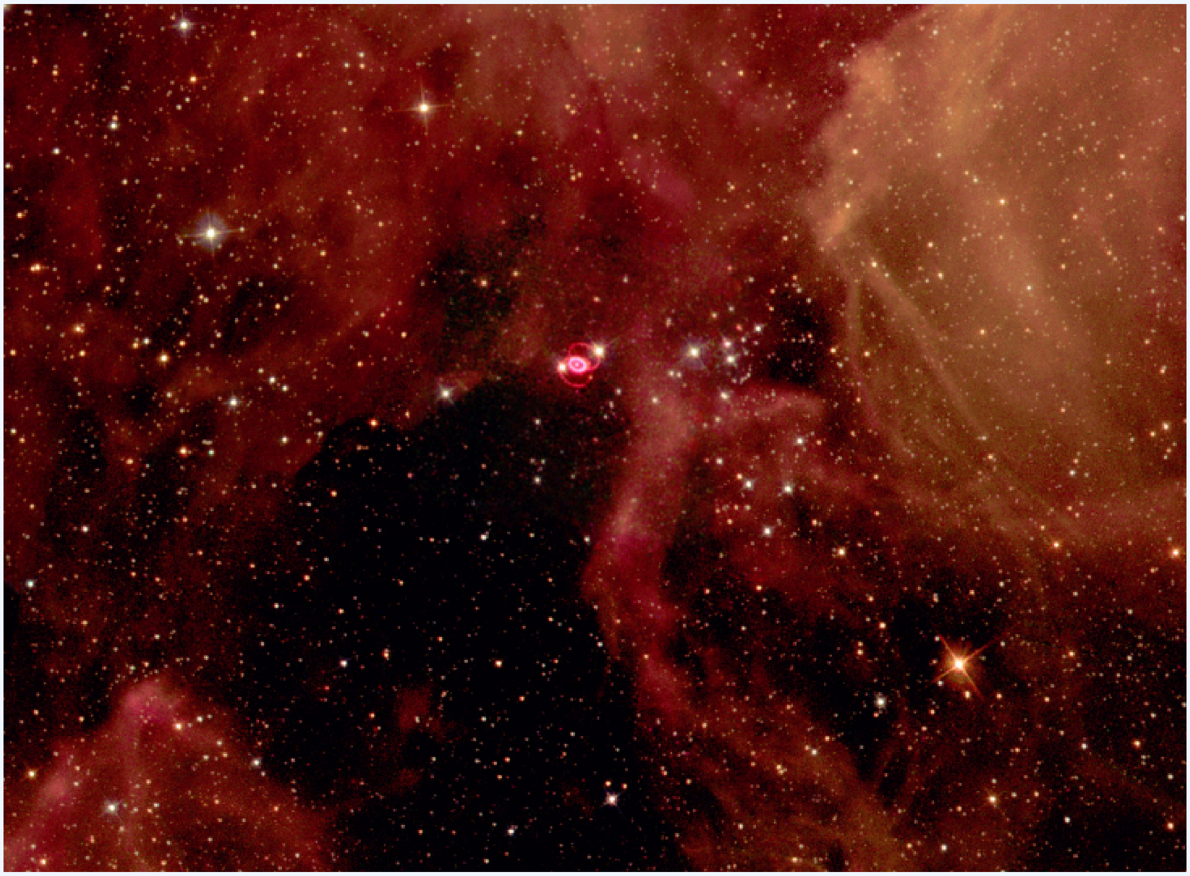


Figure 11.25. The supernova remnant with its inner and outer red rings of material is located in the Large Magellanic Cloud. This image is a composite of several images taken in 1994, 1996, and 1997—about a decade after supernova 1987A was first observed.
SN 1987a in the Large Magellanic Cloud by the Hubble Heritage Team (AURA/STScI/NASA/ESA), ESA Standard Licence.

Now known as SN 1987A, since it was the first supernova discovered in 1987, this brilliant newcomer to the southern sky gave astronomers their first opportunity to study the death of a relatively nearby star with modern instruments. It was also the first time astronomers had observed a star *before* it became a supernova. The star that blew up had been included in earlier surveys of the Large Magellanic Cloud, and as a result, we know the star was a blue supergiant just before the explosion.

By combining theory and observations at many different wavelengths, astronomers have reconstructed the life story of the star that became SN 1987A. Formed about 10 million years ago, it originally had a mass of about 20 M_{Sun}. For 90% of its life, it lived quietly on the main sequence, converting hydrogen into helium. At this time, its luminosity was about 60,000 times that of the Sun (L_{Sun}), and its spectral type was O. When the hydrogen in the centre of the star was exhausted, the core contracted and ultimately became hot enough to fuse helium. By this time, the star was a red supergiant, emitting about 100,000 times more energy than the Sun. While in this stage, the star lost some of its mass.

This lost material has actually been detected by observations with the Hubble Space Telescope as shown in Figure 11.26. The gas driven out into space by the subsequent supernova explosion is currently colliding with the material the star left behind when it was a red giant. As the two collide, we see a glowing ring.

Ring around Supernova 1987A

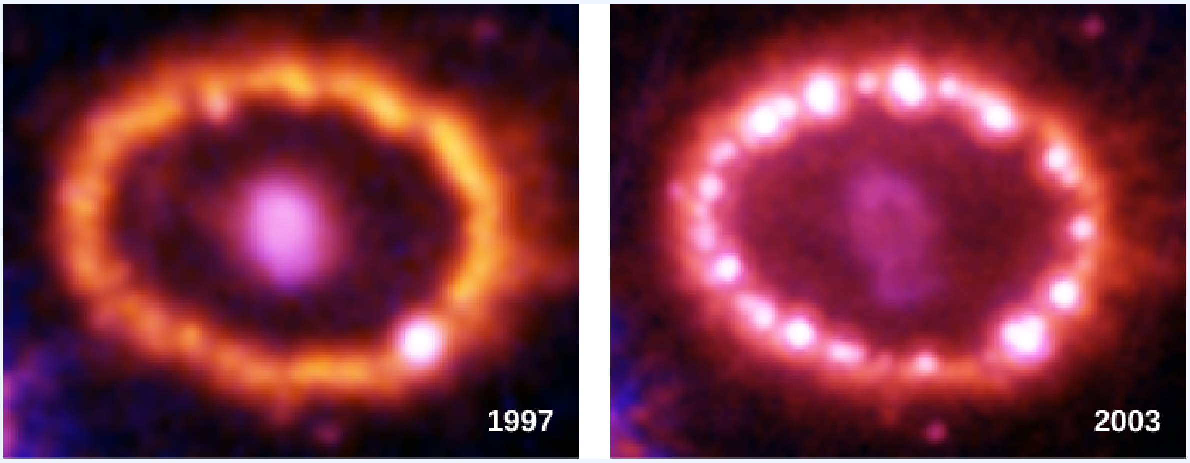


Figure 11.26. These two images show a ring of gas expelled about 30,000 years ago when the star that exploded in 1987 was a red giant. The supernova, which has been artificially dimmed, is located at the centre of the ring. The left-hand image was taken in 1997 and the right-hand image in 2003. Note that the number of bright spots has increased from 1 to more than 15 over this time interval. These spots occur where high-speed gas ejected by the supernova and moving at millions of miles per hour has reached the ring and blasted into it. The collision has heated the gas in the ring and caused it to glow more brightly. The fact that we see individual spots suggests that material ejected by the supernova is first hitting narrow, inward-projecting columns of gas in the clumpy ring. The hot spots are the first signs of a dramatic and violent collision between the new and old material that will continue over the next few years. By studying these bright spots, astronomers can determine the composition of the ring and hence learn about the nuclear processes that build heavy elements inside massive stars.
Modification of Hubble Supernova 1987A Scrapbook (1994-2003) by NASA, P. Challis, R. Kirshner (Harvard-Smithsonian Center for Astrophysics) and B. Sugerman (STScI), ESA Standard Licence.

Helium fusion lasted only about 1 million years. When the helium was exhausted at the centre of the star, the core contracted again, the radius of the surface also decreased, and the star became a blue supergiant with a luminosity still about equal to 100,000 L_{Sun}. This is what it still looked like on the outside when, after brief periods of further fusion, it reached the iron crisis we discussed earlier and exploded.

Some key stages of evolution of the star that became SN 1987A, including the ones following helium exhaustion, are listed in Table 11.7. While we don't expect you to remember these numbers, note the patterns in the table: each stage of evolution happens more quickly than the preceding one, the temperature and pressure in the core increase, and progressively heavier elements are the source of fusion energy. Once iron was created, the collapse began. It was a catastrophic collapse, lasting only a few tenths of a second; the speed of infall in the outer portion of the iron core reached 70,000 kilometres per second, about one-fourth the speed of light.

Table 11.7. Evolution of the Star That Exploded as SN 1987A

Phase	Central Temperature (K)	Central Density (g/cm^3)	Time Spent in This Phase
Hydrogen fusion	40×10^6	5	8×10^6 years
Helium fusion	190×10^6	970	10^6 years
Carbon fusion	870×10^6	170,000	2000 years
Neon fusion	1.6×10^9	3.0×10^6	6 months
Oxygen fusion	2.0×10^9	5.6×10^6	1 year
Silicon fusion	3.3×10^9	4.3×10^7	Days
Core collapse	200×10^9	2×10^{14}	Tenths of a second

In the meantime, as the core was experiencing its last catastrophe, the outer shells of neon, oxygen, carbon, helium, and hydrogen in the star did not yet know about the collapse. Information about the physical movement of different layers travels through a star at the speed of sound and cannot reach the surface in the few tenths of a second required for the core collapse to occur. Thus, the surface layers of our star hung briefly suspended, much like a cartoon character who dashes off the edge of a cliff and hangs momentarily in space before realizing that he is no longer held up by anything.

The collapse of the core continued until the densities rose to several times that of an atomic nucleus. The resistance to further collapse then became so great that the core rebounded. Infalling material ran into the "brick wall" of the rebounding core and was thrown outward with a great shock wave. Neutrinos poured out of the core, helping the shock wave blow the star apart. The shock reached the surface of the star a few hours later, and the star began to brighten into the supernova Ian Shelton observed in 1987.

Attribution

"23.3 Supernova Observations" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.12 NEUTRON STARS AND PULSARS

Neutron stars are the densest objects in the universe; the force of gravity at their surface is 10^{11} times greater than what we experience at Earth's surface. The interior of a neutron star is composed of about 95% neutrons, with a small number of protons and electrons mixed in. In effect, a neutron star is a giant atomic nucleus, with a mass about 10^{57} times the mass of a proton. Its diameter is more like the size of a small town or an asteroid than a star. (Table 11.8 compares the properties of neutron stars and white dwarfs.) Because it is so small, a neutron star probably strikes you as the object least likely to be observed from thousands of light-years away. Yet neutron stars do manage to signal their presence across vast gulfs of space.

Table 11.8. Properties of a Typical White Dwarf and a Neutron Star

Property	White Dwarf	Neutron Star
Mass (Sun = 1)	0.6 (always <1.4)	Always >1.4 and <3
Radius	7000 km	10 km
Density	8×10^5 g/cm^3	10^{14} g/cm^3

The Discovery of Neutron Stars

In 1967, Jocelyn Bell, a research student at Cambridge University, was studying distant radio sources with a special detector that had been designed and built by her advisor Antony Hewish to find rapid variations in radio signals. The project computers spewed out reams of paper showing where the telescope had surveyed the sky, and it was the job of Hewish's graduate students to go through it all, searching for interesting phenomena. In September 1967, Bell discovered what she called "a bit of scruff"—a strange radio signal unlike anything seen before.

What Bell had found, in the constellation of Vulpecula, was a source of rapid, sharp, intense, and extremely regular pulses of radio radiation. Like the regular ticking of a clock, the pulses arrived precisely every 1.33728 seconds. Such exactness first led the scientists to speculate that perhaps they had found signals from an intelligent civilization. Radio astronomers even half-jokingly dubbed the source "LGM" for "little green men." Soon, however, three similar sources were discovered in widely separated directions in the sky.

When it became apparent that this type of radio source was fairly common, astronomers concluded that they were highly unlikely to be signals from other civilizations. By today, more than 2500 such sources have been discovered; they are now called **pulsars**, short for "pulsating radio sources."

The pulse periods of different pulsars range from a little longer than 1/1000 of a second to nearly 10 seconds. At first, the pulsars seemed particularly mysterious because nothing could be seen at their location on visible-light photographs. But then a pulsar was discovered right in the centre of the Crab Nebula, a cloud of gas produced by SN 1054, a supernova that was recorded by the Chinese in 1054, shown in Figure 11.27. The energy from the Crab Nebula pulsar arrives in sharp bursts that occur 30 times each second—with a regularity that would be the envy of a Swiss watchmaker. In addition to pulses of radio energy, we can observe pulses of visible light and X-rays from the Crab Nebula. The fact that the pulsar was just in the region of the supernova remnant where we expect the leftover neutron star to be immediately alerted astronomers that pulsars might be connected with these elusive "corpses" of massive stars.

Crab Nebula

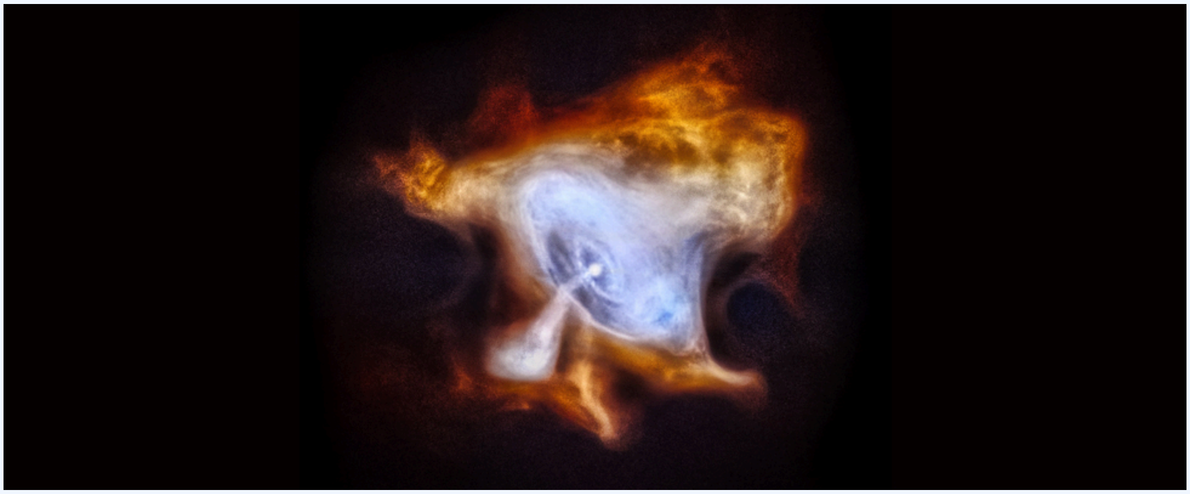


Figure 11.27. This image shows X-ray emissions from the Crab Nebula, which is about 6500 light-years away. The pulsar is the bright spot at the centre of the concentric rings. Data taken over about a year show that particles stream away from the inner ring at about half the speed of light. The jet that is perpendicular to this ring is a stream of matter and antimatter electrons also moving at half the speed of light.
Chandra Celebrates 15th Anniversary: Crab Nebula by NASA/CXC/SAO, NASA Media License.

The Crab Nebula is a fascinating object. The whole nebula glows with radiation at many wavelengths, and its overall energy output is more than 100,000 times that of the Sun—not a bad trick for the remnant of a supernova that exploded almost a thousand years ago. Astronomers soon began to look for a connection between the pulsar and the large energy output of the surrounding nebula.

By applying a combination of theory and observation, astronomers eventually concluded that pulsars must be *spinning neutron stars*. According to this model, a neutron star is something like a lighthouse on a rocky coast as pictured in Figure 11.28. To warn ships in all directions and yet not cost too much to operate, the light in a modern lighthouse turns, sweeping its beam across the dark sea. From the vantage point of a ship, you see a pulse of light each time the beam points in your direction. In the same way, radiation from a small region on a neutron star sweeps across the oceans of space, giving us a pulse of radiation each time the beam points toward Earth.

Lighthouse

Figure 11.28. A lighthouse in California warns ships on the ocean not to approach too close to the dangerous shoreline. The lighted section at the top rotates so that its beam can cover all directions. Pigeon Point Lighthouse by Anita Ritenour, CC BY-4.0.

Neutron stars are ideal candidates for such a job because the collapse has made them so small that they can turn very rapidly.

Even if the parent star was rotating very slowly when it was on the main sequence, its rotation had to speed up as it collapsed to form a neutron star. With a diameter of only 10 to 20 kilometres, a neutron star can complete one full spin in only a fraction of a second. This is just the sort of time period we observe between pulsar pulses.

Any magnetic field that existed in the original star will be highly compressed when the core collapses to a neutron star. At the surface of the neutron star, in the outer layer consisting of ordinary matter (and not just pure neutrons), protons and electrons are caught up in this spinning field and accelerated nearly to the speed of light. In only two places—the north and south magnetic poles—can the trapped particles escape the strong hold of the magnetic field as pictured in Figure 11.29. The same effect can be seen (in reverse) on Earth, where charged particles from space are *kept out* by our planet's magnetic field everywhere except near the poles. As a result, Earth's auroras (caused when charged particles hit the atmosphere at high speed) are seen mainly near the poles.

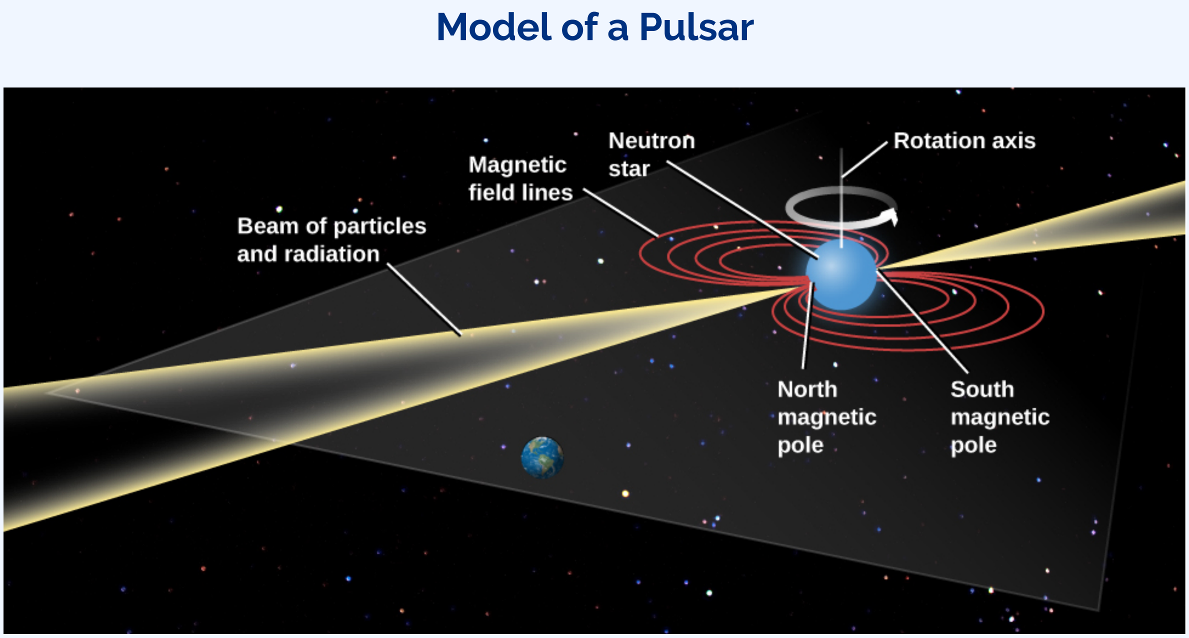


Figure 11.29. A diagram showing how beams of radiation at the magnetic poles of a neutron star can give rise to pulses of emission as the star rotates. As each beam sweeps over Earth, like a lighthouse beam sweeping over a distant ship, we see a short pulse of radiation. This model requires that the magnetic poles be located in different places from the rotation poles.
Modification of Stars La Palma by Tony Hisgett, CC-BY 4.0.

Note that in a neutron star, the magnetic north and south poles do not have to be anywhere close to the north and south poles defined by the star's rotation. Figure 11.29 shows the poles of the magnetic field perpendicular to the poles of rotation, but the two kinds of poles could make any angle.

In fact, the misalignment of the rotational axis with the magnetic axis plays a crucial role in the generation of the observed pulses in this model. At the two magnetic poles, the particles from the neutron star are focused into a narrow beam and come streaming out of the whirling magnetic region at enormous speeds. They emit energy over a broad range of the electromagnetic spectrum. The radiation itself is also confined to a narrow beam, which explains why the pulsar acts like a lighthouse. As the rotation carries first one and then the other magnetic pole of the star into our view, we see a pulse of radiation each time.

The Evolution of Pulsars

From observations of the pulsars discovered so far, astronomers have concluded that one new pulsar is born somewhere in the Galaxy every 25 to 100 years, the same rate at which supernovae are estimated to occur. Calculations suggest that the typical lifetime of a pulsar is about 10 million years; after that, the neutron star no longer rotates fast enough to produce significant beams of particles and energy, and is no longer observable. We estimate that there are about 100 million neutron stars in our Galaxy, most of them rotating too slowly to come to our notice.

The Crab pulsar is rather young (only about 960 years old) and has a short period, whereas other, older pulsars have already slowed to longer periods. Pulsars thousands of years old have lost too much energy to emit appreciably in the visible and X-ray wavelengths, and they are observed only as radio pulsars; their periods are a second or longer.

There is one other reason we can see only a fraction of the pulsars in the Galaxy. Consider our lighthouse model again. On Earth, all ships approach on the same plane—the surface of the ocean—so the lighthouse can be built to sweep its beam over that surface. But in space, objects can be anywhere in three dimensions. As a given pulsar's beam sweeps over a circle in space, there is absolutely no guarantee that this circle will include the direction of Earth. In fact, if you think about it, many more circles in space will *not* include Earth than will include it. Thus, we estimate that we are unable to observe a large number of neutron stars because their pulsar beams miss us entirely.

At the same time, it turns out that only a few of the pulsars discovered so far are embedded in the visible clouds of gas that mark the remnant of a supernova. This might at first seem mysterious, since we know that supernovae give rise to neutron stars and we should expect each pulsar to have begun its life in a supernova explosion. But the lifetime of a pulsar turns out to be about 100 times longer than the length of time required for the expanding gas of a supernova remnant to disperse into interstellar space. Thus, most pulsars are found with no other trace left of the explosion that produced them.

In addition, some pulsars are ejected by a supernova explosion that is not the same in all directions. If the supernova explosion is stronger on one side, it can kick the pulsar entirely out of the supernova remnant (some astronomers call this “getting a birth kick”). We know such kicks happen because we see a number of young supernova remnants in nearby galaxies where the pulsar is to one side of the remnant and racing away at several hundred miles per second, pictured in Figure 11.30.

Speeding Pulsar

Figure 11.30. This intriguing image (which combines X-ray, visible, and radio observations) shows the jet trailing behind a pulsar (at bottom right, lined up between the two bright stars). With a length of 37 light-years, the jet trail (seen in purple) is the longest ever observed from an object in the Milky Way. (There is also a mysterious shorter, comet-like tail that is almost perpendicular to the purple jet.) Moving at a speed between 2.5 and 5 million miles per hour, the pulsar is travelling away from the core of the supernova remnant where it originated.
Composite image of pulsar IGR J11014-6103 by NASA, NASA Media License.

Touched by a Neutron Star

On December 27, 2004, Earth was bathed with a stream of X-ray and gamma-ray radiation from a neutron star known as SGR 1806-20. What made this event so remarkable was that, despite the distance of the source, its tidal wave of radiation had measurable effects on Earth's atmosphere. The apparent brightness of this gamma-ray flare was greater than any historical star explosion.

The primary effect of the radiation was on a layer high in Earth's atmosphere called the *ionosphere*. At night, the ionosphere is normally at a height of about 85 kilometres, but during the day, energy from the Sun ionizes more molecules and lowers the boundary of the ionosphere to a height of about 60 kilometres. The pulse of X-ray and gamma-ray radiation produced about the same level of ionization as the daytime Sun. It also caused some sensitive satellites above the atmosphere to shut down their electronics.

Measurements by telescopes in space indicate that SGR 1806-20 was a special type of fast-spinning neutron star called a *magnetar*. Astronomers Robert Duncan and Christopher Thomson gave them this name because their magnetic fields are stronger than that of any other type of astronomical source—in this case, about 800 trillion times stronger than the magnetic field of Earth.

A magnetar is thought to consist of a superdense core of neutrons surrounded by a rigid crust of atoms about a mile deep with a surface made of iron. The magnetar's field is so strong that it creates huge stresses inside that can sometimes crack open the hard crust, causing a starquarke. The vibrating crust produces an enormous blast of radiation. An astronaut 0.1 light-year from this particular magnetar would have received a fatal does from the blast in less than a second.

Fortunately, we were far enough away from magnetar SGR 1806-20 to be safe. Could a magnetar ever present a real danger to Earth? To produce enough energy to disrupt the ozone layer, a magnetar would have to be located within the cloud of comets that surround the solar system, and we know no magnetars are that close. Nevertheless, it is a fascinating discovery that events on distant star corpses can have measurable effects on Earth.

Attribution

11.13 BLACK HOLES

Most stars end their lives as white dwarfs or neutron stars. When a *very* massive star collapses at the end of its life, however, not even the mutual repulsion between densely packed neutrons can support the core against its own weight. If the remaining mass of the star's core is more than about three times that of the Sun (M_{Sun}), our theories predict that *no known force can stop it from collapsing forever!* Gravity simply overwhelms all other forces and crushes the core until it occupies an infinitely small volume. A star in which this occurs may become one of the strangest objects ever predicted by theory—a **black hole**.

Classical Collapse

Let's begin with a thought experiment. We want to know what speeds are required to escape from the gravitational pull of different objects. A rocket must be launched from the surface of Earth at a very high speed if it is to escape the pull of Earth's gravity. In fact, any object—rocket, ball, astronomy book—that is thrown into the air with a velocity less than 11 kilometres per second will soon fall back to Earth's surface. Only those objects launched with a speed greater than this *escape velocity* can get away from Earth.

The escape velocity from the surface of the Sun is higher yet—618 kilometres per second. Now imagine that we begin to compress the Sun, forcing it to shrink in diameter. Recall that the pull of gravity depends on both the mass that is pulling you and your distance from the centre of gravity of that mass. If the Sun is compressed, its *mass* will remain the same, but the *distance* between a point on the Sun's surface and the centre will get smaller and smaller. Thus, as we compress the star, the pull of gravity for an object on the shrinking surface will get stronger and stronger as depicted in Figure 11.31.

Formation of a Black Hole

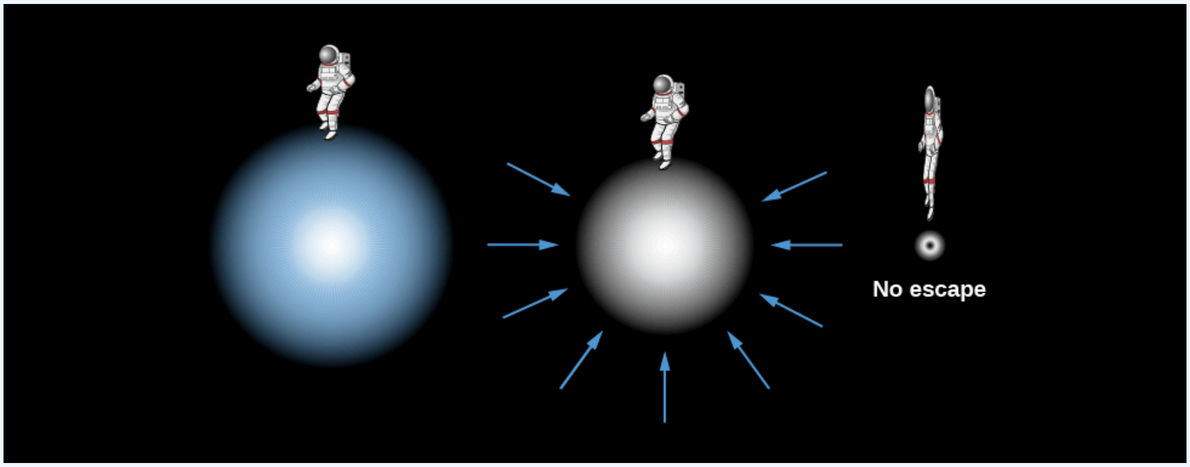


Figure 11.31. At left, an imaginary astronaut floats near the surface of a massive star-core about to collapse. As the same mass falls into a smaller sphere, the gravity at its surface goes up, making it harder for anything to escape from the stellar surface. Eventually the mass collapses into so small a sphere that the escape velocity exceeds the speed of light and nothing can get away. Note that the size of the astronaut has been exaggerated. In the last picture, the astronaut is just outside the sphere we will call the event horizon and is stretched and squeezed by the strong gravity.

When the shrinking Sun reaches the diameter of a neutron star (about 20 kilometres), the velocity required to escape its gravitational pull will be about half the speed of light. Suppose we continue to compress the Sun to a smaller and smaller diameter. (We saw this can't happen to a star like our Sun in the real world because of electron degeneracy, i.e., the mutual repulsion between tightly packed electrons; this is just a quick "thought experiment" to get our bearings).

Ultimately, as the Sun shrinks, the escape velocity near the surface would exceed the speed of light. If the speed you need to get away is faster than the fastest possible speed in the universe, then nothing, not even light, is able to escape. An object with such large escape velocity emits no light, and anything that falls into it can never return.

In modern terminology, we call an object from which light cannot escape a black hole, a name popularized by the America scientist John Wheeler starting in the late 1960s, pictured in Figure 11.32. The idea that such objects might exist is, however, not a new one. Cambridge professor and amateur astronomer John Michell wrote a paper in 1783 about the possibility that stars with escape velocities exceeding that of light might exist. And in 1796, the French mathematician Pierre-Simon, marquis de Laplace, made similar calculations using Newton's theory of gravity; he called the resulting objects "dark bodies."

John Wheeler (1911–2008)

Figure 11.32. This brilliant physicist did much pioneering work in general relativity theory and popularized the term black hole starting in the late 1960s.
John Archibald Wheeler by Ulli Steltzer, courtesy of AIP Emilio Segrè Visual Archives, Public Domain.

While these early calculations provided strong hints that something strange should be expected if very massive objects collapse under their own gravity, we really need **general relativity theory** to give an adequate description of what happens in such a situation.

Collapse with Relativity

General relativity tells us that gravity is really a curvature of spacetime. As gravity increases (as in the collapsing Sun of our thought experiment), the curvature gets larger and larger. Eventually, if the Sun could shrink down to a diameter of about 6 kilometres, only light beams sent out perpendicular to the surface would escape. All others would fall back onto the star as shown in Figure 11.33. If the Sun could then shrink just a little more, even that one remaining light beam would no longer be able to escape.

Light Paths near a Massive Object

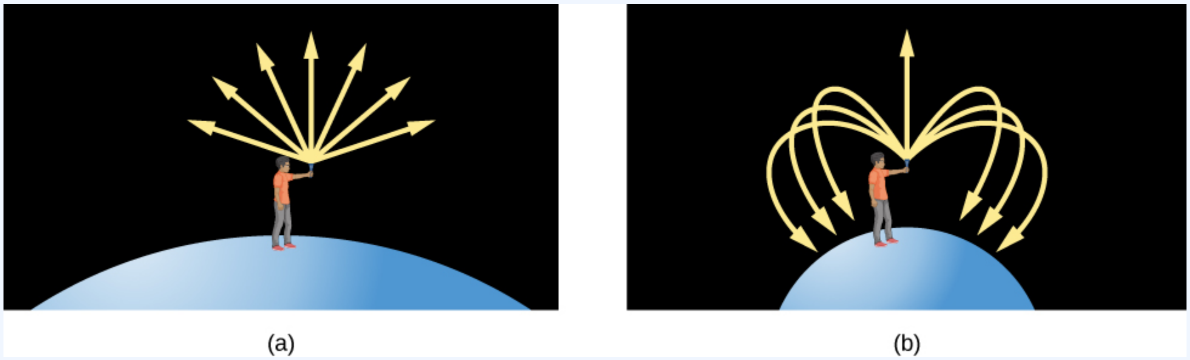


Figure 11.33. Suppose a person could stand on the surface of a normal star with a flashlight. The light leaving the flashlight travels in a straight line no matter where the flashlight is pointed. Now consider what happens if the star collapses so that it is just a little larger than a black hole. All the light paths, except the one straight up, curve back to the surface. When the star shrinks inside the event horizon and becomes a black hole, even a beam directed straight up returns.

Keep in mind that gravity is not pulling on the light. The concentration of matter has curved spacetime, and light (like the trained ant of our earlier example) is "doing its best" to go in a straight line, yet is now confronted with a world in which straight lines that used to go outward have become curved paths that lead back in. The collapsing star is a *black hole* in this view, because the very concept of "out" has no geometrical meaning. The star has become trapped in its own little pocket of spacetime, from which there is no escape.

The star's geometry cuts off communication with the rest of the universe at precisely the moment when, in our earlier picture, the escape velocity becomes equal to the speed of light. The size of the star at this moment defines a surface that we call the **event horizon**. It's a wonderfully descriptive name: just as objects that sink below our horizon cannot be seen on Earth, so anything happening inside the event horizon can no longer interact with the rest of the universe.

Imagine a future spacecraft foolish enough to land on the surface of a massive star just as it begins to collapse in the way we have been describing. Perhaps the captain is asleep at the gravity meter, and before the crew can say "Albert Einstein," they have collapsed with the star inside the event horizon. Frantically, they send an escape pod straight outward. But paths outward twist around to become paths inward, and the pod turns around and falls toward the centre of the black hole. They send a radio message to their loved ones, bidding good-bye. But radio waves, like light, must travel through spacetime, and curved spacetime allows nothing to get out. Their final message remains unheard. Events inside the event horizon can never again affect events outside it.

The characteristics of an event horizon were first worked out by astronomer and mathematician Karl Schwarzschild, pictured in Figure 11.34. A member of the German army in World War I, he died in 1916 of an illness he contracted while doing artillery shell calculations on the Russian front. His paper on the theory of event horizons was among the last things he finished as he was dying; it was the first exact solution to Einstein's equations of general relativity. The radius of the event horizon is called the **Schwarzschild radius** in his memory.

Karl Schwarzschild (1873–1916)

Figure 11.34. This German scientist was the first to demonstrate mathematically that a black hole is possible and to determine the size of a nonrotating black hole's event horizon.

The event horizon is the boundary of the black hole; calculations show that it does not get smaller once the whole star has collapsed inside it. It is the region that separates the things trapped inside it from the rest of the universe. Anything coming from the outside is also trapped once it comes inside the event horizon. The horizon's size turns out to depend only on the mass inside it. If the Sun, with its mass of 1 M_{Sun}, were to become a black hole (fortunately, it can't—this is just a thought experiment), the Schwarzschild radius would be about 3 kilometres; thus, the entire black hole would be about one-third the size of a neutron star of that same mass. Feed the black hole some mass, and the horizon will grow—but not very much. Doubling the mass will make the black hole 6 kilometres in radius, still very tiny on the cosmic scale.

The event horizons of more massive black holes have larger radii. For example, if a globular cluster of 100,000 stars (solar masses) could collapse to a black hole, it would be 300,000 kilometres in radius, a little less than half the radius of the Sun. If the entire Galaxy could collapse to a black hole, it would be only about 10^{12} kilometres in radius—about a tenth of a light year. Smaller masses have correspondingly smaller horizons: for Earth to become a black hole, it would have to be compressed to a radius of only 1 centimetre—less than the size of a grape. A typical asteroid, if crushed to a small enough size to be a black hole, would have the dimensions of an atomic nucleus.

Example 11.2

The Milky Way's Black Hole

The size of the event horizon of a black hole depends on the mass of the black hole. The greater the mass, the larger the radius of the event horizon. General relativity calculations show that the formula for the Schwarzschild radius (R_S) of the event horizon is

$$R_\text{S} = \frac{2GM}{c^2}$$

where c is the speed of light, G is the gravitational constant, and M is the mass of the black hole. Note that in this formula, 2, G, and c are all constant; only the mass changes from black hole to black hole.

Astronomers have traced the paths of several stars near the centre of our Galaxy and found that they seem to be orbiting an unseen object—dubbed Sgr A* (pronounced "Sagittarius A-star")—with a mass of about 4 million solar masses. What is the size of its Schwarzschild radius?

Solution

We can substitute data for G, M, and c (from Table 2.4) directly into the equation:

$$\begin{aligned} R_S &= \frac{2GM}{c^2} = \frac{2\left(6.67 \times 10^{-11} \mathrm{N{\cdot}m^2/kg^2}\right)\left(4 \times 10^6\right)\left(1.99 \times 10^{30}\ \mathrm{kg}\right)}{\left(3.00 \times 10^8\ \mathrm{m/s}\right)^2} \\ &= 1.18\ \times\ 10^{10}\ \mathrm{m} \end{aligned}$$

This distance is about one-fifth of the radius of Mercury's orbit around the Sun, yet the object contains 4 million solar masses and cannot be seen with our largest telescopes. You can see why astronomers are convinced this object is a black hole.

Exercise 11.2

What would be the size of a black hole that contained only as much mass as a typical pickup truck (about 3000 kg)? (Note that something with so little mass could never actually form a black hole, but it's interesting to think about the result.)

Solution

Substituting the data into our equation gives

$$R_{\mathrm{S}} = \frac{2GM}{c^2} = \frac{2\left(6.67\ \times\ 10^{-11}\mathrm{N}\cdot\mathrm{m}^2/\mathrm{kg}^2\right)(3000\,\mathrm{kg})}{\left(3.00\ \times\ 10^8\ \mathrm{m/s}\right)^2} = 1.33\ \times\ 10^{-23}\ \mathrm{m}.$$

For comparison, the size of a proton is usually considered to be about 8×10^{-16} m, which would be about ten million times larger.

A Trip into a Black Hole

The fact that scientists cannot see inside black holes has not kept them from trying to calculate what they are like. One of the first things these calculations showed was that the formation of a black hole obliterates nearly all information about the star that collapsed to form it. Physicists like to say "black holes have no hair," meaning that nothing sticks out of a black hole to give us clues about what kind of star produced it or what material has fallen inside. The only information a black hole can reveal about itself is its mass, its spin (rotation), and whether it has any electrical charge.

What happens to the collapsing star-core that made the black hole? Our best calculations predict that the material will continue to collapse under its own weight, forming an infinitely *squozen* point—a place of zero volume and infinite density—to which we give the name **singularity**. At the singularity, spacetime ceases to exist. The laws of physics as we know them break down. We do not yet have the physical understanding or the mathematical tools to describe the singularity itself, or even if singularities actually occur. From the outside, however, the entire structure of a basic black hole (one that is not rotating) can be described as a singularity surrounded by an event horizon. Compared to humans, black holes are really very simple objects.

Scientists have also calculated what would happen if an astronaut were to fall into a black hole. Let's take up an observing position a long, safe distance away from the event horizon and watch this astronaut fall toward it. At first he falls away from us, moving ever faster, just as though he were approaching any massive star. However, as he nears the event horizon of the black hole, things change. The strong gravitational field around the black hole will make his clocks run more slowly, when seen from our outside perspective.

If, as he approaches the event horizon, he sends out a signal once per second according to his clock, we will see the spacing between his signals grow longer and longer until it becomes infinitely long when he reaches the event horizon. (Recalling our discussion of gravitational redshift, we could say that if the infalling astronaut uses a blue light to send his signals every second, we will see the light get redder and redder until its wavelength is nearly infinite.) As the spacing between clock ticks approaches infinity, it will appear to us that the astronaut is slowly coming to a stop, frozen in time at the event horizon.

In the same way, all matter falling into a black hole will also appear to an outside observer to stop at the event horizon, frozen in place and taking an infinite time to fall through it. But don't think that matter falling into a black hole will therefore be easily visible at the event horizon. The tremendous redshift will make it very difficult to observe any radiation from the "frozen" victims of the black hole.

This, however, is only how we, located far away from the black hole, see things. To the astronaut, his time goes at its normal rate and he falls right on through the event horizon into the black hole. (Remember, this horizon is not a physical barrier, but only a region in space where the curvature of spacetime makes escape impossible.)

You may have trouble with the idea that you (watching from far away) and the astronaut (falling in) have such different ideas about what has happened. This is the reason Einstein's ideas about space and time are called theories of *relativity*. What each observer measures about the world depends on (is relative to) his or her frame of reference. The observer in strong gravity measures time and space differently from the one sitting in weaker gravity. When Einstein proposed these ideas, many scientists also had difficulty with the idea that two such different views of the same event could be correct, each in its own "world," and they tried to find a mistake in the calculations. There were no mistakes: we and the astronaut really would see him fall into a black hole very differently.

For the astronaut, there is no turning back. Once inside the event horizon, the astronaut, along with any signals from his radio transmitter, will remain hidden forever from the universe outside. He will, however, not have a long time (from his perspective) to feel sorry for himself as he approaches the black hole. Suppose he is falling feet first. The force of gravity that the singularity exerts on his feet is greater than on his head, so he will be stretched slightly. Because the singularity is a point, the left side of his body will be pulled slightly toward the right, and the right slightly toward the left, bringing each side closer to the singularity. The astronaut will therefore be slightly squeezed in one direction and stretched in the other. Some scientists like to call this process of stretching and narrowing **spaghettification**. The point at which the astronaut becomes so stretched that he perishes depends on the size of the black hole. For black holes with masses billions of times the mass of the Sun, such as those found at the centres of galaxies, the spaghettification becomes significant only after the astronaut passes through the event horizon. For black holes with masses of a few solar masses, the astronaut will be stretched and ripped apart even before he reaches the event horizon.

Earth exerts similar *tidal forces* on an astronaut performing a spacewalk. In the case of Earth, the tidal forces are so small that they pose no threat to the health and safety of the astronaut. Not so in the case of a black hole. Sooner or later, as the astronaut approaches the black hole, the tidal forces will become so great that the astronaut will be ripped apart, eventually reduced to a collection of individual atoms that will continue their inexorable fall into the singularity.

Requirements for a Black Hole

So, here is a prescription for finding a black hole: start by looking for a star whose motion (determined from the Doppler shift of its spectral lines) shows it to be a member of a binary star system. If both stars are visible, neither can be a black hole, so focus your attention on just those systems where only one star of the pair is visible, even with our most sensitive telescopes.

Being invisible is not enough, however, because a relatively faint star might be hard to see next to the glare of a brilliant companion or if it is shrouded by dust. And even if the star really is invisible, it could be a neutron star. Therefore, we must also have evidence that the unseen star has a mass too high to be a neutron star and that it is a collapsed object—an extremely small stellar remnant.

We can use Kepler’s law and our knowledge of the visible star to measure the mass of the invisible member of the pair. If the mass is greater than about 3 M_{Sun}, then we are likely seeing (or, more precisely, not seeing) a black hole—as long as we can make sure the object really is a collapsed star.

If matter falls toward a compact object of high gravity, the material is accelerated to high speed. Near the event horizon of a black hole, matter is moving at velocities that approach the speed of light. As the atoms whirl chaotically toward the event horizon, they rub against each other; internal friction can heat them to temperatures of 100 million K or more. Such hot matter emits radiation in the form of flickering X-rays. The last part of our prescription, then, is to look for a source of X-rays associated with the binary system. Since X-rays do not penetrate Earth’s atmosphere, such sources must be found using X-ray telescopes in space.

In our example, the infalling gas that produces the X-ray emission comes from the black hole’s companion star. Stars in close binary systems can exchange mass, especially as one of the members expands into a red giant. Suppose that one star in a double-star system has evolved to a black hole and that the second star begins to expand. If the two stars are not too far apart, the outer layers of the expanding star may reach the point where the black hole exerts more gravitational force on them than do the inner layers of the red giant to which the atmosphere belongs. The outer atmosphere then passes through the point of no return between the stars and falls toward the black hole.

The mutual revolution of the giant star and the black hole causes the material falling toward the black hole to spiral around it rather than flow directly into it. The infalling gas whirls around the black hole in a pancake of matter called an **accretion disk**. It is within the inner part of this disk that matter is revolving about the black hole so fast that internal friction heats it up to X-ray–emitting temperatures (see the image below).

Another way to form an accretion disk in a binary star system is to have a powerful stellar wind come from the black hole's companion. Such winds are a characteristic of several stages in a star's life. Some of the ejected gas in the wind will then flow close enough to the black hole to be captured by it into the disk, shown in Figure 11.35.

Binary Black Hole

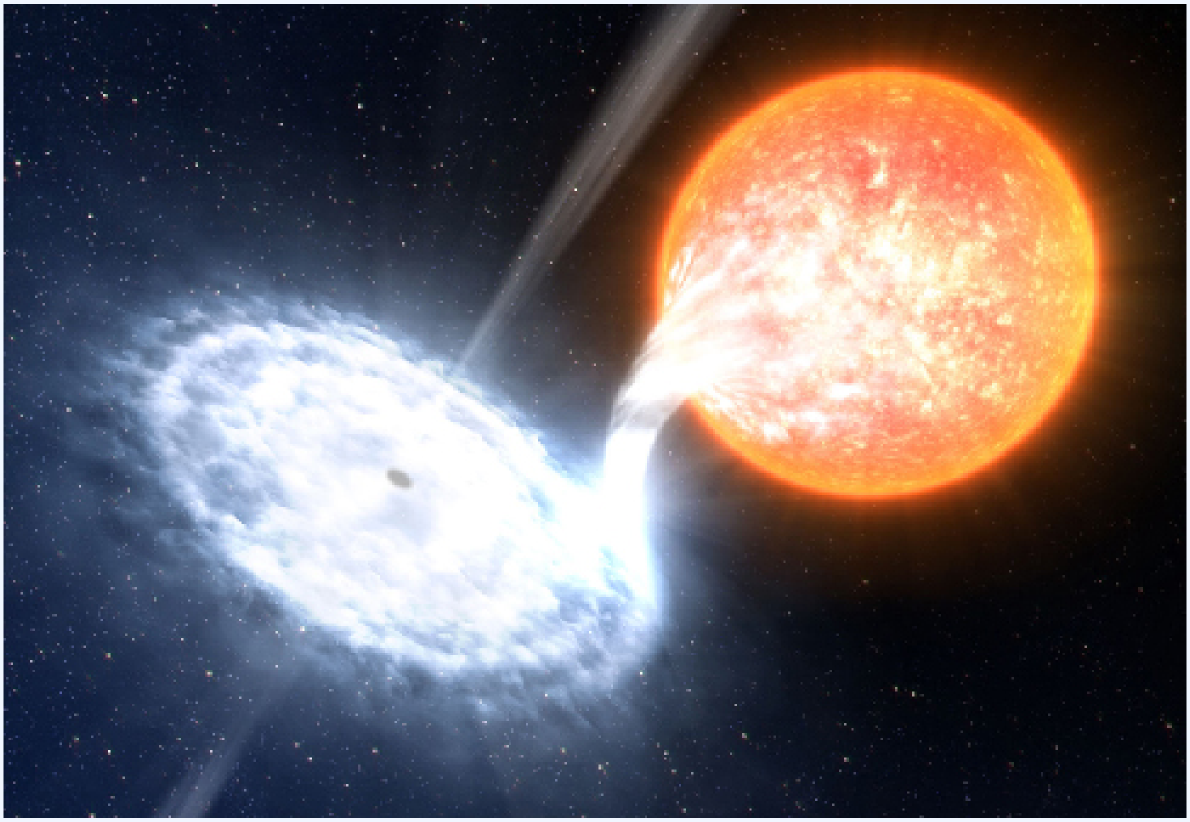


Figure 11.35. This artist's rendition shows a black hole and star (red). As matter streams from the star, it forms a disk around the black hole. Some of the swirling material close to the black hole is pushed outward perpendicular to the disk in two narrow jets.
Modification of image by ESO/L. Calçada, CC BY-4.0.

We should point out that, as often happens, the measurements we have been discussing are not quite as simple as they are described in introductory textbooks. In real life, Kepler's law allows us to calculate only the combined mass of the two stars in the binary system. We must learn more about the visible star of the pair and its history to ascertain the distance to the binary pair, the true size of the visible star's orbit, and how the orbit of the two stars is tilted toward Earth, something we can rarely measure. And neutron stars can also have accretion disks that produce X-rays, so astronomers must study the properties of these X-rays carefully when trying to determine what kind of object is at the centre of the disk. Nevertheless, a number of systems that clearly contain black holes have now been found.

Attribution

“24.1 Introducing General Relativity“, “24.5 Black Holes“, and “24.6 Evidence for Black Holes” from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, are licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

11.14 KEY TERMS

Accretion disk: the disk of gas and dust found orbiting newborn stars, as well as compact stellar remnants such as white dwarfs, neutron stars, and black holes when they are in binary systems and are sufficiently close to their binary companions to draw off material. 11.13

Balmer lines: the hydrogen lines in the visible part of the spectrum. 11.2

Black dwarf: a former white dwarf star that has become a cold stellar corpse with the mass of a star and the size of a planet after many billions of years. 11.6

Black hole: a region in spacetime where gravity is so strong that nothing—not even light—can escape. 11.13

Cataclysmic Variables: binary stars which consist of a white dwarf primary and an orbiting secondary star. 11.7

Cepheid Variables: an important class of pulsating variable stars. 11.7

Chandrasekhar limit: the upper limit to the mass of a white dwarf (equals 1.4 times the mass of the Sun). 11.6

Degenerate gas: a gas that resists further compression because no two electrons can be in the same place at the same time doing the same thing (Pauli exclusion principle). 11.6

Diamond-like star: a degenerate (white dwarf) star that cools down to the point where the atoms inside it in essence "solidify" into a giant, highly compact lattice (organized rows of atoms, just like in a crystal). 11.6

Dwarf stars: stars with radii equal to, or less than, that of the Sun. 11.3

Event horizon: a boundary in spacetime such that events inside the boundary can have no effect on the world outside it—that is, the boundary of the region around a black hole where the curvature of spacetime no longer provides any way out. 11.13

General theory of relativity: Einstein's theory relating gravity and the structure (geometry) of space and time. 11.13

Giant stars: stars with radii between 10 and 100 times that of the Sun. 11.3

Hertzsprung–Russell (H–R) diagram: diagram that summarizes the relationship between the luminosity and surface temperature of stars. 11.4

High-mass star: any star with mass exceeding four times that of the Sun. 11.3

Ia supernovae: a white dwarf explosion where the white dwarf is completely destroyed, leaving behind no remnant. 11.8

Initial mass: the mass a star possesses at the beginning of its lifetime before it begins to lose mass in the process of aging and dying. 11.5

Intermediate-mass stars: stars with a mass between 1.33 and 4 times the mass of the Sun. 11.3

Low-mass stars: the Sun, as well as all other stars with a mass less than 1.33 times the mass of the Sun. 11.3

Neutrino: ghostly subatomic particles that carry away some of the nuclear energy when an electron and a proton in the star's core merge to make a neutron. 11.10

Neutron star: a compact object of extremely high density composed almost entirely of neutrons. 11.10

Novae: original name used for cataclysmic variables as the star seemingly appeared out of nowhere. 11.7

Nucleosynthesis: the building up of heavy elements from lighter ones by nuclear fusion. 11.9

Pulsar: a variable radio source of small physical size that emits very rapid radio pulses in very regular periods that range from fractions of a second to several seconds; now understood to be a rotating, magnetic neutron star that is energetic enough to produce a detectable beam of radiation and particles. 11.12

Pulsating Variables: stars that swell and shrink, which affects the star's brightness. 11.7

Schwarzschild radius: the radius of the event horizon. 11.13

Singularity: the point of zero volume and infinite density to which any object that becomes a black hole must collapse, according to the theory of general relativity. 11.13

Spaghettification: the process of stretching and narrowing that occurs due to the force of gravity that a singularity exerts. 11.13

Spectral class: the classification of stars according to their temperatures using the characteristics of their spectra; the types are O, B, A, F, G, K, and M with L, T, and Y added recently for cooler star-like objects that recent survey have revealed. 11.2

Supergiant stars: stars with radii more than 100 times than that of the Sun. 11.3

Supernova: an explosion when an enormous amount of fusion (especially of carbon) takes place all at once. 11.8

Triple-alpha process: the process by which three helium atoms can begin to fuse to form a single carbon nucleus once the temperature in a star's core reaches that of 100 million K (but not before such point). 11.5

Type II supernova: a stellar explosion produced at the endpoint of the evolution of stars whose mass exceeds roughly 10 times the mass of the Sun. 11.10

Variable stars (variables): when a star changes luminosity due to some sort of physical characteristic, not due to an eclipse. 11.7

White dwarf: a star that has achieved a final state of equilibrium after the helium in its core has been exhausted. 11.6

Zero-age: the time when a star stops contracting, settles onto the main sequence, and begins to fuse hydrogen in its core. 11.5

CHAPTER 12: GALAXIES: THE MILKY WAY AND ITS NEIGHBOURHOOD

Chapter Overview

12.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Define the major components of the Milky Way Galaxy and its position within the universe.
- Explain the significance of dark matter in the Milky Way Galaxy and how it affects the motions of luminous matter.
- Analyze the similarities and differences between spiral galaxies and elliptical galaxies, using specific examples.
- Interpret the evidence supporting the existence of tidal streams and their role in shaping the structure of the Milky Way.
- Compare and contrast the Monolithic Collapse Model and the Multiple Merger Model for galaxy formation.

12.1 THE MILKY WAY GALAXY

Milky Way Galaxy

Figure 12.1. The Milky Way rises over Square Tower, an ancestral pueblo building at Hovenweep National Monument in Utah. Many stars and dark clouds of dust combine to make a spectacular celestial sight of our home Galaxy. The location has been designated an International Dark Sky Park by the International Dark Sky Association.
The Milky Way above Hovenweep Castle by NPS/Jacob W. Frank, Public Domain

Today, we know that our Sun is just one of the many billions of stars that make up the huge cosmic island we call the **Milky Way Galaxy**. How can we "weigh" such an enormous system of stars and measure its total mass?

One of the most striking features you can see in a truly dark sky—one without light pollution—is the band of faint white light called the Milky Way, which stretches from one horizon to the other. The name comes from an ancient Greek legend that compared its faint white splash of light to a stream of spilled milk. But folktales differ from culture to culture: one East African tribe thought of the hazy band as the smoke of ancient campfires, several Native American stories tell of a path across the sky traveled by sacred animals, and in Siberia, the diffuse arc was known as the seam of the tent of the sky.

In 1610, Galileo made the first telescopic survey of the Milky Way and discovered that it is composed of a multitude of individual stars. Today, we know that the Milky Way comprises our view inward of the huge cosmic pinwheel that we call the Milky Way Galaxy and that is our home. Moreover, our Galaxy is now recognized as just one galaxy among many billions of other galaxies in the cosmos.

The Milky Way Galaxy surrounds us, and you might think it is easy to study because it is so close. However, the very fact that we are embedded within it presents a difficult challenge. Suppose you were given the task of mapping New York City. You could do a much better job from a helicopter flying over the city than you could if you were standing in Times Square. Similarly, it would be easier to map our Galaxy if we could only get a little way outside it, but instead we are trapped inside and way out in its suburbs—far from the galactic equivalent of Times Square.

With modern instruments, astronomers can now penetrate the "smog" of the Milky Way by studying radio and infrared emissions from distant parts of the Galaxy. Measurements at these wavelengths (as well as observations of other galaxies like ours) have given us a good idea of what the Milky Way would look like if we *could* observe it from a distance.

Figure 12.2 sketches what we would see if we could view the Galaxy face-on and edge-on. The brightest part of the Galaxy consists of a thin, circular, rotating disk of stars distributed across a region about 100,000 light-years in diameter and about 1000 light-years thick. (Given how thin the disk is, perhaps a CD is a more appropriate analogy than a wheel.) In addition to stars, the dust and gas from which stars form are also found mostly in the thin disk of the Galaxy. The mass of the interstellar matter is about 15% of the mass of the stars in this disk.

Schematic Representation of the Galaxy

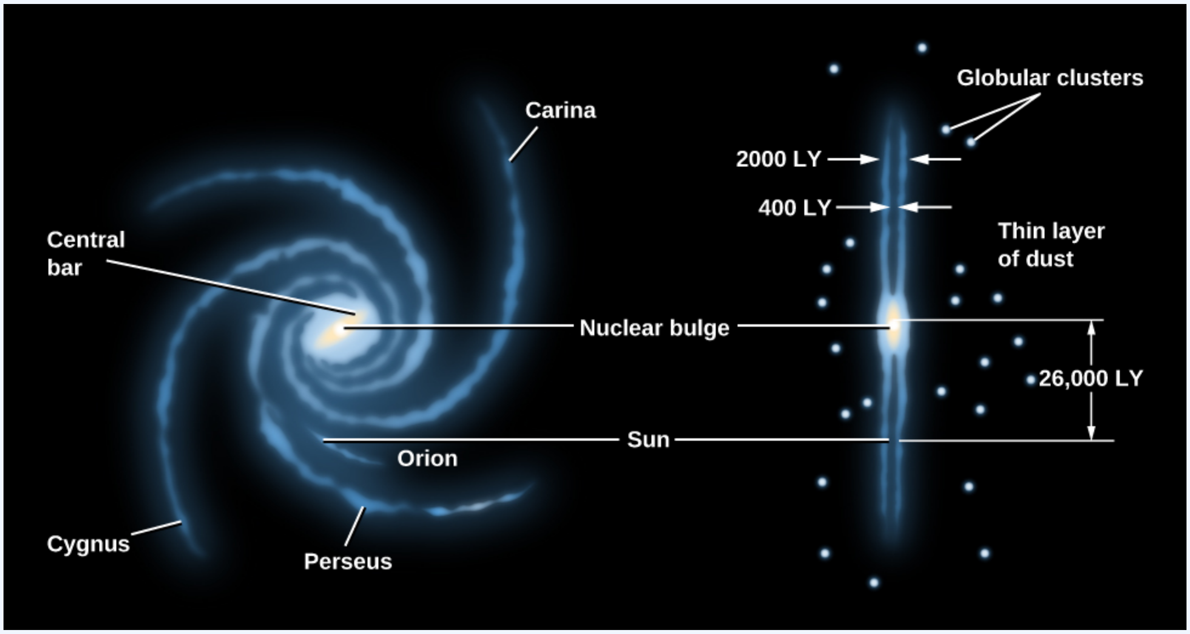


Figure 12.2. The left image shows the face-on view of the spiral disk; the right image shows the view looking edge-on along the disk. The major spiral arms are labelled. The Sun is located on the inside edge of the short Orion spur.

As the diagram in Figure 12.2 shows, the stars, gas, and dust are not spread evenly throughout the disk but are concentrated into a central bar and a series of spiral arms. Recent infrared observations have confirmed that the central bar is composed mostly of old yellow-red stars. The two main spiral arms appear to connect with the ends of the bar. They are highlighted by the blue light from young hot stars. We know many other spiral galaxies that also have bar-shaped concentrations of stars in their central regions; for that reason they are called **barred spirals**. Figure 12.3 shows two other galaxies—one without a bar and one with a strong bar—to give you a basis for comparison to our own. We will describe our spiral structure in more detail shortly. The Sun is located about halfway between the centre of the Galaxy and the edge of the disk and only about 70 light-years above its central plane.

Unbarred and Barred Spiral Galaxies

Figure 12.3. (a) This image shows the unbarred spiral galaxy M74. It contains a small central bulge of mostly old yellow-red stars, along with spiral arms that are highlighted with the blue light from young hot stars. (b) This image shows the strongly barred spiral galaxy NGC 1365. The bulge and the fainter bar both appear yellowish because the brightest stars in them are mostly old yellow and red giants. Two main spiral arms project from the ends of the bar. As in M74, these spiral arms are populated with blue stars and red patches of glowing gas—hallmarks of recent star formation. The Milky Way Galaxy is thought to have a barred spiral structure that is intermediate between these two examples.
Credit a: PESSTO snaps Supernova in Messier 74 by ESO/PESSTO/S. Smartt, CC-BY-4.0.
Credit b: Fine Details in a Barred Galaxy by ESO, CC BY-4.0.

Our thin disk of young stars, gas, and dust is embedded in a thicker but more diffuse disk of older stars; this thicker disk extends about 3000 light-years above and below the midplane of the thin disk and contains only about 5% as much mass as the thin disk.

Close in to the galactic centre (within about 10,000 light-years), the stars are no longer confined to the disk but form a **central bulge** (or nuclear bulge). When we observe with visible light, we can glimpse the stars in the bulge only in those rare directions where there happens to be relatively little interstellar dust. The first picture that actually succeeded in showing the bulge as a whole was taken at infrared wavelengths as shown in Figure 12.4.

Inner Part of the Milky Way Galaxy

Figure 12.4. This beautiful infrared map, showing half a billion stars, was obtained as part of the Two Micron All Sky Survey (2MASS). Because interstellar dust does not absorb infrared as strongly as visible light, this view reveals the previously hidden bulge of old stars that surrounds the centre of our Galaxy, along with the Galaxy's thin disk component.
Two Micron All Sky Survey by NASA/2MASS, NASA Media License.

The fact that much of the bulge is obscured by dust makes its shape difficult to determine. For a long time, astronomers assumed it was spherical. However, infrared images and other data indicate that the bulge is about two times longer than it is wide, and shaped rather like a peanut. The relationship between this elongated inner bulge and the larger bar of stars remains uncertain. At the very centre of the nuclear bulge is a tremendous concentration of matter, which we will discuss later in this chapter.

In our Galaxy, the thin and thick disks and the nuclear bulge are embedded in a spherical **halo** of very old, faint stars that extends to a distance of at least 150,000 light-years from the galactic centre. Most of the globular clusters are also found in this halo.

The mass in the Milky Way extends even farther out, well beyond the boundary of the luminous stars to a distance of at least 200,000 light-years from the centre of the Galaxy. This invisible mass has been give the name *dark matter* because it emits no light and cannot be seen with any telescope. Its composition is unknown, and it can be detected only because of its gravitational effects on the motions of luminous matter that we can see. We know that this extensive **dark matter halo** exists because of its effects on the orbits of distant star clusters and other dwarf galaxies that are associated with the Galaxy.

Some vital statistics of the thin and thick disks and the stellar halo are given in Table 12.1, with an illustration in Figure 12.5. Note particularly how the ages of stars correlate with where they are found. As we shall see, this information holds important clues to how the Milky Way Galaxy formed.

Table 12.1. Characteristics of the Milky Way Galaxy

Property	Thin Disk	Thick Disk	Stellar Halo (Excludes Dark Matter)
Stellar mass	$4 \times 10^{10} M_{\text{Sun}}$	A few percent of the thin disk mass	$10^{10} M_{\text{Sun}}$
Luminosity	$3 \times 10^{10} L_{\text{Sun}}$	A few percent of the thin disk luminosity	$8 \times 10^{8} L_{\text{Sun}}$
Typical age of stars	1 million to 10 billion years	11 billion years	13 billion years
Heavier-element abundance	High	Intermediate	Very low
Rotation	High	Intermediate	Very low

Major Parts of the Milky Way Galaxy

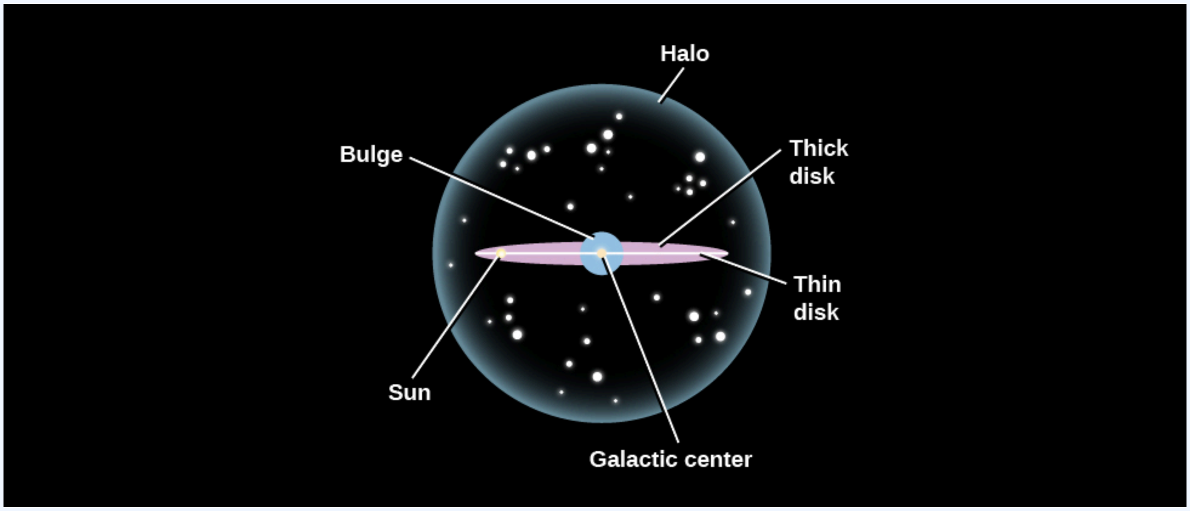


Figure 12.5. This schematic shows the major components of our Galaxy.

Establishing this overall picture of the Galaxy from our dust-shrouded viewpoint inside the thin disk has been one of the great achievements of modern astronomy (and one that took decades of effort by astronomers working with a wide range of telescopes). One thing that helped enormously was the discovery that our Galaxy is not unique in its characteristics. There are many other flat, spiral-shaped islands of stars, gas, and dust in the universe. For example, the Milky Way somewhat resembles the Andromeda galaxy, which, at a distance of about 2.3 million light-years, is our nearest neighbouring giant spiral galaxy. Just as you can get a much better picture of yourself if someone else takes the photo from a distance away, pictures and other diagnostic observations of nearby galaxies that resemble ours have been vital to our understanding of the properties of the Milky Way.

Our radio observations of the disk's gaseous component indicate that the Galaxy has two major spiral arms that emerge from the bar and several fainter arms and shorter spurs. You can see a recently assembled map of our Galaxy's arm structure—derived from studies in the infrared—in Figure 12.6.

Milky Way Bar and Arms

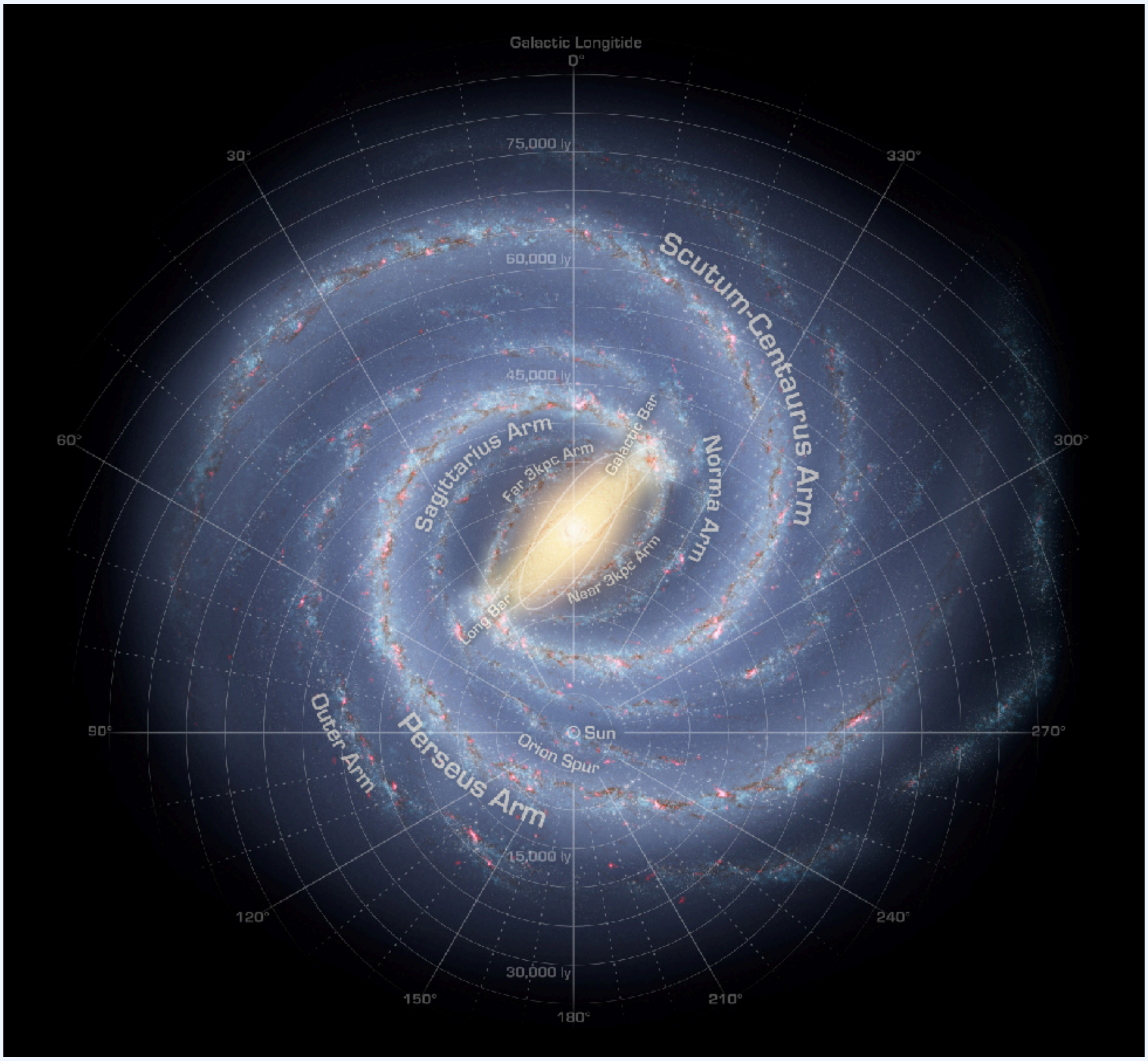


Figure 12.6. Here, we see the Milky Way Galaxy as it would look from above. This image, assembled from data from NASA's WISE mission, shows that the Milky Way Galaxy has a modest bar in its central regions. Two spiral arms, Scutum-Centaurus and Perseus, emerge from the ends of the bar and wrap around the bulge. The Sagittarius and Outer arms have fewer stars than the other two arms.
The Milky Way Galaxy by NASA/JPL-Caltech/R. Hurt (SSC/Caltech), NASA Media License.

The Sun is near the inner edge of a short arm called the Orion Spur, which is about 10,000 light-years long and contains such conspicuous features as the Cygnus Rift (the great dark nebula in the summer Milky Way) and the bright Orion Nebula. Figure 12.7 shows a few other objects that share this small section of the Galaxy with us and are easy to see. Remember, the farther away we try to look from our own arm, the more the dust in the Galaxy builds up and makes it hard to see with visible light.

Orion Spur

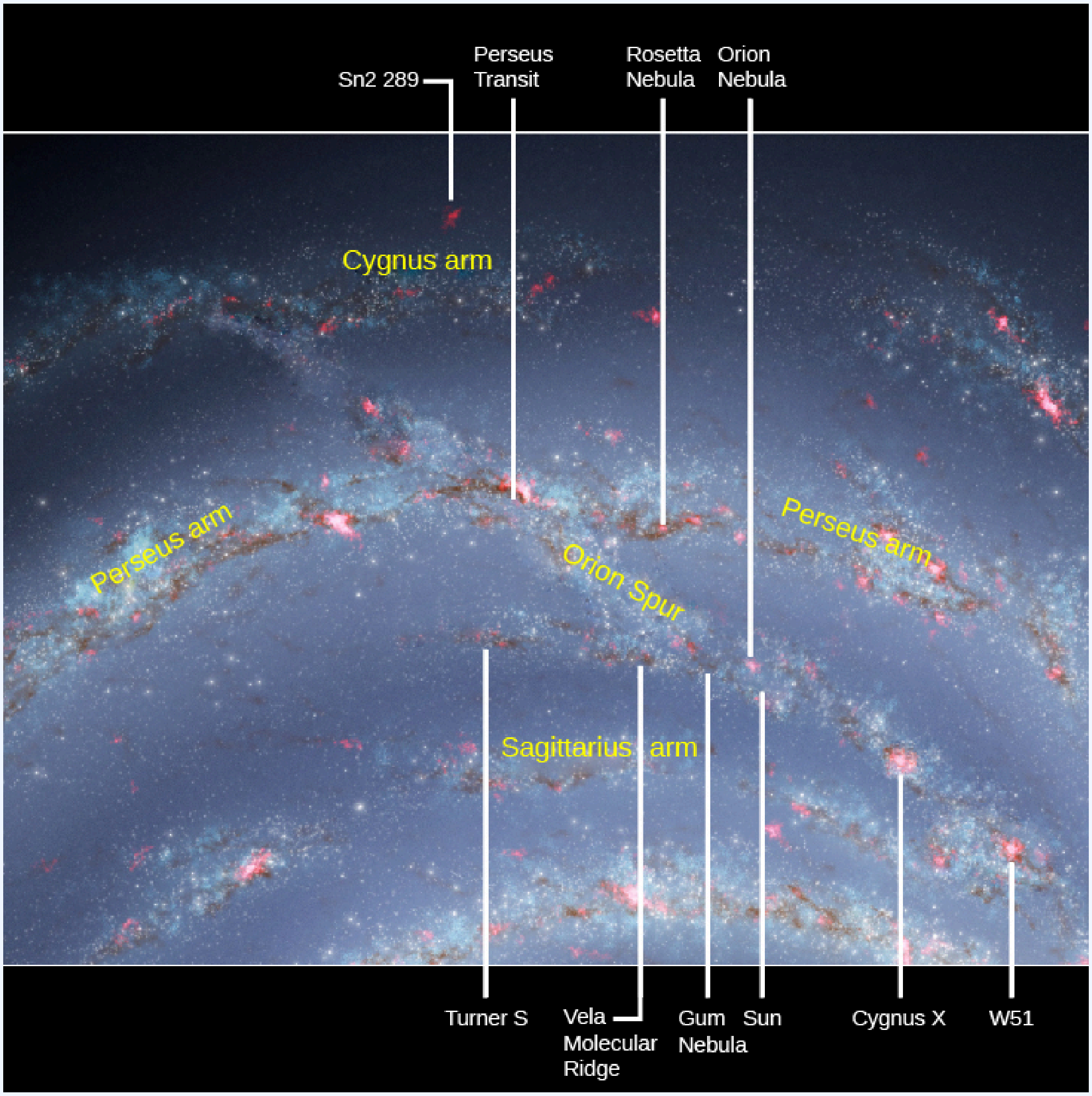


Figure 12.7. The Sun is located in the Orion Spur, which is a minor spiral arm located between two other arms. In this diagram, the white lines point to some other noteworthy objects that share this feature of the Milky Way Galaxy with the Sun.
Modification of The Milky Way by NASA/JPL-Caltech, NASA Media License.

Attribution

"25.0 Thinking Ahead", "25.1 The Architecture of the Galaxy", and "25.2 Spiral Structure" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

12.2 THE DISTRIBUTION OF GALAXIES IN SPACE

The region of the universe for which we have the most detailed information is, as you would expect, our own local neighbourhood. It turns out that the Milky Way Galaxy is a member of a small group of galaxies called, not too imaginatively, the **Local Group**. It is spread over about 3 million light-years and contains more than 54 members. There are three large spiral galaxies (our own, the Andromeda galaxy, and M33), two intermediate ellipticals, and many dwarf ellipticals and irregular galaxies.

Several new dwarf galaxies have also been found near the Andromeda galaxy. Such dwarf galaxies are difficult to find because they typically contain relatively few stars, and it is hard to distinguish them from the foreground stars in our own Milky Way.

Figure 12.8 is a rough sketch showing where the brighter members of the Local Group are located. The average of the motions of all the galaxies in the Local Group indicates that its total mass is about $4 \times 10^{12} M_{\text{Sun}}$, and at least half of this mass is contained in the two giant spirals—the Andromeda galaxy and the Milky Way Galaxy. And bear in mind that a substantial amount of the mass in the Local Group is in the form of dark matter.

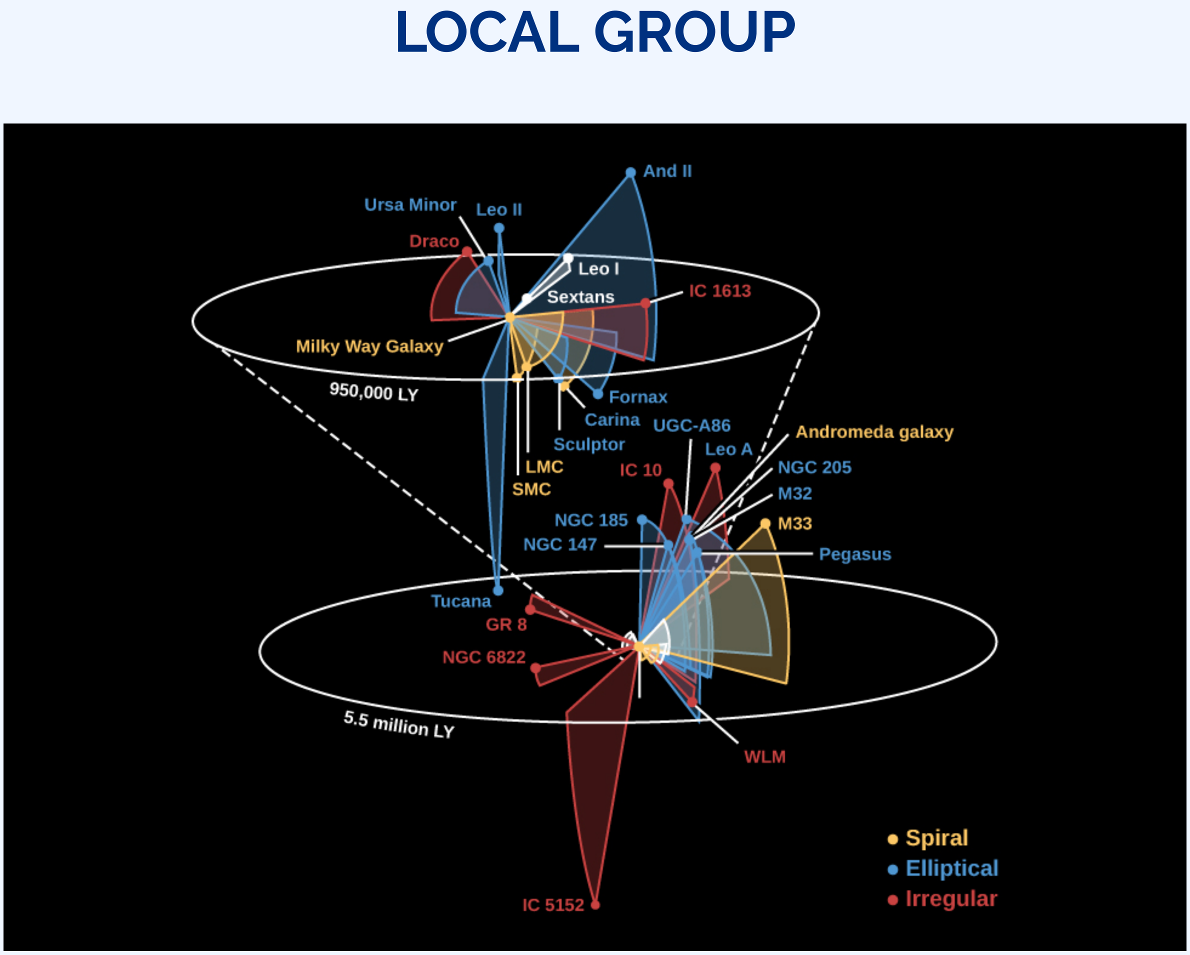


Figure 12.8. This illustration shows some members of the Local Group of galaxies, with our Milky Way at the centre. The exploded view at the top shows the region closest to the Milky Way and fits into the bigger view at the bottom as shown by the dashed lines. The three largest galaxies among the three dozen or so members of the Local Group are all spirals; the others are small irregular galaxies and dwarf ellipticals. A number of new members of the group have been found since this map was made.

Small galaxy groups like ours are hard to notice at larger distances. However, there are much more substantial groups called galaxy clusters that are easier to spot even many millions of light-years away. Such clusters are described as *poor* or *rich* depending on how many galaxies they contain. Rich clusters have thousands or even tens of thousands of galaxies, although many of the galaxies are quite faint and hard to detect.

The nearest moderately rich galaxy cluster is called the Virgo Cluster, after the constellation in which it is seen. It is about 50 million light-years away and contains thousands of members, of which a few are shown in Figure 12.9. The giant elliptical (and very active) galaxy M87 belongs to the Virgo Cluster.

Central Region of the Virgo Cluster

Figure 12.9. Virgo is the nearest rich cluster and is at a distance of about 50 million light-years. It contains hundreds of bright galaxies. In this picture you can see only the central part of the cluster, including the giant elliptical galaxy M87, just below centre. Other spirals and ellipticals are visible; the two galaxies to the top right are known as “The Eyes.”
Giant Galaxy is Still Growing by Chris Mihos (Case Western Reserve University)/ESO, CC BY 4.0.

Attribution

12.3 THE CLASSIFICATION OF GALAXIES

Our own Galaxy and the Andromeda galaxy are typical, large spiral galaxies (see images below). They consist of a central bulge, a halo, a disk, and spiral arms. Interstellar material is usually spread throughout the disks of spiral galaxies. Bright emission nebulae and hot, young stars are present, especially in the spiral arms, showing that new star formation is still occurring. The disks are often dusty, which is especially noticeable in those systems that we view almost edge on as shown in Figure 12.10.

Spiral Galaxies

(a) (b)

Figure 12.10. (a) The spiral arms of M100, shown here, are bluer than the rest of the galaxy, indicating young, high-mass stars and star-forming regions. (b) We view this spiral galaxy, NGC 4565, almost exactly edge on, and from this angle, we can see the dust in the plane of the galaxy; it appears dark because it absorbs the light from the stars in the galaxy.
Credit a: Messier 100 galaxy Hubble by Hubble Legacy Archive, NASA, ESA, and Judy Schmidt, ESA Standard License.
Credit b: NGC 4565 and 4562 by "Jschulman555"/ Wikimedia, CC BY 4.0.

In galaxies that we see face on, the bright stars and emission nebulae make the arms of spirals stand out like those of a pinwheel on the fourth of July. Open star clusters can be seen in the arms of nearer spirals, and globular clusters are often visible in their halos. Spiral galaxies contain a mixture of young and old stars, just as the Milky Way does. All spirals rotate, and the direction of their spin is such that the arms appear to trail much like the wake of a boat.

About two-thirds of the nearby spiral galaxies have boxy or peanut-shaped bars of stars running through their centres as shown in Figure 12.11. Showing great originality, astronomers call these galaxies barred spirals.

Barred Spiral Galaxy

Figure 12.11. NGC 1300, shown here, is a barred spiral galaxy. Note that the spiral arms begin at the ends of the bar.
Barred Spiral Galaxy NGC 1300 by NASA, ESA, and the Hubble Heritage Team(STScI/AURA), NASA Media License.

In both barred and unbarred spiral galaxies, we observe a range of different shapes. At one extreme, the central bulge is large and luminous, the arms are faint and tightly coiled, and bright emission nebulae and supergiant stars are inconspicuous. Hubble, who developed a system of classifying galaxies by shape, gave these galaxies the designation Sa. Galaxies at this extreme may have no clear spiral arm structure, resulting in a lens-like appearance (they are sometimes referred to as lenticular galaxies). These galaxies seem to share as many properties with **elliptical galaxies** as they do with spiral galaxies

At the other extreme, the central bulge is small and the arms are loosely wound. In these Sc galaxies, luminous stars and emission nebulae are very prominent. Our Galaxy and the Andromeda galaxy are both intermediate between the two extremes. Photographs of spiral galaxies, illustrating the different types, are shown in Figure 12.12, along with elliptical galaxies for comparison.

Hubble Classification of Galaxies

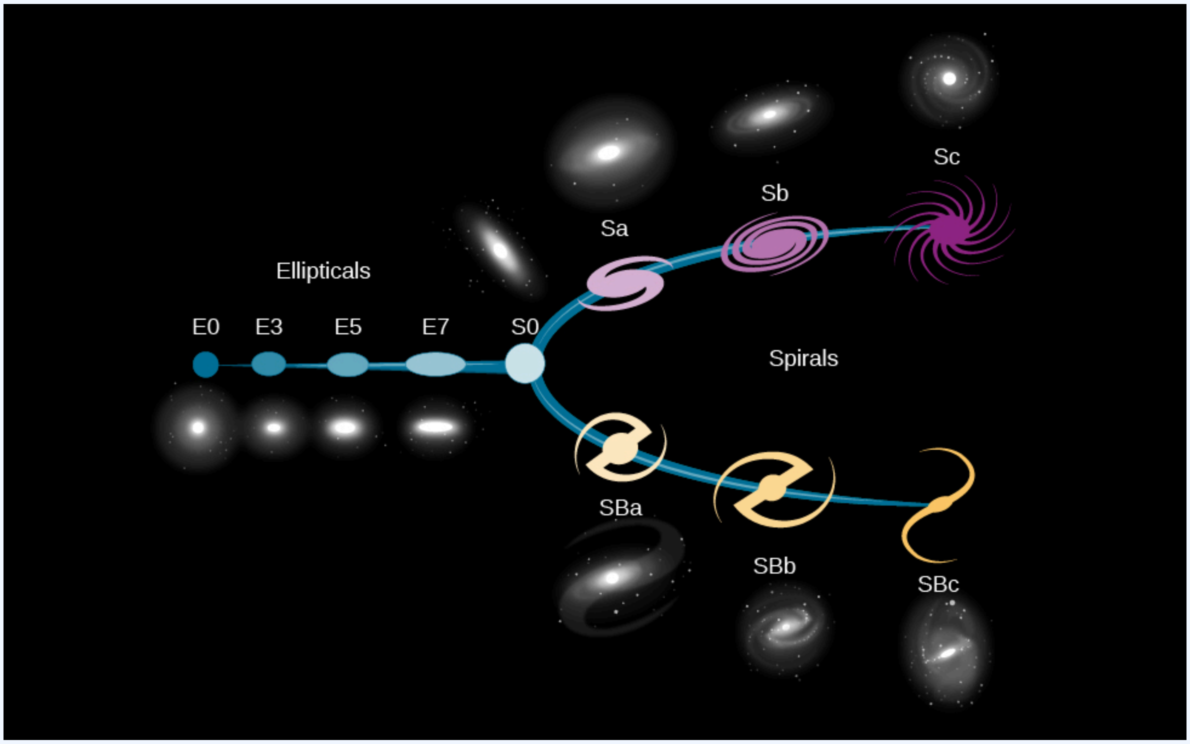


Figure 12.12. This figure shows Edwin Hubble's original classification of galaxies. Elliptical galaxies are on the left. On the right, you can see the basic spiral shapes illustrated, alongside images of actual barred and unbarred spirals.
The Hubble Tuning Fork by NASA, ESA, ESA Standard License.

The luminous parts of spiral galaxies appear to range in diameter from about 20,000 to more than 100,000 light-years. Recent studies have found that there is probably a large amount of galactic material that extends well beyond the apparent edge of galaxies. This material appears to be thin, cold gas that is difficult to detect in most observations.

From the observational data available, the masses of the visible portions of spiral galaxies are estimated to range from 1 billion to 1 trillion Suns (10^9 to $10^{12} M_{\text{Sun}}$). The total luminosities of most spirals fall in the range of 100 million to 100 billion times the luminosity of our Sun (10^8 to $10^{11} L_{\text{Sun}}$). Our Galaxy and M31 are relatively large and massive, as spirals go. There is also considerable dark matter in and around the galaxies, just as there is in the Milky Way; we deduce its presence from how fast stars in the outer parts of the Galaxy are moving in their orbits.

Elliptical galaxies consist almost entirely of old stars and have shapes that are spheres or ellipsoids (somewhat squashed spheres) as shown in Figure 12.13 They contain no trace of spiral arms. Their light is dominated by older reddish stars. In the larger nearby ellipticals, many globular clusters can be identified. Dust and emission nebulae are not conspicuous in elliptical galaxies, but many do contain a small amount of interstellar matter.

Elliptical Galaxies

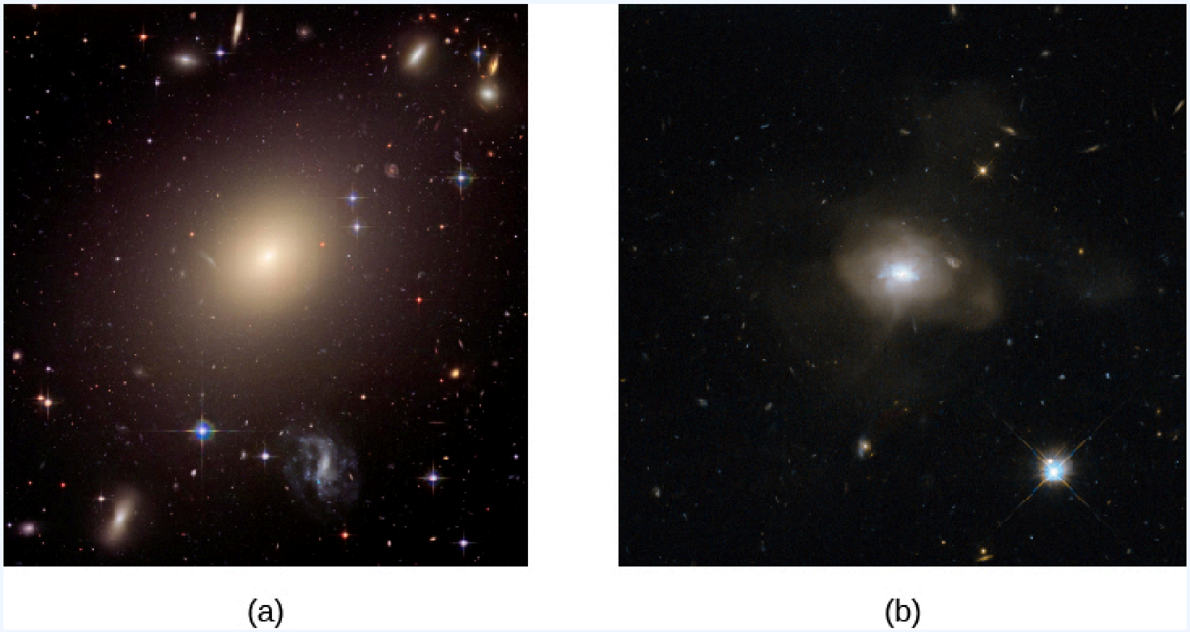


Figure 12.13. (a) ESO 325-G004 is a giant elliptical galaxy. Other elliptical galaxies can be seen around the edges of this image. (b) This elliptical galaxy probably originated from the collision of two spiral galaxies.
Credit a: Hubble Illuminates Cluster of Diverse Galaxies by NASA, ESA, and The Hubble Heritage Team (STScI/AURA), NASA Media License.
Credit b: Hubble Finds the Calm after the Galactic Storm by NASA/Hubble, NASA, NASA Media Licence.

Elliptical galaxies show various degrees of flattening, ranging from systems that are approximately spherical to those that approach the flatness of spirals. The rare giant ellipticals (for example, ESO 325-G004 in Figure 12.13) reach luminosities of $10^{11} L_{\text{Sun}}$. The mass in a giant elliptical can be as large as $10^{13} M_{\text{Sun}}$. The diametres of these large galaxies extend over several hundred thousand light-years and are considerably larger than the largest spirals. Although individual stars orbit the centre of an elliptical galaxy, the orbits are not all in the same direction, as occurs in spirals. Therefore, ellipticals don't appear to rotate in a systematic way, making it difficult to estimate how much dark matter they contain.

We find that elliptical galaxies range all the way from the giants, just described, to dwarfs, which may be the most common kind of galaxy. *Dwarf ellipticals* (sometimes called dwarf spheroidals) escaped our notice for a long time because they are very faint and difficult to see. An example of a dwarf elliptical is the Leo I Dwarf Spheroidal galaxy shown in Figure 12.14. The luminosity of this typical dwarf is about equal to that of the brightest globular clusters.

Intermediate between the giant and dwarf elliptical galaxies are systems such as M32 and M110, the two companions of the Andromeda galaxy. While they are often referred to as dwarf ellipticals, these galaxies are significantly larger than galaxies such as Leo I.

Dwarf Elliptical Galaxy

Figure 12.14. M32, a dwarf elliptical galaxy and one of the companions to the giant Andromeda galaxy M31. M32 is a dwarf by galactic standards, as it is only 2400 light-years across.
M32, NGC 221 by NOAO/AURA/NSF, CC BY 4.0.

Hubble classified galaxies that do not have the regular shapes associated with the categories we just described into the catchall bin of an **irregular galaxy**, and we continue to use his term. Typically, irregular galaxies have lower masses and luminosities than spiral galaxies. Irregular galaxies often appear disorganized, and many are undergoing relatively intense star formation activity. They contain both young population I stars and old population II stars.

The two best-known irregular galaxies are the Large Magellanic Cloud and Small Magellanic Cloud, as shown in Figure 12.15, which are at a distance of a little more than 160,000 light-years away and are among our nearest extragalactic neighbours. Their names reflect the fact that Ferdinand Magellan and his crew, making their round-the-world journey, were the first European travellers to notice them. Although not visible from the United States and Europe, these two systems are prominent from the Southern Hemisphere, where they look like wispy clouds in the night sky. Since they are only about one-tenth as distant as the Andromeda galaxy, they present an excellent opportunity for astronomers to study nebulae, star clusters, variable stars, and other key objects in the setting of another galaxy. For example, the Large Magellanic Cloud contains the 30 Doradus complex (also known as the Tarantula Nebula), one of the largest and most luminous groups of supergiant stars known in any galaxy.

4-Meter Telescope at Cerro Tololo Inter-American Observatory Silhouetted against the Southern Sky

Figure 12.15 The Milky Way is seen to the right of the dome, and the Large and Small Magellanic Clouds are seen to the left.
Blanco Telescope silhouetted against night sky by Roger Smith/NOAO/AURA/NSF, CC BY 4.0.

The Small Magellanic Cloud is considerably less massive than the Large Magellanic Cloud, and it is six times longer than it is wide. This narrow wisp of material points directly toward our Galaxy like an arrow. The Small Magellanic Cloud was most likely contorted into its current shape through gravitational interactions with the

Milky Way. A large trail of debris from this interaction between the Milky Way and the Small Magellanic Cloud has been strewn across the sky and is seen as a series of gas clouds moving at abnormally high velocity, known as the Magellanic Stream. We will see that this kind of interaction between galaxies will help explain the irregular shapes of this whole category of small galaxies,

View this beautiful album showcasing the different types of galaxies that have been photographed by the Hubble Space Telescope. Direct link: https://hubblesite.org/resource-gallery .

Attribution

"26.2 Types of Galaxies" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

12.4 THE FORMATION OF THE MILKY WAY

Information about stellar populations holds vital clues to how our Galaxy was built up over time. The flattened disk shape of the Galaxy suggests that it formed through a process similar to the one that leads to the formation of a protostar. Building on this idea, astronomers first developed models that assumed the Galaxy formed from a single rotating cloud. But, as we shall see, this turns out to be only part of the story.

The Protogalactic Cloud and the Monolithic Collapse Model

Because the oldest stars—those in the halo and in globular clusters—are distributed in a sphere centered on the nucleus of the Galaxy, it makes sense to assume that the ***protogalactic* cloud** that gave birth to our Galaxy was roughly spherical. The oldest stars in the halo have ages of 12 to 13 billion years, so we estimate that the formation of the Galaxy began about that long ago. Then, just as in the case of star formation, the protogalactic cloud collapsed and formed a thin rotating disk. Stars born before the cloud collapsed did not participate in the collapse, but have continued to orbit in the halo to the present day as shown in Figure 12.16.

Monolithic Collapse Model for the Formation of the Galaxy

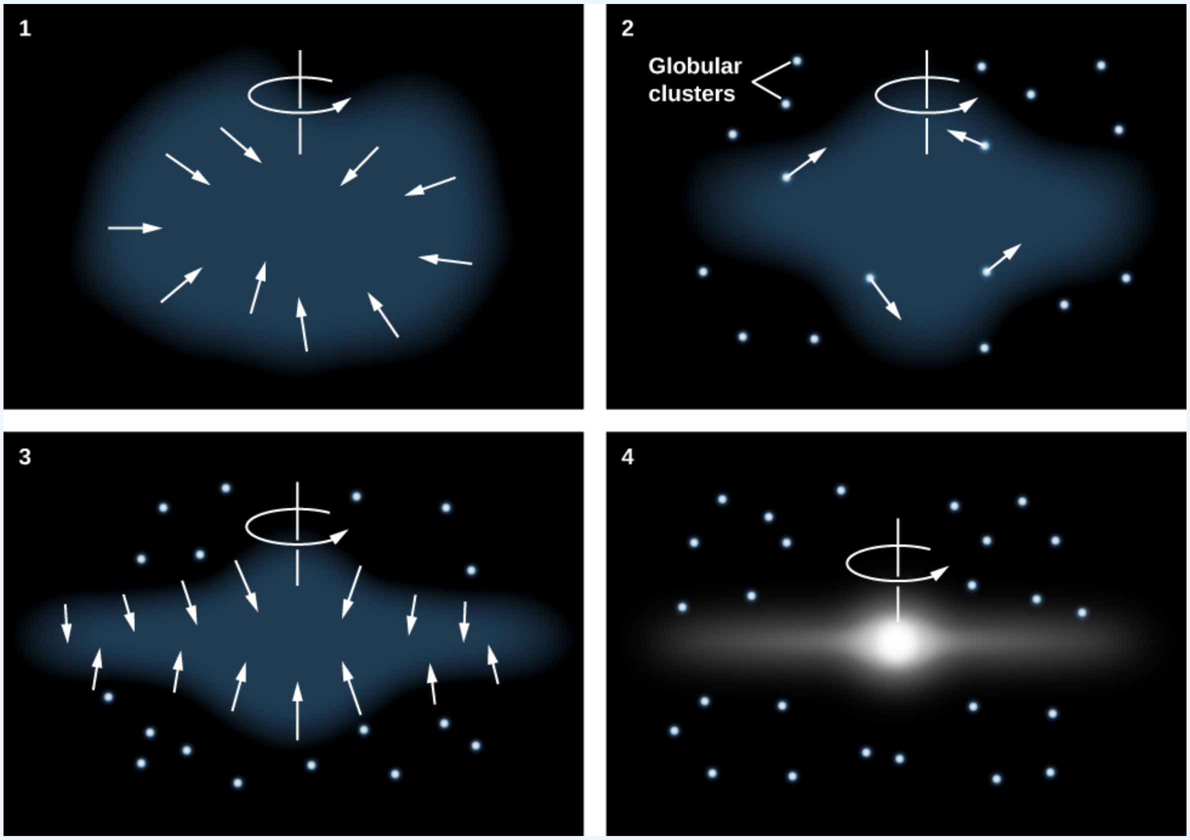


Figure 12.16. According to this model, the Milky Way Galaxy initially formed from a rotating cloud of gas that collapsed due to gravity. Halo stars and globular clusters either formed prior to the collapse or were formed elsewhere. Stars in the disk formed later, when the gas from which they were made was already "contaminated" with heavy elements produced in earlier generations of stars.

Gravitational forces caused the gas in the thin disk to fragment into clouds or clumps with masses like those of star clusters. These individual clouds then fragmented further to form stars. Since the oldest stars in the disk are nearly as old as the youngest stars in the halo, the collapse must have been rapid (astronomically speaking), requiring perhaps no more than a few hundred million years.

Collision Victims and the Multiple Merger Model

In past decades, astronomers have learned that the evolution of the Galaxy has not been quite as peaceful as this **monolithic collapse model** suggests. In 1994, astronomers discovered a small new galaxy in the direction of the constellation of Sagittarius. The Sagittarius dwarf galaxy is currently about 70,000 light-years away from Earth and 50,000 light-years from the centre of the Galaxy. It is the closest galaxy known. Pictured in Figure 12.17. It is very elongated, and its shape indicates that it is being torn apart by our Galaxy's gravitational tides—just as Comet Shoemaker-Levy 9 was torn apart when it passed too close to Jupiter in 1992.

The Sagittarius galaxy is much smaller than the Milky Way, with only about 150,000 stars, all of which seem destined to end up in the bulge and halo of our own Galaxy. But don't sound the funeral bells for the little galaxy quite yet; the ingestion of the Sagittarius dwarf will take another 100 million years or so, and the stars themselves will survive.

Sagittarius Dwarf

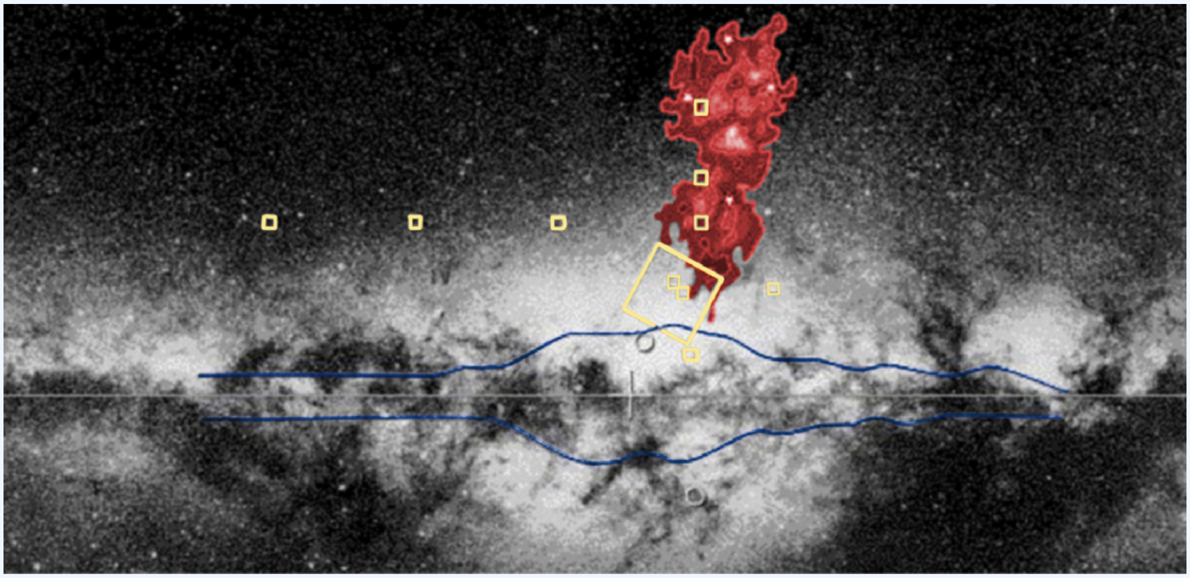


Figure 12.17. In 1994, British astronomers discovered a galaxy in the constellation of Sagittarius, located only about 50,000 light-years from the centre of the Milky Way and falling into our Galaxy. This image covers a region approximately 70° × 50° and combines a black-and-white view of the disk of our Galaxy with a red contour map showing the brightness of the dwarf galaxy. The dwarf galaxy lies on the other side of the galactic centre from us. The white stars in the red region mark the locations of several globular clusters contained within the Sagittarius dwarf galaxy. The cross marks the galactic centre. The horizontal line corresponds to the galactic plane. The blue outline on either side of the galactic plane corresponds to the infrared image in [link]. The boxes mark regions where detailed studies of individual stars led to the discovery of this galaxy.
Modification of Sagittarius Dwarf to Collide with Milky Way by R. Ibata (UBC), R. Wyse (JHU), R. Sword (IoA), NASA Media License.

Since that discovery, evidence has been found for many more close encounters between our Galaxy and other neighbour galaxies. When a small galaxy ventures too close, the force of gravity exerted by our Galaxy tugs harder on the near side than on the far side. The net effect is that the stars that originally belonged to the small galaxy are spread out into a long stream that orbits through the halo of the Milky Way as seen in Figure 12.18.

Streams in the Galactic Halo

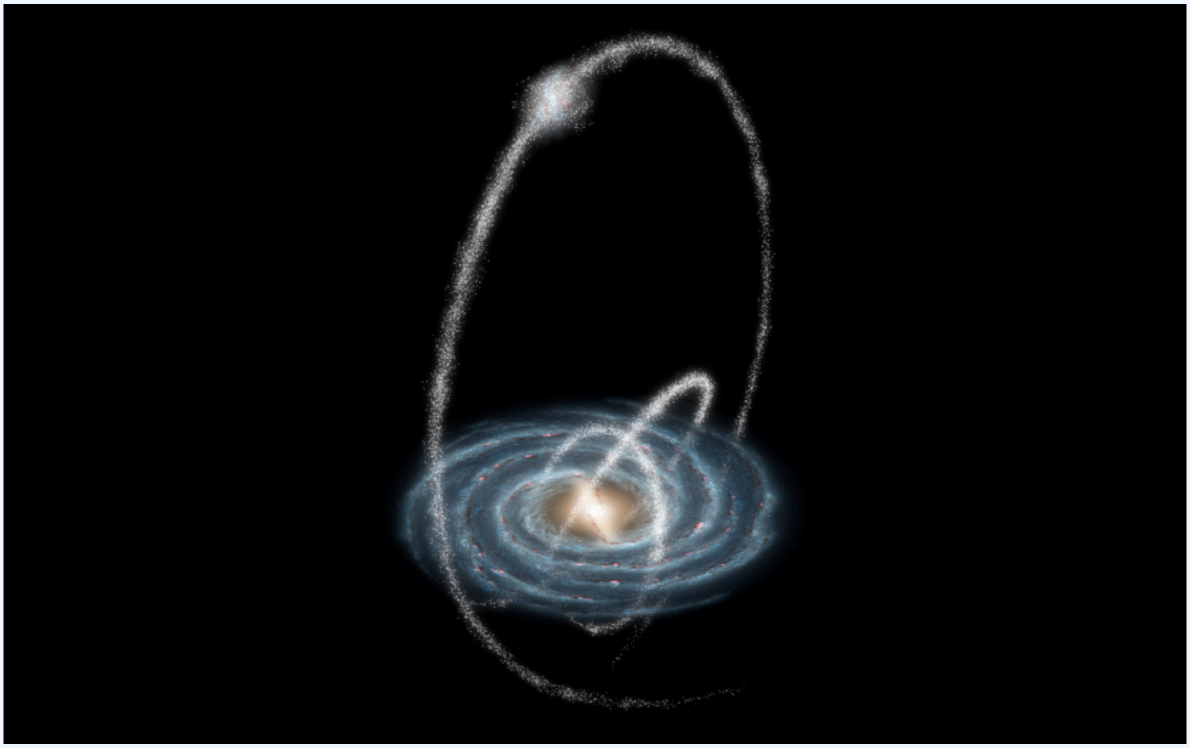


Figure 12.18. When a small galaxy is swallowed by the Milky Way, its member stars are stripped away and form streams of stars in the galactic halo. This image is based on calculations of what some of these tidal streams might look like if the Milky Way swallowed 50 dwarf galaxies over the past 10 billion years. Artist´s impression of the star and dust tail from the torn-to-pieces Sagittarius dwarf galaxy, currently being engulfed by the Milky Way by NASA/JPL-Caltech/R. Hurt (SSC/Caltech), Public Domain

Such a tidal stream can maintain its identity for billions of years. To date, astronomers have now identified streams originating from 12 small galaxies that ventured too close to the much larger Milky Way. Six more streams are associated with globular clusters. It has been suggested that large globular clusters, like Omega Centauri, are actually dense nuclei of cannibalized dwarf galaxies. The globular cluster M54 is now thought to be the nucleus of the Sagittarius dwarf we discussed earlier, which is currently merging with the Milky Way as shown in Figure 12.19. The stars in the outer regions of such galaxies are stripped off by the gravitational pull of the Milky Way, but the central dense regions may survive.

Globular Cluster M54

Figure 12.19. This beautiful Hubble Space Telescope image shows the globular cluster that is now believed to be the nucleus of the Sagittarius Dwarf Galaxy.
Messier 54 by ESA/Hubble & NASA, NASA Media License.

Calculations indicate that the Galaxy's thick disk may be a product of one or more such collisions with other galaxies. Accretion of a satellite galaxy would stir up the orbits of the stars and gas clouds originally in the thin disk and cause them to move higher above and below the mid-plane of the Galaxy. Meanwhile, the Galaxy's stars would add to the fluffed-up mix. If such a collision happened about 10 billion years ago, then any gas in the two galaxies that had not yet formed into stars would have had plenty of time to settle back down into the thin disk. The gas could then have begun forming subsequent generations of population I stars. This timing is also consistent with the typical ages of stars in the thick disk.

The Milky Way has more collisions in store. An example is the Canis Major dwarf galaxy, which has a mass of about 1% of the mass of the Milky Way. Already long tidal tails have been stripped from this galaxy, which have wrapped themselves around the Milky Way three times. Several of the globular clusters found in the Milky Way may also have come from the Canis Major dwarf, which is expected to merge gradually with the Milky Way over about the next billion years.

In about 3 billion years, the Milky Way itself will be swallowed up, since it and the Andromeda galaxy are on a collision course. Our computer models show that after a complex interaction, the two will merge to form a larger, more rounded galaxy, shown in Figure 12.20.

Collision of the Milky Way with Andromeda

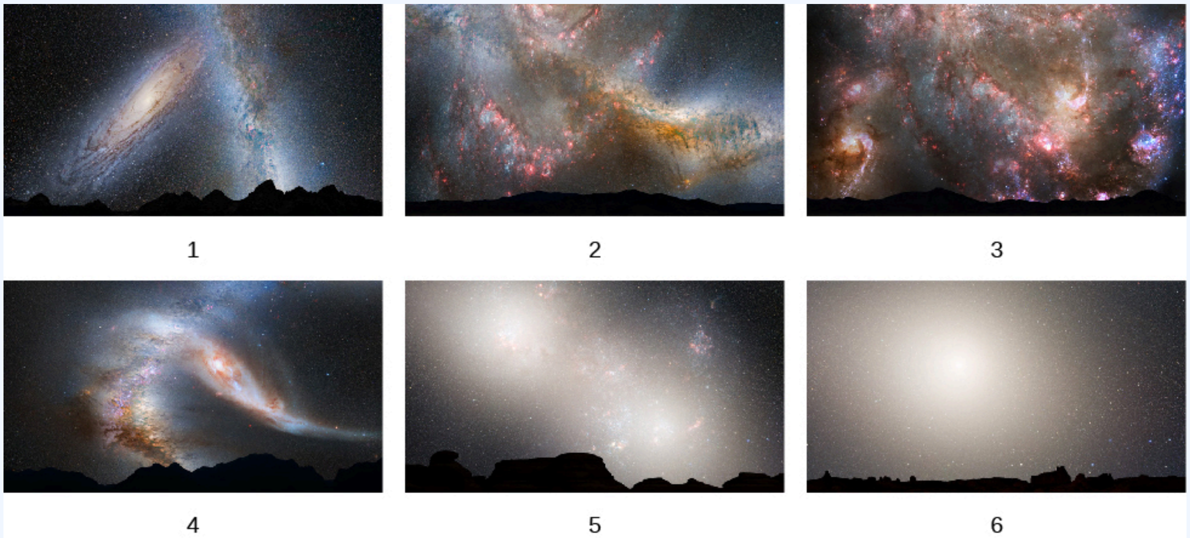


Figure 12.20. In about 3 billion years, the Milky Way Galaxy and Andromeda Galaxy will begin a long process of colliding, separating, and then coming back together to form an elliptical galaxy. The whole interaction will take 3 to 4 billion years. These images show the following sequence: (1) In 3.75 billion years, Andromeda has approached the Milky Way. (2) New star formation fills the sky 3.85 billion years from now. (3) Star formation continues at 3.9 billion years. (4) The galaxy shapes change as they interact, with Andromeda being stretched and our Galaxy becoming warped, about 4 billion years from now. (5) In 5.1 billion years, the cores of the two galaxies are bright lobes. (6) In 7 billion years, the merged galaxies form a huge elliptical galaxy whose brightness fills the night sky. This artist's illustrations show events from a vantage point 25,000 light-years from the centre of the Milky Way. However, we should mention that the Sun may not be at that distance throughout the sequence of events, as the collision readjusts the orbits of many stars within each galaxy.
Illustration Sequence of the Milky Way and Andromeda Galaxy Colliding by NASA,ESA, Z. Levay, R. van der Marel, STScI, T. Hallas, and A. Mellinger, NASA Media License.

We are thus coming to realize that "environmental influences" (and not just a galaxy's original characteristics) play an important role in determining the properties and development of our Galaxy. In future chapters we will see that collisions and mergers are a major factor in the evolution of many other galaxies as well.

At the Sun's distance from its centre, the Galaxy does not rotate like a solid wheel or a CD inside your player. Instead, the way individual objects turn around the centre of the Galaxy is more like the solar system. Stars, as well as the clouds of gas and dust, obey Kepler's third law. Objects farther from the centre take longer to complete an orbit around the Galaxy than do those closer to the centre. In other words, stars (and interstellar matter) in larger orbits in the Galaxy trail behind those in smaller ones. This effect is called **differential galactic rotation**.

Differential rotation would appear to explain why so much of the material in the disk of the Milky Way is concentrated into elongated features that resemble **spiral arms**. No matter what the original distribution of the material might be, the differential rotation of the Galaxy can stretch it out into spiral features. Figure 12.21 shows the development of spiral arms from two irregular blobs of interstellar matter. Notice that as the portions of the blobs closest to the galactic centre move faster, those farther out trail behind.

Simplified Model for the Formation of Spiral Arms

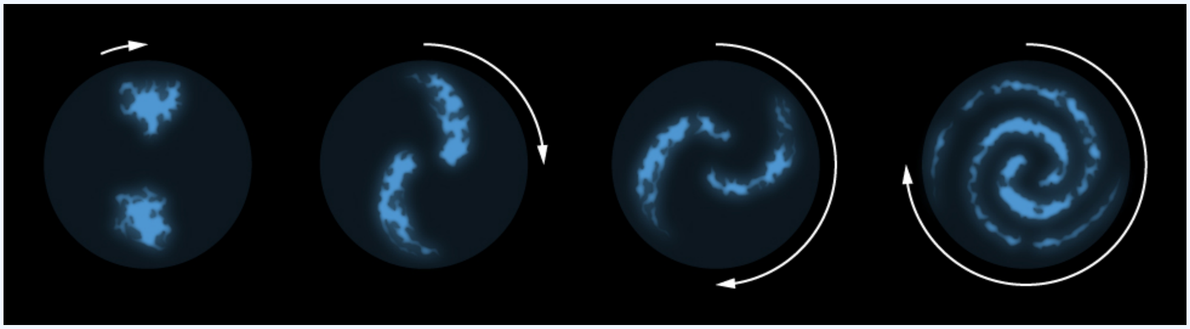


Figure 12.21. This sketch shows how spiral arms might form from irregular clouds of interstellar material stretched out by the different rotation rates throughout the Galaxy. The regions farthest from the galactic centre take longer to complete their orbits and thus lag behind the inner regions. If this were the only mechanism for creating spiral arms, then over time the spiral arms would completely wind up and disappear. Since many galaxies have spiral arms, they must be long-lived, and there must be other processes at work to maintain them.

But this picture of spiral arms presents astronomers with an immediate problem. If that's all there were to the story, differential rotation—over the roughly 13-billion-year history of the Galaxy—would have wound the Galaxy's arms tighter and tighter until all semblance of spiral structure had disappeared. But did the Milky Way actually have spiral arms when it formed 13 billion years ago? And do spiral arms, once formed, last for that long a time?

With the advent of the Hubble Space Telescope, it has become possible to observe the structure of very distant galaxies and to see what they were like shortly after they began to form more than 13 billion years ago. What the observations show is that galaxies in their infancy had bright, clumpy star-forming regions, but no regular spiral structure.

Over the next few billion years, the galaxies began to "settle down." The galaxies that were to become spirals lost their massive clumps and developed a central bulge. The turbulence in these galaxies decreased, rotation began to dominate the motions of the stars and gas, and stars began to form in a much quieter disk. Smaller star-forming clumps began to form fuzzy, not-very-distinct spiral arms. Bright, well-defined spiral arms began to appear only when the galaxies were about 3.6 billion years old. Initially, there were two well-defined arms. Multi-armed structures in galaxies like we see in the Milky Way appeared only when the universe was about 8 billion years old.

Scientists have used supercomputer calculations to model the formation and evolution of the arms. These calculations follow the motions of up to 100 million "star particles" to see whether gravitational forces can cause them to form spiral structure. What these calculations show is that giant molecular clouds have enough gravitational influence over their surroundings to initiate the formation of structures that look like spiral arms. These arms then become self-perpetuating and can survive for at least several billion years. The arms may change their brightness over time as star formation comes and goes, but they are not temporary features. The concentration of matter in the arms exerts sufficient gravitational force to keep the arms together over long periods of time.

Attribution

12.5 KEY TERMS

Barred spirals: spiral galaxies that also have bar-shaped concentrations of stars in their central regions. 12.1

Central bulge (nuclear bulge): the central (round) part of the Milky Way or a similar galaxy. 12.1

Dark matter halo: the mass in the Milky Way that extends well beyond the boundary of the luminous stars to a distance of at least 200,000 light-years from the centre of the Galaxy; although we deduce its existence from its gravity, the composition of this matter remains a mystery. 12.1

Differential galactic rotation: the idea that different parts of the Galaxy turn at different rates, since the parts of the Galaxy follow Kepler's third law: more distant objects take longer to complete one full orbit around the centre of the Galaxy. 12.4

Elliptical galaxies: galaxies whose shape is an ellipse and that contain no conspicuous interstellar material. 12.3

Halo: the outermost extent of our Galaxy (or another galaxy), containing a sparse distribution of stars and globular clusters in a more or less spherical distribution. 12.1

Irregular galaxies: galaxies that do not have the regular shapes associated with either spiral or elliptical galaxies. 12.3

Local Group: a small cluster of galaxies to which the Milky Way Galaxy belongs. 12.2

Milky Way Galaxy: the band of light encircling the sky, which is due to the many stars and diffuse nebulae lying near the plane of the Milky Way Galaxy. 12.1

Monolithic Collapse Model: according to this model, the Milky Way Galaxy initially formed from a rotating cloud of gas that collapsed due to gravity. 12.4

Protogalactic cloud: a rotating cloud of gas like the one that gave birth to the Milky Way Galaxy. 12.4

Spiral arm: a spiral-shaped region, characterized by relatively dense interstellar material and young stars, that is observed in the disks of spiral galaxies. 12.4

CHAPTER 13: GALAXIES: COLLISIONS, MERGERS, AND EVOLUTION

Chapter Overview

13.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify the major techniques used by astronomers to study galaxy evolution and cosmic time, including the use of telescopes, spectra analysis, and look-back time calculations.
- Explain the concept of a "cosmic time machine" and how it enables astronomers to observe galaxies as they were billions of years ago, allowing the reconstruction of galactic evolution over time.
- Apply the knowledge of galaxy collisions and mergers to understand the formation of elliptical and spiral galaxies, and how collisions can trigger starbursts and intense star formation.
- Analyze the differences between galaxy collisions and star collisions, explaining why galaxies can collide while stars within a galaxy rarely do.
- Explain the evidence for the presence of a supermassive black hole at the center of our Galaxy, including the observations of stars' orbits and the compact radio source Sagittarius A*.
- Explain the difference between quasars and stars in terms of their energy sources, and how the presence of supermassive black holes can account for the observed properties of quasars and AGNs.

13.1 INTRODUCTION

Colliding Galaxies

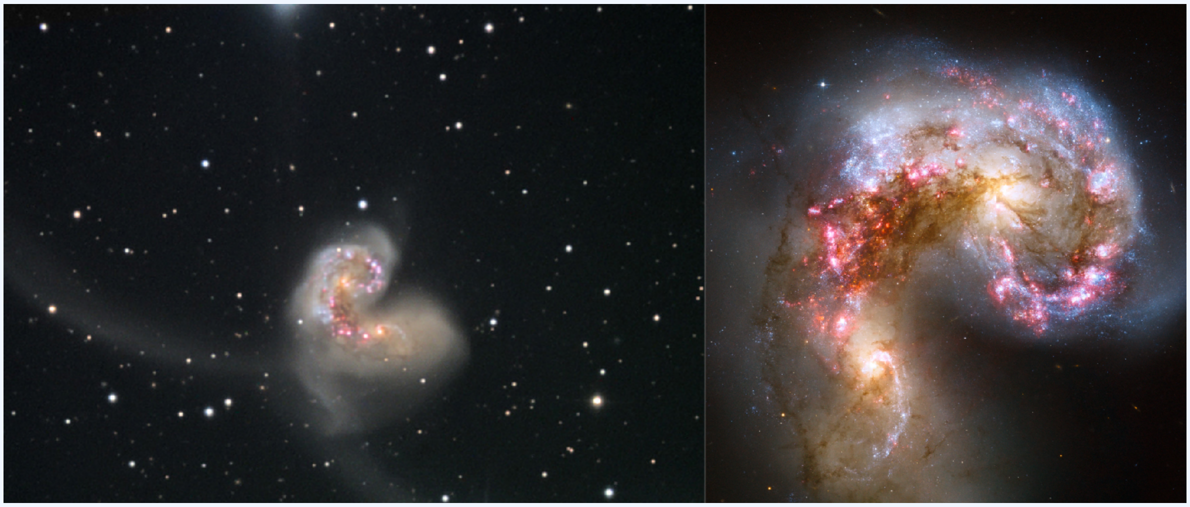


Figure 13.1. Collisions and mergers of galaxies strongly influence their evolution. On the left is a ground-based image of two colliding galaxies (NCG 4038 and 4039), sometimes nicknamed the Antennae galaxies. The long, luminous tails are material torn out of the galaxies by tidal forces during the collision. The right image shows the inner regions of these two galaxies, as taken by the Hubble Space Telescope. The cores of the twin galaxies are the orange blobs to the lower left and upper right of the center of the image. Note the dark lanes of dust crossing in front of the bright regions. The bright pink and blue star clusters are the result of a burst of star formation stimulated by the collision.
Credit left: NGC 4038 by Bob and Bill Twardy/Adam Block/NOAO/AURA/NSF, CC BY 4.0
Credit right: The Antennae Galaxies/NGC 4038-4039 by NASA, ESA, and the Hubble Heritage Team (STScI/AURA)-ESA/Hubble Collaboration, NASA Media License.

How and when did galaxies like our Milky Way form? Which formed first: stars or galaxies? Can we see direct evidence of the changes galaxies undergo over their lifetimes? If so, what determines whether a galaxy will “grow up” to be spiral or elliptical? And what is the role of “nature versus nurture”? That is to say, how much of a galaxy’s development is determined by what it looks like when it is born and how much is influenced by its environment?

Astronomers today have the tools needed to explore the universe almost back to the time it began. The huge new telescopes and sensitive detectors built in the last decades make it possible to obtain both images and spectra of galaxies so distant that their light has traveled to reach us for more than 13 billion years—more than 90% of the way back to the Big Bang: we can use the finite speed of light and the vast size of the universe as a cosmic time machine to peer back and observe how galaxies formed and evolved over time. Studying galaxies so far away in any detail is always a major challenge, largely because their distance makes them appear very faint. However, today's large telescopes on the ground and in space are finally making such a task possible.

Attribution

13.2 THE COSMIC TIME MACHINE

Let's begin by exploring some techniques astronomers use to study how galaxies are born and change over cosmic time. Suppose you wanted to understand how adult humans got to be the way they are. If you were very dedicated and patient, you could actually observe a sample of babies from birth, following them through childhood, adolescence, and into adulthood, and making basic measurements such as their heights, weights, and the proportional sizes of different parts of their bodies to understand how they change over time.

Unfortunately, we have no such possibility for understanding how galaxies grow and change over time: in a human lifetime—or even over the entire history of human civilization—individual galaxies change hardly at all. We need other tools than just patiently observing single galaxies in order to study and understand those long, slow changes.

We do, however, have one remarkable asset in studying galactic evolution. As we have seen, the universe itself is a kind of time machine that permits us to observe remote galaxies as they were long ago. For the closest galaxies, like the Andromeda galaxy, the time the light takes to reach us is on the order of a few hundred thousand to a few million years. Typically not much changes over times that short—individual stars in the galaxy may be born or die, but the overall structure and appearance of the galaxy will remain the same. But we have observed galaxies so far away that we are seeing them as they were when the light left them more than 10 billion years ago.

By observing more distant objects, we look further back toward a time when both galaxies and the universe were young, shown in Figure 13.2. This is a bit like getting letters in the mail from several distant friends: the farther the friend was when she mailed the letter to you, the longer the letter must have been in transit, and so the older the news is when it arrives in your mailbox; you are learning something about her life at an earlier time than when you read the letter.

Astronomical Time Travel

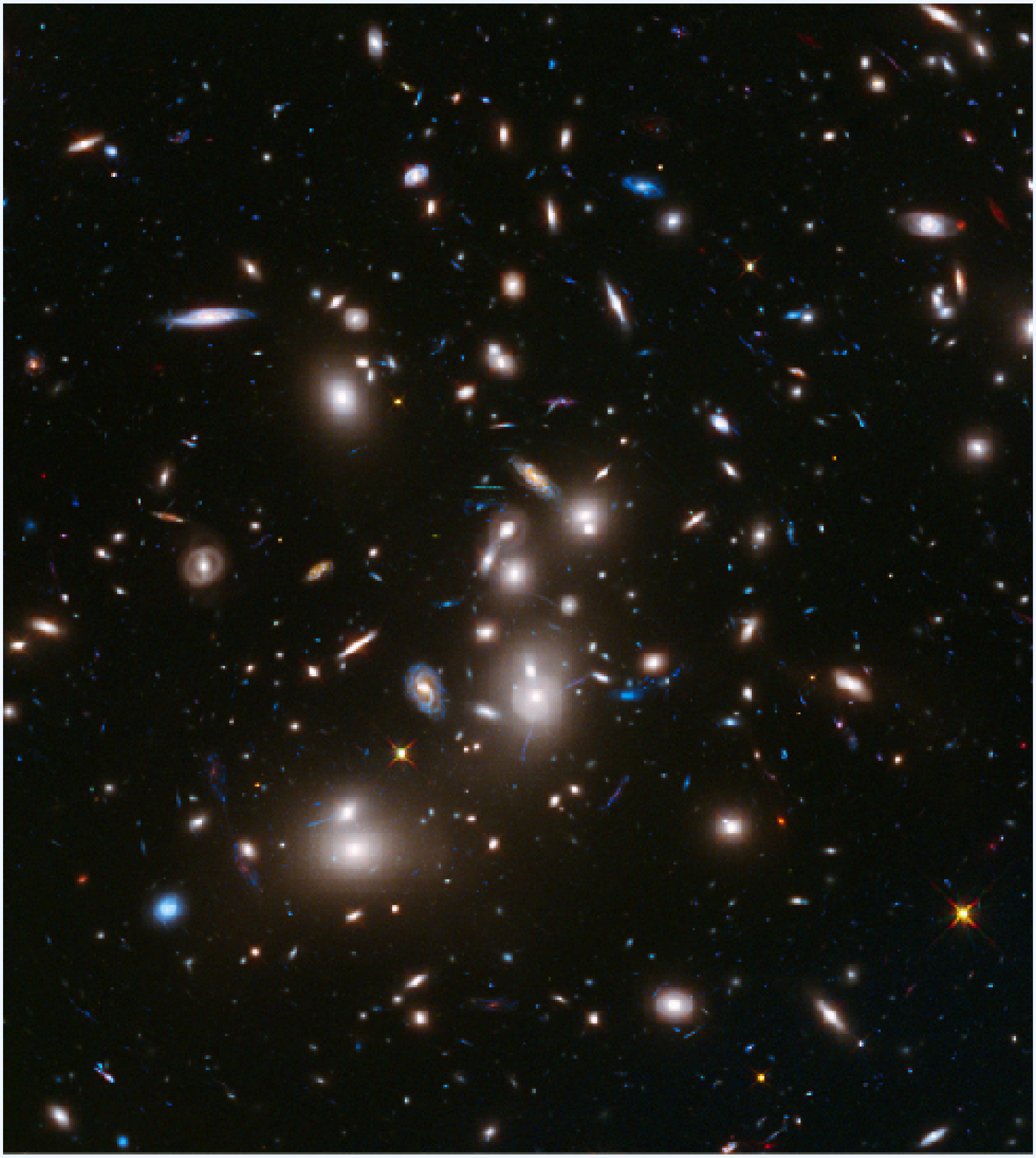


Figure 13.2. This true-colour, long-exposure image, made during 70 orbits of Earth with the Hubble Space Telescope, shows a small area in the direction of the constellation Sculptor. The massive cluster of galaxies named Abell 2744 appears in the foreground of this image. It contains several hundred galaxies, and we are seeing them as they looked 3.5 billion years ago. The immense gravity in Abell 2744 acts as a gravitational lens (see the Astronomy Basics feature box on Gravitational Lensing later in this chapter) to warp space and brighten and magnify images of nearly 3000 distant background galaxies. The more distant galaxies (many of them quite blue) appear as they did more than 12 billion years ago, not long after

the Big Bang. Blue galaxies were much more common in that earlier time than they are today. These galaxies appear blue because they are undergoing active star formation and making hot, bright blue stars. Abell 2744 by NASA, ESA, STScI, NASA Media License.

If we can't directly detect the changes over time in individual galaxies because they happen too slowly, how then can we ever understand those changes and the origins of galaxies? The solution is to observe many galaxies at many different cosmic distances and, therefore, look-back times (how far back in time we are seeing the galaxy). If we can study a thousand very distant "baby" galaxies when the universe was 1 billion years old, and another thousand slightly closer "toddler" galaxies when it was 2 billion years old, and so on until the present 13.8-billion-year-old universe of mature "adult" galaxies near us today, then maybe we can piece together a coherent picture of how the whole ensemble of galaxies evolves over time. This allows us to reconstruct the "life story" of galaxies since the universe began, even though we can't follow a single galaxy from infancy to old age.

Fortunately, there is no shortage of galaxies to study. Hold up your pinky at arm's length: the part of the sky blocked by your fingernail contains about one million galaxies, layered farther and farther back in space and time. In fact, the sky is filled with galaxies, all of them, except for Andromeda and the Magellanic Clouds, too faint to see with the naked eye—more than 100 billion galaxies in the observable universe, each one with about 100 billion stars.

This cosmic time machine, then, lets us peer into the past to answer fundamental questions about where galaxies come from and how they got to be the way they are today. Astronomers call those galactic changes over cosmic time evolution, a word that recalls the work of Darwin and others on the development of life on Earth. But note that **galaxy evolution** refers to the changes in *individual* galaxies over time, while the kind of evolution biologists study is changes in *successive generations* of living organisms over time.

Astronomy is one of the few sciences in which all measurements must be made at a distance. Geologists can take samples of the objects they are studying; chemists can conduct experiments in their laboratories to determine what a substance is made of; archeologists can use carbon dating to determine how old something is. But astronomers can't pick up and play with a star or galaxy. As we have seen throughout this book, if they want to know what galaxies are made of and how they have changed over the lifetime of the universe, they must decode the messages carried by the small number of photons that reach Earth.

Fortunately (as you have learned) electromagnetic radiation is a rich source of information. The distance to a galaxy is derived from its **redshift** (how much the lines in its spectrum are shifted to the red because of the expansion of the universe). The conversion of redshift to a distance depends on certain properties of the universe, including the value of the Hubble constant and how much mass it contains. We will describe the currently accepted model of the universe in a later chapter. For the purposes of this chapter, it is enough to know that the current best estimate for the age of the universe is 13.8 billion years. In that case, if we see an object that emitted its light 6 billion light-years ago, we are seeing it as it was when the universe was almost 8 billion years old. If we see something that emitted its light 13 billion years ago, we are seeing it as it was when the universe was less than a billion years old. So astronomers measure a galaxy's redshift from its spectrum, use the Hubble constant plus a model of the universe to turn the redshift into a distance, and use the distance and the constant speed of light to infer how far back in time they are seeing the galaxy—the look-back time.

Another important clue to the nature of a galaxy is its shape. Spiral galaxies can be distinguished from elliptical galaxies by shape. Observations show that spiral galaxies contain young stars and large amounts of interstellar matter, while elliptical galaxies have mostly old stars and very little or no star formation. Elliptical galaxies turned most of their interstellar matter into stars many billions of years ago, while star formation has continued until the present day in spiral galaxies.

If we can count the number of galaxies of each type during each epoch of the universe, it will help us understand how the pace of star formation changes with time. As we will see later in this chapter, galaxies in the distant universe—that is, young galaxies—look very different from the older galaxies that we see nearby in the present-day universe.

In addition to looking at the most distant galaxies we can find, astronomers look at the oldest stars (what we might call the fossil record) of our own Galaxy to probe what happened in the early universe. Since stars are the source of nearly all the light emitted by galaxies, we can learn a lot about the evolution of galaxies by studying the stars within them. What we find is that nearly all galaxies contain at least some very old stars. For example, our own Galaxy contains globular clusters with stars that are at least 13 billion years old, and some may be even older than that. Therefore, if we count the age of the Milky Way as the age of its oldest constituents, the Milky Way must have been born at least 13 billion years ago.

Several other observations also establish that star formation in the cosmos began very early. Astronomers have used spectra to determine the composition of some elliptical galaxies that are so far away that the light we see left them when the universe was only half as old as it is now. Yet these ellipticals contain old red stars, which must have formed billions of years earlier still.

When we make computer models of how such galaxies evolve with time, they tell us that star formation in elliptical galaxies began less than a billion years or so after the universe started its expansion, and new stars continued to form for a few billion years. But then star formation apparently stopped. When we compare distant elliptical galaxies with ones nearby, we find that ellipticals have not changed very much since the universe reached about half its current age. We'll return to this idea later in the chapter.

Observations of the most luminous galaxies take us even further back in time. Recently, as we have already noted, astronomers have discovered a few galaxies that are so far away that the light we see now left them less than a billion years or so after the beginning as seen in Figure 13.3. Yet the spectra of some of these galaxies already contain lines of heavy elements, including carbon, silicon, aluminum, and sulphur. These elements were not present when the universe began but had to be manufactured in the interiors of stars. This means that when the light from these galaxies was emitted, an entire generation of stars had already been born, lived out their lives, and died—spewing out the new elements made in their interiors through supernova explosions—even before the universe was a billion years old. And it wasn't just a few stars in each galaxy that got started this way. Enough had to live and die to affect the overall composition of the galaxy, in a way that we can still measure in the spectrum from far away.

In August of 2018, United Kingdon astronomers published results showing that there were galaxies that were formed only a few hundred million years after the Big Bang. https://www.bbc.com/news/science-environment-45198764

Very Distant Galaxy

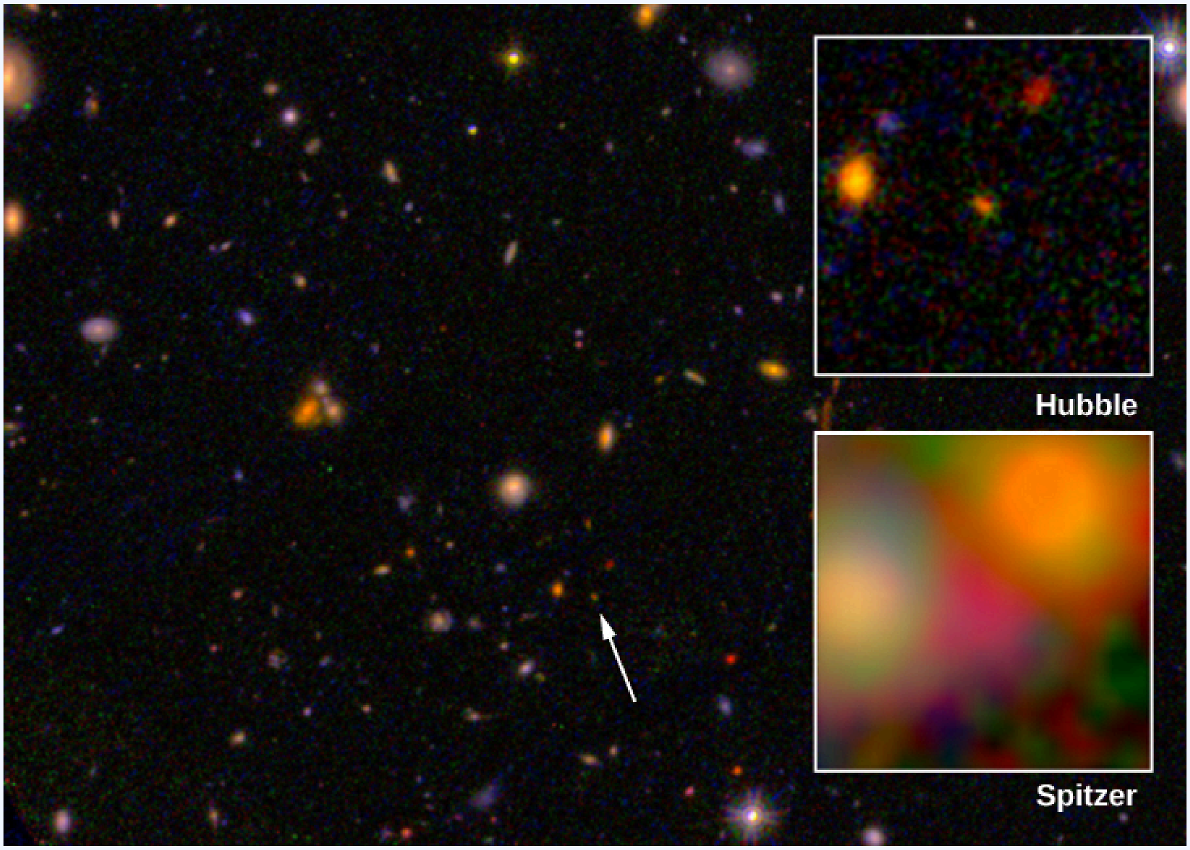


Figure 13.3. This image was made with the Hubble Space Telescope and shows the field around a luminous galaxy at a redshift z = 8.68, corresponding to a distance of about 13.2 billion light-years at the time when the light was emitted (indicated by the arrow and shown in the upper inset). Long exposures in the far-red and infrared wavelengths were combined to make the image, and additional infrared exposures with the Spitzer Space Telescope, which has lower spatial resolution than the Hubble (lower inset), show the redshifted light of normal stars. The very distant galaxy was detected because it has a strong emission line of hydrogen. This line is produced in regions where the formation of hot, young stars is taking place. EGSY8p7 by I. Labbé (Leiden University), NASA/ESA/JPL-Caltech, Public Domain

Observations of **quasars** (galaxies whose centres contain a supermassive black hole) support this conclusion. We can measure the abundances of heavy elements in the gas near quasar black holes. The composition of this gas in quasars that emitted their light 12.5 billion light-years ago is very similar to that of the Sun. This means that a large portion of the gas surrounding the black holes must have already been cycled through stars during the first 1.3 billion years after the expansion of the universe began. If we allow time for this cycling, then their first stars must have formed when the universe was only a few hundred million years old.

Attribution

13.3 COLLISIONS AND MERGERS

One of the conclusions astronomers have reached from studying distant galaxies is that collisions and mergers of whole galaxies play a crucial role in determining how galaxies acquired the shapes and sizes we see today. Only a few of the nearby galaxies are currently involved in collisions, but detailed studies of those tell us what to look for when we seek evidence of mergers in very distant and very faint galaxies. These in turn give us important clues about the different evolutionary paths galaxies have taken over cosmic time. Let's examine in more detail what happens when two galaxies collide.

The images below show a dynamic views of two galaxies that are colliding. The stars themselves in this pair of galaxies will not be affected much by this cataclysmic event. (See the Astronomy Basics feature box "Why Galaxies Collide but Stars Rarely Do" below the image) Since there is a lot of space between the stars, a direct collision between two stars is very unlikely. However, the *orbits* of many of the stars will be changed as the two galaxies move through each other, and the change in orbits can totally alter the appearance of the interacting galaxies. A gallery of interesting colliding galaxies is shown in Figure 13.4. Great rings, huge tendrils of stars and gas, and other complex structures can form in such cosmic collisions. Indeed, these strange shapes are the signposts that astronomers use to identify colliding galaxies.

Gallery of Interacting Galaxies

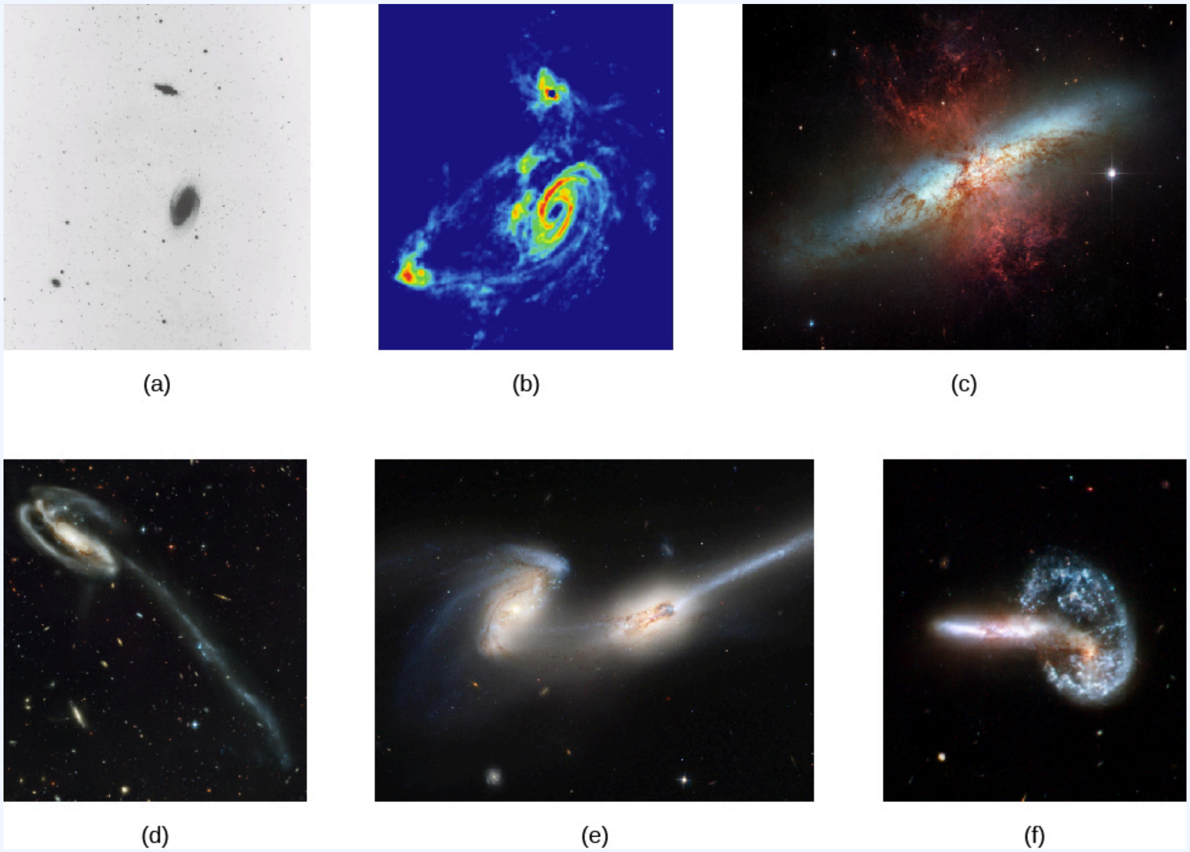


Figure 13.4. (a and b) M82 (smaller galaxy at top) and M83 (spiral) are seen (a) in a black-and-white visible light image and (b) in radio waves given off by cold hydrogen gas. The hydrogen image shows that the two galaxies are wrapped in a common shroud of gas that is being tugged and stretched by the gravity of the two galaxies. (c) This close-up view by the Hubble Space Telescope shows some of the effects of this interaction on galaxy M82, including gas streaming outward (red tendrils) powered by supernovae explosions of massive stars formed in the burst of star formation that was a result of the collision. (d) Galaxy UGC 10214 ("The Tadpole") is a barred spiral galaxy 420 million light-years from the Milky Way that has been disrupted by the passage of a smaller galaxy. The interloper's gravity pulled out the long tidal tail, which is about 280,000 light-years long, and triggered bursts of star formation seen as blue clumps along the tail. (e) Galaxies NGC 4676 A and B are nicknamed "The Mice." In this Hubble Space Telescope image, you can see the long, narrow tails of stars pulled away from the galaxies by the interactions of the two spirals. (f) Arp 148 is a pair of galaxies that are caught in the act of merging to become one new galaxy. The two appear to have already passed through each other once, causing a shockwave that reformed one into a bright blue ring of star formation, like the ripples from a stone tossed into a pond.
Credit a, b: Figure 8.11 by NRAO/AUI, CC BY 4.0.
Credit c: Messier 82 (The Cigar Galaxy) by NASA, ESA, and The Hubble Heritage Team (STScI/AURA), NASA Media License.
Credit d: Rich Background of Galaxies Behind the Tadpole Galaxy by NASA, H. Ford (JHU), G. Illingworth (UCSC/LO), M.Clampin (STScI), G. Hartig (STScI), the ACS Science Team, and ESA, ESA Standard License.
Credit e: The Mice (NGC 4676): Colliding Galaxies With Tails of Stars and Gas by NASA, H. Ford (JHU), G. Illingworth (UCSC/LO), M.Clampin (STScI), G. Hartig (STScI), the ACS Science Team, and ESA, NASA Media License.

Credit f: Arp 148 by NASA, ESA, the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration, and A. Evans (University of Virginia, Charlottesville/NRAO/Stony Brook University), NASA Media License.

Why Galaxies Collide but Stars Rarely Do

Throughout this book we have emphasized the large distances between objects in space. You might therefore have been surprised to hear about collisions between galaxies. Yet (except at the very cores of galaxies) we have not worried at all about stars inside a galaxy colliding with each other. Let's see why there is a difference.

The reason is that stars are pitifully small compared to the distances between them. Let's use our Sun as an example. The Sun is about 1.4 million kilometres wide, but is separated from the closest other star by about 4 light-years, or about 38 trillion kilometres. In other words, the Sun is 27 million of its own diametres from its nearest neighbour. If the Sun were a grapefruit in New York City, the nearest star would be another grapefruit in San Francisco. This is typical of stars that are not in the nuclear bulge of a galaxy or inside star clusters. Let's contrast this with the separation of galaxies.

The visible disk of the Milky Way is about 100,000 light-years in diameter. We have three satellite galaxies that are just one or two Milky Way diametres away from us (and will probably someday collide with us). The closest major spiral is the Andromeda Galaxy (M31), about 2.4 million light-years away. If the Milky Way were a pancake at one end of a big breakfast table, M31 would be another pancake at the other end of the same table. Our nearest large galaxy neighbour is only 24 of our Galaxy's diametres from us, and it will begin to crash into the Milky Way in about 3 billion years.

Galaxies in rich clusters are even closer together than those in our neighbourhood. Thus, the chances of galaxies colliding are far greater than the chances of stars in the disk of a galaxy colliding. And we should note that the difference between the separation of galaxies and stars also means that when galaxies do collide, their stars almost always pass right by each other like smoke passing through a screen door.

The details of galaxy collisions are complex, and the process can take hundreds of millions of years. Thus, collisions are best simulated on a computer as shown in Figure 13.5, where astronomers can calculate the slow interactions of stars, and clouds of gas and dust, via gravity. These calculations show that if the collision is slow, the colliding galaxies may coalesce to form a single galaxy.

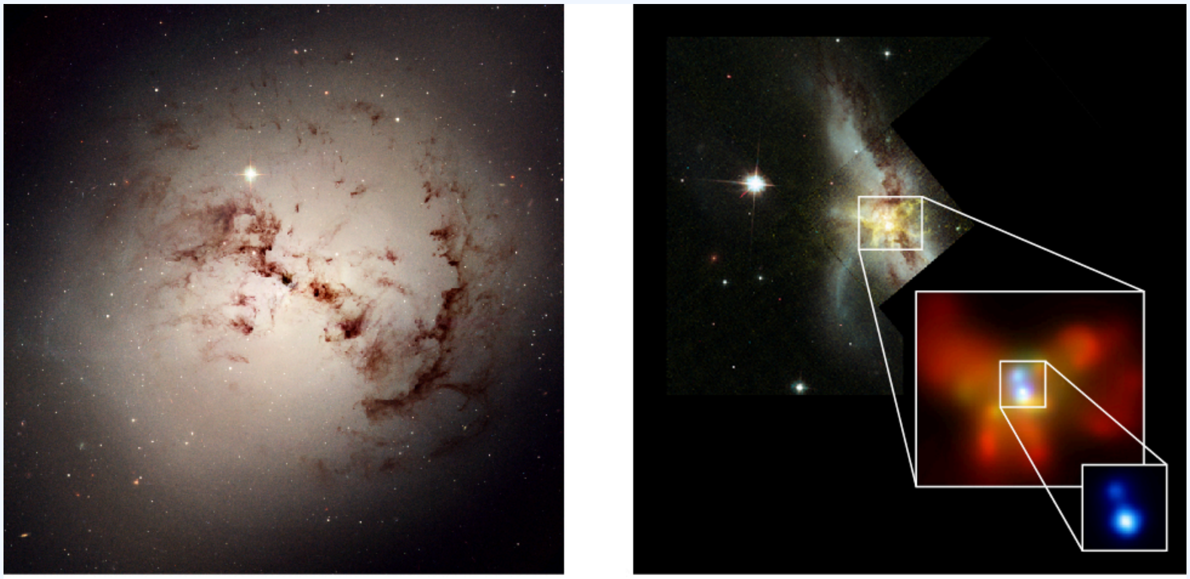
Computer Simulation of a Galaxy Collision

Figure 13.5. This computer simulation starts with two spiral galaxies merging and ends with a single elliptical galaxy. The colours show the colours of stars in the system; note the bursts of blue colour as copious star formation gets triggered by the interaction. The timescale from start to finish in this sequence is about a billion years.
Simulated Images of Merging Galaxies by P. Jonsson (Harvard-Smithsonian Center for Astrophysics), G. Novak (Princeton University), and T. J. Cox (Carnegie Observatories), NASA Media License.

When two galaxies of equal size are involved in a collision, we call such an interaction a **merger** (the term applied in the business world to two equal companies that join forces). But small galaxies can also be swallowed by larger ones—a process astronomers have called, with some relish, **galactic cannibalism**, pictured in Figure 13.6.

Modern personal computers are more than powerful enough to compute what happens when galaxies collide. There are many websites that will let you try your own hand at crashing two spiral galaxies together from the comfort of your own home or dorm room. By changing a few basic controls such as the relative masses, their separation, and the orientation of each galaxy's disk, you can create a wide range of resulting merger results. (You can also download a similar app for your iPhone or iPad at https://apps.apple.com/us/app/galaxy-collider/id301086225)

GALACTIC CANNIBALISM



(a) (b)

Figure 13.6. (a) This Hubble image shows the eerie silhouette of dark dust clouds against the glowing nucleus of the elliptical galaxy NGC 1316. Elliptical galaxies normally contain very little dust. These clouds are probably the remnant of a small companion galaxy that was cannibalized (eaten) by NGC 1316 about 100 million years ago. (b) The highly disturbed galaxy NGC 6240, imaged by Hubble Space Telescope (background image) and Chandra X-ray Telescope (both insets) is apparently the product of a merger between two gas-rich spiral galaxies. The X-ray images show that there is not one but two nuclei, both glowing brightly in X-rays and separated by only 4000 light-years. These are likely the locations of two supermassive black holes that inhabited the cores of the two galaxies pre-merger; here they are participating in a kind of "death spiral," in which the two black holes themselves will merge to become one.

Credit a: The Dusty Galaxy NGC 1316 by NASA, ESA, and The Hubble Heritage Team (STScI/AURA), NASA Media License.
Credit b: The Supermassive Black Holes of NGC 6240 and NGC 6240 by X-ray: NASA/CXC/MPE/S.Komossa et al.; Optical: NASA/STScI/R.P.van der Marel & J.Gerssen), NASA Media License.

The very large elliptical galaxies probably form by cannibalizing a variety of smaller galaxies in their clusters. These "monster" galaxies frequently possess more than one nucleus and have probably acquired their unusually high luminosities by swallowing nearby galaxies. The multiple nuclei are the remnants of their victims ([image above of the cannibal galaxies). Many of the large, peculiar galaxies that we observe also owe their chaotic shapes to past interactions. Slow collisions and mergers can even transform two or more spiral galaxies into a single elliptical galaxy.

A change in shape is not all that happens when galaxies collide. If either galaxy contains interstellar matter, the collision can compress the gas and trigger an increase in the rate at which stars are being formed—by as much as a factor of 100. Astronomers call this abrupt increase in the number of stars being formed a **starburst**, and the galaxies in which the increase occurs are termed starburst galaxies, shown in Figure 13.7. In some interacting galaxies, star formation is so intense that all the available gas is exhausted in only a few million years; the burst of star formation is clearly only a temporary phenomenon. While a starburst is going on, however, the galaxy where it is taking place becomes much brighter and much easier to detect at large distances.

Starburst Associated with Colliding Galaxies

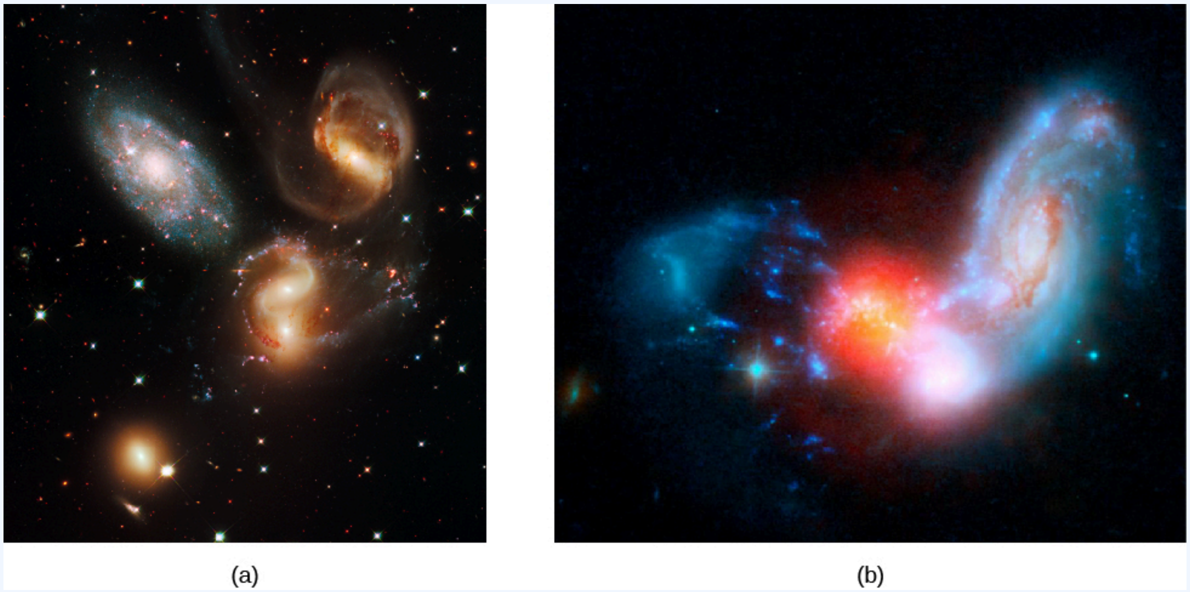


Figure 13.7. (a) Three of the galaxies in the small group known as Stephan's Quintet are interacting gravitationally with each other (the galaxy at upper left is actually much closer than the other three and is not part of this interaction), resulting in the distorted shapes seen here. Long strings of young, massive blue stars and hundreds of star formation regions glowing in the pink light of excited hydrogen gas are also results of the interaction. The ages of the star clusters range from 2 million to 1 billion years old, suggesting that there have been several different collisions within this group of galaxies, each leading to bursts of star formation. The three interacting members of Stephan's Quintet are located at a distance of 270 million light-years. (b) Most galaxies form new stars at a fairly slow rate, but members of a rare class known as starburst galaxies blaze with extremely active star formation. The galaxy II Zw 096 is one such starburst galaxy, and this combined image using both Hubble and Spitzer Space Telescope data shows that it is forming bright clusters of new stars at a prodigious rate. The blue colours show the merging galaxies in visible light, while the red colours show infrared radiation from the dusty region where star formation is happening. This galaxy is at a distance of 500 million light-years and has a diameter of about 50,000 light-years, about half the size of the Milky Way.
Credit a: Stephan's Quintet by NASA, ESA, and the Hubble SM4 ERO Team, NASA Media License.
Credit b: Image by NASA/JPL-Caltech/STScI, NASA Media License.

When astronomers finally had the tools to examine a significant number of galaxies that emitted their light 11 to 12 billion years ago, they found that these very young galaxies often resemble nearby starburst galaxies that are involved in mergers: they also have multiple nuclei and peculiar shapes, they are usually clumpier than normal galaxies today, with multiple intense knots and lumps of bright starlight, and they have higher rates of star formation than isolated galaxies. They also contain lots of blue, young, type O and B stars, as do nearby merging galaxies.

Galaxy mergers in today's universe are rare. Only about five percent of nearby galaxies are currently involved in interactions. Interactions were much more common billions of years ago, some shown in Figure 13.8, and helped build up the "more mature" galaxies we see in our time. Clearly, interactions of galaxies have played a crucial role in their evolution.

Collisions of Galaxies in a Distant Cluster

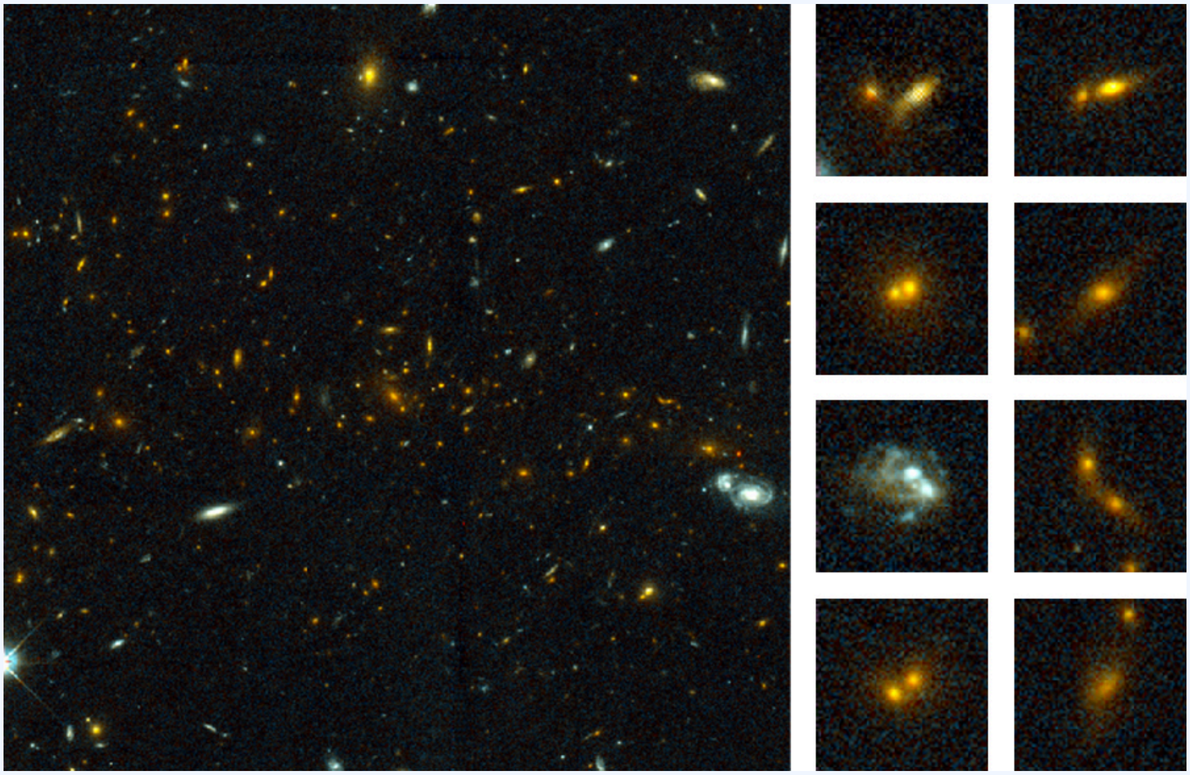


Figure 13.8. The large picture on the left shows the Hubble Space Telescope image of a cluster of galaxies at a distance of about 8 billion light-years. Among the 81 galaxies in the cluster that have been examined in some detail, 13 are the result of recent collisions of pairs of galaxies. The eight smaller images on the right are close-ups of some of the colliding galaxies. The merger process typically takes a billion years or so.
MS 1054-03 & merging galaxies by Pieter van Dokkum, Marijn Franx (University of Groningen/Leiden), ESA and NASA, NASA Media License.

Attribution

13.4 THE GALACTIC CENTER

Let’s take a voyage to the mysterious heart of our Galaxy and see what’s there. Figure 13.9 is a radio image of a region about 1500 light-years across, centred on Sagittarius A, a bright radio source that contains the smaller Sagittarius A*. Much of the radio emission comes from hot gas heated either by clusters of hot stars (the stars themselves do not produce radio emission and can’t be seen in the image) or by supernova blast waves. Most of the hollow circles visible on the radio image are supernova remnants. The other main source of radio emission is from electrons moving at high speed in regions with strong magnetic fields. The bright thin arcs and “threads” on the figure show us where this type of emission is produced.

Radio Image of Galactic Centre Region

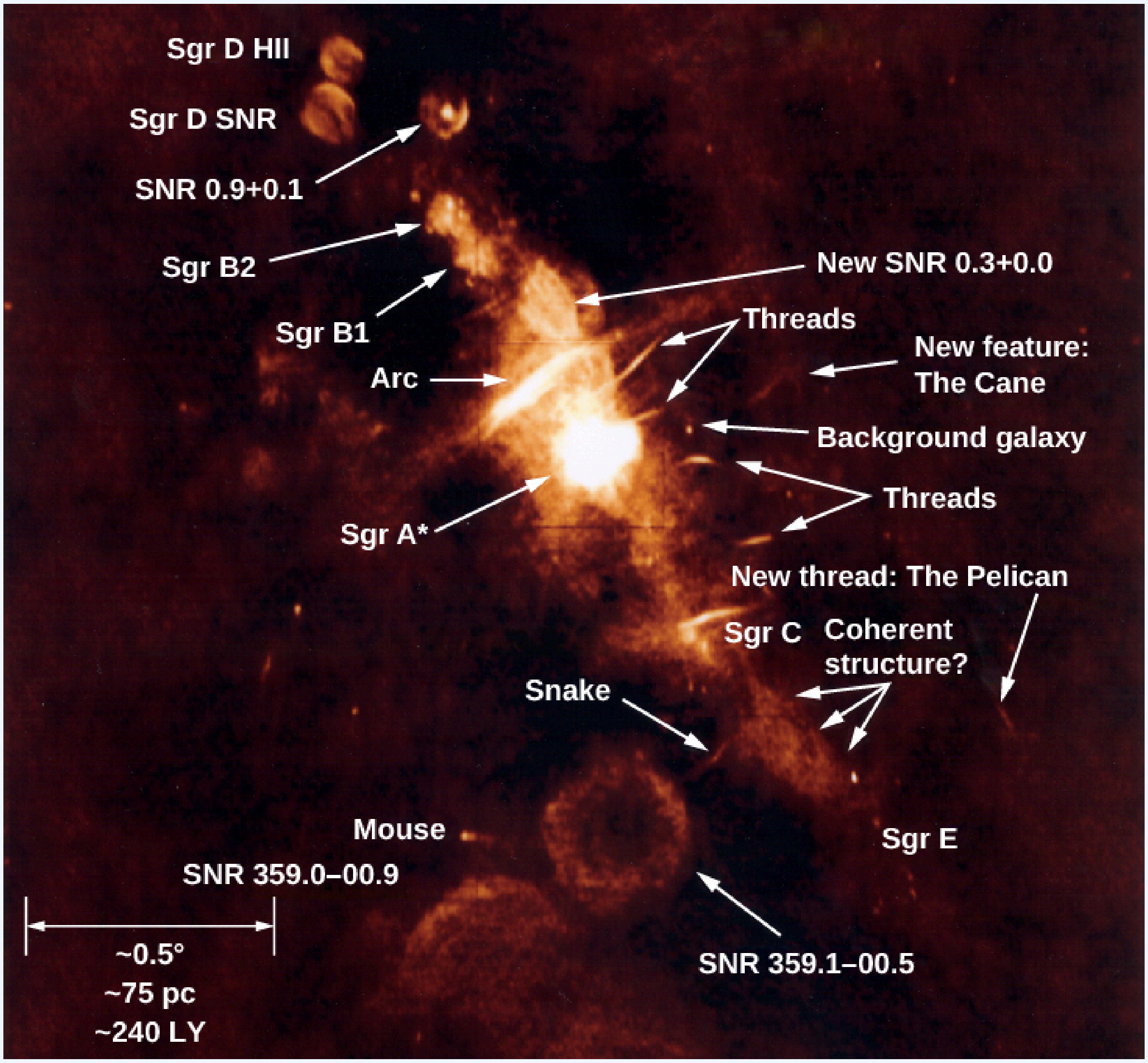


Figure 13.9. This radio map of the centre of the Galaxy (at a wavelength of 90 centimetres) was constructed from data obtained with the Very Large Array (VLA) of radio telescopes in Socorro, New Mexico. Brighter regions are more intense in radio waves. The galactic centre is inside the region labelled Sagittarius A. Sagittarius B1 and B2 are regions of active star formation. Many filaments or threadlike features are seen, as well as a number of shells (labelled SNR), which are supernova remnants. The scale bar at the bottom left is about 240 light-years long. Notice that radio astronomers also give fanciful animal names to some of the structures, much as visible-light nebulae are sometimes given the names of animals they resemble.

Galactic Center Radio Image by N. E. Kassim, D. S. Briggs, T. J. W. Lazio, T. N. LaRosa, and J. Imamura (NRL/RSD), Public Domain

Now let's focus in on the central region using a more energetic form of electromagnetic radiation. Figure 13.10 shows the X-ray emission from a smaller region 400 light-years wide and 900 light-years across centred in Sagittarius A*. Seen in this picture are hundreds of hot white dwarfs, neutron stars, and stellar black holes with accretion disks glowing with X-rays. The diffuse haze in the picture is emission from gas that lies among the stars and is at a temperature of 10 million K.

Galactic Centre in X-Rays

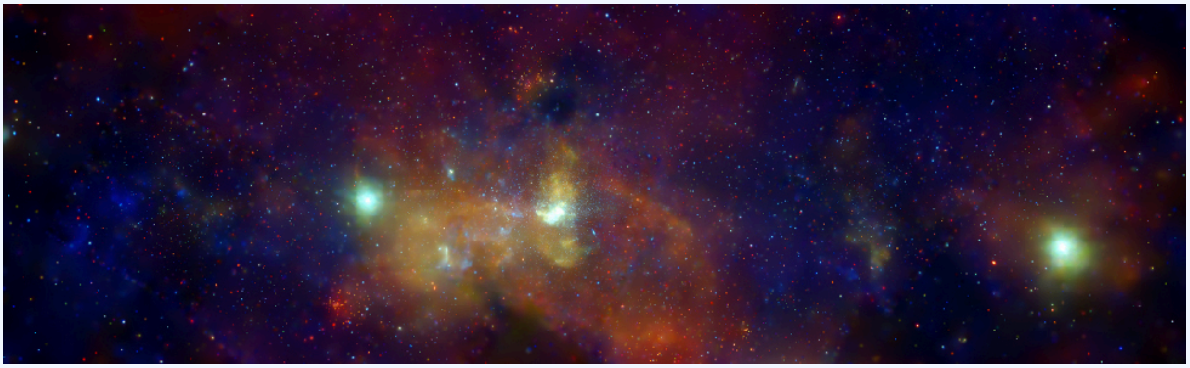


Figure 13.10. This artificial-colour mosaic of 30 images taken with the Chandra X-ray satellite shows a region 400 × 900 light-years in extent and centred on Sagittarius A*, the bright white source in the centre of the picture. The X-ray-emitting point sources are white dwarfs, neutron stars, and stellar black holes. The diffuse "haze" is emission from gas at a temperature of 10 million K. This hot gas is flowing away from the centre out into the rest of the Galaxy. The colours indicate X-ray energy bands: red (low energy), green (medium energy), and blue (high energy).
Galactic Center: New Vista of Milky Way Center Unveiled by NASA/CXC/ UMass/D. Wang et al., NASA Media License.

As we approach the centre of the Galaxy, we find the **supermassive black hole** Sagittarius A*. There are also thousands of stars within a parsec of Sagittarius A*. Most of these are old, reddish main-sequence stars. But there are also about a hundred hot OB stars that must have formed within the last few million years. There is as yet no good explanation for how stars could have formed recently so close to a supermassive black hole. Perhaps they formed in a dense cluster of stars that was originally at a larger distance from the black hole and subsequently migrated closer.

There is currently no star formation at the galactic centre, but there is lots of dust and molecular gas that is revolving around the black hole, along with some ionized gas streamers that are heated by the hot stars. Figure 13.11 is a radio map that shows these gas streamers.

Sagittarius A

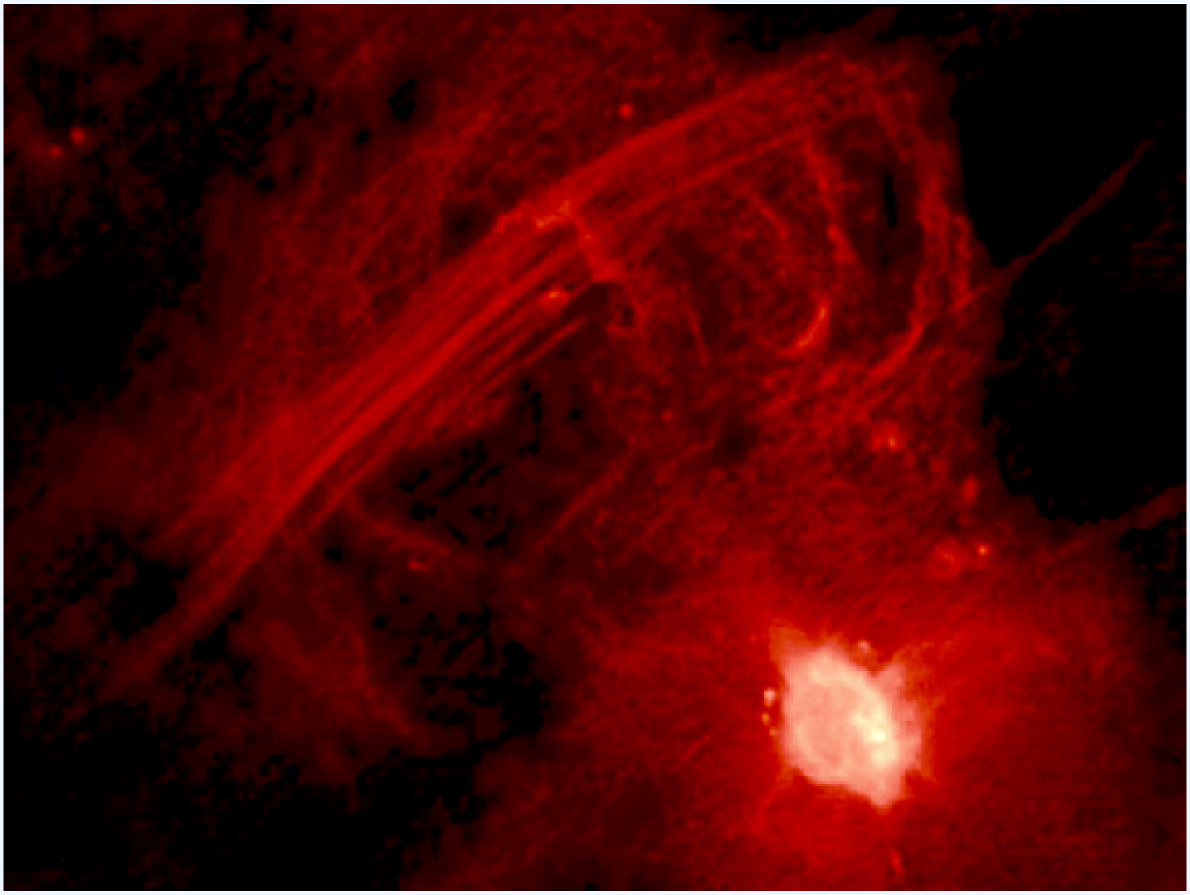


Figure 13.11. This image, taken with the Very Large Array of radio telescopes, shows the radio emission from hot, ionized gas in the centre of the Milky Way. The lines slanting across the top of the image are gas streamers. Sagittarius A* is the bright spot in the lower right.
The Galactic Center Radio Arc work by Farhad Zadeh et al. (Northwestern), VLA, NRAO, NASA Media License.

Just what is Sagittarius A*, which lies right at the centre our Galaxy? To establish that there really is a black hole there, we must show that there is a very large amount of mass crammed into a very tiny volume. Proving that a black hole exists is a challenge because the black hole itself emits no radiation. that a black hole exists is a challenge because the black hole itself emits no radiation. What astronomers must do is prove that a black hole is the only possible explanation for our observations—that a small region contains far more mass than could be accounted for by a very dense cluster of stars or something else made of ordinary matter.

To put some numbers with this discussion, the radius of the event horizon of a *galactic black hole* with a mass of about 4 million M_{Sun} would be only about 17 times the size of the Sun—the equivalent of a single red giant star. The corresponding density within this region of space would be much higher than that of any star cluster or any other ordinary astronomical object. Therefore, we must measure both the diameter of Sagittarius A* and its mass. Both radio and infrared observations are required to give us the necessary evidence.

First, let's look at how the mass can be measured. If we zero in on the inner few light-days of the Galaxy with an infrared telescope equipped with adaptive optics, we see a region crowded with individual stars as shown in Figure 13.12. These stars have now been observed for almost two decades, and astronomers have detected their rapid orbital motions around the very centre of the Galaxy.

Near-Infrared View of the Galactic Centre

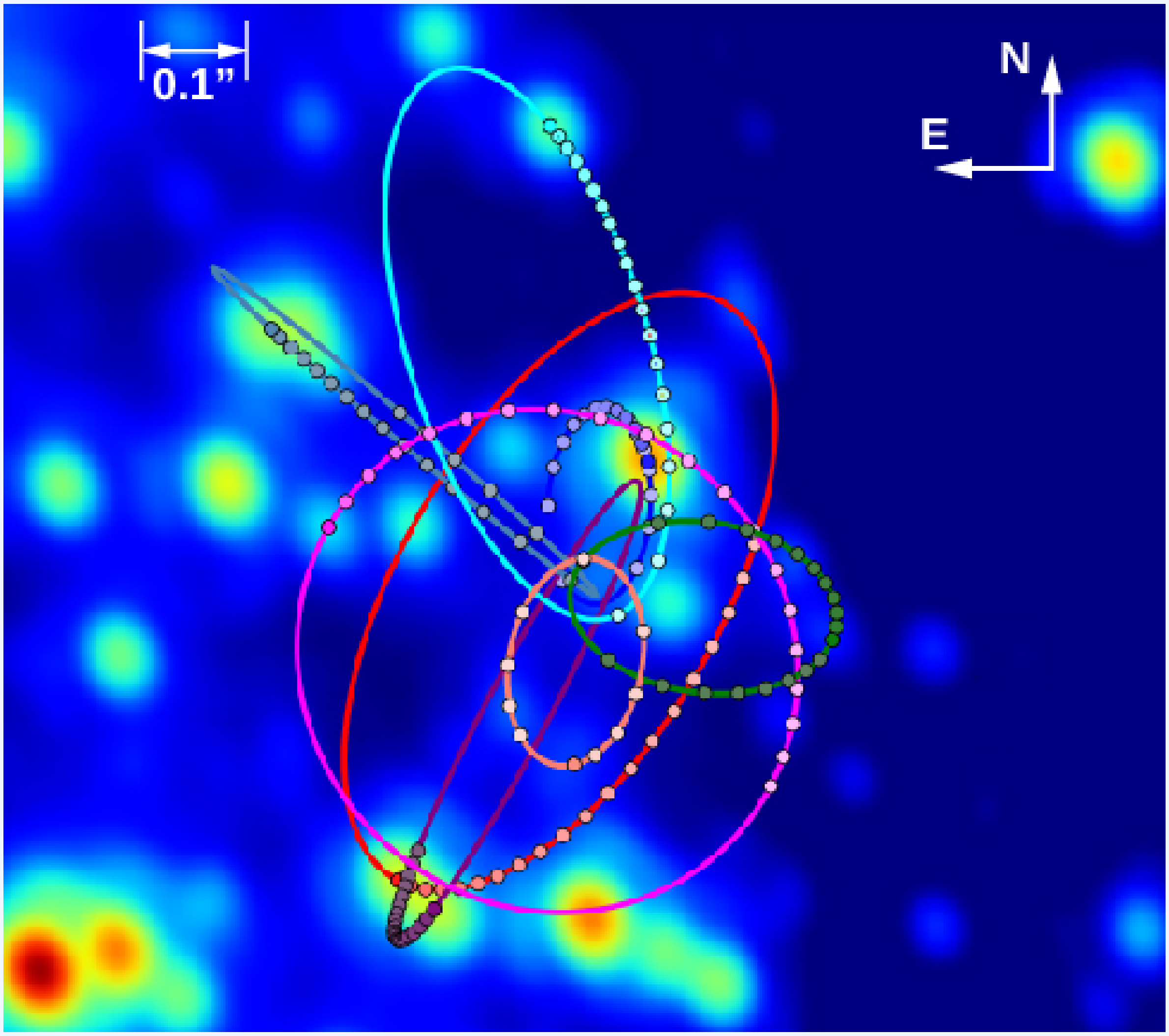


Figure 13.12. This image shows the inner 1 arcsecond, or 0.13 light-year, at the centre of the Galaxy, as observed with the giant Keck Telescope. Tracks of the orbiting stars measured from 1995 to 2014 have been added to this “snapshot.” The stars are moving around the centre very fast, and their tracks are all consistent with a single massive “gravitator” that resides in the very centre of this image. Stellar Orbits about the Galactic Center 1995-2014 by Andrea Ghez, UCLA Galactic Center Group, W.M. Keck Observatory Laser Team, UCLA Galactic Center Group Image Use Policy

If we combine observations of their periods and the size of their orbits with Kepler’s third law, we can estimate the mass of the object that keeps them in their orbits. One of the stars has been observed for its full orbit of 15.6 years. Its closest approach takes it to a distance of only 124 AU or about 17 light-hours from the black hole. This orbit, when combined with observations of other stars close to the galactic centre, indicates that a mass of 4.6 million M_{Sun} must be concentrated inside the orbit—that is, within 17 light-hours of the centre of the Galaxy.

Even tighter limits on the size of the concentration of mass at the centre of the Galaxy come from radio astronomy, which provided the first clue that a black hole might lie at the centre of the Galaxy. As matter spirals inward toward the event horizon of a black hole, it is heated in a whirling *accretion disk* and produces radio radiation. Measurements of the size of the accretion disk with the Very Long Baseline Array, which provides very high spatial resolution, show that the diameter of the radio source Sagittarius A^* is no larger than about 0.3 AU, or about the size of Mercury's orbit. (In light units, that's only 2.5 light-*minutes*!)

The observations thus show that 4.6 million solar masses are crammed into a volume that has a diameter that is no larger than the orbit of Mercury. If this were anything other than a supermassive black hole—low-mass stars that emit very little light or neutron stars or a very large number of small black holes— calculations show that these objects would be so densely packed that they would collapse to a single black hole within a hundred thousand years. That is a very short time compared with the age of the Galaxy, which probably began forming more than 13 billion years ago. Since it seems very unlikely that that we would have caught such a complex cluster of objects just before it collapsed, the evidence for a supermassive black hole at the centre of the Galaxy is convincing indeed.

Where did our galactic black hole come from? The origin of supermassive black holes in galaxies like ours is currently an active field of research. One possibility is that a large cloud of gas near the centre of the Milky Way collapsed directly to form a black hole. Since we find large black holes at the centres of most other large galaxies—even ones that are very young—this collapse probably would have taken place when the Milky Way was just beginning to take shape. The initial mass of this black hole might have been only a few tens of solar masses. Another way it could have started is that a massive star might have exploded to leave behind a seed black hole, or a dense cluster of stars might have collapsed into a black hole.

Once a black hole exists at the centre of a galaxy, it can grow over the next several billion years by devouring nearby stars and gas clouds in the crowded central regions. It can also grow by merging with other black holes.

It appears that the monster black hole at the centre of our Galaxy is not finished "eating." At the present time, we observe clouds of gas and dust falling into the galactic centre at the rate of about 1 M_{Sun} per thousand years. Stars are also on the black hole's menu. The density of stars near the galactic centre is high enough that we would expect a star to pass near the black hole and be swallowed by it every ten thousand years or so. As this happens, some of the energy of infall is released as radiation. As a result, the centre of the Galaxy might flare up and even briefly outshine all the stars in the Milky Way. Other objects might also venture too close to the black hole and be pulled in. How great a flare we observe would depend on the mass of the object falling in.

In 2013, the Chandra X-ray satellite detected a flare from the centre of our Galaxy that was 400 times brighter than the usual output from Sagittarius A*. A year later, a second flare, only half as bright, was also detected. This is much less energy than swallowing a whole star would produce. There are two theories to account for the flares. First, an asteroid might have ventured too close to the black hole and been heated to a very high temperature before being swallowed up. Alternatively, the flares might have involved interactions of the magnetic fields near the galactic centre in a process similar to the one described for solar flares. Astronomers continue to monitor the galactic centre area for flares or other activity. Although the monster in the centre of the Galaxy is not close enough to us to represent any danger, we still want to keep our eyes on it.

Andrea Ghez

A lover of puzzles, Andrea Ghez has been pursuing one of the greatest mysteries in astronomy: what strange entity lurks within the centre of our Milky Way Galaxy?

Andrea Ghez

Figure 13.13. Research by Ghez and her team has helped shape our understanding of supermassive black holes.
Andrea Ghez by John D. and Catherine T. MacArthur Foundation, CC BY 4.0.

As a child living in Chicago during the late 1960s, Andrea Ghez, pictured in Figure 13.13, was fascinated by the Apollo Moon landings. But she was also drawn to ballet and to solving all sorts of puzzles. By high school, she had lost the ballet bug in favour of competing in field hockey, playing the flute, and digging deeper into academics. Her undergraduate years at MIT were punctuated by a number of changes in her major—from mathematics to chemistry, mechanical engineering, aerospace engineering, and finally physics—where she felt her options were most open. As a physics major, she became involved in astronomical research under the guidance of one of her instructors. Once she got to do some actual observing at Kitt Peak National Observatory in Arizona, and later at Cerro Tololo Inter-American Observatory in Chile, Ghez had found her calling.

Pursuing her graduate studies at Caltech, she stuck with physics but oriented her efforts toward observational astrophysics, an area where Caltech had access to cutting-edge facilities. Though initially attracted to studying the black holes that were suspected of dwelling inside most massive galaxies, Ghez ended up spending most of her graduate study and later postdoctoral research at the University of Arizona studying stars in formation. By taking very high-resolution (detailed) imaging of regions where new stars are born, she discovered that most stars form as members of binary systems. As technologies advanced, she was able to track the orbits danced by these stellar pairings and thereby could ascertain their respective masses.

Now an astronomy professor at UCLA, Ghez has since used similar high-resolution imaging techniques to study the orbits of stars in the innermost core of the Milky Way. These orbits take years to delineate, so Ghez and her science team have logged more than 20 years of taking super-resolution infrared images with the giant Keck telescopes in Hawaii. Based on the resulting stellar orbits, the UCLA Galactic Center Group has settled (as we saw) on a gravitational solution that requires the presence of a supermassive black hole with a mass equivalent to 4.6 million Suns—all nestled within a space smaller than that occupied by our solar system. Ghez's achievements have been recognized with one of the "genius" awards given by the MacArthur Foundation. More recently, her team discovered glowing clouds of warm ionized gas that co-orbit with the stars but may be more vulnerable to the disruptive effects of the central black hole. By monitoring these clouds, the team hopes to better understand the evolution of supermassive black holes and their immediate environs. They also hope to test Einstein's theory of general relativity by carefully scrutinizing the orbits of stars that careen closest to the intensely gravitating black hole.

Besides her pioneering work as an astronomer, Ghez competes as a master swimmer, enjoys family life as a mother of two children, and actively encourages other women to pursue scientific careers.

Attribution

13.5 QUASARS

The name “**quasars**” started out as short for “quasi-stellar radio sources” (here “quasi-stellar” means “sort of like stars”). The discovery of radio sources that appeared point-like, just like stars, came with the use of surplus World War II radar equipment in the 1950s. Although few astronomers would have predicted it, the sky turned out to be full of strong sources of radio waves. As they improved the images that their new radio telescopes could make, scientists discovered that some radio sources were in the same location as faint blue “stars.” No known type of star in our Galaxy emits such powerful radio radiation. What then were these “quasi-stellar radio sources”?

The answer came when astronomers obtained visible-light spectra of two of those faint “blue stars” that were strong sources of radio waves as shown in Figure 13.14. Spectra of these radio “stars” only deepened the mystery: they had emission lines, but astronomers at first could not identify them with any known substance. By the 1960s, astronomers had a century of experience in identifying elements and compounds in the spectra of stars. Elaborate tables had been published showing the lines that each element would produce under a wide range of conditions. A “star” with unidentifiable lines in the ordinary visible light spectrum had to be something completely new.

Typical Quasar

Figure 13.14. The arrow in this image marks the quasar known by its catalog number, PKS 1117-248. Note that nothing in this image distinguishes the quasar from an ordinary star. Its spectrum, however, shows that it is moving away from us at a speed of 36% the speed of light, or 67,000 miles per second. In contrast, the maximum speed observed for any star is only a few hundred miles per second. QSO H1821+643 Indicates a Universe Filled with Hydrogen by Todd M. Tripp (Princeton) et al. WIYN Observatory, NOAO, NSF; & HST, NASA, NASA Media License.

In 1963 at Caltech's Palomar Observatory, Maarten Schmidt was puzzling over the spectrum of one of the radio stars, which was named 3C 273 because it was the 273rd entry in the third Cambridge catalog of radio sources (Figure 13.15). There were strong emission lines in the spectrum, and Schmidt recognized that they had the same spacing between them as the Balmer lines of hydrogen. But the lines in 3C 273 were shifted far to the red of the wavelengths at which the Balmer lines are normally located. Indeed, these lines were at such long wavelengths that if the redshifts were attributed to the Doppler effect, 3C 273 was receding from us at a speed of 45,000 kilometres per second, or about 15% the speed of light! Since stars don't show Doppler shifts this large, no one had thought of considering high redshifts to be the cause of the strange spectra.

Quasar 3C 273

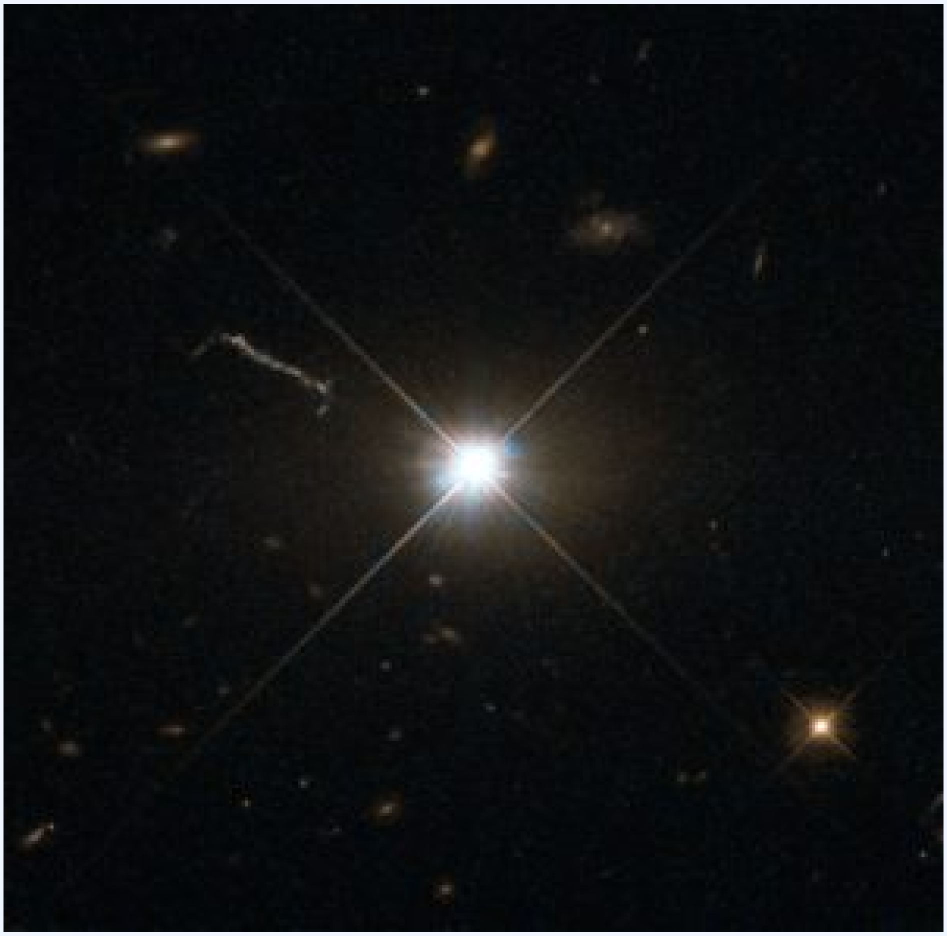


Figure 13.15. This is the first quasar for which a redshift was measured. The redshift showed that the light from it took about 2.5 billion years to reach us. Despite this great distance, it is still one of the quasars closest to the Milky Way Galaxy. Note also the faint streak going toward the upper left from the quasar. Some quasars, like 3C 273, eject super-fast jets of material. The jet from 3C 273 is about 200,000 light-years long.
Best image of bright quasar 3C 273 by ESA/Hubble/NASA, ESA Standard License.

The puzzling emission lines in other star-like radio sources were then reexamined to see if they, too, might be well-known lines with large redshifts. This proved to be the case, but the other objects were found to be receding from us at even greater speeds. Their astounding speeds showed that the radio "stars" could not possibly be stars in our own Galaxy. Any true star moving at more than a few hundred kilometres per second would be able to overcome the gravitational pull of the Galaxy and completely escape from it. (As we shall see later in this chapter, astronomers eventually discovered that there was also more to these "stars" than just a point of light.)

It turns out that these high-velocity objects only look like stars because they are compact and very far away. Later, astronomers discovered objects with large redshifts that appear star-like but have no radio emission. Observations also showed that quasars were bright in the infrared and X-ray bands too, and not all these X-ray or infrared-bright quasars could be seen in either the radio or the visible-light bands of the spectrum. Today, all these objects are referred to as *quasi-stellar objects* (*QSOs*), or, as they are more popularly known, quasars. (The name was also soon appropriated by a manufacturer of home electronics.)

Read an interview with Maarten Schmidt on the fiftieth anniversary of his insight about the spectrum of quasars and their redshifts. Direct link: https://www.space.com/20244-quasar-mystery-discoverer-interview.html

Over a million quasars have now been discovered, and spectra are available for over a hundred thousand. All these spectra show redshifts, none show blueshifts, and their redshifts can be very large. Yet in a photo they look just like stars as shown in Figure 13.16.

Typical Quasar Imaged by the Hubble Space Telescope

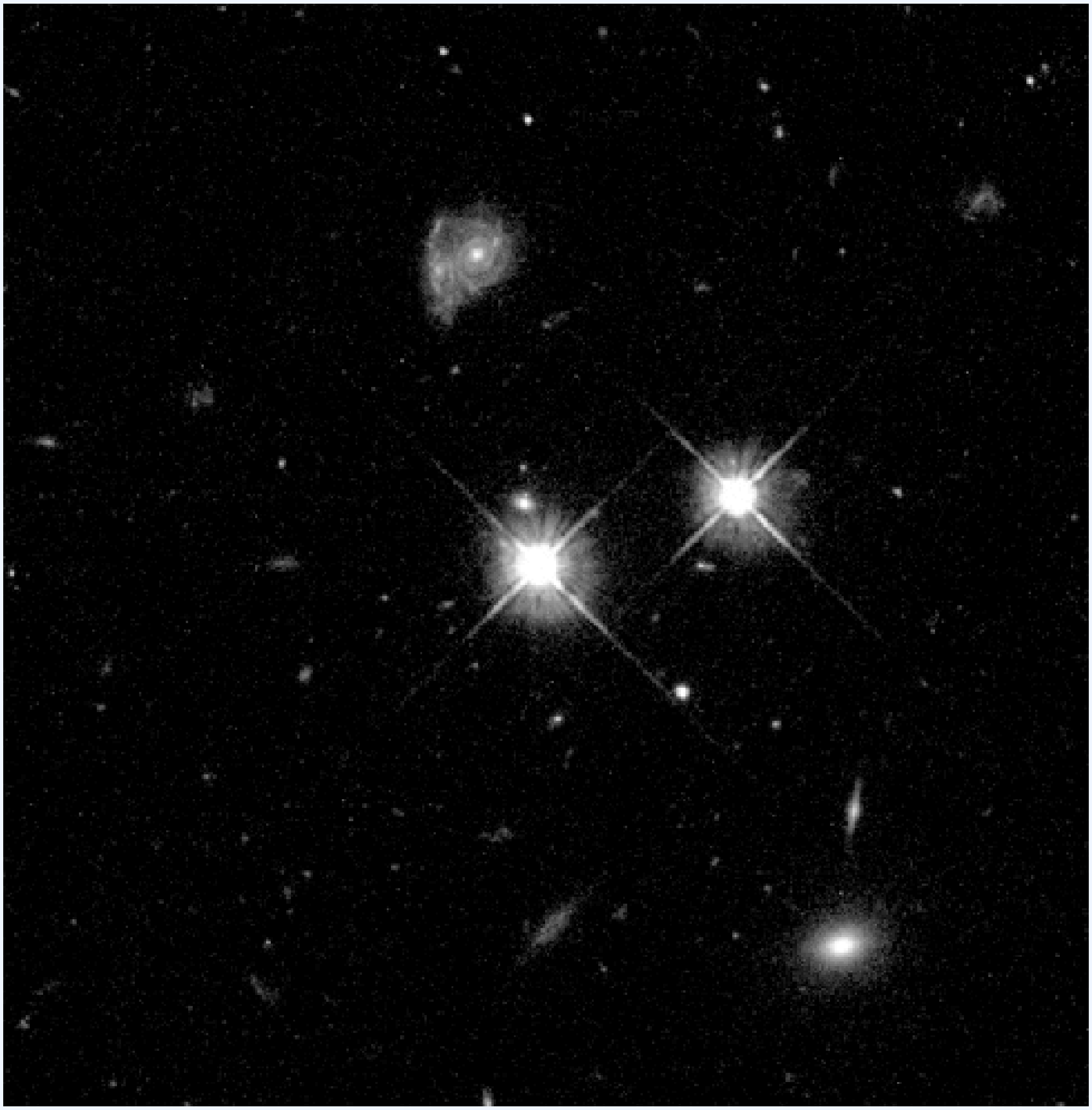


Figure 13.16. One of these two bright “stars” in the middle is in our Galaxy, while the other is a quasar 9 billion light-years away. From this picture alone, there’s no way to say which is which. (The quasar is the one in the center of the picture.)
Hubble’s 100,000th Exposure by Charles Steidel (CIT)/NASA/ESA, ESA Standard License.

In the record-holding quasars, the first Lyman series line of hydrogen, with a laboratory wavelength of 121.5

nanometres in the ultraviolet portion of the spectrum, is shifted all the way through the visible region to the infrared. At such high redshifts, the simple formula for converting a Doppler shift to speed must be modified to take into account the effects of the theory of relativity. If we apply the relativistic form of the Doppler shift formula, we find that these redshifts correspond to velocities of about 96% of the speed of light.

Given their large distances, quasars have to be extremely luminous to be visible to us at all—far brighter than any normal galaxy. In visible light alone, most are far more energetic than the brightest elliptical galaxies. But, as we saw, quasars also emit energy at X-ray and ultraviolet wavelengths, and some are radio sources as well. When all their radiation is added together, some QSOs have total luminosities as large as a hundred trillion Suns ($10^{14}L_{\text{Sun}}$), which is 10 to 100 times the brightness of luminous elliptical galaxies.

Finding a mechanism to produce the large amount of energy emitted by a quasar would be difficult under any circumstances. But there is an additional problem. When astronomers began monitoring quasars carefully, they found that some vary in luminosity on time scales of months, weeks, or even, in some cases, days. This variation is irregular and can change the brightness of a quasar by a few tens of percent in both its visible light and radio output.

Think about what such a change in luminosity means. A quasar at its dimmest is still more brilliant than any normal galaxy. Now imagine that the brightness increases by 30% in a few weeks. Whatever mechanism is responsible must be able to release new energy at rates that stagger our imaginations. The most dramatic changes in quasar brightness are equivalent to the energy released by 100,000 billion Suns. To produce this much energy we would have to convert the total mass of about ten Earths into energy every minute.

Moreover, because the fluctuations occur in such short times, the part of a quasar that is varying must be smaller than the distance light travels in the time it takes the variation to occur—typically a few months. To see why this must be so, let's consider a cluster of stars 10 light-years in diameter at a very large distance from Earth (see Figure 13.17 in which Earth is off to the right). Suppose every star in this cluster somehow brightens simultaneously and remains bright. When the light from this event arrives at Earth, we would first see the brighter light from stars on the near side; 5 years later we would see increased light from stars at the center. Ten years would pass before we detected more light from stars on the far side.

How the Size of a Source Affects the Timescale of Its Variability

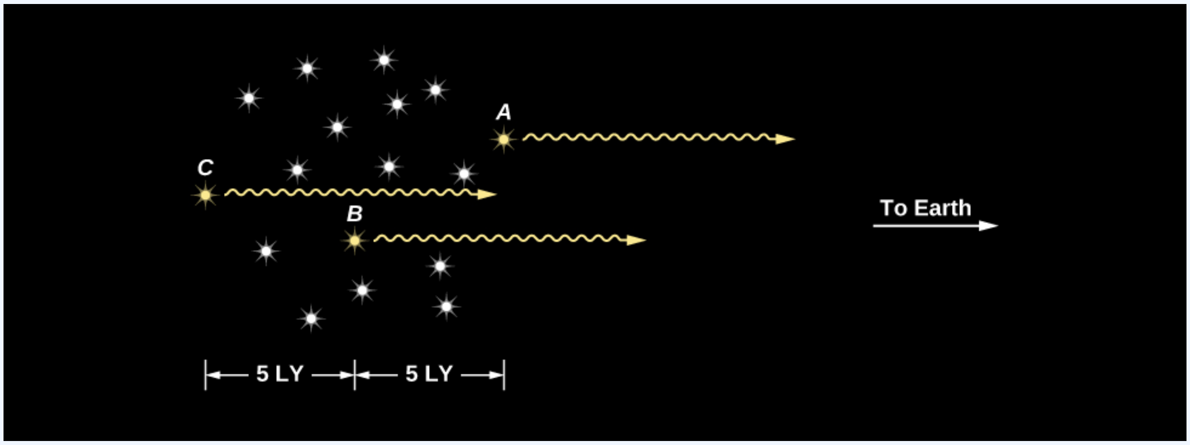


Figure 13.17. This diagram shows why light variations from a large region in space appear to last for an extended period of time as viewed from Earth. Suppose all the stars in this cluster, which is 10 light-years across, brighten simultaneously and instantaneously. From Earth, star A will appear to brighten 5 years before star B, which in turn will appear to brighten 5 years earlier than star C. It will take 10 years for an Earth observer to get the full effect of the brightening.

Even though all stars in the cluster brightened at the same time, the fact that the cluster is 10 light-years wide means that 10 years must elapse before the increased light from every part of the cluster reaches us. From Earth we would see the cluster get brighter and brighter, as light from more and more stars began to reach us. Not until 10 years after the brightening began would we see the cluster reach maximum brightness. In other words, if an extended object suddenly flares up, it will seem to brighten over a period of time equal to the time it takes light to travel across the object from its far side.

We can apply this idea to brightness changes in quasars to estimate their diametres. Because quasars typically vary (get brighter and dimmer) over periods of a few months, the region where the energy is generated can be no larger than a few light-months across. If it were larger, it would take longer than a few months for the light from the far side to reach us.

How large is a region of a few light-months? Pluto, usually the outermost (dwarf) planet in our solar system, is about 5.5 light-hours from us, while the nearest star is 4 light-years away. Clearly a region a few light months across is tiny relative to the size of the entire Galaxy. And some quasars vary even more rapidly, which means their energy is generated in an even smaller region. Whatever mechanism powers the quasars must be able to generate more energy than that produced by an entire galaxy in a volume of space that, in some cases, is not much larger than our solar system.

Even before the discovery of quasars, there had been hints that something very strange was going on in the centers of at least some galaxies. Back in 1918, American astronomer Heber Curtis used the large Lick Observatory telescope to photograph the galaxy Messier 87 in the constellation Virgo. On that photograph, he saw what we now call a jet coming from the centre, or nucleus, of the galaxy, pictured in Figure 13.18. This jet literally and figuratively pointed to some strange activity going on in that galaxy nucleus. But he had no idea what it was. No one else knew what to do with this space oddity either.

The random factoid that such a central jet existed lay around for a quarter century, until Carl Seyfert, a young astronomer at Mount Wilson Observatory, also in California, found half a dozen galaxies with extremely bright nuclei that were almost stellar, rather than fuzzy in appearance like most galaxy nuclei. Using spectroscopy, he found that these nuclei contain gas moving at up to two percent the speed of light. That may not sound like much, but it is 6 million miles per hour, and more than 10 times faster than the typical motions of stars in galaxies.

M87 Jet

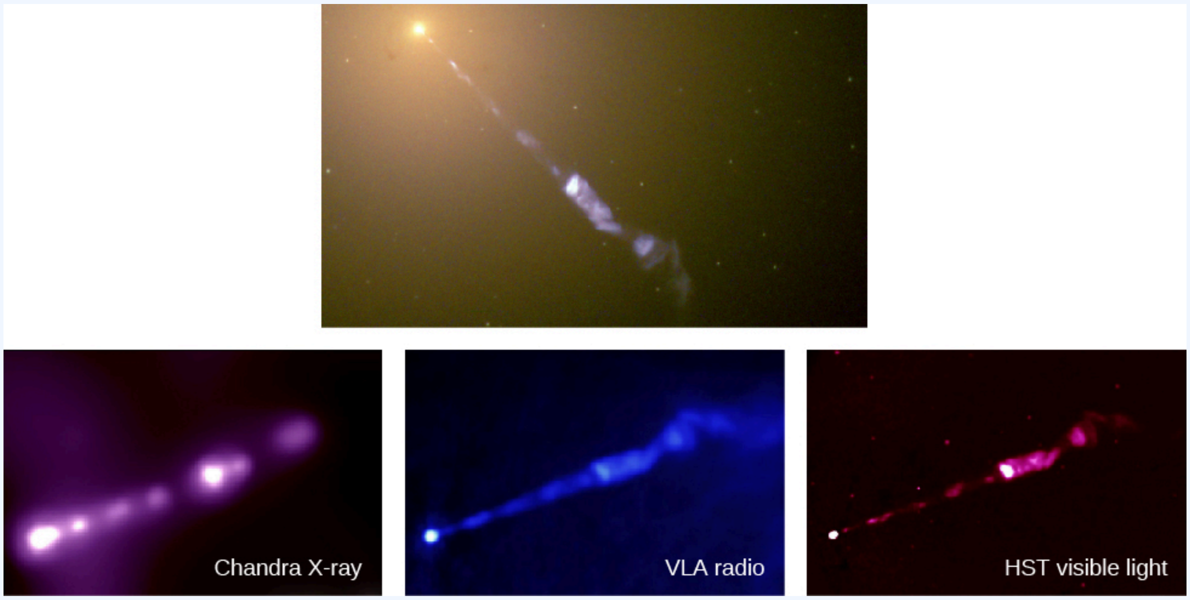


Figure 13.18. Streaming out like a cosmic searchlight from the center of the galaxy, M87 is one of nature's most amazing phenomena, a huge jet of electrons and other particles travelling at nearly the speed of light. In this Hubble Space Telescope image, the blue of the jet contrasts with the yellow glow from the combined light of billions of unseen stars and yellow, point-like globular clusters that make up the galaxy (at the upper left). As we shall see later in this chapter, the jet, which is several thousand light-years long, originates in a disk of superheated gas swirling around a giant black hole at the centre of M87. The light that we see is produced by electrons twisting along magnetic field lines in the jet, a process known as synchrotron radiation, which gives the jet its bluish tint. The jet in M87 can be observed in X-ray, radio, and visible light, as shown in the bottom three images. At the extreme left of each bottom image, we see the bright galactic nucleus harbouring a supermassive black hole.
Credit top: M87 jet by NASA, The Hubble Heritage Team(STScI/AURA), NASA Media License.
Credit bottom: M87 Jet: Chandra Sheds Light on the Knotty Problem of the M87 Jet by X-ray: H. Marshall (MIT), et al., CXC, NASA; Radio: F. Zhou, F. Owen (NRAO), J. Biretta (STScI); Optical: E. Perlman (UMBC), et al., NASA Media License.

After decades of study, astronomers identified many other strange objects beyond our Milky Way Galaxy; they populate a whole "zoo" of what are now called active galaxies or **active galactic nuclei (AGN)**. Astronomers first called them by many different names, depending on what sorts of observations discovered each category, but now we know that we are always looking at the same basic mechanism. What all these galaxies have in common is some activity in their nuclei that produces an enormous amount of energy in a very small volume of space. In the next section, we describe a model that explains all these galaxies with strong central activity—both the AGNs and the QSOs.

To see a jet for yourself, check out a time-lapse video of the jet ejected from NGC 3862. Direct link: https://hubblesite.org/video/782/category/10-black-holes

Attribution

13.6 SUPERMASSIVE BLACK HOLES

In order to find a common model for quasars (and their cousins, the AGNs), let's first list the common characteristics we have been describing—and add some new ones:

- Quasars are hugely powerful, emitting more power in radiated light than all the stars in our Galaxy combined.
- Quasars are tiny, about the size of our solar system (to astronomers, that is really small!).
- Some quasars are observed to be shooting out pairs of straight jets at close to the speed of light, in a tight beam, to distances far beyond the galaxies they live in. These jets are themselves powerful sources of radio and gamma-ray radiation.
- Because quasars put out so much power from such a small region, they can't be powered by nuclear fusion the way stars are; they must use some process that is far more efficient.
- As we shall see later in this chapter, quasars were much more common when the universe was young than they are today. That means they must have been able to form in the first billion years or so after the universe began to expand.

The readers of this text are in a much better position than the astronomers who discovered quasars in the 1960s to guess what powers the quasars. That's because the key idea in solving the puzzle came from observations of the black holes. The discovery of the first stellar mass black hole in the binary system Cygnus X-1 was announced in 1971, several years after the discovery of quasars. Proof that there is a black hole at the centre of our own Galaxy came even later. Back when astronomers first began trying to figure out what powered quasars, black holes were simply one of the more exotic predictions of the general theory of relativity that still waited to be connected to the real world.

It was only as proof of the existence of black holes accumulated over several decades that it became clearer that only supermassive black holes could account for all the observed properties of quasars and AGNs. As we saw in earlier chapters, our own Galaxy has a black hole in its centre, and the energy is emitted from a small central region. While our black hole doesn't have the mass or energy of the quasar black holes, the mechanism that powers them is similar. The evidence now shows that most—and probably all—elliptical galaxies and all spirals with nuclear bulges have black holes at their centres. The amount of energy emitted by material near the black hole depends on two things: the mass of the black hole and the amount of matter that is falling into it.

If a black hole with a billion Suns' worth of mass inside ($10^9 M_{\text{Sun}}$) accretes (gathers) even a relatively modest amount of additional material—say, about 10 M_{Sun} per year—then (as we shall see) it can, in the process, produce as much energy as a thousand normal galaxies. This is enough to account for the total energy of a quasar. If the mass of the black hole is smaller than a billion solar masses or the accretion rate is low, then the amount of energy emitted can be much smaller, as it is in the case of the Milky Way.

Watch a video of an artist's impression of matter accreting around a supermassive black hole.
Direct link: https://www.spacetelescope.org/videos/hubblecast43d/

Observational Evidence for Black Holes

In order to prove that a black hole is present at the centre of a galaxy, we must demonstrate that so much mass is crammed into so small a volume that no normal objects—massive stars or clusters of stars—could possibly account for it (just as we did for the black hole in the Milky Way). We already know from observations that an accreting black hole is surrounded by a hot *accretion disk* with gas and dust that swirl around the black hole before it falls in.

If we assume that the energy emitted by quasars is also produced by a hot accretion disk, then, as we saw in the previous section, the size of the disk must be given by the time the quasar energy takes to vary. For quasars, the emission in visible light varies on typical time scales of 5 to 2000 days, limiting the size of the disk to that many light-days.

In the X-ray band, quasars vary even more rapidly, so the light travel time argument tells us that this more energetic radiation is generated in an even smaller region. Therefore, the mass around which the accretion disk is swirling must be confined to a space that is even smaller. If the quasar mechanism involves a great deal of mass, then the only astronomical object that can confine a lot of mass into a very small space is a black hole. In a few cases, it turns out that the X-rays are emitted from a region just a few times the size of the black hole event horizon.

The next challenge, then, is to "weigh" this central mass in a quasar. In the case of our own Galaxy, we used observations of the orbits of stars very close to the galactic centre, along with Kepler's third law, to estimate the mass of the central black hole. In the case of distant galaxies, we cannot measure the orbits of individual stars, but we can measure the orbital speed of the gas in the rotating accretion disk. The Hubble Space Telescope is especially well suited to this task because it is above the blurring of Earth's atmosphere and can obtain spectra very close to the bright central regions of active galaxies. The Doppler effect is then used to measure radial velocities of the orbiting material and so derive the speed with which it moves around.

One of the first galaxies to be studied with the Hubble Space Telescope is our old favourite, the giant elliptical M87. Hubble Space Telescope images showed that there is a disk of hot (10,000 K) gas swirling around the centre of M87 as shown in Figure 13.19. It was surprising to find hot gas in an elliptical galaxy because this type of galaxy is usually devoid of gas and dust. But the discovery was extremely useful for pinning down the existence of the black hole. Astronomers measured the Doppler shift of spectral lines emitted by this gas, found its speed of rotation, and then used the speed to derive the amount of mass inside the disk—applying Kepler's third law.

Evidence for a Black Hole at the Centre of M87

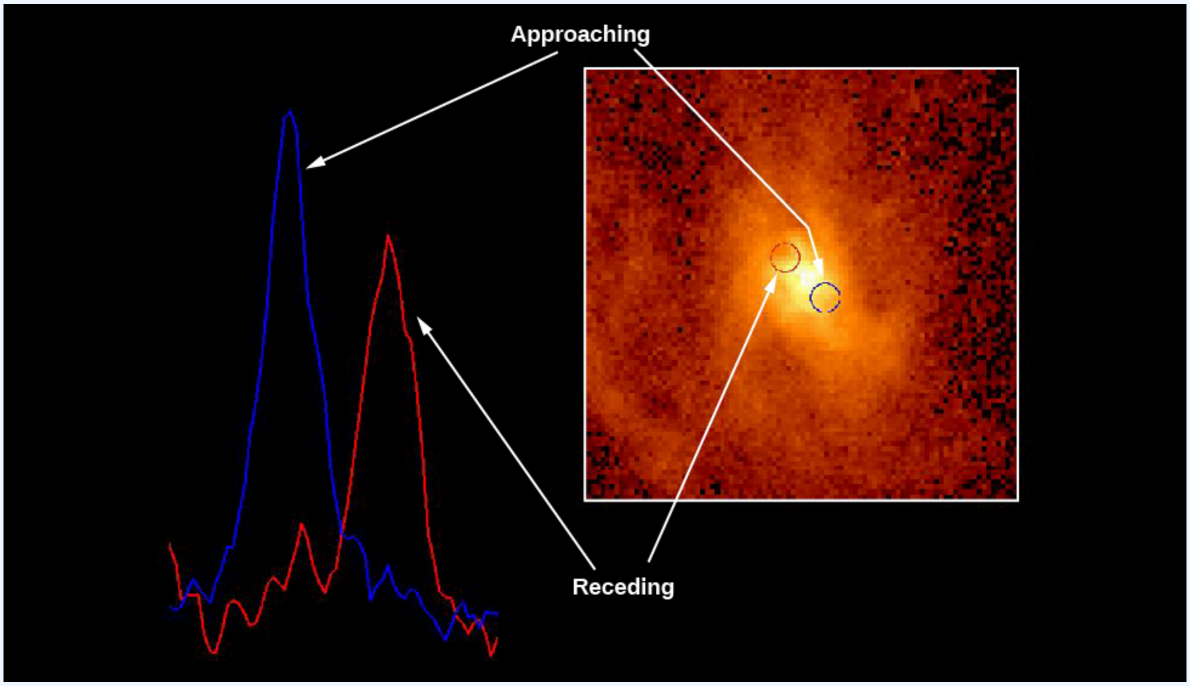


Figure 13.19. The disk of whirling gas at right was discovered at the centre of the giant elliptical galaxy M87 with the Hubble Space Telescope. Observations made on opposite sides of the disk show that one side is approaching us (the spectral lines are blueshifted by the Doppler effect) while the other is receding (lines redshifted), a clear indication that the disk is rotating. The rotation speed is about 550 kilometres per second or 1.2 million miles per hour. Such a high rotation speed is evidence that there is a very massive black hole at the centre of M87.
Spectrum of gas disc in active galaxy M87 by Holland Ford, Space Telescope Science Institute/Johns Hopkins University; Richard Harms, Applied Research Corp.; Zlatan Tsvetanov, Arthur Davidsen, and Gerard Kriss at Johns Hopkins; Ralph Bohlin and GeorgeHartig at Space Telescope Science Institute; Linda Dressel and Ajay K.Kochhar at Applied Research Corp. in Landover, Md.; and Bruce Margon fromthe University of Washington in Seattle NASA/ESA, NASA Media License.

Modern estimates show that there is a mass of at least 3.5 billion M_{Sun} concentrated in a tiny region at the very centre of M87. So much mass in such a small volume of space must be a black hole. Let's stop for a moment and take in this figure: a single black hole that has swallowed enough material to make 3.5 billion stars like the

Sun. Few astronomical measurements have ever led to so mind-boggling a result. What a strange environment the neighbourhood of such a supermassive black hole must be.

Another example is shown in Figure 13.20. Here, we see a disk of dust and gas that surrounds a 300-million-M_{Sun} black hole in the centre of an elliptical galaxy. (The bright spot in the centre is produced by the combined light of stars that have been pulled close together by the gravitational force of the black hole.) The mass of the black hole was again derived from measurements of the rotational speed of the disk. The gas in the disk is moving around at 155 kilometres per second at a distance of only 186 light-years from its centre. Given the pull of the mass at the centre, we expect that the whole dust disk should be swallowed by the black hole in several billion years.

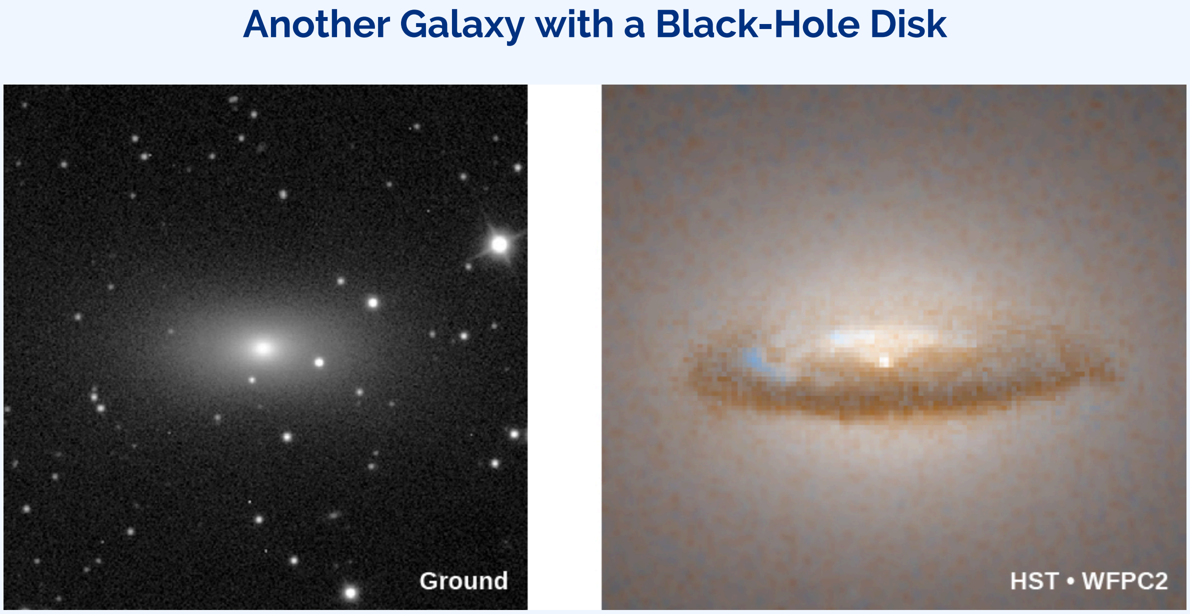


Figure 13.20. The ground-based image shows an elliptical galaxy called NGC 7052 located in the constellation of Vulpecula, almost 200 million light-years from Earth. At the galaxy's centre (right) is a dust disk roughly 3700 light-years in diameter. The disk rotates like a giant merry-go-round: gas in the inner part (186 light-years from the centre) whirls around at a speed of 155 kilometres per second (341,000 miles per hour). From these measurements and Kepler's third law, it is possible to estimate that the disk is orbiting around a central black hole with a mass of 300 million Suns.
Modification of Disk around a Black Hole in Galaxy NGC 7052 by Roeland P. van der Marel (STScI), Frank C. van den Bosch (Univ. of Washington), and NASA/ESA, ESA Standard License.

But do we *have* to accept black holes as the only explanation of what lies at the centre of these galaxies? What else could we put in such a small space other than a giant black hole? The alternative is stars. But to explain the masses in the centres of galaxies without a black hole we need to put at least a million stars in a region the size of the solar system. To fit, they would have be only 2 star diametres apart. Collisions between stars would happen all the time. And these collisions would lead to mergers of stars, and very soon the one giant star that they form would collapse into a black hole. So there is really no escape: only a black hole can fit so much mass into so small a space.

As we saw earlier, observations now show that all the galaxies with a spherical concentration of stars—either elliptical galaxies or spiral galaxies with nuclear bulges—harbour one of these giant black holes at their centres. Among them is our neighbour spiral galaxy, the Andromeda galaxy, M31. The masses of these central black holes range from a just under a million up to at least 30 billion times the mass of the Sun. Several black holes may be even more massive, but the mass estimates have large uncertainties and need verification. We call these black holes "supermassive" to distinguish them from the much smaller black holes that form when some stars die. So far, the most massive black holes from stars—those detected through gravitational waves detected by LIGO—have masses only a little over 30 solar masses.

Energy Production around a Black Hole

By now, you may be willing to entertain the idea that huge black holes lurk at the centres of active galaxies. But we still need to answer the question of how such a black hole can account for one of the most powerful sources of energy in the universe. A black hole itself can radiate no energy. Any energy we detect from it must come from material very close to the black hole, but not inside its event horizon.

In a galaxy, a central black hole (with its strong gravity) attracts matter—stars, dust, and gas—orbiting in the dense nuclear regions. This matter spirals in toward the spinning black hole and forms an accretion disk of material around it. As the material spirals ever closer to the black hole, it accelerates and becomes compressed, heating up to temperatures of millions of degrees. Such hot matter can radiate prodigious amounts of energy as it falls in toward the black hole.

To convince yourself that falling into a region with strong gravity can release a great deal of energy, imagine dropping a printed version of your astronomy textbook out the window of the ground floor of the library. It will land with a thud, and maybe give a surprised pigeon a nasty bump, but the energy released by its fall will not be very great. Now take the same book up to the fifteenth floor of a tall building and drop it from there. For anyone below, astronomy could suddenly become a deadly subject; when the book hits, it does so with a great deal of energy.

Dropping things from far away into the much stronger gravity of a black hole is much more effective in turning the energy released by infall into other forms of energy. Just as the falling book can heat up the air, shake the ground, or produce sound energy that can be heard some distance away, so the energy of material falling toward a black hole can be converted to significant amounts of electromagnetic radiation.

What a black hole has to work with is not textbooks but streams of infalling gas. If a dense blob of gas moves through a thin gas at high speed, it heats up as it slows by friction. As it slows down, kinetic (motion) energy is turned into heat energy. Just like a spaceship reentering the atmosphere as shown in Figure 13.21, gas approaching a black hole heats up and glows where it meets other gas. But this gas, as it approaches the event horizon, reaches speeds of 10% the speed of light and more. It therefore gets far, far hotter than a spaceship, which reaches no more than about 1500 K. Indeed, gas near a supermassive black hole reaches a temperature of about 150,000 K, about 100 times hotter than a spaceship returning to Earth. It can even get so hot—millions of degrees—that it radiates X-rays.

Friction in Earth's Atmosphere

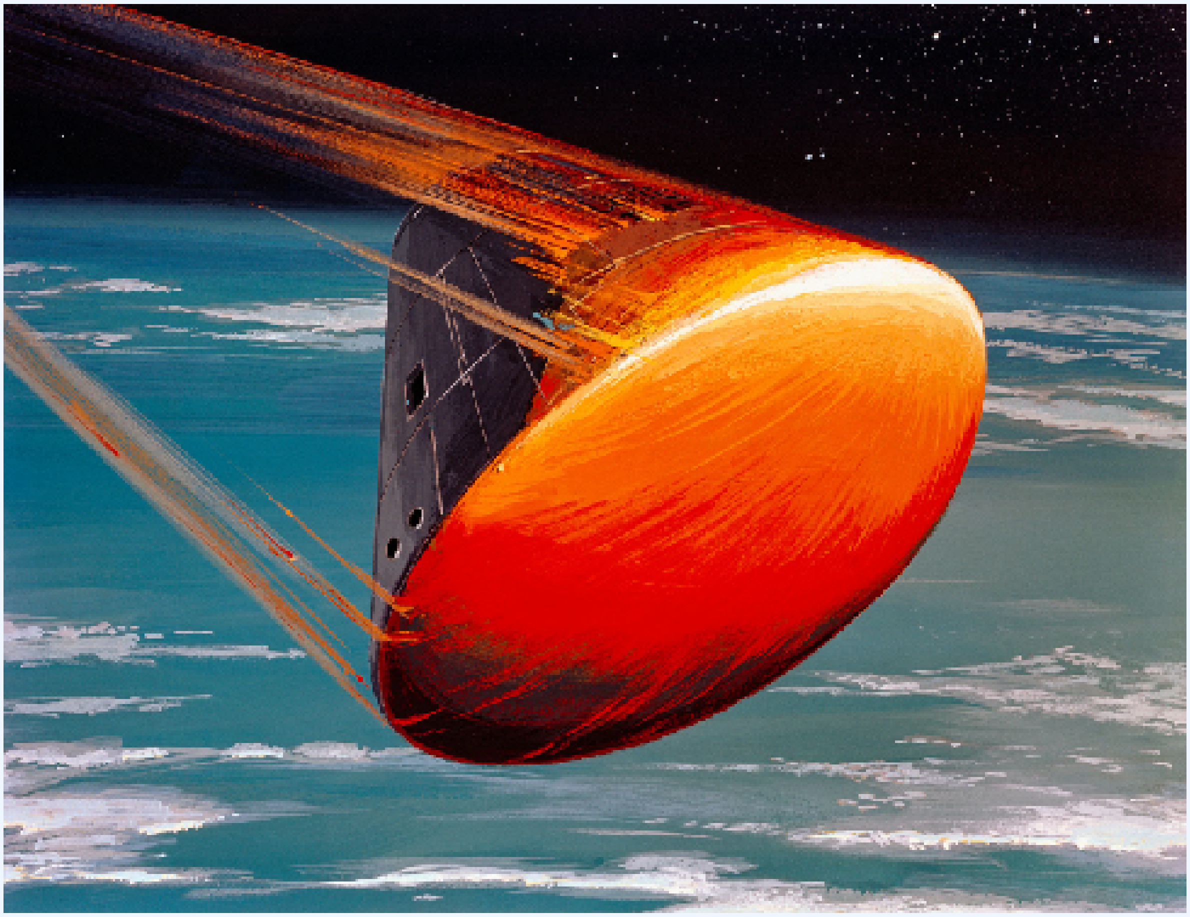


Figure 13.21. In this artist's impression, the rapid motion of a spacecraft (the Apollo mission reentry capsule) through the atmosphere compresses and heats the air ahead of it, which heats the spacecraft in turn until it glows red hot. Pushing on the air slows down the spacecraft, turning the kinetic energy of the spacecraft into heat. Fast-moving gas falling into a quasar heats up in a similar way.
Apollo cm by NASA, NASA Media License.

The amount of energy that can be liberated this way is enormous. Einstein showed that mass and energy are interchangeable with his famous formula $E = mc^2$. A hydrogen bomb releases just 1% of that energy, as does a star. Quasars are much more efficient than that. The energy released falling to the event horizon of a black hole can easily reach 10% or, in the extreme theoretical limit, 32%, of that energy. (Unlike the hydrogen atoms in a bomb or a star, the gas falling into the black hole is not actually losing mass from its atoms to free up the energy; the energy is produced just because the gas is falling closer and closer to the black hole.) This huge energy release explains how a tiny volume like the region around a black hole can release as much power as a whole galaxy. But to radiate all that energy, instead of just falling inside the event horizon with barely a peep, the hot gas must take the time to swirl around the star in the accretion disk and emit some of its energy.

Most black holes don't show any signs of quasar emission. We call them "quiescent." But, like sleeping dragons, they can be woken up by being roused with a fresh supply of gas. Our own Milky Way black hole is currently quiescent, but it may have been a quasar just a few million years ago as shown in Figure 13.22. Two giant bubbles that extend 25,000 light-years above and below the galactic centre are emitting gamma rays. Were these produced a few million years ago when a significant amount of matter fell into the black hole at the centre of the galaxy? Astronomers are still working to understand what remarkable event might have formed these enormous bubbles.

Fermi Bubbles in the Galaxy

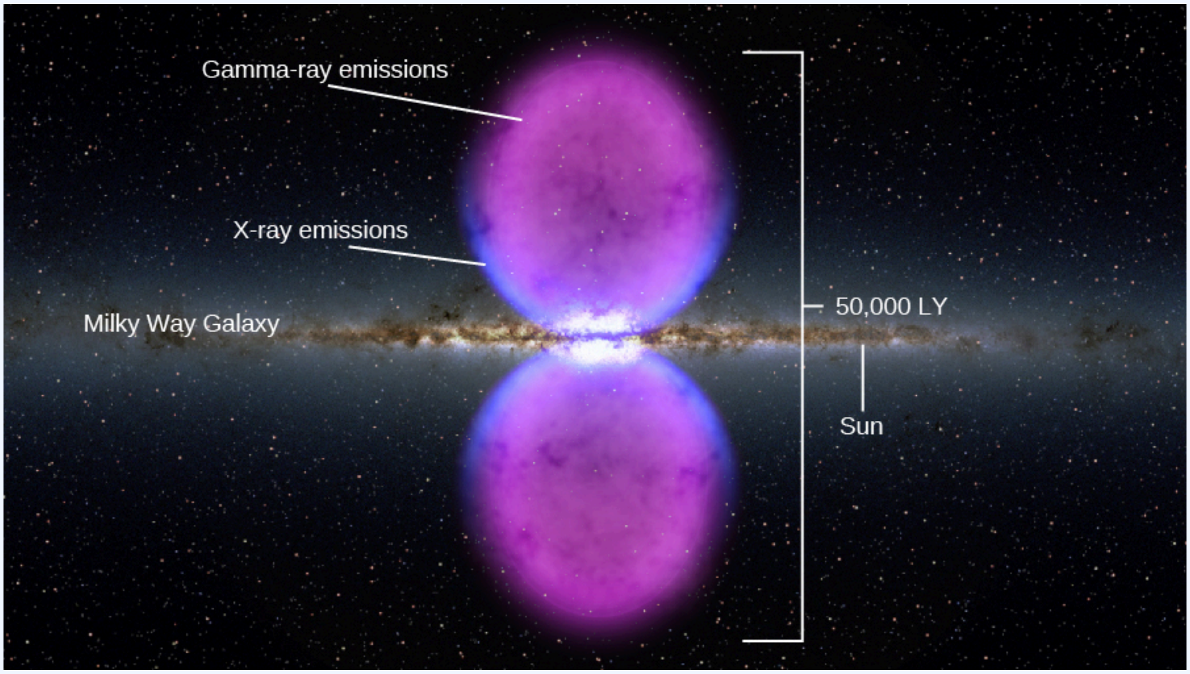


Figure 13.22. Giant bubbles shining in gamma-ray light lie above and below the centre of the Milky Way Galaxy, as seen by the Fermi satellite. (The gamma-ray and X-ray image is superimposed on a visible-light image of the inner parts of our Galaxy.) The bubbles may be evidence that the supermassive black hole at the centre of our Galaxy was a quasar a few million years ago.
Galactic Lobes by NASA's Goddard Space Flight Center, NASA Media License.

The physics required to account for the exact way in which the energy of infalling material is converted to radiation near a black hole is far more complicated than our simple discussion suggests. To understand what happens in the "rough and tumble" region around a massive black hole, astronomers and physicists must resort to computer simulations (and they require supercomputers, fast machines capable of awesome numbers of calculations per second). The details of these models are beyond the scope of our book, but they support the basic description presented here.

Radio Jets

So far, our model seems to explain the central energy source in quasars and active galaxies. But, as we have seen, there is more to quasars and other active galaxies than the point-like energy source. They can also have long jets that glow with radio waves, light, and sometimes even X-rays, and that extend far beyond the limits of the parent galaxy. Can we find a way for our black hole and its accretion disk to produce these jets of energetic particles as well?

Many different observations have now traced these jets to within 3 to 30 light-years of the parent quasar or galactic nucleus. While the black hole and accretion disk are typically smaller than 1 light-year, we nevertheless presume that if the jets come this close, they probably originate in the vicinity of the black hole. Another characteristic of the jets we need to explain is that they contain matter moving close to the speed of light.

Why are energetic electrons and other particles near a supermassive black hole ejected into jets, and often into two oppositely directed jets, rather than in all directions? Again, we must use theoretical models and supercomputer simulations of what happens when a lot of material whirls inward in a crowded black hole accretion disk. Although there is no agreement on exactly how jets form, it has become clear that any material escaping from the neighbourhood of the black hole has an easier time doing so *perpendicular to* the disk.

In some ways, the inner regions of black hole accretion disks resemble a baby that is just learning to eat by herself. As much food as goes into the baby's mouth can sometimes wind up being spit out in various directions. In the same way, some of the material whirling inward toward a black hole finds itself under tremendous pressure and orbiting with tremendous speed. Under such conditions, simulations show that a significant amount of material can be flung outward—not back along the disk, where more material is crowding in, but above and below the disk. If the disk is thick (as it tends to be when a lot of material falls in quickly), it can channel the outrushing material into narrow beams perpendicular to the disk as shown in Figure 13.23.

Models of Accretion Disks

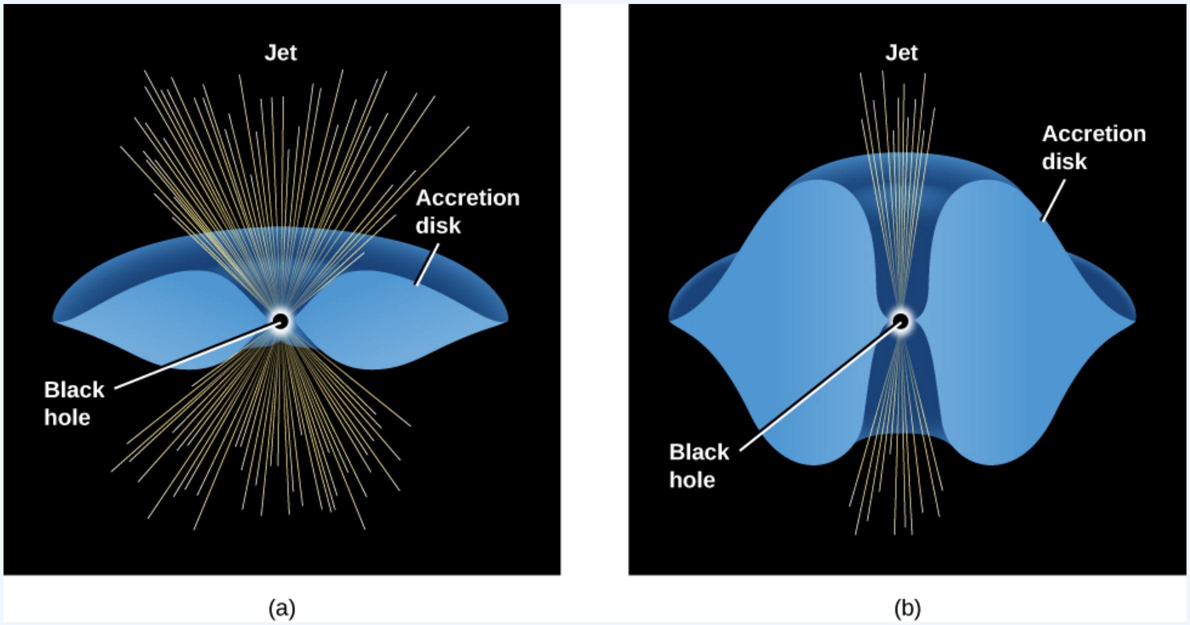


Figure 13.23. These schematic drawings show what accretion disks might look like around large black holes for (a) a thin accretion disk and (b) a “fat” disk—the type needed to account for channeling the outflow of hot material into narrow jets oriented perpendicular to the disk.

Figure 13.24 shows observations of an elliptical galaxy that behaves in exactly this way. At the centre of this active galaxy, there is a ring of dust and gas about 400 light-years in diameter, surrounding a 1.2-billion-M_{Sun} black hole. Radio observations show that two jets emerge in a direction perpendicular to the ring, just as the model predicts.

Jets and Disk in an Active Galaxy

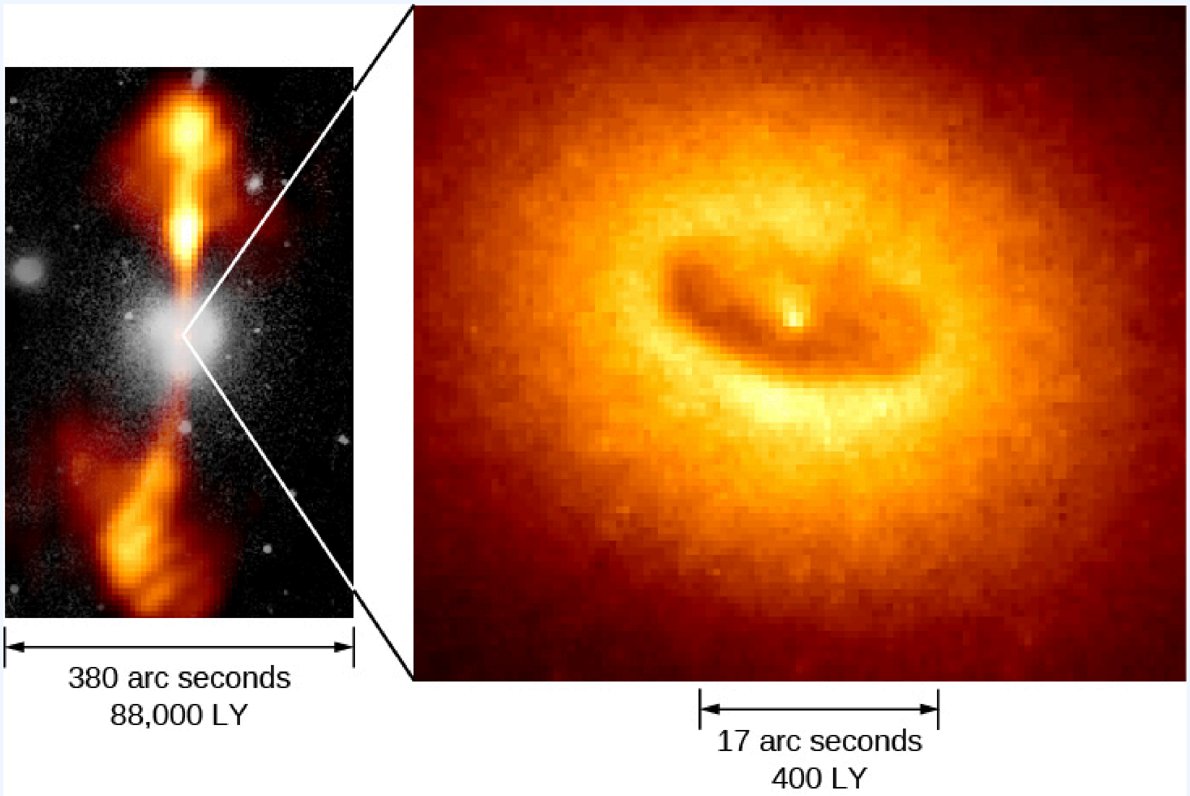


Figure 13.24. The picture on the left shows the active elliptical galaxy NGC 4261, which is located in the Virgo Cluster at a distance of about 100 million light-years. The galaxy itself—the white circular region in the centre—is shown the way it looks in visible light, while the jets are seen at radio wavelengths. A Hubble Space Telescope image of the central portion of the galaxy is shown on the right. It contains a ring of dust and gas about 800 light-years in diameter, surrounding a supermassive black hole. Note that the jets emerge from the galaxy in a direction perpendicular to the plane of the ring.
Core of Galaxy NGC 4261 by ESA/HST, ESA Standard License.

Quasars and the Attitudes of Astronomers

The discovery of quasars in the early 1960s was the first in a series of surprises astronomers had in store. Within another decade they would find neutron stars (in the form of pulsars), the first hints of black holes (in binary X-ray sources), and even the radio echo of the Big Bang itself. Many more new discoveries lay ahead.

As Maarten Schmidt reminisced in 1988, “This had, I believe, a profound impact on the conduct of those practicing astronomy. Before the 1960s, there was much authoritarianism in the field. New ideas expressed at meetings would be instantly judged by senior astronomers and rejected if too far out.” We saw a good example of this in the trouble Chandrasekhar had in finding acceptance for his ideas about the death of stars with cores greater than 1.4 M_{Sun} (see the feature box on Subrahmanyan Chandrasekhar in an earlier chapter).

“The discoveries of the 1960s,” Schmidt continued, “were an embarrassment, in the sense that they were totally unexpected and could not be evaluated immediately. In reaction to these developments, an attitude has evolved where even outlandish ideas in astronomy are taken seriously. Given our lack of solid knowledge in extragalactic astronomy, this is probably to be preferred over authoritarianism.”[1]

That is not to say that astronomers (being human) don’t continue to have prejudices and preferences. For example, a small group of astronomers who thought that the redshifts of quasars were not connected with their distances (which was definitely a minority opinion) often felt excluded from meetings or from access to telescopes in the 1960s and 1970s. It’s not so clear that they actually *were* excluded, as much as that they felt the very difficult pressure of knowing that most of their colleagues strongly disagreed with them. As it turned out, the evidence—which must ultimately decide all scientific questions—was not on their side either.

But today, as better instruments bring solutions to some problems and starkly illuminate our ignorance about others, the entire field of astronomy seems more open to discussing unusual ideas. Of course, before any hypotheses become accepted, they must be tested—again and again—against the evidence that nature itself reveals. Still, the many strange proposals published about what dark matter might be attest to the new openness that Schmidt described.

With this black hole model, we have come a long way toward understanding the quasars and active galaxies that seemed very mysterious only a few decades ago. As often happens in astronomy, a combination of better instruments (making better observations) and improved theoretical models enabled us to make significant progress on a puzzling aspect of the cosmos.

Attribution

13.7 EVOLUTION OF THE UNIVERSE IN ITS EARLY STAGES

The quasars' brilliance and large distance make them ideal probes of the far reaches of the universe and its remote past. Recall that when first introducing quasars, we mentioned that they generally tend to be far away. When we see extremely distant objects, we are seeing them as they were long ago. Radiation from a quasar 8 billion light-years away is telling us what that quasar and its environment were like 8 billion years ago, much closer to the time that the galaxy that surrounds it first formed. Astronomers have now detected light emitted from quasars that were already formed only a few hundred million years after the universe began its expansion 13.8 billion years ago. Thus, they give us a remarkable opportunity to learn about the time when large structures were first assembling in the cosmos.

The Evolution of Quasars

Quasars provide compelling evidence that we live in an evolving universe—one that changes with time. They tell us that astronomers living billions of years ago would have seen a universe that is very different from the universe of today. Counts of the number of quasars at different redshifts (and thus at different times in the evolution of the universe) show us how dramatic these changes are, as shown in Figure 13.25. We now know that the number of quasars was greatest at the time when the universe was only 20% of its present age.

Relative Number of Quasars and Rate at Which Stars Formed as a Function of the Age of the Universe

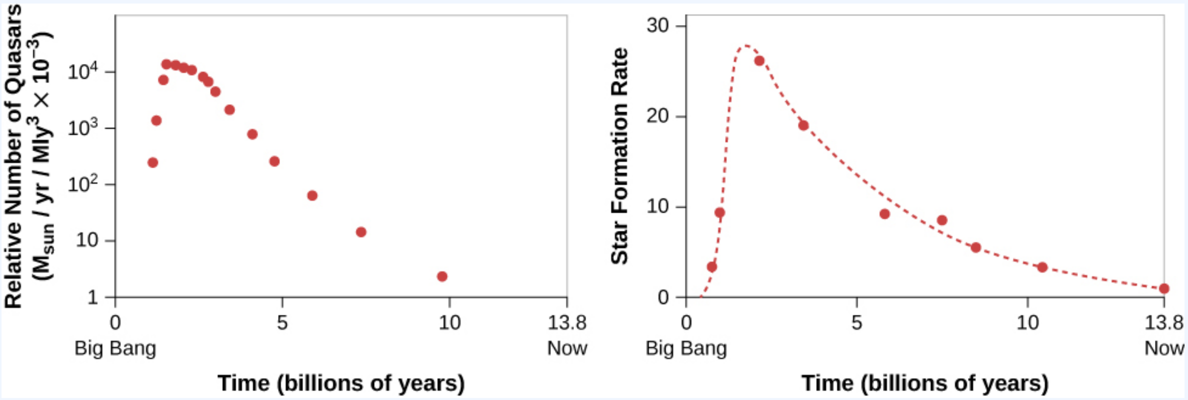


Figure 13.25. An age of 0 on the plots corresponds to the beginning of the universe; an age of 13.8 corresponds to the present time. Both the number of quasars and the rate of star formation were at a peak when the universe was about 20% as old as it is now.

As you can see, the drop-off in the numbers of quasars as time gets nearer to the present day is quite abrupt. Observations also show that the emission from the accretion disks around the most massive black holes peaks early and then fades. The most powerful quasars are seen only at early times. In order to explain this result, we make use of our model of the energy source of the quasars—namely that quasars are black holes with enough fuel to make a brilliant accretion disk right around them.

The fact that there were more quasars long ago (far away) than there are today (nearby) could be explained if there was more material available to be accreted by black holes early in the history of the universe. You might say that the quasars were more active when their black holes had fuel for their “energy-producing engines.” If that fuel was mostly consumed in the first few billion years after the universe began its expansion, then later in its life, a “hungry” black hole would have very little left with which to light up the galaxy’s central regions.

In other words, if matter in the accretion disk is continually being depleted by falling into the black hole or being blown out from the galaxy in the form of jets, then a quasar can continue to radiate only as long as new gas is available to replenish the accretion disk.

In fact, there *was* more gas around to be accreted early in the history of the universe. Back then, most gas had not yet collapsed to form stars, so there was more fuel available for both the feeding of black holes and the forming of new stars. Much of that fuel was subsequently consumed in the formation of stars during the first few billion years after the universe began its expansion. Later in its life, a galaxy would have little left to feed a hungry black hole or to form more new stars. As we see from Figure 13.25, both star formation and black hole growth peaked together when the universe was about 2 billion years old. Ever since, both have been in sharp decline. We are late to the party of the galaxies and have missed some of the early excitement.

Observations of nearer galaxies (seen later in time) indicate that there is another source of fuel for the central black holes—the collision of galaxies. If two galaxies of similar mass collide and merge, or if a smaller galaxy is pulled into a larger one, then gas and dust from one may come close enough to the black hole in the other to be devoured by it and so provide the necessary fuel. Astronomers have found that collisions were also much more common early in the history of the universe than they are today. There were more small galaxies in those early times because over time small galaxies tend to combine into larger ones. Again, this means that we would expect to see more quasars long ago (far away) than we do today (nearby)—as we in fact do.

Codependence of Black Holes and Galaxies

Once black hole masses began to be measured reliably in the late 1990s, they posed an enigma. It looked as though the mass of the central black hole depended on the mass of the galaxy. The black holes in galaxies always seem to be just 1/200 the mass of the galaxy they live in. This result is shown schematically in Figure 13.26, and some of the observations are plotted in Figure 13.27.

Relationship between Black Hole Mass and the Mass of the Host Galaxy

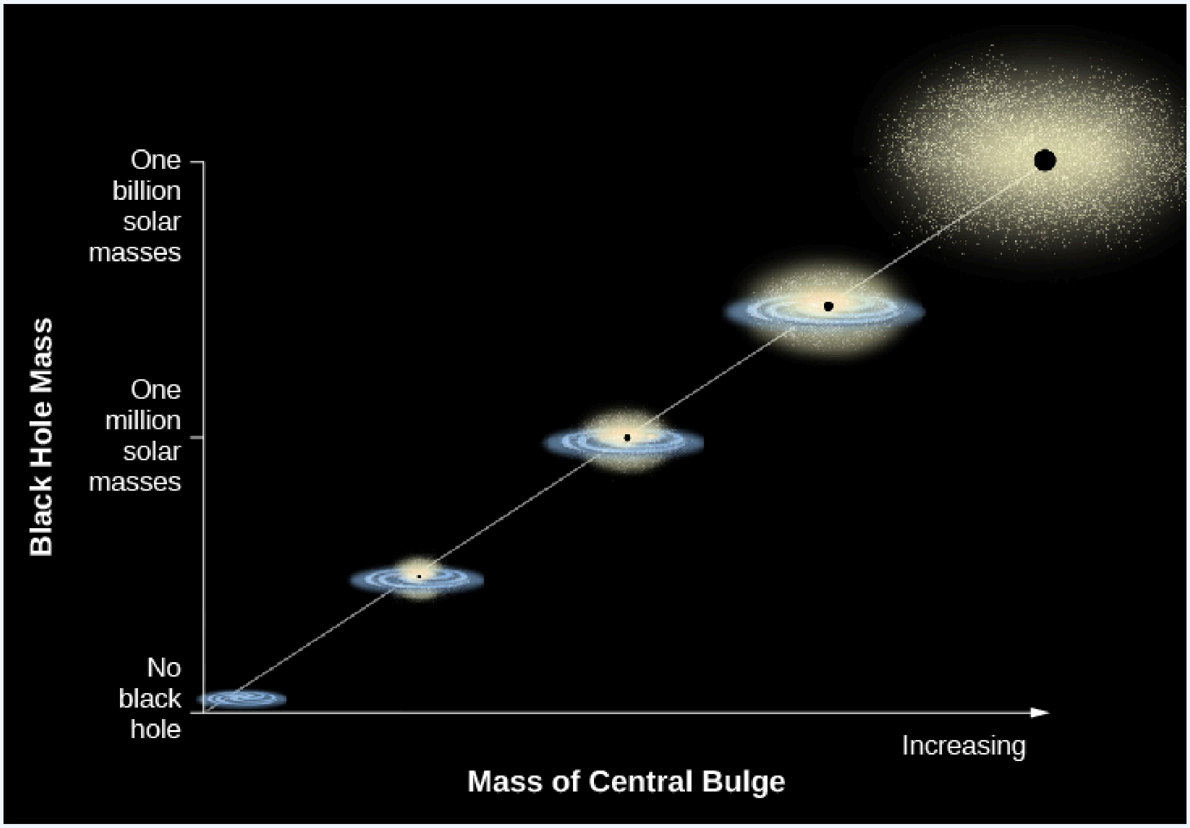


Figure 13.26. Observations show that there is a close correlation between the mass of the black hole at the centre of a galaxy and the mass of the spherical distribution of stars that surrounds the black hole. That spherical distribution may be in the form of either an elliptical galaxy or the central bulge of a spiral galaxy. Correlation of black hole Mass and bulge mass/brightness by K. Cordes, S. Brown (STScI), ESA Standard License.

Correlation between the Mass of the Central Black Hole and the Mass Contained within the Bulge of Stars Surrounding the Black Hole, Using Data from Real Galaxies

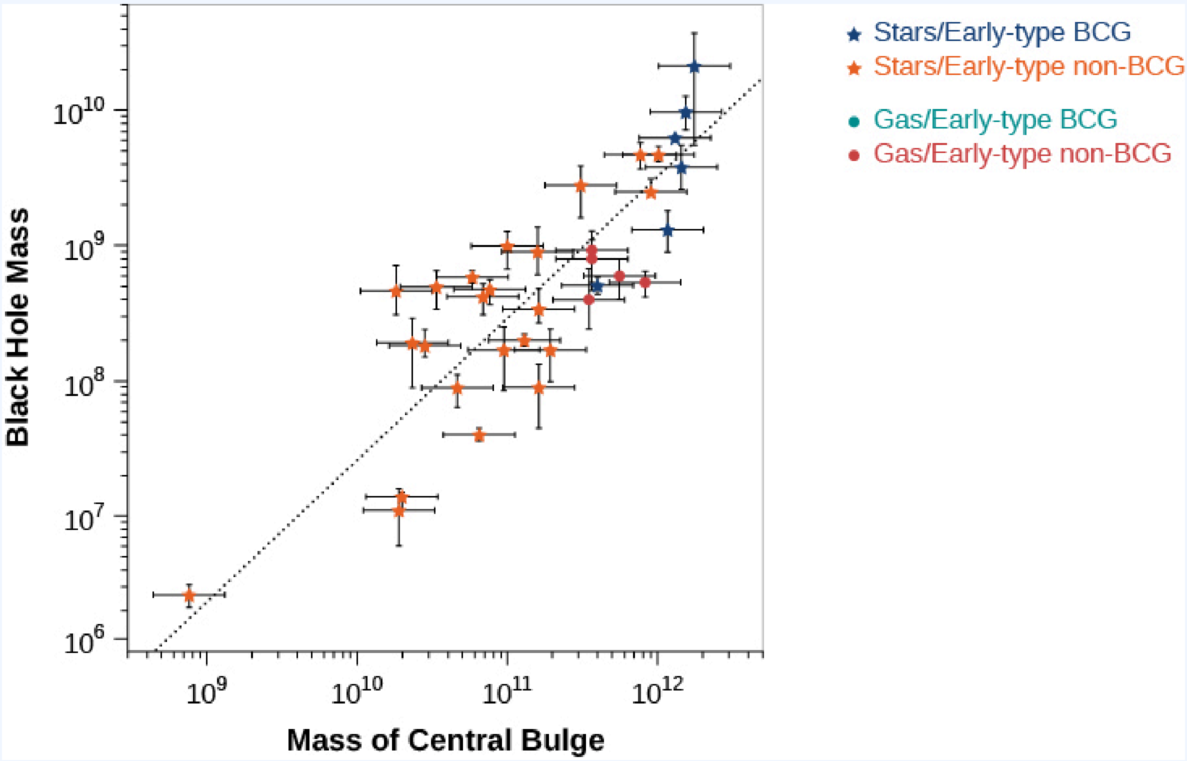


Figure 13.27. The black hole always turns out to be about 1/200 the mass of the stars surrounding it. The horizontal and vertical bars surrounding each point show the uncertainty of the measurement. Figure 1 from Nicholas J. McConnell, Chung-Pei Ma, “Revisiting the Scaling Relations of Black Hole Masses and Host Galaxy Properties,” The Astrophysical Journal, 764:184 (14 pp.), February 20, 2013, used under fair dealing.

Somehow black hole mass and the mass of the surrounding bulge of stars are connected. But why does this correlation exist? Unfortunately, astronomers do not yet know the answer to this question. We do know, however, that the black hole can influence the rate of star formation in the galaxy, and that the properties of the surrounding galaxy can influence how fast the black hole grows. Let’s see how these processes work.

How a Galaxy Can Influence a Black Hole in Its Center

Let's look first at how the surrounding galaxy might influence the growth and size of the black hole. Without large quantities of fresh "food," the surroundings of black holes glow only weakly as bits of local material spiral inward toward the black hole. So somehow large amounts of gas have to find their way to the black hole from the galaxy in order to feed the quasar and make it grow and give off the energy to be noticed. Where does this "food" for the black hole come from originally and how might it be replenished? The jury is still out, but the options are pretty clear.

One obvious source of fuel for the black hole is matter from the host galaxy itself. Galaxies start out with large amounts of interstellar gas and dust, and at least some of this interstellar matter is gradually converted into stars as the galaxy evolves. On the other hand, as stars go through their lives and die, they lose mass all the time into the space between them, thereby returning some of the gas and dust to the interstellar medium. We expect to find more gas and dust in the central regions early in a galaxy's life than later on, when much of it has been converted into stars. Any of the interstellar matter that ventures too close to the black hole may be accreted by it. This means that we would expect that the number and luminosity of quasars powered in this way would decline with time. And as we have seen, that is just what we find.

Today both *elliptical galaxies* and the *nuclear bulges of spiral galaxies* have very little raw material left to serve as a source of fuel for the black hole. And most of the giant black holes in nearby galaxies, including the one in our own Milky Way, are now dark and relatively quiet—mere shadows of their former selves. So that fits with our observations.

We should note that even if you have a quiescent supermassive black hole, a star in the area could occasionally get close to it. Then the powerful tidal forces of the black hole can pull the whole star apart into a stream of gas. This stream quickly forms an accretion disk that gives off energy in the normal way and makes the black hole region into a temporary quasar. However, the material will fall into the black hole after only a few weeks or months. The black hole then goes back into its lurking, quiescent state, until another victim wanders by.

This sort of "cannibal" event happens only once every 100,000 years or so in a typical galaxy. But we can monitor millions of galaxies in the sky, so a few of these "tidal disruption events" are found each year, shown in Figure 13.28 However, these individual events, dramatic as they are, are too rare to account for the huge masses of the central black holes.

A Black Hole Snacks on a Star

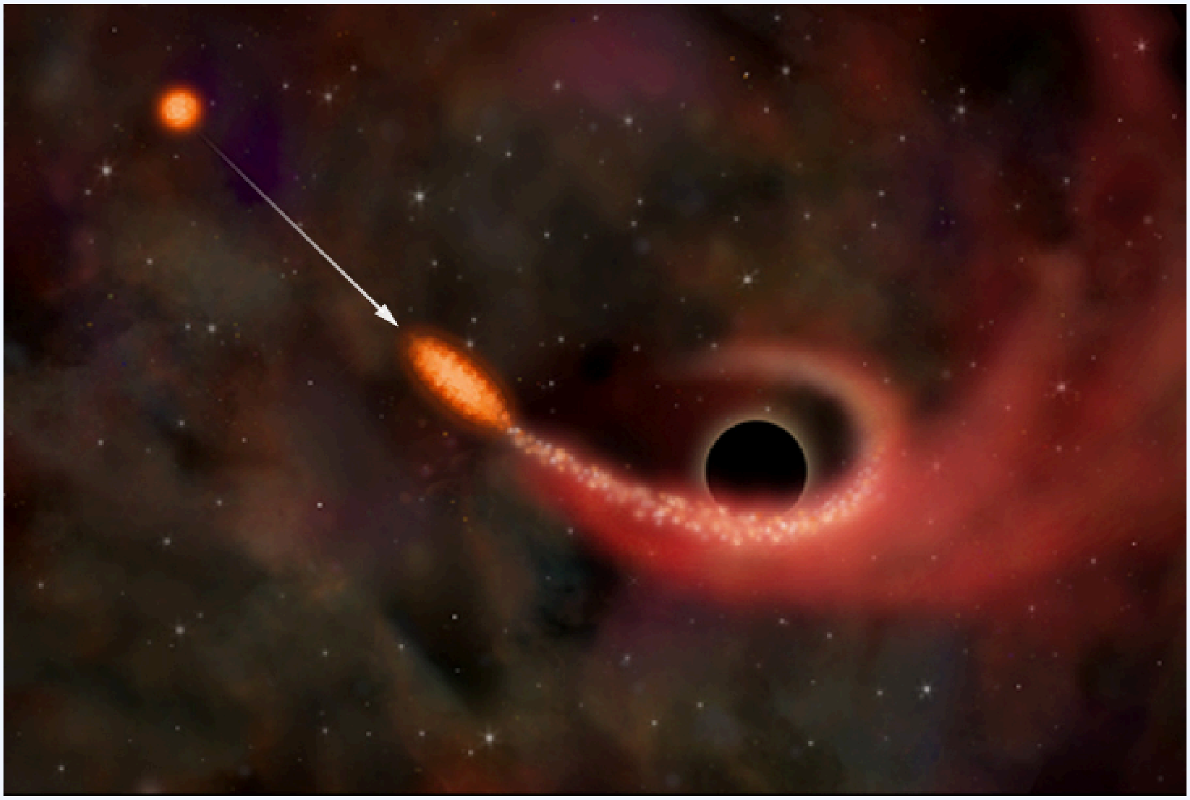


Figure 13.28. This artist's impression shows three stages of a star (red) swinging too close to a giant black hole (black circle). The star starts off (top left) in its normal spherical shape, then begins to be pulled into a long football shape by tides raised by the black hole (center). When the star gets closer still, the tides become stronger than the gravity holding the star together, and it breaks up into a streamer (right). Much of the star's matter forms a temporary accretion disk that lights up as a quasar for a few weeks or months.
Illustration Explaining Tidal Disruption by NASA/CXC/M. Weiss, NASA Media License.

Another source of fuel for the black hole is the collision of its host galaxy with another galaxy. Some of the brightest galaxies turn out, when a detailed picture is taken, to be pairs of colliding galaxies. And most of them have quasars inside them, not easily visible to us because they are buried by enormous amounts of dust and gas.

A collision between two cars creates quite a mess, pushing parts out of their regular place. In the same way, if two galaxies collide and merge, then gas and dust (though not so much the stars) can get pushed out of their regular orbits. Some may veer close enough to the black hole in one galaxy or the other to be devoured by it and so provide the necessary fuel to power a quasar. As we saw, galaxy collisions and mergers happened most frequently when the universe was young and probably help account for the fact that quasars were most common when the universe was only about 20% of its current age.

Collisions in today's universe are less frequent, but they do happen. Once a galaxy reaches the size of the Milky Way, most of the galaxies it merges with will be much smaller galaxies—*dwarf galaxies*. These don't disrupt the big galaxy much, but they can supply some additional gas to its black hole.

By the way, if two galaxies, each of which contains a black hole, collide, then the two black holes may merge and form an even larger black hole, shown in Figure 13.29. In this process they will emit a burst of gravitational waves. One of the main goals of the European Space Agency's planned LISA (Laser Interferometer Space Antenna) mission is to detect the gravitational wave signals from the merging of supermassive black holes.

Colliding Galaxies with Two Black Holes

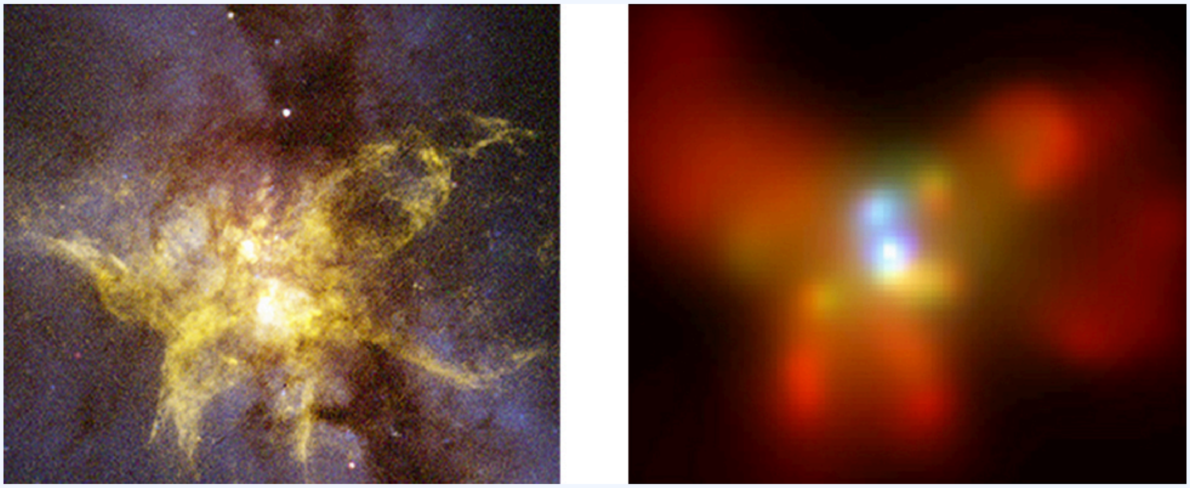


Figure 13.29. We compare Hubble Space Telescope visible-light (left) and Chandra X-ray (right) images of the central regions of NGC 6240, a galaxy about 400 million light-years away. It is a prime example of a galaxy in which stars are forming, evolving, and exploding at an exceptionally rapid rate due to a relatively recent merger (30 million years ago). The Chandra image shows two bright X-ray sources, each produced by hot gas surrounding a black hole. Over the course of the next few hundred million years, the two supermassive black holes, which are about 3000 light-years apart, will drift toward each other and merge to form an even larger black hole. This detection of a binary black hole supports the idea that black holes can grow to enormous masses in the centers of galaxies by merging with nearby galaxies. Optical & X-ray Comparison of NGC 6240 by NASA/CXC/MPE/S.Komossa et al; credit right: NASA/STScI/R. P. van der Marel, J. Gerssen, NASA Media License.

How Does the Black Hole Influence the Formation of Stars in the Galaxy?

We have seen that the material in galaxies can influence the growth of the black hole. The black hole in turn can also influence the galaxy in which it resides. It can do so in three ways: through its jets, through winds of particles that manage to stream away from the accretion disk, and through radiation from the accretion disk. As they stream away from the black hole, all three can either promote star formation by compressing the surrounding gas and dust—or instead suppress star formation by heating the surrounding gas and shredding molecular clouds, thereby inhibiting or preventing star formation. The outflowing energy can even be enough to halt the accretion of new material and starve the black hole of fuel. Astronomers are still trying to evaluate the relative importance of these effects in determining the overall evolution of galactic bulges and the rates of star formation.

In summary, we have seen how galaxies and supermassive black holes can each influence the evolution of the other: the galaxy supplies fuel to the black hole, and the quasar can either support or suppress star formation. The balance of these processes probably helps account for the correlation between black hole and bulge masses, but there are as yet no theories that explain quantitatively and in detail why the correlation between black hole and bulge masses is as tight as it is or why the black hole mass is always about 1/200 times the mass of the bulge.

The Birth of Black Holes and Galaxies

While the connection between quasars and galaxies is increasingly clear, the biggest puzzle of all—namely, how the supermassive black holes in galaxies got started—remains unsolved. Observations show that they existed when the universe was very young. One dramatic example is the discovery of a quasar that was already shining when the universe was only 700 million years old. What does it take to create a large black hole so quickly? A related problem is that in order to eventually build black holes containing more than 2 billion solar masses, it is necessary to have giant “seed” black holes with masses at least 2000 times the mass of the Sun—and they must somehow have been created shortly after the expansion of the universe began.

Astronomers are now working actively to develop models for how these seed black holes might have formed. Theories suggest that galaxies formed from collapsing clouds of dark matter and gas. Some of the gas formed stars, but perhaps some of the gas settled to the center where it became so concentrated that it formed a black hole. If this happened, the black hole could form right away—although this requires that the gas should not be rotating very much initially.

A more likely scenario is that the gas will have some angular momentum (rotation) that will prevent direct collapse to a black hole. In that case, the very first generation of stars will form, and some of them, according to calculations, will have masses hundreds of times that of the Sun. When these stars finish burning hydrogen, just a few million years later, the supernovae they end with will create black holes a hundred or so times the mass of the Sun. These can then merge with others or accrete the rich gas supply available at these early times.

The challenge is growing these smaller black holes quickly enough to make the much larger black holes we see a few hundred million years later. It turns out to be difficult because there are limits on how fast they can accrete matter. These should make sense to you from what we discussed earlier in the chapter. If the rate of accretion becomes too high, then the energy streaming outward from the black hole's accretion disk will become so strong as to blow away the infalling matter.

What if, instead, a collapsing gas cloud doesn't form a black hole directly or break up and form a group of regular stars, but stays together and makes one fairly massive star embedded within a dense cluster of thousands of lower mass stars and large quantities of dense gas? The massive star will have a short lifetime and will soon collapse to become a black hole. It can then begin to attract the dense gas surrounding it. But calculations show that the gravitational attraction of the many nearby stars will cause the black hole to zigzag randomly within the cluster and will prevent the formation of an accretion disk. If there is no accretion disk, then matter can fall freely into the black hole from all directions. Calculations suggest that under these conditions, a black hole even as small as 10 times the mass of the Sun could grow to more than 10 billion times the mass of the Sun by the time the universe is a billion years old.

Scientists are exploring other ideas for how to form the seeds of supermassive black holes, and this remains a very active field of research. Whatever mechanism caused the rapid formation of these supermassive black holes, they do give us a way to observe the youthful universe when it was only about five percent as old as it is now.

Attribution

"27.3 Quasars as Probes of Evolution in the Universe" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

13.8 KEY TERMS

Active galactic nuclei (AGN): galaxies that are almost as luminous as quasars and share many of their properties, although to a less spectacular degree; abnormal amounts of energy are produced in their centres. 13.5

Galactic cannibalism: a process by which a larger galaxy strips material from or completely swallows a smaller one. 13.3

Galaxy evolution: changes in individual galaxies over cosmic time, inferred by observing snapshots of many different galaxies at different times in their lives. 13.2

Merger: a collision between galaxies (of roughly comparable size) that combine to form a single new structure. 13.3

Quasar: an object of very high redshift that looks like a star but is extragalactic and highly luminous; also called a quasi-stellar object, or QSO. 13.5

Redshift: how much the lines in a galaxy's spectrum are shifted to the red because of the expansion of the universe. 13.2

Starburst: a galaxy or merger of multiple galaxies that turns gas into stars much faster than usual. 13.3

Supermassive black hole: the object in the centre of most large galaxies that is so massive and compact that light cannot escape from it; the Milky Way's supermassive black hole contains 4.6 millions of Suns' worth of mass. 13.4

CHAPTER 14: THE STRUCTURE OF THE UNIVERSE

Chapter Overview

14.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Explain the fundamental question in astronomy and cosmology, “How did galaxies and large-scale structures in the universe form and evolve?”
- Describe Edwin Hubble’s two crucial discoveries from his survey of galaxies and their implications for understanding the large-scale structure of the universe.
- Differentiate and compare the distribution of galaxies in various regions of the universe, including clusters, voids, and filamentary superclusters, and interpret the implications of these structures on the large-scale structure of the cosmos.
- Summarize the evidence supporting the existence of dark matter in galaxies and galaxy clusters.
- Evaluate the challenges and limitations of mapping the universe.

14.1 INTRODUCTION

In the preceding section, we emphasized the role of mergers in shaping the evolution of galaxies. In order to collide, galaxies must be fairly close together. To estimate how often collisions occur and how they affect galaxy evolution, astronomers need to know how galaxies are distributed in space and over cosmic time. Are most of them isolated from one another or do they congregate in groups? If they congregate, how large are the groups and how and when did they form? And how, in general, are galaxies and their groups arranged in the cosmos? Are there as many in one direction of the sky as in any other, for example? How did galaxies get to be arranged the way we find them today?

Edwin Hubble found answers to some of these questions only a few years after he first showed that the spiral nebulae were galaxies and not part of our Milky Way. As he examined galaxies all over the sky, Hubble made two discoveries that turned out to be crucial for studies of the evolution of the universe.

Attribution

"28.3 The Distribution of Galaxies in Space" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

14.2 THE COSMOLOGICAL PRINCIPLE

Hubble made his observations with what were then the world's largest telescopes—the 100-inch and 60-inch reflectors on Mount Wilson. These telescopes have small fields of view: they can see only a small part of the heavens at a time. To photograph the entire sky with the 100-inch telescope, for example, would have taken longer than a human lifetime. So instead, Hubble sampled the sky in many regions, much as Herschel did with his star gauging. In the 1930s, Hubble photographed 1283 sample areas, and on each print, he carefully counted the numbers of galaxy images. He is pictured in Figure 14.1.

The first discovery Hubble made from his survey was that the number of galaxies visible in each area of the sky is about the same. (Strictly speaking, this is true only if the light from distant galaxies is not absorbed by dust in our own Galaxy, but Hubble made corrections for this absorption.) He also found that the numbers of galaxies increase with faintness, as we would expect if the density of galaxies is about the same at all distances from us.

To understand what we mean, imagine you are taking snapshots in a crowded stadium during a sold-out concert. The people sitting near you look big, so only a few of them will fit into a photo. But if you focus on the people sitting in seats way on the other side of the stadium, they look so small that many more will fit into your picture. If all parts of the stadium have the same seat arrangements, then as you look farther and farther away, your photo will get more and more crowded with people. In the same way, as Hubble looked at fainter and fainter galaxies, he saw more and more of them.

Hubble at Work

Figure 14.1. Edwin Hubble at the 100-inch telescope on Mount Wilson.
Edwin Hubble by NASA/STScI, NASA Media License.

Hubble's findings are enormously important, for they indicate that the universe is both **isotropic** and **homogeneous**—it looks the same in all directions, and a large volume of space at any given redshift or distance is much like any other volume at that redshift. If that is so, it does not matter what section of the universe we observe (as long as it is a sizeable portion): any section will look the same as any other.

Hubble's results—and many more that have followed in the nearly 100 years since then—imply not only that the universe is about the same everywhere (apart from changes with time) but also that aside from small-scale local differences, the part we can see around us is representative of the whole. The idea that the universe is the same everywhere is called the **cosmological principle** and is the starting assumption for nearly all theories that describe the entire universe.

Without the cosmological principle, we could make no progress at all in studying the universe. Suppose our own local neighbourhood were unusual in some way. Then we could no more understand what the universe is like than if we were marooned on a warm south-sea island without outside communication and were trying to understand the geography of Earth. From our limited island vantage point, we could not know that some parts of the planet are covered with snow and ice, or that large continents exist with a much greater variety of terrain than that found on our island.

Hubble merely counted the numbers of galaxies in various directions without knowing how far away most of them were. With modern instruments, astronomers have measured the velocities and distances of hundreds of thousands of galaxies, and so built up a meaningful picture of the large-scale structure of the universe. In the rest of this section, we describe what we know about the distribution of galaxies, beginning with those that are nearby.

Attribution

"28.3 The Distribution of Galaxies in Space" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

14.3 SUPERCLUSTERS AND VOIDS

After astronomers discovered clusters of galaxies, they naturally wondered whether there were still larger structures in the universe. Do clusters of galaxies gather together? To answer this question, we must be able to map large parts of the universe in three dimensions. We must know not only the position of each galaxy on the sky (that's two dimensions) but also its distance from us (the third dimension).

This means we must be able to measure the redshift of each galaxy in our map. Taking a spectrum of each individual galaxy to do this is a much more time-consuming task than simply counting galaxies seen in different directions on the sky, as Hubble did. Today, astronomers have found ways to get the spectra of many galaxies in the same field of view (sometimes hundreds or even thousands at a time) to cut down the time it takes to finish their three-dimensional maps. Larger telescopes are also able to measure the redshifts—and therefore the distances—of much more distant galaxies and (again) to do so much more quickly than previously possible.

Another challenge astronomers faced in deciding how to go about constructing a map of the universe is similar to that confronted by the first team of explorers in a huge, uncharted territory on Earth. Since there is only one band of explorers and an enormous amount of land, they have to make choices about where to go first. One strategy might be to strike out in a straight line in order to get a sense of the terrain. They might, for example, cross some mostly empty prairies and then hit a dense forest. As they make their way through the forest, they learn how thick it is in the direction they are travelling, but not its width to their left or right. Then a river crosses their path; as they wade across, they can measure its width but learn nothing about its length. Still, as they go on in their straight line, they begin to get some sense of what the landscape is like and can make at least part of a map. Other explorers, striking out in other directions, will someday help fill in the remaining parts of that map.

Astronomers have traditionally had to make the same sort of choices. We cannot explore the universe in every direction to infinite "depth" or sensitivity: there are far too many galaxies and far too few telescopes to do the job. But we can pick a single direction or a small slice of the sky and start mapping the galaxies. Margaret Geller, the late John Huchra, and their students at the Harvard-Smithsonian Center for Astrophysics pioneered this technique, and several other groups have extended their work to cover larger volumes of space.

Margaret Geller: Cosmic Surveyor

Born in 1947, Margaret Geller is the daughter of a chemist who encouraged her interest in science and helped her visualize the three-dimensional structure of molecules as a child. (It was a skill that would later come in very handy for visualizing the three-dimensional structure of the universe.) She remembers being bored in elementary school, but she was encouraged to read on her own by her parents. Her recollections also include subtle messages from teachers that mathematics (her strong early interest) was not a field for girls, but she did not allow herself to be deterred.

Geller obtained a BA in physics from the University of California at Berkeley and became the second woman to receive a PhD in physics from Princeton. There, while working with James Peebles, one of the world's leading cosmologists, she became interested in problems relating to the large-scale structure of the universe. In 1980, she accepted a research position at the Harvard-Smithsonian Center for Astrophysics, one of the nation's most dynamic institutions for astronomy research. She saw that to make progress in understanding how galaxies and clusters are organized, a far more intensive series of surveys was required. Although it would not bear fruit for many years, Geller and her collaborators began the long, arduous task of mapping the galaxies, she is pictured in Figure 14.2.

Margaret Geller

Figure 14.2. Geller's work mapping and researching galaxies has helped us to better understand the structure of the universe.
Margaret J. Geller by Massimo Ramella, Smithsonian Institution Open Access (CC0).

Her team was fortunate to be given access to a telescope that could be dedicated to their project, the 60-inch reflector on Mount Hopkins, near Tucson, Arizona, where they and their assistants took spectra to determine galaxy distances. To get a slice of the universe, they pointed their telescope at a predetermined position in the sky and then let the rotation of Earth bring new galaxies into their field of view. In this way, they measured the positions and redshifts of over 18,000 galaxies and made a wide range of interesting maps to display their data. Their surveys now include "slices" in both the Northern and Southern Hemispheres.

As news of her important work spread beyond the community of astronomers, Geller received a MacArthur Foundation Fellowship in 1990. These fellowships, popularly called "genius awards," are designed to recognize truly creative work in a wide range of fields. Geller continues to have a strong interest in visualization and has (with filmmaker Boyd Estus) made several award-winning videos explaining her work to nonscientists (one is titled *So Many Galaxies . . . So Little Time*). She has appeared on a variety of national news and documentary programs, including the *MacNeil/Lehrer NewsHour*, *The Astronomers*, and *The Infinite Voyage*. Energetic and outspoken, she has given talks on her work to many audiences around the country, and works hard to find ways to explain the significance of her pioneering surveys to the public.

"It's exciting to discover something that nobody's seen before. [To be] one of the first three people to ever see that slice of the universe [was] sort of being like Columbus. . . . Nobody expected such a striking pattern!"—Margaret Geller

The largest universe mapping project to date is the Sloan Digital Sky Survey (see the feature box Astronomy and Technology: The Sloan Digital Sky Survey at the end of this section). A plot of the distribution of galaxies mapped by the Sloan survey is shown in Figure 14.3. To the surprise of astronomers, maps like the one in the figure showed that clusters of galaxies are not arranged uniformly throughout the universe, but are found in huge filamentary **superclusters** that look like great arcs of inkblots splattered across a page. The superclusters resemble an irregularly torn sheet of paper or a pancake in shape—they can extend for hundreds of millions of light-years in two dimensions, but are only 10 to 20 million light-years thick in the third dimension. Detailed study of some of these structures shows that their masses are a few times $10^{16} M_{\text{Sun}}$, which is 10,000 times more massive than the Milky Way Galaxy.

Sloan Digital Sky Survey Map of the Large-Scale Structure of the Universe

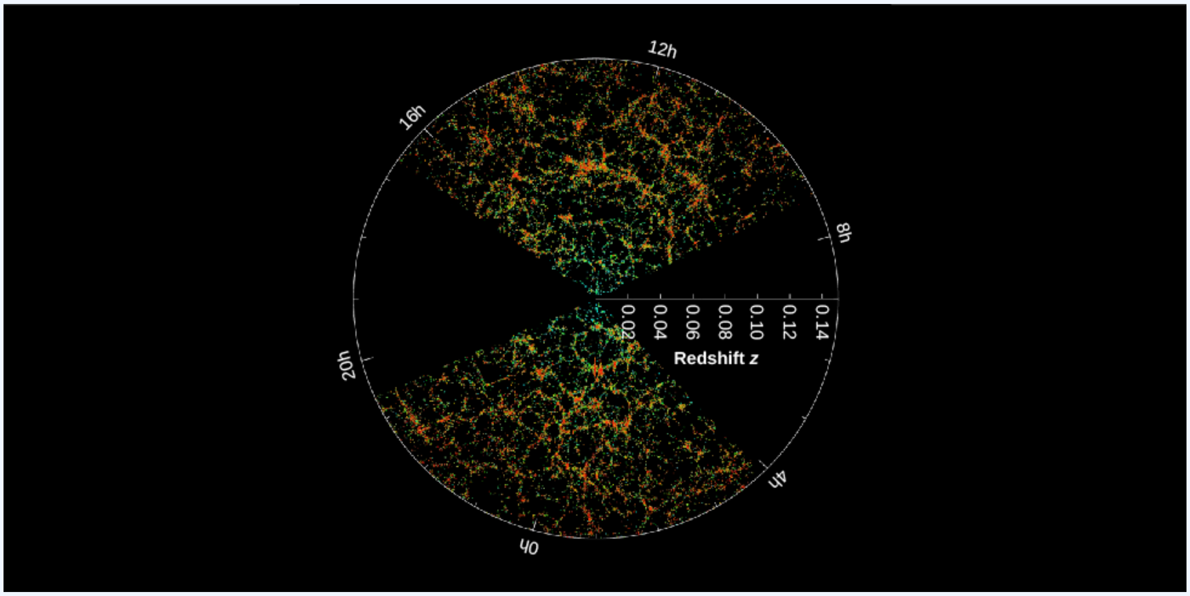


Figure 14.3. This image shows slices from the SDSS map. The point at the centre corresponds to the Milky Way and might say "You Are Here!" Points on the map moving outward from the centre are farther away. The distance to the galaxies is indicated by their redshifts (following Hubble's law), shown on the horizontal line going right from the centre. The redshift z = Δλ/λ, where Δλ is the difference between the observed wavelength and the wavelength λ emitted by a nonmoving source in the laboratory. Hour angle on the sky is shown around the circumference of the circular graph. The colours of the galaxies indicate the ages of their stars, with the redder colour showing galaxies that are made of older stars. The outer circle is at a distance of two billion light-years from us. Note that red (older stars) galaxies are more strongly clustered than blue galaxies (young stars). The unmapped areas are where our view of the universe is obstructed by dust in our own Galaxy.
A slice through the SDSS 3-dimensional map of the distribution of galaxies by M. Blanton and the Sloan Digital Sky Survey, NASA Media License.

Separating the filaments and sheets in a supercluster are **voids**, which look like huge empty bubbles walled in by the great arcs of galaxies. They have typical diametres of 150 million light-years, with the clusters of galaxies concentrated along their walls. The whole arrangement of filaments and voids reminds us of a sponge, the inside of a honeycomb, or a hunk of Swiss cheese with very large holes. If you take a good slice or cross-section through any of these, you will see something that looks roughly like Figure 14.3.

Before these voids were discovered, most astronomers would probably have predicted that the regions between giant clusters of galaxies were filled with many small groups of galaxies, or even with isolated individual galaxies. Careful searches within these voids have found few galaxies of any kind. Apparently, 90 percent of the galaxies occupy less than 10 percent of the volume of space.

Example 14.1

Galaxy Distribution

To determine the distribution of galaxies in three-dimensional space, astronomers have to measure their positions and their redshifts. The larger the volume of space surveyed, the more likely the measurement is a fair sample of the universe as a whole. However, the work involved increases very rapidly as you increase the volume covered by the survey.

Let's do a quick calculation to see why this is so.

Suppose that you have completed a survey of all the galaxies within 30 million light-years and you now want to survey to 60 million light-years. What volume of space is covered by your second survey? How much larger is this volume than the volume of your first survey? Remember that the volume of a sphere, V, is given by the formula $V = {}^{4}/_{3}\,\pi R^3$, where R is the radius of the sphere.

Solution

Since the volume of a sphere depends on R^3 and the second survey reaches twice as far in distance, it will cover a volume that is $2^3 = 8$ times larger. The total volume covered by the second survey will be $(4/3)\pi \times (60 \text{ million light-years})^3 = 9 \times 10^{23}$ light-years^3.

Exercise 14.1

Suppose you now want to expand your survey to 90 million light-years. What volume of space is covered, and how much larger is it than the volume of the second survey?

Solution

The total volume covered is (4/3)π × (90 million light-years)^3 = 3.05 × 10^{24} light-years^3. The survey reaches 3 times as far in distance, so it will cover a volume that is 3^3 = 27 times larger.

Even larger, more sensitive telescopes and surveys are currently being designed and built to peer farther and farther out in space and back in time. The new 50-meter Large Millimeter Telescope in Mexico and the Atacama Large Millimeter Array in Chile can detect far-infrared and millimeter-wave radiation from massive starbursting galaxies at redshifts and thus distances more than 90% of the way back to the Big Bang. These cannot be observed with visible light because their star formation regions are wrapped in clouds of thick dust. And in 2018, the 6.5-meter-diameter James Webb Space Telescope is scheduled to launch. It will be the first new major visible light and near-infrared telescope in space since Hubble was launched more than 25 years earlier. One of the major goals of this telescope is to observe directly the light of the first galaxies and even the first stars to shine, less than half a billion years after the Big Bang.

At this point, you may be wondering what exactly in the Sloan Sky Map above is expanding. We know that the galaxies and clusters of galaxies are held together by their gravity and do not expand as the universe does. However, the voids do grow larger and the filaments move farther apart as space stretches.

Astronomy and Technology: The Sloan Digital Sky Survey

In Edwin Hubble's day, spectra of galaxies had to be taken one at a time. The faint light of a distant galaxy gathered by a large telescope was put through a slit, and then a spectrometer (also called a spectrograph) was used to separate the colours and record the spectrum. This was a laborious process, ill suited to the demands of making large-scale maps that require the redshifts of many thousands of galaxies.

But new technology has come to the rescue of astronomers who seek three-dimensional maps of the universe of galaxies. One ambitious survey of the sky was produced using a special telescope, camera, and spectrograph atop the Sacramento Mountains of New Mexico. Called the Sloan Digital Sky Survey (SDSS), after the foundation that provided a large part of the funding, the program used a 2.5-meter telescope (about the same aperture as the Hubble) as a wide-angle astronomical camera. During a mapping program lasting more than ten years, astronomers used the SDSS's 30 charge-coupled devices (CCDs)—sensitive electronic light detectors similar to those used in many digital cameras and cell phones—to take images of over 500 million objects and spectra of over 3 million, covering more than one-quarter of the celestial sphere. Like many large projects in modern science, the Sloan Survey involved scientists and engineers from many different institutions, ranging from universities to national laboratories.

Every clear night for more than a decade, astronomers used the instrument to make images recording the position and brightness of celestial objects in long strips of the sky. The information in each strip was digitally recorded and preserved for future generations. When the seeing was only adequate, the telescope was used for taking spectra of galaxies and quasars—but it did so for up to *640 objects at a time*.

The key to the success of the project was a series of **optical fibers**, thin tubes of flexible glass that can transmit light from a source to the CCD that then records the spectrum. After taking images of a part of the sky and identifying which objects are galaxies, project scientists drilled an aluminum plate with holes for attaching fibers at the location of each galaxy. The telescope was then pointed at the right section of the sky, and the fibers led the light of each galaxy to the spectrometer for individual recording, shown in Figure 14.4.

Sloan Digital Sky Survey

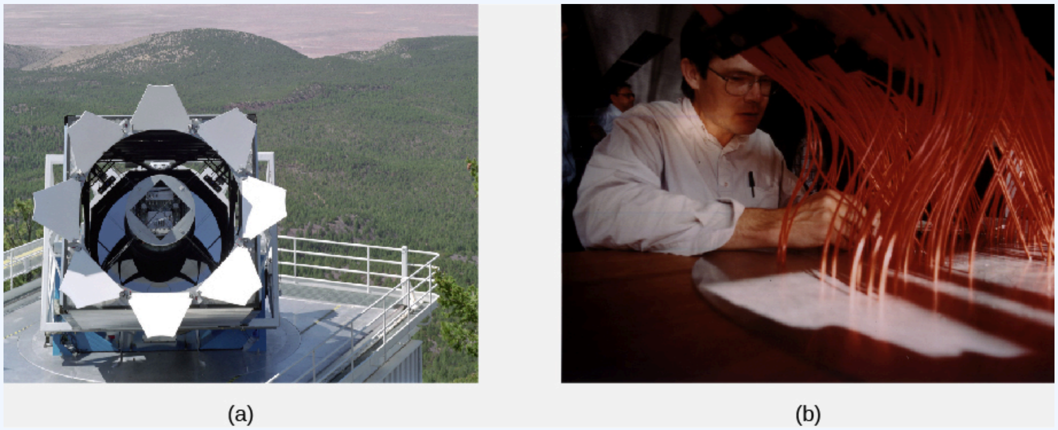


Figure 14.4 (a) The Sloan Digital Sky Survey telescope is seen here in front of the Sacramento Mountains in New Mexico. (b) Astronomer Richard Kron inserts some of the optical fibers into the pre-drilled plate to enable the instruments to make many spectra of galaxies at the same time.
Credit a: Telescope and mountains by Sloan Digital Sky Survey, SDSS Image Usage Policy.
Credit b: Plug Plate Array by Sloan Digital Sky Survey, SDSS Image Usage Policy.

About an hour was sufficient for each set of spectra, and the pre-drilled aluminum plates could be switched quickly. Thus, it was possible to take as many as 5000 spectra in one night (provided the weather was good enough).

The galaxy survey led to a more comprehensive map of the sky than has ever before been possible, allowing astronomers to test their ideas about large-scale structure and the evolution of galaxies against an impressive array of real data.

The information recorded by the Sloan Survey staggers the imagination. The data came in at 8 megabytes per second (this means 8 million individual numbers or characters every second). Over the course of the project, scientists recorded over 15 terabytes, or 15 thousand billion bytes, which they estimate is comparable to the information contained in the Library of Congress. Organizing and sorting this volume of data and extracting the useful scientific results it contains is a formidable challenge, even in our information age. Like many other fields, astronomy has now entered an era of "Big Data," requiring supercomputers and advanced computer algorithms to sift through all those terabytes of data efficiently.

One very successful solution to the challenge of dealing with such large datasets is to turn to "citizen science," or crowd-sourcing, an approach the SDSS helped pioneer. The human eye is very good at recognizing subtle differences among shapes, such as between two different spiral galaxies, while computers often fail at such tasks. When Sloan project astronomers wanted to catalog the shapes of some of the millions of galaxies in their new images, they launched the "Galaxy Zoo" project: volunteers around the world were given a short training course online, then were provided with a few dozen galaxy images to classify by eye. The project was wildly successful, resulting in over 40 million galaxy classifications by more than 100,000 volunteers and the discovery of whole new types of galaxies.

Learn more about how you can be part of the project of classifying galaxies in this citizen science effort. This program is part of a whole series of "citizen science" projects that enable people in all walks of life to be part of the research that professional astronomers (and scholars in a growing number of fields) need help with. Direct link: https://www.zooniverse.org/projects . Direct link for citizen science projects in Canada: http://science.gc.ca/eic/site/063.nsf/eng/h_97169.html

Attribution

14.4 THE CHALLENGE OF DARK MATTER

So far this chapter has focused almost entirely on matter that radiates electromagnetic energy—stars, planets, gas, and dust. But, as we have pointed out in several earlier chapters, it is now clear that galaxies contain large amounts of dark matter as well. There is much more dark matter, in fact, than matter we can see—which means it would be foolish to ignore the effect of this unseen material in our theories about the structure of the universe. (As many a ship captain in the polar seas found out too late, the part of the iceberg visible above the ocean's surface was not necessarily the only part he needed to pay attention to.) Dark matter turns out to be extremely important in determining the evolution of galaxies and of the universe as a whole.

The idea that much of the universe is filled with dark matter may seem like a bizarre concept, but we can cite a historical example of "dark matter" much closer to home. In the mid-nineteenth century, measurements showed that the planet Uranus did not follow exactly the orbit predicted from Newton's laws if one added up the gravitational forces of all the known objects in the solar system. Some people worried that Newton's laws may simply not work so far out in our solar system. But the more straightforward interpretation was to attribute Uranus' orbital deviations to the gravitational effects of a new planet that had not yet been seen. Calculations showed where that planet had to be, and Neptune was discovered just about in the predicted location.

In the same way, astronomers now routinely determine the location and amount of dark matter in galaxies by measuring its gravitational effects on objects we can see. And, by measuring the way that galaxies move in clusters, scientists have discovered that dark matter is also distributed among the galaxies in the clusters. Since the environment surrounding a galaxy is important in its development, dark matter must play a central role in galaxy evolution as well. Indeed, it appears that dark matter makes up most of the matter in the universe. But what *is* dark matter? What is it made of? We'll look next at the search for dark matter and the quest to determine its nature.

Dark Matter in the Local Neighbourhood

Is there dark matter in our own solar system? Astronomers have examined the orbits of the known planets and of spacecraft as they journey to the outer planets and beyond. No deviations have been found from the orbits predicted on the basis of the masses of objects already discovered in our solar system and the theory of gravity. We therefore conclude that there is no evidence that there are large amounts of dark matter nearby.

Astronomers have also looked for evidence of dark matter in the region of the Milky Way Galaxy that lies within a few hundred light-years of the Sun. In this vicinity, most of the stars are restricted to a thin disk. It is possible to calculate how much mass the disk must contain in order to keep the stars from wandering far above or below it. The total matter that must be in the disk is less than twice the amount of luminous matter. This means that no more than half of the mass in the region near the Sun can be dark matter.

Dark Matter in and around Galaxies

In contrast to our local neighbourhood near the Sun and solar system, there is ample evidence strongly suggesting that about 90% of the mass in the entire galaxy is in the form of a halo of dark matter. In other words, there is apparently about nine times more dark matter than visible matter. Astronomers have found some stars in the outer regions of the Milky Way beyond its bright disk, and these stars are revolving very rapidly around its centre. The mass contained in all the stars and all the interstellar matter we can detect in the galaxy does not exert enough gravitational force to explain how those fast-moving stars remain in their orbits and do not fly away. Only by having large amounts of unseen matter could the galaxy be holding on to those fast-moving outer stars. The same result is found for other spiral galaxies as well.

Figure 14.5 is an example of the kinds of observations astronomers are making, for the Andromeda galaxy, a member of our Local Group. The observed rotation of spiral galaxies like Andromeda is usually seen in plots, known as **rotation curves**, that show velocity versus distance from the galaxy centre. Such plots suggest that the dark matter is found in a large halo surrounding the luminous parts of each galaxy. The radius of the halos around the Milky Way and Andromeda may be as large as 300,000 light-years, much larger than the visible size of these galaxies.

Rotation Indicates Dark Matter

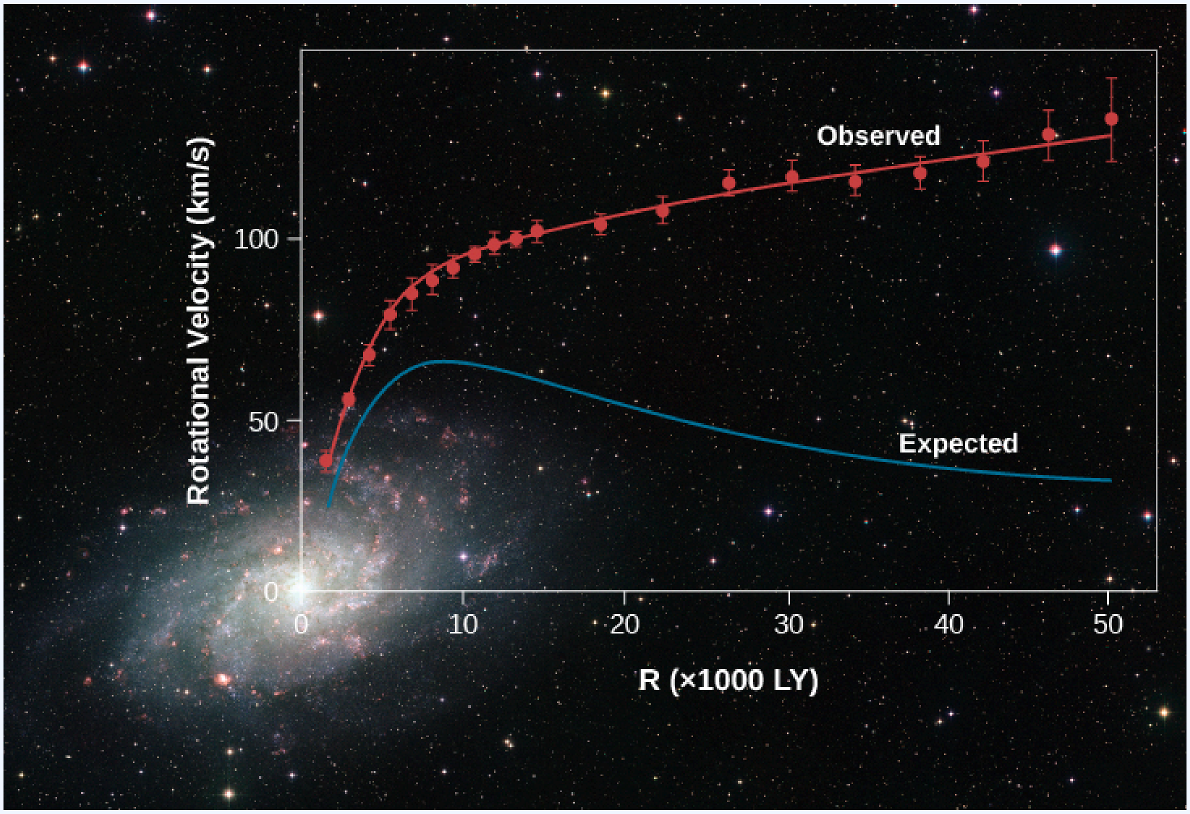


Figure 14.5. We see the Milky Way's sister, the spiral Andromeda galaxy, with a graph that shows the velocity at which stars and clouds of gas orbit the galaxy at different distances from the centre (red line). As is true of the Milky Way, the rotational velocity (or orbital speed) does not decrease with distance from the centre, which is what you would expect if an assembly of objects rotates around a common centre. A calculation (blue line) based on the total mass visible as stars, gas, and dust predicts that the velocity should be much lower at larger distances from the centre. The discrepancy between the two curves implies the presence of a halo of massive dark matter extending outside the boundary of the luminous matter. This dark matter causes everything in the galaxy to orbit faster than the observed matter alone could explain.
Background: modification of work by ESO, CC BY 4.0.

Dark Matter in Clusters of Galaxies

Galaxies in clusters also move around: they orbit the cluster's centre of mass. It is not possible for us to follow a galaxy around its entire orbit because that typically takes about a billion years. It is possible, however, to measure the velocities with which galaxies in a cluster are moving, and then estimate what the total mass in the cluster must be to keep the individual galaxies from flying out of the cluster. The observations indicate that the mass of the galaxies alone cannot keep the cluster together—some other gravity must again be present. The total amount of dark matter in clusters exceeds by more than ten times the luminous mass contained within the galaxies themselves, indicating that dark matter exists between galaxies as well as inside them.

There is another approach we can take to measuring the amount of dark matter in clusters of galaxies. As we saw, the universe is expanding, but this expansion is not perfectly uniform, thanks to the interfering hand of gravity. Suppose, for example, that a galaxy lies outside but relatively close to a rich cluster of galaxies. The gravitational force of the cluster will tug on that neighbouring galaxy and slow down the rate at which it moves away from the cluster due to the expansion of the universe.

Consider the Local Group of galaxies, lying on the outskirts of the Virgo Supercluster. The mass concentrated at the centre of the Virgo Cluster exerts a gravitational force on the Local Group. As a result, the Local Group is moving away from the centre of the Virgo Cluster at a velocity a few hundred kilometres per second slower than the Hubble law predicts. By measuring such deviations from a smooth expansion, astronomers can estimate the total amount of mass contained in large clusters.

There are two other very useful methods for measuring the amount of dark matter in galaxy clusters, and both of them have produced results in general agreement with the method of measuring galaxy velocities: gravitational lensing and X-ray emission. Let's take a look at both.

As Albert Einstein showed in his theory of general relativity, the presence of mass bends the surrounding fabric of spacetime. Light follows those bends, so very massive objects can bend light significantly. Visible galaxies are not the only possible gravitational lenses. Dark matter can also reveal its presence by producing this effect. Figure 14.6 shows a galaxy cluster that is acting like a gravitational lens; the streaks and arcs you see on the picture are lensed images of more distant galaxies. Gravitational lensing is well enough understood that astronomers can use the many ovals and arcs seen in this image to calculate detailed maps of how much matter there is in the cluster and how that mass is distributed. The result from studies of many such gravitational lens clusters shows that, like individual galaxies, galaxy clusters contain more than ten times as much dark matter as luminous matter.

Cluster Abell 2218

Figure 14.6. This view from the Hubble Space Telescope shows the massive galaxy cluster Abell 2218 at a distance of about 2 billion light-years. Most of the yellowish objects are galaxies belonging to the cluster. But notice the numerous long, thin streaks, many of them blue; those are the distorted and magnified images of even more distant background galaxies, gravitationally lensed by the enormous mass of the intervening cluster. By carefully analyzing the lensed images, astronomers can construct a map of the dark matter that dominates the mass of the cluster.
Image by NASA, ESA, and Johan Richard (Caltech), ESA Standard License.

The third method astronomers use to detect and measure dark matter in galaxy clusters is to image them in the light of X-rays. When the first sensitive X-ray telescopes were launched into orbit around Earth in the 1970s and trained on massive galaxy clusters, it was quickly discovered that the clusters emit copious X-ray radiation as seen in Figure 14.6. Most stars do not emit much X-ray radiation, and neither does most of the gas or dust between the stars inside galaxies. What could be emitting the X-rays seen from virtually all massive galaxy clusters?

It turns out that just as galaxies have gas distributed between their stars, clusters of galaxies have gas distributed between their galaxies. The particles in these huge reservoirs of gas are not just sitting still; rather, they are constantly moving, zooming around under the influence of the cluster's immense gravity like mini planets around a giant sun. As they move and bump against each other, the gas heats up hotter and hotter until, at temperatures as high as 100 million K, it shines brightly at X-ray wavelengths. The more mass the cluster has, the faster the motions, the hotter the gas, and the brighter the X-rays. Astronomers calculate that the mass present to induce those motions must be about ten times the mass they can see in the clusters, including all the galaxies and all the gas. Once again, this is evidence that the galaxy clusters are seen to be dominated by dark matter.

X-Ray Image of a Galaxy Cluster

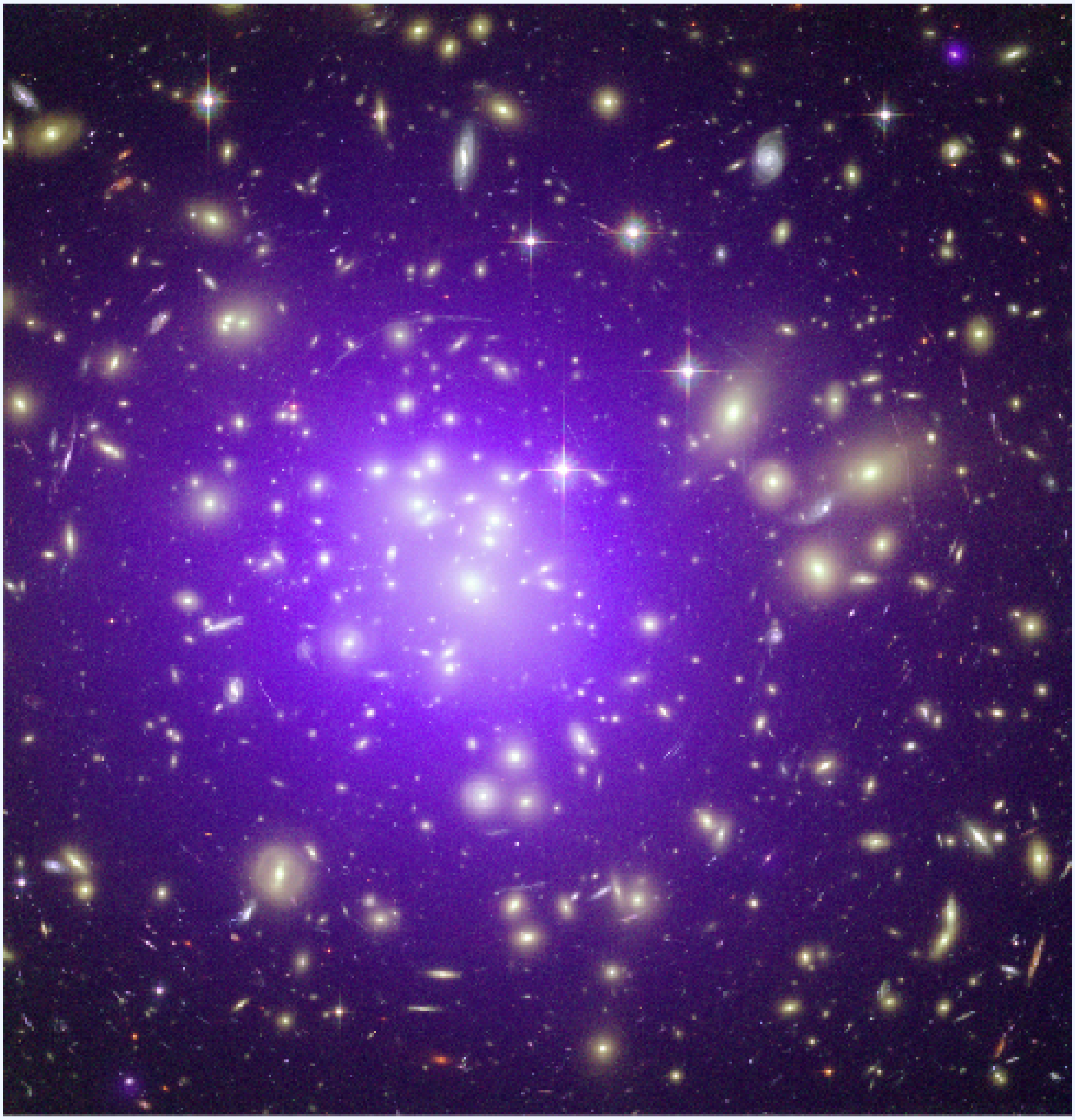


Figure 14.7. This composite image shows the galaxy cluster Abell 1689 at a distance of 2.3 billion light-years. The finely detailed views of the galaxies, most of them yellow, are in visible and near-infrared light from the Hubble Space Telescope, while the diffuse purple haze shows X-rays as seen by Chandra X-ray Observatory. The abundant X-rays, the gravitationally lensed images (thin curving arcs) of background galaxies, and the measured velocities of galaxies in the clusters all show that the total mass of Abell 1689—most of it dark matter—is about 1015 solar masses.
Abell 1689: A Galaxy Cluster Makes Its Mark by NASA/ESA/JPL-Caltech/Yale/CNRS, NASA Media License.

What Is the Dark Matter?

How do we go about figuring out what the dark matter consists of? The technique we might use depends on its composition. Let's consider the possibility that some of the dark matter is made up of normal particles: protons, neutrons, and electrons. Suppose these particles were assembled into black holes, brown dwarfs, or white dwarfs. If the black holes had no accretion disks, they would be invisible to us. White and brown dwarfs do emit some radiation but have such low luminosities that they cannot be seen at distances greater than a few thousand light-years.

We can, however, look for such compact objects because they can act as gravitational lenses. Suppose the dark matter in the halo of the Milky Way were made up of black holes, brown dwarfs, and white dwarfs. These objects have been whimsically dubbed MACHOs (MAssive Compact Halo Objects). If an invisible MACHO passes directly between a distant star and Earth, it acts as a gravitational lens, focusing the light from the distant star. This causes the star to appear to brighten over a time interval of a few hours to several days before returning to its normal brightness. Since we can't predict when any given star might brighten this way, we have to monitor huge numbers of stars to catch one in the act. There are not enough astronomers to keep monitoring so many stars, but today's automated telescopes and computer systems can do it for us.

Research teams making observations of millions of stars in the nearby galaxy called the Large Magellanic Cloud have reported several examples of the type of brightening expected if MACHOs are present in the halo of the Milky Way, as pictured in Figure 14.8. However, there are not enough MACHOs in the halo of the Milky Way to account for the mass of the dark matter in the halo.

Large and Small Magellanic Clouds

Figure 14.8. Here, the two small galaxies we call the Large Magellanic Cloud and Small Magellanic Cloud can be seen above the auxiliary telescopes for the Very Large Telescope Array on Cerro Paranal in Chile. You can see from the number of stars that are visible that this is a very dark site for doing astronomy. A Starry Combination by ESO/J. Colosimo, CC BY 4.0.

This result, along with a variety of other experiments, leads us to conclude that the types of matter we are familiar with can make up only a tiny portion of the dark matter. Another possibility is that dark matter is composed of some new type of particle—one that researchers are now trying to detect in laboratories here on Earth.

The kinds of dark matter particles that astronomers and physicists have proposed generally fall into two main categories: hot and cold dark matter. The terms *hot* and *cold* don't refer to true temperatures, but rather to the average velocities of the particles, analogous to how we might think of particles of air moving in your room right now. In a cold room, the air particles move more slowly on average than in a warm room.

In the early universe, if dark matter particles easily moved fast and far compared to the lumps and bumps of ordinary matter that eventually became galaxies and larger structures, we call those particles **hot dark matter**.

In that case, smaller lumps and bumps would be smeared out by the particle motions, meaning fewer small galaxies would get made.

On the other hand, if the dark matter particles moved slowly and covered only small distances compared to the sizes of the lumps in the early universe, we call that **cold dark matter**. Their slow speeds and energy would mean that even the smaller lumps of ordinary matter would survive to grow into small galaxies. By looking at when galaxies formed and how they evolve, we can use observations to distinguish between the two kinds of dark matter. So far, observations seem most consistent with models based on cold dark matter.

Solving the dark matter problem is one of the biggest challenges facing astronomers. After all, we can hardly understand the evolution of galaxies and the long-term history of the universe without understanding what its most massive component is made of. For example, we need to know just what role dark matter played in starting the higher-density "seeds" that led to the formation of galaxies. And since many galaxies have large halos made of dark matter, how does this affect their interactions with one another and the shapes and types of galaxies that their collisions create?

Astronomers armed with various theories are working hard to produce models of galaxy structure and evolution that take dark matter into account in just the right way. Even though we don't know what the dark matter is, we do have some clues about how it affected the formation of the very first galaxies. As we will see in the next chapter, careful measurements of the microwave radiation left over after the Big Bang have allowed astronomers to set very tight limits on the actual sizes of those early seeds that led to the formation of the large galaxies that we see in today's universe. Astronomers have also measured the relative numbers and distances between galaxies and clusters of different sizes in the universe today. So far, most of the evidence seems to weigh heavily in favour of cold dark matter, and most current models of galaxy and large-scale structure formation use cold dark matter as their main ingredient.

As if the presence of dark matter—a mysterious substance that exerts gravity and outweighs all the known stars and galaxies in the universe but does not emit or absorb light—were not enough, there is an even more baffling and equally important constituent of the universe that has only recently been discovered: we have called it **dark energy** in parallel with dark matter. We will say more about it and explore its effects on the evolution of the universe in the next chapter. For now, we can complete our inventory of the contents of the universe by noting that it appears that the entire universe contains some mysterious energy that pushes spacetime apart, taking galaxies and the larger structures made of galaxies along with it. Observations show that dark energy becomes more and more important relative to gravity as the universe ages. As a result, the expansion of the universe is accelerating, and this acceleration seems to be happening mostly since the universe was about half its current age.

What we see when we peer out into the universe—the light from trillions of stars in hundreds of billions of galaxies wrapped in intricate veils of gas and dust—is therefore actually only a sprinkling of icing on top of the cake: as we will see in the next chapter, when we look outside galaxies and clusters of galaxies at the universe as a whole, astronomers find that for every gram of luminous normal matter, such as protons, neutrons, electrons, and atoms in the universe, there are about 4 grams of nonluminous normal matter, mainly intergalactic hydrogen and helium. There are about 27 grams of dark matter, and the energy equivalent (remember Einstein's famous $E = mc^2$) of about 68 grams of dark energy. Dark matter, and (as we will see) even more so dark energy, are dramatic demonstrations of what we have tried to emphasize throughout this book: science is always a "progress report," and we often encounter areas where we have more questions than answers.

Let's next put together all these clues to trace the life history of galaxies and large-scale structure in the universe. What follows is the current consensus, but research in this field is moving rapidly, and some of these ideas will probably be modified as new observations are made.

Attribution

14.5 THE FORMATION AND EVOLUTION OF GALAXIES AND STRUCTURE IN THE UNIVERSE

As with most branches of natural science, astronomers and cosmologists always want to know the answer to the question, "How did it get that way?" What made galaxies and galaxy clusters, superclusters, voids, and filaments look the way they do? The existence of such large filaments of galaxies and voids is an interesting puzzle because we have evidence (to be discussed in the next chapter) that the universe was extremely smooth even a few hundred thousand years after forming. The challenge for theoreticians is to understand how a nearly featureless universe changed into the complex and lumpy one that we see today. Armed with our observations and current understanding of galaxy evolution over cosmic time, dark matter, and large-scale structure, we are now prepared to try to answer that question on some of the largest possible scales in the universe. As we will see, the short answer to how the universe got this way is "dark matter + gravity + time."

How Galaxies Form and Grow

We've already seen that galaxies were more numerous, but smaller, bluer, and clumpier, in the distant past than they are today, and that galaxy mergers play a significant role in their evolution. At the same time, we have observed quasars and galaxies that emitted their light when the universe was less than a billion years old—so we know that large condensations of matter had begun to form at least that early. We also saw in the previous chapter that many quasars are found in the centres of elliptical galaxies. This means that some of the first large concentrations of matter must have evolved into the elliptical galaxies that we see in today's universe. It seems likely that the supermassive black holes in the centres of galaxies and the spherical distribution of ordinary matter around them formed at the same time and through related physical processes.

Dramatic confirmation of that picture arrived only in the last decade, when astronomers discovered a curious empirical relationship: as we saw in the previous chapter, the more massive a galaxy is, the more massive its central black hole is. Somehow, the black hole and the galaxy "know" enough about each other to match their growth rates.

There have been two main types of galaxy formation models to explain all those observations. The first asserts that massive elliptical galaxies formed in a single, rapid collapse of gas and dark matter, during which virtually all the gas was turned quickly into stars. Afterward the galaxies changed only slowly as the stars evolved. This is what astronomers call a "top-down" scenario.

The second model suggests that today's giant ellipticals were formed mostly through mergers of smaller galaxies that had already converted at least some of their gas into stars—a “bottom-up” scenario. In other words, astronomers have debated whether giant ellipticals formed most of their stars in the large galaxy that we see today or in separate small galaxies that subsequently merged.

Since we see some luminous quasars from when the universe was less than a billion years old, it is likely that at least some giant ellipticals began their evolution very early through the collapse of a single cloud. However, the best evidence also seems to show that mature *giant* elliptical galaxies like the ones we see nearby were rare before the universe was about 6 billion years old and that they are much more common today than they were when the universe was young. Observations also indicate that most of the gas in elliptical galaxies was converted to stars by the time the universe was about 3 billion years old, so it appears that elliptical galaxies have not formed many new stars since then. They are often said to be “red and dead”—that is, they mostly contain old, cool, red stars, and there is little or no new star formation going on.

These observations (when considered together) suggest that the giant elliptical galaxies that we see nearby formed from a combination of both top-down and bottom-up mechanisms, with the most massive galaxies forming in the densest clusters where both processes happened very early and quickly in the history of the universe.

The situation with spiral galaxies is apparently very different. The bulges of these galaxies formed early, like the elliptical galaxies shown in Figure 14.9. However, the disks formed later (remember that the stars in the disk of the Milky Way are younger than the stars in the bulge and the halo) and still contain gas and dust. However, the rate of star formation in spirals today is about ten times lower than it was 8 billion years ago. The number of stars being formed drops as the gas is used up. So spirals seem to form mostly “bottom up” but over a longer time than ellipticals and in a more complex way, with at least two distinct phases.

Growth of Spiral Bulges

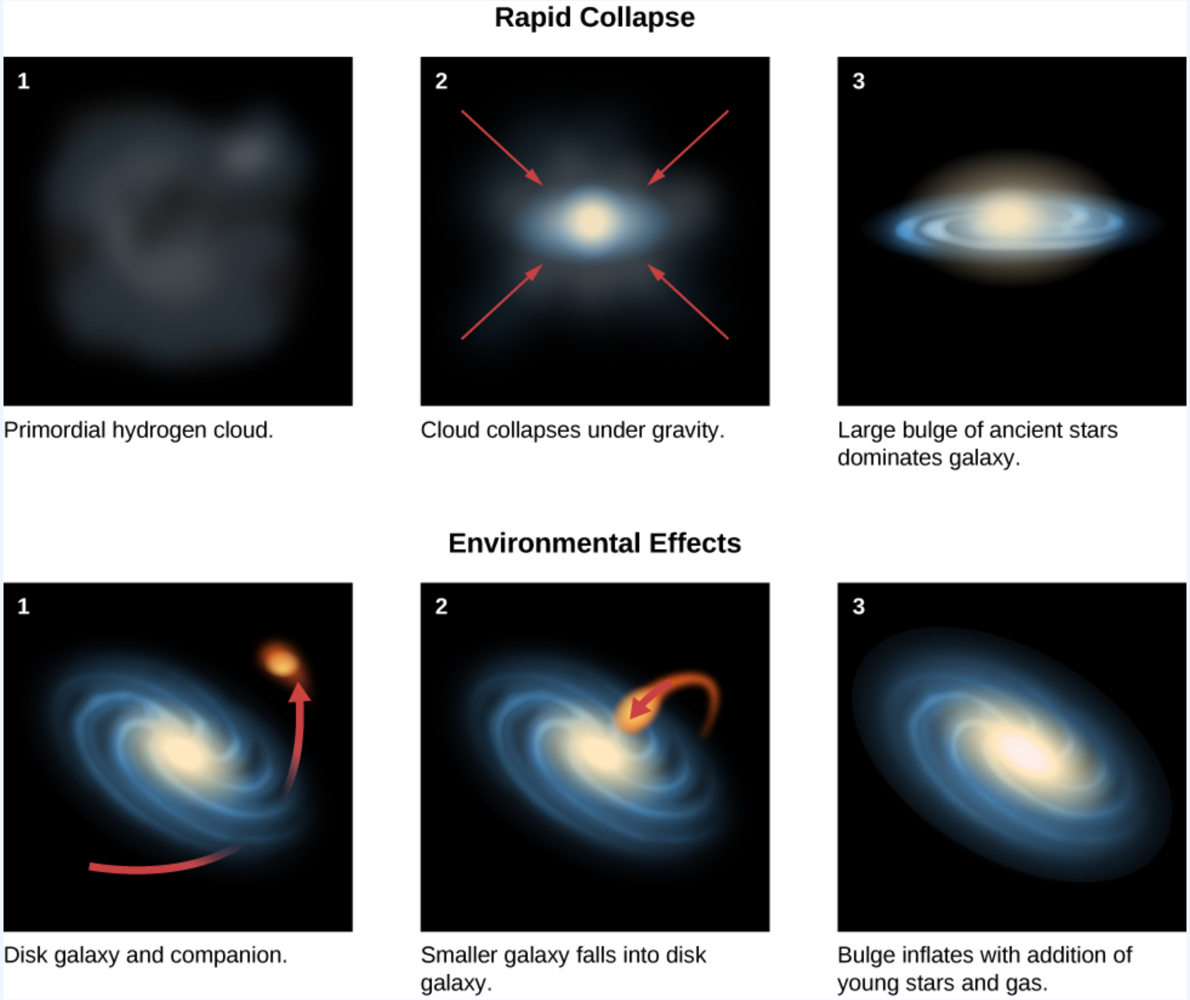


Figure 14.9. The nuclear bulges of some spiral galaxies formed through the collapse of a single protogalactic cloud (top row). Others grew over time through mergers with other smaller galaxies (bottom row).

Hubble originally thought that elliptical galaxies were young and would eventually turn into spirals, an idea we now know is not true. In fact, as we saw above, it's more likely the other way around: two spirals that crash together under their mutual gravity can turn into an elliptical.

Despite these advances in our understanding of how galaxies form and evolve, many questions remain. For example, it's even possible, given current evidence, that spiral galaxies could lose their spiral arms and disks in a merger event, making them look more like an elliptical or irregular galaxy, and then regain the disk and arms again later if enough gas remains available. The story of how galaxies assume their final shapes is still being written as we learn more about galaxies and their environment.

Forming Galaxy Clusters, Superclusters, Voids, and Filaments

If individual galaxies seem to grow mostly by assembling smaller pieces together gravitationally over cosmic time, what about the clusters of galaxies and larger structures? How do we explain the large-scale maps that show galaxies distributed on the walls of huge sponge- or bubble-like structures spanning hundreds of millions of light-years?

As we saw, observations have found increasing evidence for concentrations, filaments, clusters, and superclusters of galaxies when the universe was less than 3 billion years old, this is pictured in Figure 14.10. This means that large concentrations of galaxies had already come together when the universe was less than a quarter as old as it is now.

Merging Galaxies in a Distant Cluster

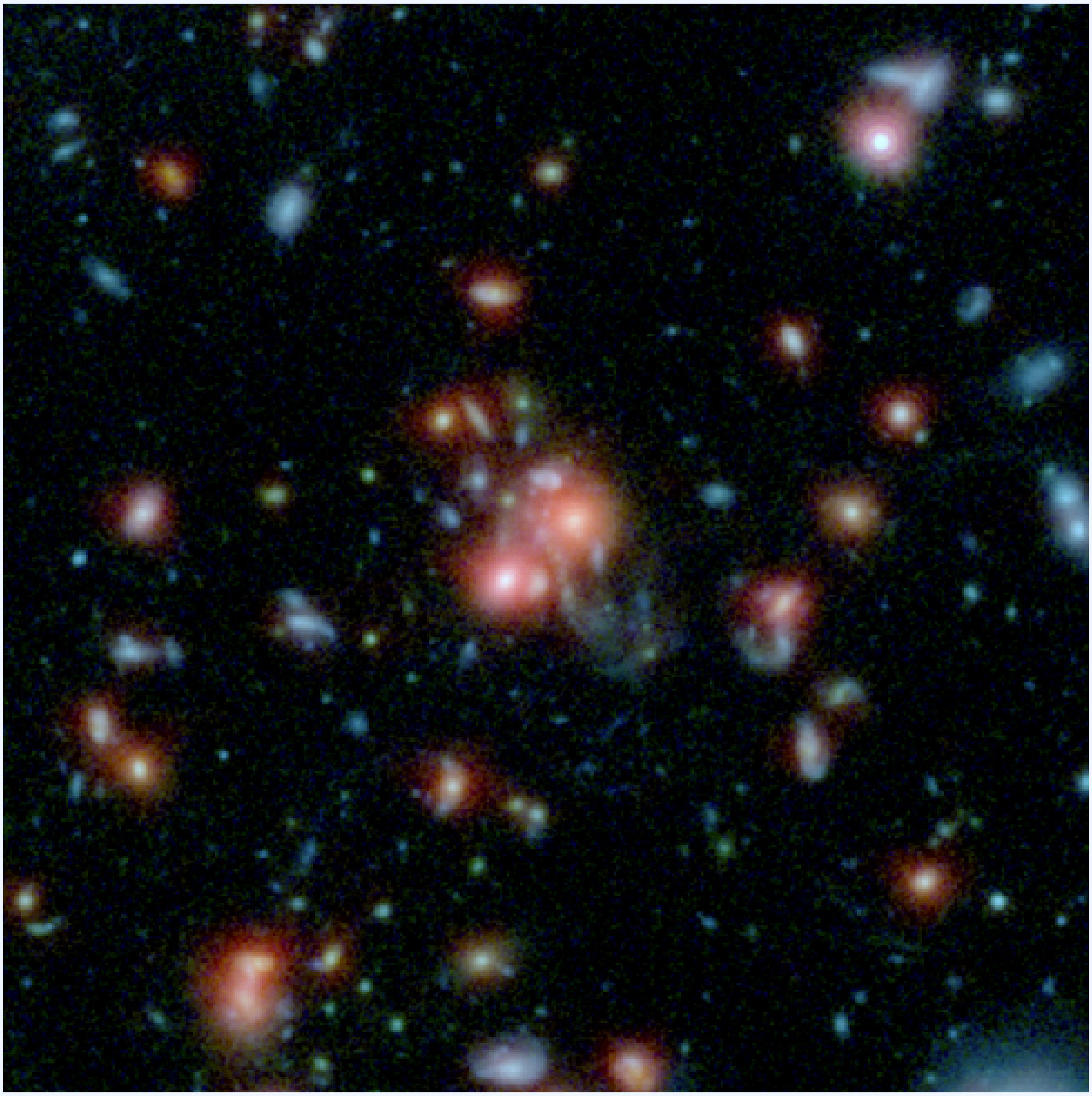


Figure 14.10. This Hubble image shows the core of one of the most distant galaxy clusters yet discovered, SpARCS 1049+56; we are seeing it as it was nearly 10 billion years ago. The surprise delivered by the image was the "train wreck" of chaotic galaxy shapes and blue tidal tails: apparently there are several galaxies right in the core that are merging together, the probable cause of a massive burst of star formation and bright infrared emission from the cluster.

SpARCS1049+56 by NASA/STScI/ESA/JPL-Caltech/McGill, NASA Media License.

Almost all the currently favoured models of how large-scale structure formed in the universe tell a story similar to that for individual galaxies: tiny dark matter "seeds" in the hot cosmic soup after the Big Bang grew by gravity into larger and larger structures as cosmic time ticked on as depicted in Figure 14.11. The final models we construct will need to be able to explain the size, shape, age, number, and spatial distribution of galaxies, clusters, and filaments—not only today, but also far back in time. Therefore, astronomers are working hard to measure and then to model those features of large-scale structure as accurately as possible. So far, a mixture of 5% normal atoms, 27% cold dark matter, and 68% dark energy seems to be the best way to explain all the evidence currently available (see the next chapter).

Growth of Large-Scale Structure as Calculated by Supercomputers

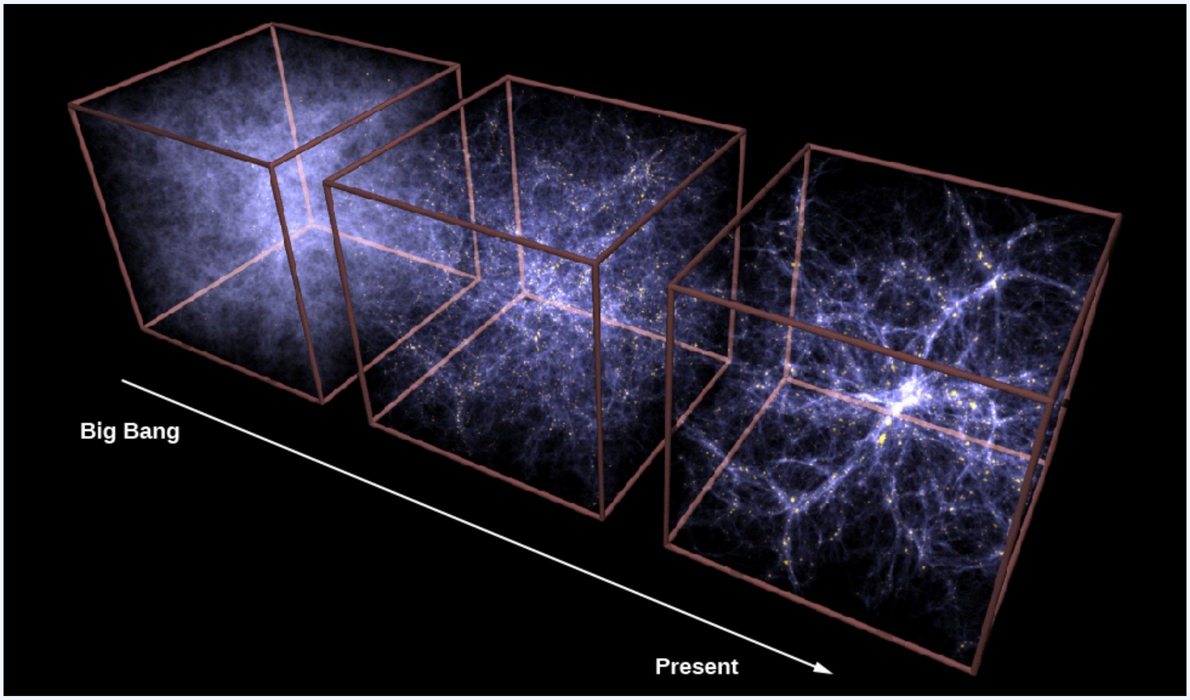


Figure 14.11. The boxes show how filaments and superclusters of galaxies grow over time, from a relatively smooth distribution of dark matter and gas, with few galaxies formed in the first 2 billion years after the Big Bang, to the very clumpy strings of galaxies with large voids today.
Dark Energy Found Stifling Growth in Universe by CXC/MPE/V.Springel, NASA Media License.

The box at left is labelled "Big Bang," the box at centre is unlabelled and the box at right is labelled "Present". A white arrow points from left to right representing the direction of time.

Scientists even have a model to explain how a nearly uniform, hot "soup" of particles and energy at the beginning of time acquired the Swiss-cheese-like structure that we now see on the largest scales. As we will see in the next chapter, when the universe was only a few hundred thousand years old, *everything* was at a temperature of a few thousand degrees. Theorists suggest that at that early time, all the hot gas was vibrating, much as sound waves vibrate the air of a nightclub with an especially loud band. This vibrating could have concentrated matter into high-density peaks and created emptier spaces between them. When the universe cooled, the concentrations of matter were "frozen in," and galaxies ultimately formed from the matter in these high-density regions.

The Big Picture

To finish this chapter, let's put all these ideas together to tell a coherent story of how the universe came to look the way it does. Initially, as we said, the distribution of matter (both luminous and dark) was nearly, but not quite exactly, smooth and uniform. That "not quite" is the key to everything. Here and there were lumps where the density of matter (both luminous and dark) was ever so slightly higher than average.

Initially, each individual lump expanded because the whole universe was expanding. However, as the universe continued to expand, the regions of higher density acquired still more mass because they exerted a slightly larger than average gravitational force on surrounding material. If the inward pull of gravity was high enough, the denser individual regions ultimately stopped expanding. They then began to collapse into irregularly shaped blobs (that's the technical term astronomers use!). In many regions the collapse was more rapid in one direction, so the concentrations of matter were not spherical but came to resemble giant clumps, pancakes, and rope-like filaments—each much larger than individual galaxies.

These elongated clumps existed throughout the early universe, oriented in different directions and collapsing at different rates. The clumps provided the framework for the large-scale filamentary and bubble-like structures that we see preserved in the universe today.

The universe then proceeded to "build itself" from the bottom up. Within the clumps, smaller structures formed first, then merged to build larger ones, like Lego pieces being put together one by one to create a giant Lego metropolis. The first dense concentrations of matter that collapsed were the size of small dwarf galaxies or globular clusters—which helps explain why globular clusters are the oldest things in the Milky Way and most other galaxies. These fragments then gradually assembled to build galaxies, galaxy clusters, and, ultimately, superclusters of galaxies.

According to this picture, small galaxies and large star clusters first formed in the highest density regions of all—the filaments and nodes where the pancakes intersect—when the universe was about two percent of its current age. Some stars may have formed even before the first star clusters and galaxies came into existence. Some galaxy-galaxy collisions triggered massive bursts of star formation, and some of these led to the formation of black holes. In that rich, crowded environment, black holes found constant food and grew in mass. The development of massive black holes then triggered quasars and other active galactic nuclei whose powerful outflows of energy and matter shut off the star formation in their host galaxies. The early universe must have been an exciting place!

Clusters of galaxies then formed as individual galaxies congregated, drawn together by their mutual gravitational attraction as illustrated in Figure 14.12. First, a few galaxies came together to form groups, much like our own Local Group. Then the groups began combining to form clusters and, eventually, superclusters. This model predicts that clusters and superclusters should still be in the process of gathering together, and observations do in fact suggest that clusters are still gathering up their flocks of galaxies and collecting more gas as it flows in along filaments. In some instances we even see entire clusters of galaxies merging together.

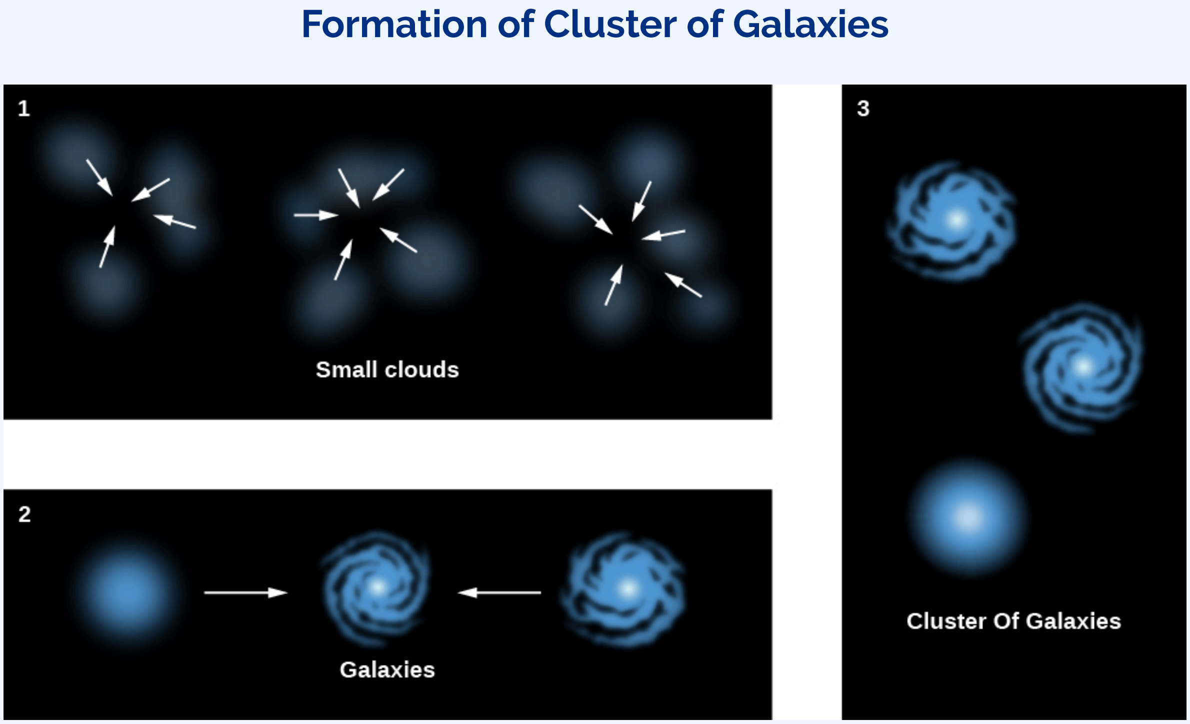


Figure 14.12. This schematic diagram shows how galaxies might have formed if small clouds formed first and then congregated to form galaxies and then clusters of galaxies.

Most giant elliptical galaxies formed through the collision and merger of many smaller fragments. Some spiral galaxies may have formed in relatively isolated regions from a single cloud of gas that collapsed to make a flattened disk, but others acquired additional stars, gas, and dark matter through collisions, and the stars acquired through these collisions now populate their halos and bulges. As we have seen, our Milky Way is still capturing small galaxies and adding them to its halo, and probably also pulling fresh gas from these galaxies into its disk.

Attribution

14.6 KEY TERMS

Cold dark matter: slow-moving massive particles, not yet identified, that don't absorb, emit, or reflect light or other electromagnetic radiation, and that make up most of the mass of galaxies and galaxy clusters. 14.4

Cosmological principle: the assumption that, on the large scale, the universe at any given time is the same everywhere—isotropic and homogeneous. 14.2

Dark energy: an energy that is causing the expansion of the universe to accelerate; the source of this energy is not yet understood. 14.4

Homogeneous: having a consistent and even distribution of matter that is the same everywhere. 14.2

Hot dark matter: massive particles, not yet identified, that don't absorb, emit, or reflect light or other electromagnetic radiation, and that make up most of the mass of galaxies and galaxy clusters; hot dark matter is faster-moving material than cold dark matter. 14.4

Isotropic: the same in all directions. 14.2

Optical fibers: thin tubes of flexible glass that can transmit light from a source to the CCD that then records the spectrum. 14.3

Rotation curves: plots that show velocity versus distance from the galaxy centre for the observed rotation of spiral galaxies. 14.4

Supercluster: a large region of space (more than 100 million light-years across) where groups and clusters of galaxies are more concentrated; a cluster of clusters of galaxies. 14.3

Void: a region between clusters and superclusters of galaxies that appears relatively empty of galaxies. 14.3

CHAPTER 15: THE ORIGIN AND EVOLUTION OF THE UNIVERSE

Chapter Overview

15.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Summarize the process of how astronomers measured the spectra of galaxies and explain how the redshift phenomenon led to the discovery of the expanding universe.
- Calculate the age of the universe based on measurements of galaxy distances and velocities, applying the Hubble constant concept.
- Evaluate the evidence and reasons behind astronomers' conclusions about the fate of the universe, considering factors like density, dark energy, and the observed acceleration of expansion.
- Explain the concept of Cosmic Microwave Background (CMB) radiation and its significance in supporting the Big Bang theory.
- Identify the key components of the standard model of cosmology, including the assumptions, observations, and supporting evidence.
- Describe the composition of the universe, distinguishing between luminous matter, dark matter, and dark energy, and how they contribute to the critical density.

15.1 THE EXPANDING UNIVERSE

We now come to one of the most important discoveries ever made in astronomy—the fact that the universe is expanding. Before we describe how the discovery was made, we should point out that the first steps in the study of galaxies came at a time when the techniques of spectroscopy were also making great strides. Astronomers using large telescopes could record the spectrum of a faint star or galaxy on photographic plates, guiding their telescopes so they remained pointed to the same object for many hours and collected more light. The resulting spectra of galaxies contained a wealth of information about the composition of the galaxy and the velocities of these great star systems.

Slipher's Pioneering Observations

Curiously, the discovery of the expansion of the universe began with the search for Martians and other solar systems. In 1894, the controversial (and wealthy) astronomer Percival Lowell established an observatory in Flagstaff, Arizona, to study the planets and search for life in the universe. Lowell thought that the spiral nebulae might be solar systems in the process of formation. He therefore asked one of the observatory's young astronomers, Vesto M. Slipher, pictured in Figure 15.1, to photograph the spectra of some of the spiral nebulae to see if their spectral lines might show chemical compositions like those expected for newly forming planets.

Vesto M. Slipher (1875–1969)

Figure 15.1. Slipher spent his entire career at the Lowell Observatory, where he discovered the large radial velocities of galaxies. (credit: Lowell Observatory, used under fair dealing. All rights reserved.)

The Lowell Observatory's major instrument was a 24-inch refracting telescope, which was not at all well suited to observations of faint spiral nebulae. With the technology available in those days, photographic plates had to be exposed for 20 to 40 hours to produce a good spectrum (in which the positions of the lines could reveal a galaxy's motion). This often meant continuing to expose the same photograph over several nights. Beginning in 1912, and making heroic efforts over a period of about 20 years, Slipher managed to photograph the spectra of more than 40 of the spiral nebulae (which would all turn out to be galaxies).

To his surprise, the spectral lines of most galaxies showed an astounding redshift. By “redshift” we mean that the lines in the spectra are displaced toward longer wavelengths (toward the red end of the visible spectrum). A redshift is seen when the source of the waves is moving away from us. Slipher’s observations showed that most spirals are racing away at huge speeds; the highest velocity he measured was 1800 kilometres per second.

Only a few spirals—such as the Andromeda and Triangulum Galaxies and M81—all of which are now known to be our close neighbours, turned out to be approaching us. All the other galaxies were moving away. Slipher first announced this discovery in 1914, years before Hubble showed that these objects were other galaxies and before anyone knew how far away they were. No one at the time quite knew what to make of this discovery.

Hubble’s Law (Now the Hubble-Lemaître Law)

The profound implications of Slipher’s work became apparent only during the 1920s. Georges Lemaître was a Belgian priest and a trained astronomer. In 1927, he published a paper in French in an obscure Belgian journal in which he suggested that we live in an expanding universe. The title of the paper (translated into English) is “A Homogenous Universe of Constant Mass and Growing Radius Accounting for the Radial Velocity of Extragalactic Nebulae.” Lemaître had discovered that Einstein’s equations of relativity were consistent with an expanding universe (as had the Russian scientist Alexander Friedmann independently in 1922). Lemaître then went on to use Slipher’s data to support the hypothesis that the universe actually is expanding and to estimate the rate of expansion. Initially, scientists paid little attention to this paper, perhaps because the Belgian journal was not widely available.

In the meantime, Hubble was making observations of galaxies with the 2.5-meter telescope on Mt. Wilson, which was then the world’s largest. Hubble carried out the key observations in collaboration with a remarkable man, Milton Humason, who dropped out of school in the eighth grade and began his astronomical career by driving a mule train up the trail on Mount Wilson to the observatory. He is shown in Figure 15.2. In those early days, supplies had to be brought up that way; even astronomers hiked up to the mountaintop for their turns at the telescope. Humason became interested in the work of the astronomers and, after marrying the daughter of the observatory’s electrician, took a job as janitor there. After a time, he became a night assistant, helping the astronomers run the telescope and record data. Eventually, he made such a mark that he became a full astronomer at the observatory.

Milton Humason (1891–1972)

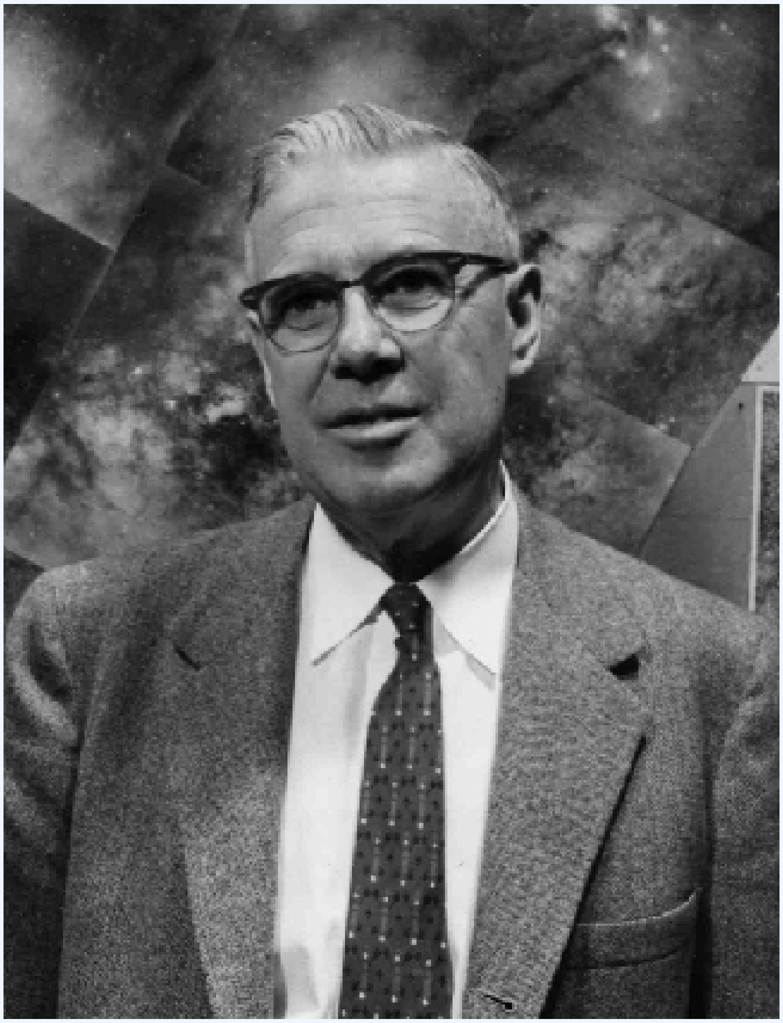
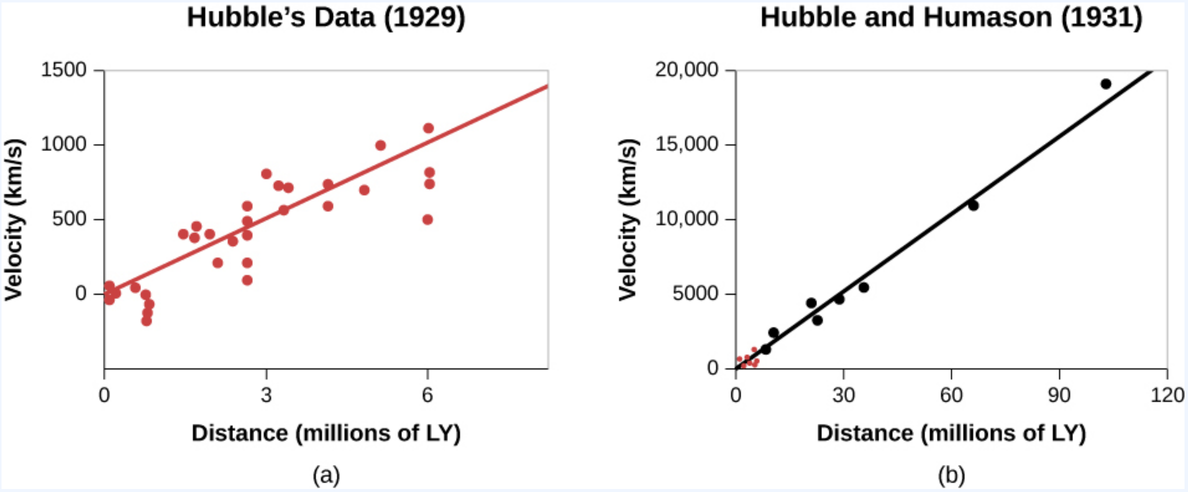


Figure 15.2. Humason was Hubble's collaborator on the great task of observing, measuring, and classifying the characteristics of many galaxies. (credit: Caltech Archives, used under fair dealing. All rights reserved.)

By the late 1920s, Humason was collaborating with Hubble by photographing the spectra of faint galaxies with the 2.5-meter telescope. (By then, there was no question that the spiral nebulae were in fact galaxies.) Hubble had found ways to improve the accuracy of the estimates of distances to spiral galaxies, and he was able to measure much fainter and more distant galaxies than Slipher could observe with his much-smaller telescope. When Hubble laid his own distance estimates next to measurements of the recession velocities (the speed with which the galaxies were moving away), he found something stunning: there was a relationship between distance and velocity for galaxies. *The more distant the galaxy, the faster it was receding from us.*

In 1931, Hubble and Humason jointly published the seminal paper where they compared distances and velocities of remote galaxies moving away from us at speeds as high as 20,000 kilometres per second and were able to show that the recession velocities of galaxies are directly proportional to their distances from us as shown in Figure 15.3, just as Lemaître had suggested. In August 2018 the International Astronomical Union voted to rename this as the **Hubble-Lemaître Law** (https://www.iau.org/news/pressreleases/detail/iau1812/).

Hubble's Law



(a) (b)

Figure 15.3. (a) These data show Hubble's original velocity-distance relation, adapted from his 1929 paper in the Proceedings of the National Academy of Sciences. (b) These data show Hubble and Humason's velocity-distance relation, adapted from their 1931 paper in The Astrophysical Journal. The red dots at the lower left are the points in the diagram in the 1929 paper. Comparison of the two graphs shows how rapidly the determination of galactic distances and redshifts progressed in the 2 years between these publications.

We now know that this relationship holds for every galaxy except a few of the nearest ones. Nearly all of the galaxies that are approaching us turn out to be part of the Milky Way's own group of galaxies, which have their own individual motions, just as birds flying in a group may fly in slightly different directions at slightly different speeds even though the entire flock travels through space together.

Written as a formula, the relationship between velocity and distance is

$$V = H \times d$$

where v is the recession speed, d is the distance, and H is a number called the **Hubble constant**. This equation is now known as Hubble's law.

Constants of Proportionality

Mathematical relationships such as Hubble's law are pretty common in life. To take a simple example, suppose your college or university hires you to call rich alumni and ask for donations. You are paid \$ 2.50 for each call; the more calls you can squeeze in between studying astronomy and other courses, the more money you take home. We can set up a formula that connects p, your pay, and n , the number of calls.

p = A\times n

where A is the alumni constant, with a value of $2.50.

If you make 20 calls, you will earn \$2.50 times 20, or \$50.00. Suppose your boss forgets to tell you what you will get paid for each call. You can calculate the alumni constant that governs your pay by keeping track of how many calls you make and noting your gross pay each week. If you make 100 calls the first week and are paid \$ 250, you can deduce that the constant is \$ 2.50, in units of dollars per call. Hubble, of course, had no "boss" to tell him what his constant would be—he had to calculate its value from the measurements of distance and velocity.

Astronomers express the value of Hubble's constant in units that relate to how they measure speed and velocity for galaxies. In this book, we will use kilometres per second per million light-years as that unit. For many years, estimates of the value of the Hubble constant have been in the range of 15 to 30 kilometres per second per million light-years The most recent work appears to be converging on a value near 22 kilometres per second per million light-years If H is 22 kilometres per second per million light-years, a galaxy moves away from us at a speed of 22 kilometres per second for every million light-years of its distance. As an example, a galaxy 100 million light-years away is moving away from us at a speed of 2200 kilometres per second.

Hubble's law tells us something fundamental about the universe. Since all but the nearest galaxies appear to be in motion away from us, with the most distant ones moving the fastest, we must be living in an expanding universe. We will explore the implications of this idea shortly, as well as in the final chapters of this text. For now, we will just say that Hubble's observation underlies all our theories about the origin and evolution of the universe.

Hubble's Law and Distances

The regularity expressed in Hubble's law has a built-in bonus: it gives us a new way to determine the distances to remote galaxies. First, we must reliably establish Hubble's constant by measuring both the distance and the velocity of many galaxies in many directions to be sure Hubble's law is truly a universal property of galaxies. But once we have calculated the value of this constant and are satisfied that it applies everywhere, much more of the universe opens up for distance determination. Basically, if we can obtain a spectrum of a galaxy, we can immediately tell how far away it is.

The procedure works like this. We use the spectrum to measure the speed with which the galaxy is moving away from us. If we then put this speed and the Hubble constant into Hubble's law equation, we can solve for the distance.

Example 15.1

Hubble's Law

Hubble's law ($v = H \times d$) allows us to calculate the distance to any galaxy. Here is how we use it in practice.

We have measured Hubble's constant to be 22 km/s per million light-years. This means that if a galaxy is 1 million light-years farther away, it will move away 22 km/s faster. So, if we find a galaxy that is moving away at 18,000 km/s, what does Hubble's law tells us about the distance to the galaxy?

Solution

$$d = \frac{v}{H} = \frac{18,000\ \text{km/s}}{\frac{22\ \text{km/s}}{1\ \text{million light-years}}} = \frac{18,000}{22} \times \frac{1\ \text{million light-years}}{1} = 818\text{million l.y.}$$

Note how we handled the units here: the km/s in the numerator and denominator cancel, and the factor of million light-years in the denominator of the constant must be divided correctly before we get our distance of 818 million light-years.

Exercise 15.1

Using 22 km/s/million light-years for Hubble's constant, what recessional velocity do we expect to find if we observe a galaxy at 500 million light-years? The answer is 11,000 km/s as shown below.

Solution

$$v = d \times H = 500 \text{ million light years} \times \frac{22 \text{ km/s}}{1 \text{ million light year}} = 11,000 \text{ km/s}$$

Variation of Hubble's Constant

The use of redshift is potentially a very important technique for determining distances because as we have seen, most of our methods for determining galaxy distances are limited to approximately the nearest few hundred million light-years (and they have large uncertainties at these distances). The use of Hubble's law as a distance indicator requires only a spectrum of a galaxy and a measurement of the Doppler shift, and with large telescopes and modern spectrographs, spectra can be taken of extremely faint galaxies.

But, as is often the case in science, things are not so simple. This technique works if, and only if, the Hubble constant has been truly constant throughout the entire life of the universe. When we observe galaxies billions of light-years away, we are seeing them as they were billions of years ago. What if the Hubble "constant" was different billions of years ago? Before 1998, astronomers thought that, although the universe is expanding, the expansion should be slowing down, or decelerating, because the overall gravitational pull of all matter in the universe would have a dominant, measureable effect. If the expansion is decelerating, then the Hubble constant should be decreasing over time.

The discovery that type Ia supernovae are standard bulbs gave astronomers the tool they needed to observe extremely distant galaxies and measure the rate of expansion billions of years ago. The results were completely unexpected. It turns out that the expansion of the universe is *accelerating* over time! What makes this result so astounding is that there is no way that existing physical theories can account for this observation. While a decelerating universe could easily be explained by gravity, there was no force or property in the universe known to astronomers that could account for the acceleration. Later in this chapter, we will look in more detail at the observations that led to this totally unexpected result and explore its implications for the ultimate fate of the universe.

In any case, if the Hubble constant is not really a constant when we look over large spans of space and time, then the calculation of galaxy distances using the Hubble constant won't be accurate. As we shall see later in this chapter, the accurate calculation of distances requires a model for how the Hubble constant has changed over time. The farther away a galaxy is (and the longer ago we are seeing it), the more important it is to include the effects of the change in the Hubble constant. For galaxies within a few billion light-years, however, the assumption that the Hubble constant is indeed constant gives good estimates of distance.

Models for an Expanding Universe

At first, thinking about Hubble's law and being a fan of the work of Copernicus and Harlow Shapley, you might be shocked. Are all the galaxies really moving *away from us*? Is there, after all, something special about our position in the universe? Worry not; the fact that galaxies are receding from us and that more distant galaxies are moving away more rapidly than nearby ones shows only that the universe is expanding uniformly.

A uniformly expanding universe is one that is expanding at the same rate everywhere. In such a universe, we and all other observers, no matter where they are located, must observe a proportionality between the velocities and distances of equivalently remote galaxies. (Here, we are ignoring the fact that the Hubble constant is not constant over all time, but if at any given time in the evolution of the universe the Hubble constant has the same value everywhere, this argument still works.)

To see why, first imagine a ruler made of stretchable rubber, with the usual lines marked off at each centimetre. Now suppose someone with strong arms grabs each end of the ruler and slowly stretches it so that, say, it doubles in length in 1 minute as illustrated in Figure 14.4. Consider an intelligent ant sitting on the mark at 2 centimetres—a point that is not at either end nor in the middle of the ruler. He measures how fast other ants, sitting at the 4-, 7-, and 12-centimetre marks, move away from him as the ruler stretches.

Stretching a Ruler

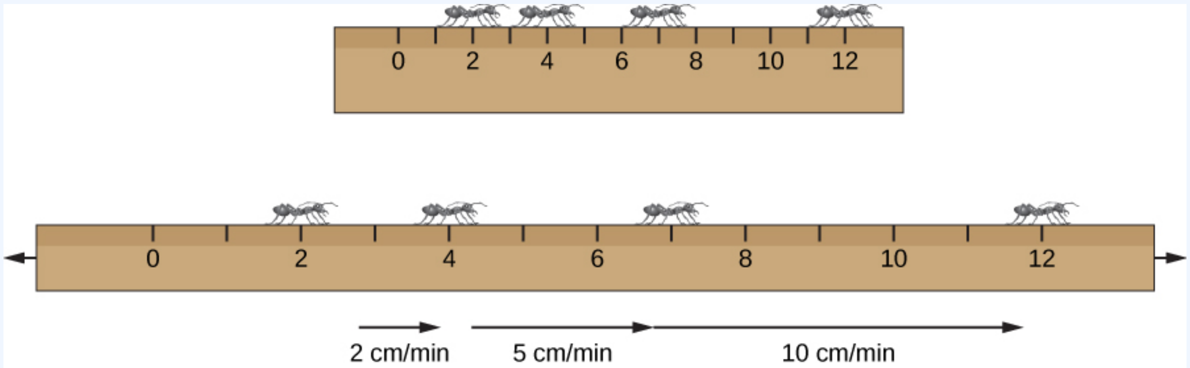


Figure 15.4. Ants on a stretching ruler see other ants move away from them. The speed with which another ant moves away is proportional to its distance.

The ant at 4 centimetres, originally 2 centimetres away from our ant, has doubled its distance in 1 minute; it therefore moved away at a speed of 2 centimetres per minute. The ant at the 7-centimetres mark, which was originally 5 centimetres away from our ant, is now 10 centimetres away; it thus had to move at 5 centimetres per minute. The one that started at the 12-centimetres mark, which was 10 centimetres away from the ant doing the counting, is now 20 centimetres away, meaning it must have raced away at a speed of 10 centimetres per minute. Ants at different distances move away at different speeds, and their speeds are proportional to their distances (just as Hubble's law indicates for galaxies). Yet, notice in our example that all the ruler was doing was stretching uniformly. Also, notice that none of the ants were actually moving of their own accord, it was the stretching of the ruler that moved them apart.

Now let's repeat the analysis, but put the intelligent ant on some other mark—say, on 7 or 12 centimetres. We discover that, as long as the ruler stretches uniformly, this ant also finds every other ant moving away at a speed proportional to its distance. In other words, the kind of relationship expressed by Hubble's law can be explained by a uniform stretching of the "world" of the ants. And all the ants in our simple diagram will see the other ants moving away from them as the ruler stretches.

For a three-dimensional analogy, let's look at the loaf of raisin bread in Figure 15.5. The chef has accidentally put too much yeast in the dough, and when she sets the bread out to rise, it doubles in size during the next hour, causing all the raisins to move farther apart. On the figure, we again pick a representative raisin (that is not at the edge or the centre of the loaf) and show the distances from it to several others in the figure (before and after the loaf expands).

Expanding Raisin Bread

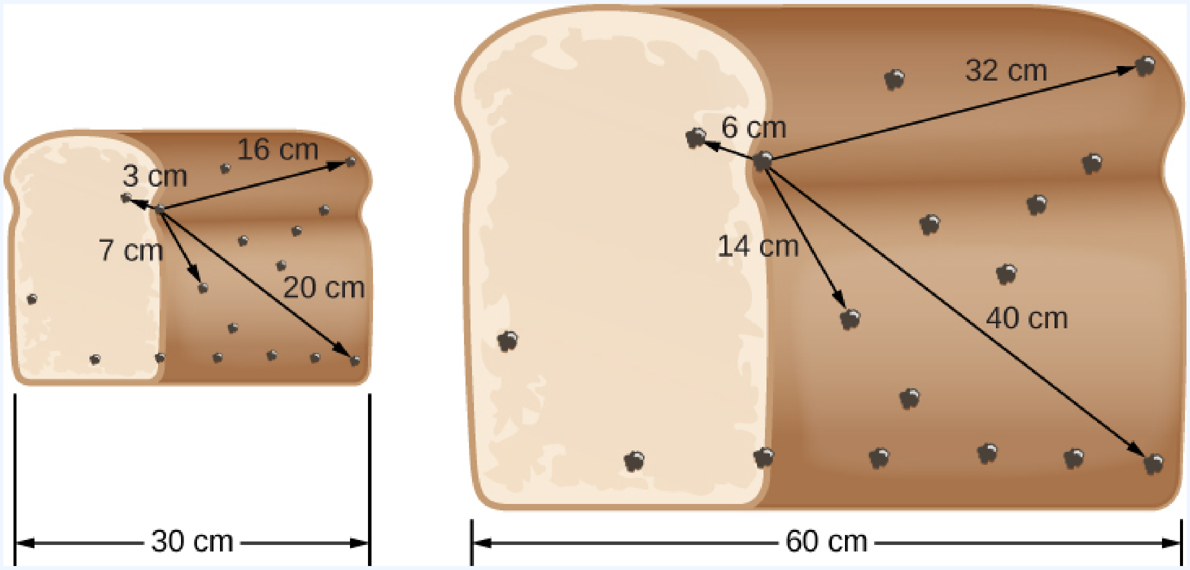


Figure 15.5. As the raisin bread rises, the raisins "see" other raisins moving away. More distant raisins move away faster in a uniformly expanding bread.

Measure the increases in distance and calculate the speeds for yourself on the raisin bread, just like we did for the ruler. You will see that, since each distance doubles during the hour, each raisin moves away from our selected raisin at a speed proportional to its distance. The same is true no matter which raisin you start with.

Our two analogies are useful for clarifying our thinking, but you must not take them literally. On both the ruler and the raisin bread, there are points that are at the end or edge. You can use these to pinpoint the middle of the ruler and the loaf. While our models of the universe have some resemblance to the properties of the ruler and the loaf, the universe has no boundaries, no edges, and no centre (all mind-boggling ideas that we will discuss in a later chapter).

What is useful to notice about both the ants and the raisins is that they themselves did not "cause" their motion. It isn't as if the raisins decided to take a trip away from each other and then hopped on a hoverboard to get away. No, in both our analogies, it was the stretching of the medium (the ruler or the bread) that moved the ants or the raisins farther apart. In the same way, we will see later in this chapter that the galaxies don't have rocket motors propelling them away from each other. Instead, they are passive participants in the *expansion of space*. As space stretches, the galaxies are carried farther and farther apart much as the ants and the raisins were.

The expansion of the universe, by the way, does not imply that the individual galaxies and clusters of galaxies themselves are expanding. Neither raisins nor the ants in our analogy grow in size as the loaf expands. Similarly, gravity holds galaxies and clusters of galaxies together, and they get farther away from each other—without themselves changing in size—as the universe expands.

Attribution

15.2 THE AGE OF THE UNIVERSE

To explore the history of the universe, we will follow the same path that astronomers followed historically—beginning with studies of the nearby universe and then probing ever-more-distant objects and looking further back in time.

The realization that the universe changes with time came in the 1920s and 1930s when measurements of the redshifts of a large sample of galaxies became available. With hindsight, it is surprising that scientists were so shocked to discover that the universe is expanding. In fact, our theories of gravity demand that the universe must be either expanding or contracting. To show what we mean, let's begin with a universe of finite size—say a giant ball of a thousand galaxies. All these galaxies attract each other because of their gravity. If they were initially stationary, they would inevitably begin to move closer together and eventually collide. They could avoid this collapse only if for some reason they happened to be moving away from each other at high speeds. In just the same way, only if a rocket is launched at high enough speed can it avoid falling back to Earth.

The problem of what happens in an infinite universe is harder to solve, but Einstein (and others) used his theory of general relativity to show that even infinite universes cannot be static. Since astronomers at that time did not yet know the universe was expanding (and Einstein himself was philosophically unwilling to accept a universe in motion), he changed his equations by introducing an arbitrary new term (we might call it a fudge factor) called the **cosmological constant**. This constant represented a hypothetical force of repulsion that could balance gravitational attraction on the largest scales and permit galaxies to remain at fixed distances from one another. That way, the universe could remain still.

Einstein and Hubble

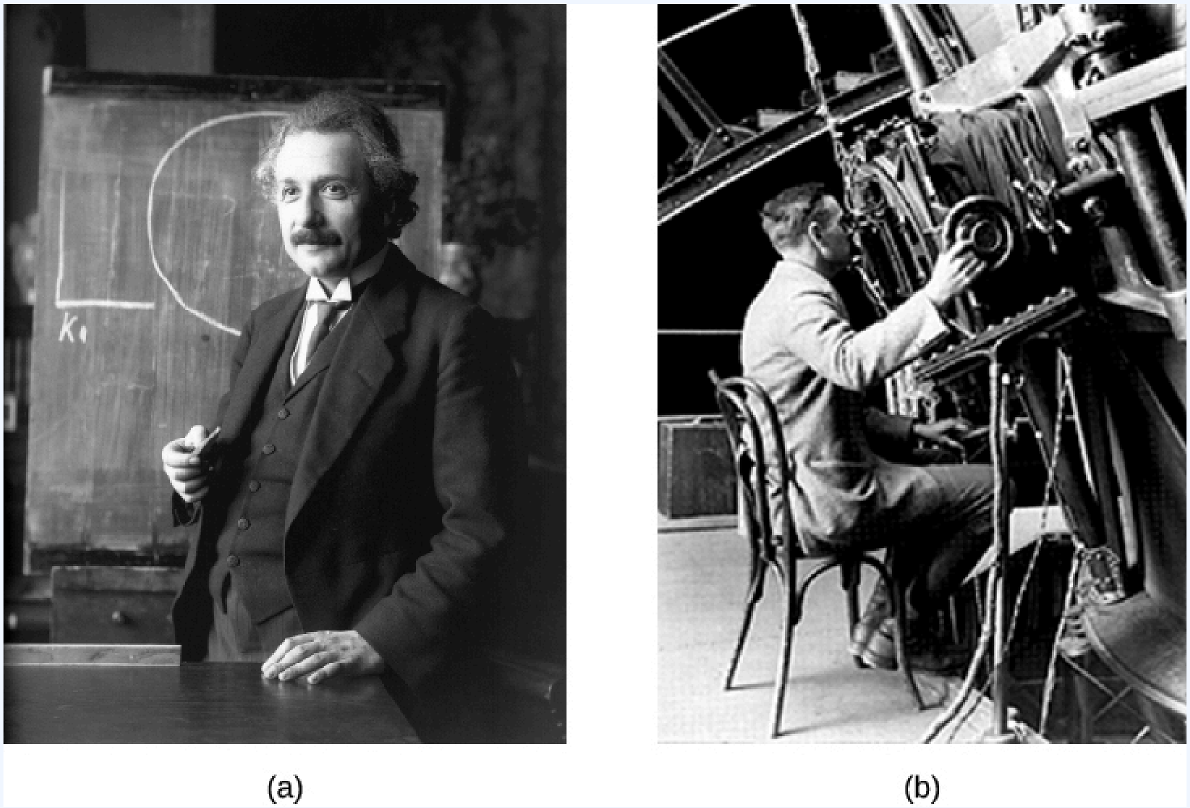


Figure 15.6. (a) Albert Einstein is shown in a 1921 photograph.
Albert Einstein 1921 by F Schmutzer by Creative Commons, CC0 (b) Edwin Hubble at work in the Mt. Wilson Observatory.
Edwin Hubble by NASA, NASA Licence

About a decade later, Hubble, and his coworkers reported that the universe is expanding, so that no mysterious balancing force is needed. Einstein is reported to have said that the introduction of the cosmological constant was “the biggest blunder of my life.” As we shall see later in this chapter, however, relatively recent observations indicate that the expansion is *accelerating*. Observations are now being carried out to determine whether this acceleration is consistent with a cosmological constant. In a way, it may turn out that Einstein was right after all.

View this web exhibit on the history of our thinking about **cosmology**, with images and biographies, from the American Institute of Physics Center for the History of Physics. Direct link: https://history.aip.org/history/exhibits/cosmology/index.htm

The Hubble Time

If we had a movie of the expanding universe and ran the film *backward*, what would we see? The galaxies, instead of moving apart, would move *together* in our movie—getting closer and closer all the time. Eventually, we would find that all the matter we can see today was once concentrated in an infinitesimally small volume. Astronomers identify this time with the *beginning of the universe*. The explosion of that concentrated universe at the beginning of time is called the **Big Bang** (not a bad term, since you can't have a bigger bang than one that creates the entire universe). But when did this bang occur?

We can make a reasonable estimate of the time since the universal expansion began. To see how astronomers do this, let's begin with an analogy. Suppose your astronomy class decides to have a party (a kind of "Big Bang") at someone's home to celebrate the end of the semester. Unfortunately, everyone is celebrating with so much enthusiasm that the neighbours call the police, who arrive and send everyone away at the same moment. You get home at 2 a.m., still somewhat upset about the way the party ended, and realize you forgot to look at your watch to see what time the police got there. But you use a map to measure that the distance between the party and your house is 40 kilometres. And you also remember that you drove the whole trip at a steady speed of 80 kilometres/hour (since you were worried about the police cars following you). Therefore, the trip must have taken:

$$\text{time} = \frac{\text{distance}}{\text{velocity}} = \frac{40 \text{ kilometres}}{80 \text{ kilometres/hour}} = 0.5 \text{ hours}$$

So the party must have broken up at 1:30 a.m.

No humans were around to look at their watches when the universe began, but we can use the same technique to estimate when the galaxies began moving away from each other. (Remember that, in reality, it is space that is expanding, not the galaxies that are moving through static space.) If we can measure how far apart the galaxies are now, and how fast they are moving, we can figure out how long a trip it's been.

Let's call the age of the universe measured in this way T_0. Let's first do a simple case by assuming that the expansion has been at a constant rate ever since the expansion of the universe began. In this case, the time it has taken a galaxy to move a distance, d, away from the Milky Way (remember that at the beginning the galaxies were all together in a very tiny volume) is (as in our example)

$$T_0 = d/v$$

where v is the velocity of the galaxy. If we can measure the speed with which galaxies are moving away, and also the distances between them, we can establish how long ago the expansion began.

Making such measurements should sound very familiar. This is just what Hubble and many astronomers after him needed to do in order to establish the Hubble law and the Hubble constant. We learned in a previous chapter that a galaxy's distance and its velocity in the expanding universe are related by

$$V = H \times d$$

where H is the Hubble constant. Combining these two expressions gives us

$$T_0 = \frac{d}{v} = \frac{d}{(H \times d)} = \frac{1}{H}$$

We see, then, that the work of calculating this time was already done for us when astronomers measured the Hubble constant. The age of the universe estimated in this way turns out to be just the *reciprocal of the Hubble constant* (that is, $1/H$). This age estimate is sometimes called the Hubble time. For a Hubble constant of 20 kilometres/second per million light-years, the Hubble time is about 15 billion years. The unit used by astronomers for the Hubble constant is kilometres/second per million parsecs. In these units, the Hubble constant is equal to about 70 kilometres/second per million parsecs, again with an uncertainty of about 5%.

To make numbers easier to remember, we have done some rounding here. Estimates for the Hubble constant are actually closer to 21 or 22 kilometres/second per million light-years, which would make the age closer to 14 billion years. But there is still about a 5% uncertainty in the Hubble constant, which means the age of the universe estimated in this way is also uncertain by about 5%.

To put these uncertainties in perspective, however, you should know that 50 years ago, the uncertainty was a factor of 2. Remarkable progress toward pinning down the Hubble constant has been made in the last couple of decades.

The Role of Deceleration

The Hubble time is the right age for the universe only if the expansion rate has been constant throughout the time since the expansion of the universe began. Continuing with our end-of-the-semester-party analogy, this is equivalent to assuming that you traveled home from the party at a constant rate, when in fact this may not have been the case. At first, mad about having to leave, you may have driven fast, but then as you calmed down—and thought about police cars on the highway—you may have begun to slow down until you were driving at a more socially acceptable speed (such as 80 kilometres/hour). In this case, given that you were driving faster at the beginning, the trip home would have taken less than a half-hour.

In the same way, in calculating the Hubble time, we have assumed that *H* has been constant throughout all of time. It turns out that this is not a good assumption. Earlier in their thinking about this, astronomers expected that the rate of expansion should be slowing down. We know that matter creates gravity, whereby all objects pull on all other objects. The mutual attraction between galaxies was expected to slow the expansion as time passed. This means that, if gravity were the only force acting (a big *if*, as we shall see in the next section), then the rate of expansion must have been faster in the past than it is today. In this case, we would say the universe has been *decelerating* since the beginning.

How much it has decelerated depends on the importance of gravity in slowing the expansion. If the universe were nearly empty, the role of gravity would be minor. Then the deceleration would be close to zero, and the universe would have been expanding at a constant rate. But in a universe with any significant density of matter, the pull of gravity means that the rate of expansion should be slower now than it used to be. If we use the current rate of expansion to estimate how long it took the galaxies to reach their current separations, we will overestimate the age of the universe—just as we may have overestimated the time it took for you to get home from the party.

A Universal Acceleration

Astronomers spent several decades looking for evidence that the expansion was decelerating, but they were not successful. What they needed were 1) larger telescopes so that they could measure the redshifts of more distant galaxies and 2) a very luminous **standard bulb** (or standard candle), that is, some astronomical object with known luminosity that produces an enormous amount of energy and can be observed at distances of a billion light-years or more.

If we compare how luminous a standard bulb is supposed to be and how dim it actually looks in our telescopes, the difference allows us to calculate its distance. The redshift of the galaxy such a bulb is in can tell us how fast it is moving in the universe. So we can measure its distance and motion independently.

These two requirements were finally met in the 1990s. Astronomers showed that supernovae of type Ia, with some corrections based on the shapes of their light curves, are standard bulbs. This type of supernova occurs when a white dwarf accretes enough material from a companion star to exceed the Chandrasekhar limit and then collapses and explodes. At the time of maximum brightness, these dramatic supernovae can briefly outshine the galaxies that host them, and hence, they can be observed at very large distances. Large 8- to 10-meter telescopes can be used to obtain the spectra needed to measure the redshifts of the host galaxies as shown in Figure 15.7.

Five Supernovae and Their Host Galaxies

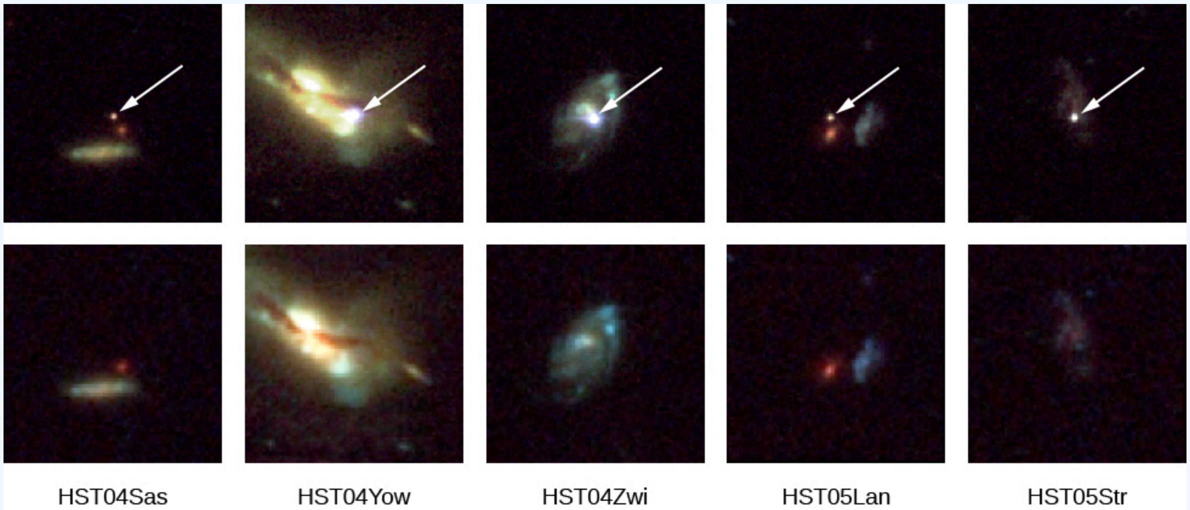


Figure 15.7. The top row shows each galaxy and its supernova (arrow). The bottom row shows the same galaxies either before or after the supernovae exploded.
Hubble Catches Up With Distant Supernovae by NASA, NASA Licence

The result of painstaking, careful study of these supernovae in a range of galaxies, carried out by two groups of researchers, was published in 1998. It was shocking—and so revolutionary that their discovery received the 2011 Nobel Prize in Physics. What the researchers found was that these type Ia supernovae in distant galaxies were fainter than expected from Hubble's law, given the measured redshifts of their host galaxies. In other words, distances estimated from the supernovae used as standard bulbs disagreed with the distances measured from the redshifts.

If the universe were decelerating, we would expect the far-away supernovae to be *brighter* than expected. The slowing down would have kept them closer to us. Instead, they were *fainter*, which at first seemed to make no sense.

Before accepting this shocking development, astronomers first explored the possibility that the supernovae might not really be as useful as standard bulbs as they thought. Perhaps the supernovae appeared too faint because dust along our line of sight to them absorbed some of their light. Or perhaps the supernovae at large distances were for some reason intrinsically less luminous than nearby supernovae of type Ia.

A host of more detailed observations ruled out these possibilities. Scientists then had to consider the alternative that the distance estimated from the redshift was incorrect. Distances derived from redshifts assume that the Hubble constant has been truly constant for all time. We saw that one way it might not be constant is that the expansion is slowing down. But suppose neither assumption is right (steady speed or slowing down.)

Suppose, instead, that the universe is *accelerating*. If the universe is expanding faster now than it was billions of years ago, our motion away from the distant supernovae has sped up since the explosion occurred, sweeping us farther away from them. The light of the explosion has to travel a greater distance to reach us than if the expansion rate were constant. The farther the light travels, the fainter it appears. This conclusion would explain the supernova observations in a natural way, and this has now been substantiated by many additional observations over the last couple of decades. It really seems that *the expansion of the universe is accelerating*, a notion so unexpected that astronomers at first resisted considering it.

How can the expansion of the universe be speeding up? If you want to accelerate your car, you must supply energy by stepping on the gas. Similarly, energy must be supplied to accelerate the expansion of the universe. The discovery of the acceleration was shocking because scientists still have no idea what the source of the energy is. Scientists call whatever it is dark energy, which is a clear sign of how little we understand it.

Note that this new component of the universe is not the dark matter we talked about in earlier chapters. Dark energy is something else that we have also not yet detected in our laboratories on Earth.

What is dark energy? One possibility is that it is the cosmological constant, which is an energy associated with the vacuum of "empty" space itself. Quantum mechanics (the intriguing theory of how things behave at the atomic and subatomic levels) tells us that the source of this vacuum energy might be tiny elementary particles that flicker in and out of existence everywhere throughout the universe. Various attempts have been made to calculate how big the effects of this vacuum energy should be, but so far these attempts have been unsuccessful. In fact, the order of magnitude of theoretical estimates of the vacuum energy based on the quantum mechanics of matter and the value required to account for the acceleration of the expansion of the universe differ by an incredible factor of at least 10^{120} (that is a 1 followed by 120 zeros)! Various other theories have been suggested, but the bottom line is that, although there is compelling evidence that dark energy exists, we do not yet know the source of that energy.

Whatever the dark energy turns out to be, we should note that the discovery that the rate of expansion has not been constant since the beginning of the universe complicates the calculation of the age of the universe. Interestingly, the acceleration seems not to have started with the Big Bang. During the first several billion years after the Big Bang, when galaxies were close together, gravity was strong enough to slow the expansion. As galaxies moved farther apart, the effect of gravity weakened. Several billion years after the Big Bang, dark energy took over, and the expansion began to accelerate, illustrated in Figure 15.8.

Changes in the Rate of Expansion of the Universe Since Its

Beginning 13.8 Billion Years Ago

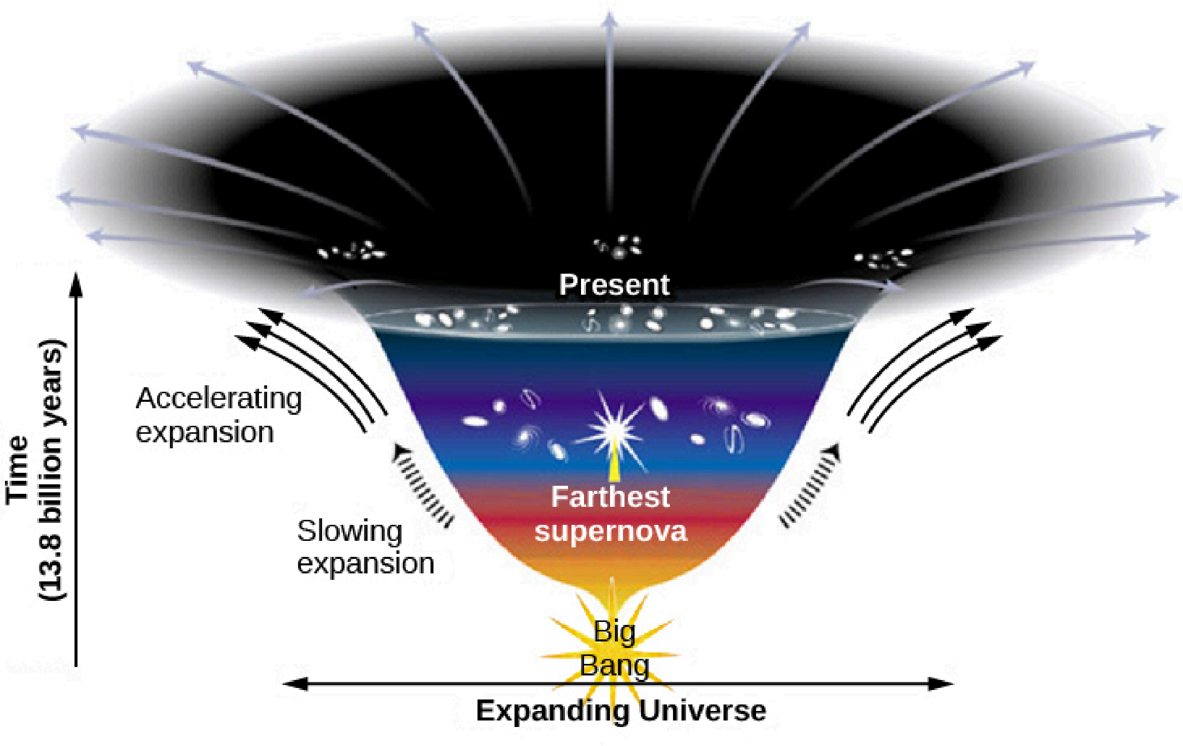


Figure 15.8. The more the diagram spreads out horizontally, the faster the change in the velocity of expansion. After a period of very rapid expansion at the beginning, which scientists call inflation and which we will discuss later in this chapter, the expansion began to decelerate. Galaxies were then close together, and their mutual gravitational attraction slowed the expansion. After a few billion years, when galaxies were farther apart, the influence of gravity began to weaken. Dark energy then took over and caused the expansion to accelerate. (credit: modification of work by Ann Feild (STScI))
The Expanding Universe by NASA, NASA Licence

Deceleration works to make the age of the universe estimated by the simple relation $T_0 = 1/H$ seem older than it really is, whereas acceleration works to make it seem younger. By happy coincidence, our best estimates of how much deceleration and acceleration occurred lead to an answer for the age very close to $T_0 = 1/H$. The best current estimate is that the universe is 13.8 billion years old with an uncertainty of only about 100 million years.

Throughout this chapter, we have referred to the Hubble *constant*. We now know that the Hubble constant does change with time. It is, however, constant everywhere in the universe at any given time. When we say the Hubble constant is about 70 kilometres/second/million parsecs, we mean that this is the value of the Hubble constant at the current time.

Comparing Ages

We now have one estimate for the age of the universe from its expansion. Is this estimate consistent with other observations? For example, are the oldest stars or other astronomical objects younger than 13.8 billion years? After all, the universe has to be at least as old as the oldest objects in it.

In our Galaxy and others, the oldest stars are found in the globular clusters as shown in Figure 15.9, which can be dated using the models of stellar evolution described in an earlier chapter.

Globular Cluster 47 Tucanae

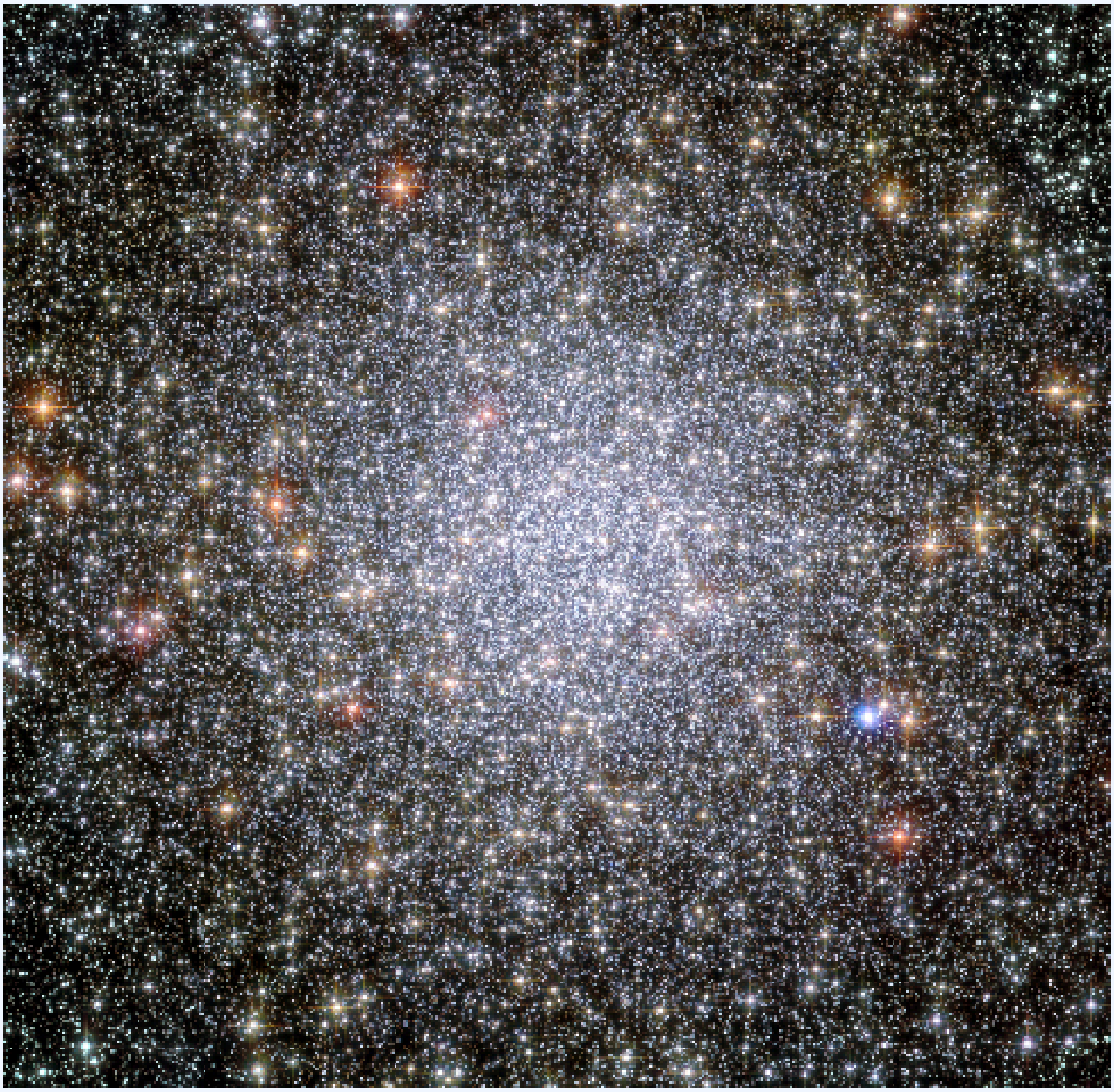


Figure 15.9. This NASA/ESA Hubble Space Telescope image shows a globular cluster known as 47 Tucanae, since it is in the constellation of Tucana (The Toucan) in the southern sky. The second-brightest globular cluster in the night sky, it includes hundreds of thousands of stars. Globular clusters are among the oldest objects in our Galaxy and can be used to estimate its age. (credit: NASA, ESA, and the Hubble Heritage (STScI/AURA)-ESA/Hubble Collaboration)
Hubble Admires a Youthful Globular Star Cluster by NASA, NASA Licence

The accuracy of the age estimates of the globular clusters has improved markedly in recent years for two reasons. First, models of interiors of globular cluster stars have been improved, mainly through better information about how atoms absorb radiation as they make their way from the centre of a star out into space. Second, observations from satellites have improved the accuracy of our measurements of the distances to these clusters. The conclusion is that the oldest stars formed about 12–13 billion years ago.

This age estimate has recently been confirmed by the study of the spectrum of uranium in the stars. The isotope uranium-238 is radioactive and decays (changes into another element) over time. (Uranium-238 gets its designation because it has 92 protons and 146 neutrons.) We know (from how stars and supernovae make elements) how much uranium-238 is generally made compared to other elements. Suppose we measure the amount of uranium relative to nonradioactive elements in a very old star and in our own Sun, and compare the abundances. With those pieces of information, we can estimate how much longer the uranium has been decaying in the very old star because we know from our own Sun how much uranium decays in 4.5 billion years.

The line of uranium is very weak and hard to make out even in the Sun, but it has now been measured in one extremely old star using the European Very Large Telescope as seen in Figure 15.10. Comparing the abundance with that in the solar system, whose age we know, astronomers estimate the star is 12.5 billion years old, with an uncertainty of about 3 billion years. While the uncertainty is large, this work is important confirmation of the ages estimated by studies of the globular cluster stars. Note that the uranium age estimate is completely independent; it does not depend on either the measurement of distances or on models of the interiors of stars.

European Extremely Large Telescope, European Very Large Telescope, and the Colosseum

Figure 15.10. The European Extremely Large Telescope (E-ELT) is currently under construction in Chile. This image compares the size of the E-ELT (left) with the four 8-meter telescopes of the European Very Large Telescope (centre) and with the Colosseum in Rome (right). The mirror of the E-ELT will be 39 metres in diameter. Astronomers are building a new generation of giant telescopes in order to observe very distant galaxies and understand what they were like when they were newly formed and the universe was young. (credit: modification of work by ESO)
ELT and VLT vs Colosseum by ESO, ESO Licence

As we shall see later in this chapter, the globular cluster stars probably did not form until the expansion of the universe had been underway for at least a few hundred million years. Accordingly, their ages are consistent with the 13.8 billion-year age estimated from the expansion rate.

Attribution

"29.1 The Age of the Universe" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

15.3 A MODEL OF THE UNIVERSE

Let's now use the results about the expansion of the universe to look at how these ideas might be applied to develop a model for the evolution of the universe as a whole. With this model, astronomers can make predictions about how the universe has evolved so far and what will happen to it in the future.

The Expanding Universe

Every model of the universe must include the expansion we observe. Another key element of the models is that the cosmological principle is valid: on the large scale, the universe at any given time is the same everywhere (homogeneous and isotropic). As a result, the expansion rate must be the same everywhere during any epoch of cosmic time. If so, we don't need to think about the entire universe when we think about the expansion, we can just look at any sufficiently large portion of it. (Some models for dark energy would allow the expansion rate to be different in different directions, and scientists are designing experiments to test this idea. However, until such evidence is found, we will assume that the cosmological principle applies throughout the universe.)

In an earlier chapter, we hinted that when we think of the expansion of the universe, it is more correct to think of space itself stretching rather than of galaxies moving through static space. Nevertheless, we have since been discussing the redshifts of galaxies as if they resulted from the motion of the galaxies themselves.

Now, however, it is time to finally put such simplistic notions behind us and take a more sophisticated look at the cosmic expansion. Recall from our discussion of Einstein's theory of general relativity that space—or, more precisely, spacetime—is not a mere backdrop to the action of the universe, as Newton thought. Rather, it is an active participant—affected by and in turn affecting the matter and energy in the universe.

Since the expansion of the universe is the stretching of all spacetime, all points in the universe are stretching together. Thus, the expansion began *everywhere at once*. Unfortunately for tourist agencies of the future, there is no location you can visit where the stretching of space began or where we can say that the Big Bang happened.

To describe just how space stretches, we say the cosmic expansion causes the universe to undergo a uniform change in *scale* over time. By scale we mean, for example, the distance between two clusters of galaxies. It is customary to represent the scale by the factor R; if R doubles, then the distance between the clusters has doubled. Since the universe is expanding at the same rate everywhere, the change in R tells us how much it has expanded (or contracted) at any given time. For a static universe, R would be constant as time passes. In an expanding universe, R increases with time.

If it is space that is stretching rather than galaxies moving through space, then why do the galaxies show redshifts in their spectra? When you were young and naïve—a few chapters ago—it was fine to discuss the redshifts of distant galaxies as if they resulted from their motion away from us. But now that you are an older and wiser student of cosmology, this view will simply not do.

A more accurate view of the redshifts of galaxies is that the light waves are stretched by the stretching of the space they travel through. Think about the light from a remote galaxy. As it moves away from its source, the light has to travel through space. If space is stretching during all the time the light is travelling, the light waves will be stretched as well. A redshift is a stretching of waves—the wavelength of each wave increases as illustrated in Figure 15.11. Light from more distant galaxies travels for more time than light from closer ones. This means that the light has stretched more than light from closer ones and thus shows a greater redshift.

Expansion and Redshift

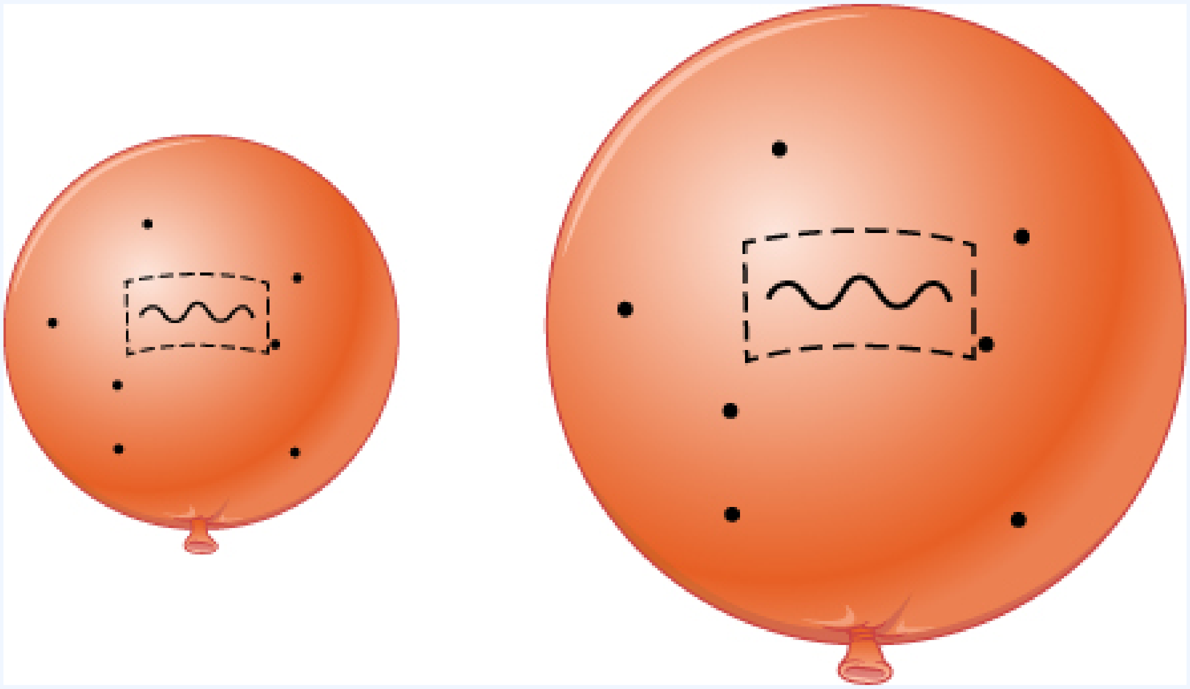


Figure 15.11. As an elastic surface expands, a wave on its surface stretches. For light waves, the increase in wavelength would be seen as a redshift.

Thus, what the measured redshift of light from an object is telling us is how much the universe has expanded since the light left the object. If the universe has expanded by a factor of 2, then the wavelength of the light (and all electromagnetic waves from the same source) will have doubled.

Models of the Expansion

Before astronomers knew about dark energy or had a good measurement of how much matter exists in the universe, they made speculative models about how the universe might evolve over time. The four possible scenarios are shown in Figure 15.12. In this diagram, time moves forward from the bottom upward, and the scale of space increases by the horizontal circles becoming wider.

Four Possible Models of the Universe

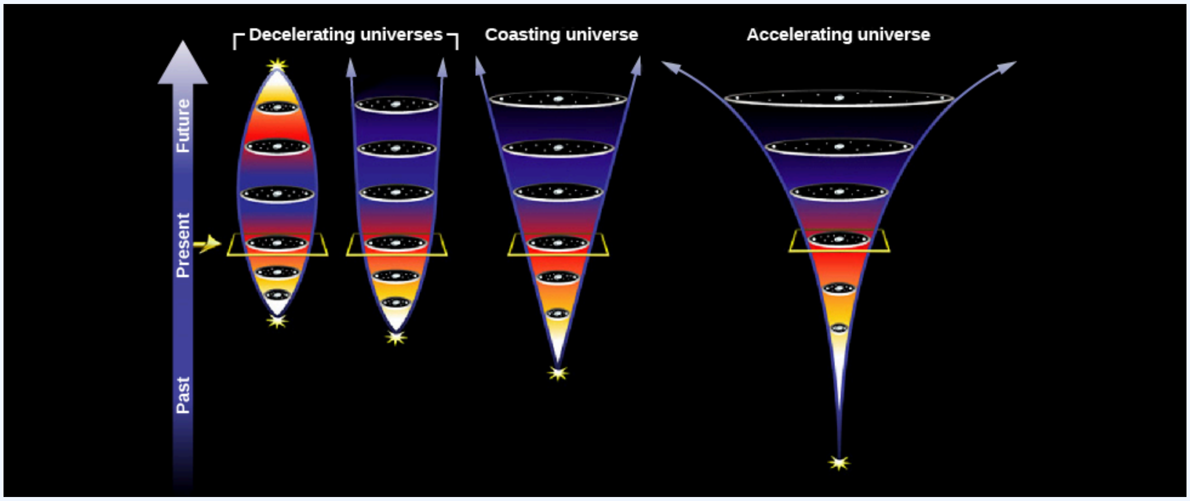


Figure 15.12. The yellow square marks the present in all four cases, and for all four, the Hubble constant is equal to the same value at the present time. Time is measured in the vertical direction. The first two universes on the left are ones in which the rate of expansion slows over time. The one on the left will eventually slow, come to a stop and reverse, ending up in a "big crunch," while the one next to it will continue to expand forever, but ever-more slowly as time passes. The "coasting" universe is one that expands at a constant rate given by the Hubble constant throughout all of cosmic time. The accelerating universe on the right will continue to expand faster and faster forever.
Possible models of the expanding Universe by ESA Hubble, Usage of ESA/Hubble images

The simplest scenario of an expanding universe would be one in which R increases with time at a constant rate. But you already know that life is not so simple. The universe contains a great deal of mass and its gravity decelerates the expansion—by a large amount if the universe contains a lot of matter, or by a negligible amount if the universe is nearly empty. Then there is the observed acceleration, which astronomers blame on a kind of dark energy.

Let's first explore the range of possibilities with models for different amounts of mass in the universe and for different contributions by dark energy. In some models—as we shall see—the universe expands forever. In others, it stops expanding and starts to contract. After looking at the extreme possibilities, we will look at recent observations that allow us to choose the most likely scenario.

We should perhaps pause for a minute to note how remarkable it is that we can do this at all. Our understanding of the principles that underlie how the universe works on the large scale and our observations of how the objects in the universe change with time allow us to model the evolution of the entire cosmos these days. It is one of the loftiest achievements of the human mind.

What astronomers look at in practice, to determine the kind of universe we live in, is the **average density** of the universe. This is the mass of matter (including the equivalent mass of energy)[1] that would be contained in each unit of volume (say, 1 cubic centimetre) if all the stars, galaxies, and other objects were taken apart, atom by atom, and if all those particles, along with the light and other energy, were distributed throughout all of space with absolute uniformity. If the average density is low, there is less mass and less gravity, and the universe will not decelerate very much. It can therefore expand forever. Higher average density, on the other hand, means there is more mass and more gravity and that the stretching of space might slow down enough that the expansion will eventually stop. An extremely high density might even cause the universe to collapse again.

For a given rate of expansion, there is a **critical density**—the mass per unit volume that will be just enough to slow the expansion to zero at some time infinitely far in the future. If the actual density is higher than this critical density, then the expansion will ultimately reverse and the universe will begin to contract. If the actual density is lower, then the universe will expand forever.

These various possibilities are illustrated in Figure 15.13. In this graph, one of the most comprehensive in all of science, we chart the development of the scale of space in the cosmos against the passage of time. Time increases to the right, and the scale of the universe, R, increases upward in the figure. Today, at the point marked "present" along the time axis, R is increasing in each model. We know that the galaxies are currently expanding away from each other, no matter which model is right. (The same situation holds for a baseball thrown high into the air. While it may eventually fall back down, near the beginning of the throw it moves upward most rapidly.)

The various lines moving across the graph correspond to different models of the universe. The straight dashed line corresponds to the empty universe with no deceleration; it intercepts the time axis at a time, T_0 (the Hubble time), in the past. This is not a realistic model but gives us a measure to compare other models to. The curves below the dashed line represent models with no dark energy and with varying amounts of deceleration, starting from the Big Bang at shorter times in the past. The curve above the dashed line shows what happens if the expansion is accelerating. Let's take a closer look at the future according to the different models.

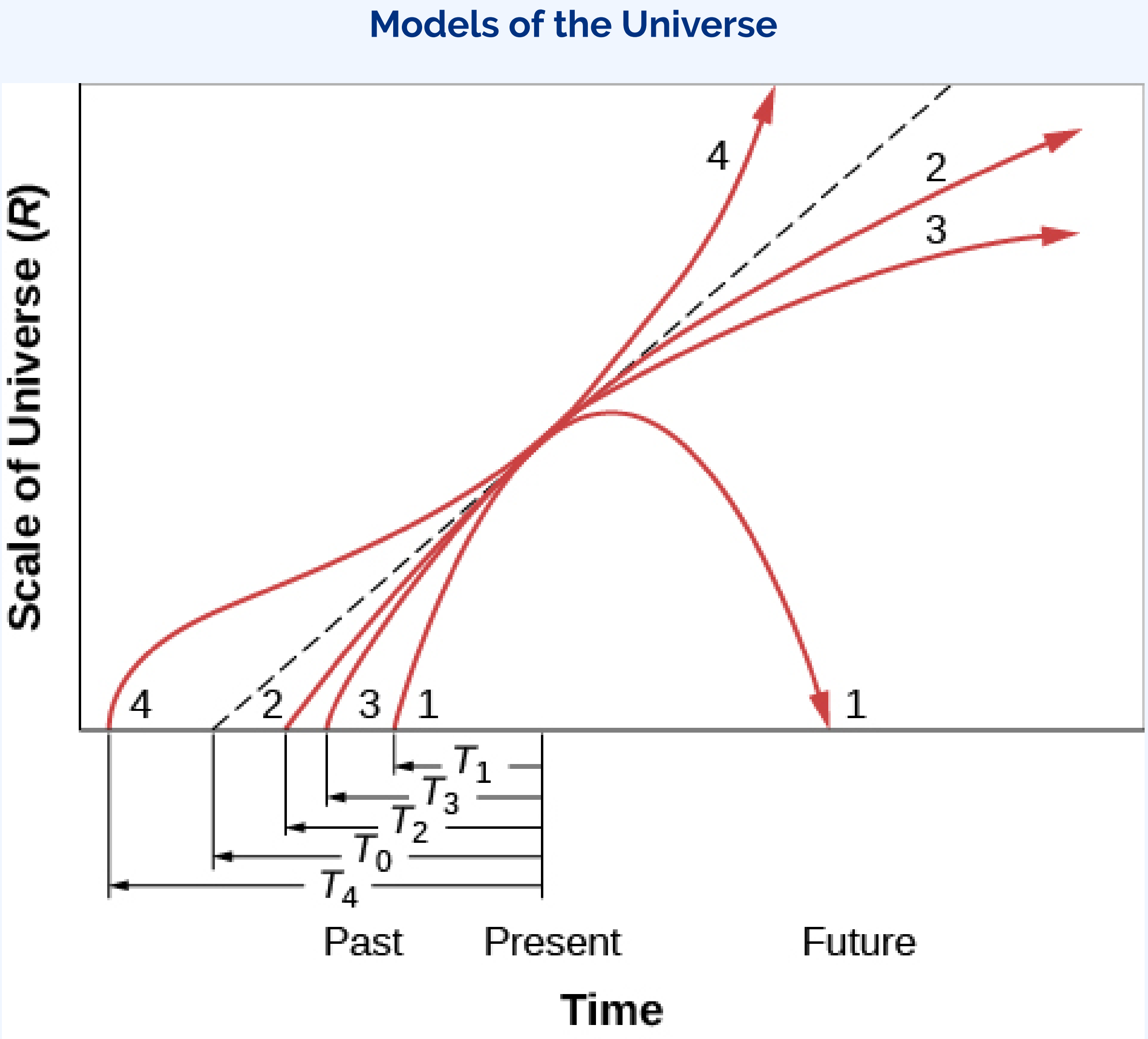


Figure 15.13. This graph plots R, the scale of the universe, against time for various cosmological models. Curve 1 represents a universe where the density is greater than the critical value; this model predicts that the universe will eventually collapse. Curve 2 represents a universe with a density lower than critical; the universe will continue to expand but at an ever-slower rate. Curve 3 is a critical-density universe; in this universe, the expansion will gradually slow to a stop infinitely far in the future. Curve 4 represents a universe that is accelerating because of the effects of dark energy. The dashed line is for an empty universe, one in which the expansion is not slowed by gravity or accelerated by dark energy. Time is very compressed on this graph.

Let's start with curve 1 in Figure 15.13. In this case, the actual density of the universe is higher than the critical density and there is no dark energy. This universe will stop expanding at some time in the future and begin contracting. This model is called a **closed universe** and corresponds to the universe on the left in Figure 15.12.

Eventually, the scale drops to zero, which means that space will have shrunk to an infinitely small size. The noted physicist John Wheeler called this the "big crunch," because matter, energy, space, and time would all be crushed out of existence. Note that the "big crunch" is the opposite of the Big Bang—it is an *implosion*. The universe is not expanding but rather collapsing in upon itself.

Some scientists speculated that another Big Bang might follow the crunch, giving rise to a new expansion phase, and then another contraction—perhaps oscillating between successive Big Bangs and big crunches indefinitely in the past and future. Such speculation was sometimes referred to as the oscillating theory of the universe. The challenge for theorists was how to describe the transition from collapse (when space and time themselves disappear into the big crunch) to expansion. With the discovery of dark energy, however, it does not appear that the universe will experience a big crunch, so we can put worrying about it on the back burner.

If the density of the universe is less than the critical density (curve 2 in Figure 15.13 and the universe second from the left in Figure 15.12), gravity is never important enough to stop the expansion, and so the universe expands forever. Such a universe is infinite and this model is called an **open universe**. Time and space begin with the Big Bang, but they have no end; the universe simply continues expanding, always a bit more slowly as time goes on. Groups of galaxies eventually get so far apart that it would be difficult for observers in any of them to see the others.

At the critical density (curve 3), the universe can just barely expand forever. The critical-density universe has an age of exactly two-thirds T_0, where T_0 is the age of the empty universe. Universes that will someday begin to contract have ages less than two-thirds T_0.

In an empty universe (the dashed line Figure 15.13 and the coasting universe in Figure 15.12), neither gravity nor dark energy is important enough to affect the expansion rate, which is therefore constant throughout all time.

In a universe with dark energy, the rate of the expansion will increase with time, and the expansion will continue at an ever-faster rate. Curve 4 in Figure 15.13, which represents this universe, has a complicated shape. In the beginning, when the matter is all very close together, the rate of expansion is most influenced by gravity. Dark energy appears to act only over large scales and thus becomes more important as the universe grows larger and the matter begins to thin out. In this model, at first the universe slows down, but as space stretches, the acceleration plays a greater role and the expansion speeds up.

The Cosmic Tug of War

We might summarize our discussion so far by saying that a "tug of war" is going on in the universe between the forces that push everything apart and the gravitational attraction of matter, which pulls everything together. If we can determine who will win this tug of war, we will learn the ultimate fate of the universe.

The first thing we need to know is the density of the universe. Is it greater than, less than, or equal to the critical density? The critical density today depends on the value of the expansion rate today, H_0. If the Hubble constant is around 20 kilometres/second per million light-years, the critical density is about 10^{-26} kg/m^3. Let's see how this value compares with the actual density of the universe.

We can start our survey of how dense the cosmos is by ignoring the dark energy and just estimating the density of all matter in the universe, including ordinary matter and dark matter. Here is where the cosmological principle really comes in handy. Since the universe is the same all over (at least on large scales), we only need to measure how much matter exists in a (large) representative sample of it. This is similar to the way a representative survey of a few thousand people can tell us whom the millions of residents of the US prefer for president.

There are several methods by which we can try to determine the average density of matter in space. One way is to count all the galaxies out to a given distance and use estimates of their masses, including dark matter, to calculate the average density. Such estimates indicate a density of about 1 to 2 × 10^{-27} kg/m^3 (10 to 20% of critical), which by itself is too small to stop the expansion.

A lot of the dark matter lies outside the boundaries of galaxies, so this inventory is not yet complete. But even if we add an estimate of the dark matter outside galaxies, our total won't rise beyond about 30% of the critical density. We'll pin these numbers down more precisely later in this chapter, where we will also include the effects of dark energy.

In any case, even if we ignore dark energy, the evidence is that the universe will continue to expand forever. The discovery of dark energy that is causing the rate of expansion to speed up only strengthens this conclusion. Things definitely do not look good for fans of the closed universe (big crunch) model.

What Might the Universe Be Like in the Distant Future?

> Some say the world will end in fire,
> Some say in ice.
> From what I've tasted of desire
> I hold with those who favor fire.
> —From the poem "Fire and Ice" by Robert Frost (1923)

Given the destructive power of impacting asteroids, expanding red giants, and nearby supernovae, our species may not be around in the remote future. Nevertheless, you might enjoy speculating about what it would be like to live in a much, much older universe.

The observed acceleration makes it likely that we will have continued expansion into the indefinite future. If the universe expands forever (*R* increases without limit), the clusters of galaxies will spread ever farther apart with time. As eons pass, the universe will get thinner, colder, and darker.

Within each galaxy, stars will continue to go through their lives, eventually becoming white dwarfs, neutron stars, and black holes. Low-mass stars might take a long time to finish their evolution, but in this model, we would literally have all the time in the world. Ultimately, even the white dwarfs will cool down to be black dwarfs, any neutron stars that reveal themselves as pulsars will slowly stop spinning, and black holes with accretion disks will one day complete their "meals." The remains of stars will all be dark and difficult to observe.

This means that the light that now reveals galaxies to us will eventually go out. Even if a small pocket of raw material were left in one unsung corner of a galaxy, ready to be turned into a fresh cluster of stars, we will only have to wait until the time that their evolution is also complete. And time is one thing this model of the universe has plenty of. There will surely come a time when all the stars are out, galaxies are as dark as space, and no source of heat remains to help living things survive. Then the lifeless galaxies will just continue to move apart in their lightless realm.

If this view of the future seems discouraging (from a human perspective), keep in mind that we fundamentally do not understand why the expansion rate is currently accelerating. Thus, our speculations about the future are just that: speculations. You might take heart in the knowledge that science is always a progress report. The most advanced ideas about the universe from a hundred years ago now strike us as rather primitive. It may well be that our best models of today will in a hundred or a thousand years also seem rather simplistic and that there are other factors determining the ultimate fate of the universe of which we are still completely unaware.

Ages of Distant Galaxies

In an earlier chapter, we discussed how we can use Hubble's law to measure the distance to a galaxy. But that simple method only works with galaxies that are not too far away. Once we get to large distances, we are looking so far into the past that we must take into account changes in the rate of the expansion of the universe. Since we cannot measure these changes directly, we must assume one of the models of the universe to be able to convert large redshifts into distances.

This is why astronomers squirm when reporters and students ask them exactly how far away some newly discovered distant quasar or galaxy is. We really can't give an answer without first explaining the model of the universe we are assuming in calculating it (by which time a reporter or student is long gone or asleep).

Specifically, we must use a model that includes the change in the expansion rate with time. The key ingredients of the model are the amounts of matter, including dark matter, and the equivalent mass (according to $E = mc^2$) of the dark energy along with the Hubble constant.

Elsewhere in this book, we have estimated the mass density of ordinary matter plus dark matter as roughly 0.3 times the critical density, and the mass equivalent of dark energy as roughly 0.7 times the critical density. We will refer to these values as the "standard model of the universe." The latest (slightly improved) estimates for these values and the evidence for them will be given later in this chapter. Calculations also require the current value of the Hubble constant. For Table 15.1, we have adopted a Hubble constant of 67.3 kilometres/second/million parsecs (rather than rounding it to 70 kilometres/second/million parsecs), which is consistent with the 13.8 billion-year age of the universe estimated by the latest observations.

Once we assume a model, we can use it to calculate the age of the universe at the time an object emitted the light we see. As an example, Table 15.1 lists the times that light was emitted by objects at different redshifts as fractions of the current age of the universe. The times are given for two very different models so you can get a feeling for the fact that the calculated ages are fairly similar. The first model assumes that the universe has a critical density of matter and no dark energy. The second model is the standard model described in the preceding paragraph. The first column in the table is the redshift, which is given by the equation $z = \frac{\Delta\lambda}{\lambda_0}$ and is a measure of how much the wavelength of light has been stretched by the expansion of the universe on its long journey to us.

Table 15.1. Ages of the Universe at Different Redshifts

Redshift	Percent of Current Age of Universe When the Light Was Emitted (mass = critical density)	Percent of Current Age of Universe When the Light Was Emitted (mass = 0.3 critical density; dark energy = 0.7 critical density)
0	100 (now)	100 (now)
0.5	54	63
1.0	35	43
2.0	19	24
3.0	13	16
4.0	9	11
5.0	7	9
8.0	4	5
11.9	0.02	0.027
Infinite	0	0

Notice that as we find objects with higher and higher redshifts, we are looking back to smaller and smaller fractions of the age of the universe. The highest observed redshifts as this book is being written are close to 12 as shown in Figure 15.14. As Table 15.1 shows, we are seeing these galaxies as they were when the universe was only about 3% as old as it is now. They were already formed only about 700 million years after the Big Bang.

Hubble Ultra-Deep Field

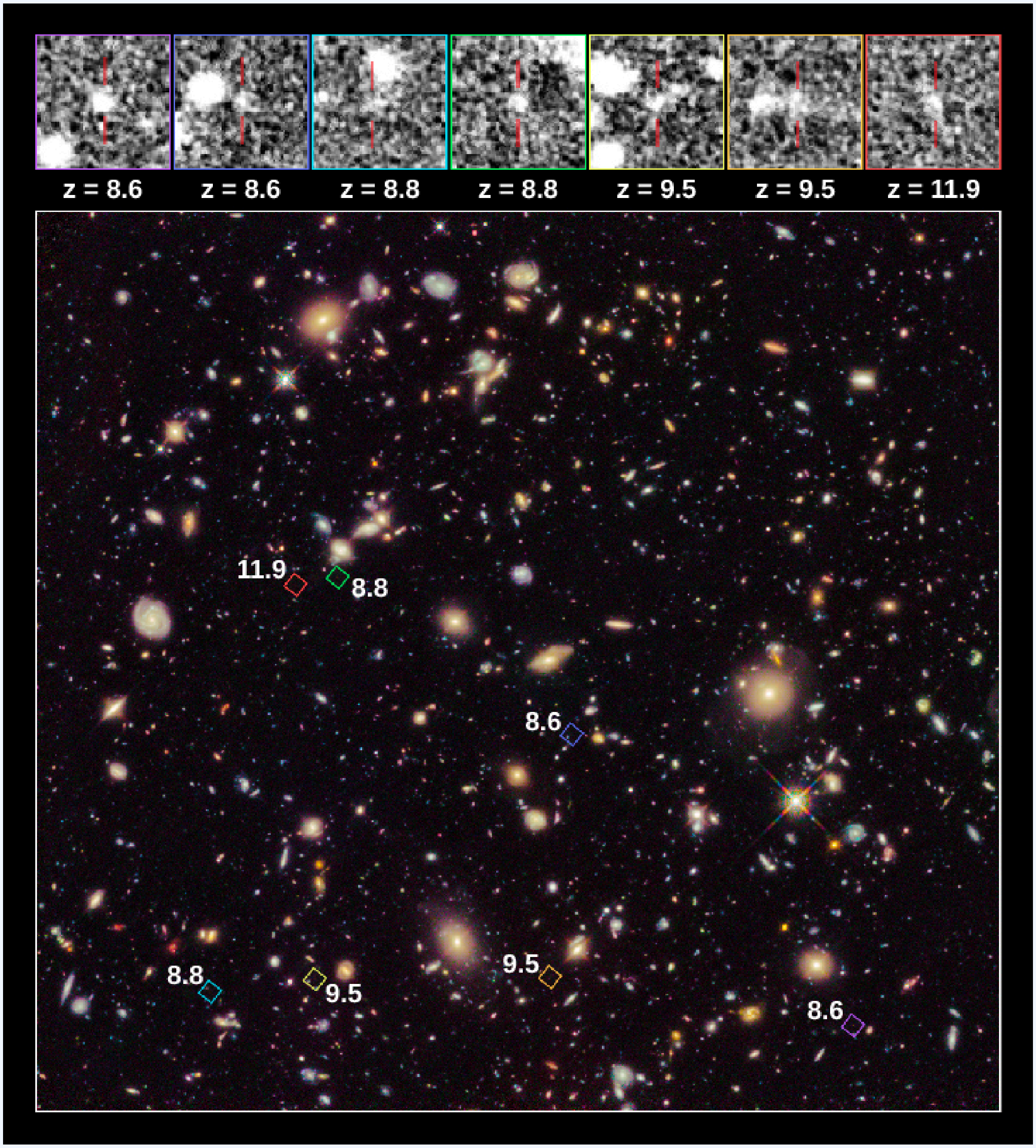


Figure 15.14. This image, called the Hubble Ultra Deep Field, shows faint galaxies, seen very far away and therefore very far back in time. The coloured squares in the main image outline the locations of the galaxies. Enlarged views of each galaxy are shown in the black-and-white images. The red lines mark each galaxy's location. The "redshift" of each galaxy is indicated below each box, denoted by the symbol "z." The redshift measures how much a galaxy's ultraviolet and visible light has been stretched to infrared wavelengths by the universe's expansion. The larger the redshift, the more distant the galaxy, and therefore the further astronomers are seeing back in time. One of the seven galaxies may be a distance

breaker, observed at a redshift of 11.9. If this redshift is confirmed by additional measurements, the galaxy is seen as it appeared only 380 million years after the Big Bang, when the universe was less than 3% of its present age. (credit: modification of work by NASA, ESA, R. Ellis (Caltech), and the UDF 2012 Team)
NASA'S Hubble Provides First Census of Galaxies Near Cosmic Dawn by NASA, NASA Licence

---

Attribution

"29.2 A Model of the Universe" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

15.4 THE COSMIC MICROWAVE BACKGROUND

If the model of the universe described in the previous section is correct, then—as we look far outward in the universe and thus far back in time—the first "afterglow" of the hot, early universe should still be detectable. Observations of it would be very strong evidence that our theoretical calculations about how the universe evolved are correct. As we shall see, we have indeed detected the radiation emitted at this photon decoupling time, when radiation began to stream freely through the universe without interacting with matter.

Cosmic Microwave Background and Clouds Compared

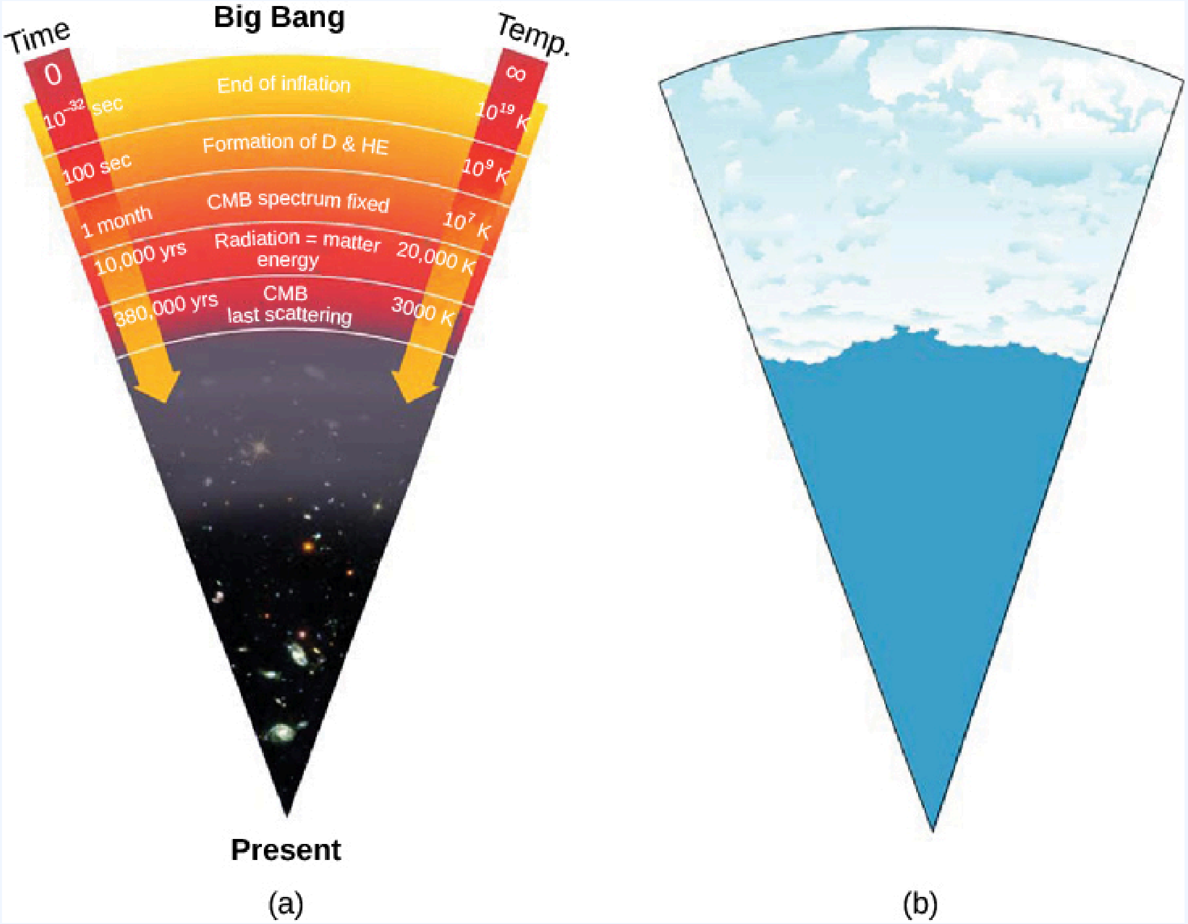


Figure 15.15. (a) Early in the universe, photons (electromagnetic energy) were scattering off the crowded, hot, charged particles and could not get very far without colliding with another particle. But after electrons and photons settled into neutral atoms, there was far less scattering, and photons could travel over vast distances. The universe became transparent. As we look out in space and back in time, we can't see back beyond this time. (b) This is similar to what happens when we see clouds in Earth's atmosphere. Water droplets in a cloud scatter light very efficiently, but clear air lets light travel over long distances. So as we look up into the atmosphere, our vision is blocked by the cloud layers and we can't see beyond them. SURFACE OF LAST SCATTERING by NASA, NASA Licence

The detection of this afterglow was initially an accident. In the late 1940s, Ralph Alpher and Robert Herman, working with George Gamow, realized that just before the universe became transparent, it must have been radiating like a blackbody at a temperature of about 3000 K—the temperature at which hydrogen atoms could begin to form. If we could have seen that radiation just after neutral atoms formed, it would have resembled radiation from a reddish star. It was as if a giant fireball filled the whole universe.

But that was nearly 14 billion years ago, and, in the meantime, the scale of the universe has increased a thousand fold. This expansion has increased the wavelength of the radiation by a factor of 1000. According to Wien's law, which relates wavelength and temperature, the expansion has correspondingly lowered the temperature by a factor of 1000.

Alpher and Herman predicted that the glow from the fireball should now be at radio wavelengths and should resemble the radiation from a blackbody at a temperature only a few degrees above absolute zero. Since the fireball was everywhere throughout the universe, the radiation left over from it should also be everywhere. If our eyes were sensitive to radio wavelengths, the whole sky would appear to glow very faintly. However, our eyes can't see at these wavelengths, and at the time Alpher and Herman made their prediction, there were no instruments that could detect the glow. Over the years, their prediction was forgotten.

In the mid-1960s, in Holmdel, New Jersey, Arno Penzias and Robert Wilson of AT&T's Bell Laboratories had built a delicate microwave antenna to measure astronomical sources, including supernova remnants like Cassiopeia A. They were plagued with some unexpected background noise, just like faint static on a radio, which they could not get rid of. The puzzling thing about this radiation was that it seemed to be coming from all directions at once. This is very unusual in astronomy: after all, most radiation has a specific direction where it is strongest—the direction of the Sun, or a supernova remnant, or the disk of the Milky Way, for example.

Robert Wilson (left) and Arno Penzias (right)

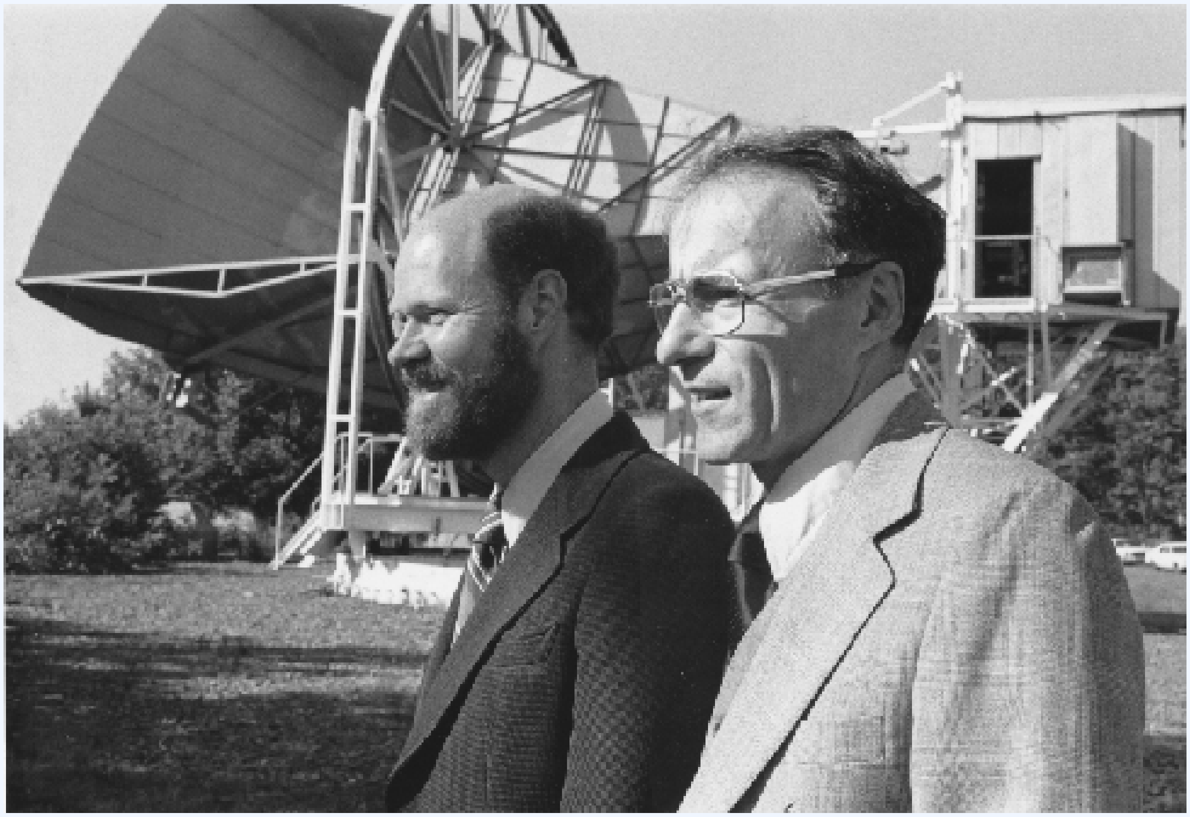


Figure 15.16. These two scientists are standing in front of the horn-shaped antenna with which they discovered the cosmic background radiation. The photo was taken in 1978, just after they received the Nobel Prize in physics.

Penzias and Wilson at first thought that any radiation appearing to come from all directions must originate from inside their telescope, so they took everything apart to look for the source of the noise. They even found that some pigeons had roosted inside the big horn-shaped antenna and had left (as Penzias delicately put it) "a layer of white, sticky, dielectric substance coating the inside of the antenna." However, nothing the scientists did could reduce the background radiation to zero, and they reluctantly came to accept that it must be real, and it must be coming from space.

Penzias and Wilson were not cosmologists, but as they began to discuss their puzzling discovery with other scientists, they were quickly put in touch with a group of astronomers and physicists at Princeton University (a short drive away). These astronomers had—as it happened—been redoing the calculations of Alpher and Herman from the 1940s and also realized that the radiation from the decoupling time should be detectable as a faint afterglow of radio waves. The different calculations of what the observed temperature would be for this cosmic microwave background (CMB)[1] were uncertain, but all predicted less than 40 K.

Penzias and Wilson found the distribution of intensity at different radio wavelengths to correspond to a temperature of 3.5 K. This is very cold—closer to absolute zero than most other astronomical measurements—and a testament to how much space (and the waves within it) has stretched. Their measurements have been repeated with better instruments, which give us a reading of 2.73 K. So Penzias and Wilson came very close. Rounding this value, scientists often refer to “the 3-degree microwave background.”

Many other experiments on Earth and in space soon confirmed the discovery by Penzias and Wilson: The radiation was indeed coming from all directions (it was isotropic) and matched the predictions of the Big Bang theory with remarkable precision. Penzias and Wilson had inadvertently observed the glow from the primeval fireball. They received the Nobel Prize for their work in 1978. And just before his death in 1966, Lemaître learned that his “vanished brilliance” had been discovered and confirmed.

One issue that worried astronomers is that Penzias and Wilson were measuring the background radiation filling space through Earth’s atmosphere. What if that atmosphere is a source of radio waves or somehow affected their measurements? It would be better to measure something this important from space.

The first accurate measurements of the CMB were made with a satellite orbiting Earth. Named the Cosmic Background Explorer (COBE), it was launched by NASA in November 1989. The data it received quickly showed that the CMB closely matches that expected from a blackbody with a temperature of 2.73 K. This is exactly the result expected if the CMB was indeed redshifted radiation emitted by a hot gas that filled all of space shortly after the universe began.

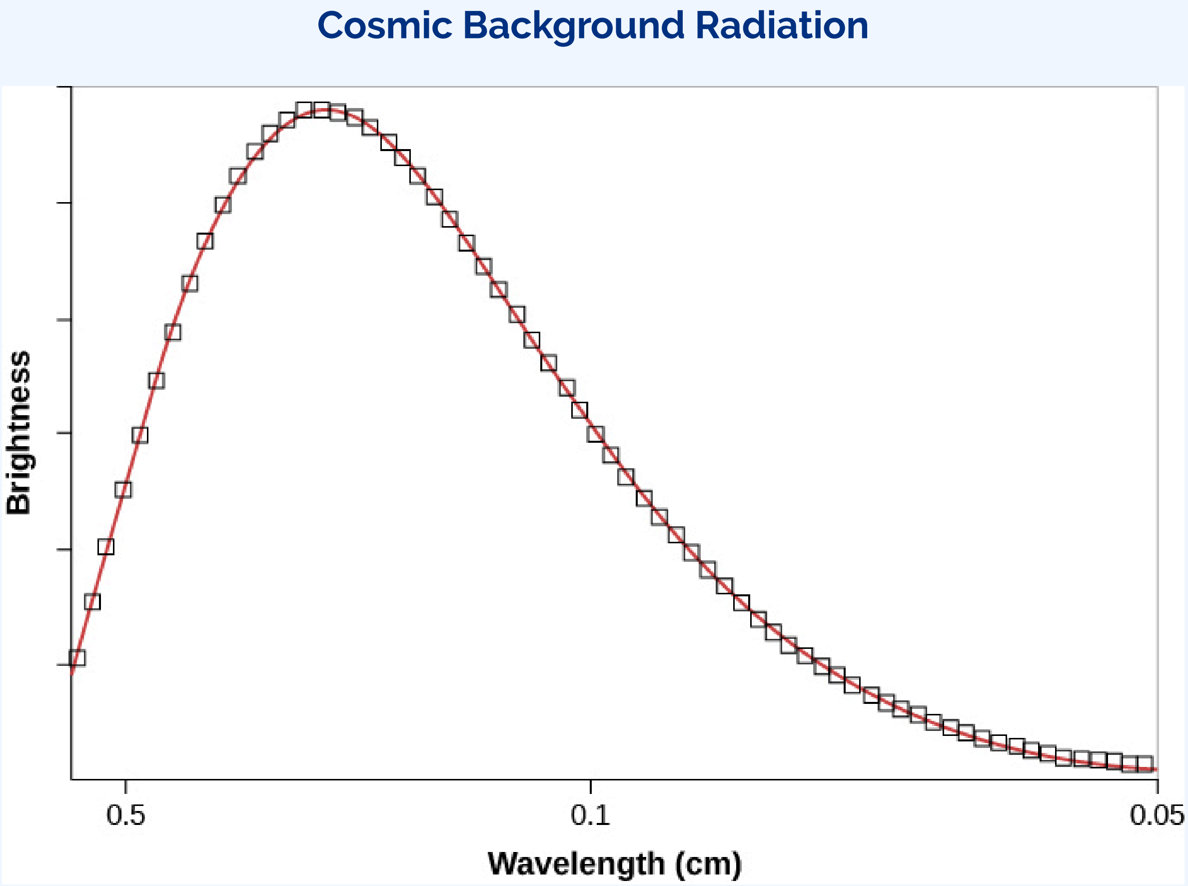


Figure 15.17. The solid line shows how the intensity of radiation should change with wavelength for a blackbody with a temperature of 2.73 K. The boxes show the intensity of the cosmic background radiation as measured at various wavelengths by COBE's instruments. The fit is perfect. When this graph was first shown at a meeting of astronomers, they gave it a standing ovation.
Intensity of radiation change by Wenzhong Zhang, CC BY 4.0

The first important conclusion from measurements of the CMB, therefore, is that the universe we have today has indeed evolved from a hot, uniform state. This observation also provides direct support for the general idea that we live in an evolving universe, since the universe is cooler today than it was in the beginning.

It was known even before the launch of COBE that the CMB is extremely *isotropic*. In fact, its uniformity in every direction is one of the best confirmations of the cosmological principle— that the universe is homogenous and isotropic.

According to our theories, however, the temperature could not have been *perfectly* uniform when the CMB was emitted. After all, the CMB is radiation that was scattered from the particles in the universe at the time of decoupling. If the radiation were completely smooth, then all those particles must have been distributed through space absolutely evenly. Yet it is those particles that have become all the galaxies and stars (and astronomy students) that now inhabit the cosmos. Had the particles been completely smoothly distributed, they could not have formed all the large-scale structures now present in the universe—the clusters and superclusters of galaxies discussed in the last few chapters.

The early universe must have had tiny density fluctuations from which such structures could evolve. Regions of higher-than-average density would have attracted additional matter and eventually grown into the galaxies and clusters that we see today. It turned out that these denser regions would appear to us to be colder spots, that is, they would have lower-than-average temperatures.

The reason that temperature and density are related can be explained this way. At the time of decoupling, photons in a slightly denser portion of space had to expend some of their energy to escape the gravitational force exerted by the surrounding gas. In losing energy, the photons became slightly colder than the overall average temperature at the time of decoupling. Vice versa, photons that were located in a slightly less dense portion of space lost less energy upon leaving it than other photons, thus appearing slightly hotter than average. Therefore, if the seeds of present-day galaxies existed at the time that the CMB was emitted, we should see some slight variations in the CMB temperature as we look in different directions in the sky.

Scientists working with the data from the COBE satellite did indeed detect very subtle temperature differences—about 1 part in 100,000—in the CMB. The regions of lower-than-average temperature come in a variety of sizes, but even the smallest of the colder areas detected by COBE is far too large to be the precursor of an individual galaxy, or even a supercluster of galaxies. This is because the COBE instrument had "blurry vision" (poor resolution) and could only measure large patches of the sky. We needed instruments with "sharper vision."

The most detailed measurements of the CMB have been obtained by two satellites launched more recently than COBE. The results from the first of these satellites, the Wilkinson Microwave Anisotropy Probe (WMAP) spacecraft, were published in 2003. In 2015, measurements from the Planck satellite extended the WMAP measurements to even-higher spatial resolution and lower noise.

CMB Observations

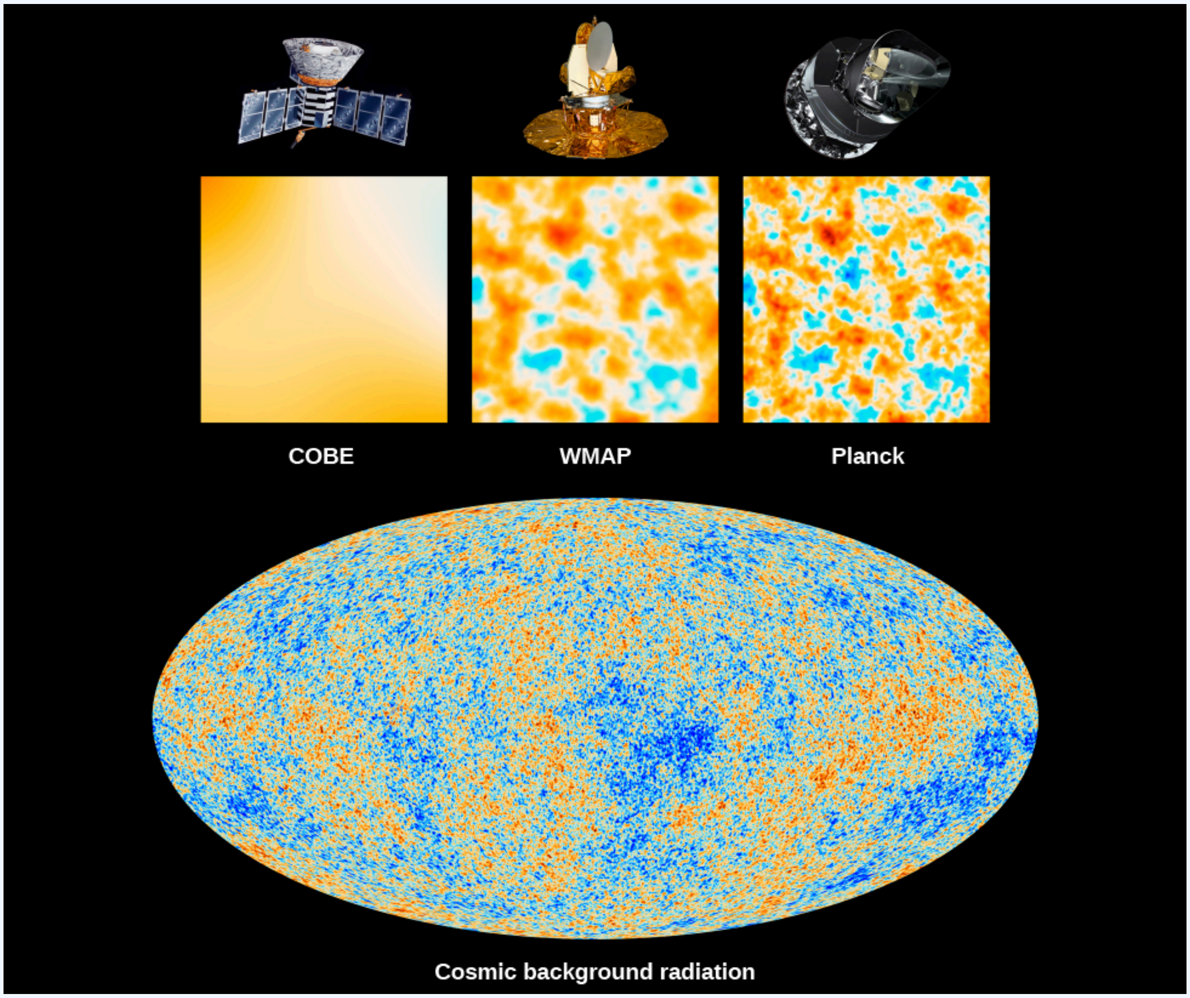


Figure 15.18. This comparison shows how much detail can be seen in the observations of three satellites used to measure the CMB. The CMB is a snapshot of the oldest light in our universe, imprinted on the sky when the universe was just about 380,000 years old. The first spacecraft, launched in 1989, is NASA's Cosmic Background Explorer, or COBE. WMAP was launched in 2001, and Planck was launched in 2009. The three panels show 10-square-degree patches of all-sky maps. This cosmic background radiation image (bottom) is an all-sky map of the CMB as observed by the Planck mission. The colours in the map represent different temperatures: red for warmer and blue for cooler. These tiny temperature fluctuations correspond to regions of slightly different densities, representing the seeds of all future structures: the stars, galaxies, and galaxy clusters of today.
Planck Mission Brings Universe Into Sharp Focus by JPL, JPL Image Use Policy

Theoretical calculations show that the sizes of the hot and cold spots in the CMB depend on the geometry of the universe and hence on its total density. (It's not at all obvious that it should do so, and it takes some pretty fancy calculations—way beyond the level of our text—to make the connection, but having such a dependence is very useful.) The total density we are discussing here includes both the amount of mass in the universe and the mass equivalent of the dark energy. That is, we must add together mass and energy: ordinary matter, dark matter, and the dark energy that is speeding up the expansion.

To see why this works, remember that with his theory of general relativity, Einstein showed that matter can curve space and that the amount of curvature depends on the amount of matter present. Therefore, the total amount of matter in the universe (including dark matter and the equivalent matter contribution by dark energy), determines the overall geometry of space. Just like the geometry of space around a black hole has a curvature to it, so the entire universe may have a curvature. Let's take a look at the possibilities.

If the density of matter is higher than the critical density, the universe will eventually collapse. In such a closed universe, two initially parallel rays of light will eventually meet. This kind of geometry is referred to as spherical geometry. If the density of matter is less than critical, the universe will expand forever. Two initially parallel rays of light will diverge, and this is referred to as hyperbolic geometry. In a critical-density universe, two parallel light rays never meet, and the expansion comes to a halt only at some time infinitely far in the future. We refer to this as a flat universe, and the kind of Euclidean geometry you learned in high school applies in this type of universe.

Picturing Space Curvature for the Entire Universe

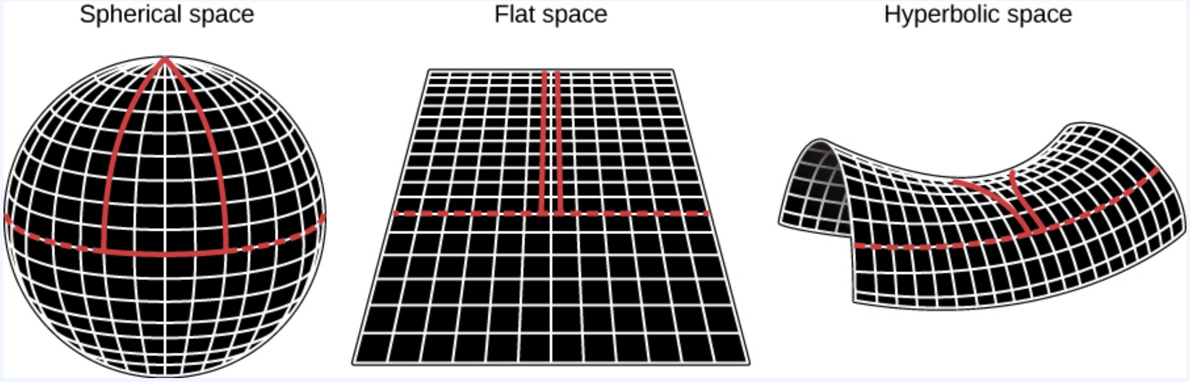


Figure 15.19. The density of matter and energy determines the overall geometry of space. If the density of the universe is greater than the critical density, then the universe will ultimately collapse and space is said to be closed like the surface of a sphere. If the density exactly equals the critical density, then space is flat like a sheet of paper; the universe will expand forever, with the rate of expansion coming to a halt infinitely far in the future. If the density is less than critical, then the expansion will continue forever and space is said to be open and negatively curved like the surface of a saddle (where more space than you expect opens up as you move farther away). Note that the red lines in each diagram show what happens in each kind of space—they are initially parallel but follow different paths depending on the curvature of space. Remember that these drawings are trying to show how space for the entire universe is "warped"—this can't be seen locally in the small amount of space that we humans occupy.

If the density of the universe is equal to the critical density, then the hot and cold spots in the CMB should typically be about a degree in size. If the density is greater than critical, then the typical sizes will be larger than one degree. If the universe has a density less than critical, then the structures will appear smaller. In Figure 15.20, you can see the differences easily. WMAP and Planck observations of the CMB confirmed earlier experiments that we do indeed live in a flat, critical-density universe.

Comparison of CMB Observations with Possible Models of the Universe

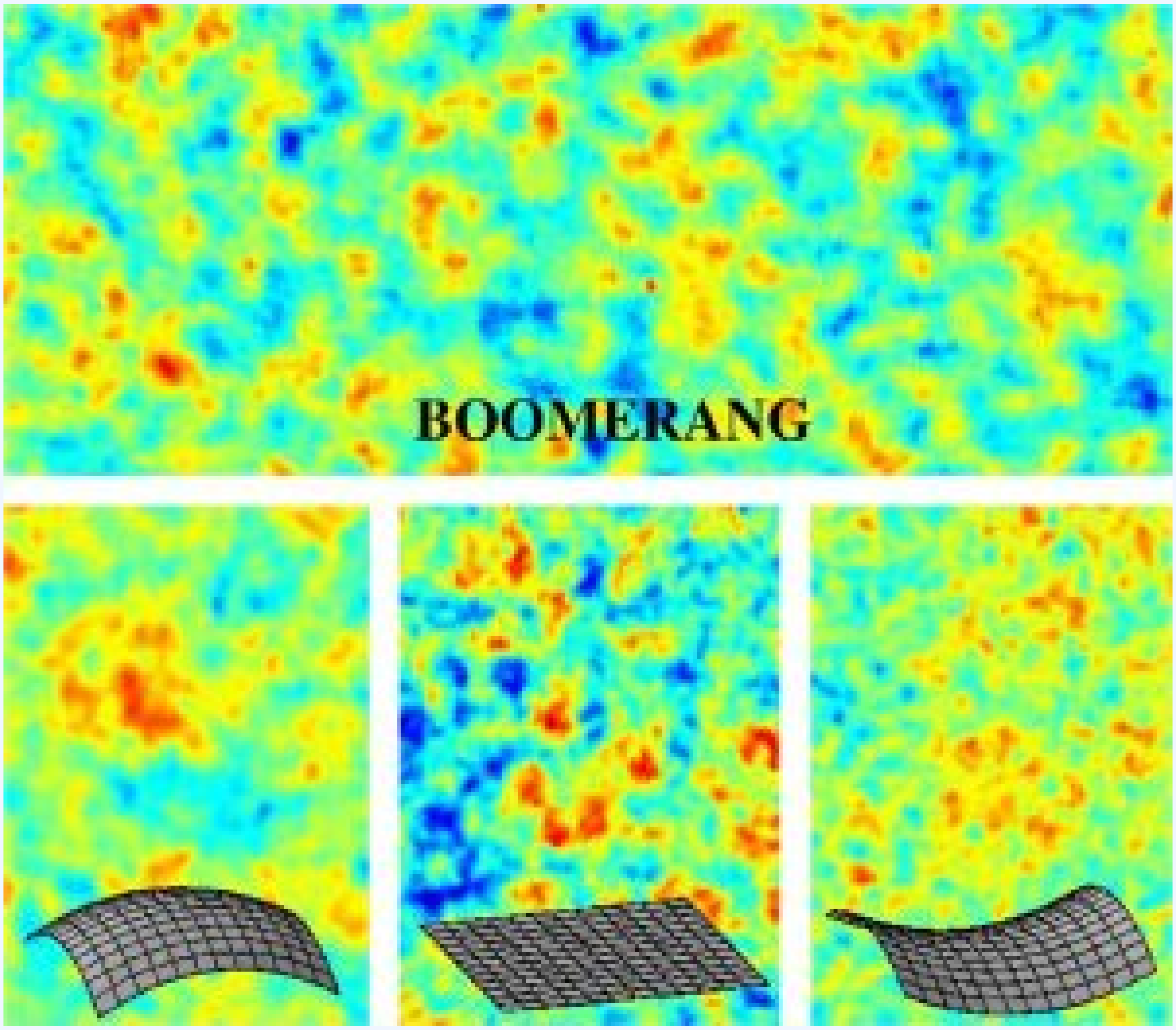


Figure 15.20. Cosmological simulations predict that if our universe has critical density, then the CMB images will be dominated by hot and cold spots of around one degree in size (bottom centre). If, on the other hand, the density is higher than critical (and the universe will ultimately collapse), then the images' hot and cold spots will appear larger than one degree (bottom left). If the density of the universe is less than critical (and the expansion will continue forever), then the structures will appear smaller (bottom right). As the measurements show, the universe is at critical density. The measurements shown were made by a balloon-borne instrument called BOOMERanG (Balloon Observations of Millimetric Extragalactic Radiation and Geophysics), which was flown in Antarctica. Subsequent satellite observations by WMAP and Planck confirm the BOOMERanG result.
Maps of the CMB by Janelle M. Bailey , CC BY 4.0

Key numbers from an analysis of the Planck data give us the best values currently available for some of the basic properties of the universe:

- Age of universe: 13.799 ± 0.038 billion years (Note: That means we know the age of the universe to within 38 million years. Amazing!)
- Hubble constant: 67.31 ± 0.96 kilometres/second/million parsecs
- Fraction of universe's content that is "dark energy": 68.5% ± 1.3%
- Fraction of the universe's content that is matter: 31.5% ± 1.3%

Note that this value for the Hubble constant is slightly smaller than the value of 70 kilometres/second/million parsecs that we have adopted in this book. In fact, the value derived from measurements of redshifts is 73 kilometres/second/million parsecs. So precise is modern cosmology these days that scientists are working hard to resolve this discrepancy. The fact that the difference between these two independent measurements is so small is actually a remarkable achievement. Only a few decades ago, astronomers were arguing about whether the Hubble constant was around 50 kilometres/second/million parsecs or 100 kilometres/second/million parsecs.

Analysis of Planck data also shows that ordinary matter (mainly protons and neutrons) makes up 4.9% of the total density. Dark matter plus normal matter add up to 31.5% of the total density. Dark energy contributes the remaining 68.5%. The age of the universe at decoupling—that is, when the CMB was emitted—was 380,000 years.

Perhaps the most surprising result from the high-precision measurements by WMAP and the even higher-precision measurements from Planck is that there were no surprises. The model of cosmology with ordinary matter at about 5%, dark matter at about 25%, and dark energy about 70% has survived since the late 1990s when cosmologists were forced in that direction by the supernovae data. In other words, the very strange universe that we have been describing, with only about 5% of its contents being made up of the kinds of matter we are familiar with here on Earth, really seems to be the universe we live in.

After the CMB was emitted, the universe continued to expand and cool off. By 400 to 500 million years after the Big Bang, the very first stars and galaxies had already formed. Deep in the interiors of stars, matter was reheated, nuclear reactions were ignited, and the more gradual synthesis of the heavier elements that we have discussed throughout this book began.

We conclude this quick tour of our model of the early universe with a reminder. You must not think of the Big Bang as a *localized* explosion *in space*, like an exploding superstar. There were no boundaries and there was no single site where the explosion happened. It was an explosion *of space* (and time and matter and energy) that happened everywhere in the universe. All matter and energy that exist today, including the particles of which you are made, came from the Big Bang. We were, and still are, in the midst of a Big Bang; it is all around us.

Attribution

"29.4 The Cosmic Microwave Background" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

15.5 THE COMPOSITION OF THE UNIVERSE

The model of the universe we described in the previous section is the simplest model that explains the observations. It assumes that general relativity is the correct theory of gravity throughout the universe. With this assumption, the model then accounts for the existence and structure of the CMB; the abundances of the light elements deuterium, helium, and lithium; and the acceleration of the expansion of the universe. All of the observations to date support the validity of the model, which is referred to as the standard (or concordance) model of cosmology.

Figure 15.21 and Table 15.2 summarize the current best estimates of the contents of the universe. Luminous matter in stars and galaxies and neutrinos contributes about 1% of the mass required to reach critical density. Another 4% is mainly in the form of hydrogen and helium in the space between stars and in intergalactic space. Dark matter accounts for about an additional 27% of the critical density. The mass equivalent of dark energy (according to $E = mc^2$) then supplies the remaining 68% of the critical density.

Composition of the Universe

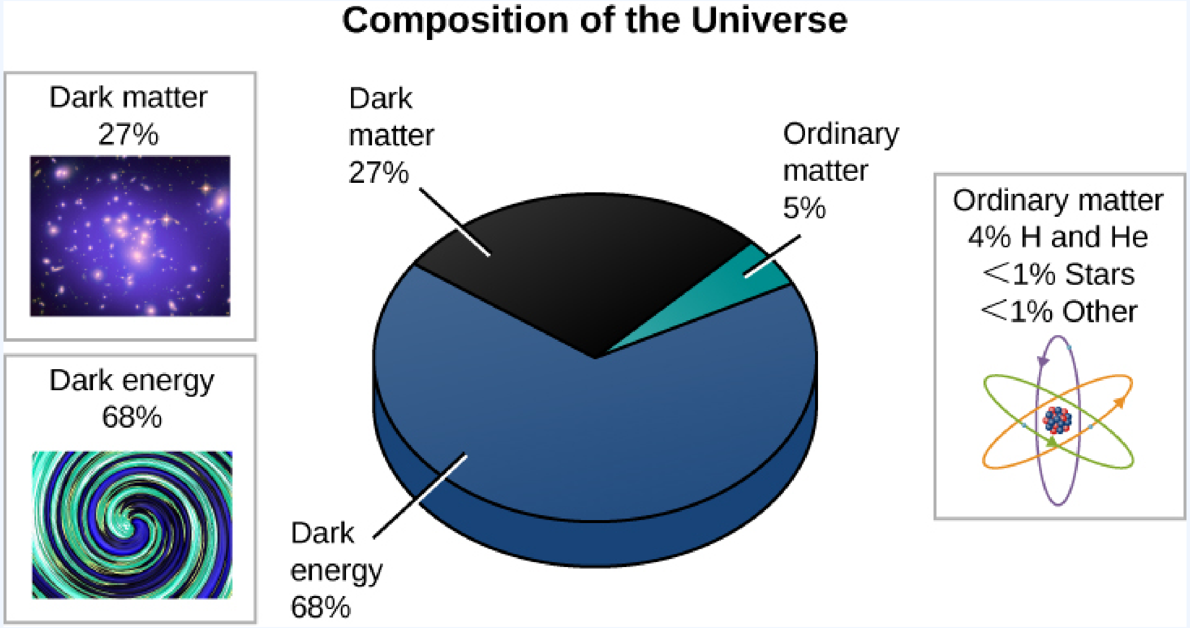


Figure 15.21. Only about 5% of all the mass and energy in the universe is matter with which we are familiar here on Earth. Most ordinary matter consists of hydrogen and helium located in interstellar and intergalactic space. Only about one-half of 1% of the critical density of the universe is found in stars. Dark matter and dark energy, which have not yet been detected in earthbound laboratories, account for 95% of the contents of the universe.

Table 15.2. What Different Kinds of Objects Contribute to the Density of the Universe

Object	Density as a Percent of Critical Density
Luminous matter (stars, etc.)	<1
Hydrogen and helium in interstellar and intergalactic space	4
Dark matter	27
Equivalent mass density of the dark energy	68

This table should shock you. What we are saying is that 95% of the stuff of the universe is either dark matter or dark energy—neither of which has ever been detected in a laboratory here on Earth. This whole textbook, which has focused on objects that emit electromagnetic radiation, has generally been ignoring 95% of what is out there. Who says there aren't big mysteries yet to solve in science!

Figure 15.22 shows how our ideas of the composition of the universe have changed over just the past three decades. The fraction of the universe that we think is made of the same particles as astronomy students has been decreasing steadily.

Changing Estimates of the Content of the Universe

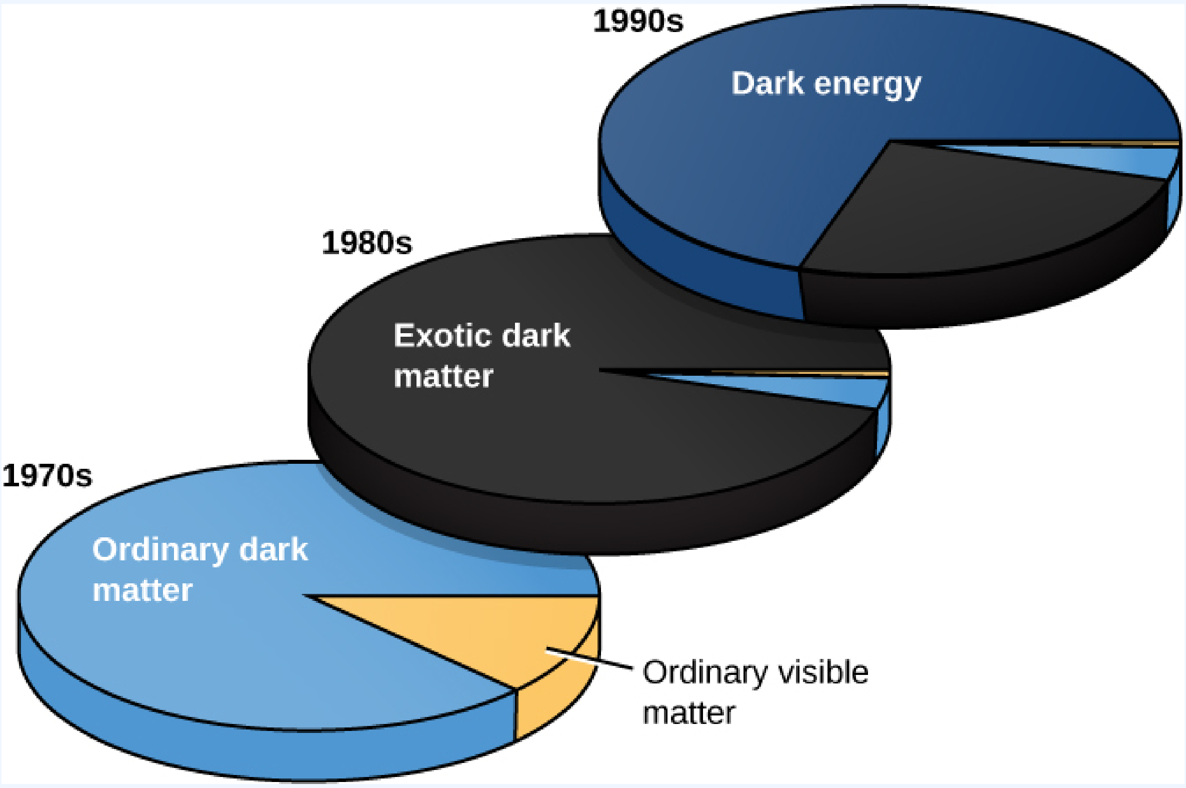


Figure 15.22. This diagram shows the changes in our understanding of the contents of the universe over the past three decades. In the 1970s, we suspected that most of the matter in the universe was invisible, but we thought that this matter might be ordinary matter (protons, neutrons, etc.) that was simply not producing electromagnetic radiation. By the 1980s, it was becoming likely that most of the dark matter was made of something we had not yet detected on Earth. By the late 1990s, a variety of experiments had shown that we live in a critical -density universe and that dark energy contributes about 70% of what is required to reach critical density. Note how the estimate of the relative importance of ordinary luminous matter (shown in yellow) has diminished over time.

What Is Dark Matter?

Many astronomers find the situation we have described very satisfying. Several independent experiments now agree on the type of universe we live in and on the inventory of what it contains. We seem to be very close to having a cosmological model that explains nearly everything. Others are not yet ready to jump on the bandwagon. They say, "show me the 96% of the universe we can't detect directly—for example, find me some dark matter!"

At first, astronomers thought that dark matter might be hidden in objects that appear dark because they emit no light (e.g., black holes) or that are too faint to be observed at large distances (e.g., planets or white dwarfs). However, these objects would be made of ordinary matter, and the deuterium abundance tells us that no more than 5% of the critical density consists of ordinary matter.

Another possible form that dark matter can take is some type of elementary particle that we have not yet detected here on Earth—a particle that has mass and exists in sufficient abundance to contribute 23% of the critical density. Some physics theories predict the existence of such particles. One class of these particles has been given the name WIMPs, which stands for **weakly interacting massive particles**. Since these particles do not participate in nuclear reactions leading to the production of deuterium, the deuterium abundance puts no limits on how many WIMPs might be in the universe. (A number of other exotic particles have also been suggested as prime constituents of dark matter, but we will confine our discussion to WIMPs as a useful example.)

If large numbers of WIMPs do exist, then some of them should be passing through our physics laboratories right now. The trick is to catch them. Since by definition they interact only weakly (infrequently) with other matter, the chances that they will have a measurable effect are small. We don't know the mass of these particles, but various theories suggest that it might be a few to a few hundred times the mass of a proton. If WIMPs are 60 times the mass of a proton, there would be about 10 million of them passing through your outstretched hand every second—with absolutely no effect on you. If that seems too mind-boggling, bear in mind that neutrinos interact weakly with ordinary matter, and yet we were able to "catch" them eventually.

Despite the challenges, more than 30 experiments designed to detect WIMPS are in operation or in the planning stages. Predictions of how many times WIMPs might actually collide with the nucleus of an atom in the instrument designed to detect them are in the range of 1 event per year to 1 event per 1000 years per kilogram of detector. The detector must therefore be large. It must be shielded from radioactivity or other types of particles, such as neutrons, passing through it, and hence these detectors are placed in deep mines. The energy imparted to an atomic nucleus in the detector by collision with a WIMP will be small, and so the detector must be cooled to a very low temperature.

The WIMP detectors are made out of crystals of germanium, silicon, or xenon. The detectors are cooled to a few thousandths of a degree—very close to absolute zero. That means that the atoms in the detector are so cold that they are scarcely vibrating at all. If a dark matter particle collides with one of the atoms, it will cause the whole crystal to vibrate and the temperature therefore to increase ever so slightly. Some other interactions may generate a detectable flash of light.

A different kind of search for WIMPs is being conducted at the Large Hadron Collider (LHC) at CERN, Europe's particle physics lab near Geneva, Switzerland. In this experiment, protons collide with enough energy potentially to produce WIMPs. The LHC detectors cannot detect the WIMPs directly, but if WIMPs are produced, they will pass through the detectors, carrying energy away with them. Experimenters will then add up all the energy that they detect as a result of the collisions of protons to determine if any energy is missing.

So far, none of these experiments has detected WIMPs. Will the newer experiments pay off? Or will scientists have to search for some other explanation for dark matter? Only time will tell (Figure 15.23 below).

Dark Matter

Figure 15.23. This cartoon from NASA takes a humorous look at how little we yet understand about dark

matter.
Dark Matter by NASA, NASA Licence

Dark Matter and the Formation of Galaxies

As elusive as dark matter may be in the current-day universe, galaxies could not have formed quickly without it. Galaxies grew from density fluctuations in the early universe, and some had already formed only about 400–500 million years after the Big Bang. The observations with WMAP, Planck, and other experiments give us information on the size of those density fluctuations. It turns out that the density variations we observe are too small to have formed galaxies so soon after the Big Bang. In the hot, early universe, energetic photons collided with hydrogen and helium, and kept them moving so rapidly that gravity was still not strong enough to cause the atoms to come together to form galaxies. How can we reconcile this with the fact that galaxies *did* form and are all around us?

Our instruments that measure the CMB give us information about density fluctuations only for *ordinary matter*, which interacts with radiation. Dark matter, as its name indicates, does not interact with photons at all. Dark matter could have had much greater variations in density and been able to come together to form gravitational "traps" that could then have begun to attract ordinary matter immediately after the universe became transparent. As ordinary matter became increasingly concentrated, it could have turned into galaxies quickly thanks to these dark matter traps.

For an analogy, imagine a boulevard with traffic lights every half mile or so. Suppose you are part of a motorcade of cars accompanied by police who lead you past each light, even if it is red. So, too, when the early universe was opaque, radiation interacted with ordinary matter, imparting energy to it and carrying it along, sweeping past the concentrations of dark matter. Now suppose the police leave the motorcade, which then encounters some red lights. The lights act as traffic traps; approaching cars now have to stop, and so they bunch up. Likewise, after the early universe became transparent, ordinary matter interacted with radiation only occasionally and so could fall into the dark matter traps.

The Universe in a Nutshell

In the previous sections of this chapter, we traced the evolution of the universe progressively further back in time. Astronomical discovery has followed this path historically, as new instruments and new techniques have allowed us to probe ever closer to the beginning of time. The rate of expansion of the universe was determined from measurements of nearby galaxies. Determinations of the abundances of deuterium, helium, and lithium based on nearby stars and galaxies were used to put limits on how much ordinary matter is in the universe. The motions of stars in galaxies and of galaxies within clusters of galaxies could only be explained if there were large quantities of dark matter. Measurements of supernovae that exploded when the universe was about half as old as it is now indicated that the rate of expansion of the universe has sped up since those explosions occurred. Observations of extremely faint galaxies show that galaxies had begun to form when the universe was only 400–500 million years old. And observations of the CMB confirmed early theories that the universe was initially very hot.

But all this moving further and further backward in time might have left you a bit dizzy. So now let's instead show how the universe evolves as time moves forward.

Figure 15.24 summarizes the entire history of the observable universe from the beginning in a single diagram. The universe was very hot when it began to expand. We have fossil remnants of the very early universe in the form of neutrons, protons, electrons, and neutrinos, and the atomic nuclei that formed when the universe was 3–4 minutes old: deuterium, helium, and a small amount of lithium. Dark matter also remains, but we do not yet know what form it is in.

HISTORY OF THE UNIVERSE

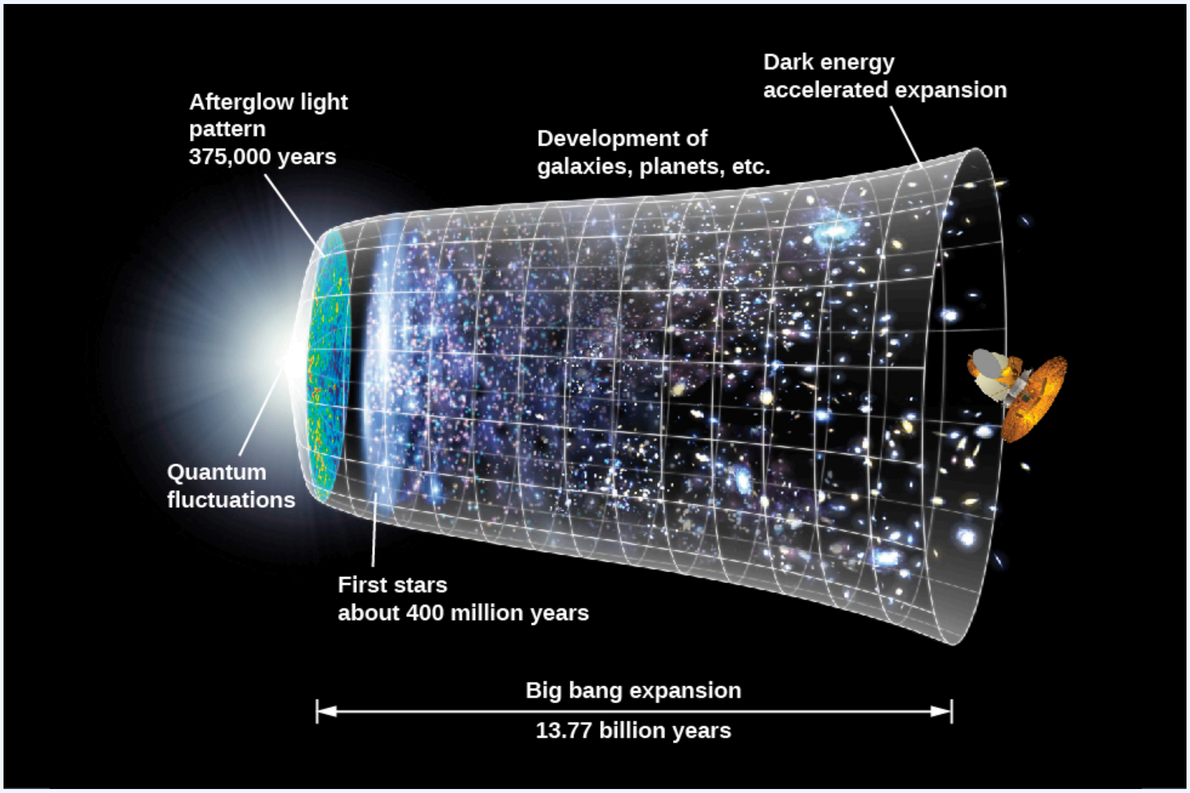


Figure 15.24. This image summarizes the changes that have occurred in the universe during the last 13.8 billion years. Protons, deuterium, helium, and some lithium were produced in the initial fireball. About 380,000 years after the Big Bang, the universe became transparent to electromagnetic radiation for the first time. COBE, WMAP, Planck, and other instruments have been used to study the radiation that was emitted at that time and that is still visible today (the CMB). The universe was then dark (except for this background radiation) until the first stars and galaxies began to form only a few hundred million years after the Big Bang. Existing space and ground-based telescopes have made substantial progress in studying the subsequent evolution of galaxies.
Making Sense of the Big Bang: Wilkinson Microwave Anisotropy Probe by NASA, NASA Licence

The universe gradually cooled; when it was about 380,000 years old, and at a temperature of about 3000 K, electrons combined with protons to form hydrogen atoms. At this point, as we saw, the universe became transparent to light, and astronomers have detected the CMB emitted at this time. The universe still contained no stars or galaxies, and so it entered what astronomers call "the dark ages" (since stars were not lighting up the darkness). During the next several hundred million years, small fluctuations in the density of the dark matter grew, forming gravitational traps that concentrated the ordinary matter, which began to form galaxies about 400–500 million years after the Big Bang.

By the time the universe was about a billion years old, it had entered its own renaissance: it was again blazing with radiation, but this time from newly formed stars, star clusters, and small galaxies. Over the next several billion years, small galaxies merged to form the giants we see today. Clusters and superclusters of galaxies began to grow, and the universe eventually began to resemble what we see nearby.

During the next 20 years, astronomers plan to build giant new telescopes both in space and on the ground to explore even further back in time. In 2022, the James Webb Space Telescope, a 6.5-meter telescope that is the successor to the Hubble Space Telescope, will be launched and assembled in space. The predictions are that with this powerful instrument we should be able to look back far enough to analyze in detail the formation of the first galaxies.

Attribution

"29.5 What Is the Universe Really Made Of?" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

15.6 KEY TERMS

Average density (of the universe): the mass of matter (including the equivalent mass of energy) that would be contained in each unit of volume (say, 1 cubic centimetre) if all the stars, galaxies, and other objects were taken apart, atom by atom, and if all those particles, along with the light and other energy, were distributed throughout all of space with absolute uniformity. 15.3

Big Bang: the theory of cosmology in which the expansion of the universe began with a primeval explosion (of space, time, matter, and energy). 15.2

Closed universe: a model in which the universe expands from a Big Bang, stops, and then contracts to a big crunch. 15.3

Cosmic microwave background (CMB): microwave radiation coming from all directions that is the redshifted afterglow of the Big Bang. 15.4

Cosmological constant: the term in the equations of general relativity that represents a repulsive force in the universe. 15.2

Cosmology: the study of the organization and evolution of the universe. 15.2

Critical density: in cosmology, the density that is just sufficient to bring the expansion of the universe to a stop after infinite time. 15.3

Flat universe: a model of the universe that has a critical density and in which the geometry of the universe is flat, like a sheet of paper. 15.4

Hubble constant: a constant of proportionality in the law relating the velocities of remote galaxies to their distances. 15.1

Hubble-Lemaître Law: a rule that the radial velocities of remove galaxies are proportional to their distances from us. 15.1

Open universe: a model in which the density of the universe is not high enough to bring the expansion of the universe to a halt. 15.3

Photon decoupling time: when radiation began to stream freely through the universe without interacting with matter. 15.4

Standard bulb (standard candle): some astronomical object with known luminosity that produces an enormous amount of energy and can be observed at distances of a billion light-years or more. 15.2

Weakly interacting massive particles (WIMPs): weakly interacting massive particles are one of the candidates for the composition of dark matter. 15.5

CHAPTER 16: LIFE IN THE UNIVERSE

Chapter Overview

16.0 LEARNING OBJECTIVES

Learning Objectives

In this chapter, students will learn how to:

- Identify key concepts and terminology related to astrobiology, such as the Copernican principle, organic molecules, photosynthesis, and habitable environments.
- Evaluate the Fermi paradox and various proposed solutions, considering the implications for the prevalence of intelligent life in the universe and the potential for interstellar communication or travel.
- Compare and contrast the ability of different microorganisms to use various chemical energy sources for survival.
- Analyze the factors that contribute to the thriving microbial life in extreme conditions, such as hot springs, acidic environments, and high-pressure regions.
- Explain the factors that define a planet's habitability, including surface temperature, atmosphere, and distance from the star.
- Utilize the Drake Equation to estimate the number of potential communicating civilizations in the Galaxy, given specific assumptions and values for its factors.
- Form a well-reasoned argument about the likelihood of finding intelligent life in the universe and its potential implications for humanity.

16.1 INTRODUCTION

Astrobiology: The Road to Life in the Universe

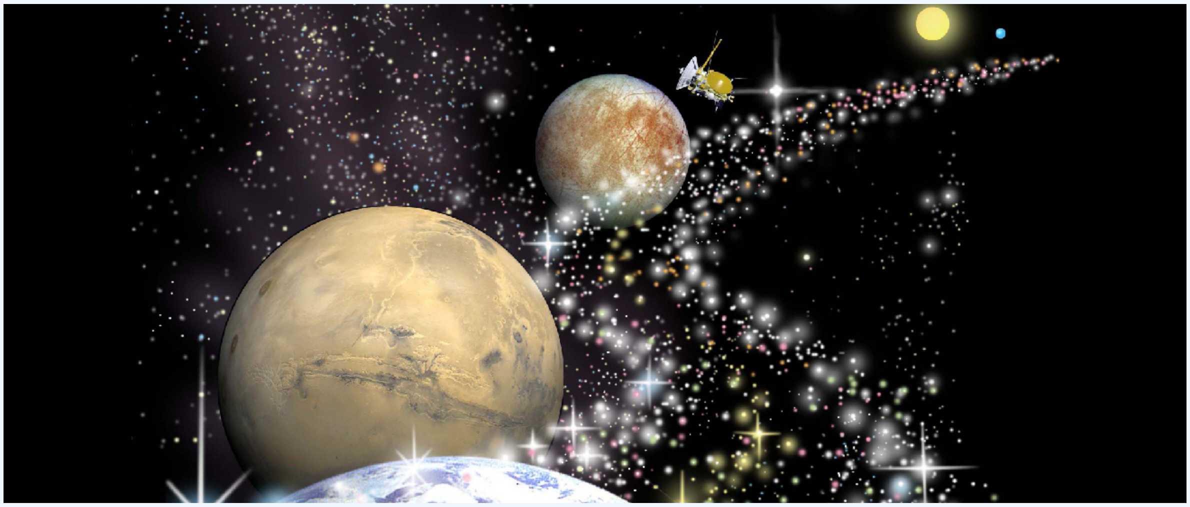


Figure 16.1 In this fanciful montage produced by a NASA artist, we see one road map for discovering life in the universe. Learning more about the origin, evolution, and properties of life on Earth aids us in searching for evidence of life beyond our planet. Our neighbour world, Mars, had warmer, wetter conditions billions of years ago that might have helped life there begin. Farther out, Jupiter's moon Europa represents the icy moons of the outer solar system. Beneath their shells of solid ice may lie vast oceans of liquid water that could support biology. Beyond our solar system are stars that host their own planets, some of which might be similar to Earth in the ability to support liquid water—and a thriving biosphere—at the planet's surface. Research is pushing actively in all these directions with the goal of proving a scientific answer to the question, "Are we alone?" (credit: modification of work by NASA)

As we have learned more about the universe, we have naturally wondered whether there might be other forms of life out there. The ancient question, "Are we alone in the universe?" connects us to generations of humans before us. While in the past, this question was in the realm of philosophy or science fiction, today we have the means to seek an answer through scientific inquiry. In this chapter, we will consider how life began on Earth, whether the same processes could have led to life on other worlds, and how we might seek evidence of life elsewhere. This is the science of astrobiology.

The search for life on other planets is not the same as the search for *intelligent* life, which (if it exists) is surely much rarer. Learning more about the origin, evolution, and properties of life on Earth aids us in searching for evidence of all kinds of life beyond that on our planet.

Attribution

16.2 THE COSMIC CONTEXT OF LIFE

We saw that the universe was born in the Big Bang about 14 billion years ago. After the initial hot, dense fireball of creation cooled sufficiently for atoms to exist, all matter consisted of hydrogen and helium (with a very small amount of lithium). As the universe aged, processes within stars created the other elements, including those that make up Earth (such as iron, silicon, magnesium, and oxygen) and those required for life as we know it, such as carbon, oxygen, and nitrogen. These and other elements combined in space to produce a wide variety of compounds that form the basis of life on Earth. In particular, life on Earth is based on the presence of a key unit known as an **organic molecule**, a molecule that contains carbon. Especially important are the hydrocarbons, chemical compounds made up entirely of hydrogen and carbon, which serve as the basis for our biological chemistry, or *biochemistry*. While we do not understand the details of how life on Earth began, it is clear that to make creatures like us possible, events like the ones we described must have occurred, resulting in what is called the *chemical evolution* of the universe.

What Made Earth Hospitable to Life?

About 5 billion years ago, a cloud of gas and dust in this cosmic neighbourhood began to collapse under its own weight. Out of this cloud formed the Sun and its planets, together with all the smaller bodies, such as comets, that also orbit the Sun (see Figure 16.2 below). The third planet from the Sun, as it cooled, eventually allowed the formation of large quantities of liquid water on its surface.

Comet Hyakutake

Figure 16.2 This image was captured in 1996 by NASA photographer Bill Ingalls. Comet impacts can deliver both water and a variety of interesting chemicals, including some organic chemicals, to Earth.
Image of Comet Hyakutake by NASA photographer Bill Ingalls. by NASA, NASA Licence

The chemical variety and moderate conditions on Earth eventually led to the formation of molecules that could make copies of themselves (reproduce), which is essential for beginning life. Over the billions of years of Earth history, life evolved and became more complex. The course of evolution was punctuated by occasional planet-wide changes caused by collisions with some of the smaller bodies that did not make it into the Sun or one of its accompanying worlds. As we saw in the chapter on the terrestrial planets, mammals may owe their domination of Earth's surface to just such a collision 65 million years ago, which led to the extinction of the dinosaurs (along with the majority of other living things). The details of such mass extinctions are currently the focus of a great deal of scientific interest.

Through many twisting turns, the course of evolution on Earth produced a creature with self-consciousness, able to ask questions about its own origins and place in the cosmos (see Figure 16.3 below). Like most of Earth, this creature is composed of atoms that were forged in earlier generations of stars—in this case, assembled into both its body and brain. We might say that through the thoughts of human beings, the matter in the universe can become aware of itself.

Young Human

Figure 16.3 Human beings have the intellect to wonder about their planet and what lies beyond it. Through them (and perhaps other intelligent life), the universe becomes aware of itself.
Baby in gray sweater by Christian Bowen, Unsplash Licence

Think about those atoms in your body for a minute. They are merely on loan to you from the lending library of atoms that make up our local corner of the universe. Atoms of many kinds circulate through your body and then leave it—with each breath you inhale and exhale and the food you eat and excrete. Even the atoms that take up more permanent residence in your tissues will not be part of you much longer than you are alive. Ultimately, you will return your atoms to the vast reservoir of Earth, where they will be incorporated into other structures and even other living things in the millennia to come.

This picture of *cosmic evolution*, of our descent from the stars, has been obtained through the efforts of scientists in many fields over many decades. Some of its details are still tentative and incomplete, but we feel reasonably confident in its broad outlines. It is remarkable how much we have been able to learn in the short time we have had the instruments to probe the physical nature of the universe.

The Copernican Principle

Our study of astronomy has taught us that we have always been wrong in the past whenever we have claimed that Earth is somehow unique. Galileo, using the newly invented technology of the telescope, showed us that Earth is not the centre of the solar system, but merely one of a number of objects orbiting the Sun. Our study of the stars has demonstrated that the Sun itself is a rather undistinguished star, halfway through its long main-sequence stage like so many billions of others. There seems nothing special about our position in the Milky Way Galaxy either, and nothing surprising about our Galaxy's position in either its own group or its supercluster.

The discovery of planets around other stars confirms our idea that the formation of planets is a natural consequence of the formation of stars. We have identified thousands of exoplanets—planets orbiting around other stars, from huge ones orbiting close to their stars (informally called "hot Jupiters") down to planets smaller than Earth. A steady stream of exoplanet discoveries is leading to the conclusion that earthlike planets occur frequently—enough that there are likely many billions of "exo-Earths" in our own Milky Way Galaxy alone. From a planetary perspective, smaller planets are not unique.

Philosophers of science sometimes call the idea that there is nothing special about our place in the universe the **Copernican principle**. Given all of the above, most scientists would be surprised if life were limited to our planet and had started nowhere else. There are billions of stars in our Galaxy old enough for life to have developed on a planet around them, and there are billions of other galaxies as well. Astronomers and biologists have long conjectured that a series of events similar to those on the early Earth probably led to living organisms on many planets around other stars, and possibly even on other planets in our solar system, such as Mars.

The real scientific issue (which we do not currently know the answer to) is whether organic biochemistry is likely or unlikely in the universe at large. Are we a fortunate and exceedingly rare outcome of chemical evolution, or is organic biochemistry a regular part of the chemical evolution of the cosmos? We do not yet know the answer to this question, but data, even an exceedingly small amount (like finding "unrelated to us" living systems on a world like Europa), will help us arrive at it.

So Where Are They?

If the Copernican principle is applied to life, then biology may be rather common among planets. Taken to its logical limit, the Copernican principle also suggests that intelligent life like us might be common. Intelligence like ours has some very special properties, including an ability to make progress through the application of technology. Organic life around other (older) stars may have started a billion years earlier than we did on Earth, so they may have had a lot more time to develop advanced technology such as sending information, probes, or even life-forms between stars.

Faced with such a prospect, physicist Enrico Fermi asked a question several decades ago that is now called the **Fermi paradox**: where are they? If life and intelligence are common and have such tremendous capacity for growth, why is there not a network of galactic civilizations whose presence extends even into a "latecomer" planetary system like ours?

Several solutions have been suggested to the Fermi paradox. Perhaps life is common but intelligence (or at least technological civilization) is rare. Perhaps such a network will come about in the future but has not yet had the time to develop. Maybe there are invisible streams of data flowing past us all the time that we are not advanced enough or sensitive enough to detect. Maybe advanced species make it a practice not to interfere with immature, developing consciousness such as our own. Or perhaps civilizations that reach a certain level of technology then self-destruct, meaning there are no other civilizations now existing in our Galaxy. We do not yet know whether any advanced life is out there and, if it is, why we are not aware of it. Still, you might want to keep these issues in mind as you read the rest of this chapter.

Attribution

16.3 ASTROBIOLOGY

Scientists today take a multidisciplinary approach to studying the origin, evolution, distribution, and ultimate fate of life in the universe; this field of study is known as **astrobiology**. You may also sometimes hear this field referred to as *exobiology* or *bioastronomy.* Astrobiology brings together astronomers, planetary scientists, chemists, geologists, and biologists (among others) to work on the same problems from their various perspectives.

Among the issues that astrobiologists explore are the conditions in which life arose on Earth and the reasons for the extraordinary adaptability of life on our planet. They are also involved in identifying habitable worlds beyond Earth and in trying to understand in practical terms how to look for life on those worlds. Let's look at some of these issues in more detail.

The Building Blocks of Life

While no unambiguous evidence for life has yet been found anywhere beyond Earth, life's chemical building blocks have been detected in a wide range of extraterrestrial environments. Meteorites have been found to contain two kinds of substances whose chemical structures mark them as having an extraterrestrial origin—amino acids and sugars. **Amino acids** are **organic compounds** that are the molecular building blocks of proteins. **Proteins** are key biological molecules that provide the structure and function of the body's tissues and organs and essentially carry out the "work" of the cell. When we examine the gas and dust around comets, we also find a number of organic molecules—compounds that on Earth are associated with the chemistry of life.

Expanding beyond our solar system, one of the most interesting results of modern radio astronomy has been the discovery of organic molecules in giant clouds of gas and dust between stars. More than 100 different molecules have been identified in these reservoirs of cosmic raw material, including formaldehyde, alcohol, and others we know as important stepping stones in the development of life on Earth. Using radio telescopes and radio spectrometres, astronomers can measure the abundances of various chemicals in these clouds. We find organic molecules most readily in regions where the interstellar dust is most abundant, and it turns out these are precisely the regions where star formation (and probably planet formation) happen most easily (see Figure 16.4 below).

Cloud of Gas and Dust

Figure 16.4 The Cat's Paw Nebula, captured by NASA's Spitzer Space Telescope, resides in the Milky Way's Scorpius constellation, 4,200 to 5,500 light years distant. This image, part of the Galactic Legacy Infrared Midplane Survey using Spitzer's Infrared Array Camera, depicts gas and dust about to give birth to new stars. It also highlights denser regions of material where star formation is underway. Spanning 80-90 light years, this star-forming area produces fresh stars that heat up surrounding compressed gas, resulting in the formation of expanding "bubbles."
PIA22567: Cat's Paw Image 2 by NASA/JPL-Caltech, NASA Media Licence.

Clearly the early Earth itself produced some of the molecular building blocks of life. Since the early 1950s, scientists have tried to duplicate in their laboratories the chemical pathways that led to life on our planet. In a series of experiments known as the *Miller-Urey experiments*, pioneered by Stanley Miller and Harold Urey at the University of Chicago, biochemists have simulated conditions on early Earth and have been able to produce some of the fundamental building blocks of life, including those that form proteins and other large biological molecules known as nucleic acids (which we will discuss shortly).

Although these experiments produced encouraging results, there are some problems with them. The most interesting chemistry from a biological perspective takes place with hydrogen-rich or *reducing* gases, such as ammonia and methane. However, the early atmosphere of Earth was probably dominated by carbon dioxide (as Venus' and Mars' atmospheres still are today) and may not have contained an abundance of reducing gases comparable to that used in Miller-Urey type experiments. Hydrothermal vents—seafloor systems in which ocean water is superheated and circulated through crustal or mantle rocks before reemerging into the ocean—have also been suggested as potential contributors of organic compounds on the early Earth, and such sources would not require Earth to have an early reducing atmosphere.

Both earthly and extraterrestrial sources may have contributed to Earth's early supply of organic molecules, although we have more direct evidence for the latter. It is even conceivable that life itself originated elsewhere and was seeded onto our planet—although this, of course, does not solve the problem of how that life originated to begin with.

The Origin and Early Evolution of Life

The carbon compounds that form the chemical basis of life may be common in the universe, but it is still a giant step from these building blocks to a living cell. Even the simplest molecules of the **genes** (the basic functional units that carry the genetic, or hereditary, material in a cell) contain millions of molecular units, each arranged in a precise sequence. Furthermore, even the most primitive life required two special capabilities: a means of extracting energy from its environment, and a means of encoding and replicating information in order to make faithful copies of itself. Biologists today can see ways that either of these capabilities might have formed in a natural environment, but we are still a long way from knowing how the two came together in the first life-forms.

We have no solid evidence for the pathway that led to the origin of life on our planet except for whatever early history may be retained in the biochemistry of modern life. Indeed, we have very little direct evidence of what Earth itself was like during its earliest history—our planet is so effective at resurfacing itself through plate tectonics that very few rocks remain from this early period. In the earlier chapter on Terrestrial Planets, you learned that Earth was subjected to a heavy bombardment—a period of large impact events—some 3.8 to 4.1 billion years ago. Large impacts would have been energetic enough to heat-sterilize the surface layers of Earth, so that even if life had begun by this time, it might well have been wiped out.

When the large impacts ceased, the scene was set for a more peaceful environment on our planet. If the oceans of Earth contained accumulated organic material from any of the sources already mentioned, the ingredients were available to make living organisms. We do not understand in any detail the sequence of events that led from molecules to biology, but there is fossil evidence of microbial life in 3.5-billion-year-old rocks, and possible (debated) evidence for life as far back as 3.8 billion years.

Life as we know it employs two main molecular systems: the functional molecules known as proteins, which carry out the chemical work of the cell, and information-containing molecules of **DNA (deoxyribonucleic acid)** that store information about how to create the cell and its chemical and structural components. The origin of life is sometimes considered a "chicken and egg problem" because, in modern biology, neither of these systems works without the other. It is our proteins that assemble DNA strands in the precise order required to store information, but the proteins are created based on information stored in DNA. Which came first? Some origin of life researchers believe that prebiotic chemistry was based on molecules that could both store information and do the chemical work of the cell. It has been suggested that **RNA (ribonucleic acid)**, a molecule that aids in the flow of genetic information from DNA to proteins, might have served such a purpose. The idea of an early "RNA world" has become increasingly accepted, but a great deal remains to be understood about the origin of life.

Perhaps the most important innovation in the history of biology, apart from the origin of life itself, was the discovery of the process of **photosynthesis**, the complex sequence of chemical reactions through which some living things can use sunlight to manufacture products that store energy (such as carbohydrates), releasing oxygen as one by-product. Previously, life had to make do with sources of chemical energy available on Earth or delivered from space. But the abundant energy available in sunlight could support a larger and more productive biosphere, as well as some biochemical reactions not previously possible for life. One of these was the production of oxygen (as a waste product) from carbon dioxide, and the increase in atmospheric levels of oxygen about 2.4 billion years ago means that oxygen-producing photosynthesis must have emerged and become globally important by this time. In fact, it is likely that oxygen-producing photosynthesis emerged considerably earlier.

Some forms of chemical evidence contained in ancient rocks, such as the solid, layered rock formations known as **stromatolites**, are thought to be the fossils of oxygen-producing photosynthetic bacteria in rocks that are almost 3.5 billion years old (Figure 16.5). It is generally thought that a simpler form of photosynthesis that does not produce oxygen (and is still used by some bacteria today) probably preceded oxygen-producing photosynthesis, and there is strong fossil evidence that one or the other type of photosynthesis was functioning on Earth at least as far back as 3.4 billion years ago.

Stromatolites Preserve the Earliest Physical Representation of Life on Earth

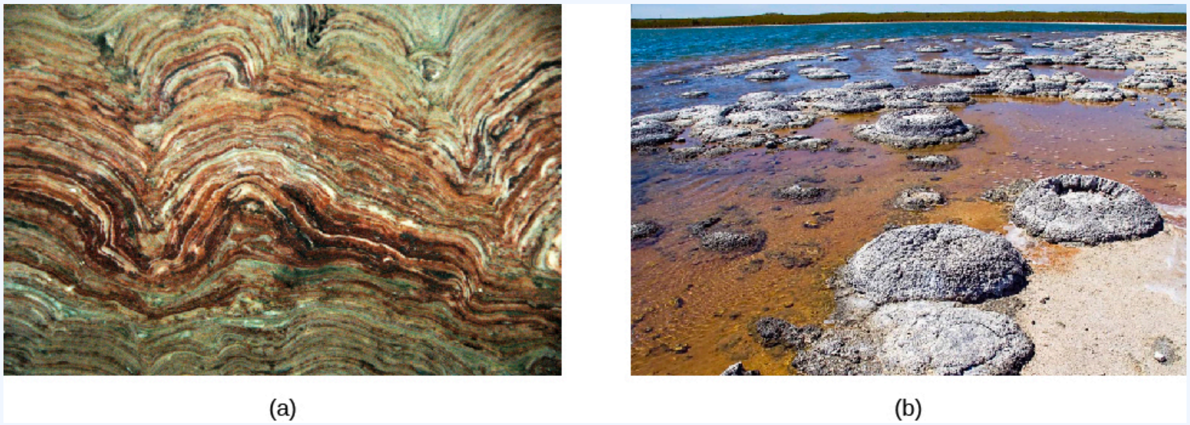


Figure 16.5 In their reach for sunlight, the single-celled microbes formed mats that trapped sediments in the water above them. Such trapped sediments fell and formed layers on top of the mats. The microbes then climbed atop the sediment layers and trapped more sediment. What is found in the rock record are (a) the solidified, curved sedimentary layers that are signatures of biological activity. The earliest known stromatolite is 3.47 billion years old and is found in Western Australia.
Stromatolites by James St. John, CC BY 2.0 (b) This more recent example is in Lake Thetis, also in Western Australia.
Stromatolites at Lake Thetis by Ruth Ellison, CC BY 2.0

The free oxygen produced by photosynthesis began accumulating in our atmosphere about 2.4 billion years ago. The interaction of sunlight with oxygen can produce ozone (which has three atoms of oxygen per molecule, as compared to the two atoms per molecule in the oxygen we breathe), which accumulated in a layer high in Earth's atmosphere. As it does on Earth today, this ozone provided protection from the Sun's damaging ultraviolet radiation. This allowed life to colonize the landmasses of our planet instead of remaining only in the ocean.

The rise in oxygen levels was deadly to some microbes because, as a highly reactive chemical, it can irreversibly damage some of the biomolecules that early life had developed in the absence of oxygen. For other microbes, it was a boon: combining oxygen with organic matter or other reduced chemicals generates a lot of energy—you can see this when a log burns, for example—and many forms of life adopted this way of living. This new energy source made possible a great proliferation of organisms, which continued to evolve in an oxygen-rich environment.

The details of that evolution are properly the subject of biology courses, but the process of evolution by natural selection (survival of the fittest) provides a clear explanation for the development of Earth's remarkable variety of life-forms. It does not, however, directly solve the mystery of life's earliest beginnings. We hypothesize that life will arise whenever conditions are appropriate, but this hypothesis is just another form of the Copernican principle. We now have the potential to address this hypothesis with observations. If a second example of life is found in our solar system or a nearby star, it would imply that life emerges commonly enough that the universe is likely filled with biology. To make such observations, however, we must first decide where to focus our search.

Habitable Environments

Among the staggering number of objects in our solar system, Galaxy, and universe, some may have conditions suitable for life, while others do not. Understanding what conditions and features make a **habitable environment**—an environment capable of hosting life—is important both for understanding how widespread habitable environments may be in the universe and for focusing a search for life beyond Earth. Here, we discuss habitability from the perspective of the life we know. We will explore the basic requirements of life and, in the following section, consider the full range of environmental conditions on Earth where life is found. While we can't entirely rule out the possibility that other life-forms might have biochemistry based on alternatives to carbon and liquid water, such life "as we don't know it" is still completely speculative. In our discussion here, we are focusing on habitability for life that is chemically similar to that on Earth.

Life requires a solvent (a liquid in which chemicals can dissolve) that enables the construction of biomolecules and the interactions between them. For life as we know it, that solvent is water, which has a variety of properties that are critical to how our biochemistry works. Water is abundant in the universe, but life requires that water be in liquid form (rather than ice or gas) in order to properly fill its role in biochemistry. That is the case only within a certain range of temperatures and pressures—too high or too low in either variable, and water takes the form of a solid or a gas. Identifying environments where water is present within the appropriate range of temperature and pressure is thus an important first step in identifying habitable environments. Indeed, a "follow the water" strategy has been, and continues to be, a key driver in the exploration of planets both within and beyond our solar system.

Our biochemistry is based on molecules made of carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur. Carbon is at the core of organic chemistry. Its ability to form four bonds, both with itself and with the other elements of life, allows for the formation of a vast number of potential molecules on which to base biochemistry. The remaining elements contribute structure and chemical reactivity to our biomolecules, and form the basis of many of the interactions among them. These "biogenic elements," sometimes referred to with the acronym CHNOPS (carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur), are the raw materials from which life is assembled, and an accessible supply of them is a second requirement of habitability.

As we learned in previous chapters on nuclear fusion and the life story of the stars, carbon, nitrogen, oxygen, phosphorus, and sulfur are all formed by fusion within stars and then distributed out into their galaxy as those stars die. But how they are distributed among the planets that form within a new star system, in what form, and how chemical, physical, and geological processes on those planets cycle the elements into structures that are accessible to biology, can have significant impacts on the distribution of life. In Earth's oceans, for example, the abundance of phytoplankton (simple organisms that are the base of the ocean food chain) in surface waters can vary by a thousand-fold because the supply of nitrogen differs from place to place (Figure 16.6). Understanding what processes control the accessibility of elements at all scales is thus a critical part of identifying habitable environments.

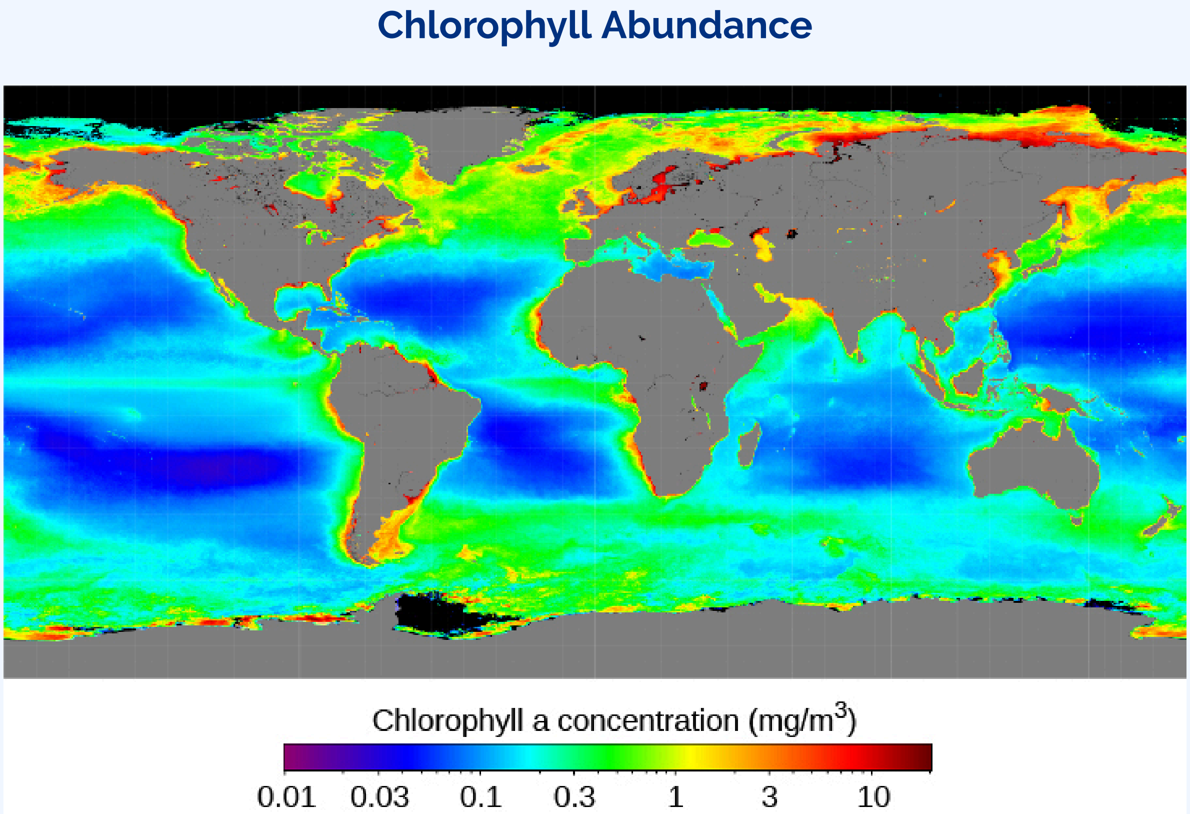


Figure 16.6. The abundance of chlorophyll (an indicator of photosynthetic bacteria and algae) varies by almost a thousand-fold across the ocean basins. That variation is almost entirely due to the availability of nitrogen—one of the major "biogenic elements" in forms that can be used by life. (credit: modification of Seasonal Chlorophyll Comparison by NASA, NASA Licence)

With these first two requirements, we have the elemental raw materials of life and a solvent in which to assemble them into the complicated molecules that drive our biochemistry. But carrying out that assembly and maintaining the complicated biochemical machinery of life takes energy. You fulfill your own requirement for energy every time you eat food or take a breath, and you would not live for long if you failed to do either on a regular basis. Life on Earth makes use of two main types of energy: for you, these are the oxygen in the air you breathe and the organic molecules in your food. But life overall can use a much wider array of chemicals and, while all animals require oxygen, many bacteria do not. One of the earliest known life processes, which still operates in some modern microorganisms, combines hydrogen and carbon dioxide to make methane, releasing energy in the process. There are microorganisms that "breathe" metals that would be toxic to us, and even some that breathe in sulfur and breathe out sulfuric acid. Plants and photosynthetic microorganisms have also evolved mechanisms to use the energy in light directly.

Water in the liquid phase, the biogenic elements, and energy are the fundamental requirements for habitability. But are there additional environmental constraints? We consider this in the next section.

Grand Prismatic Spring in Yellowstone National Park

Figure 16.7. This hot spring, where water emerges from the bluish centre at temperatures near the local boiling point (about 92 ºC), supports a thriving array of microbial life. The green, yellow, and orange colours around the edges come from thick “mats” of photosynthetic bacteria. In fact, their colouration in part demonstrates their use of light energy—some wavelengths of incoming sunlight are selectively captured for energy; the rest are reflected back. Since it lacks the captured wavelengths, this light is now different in colour than the sunlight that illuminates it. The blue part of the spring has temperatures too high to allow photosynthetic life (hence the lack of colour except that supplied by water itself), but life is still present. Here, at nearly boiling temperatures, bacteria use the chemical energy supplied by the combination of hydrogen and other chemicals with oxygen. Grand Prismatic by Mike Goad, Pixabay Content Licence

Life in Extreme Conditions

At a chemical level, life consists of many types of molecules that interact with one another to carry out the processes of life. In addition to water, elemental raw materials, and energy, life also needs an environment in which those complicated molecules are stable (don’t break down before they can do their jobs) and their interactions are possible. Your own biochemistry works properly only within a very narrow range of about 10

°C in body temperature and two-tenths of a unit in blood pH (pH is a numerical measure of acidity, or the amount of free hydrogen ions). Beyond those limits, you are in serious danger.

Life overall must also have limits to the conditions in which it can properly work but, as we will see, they are much broader than human limits. The resources that fuel life are distributed across a very wide range of conditions. For example, there is abundant chemical energy to be had in hot springs that are essentially boiling acid (see Figure 16.7 above). This provides ample incentive for evolution to fill as much of that range with life as is biochemically possible. An organism (usually a microbe) that tolerates or even thrives under conditions that most of the life around us would consider hostile, such as very high or low temperature or acidity, is known as an **extremophile** (where the suffix *-phile* means "lover of"). Let's have a look at some of the conditions that can challenge life and the organisms that have managed to carve out a niche at the far reaches of possibility.

Both high and low temperatures can cause a problem for life. As a large organism, you are able to maintain an almost constant body temperature whether it is colder or warmer in the environment around you. But this is not possible at the tiny size of microorganisms; whatever the temperature in the outside world is also the temperature of the microbe, and its biochemistry must be able to function at that temperature. High temperatures are the enemy of complexity—increasing thermal energy tends to break apart big molecules into smaller and smaller bits, and life needs to stabilize the molecules with stronger bonds and special proteins. But this approach has its limits.

Nevertheless, as noted earlier, high-temperature environments like hot springs and hydrothermal vents often offer abundant sources of chemical energy and therefore drive the evolution of organisms that can tolerate high temperatures (see Figure 16.8); such an organism is called a **thermophile**. Currently, the high temperature record holder is a methane-producing microorganism that can grow at 122 °C, where the pressure also is so high that water still does not boil. That's amazing when you think about it. We cook our food—meaning, we alter the chemistry and structure of its biomolecules—by boiling it at a temperature of 100 °C. In fact, food begins to cook at much lower temperatures than this. And yet, there are organisms whose biochemistry remains intact and operates just fine at temperatures 20 degrees higher.

Hydrothermal Vent on the Sea Floor

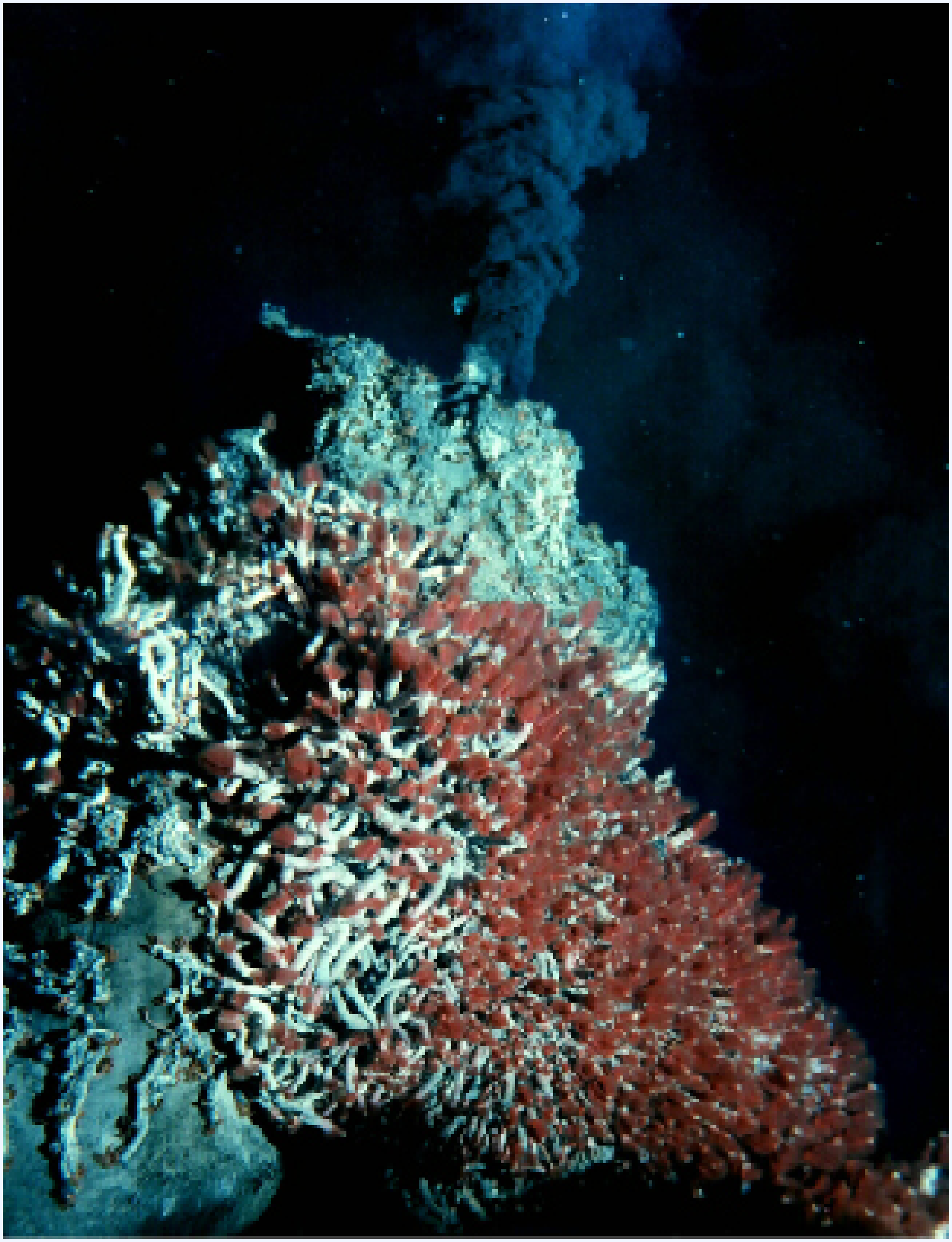


Figure 16.8. What appears to be black smoke is actually superheated water filled with minerals of metal sulfide. Hydrothermal vent fluid can represent a rich source of chemical energy, and therefore a driver for the evolution of microorganisms that can tolerate high temperatures. Bacteria feeding on this chemical energy form the base of a food chain that can support thriving communities of animals—in this case, a dense patch of red and white tubeworms growing around the base of the vent.
expl2366 by NOAA Photo Library, CC BY 2.0

Cold can also be a problem, in part because it slows down metabolism to very low levels, but also because it can cause physical changes in biomolecules. Cell membranes—the molecular envelopes that surround cells and allow their exchange of chemicals with the world outside—are basically made of fatlike molecules. And just as fat congeals when it cools, membranes crystallize, changing how they function in the exchange of materials in and out of the cell. Some cold-adapted cells (called **psychrophiles**) have changed the chemical composition of their membranes in order to cope with this problem; but again, there are limits. Thus far, the coldest temperature at which any microbe has been shown to reproduce is about –25 ºC.

Conditions that are very acidic or alkaline can also be problematic for life because many of our important molecules, like proteins and DNA, are broken down under such conditions. For example, household drain cleaner, which does its job by breaking down the chemical structure of things like hair clogs, is a very alkaline solution. The most acid-tolerant organisms (**acidophiles**) are capable of living at pH values near zero—about ten million times more acidic than your blood (Figure 16.9 below). At the other extreme, some **alkaliphiles** can grow at pH levels of about 13, which is comparable to the pH of household bleach and almost a million times more alkaline than your blood.

Spain's Rio Tinto

Figure 16.9 With a pH close to 2, Rio Tinto is literally a river of acid. Acid-loving microorganisms (acidophiles) not only thrive in these waters, their metabolic activities help generate the acid in the first place. The rusty red colour that gives the river its name comes from high levels of iron dissolved in the waters.
The Rio Tinto in 2006 by Riotinto2006, CC0 1.0

High levels of salts in the environment can also cause a problem for life because the salt blocks some cellular functions. Humans recognized this centuries ago and began to salt-cure food to keep it from spoiling—meaning, to keep it from being colonized by microorganisms. Yet some microbes have evolved to grow in water that is saturated in sodium chloride (table salt)—about ten times as salty as seawater (Figure 16.10 below).

Salt Ponds

Figure 16.10. The waters of an evaporative salt works near San Francisco are coloured pink by thriving communities of photosynthetic organisms. These waters are about ten times as salty as seawater—enough for sodium chloride to begin to crystallize out—yet some organisms can survive and thrive in these conditions.
STS111-376-3 by NASA, NASA Licence

Very high pressures can literally squeeze life's biomolecules, causing them to adopt more compact forms that do not work very well. But we still find life—not just microbial, but even animal life—at the bottoms of our ocean trenches, where pressures are more than 1000 times atmospheric pressure. Many other adaptions to environmental "extremes" are also known. There is even an organism, **Deinococcus radiodurans**, that can tolerate ionizing radiation (such as that released by radioactive elements) a thousand times more intense than you would be able to withstand. It is also very good at surviving extreme desiccation (drying out) and a variety of metals that would be toxic to humans.

From many such examples, we can conclude that life is capable of tolerating a wide range of environmental extremes—so much so that we have to work hard to identify places where life can't exist. A few such places are known—for example, the waters of hydrothermal vents at over 300 °C appear too hot to support any life—and finding these places helps define the possibility for life elsewhere. The study of extremophiles over the last few decades has expanded our sense of the range of conditions life can survive and, in doing so, has made many scientists more optimistic about the possibility that life might exist beyond Earth.

Attribution

"30.2 Astrobiology" from Douglas College Astronomy 1105 by Douglas College Department of Physics and Astronomy, is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted. Adapted from Astronomy 2e.

16.4 THE SEARCH FOR EXTRATERRESTRIAL LIFE

Astronomers and planetary scientists continue to search for life in the solar system and the universe at large. In this section, we discuss two kinds of searches. First is the direct exploration of planets within our own solar system, especially Mars and some of the icy moons of the outer solar system. Second is the even more difficult task of searching for evidence of life—a **biomarker**—on planets circling other stars. In the next section, we will examine SETI, the *search for extraterrestrial intelligence*. As you will see, the approaches taken in these three cases are very different, even though the goal of each is the same: to determine if life on Earth is unique in the universe.

Life on Mars

The possibility that Mars hosts, or has hosted, life has a rich history dating back to the "canals" that some people claimed to see on the martian surface toward the end of the nineteenth century and the beginning of the twentieth. With the dawn of the space age came the possibility to address this question up close through a progression of missions to Mars that began with the first successful flyby of a robotic spacecraft in 1964 and have led to the deployment of NASA's *Curiosity* rover, which landed on Mars' surface in 2012.

The earliest missions to Mars provided some hints that liquid water—one of life's primary requirements—may once have flowed on the surface, and later missions have strengthened this conclusion. The NASA Viking landers, whose purpose was to search directly for evidence of life on Mars, arrived on Mars in 1976. Viking's onboard instruments found no organic molecules (the stuff of which life is made), and no evidence of biological activity in the martian soils it analyzed.

This result is not particularly surprising because, despite the evidence of flowing liquid water in the past, liquid water on the surface of Mars is generally not stable today. Over much of Mars, temperatures and pressures at the surface are so low that pure water would either freeze or boil away (under very low pressures, water will boil at a much lower temperature than usual). To make matters worse, unlike Earth, Mars does not have a magnetic field and ozone layer to protect the surface from harmful solar ultraviolet radiation and energetic particles. However, Viking's analyses of the soil said nothing about whether life may have existed in Mars' distant past, when liquid water was more abundant. We do know that water in the form of ice exists in abundance on Mars, not so deep beneath its surface. Water vapor is also a constituent of the atmosphere of Mars.

Since the visit of Viking, our understanding of Mars has deepened spectacularly. Orbiting spacecraft have provided ever-more detailed images of the surface and detected the presence of minerals that could have formed only in the presence of liquid water. Two bold surface missions, the Mars Exploration Rovers *Spirit* and *Opportunity* (2004), followed by the much larger *Curiosity* Rover (2012), confirmed these remote-sensing data. All three rovers found abundant evidence for a past history of liquid water, revealed not only from the mineralogy of rocks they analyzed, but also from the unique layering of rock formations.

Curiosity has gone a step beyond evidence for water and confirmed the existence of habitable environments on ancient Mars. "Habitable" means not only that liquid water was present, but that life's requirements for energy and elemental raw materials could also have been met. The strongest evidence of an ancient habitable environment came from analyzing a very fine-grained rock called a mudstone—a rock type that is widespread on Earth but was unknown on Mars until *Curiosity* found it (see Figure 16.11 below). The mudstone can tell us a great deal about the wet environments in which they formed.

Mudstone

Figure 16.11. Shown are the first holes drilled by NASA's Curiosity Mars rover into a mudstone, with "fresh" drill-pilings around the holes. Notice the difference in colour between the red ancient martian surface and the grey newly exposed rock powder that came from the drill holes. Each drill hole is about 0.6 inch (1.6 cm) in diameter. (credit: modification of work by NASA/JPL-Caltech/MSSS)
Dust from Mars Drilling: Tailings and Discard Piles by NASA, NASA/JPL Image Use Policy.

Five decades of robotic exploration have allowed us to develop a picture of how Mars evolved through time. Early Mars had epochs of warmer and wetter conditions that would have been conducive to life at the surface. However, Mars eventually lost much of its early atmosphere and the surface water began to dry up. As that happened, the ever-shrinking reservoirs of liquid water on the martian surface became saltier and more acidic, until the surface finally had no significant liquid water and was bathed in harsh solar radiation. The surface thus became uninhabitable, but this might not be the case for the planet overall.

Reservoirs of ice and liquid water could still exist underground, where pressure and temperature conditions make it stable. There is recent evidence to suggest that liquid water (probably very salty water) can occasionally (and briefly) flow on the surface even today. Thus, Mars might even have habitable conditions in the present day, but of a much different sort than we normally think of on Earth.

Our study of Mars reveals a planet with a fascinating history—one that saw its ability to host surface life dwindle billions of years ago, but perhaps allowing life to adapt and survive in favourable environmental niches. Even if life did not survive, we expect that we might find evidence of life if it ever took hold on Mars. If it is there, it is hidden in the crust, and we are still learning how best to decipher that evidence.

Life in the Outer Solar System

The massive gas and ice giant planets of the outer solar system—Jupiter, Saturn, Uranus, and Neptune—are almost certainly not habitable for life as we know it, but some of their moons might be (see [image below]). Although these worlds in the outer solar system contain abundant water, they receive so little warming sunlight in their distant orbits that it was long believed they would be "geologically dead" balls of hard-frozen ice and rock. But, as we saw in the chapter "Other Objects in the Solar System", missions to the outer solar system have found something much more interesting.

Jupiter's moon Europa revealed itself to the Voyager and Galileo missions as an active world whose icy surface apparently conceals an ocean with a depth of tens to perhaps a hundred kilometres. As the moon orbits Jupiter, the planet's massive gravity creates tides on Europa—just as our own Moon's gravity creates our ocean tides—and the friction of all that pushing and pulling generates enough heat to keep the water in liquid form (Figure 16.12.). Similar tides act upon other moons if they orbit close to the planet. Scientists now think that six or more of the outer solar system's icy moons may harbour liquid water oceans for the same reason. Among these, Europa and Enceladus, a moon of Saturn, have thus far been of greatest interest to astrobiologists.

Jupiter's Moons

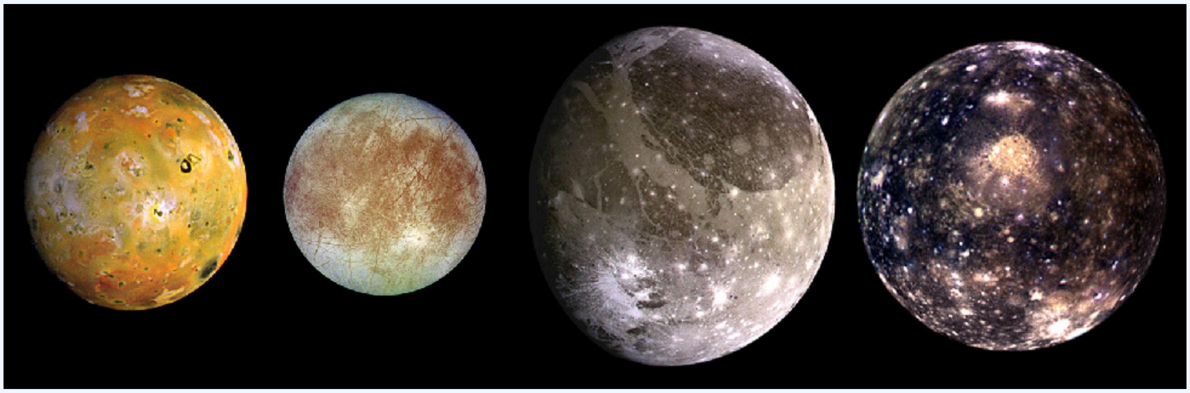


Figure 16.12. The Galilean moons of Jupiter are shown to relative scale and arranged in order of their orbital distance from Jupiter. At far left, Io orbits closest to Jupiter and so experiences the strongest tidal heating by Jupiter's massive gravity. This effect is so strong that Io is thought to be the most volcanically active body in our solar system. At far right, Callisto shows a surface scarred by billions of years' worth of craters—an indication that the moon's surface is old and that Callisto may be far less active than its sibling moons. Between these hot and cold extremes, Europa, second from left, orbits at a distance where Jupiter's tidal heating may be "just right" to sustain a liquid water ocean beneath its icy crust. PIA00601: The Galilean Satellites by NASA/JPL, JPL Image Use Policy.

Europa has probably had an ocean for most or all of its history, but habitability requires more than just liquid water. Life also requires energy, and because sunlight does not penetrate below the kilometres-thick ice crust of Europa, this would have to be chemical energy. One of Europa's key attributes from an astrobiology perspective is that its ocean is most likely in direct contact with an underlying rocky mantle, and the interaction of water and rocks—especially at high temperatures, as within Earth's hydrothermal vent systems—yields a **reducing chemistry** (where molecules tend to give up electrons readily) that is like one half of a chemical battery. To complete the battery and provide energy that could be used by life requires that an **oxidizing chemistry** (where molecules tend to accept electrons readily) also be available. On Earth, when chemically reducing vent fluids meet oxygen-containing seawater, the energy that becomes available often supports thriving communities of microorganisms and animals on the sea floor, far from the light of the Sun.

The Galileo mission found that Europa's icy surface does contain an abundance of oxidizing chemicals. This means that availability of energy to support life depends very much on whether the chemistry of the surface and the ocean can mix, despite the kilometres of ice in between. That Europa's ice crust appears geologically "young" (only tens of millions of years old, on average) and that it is active makes it tantalizing to think that such mixing might indeed occur. Understanding whether and how much exchange occurs between the surface and ocean of Europa will be a key science objective of future missions to Europa, and a major step forward in understanding whether this moon could be a cradle of life.

Jupiter's Moon Europa, as Imaged by NASA's Galileo Mission

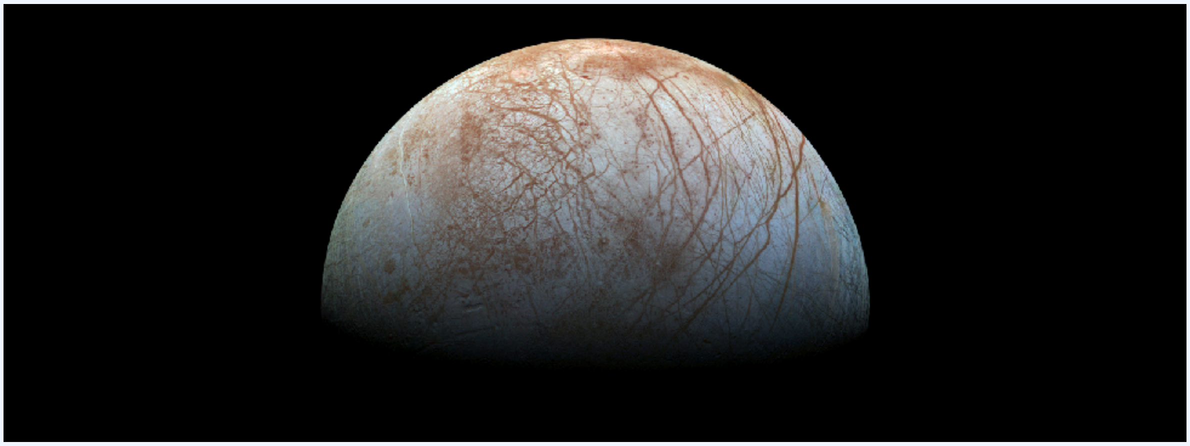


Figure 16.13. The relative scarcity of craters on Europa suggests a surface that is "geologically young," and the network of coloured ridges and cracks suggests constant activity and motion. Galileo's instruments also strongly suggested the presence of a massive ocean of salty liquid water beneath the icy crust. PIA19048: Europa's Stunning Surface by JPL, JPL Image Use Policy.

In 2005, the Cassini mission performed a close flyby of a small (500-kilometre diameter) moon of Saturn, Enceladus (Figure 16.14), and made a remarkable discovery. Plumes of gas and icy material were venting from the moon's south polar region at a collective rate of about 250 kilograms of material per second. Several observations, including the discovery of salts associated with the icy material, suggest that their source is a liquid water ocean beneath tens of kilometres of ice. Although it remains to be shown definitively whether the ocean is local or global, transient or long-lived, it does appear to be in contact, and to have reacted, with a rocky interior. As on Europa, this is probably a necessary—though not sufficient—condition for habitability. What makes Enceladus so enticing to planetary scientists, though, are those plumes of material that seem to come directly from its ocean: samples of the interior are there for the taking by any spacecraft sent flying through. For a future mission, such samples could yield evidence not only of whether Enceladus is habitable but, indeed, of whether it is home to life.

Image of Saturn's Moon Enceladus from NASA's Cassini Mission

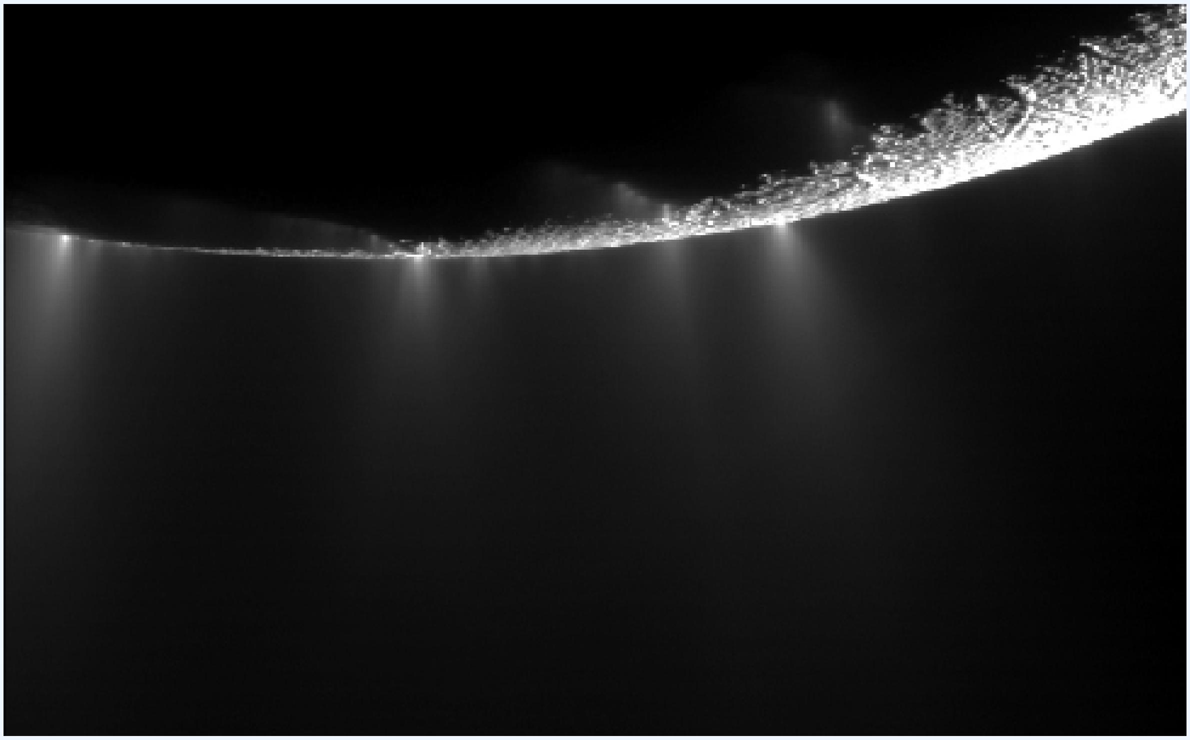


Figure 16.14. The south polar region was found to have multiple plumes of ice and gas that, combined, are venting about 250 kilograms of material per second into space. Such features suggest that Enceladus, like Europa, has a sub-ice ocean.
Bursting at the Seams: the Geyser Basin of Enceladus by NASA/JPL, JPL Image Use Policy

Saturn's big moon Titan is very different from both Enceladus and Europa (see Figure 16.15 below). Although it may host a liquid water layer deep within its interior, it is the surface of Titan and its unusual chemistry that makes this moon such an interesting place. Titan's thick atmosphere—the only one among moons in the solar system—is composed mostly of nitrogen but also of about 5% methane. In the upper atmosphere, the Sun's ultraviolet light breaks apart and recombines these molecules into more complex organic compounds that are collectively known as **tholins**. The tholins shroud Titan in an orange haze, and imagery from Cassini and from the Huygens probe that descended to Titan's surface show that heavier particles appear to accumulate on the surface, even forming "dunes" that are cut and sculpted by flows of liquid hydrocarbons (such as liquid methane). Some scientists see this organic chemical factory as a natural laboratory that may yield some clues about the solar system's early chemistry—perhaps even chemistry that could support the origin of life.

Image of Saturn's Moon Titan from NASA's Cassini Mission

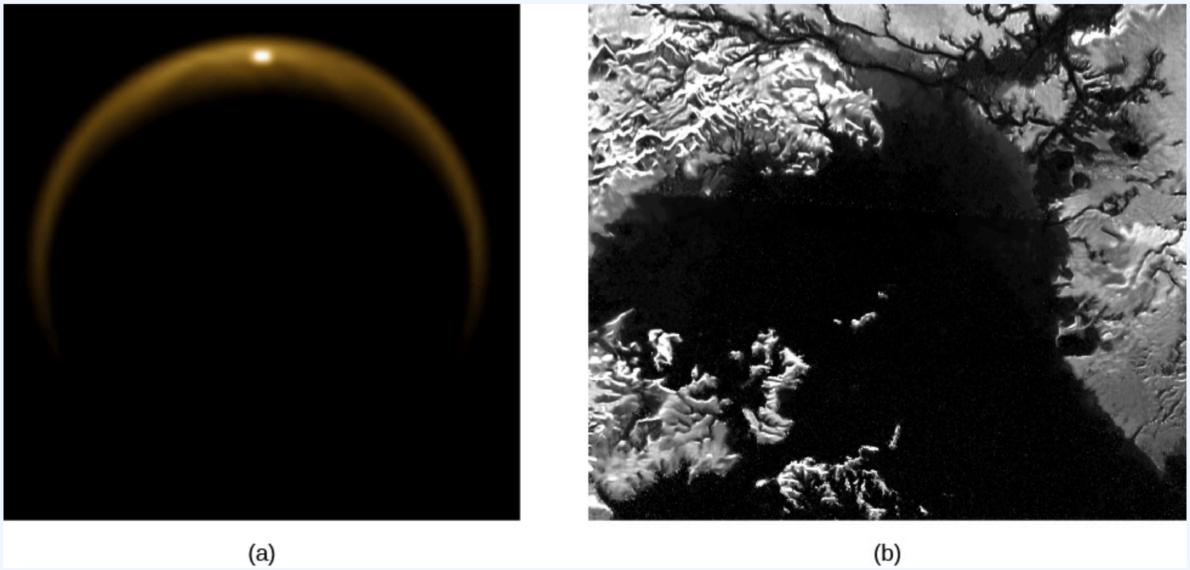


Figure 16.15. (a) The hazy orange glow comes from Titan's thick atmosphere (the only one known among the moons of the solar system). That atmosphere is mostly nitrogen but also contains methane and potentially a variety of complex organic compounds. The bright spot near the top of the image is sunlight reflected from a very flat surface—almost certainly a liquid. We see this effect, called "glint," when sunlight reflects off the surface of a lake or ocean.
Reflection of Sunlight off Titan Lake by NASA, NASA Licence (b) Cassini radar imagery shows what look very much like landforms and lakes on the surface of Titan. But the surface lakes and oceans of Titan are not water; they are probably made of liquid hydrocarbons like methane and ethane.
PIA19052: Despeckling Ligea Mare by JPL, JPL Media Licence

Habitable Planets Orbiting Other Stars

One of the most exciting developments in astronomy during the last two decades is the ability to detect exoplanets—planets orbiting other stars. As we saw in the chapter on the formation of stars and planets, since the discovery of the first exoplanet in 1995, there have been thousands of confirmed detections, and many more candidates that are not yet confirmed. These include several dozen possibly habitable exoplanets. Such numbers finally allow us to make some predictions about exoplanets and their life-hosting potential. The majority of stars with mass similar to the Sun appear to host at least one planet, with multi-planet systems like our own not unusual. How many of these planets might be habitable, and how could we search for life there?

The NASA Exoplanet Archive is an up-to-date searchable online source of data and tools on everything to do with exoplanets. Explore stellar and exoplanet parameters and characteristics, find the latest news on exoplanet discoveries, plot your own data interactively, and link to other related resources. Direct link: https://exoplanetarchive.ipac.caltech.edu/

In evaluating the prospect for life in distant planetary systems, astrobiologists have developed the idea of a **habitable zone**—a region around a star where suitable conditions might exist for life. This concept focuses on life's requirement for liquid water, and the habitable zone is generally thought of as the range of distances from the central star in which water could be present in liquid form at a planet's surface. In our own solar system, for example, Venus has surface temperatures far above the boiling point of water and Mars has surface temperatures that are almost always below the freezing point of water. Earth, which orbits between the two, has a surface temperature that is "just right" to keep much of our surface water in liquid form.

Whether surface temperatures are suitable for maintaining liquid water depends on a planet's "radiation budget" —how much starlight energy it absorbs and retains—and whether or how processes like winds and ocean circulation distribute that energy around the planet. How much stellar energy a planet receives, in turn, depends on how much and what sort of light the star emits and how far the planet is from that star,[1] how much it reflects back to space, and how effectively the planet's atmosphere can retain heat through the greenhouse effect. All of these can vary substantially, and all matter a lot. For example, Venus receives about twice as much starlight per square meter as Earth but, because of its dense cloud cover, also reflects about twice as much of that light back to space as Earth does. Mars receives only about half as much starlight as Earth, but also reflects only about half as much. Thus, despite their differing orbital distances, the three planets actually absorb comparable amounts of sunlight energy. Why, then, are they so dramatically different?

As we learned in several chapters about the planets, some of the gases that make up planetary atmospheres are very effective at trapping infrared light—the very range of wavelengths at which planets radiate thermal energy back out to space—and this can raise the planet's surface temperature quite a bit more than would otherwise be the case. This is the same "greenhouse effect" that is of such concern for global warming on our planet. Earth's natural greenhouse effect, which comes mostly from water vapour and carbon dioxide in the atmosphere, raises our average surface temperature by about 33 °C over the value it would have if there were no greenhouse gases in the atmosphere. Mars has a very thin atmosphere and thus very little greenhouse warming (about 2 °C worth), while Venus has a massive carbon dioxide atmosphere that creates very strong greenhouse warming (about 510 °C worth). These worlds are much colder and much hotter, respectively, than Earth would be if moved into their orbits. Thus, we must consider the nature of any atmosphere as well as the distance from the star in evaluating the range of habitability.

Of course, as we have learned, stars also vary widely in the intensity and spectrum (the wavelengths of light) they emit. Some are much brighter and hotter (bluer), while others are significantly dimmer and cooler (redder), and the distance of the habitable zone varies accordingly. For example, the habitable zone around M-dwarf stars is 3 to 30 times closer in than for G-type (Sun-like) stars. There is a lot of interest in whether such systems could be habitable because—although they have some potential downsides for supporting life—M-dwarf stars are by far the most numerous and long-lived in our Galaxy.

The luminosity of stars like the Sun also increases over their main-sequence lifetime, and this means that the habitable zone migrates outward as a star system ages. Calculations indicate that the power output of the Sun, for example, has increased by at least 30% over the past 4 billion years. Thus, Venus was once within the habitable zone, while Earth received a level of solar energy insufficient to keep the modern Earth (with its present atmosphere) from freezing over. In spite of this, there is plenty of geological evidence that liquid water was present on Earth's surface billions of years ago. The phenomenon of increasing stellar output and an outwardly migrating habitable zone has led to another concept: the ***continuously*** **habitable zone** is defined by the range of orbits that would remain within the habitable zone during the entire lifetime of the star system. As you might imagine, the continuously habitable zone is quite a bit narrower than the habitable zone is at any one time in a star's history. The nearest star to the Sun, Proxima Centauri, is an M star that has a planet with a mass of at least 1.3 Earth masses, taking about 11 days to orbit. At the distance for such a quick orbit (0.05 AU), the planet may be in the habitable zone of its star, although whether conditions on such a planet near such a star are hospitable for life is a matter of great scientific debate.

Even when planets orbit within the habitable zone of their star, it is no guarantee that they are habitable. For example, Venus today has virtually no water, so even if it were suddenly moved to a "just right" orbit within the habitable zone, a critical requirement for life would still be lacking.

Scientists are working to understand all the factors that define the habitable zone and the habitability of planets orbiting within that zone because this will be our primary guide in targeting exoplanets on which to seek evidence of life. As technology for detecting exoplanets has advanced, so too has our potential to find Earth-size worlds within the habitable zones of their parent stars. Of the confirmed or candidate exoplanets known at the time of writing, nearly 300 are considered to be orbiting within the habitable zone and more than 10% of those are roughly Earth-size.

Explore the habitable universe at the online Planetary Habitability Laboratory created by the University of Puerto Rico at Arecibo. See the potentially habitable exoplanets and other interesting places in the universe, watch video clips, and link to numerous related resources on astrobiology. Direct link: http://phl.upr.edu/projects/habitable-exoplanets-catalog

Biomarkers

Our observations suggest increasingly that Earth-size planets orbiting within the habitable zone may be common in the Galaxy—current estimates suggest that more than 40% of stars have at least one. But are any of them inhabited? With no ability to send probes there to sample, we will have to derive the answer from the light and other radiation that come to us from these faraway systems (Figure 16.16 below). What types of observations might constitute good evidence for life?

Earth, as Seen by NASA's Voyager 1

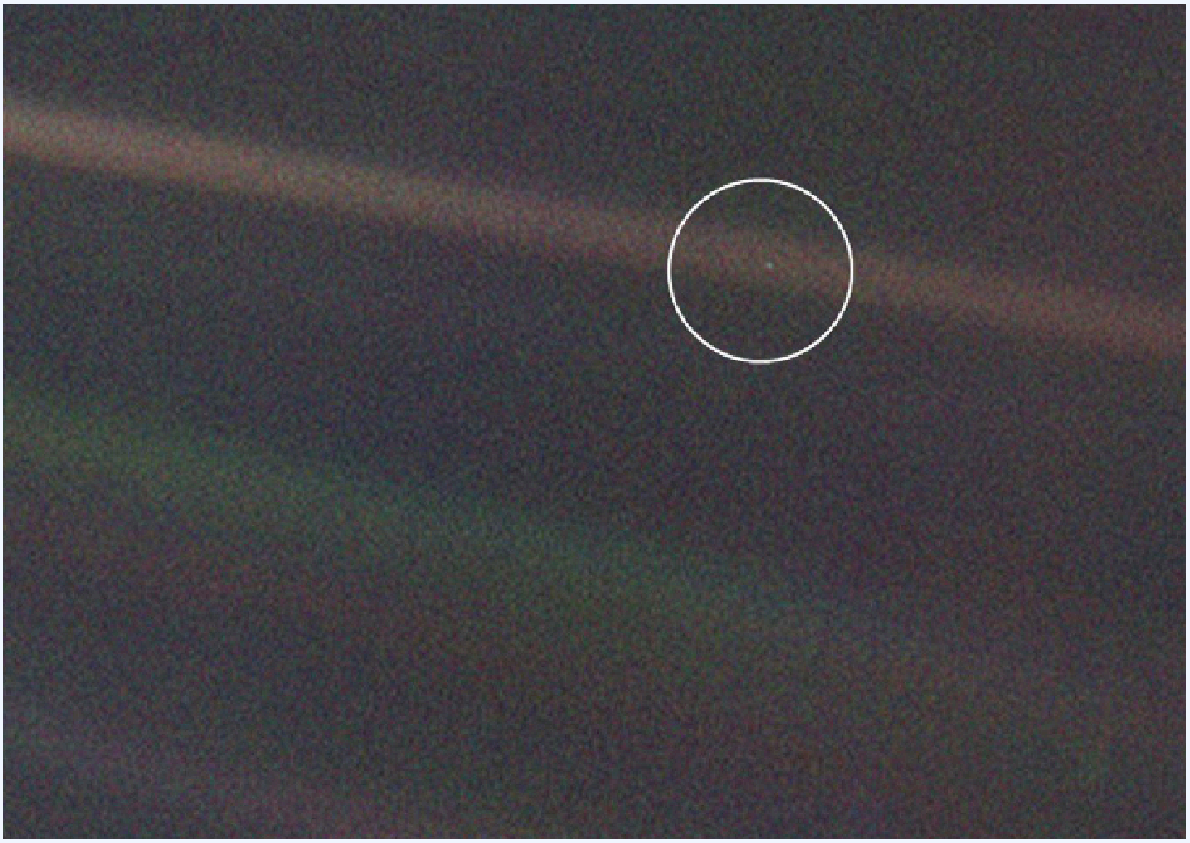


Figure 16.16. In this image, taken from 4 billion miles away, Earth appears as a "pale blue dot" representing less than a pixel's worth of light. Would this light reveal Earth as a habitable and inhabited world? Our search for life on exoplanets will depend on an ability to extract information about life from the faint light of faraway worlds.
Solar System Portrait – Earth as 'Pale Blue Dot' by NASA, NASA Licence

To be sure, we need to look for robust biospheres (atmospheres, surfaces, and/or oceans) capable of creating planet-scale change. Earth hosts such a biosphere: the composition of our atmosphere and the spectrum of light reflected from our planet differ considerably from what would be expected in the absence of life. Presently, Earth is the only body in our solar system for which this is true, despite the possibility that habitable conditions might prevail in the subsurface of Mars or inside the icy moons of the outer solar system. Even if life exists on these worlds, it is very unlikely that it could yield planet-scale changes that are both telescopically observable and clearly biological in origin.

What makes Earth "special" among the potentially habitable worlds in our solar system is that it has a photosynthetic biosphere. This requires the presence of liquid water at the planet's surface, where organisms have direct access to sunlight. The habitable zone concept focuses on this requirement for surface liquid water—even though we know that subsurface habitable conditions could prevail at more distant orbits—exactly because these worlds would have biospheres detectable at a distance.

Indeed, plants and photosynthetic microorganisms are so abundant at Earth's surface that they affect the color of the light that our planet reflects out into space—we appear greener in visible wavelengths and reflect more near-infrared light than we otherwise would. Moreover, photosynthesis has changed Earth's atmosphere at a large scale—more than 20% of our atmosphere comes from the photosynthetic waste product, oxygen. Such high levels would be very difficult to explain in the absence of life. Other gases, such as nitrous oxide and methane, when found simultaneously with oxygen, have also been suggested as possible indicators of life. When sufficiently abundant in an atmosphere, such gases could be detected by their effect on the spectrum of light that a planet emits or reflects. (As we saw in the chapter on exoplanets, astronomers today are beginning to have the capability of detecting the spectrum of the atmospheres of some planets orbiting other stars.)

Astronomers have thus concluded that, at least initially, a search for life outside our solar system should focus on exoplanets that are as much like Earth as possible—roughly Earth-size planets orbiting in the habitable zone—and look for the presence of gases in the atmosphere or colours in the visible spectrum that are hard to explain except by the presence of biology. Simple, right? In reality, the search for exoplanet life poses many challenges.

As you might imagine, this task is more challenging for planetary systems that are farther away and, in practical terms, this will limit our search to the habitable worlds closest to our own. Should we become limited to a very small number of nearby targets, it will also become important to consider the habitability of planets orbiting the M-dwarfs we discussed above.

If we manage to separate out a clean signal from the planet and find some features in the light spectrum that might be indicative of life, we will need to work hard to think of any nonbiological process that might account for them. "Life is the hypothesis of last resort," noted astronomer Carl Sagan—meaning that we must exhaust all other explanations for what we see before claiming to have found evidence of extraterrestrial biology. This requires some understanding of what processes might operate on worlds that we will know relatively little about; what we find on Earth can serve as a guide but also has potential to lead us astray (Figure 16.17 below).

Recall, for example, that it would be extremely difficult to account for the abundance of oxygen in Earth's atmosphere except by the presence of biology. But it has been hypothesized that oxygen could build up to substantial levels on planets orbiting M-dwarf stars through the action of ultraviolet radiation on the atmosphere—with no need for biology. It will be critical to understand where such “false positives” might exist in carrying out our search.

We need to understand that we might not be able to detect biospheres even if they exist. Life has flourished on Earth for perhaps 3.5 billion years, but the atmospheric “biosignatures” that, today, would supply good evidence for life to distant astronomers have not been present for all of that time. Oxygen, for example, accumulated to detectable levels in our atmosphere only a little over 2 billion years ago. Could life on Earth have been detected before that time? Scientists are working actively to understand what additional features might have provided evidence of life on Earth during that early history, and thereby help our chances of finding life beyond.

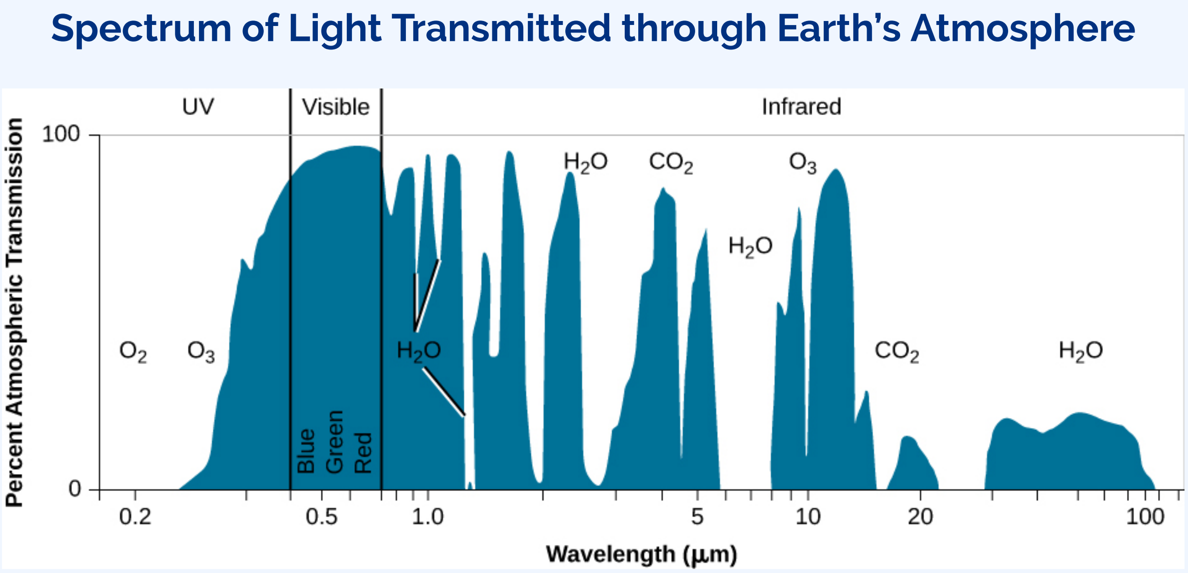


Figure 16.17. This graph shows wavelengths ranging from ultraviolet (far left) to infrared. The many downward “spikes” come from absorption of particular wavelengths by molecules in Earth’s atmosphere. Some of these compounds, like water and the combination oxygen/ozone and methane, might reveal Earth as both habitable and inhabited. We will have to rely on this sort of information to seek life on exoplanets, but our spectra will be of much poorer quality than this one, in part because we will receive so little light from the planet. (credit: modification of work by NASA)

Attribution

16.5 THE SEARCH FOR EXTRATERRESTRIAL INTELLIGENCE

Given all the developments discussed in this chapter, it seems likely that life could have developed on many planets around other stars. Even if that life is microbial, we saw that we may soon have ways to search for chemical biosignatures. This search is of fundamental importance for understanding biology, but it does not answer the question, "Are we alone?" that we raised at the beginning of this chapter. When we ask this question, many people think of other intelligent creatures, perhaps beings that have developed technology similar to our own. If any intelligent, technical civilizations have arisen, as has happened on Earth in the most recent blink of cosmic time, how could we make contact with them?

This problem is similar to making contact with people who live in a remote part of Earth. If students in the United States want to converse with students in Australia, for example, they have two choices. Either one group gets on an airplane and travels to meet the other, or they communicate by sending a message remotely. Given how expensive airline tickets are, most students would probably select the message route.

In the same way, if we want to get in touch with intelligent life around other stars, we can travel, or we can try to exchange messages. Because of the great distances involved, interstellar space travel would be very slow and prohibitively expensive. The fastest spacecraft the human species has built so far would take almost 80,000 years to get to the nearest star. While we could certainly design a faster craft, the more quickly we require it to travel, the greater the energy cost involved. To reach neighboring stars in less than a human life span, we would have to travel close to the speed of light. In that case, however, the expense would become truly astronomical.

Interstellar Travel

Bernard Oliver, an engineer with an abiding interest in life elsewhere, made a revealing calculation about the costs of rapid interstellar space travel. Since we do not know what sort of technology we (or other civilizations) might someday develop, Oliver considered a trip to the nearest star (and back again) in a spaceship with a "perfect engine"—one that would convert its fuel into energy with 100% efficiency. Even with a perfect engine, the energy cost of a single round-trip journey at 70% the speed of light turns out to be equivalent to several hundred thousand years' worth of total U.S. electrical energy consumption. The cost of such travel is literally out of this world.

This is one reason astronomers are so skeptical about claims that UFOs are spaceships from extraterrestrial civilizations. Given the distance and energy expense involved, it seems unlikely that the dozens of UFOs (and even UFO abductions) claimed each year could be visitors from other stars so fascinated by Earth civilization that they are willing to expend fantastically large amounts of energy or time to reach us. Nor does it seem credible that these visitors have made this long and expensive journey and then systematically avoided contacting our governments or political and intellectual leaders.

Not every UFO report has been explained (in many cases, the observations are sketchy or contradictory). But investigation almost always converts them to IFOs (identified flying objects) or NFOs (not-at-all flying objects). While some are hoaxes, others are natural phenomena, such as bright planets, ball lightning, fireballs (bright meteors), or even flocks of birds that landed in an oil slick to make their bellies reflective. Still others are human craft, such as private planes with some lights missing, or secret military aircraft. It is also interesting that the group of people who most avidly look at the night sky, the amateur astronomers, have never reported UFO sightings. Further, not a single UFO has ever left behind any physical evidence that can be tested in a laboratory and shown to be of nonterrestrial origin.

Another common aspect of belief that aliens are visiting Earth comes from people who have difficulty accepting human accomplishments. There are many books and TV shows, for example, that assert that humans could not have built the great pyramids of Egypt, and therefore they must have been built by aliens. The huge statues (called Moai) on Easter Island are also sometimes claimed to have been built by aliens. Some people even think that the accomplishments of space exploration today are based on alien technology.

However, the evidence from archaeology and history is clear: ancient monuments were built by ancient *people,* whose brains and ingenuity were every bit as capable as ours are today, even if they didn’t have electronic textbooks like you do.

Messages on Spacecraft

While space travel by living creatures seems very difficult, robot probes can travel over long distances and over long periods of time. Five spacecraft—two Pioneers, two Voyagers, and New Horizons—are now leaving the solar system. At their coasting speeds, they will take hundreds of thousands or millions of years to get anywhere close to another star. On the other hand, they were the first products of human technology to go beyond our home system, so we wanted to put messages on board to show where they came from.

Each Pioneer carries a plaque with a pictorial message engraved on a gold-anodized aluminum plate (Figure 16.18 below). The Voyagers, launched in 1977, have audio and video records attached, which allowed the inclusion of over 100 photographs and a selection of music from around the world. Given the enormous space between stars in our section of the Galaxy, it is very unlikely that these messages will ever be received by anyone. They are more like a note in a bottle thrown into the sea by a shipwrecked sailor, with no realistic expectation of its being found soon but a slim hope that perhaps someday, somehow, someone will know of the sender’s fate.

Interstellar Messages

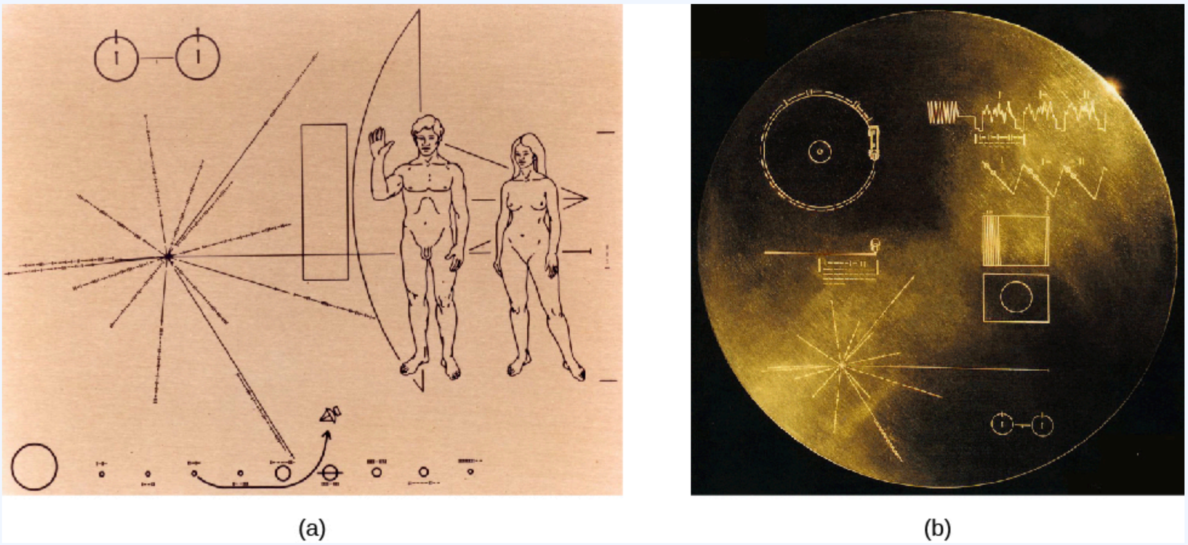


Figure 16.18. (a) This is the image engraved on the plaques aboard the Pioneer 10 and 11 spacecraft. The human figures are drawn in proportion to the spacecraft, which is shown behind them. The Sun and planets in the solar system can be seen at the bottom, with the trajectory that the spacecraft followed. The lines and markings in the left centre show the positions and pulse periods for a number of pulsars, which might help locate the spacecraft's origins in space and time.
Pioneer 10 and 11 by NASA, NASA Licence (b) Encoded onto a gold-coated copper disk, the Voyager record contains 118 photographs, 90 minutes of music from around the world, greetings in almost 60 languages, and other audio material. It is a summary of the sights and sounds of Earth.
Golden Record by JPL, JPL Media Licence

THE VOYAGER MESSAGE

An Excerpt from the Voyager Record:

> “We cast this message into the cosmos. It is likely to survive a billion years into our future, when our civilization is profoundly altered. . . . If [another] civilization intercepts Voyager and can understand these recorded contents, here is our message:
>
> This is a present from a small, distant world, a token of our sounds, our science, our images, our music, our thoughts, and our feelings. We are attempting to survive our time so we may live into yours. We hope, someday, having solved the problems we face, to join a community of galactic civilizations. This record represents our hope and our determination, and our goodwill in a vast and awesome universe.”
>
> —Jimmy Carter, President of the United States of America, June 16, 1977

Communicating with the Stars

If direct visits among stars are unlikely, we must turn to the alternative for making contact: exchanging messages. Here the news is a lot better. We already use a messenger—light or, more generally, electromagnetic waves—that moves through space at the fastest speed in the universe. Traveling at 300,000 kilometres per second, light reaches the nearest star in only 4 years and does so at a tiny fraction of the cost of sending material objects. These advantages are so clear and obvious that we assume they will occur to any other species of intelligent beings that develop technology.

However, we have access to a wide spectrum of electromagnetic radiation, ranging from the longest-wavelength radio waves to the shortest-wavelength gamma rays. Which would be the best for interstellar communication? It would not be smart to select a wavelength that is easily absorbed by interstellar gas and dust, or one that is unlikely to penetrate the atmosphere of a planet like ours. Nor would we want to pick a wavelength that has lots of competition for attention in our neighbourhood.

One final criterion makes the selection easier: we want the radiation to be inexpensive enough to produce in large quantities. When we consider all these requirements, radio waves turn out to be the best answer. Being the lowest-frequency (and lowest-energy) band of the spectrum, they are not very expensive to produce, and we already use them extensively for communications on Earth. They are not significantly absorbed by interstellar dust and gas. With some exceptions, they easily pass through Earth’s atmosphere and through the atmospheres of the other planets we are acquainted with.

The Cosmic Haystack

Having made the decision that radio is the most likely means of communication among intelligent civilizations, we still have many questions and a daunting task ahead of us. Shall we *send* a message, or try to *receive* one? Obviously, if every civilization decides to receive only, then no one will be sending, and everyone will be disappointed. On the other hand, it may be appropriate for us to *begin* by listening, since we are likely to be among the most primitive civilizations in the Galaxy who are interested in exchanging messages.

We do not make this statement to insult the human species (which, with certain exceptions, we are rather fond of). Instead, we base it on the fact that humans have had the ability to receive (or send) a radio message across interstellar distances for only a few decades. Compared to the ages of the stars and the Galaxy, this is a mere instant. If there are civilizations out there that are ahead of us in development by even a short time (in the cosmic sense), they are likely to have a technology head start of many, many years.

In other words, we, who have just started, may well be the "youngest" species in the Galaxy with this capability (see the discussion in shaded box below). Just as the youngest members of a community are often told to be quiet and listen to their elders for a while before they say something foolish, so may we want to begin our exercise in extraterrestrial communication by listening.

Even restricting our activities to listening, however, leaves us with an array of challenging questions. For example, if an extraterrestrial civilization's signal is too weak to be detected by our present-day radio telescopes, we will not detect them. In addition, it would be very expensive for an extraterrestrial civilization to broadcast on a huge number of channels. Most likely, they select one or a few channels for their particular message. Communicating on a narrow band of channels also helps distinguish an artificial message from the radio static that comes from natural cosmic processes. But the radio band contains an astronomically large number of possible channels. How can we know in advance which one they have selected, and how they have coded their message into the signal?

Table 16.1 summarizes these and other factors that scientists must grapple with when trying to tune in to radio messages from distant civilizations. Because their success depends on either guessing right about so many factors or searching through all the possibilities for each factor, some scientists have compared their quest to looking for a needle in a haystack. Thus, they like to say that the list of factors in Table 16.1 defines the *cosmic haystack problem*.

Table 16.1 The Cosmic Haystack Problem: Some Questions about an Extraterrestrial Message

Factors
From which direction (which star) is the message coming?
On what channels (or frequencies) is the message being broadcast?
How wide in frequency is the channel?
How strong is the signal (can our radio telescopes detect it)?
Is the signal continuous, or does it shut off at times (as, for example, a lighthouse beam does when it turns away from us)?
Does the signal drift (change) in frequency because of the changing relative motion of the source and the receiver?
How is the message encoded in the signal (how do we decipher it)?
Can we even recognize a message from a completely alien species? Might it take a form we don't at all expect?

Radio Searches

Although the cosmic haystack problem seems daunting, many other research problems in astronomy also require a large investment of time, equipment, and patient effort. And, of course, if we don't search, we're sure not to find anything.

The very first search was conducted by astronomer Frank Drake in 1960, using the 85-foot antenna at the National Radio Astronomy Observatory (Figure 16.19 below). Called Project Ozma, after the queen of the exotic Land of Oz in the children's stories of L. Frank Baum, his experiment involved looking at about 7200 channels and two nearby stars over a period of 200 hours. Although he found nothing, Drake demonstrated that we had the technology to do such a search, and set the stage for the more sophisticated projects that followed.

Project Ozma and the Allen Telescope Array

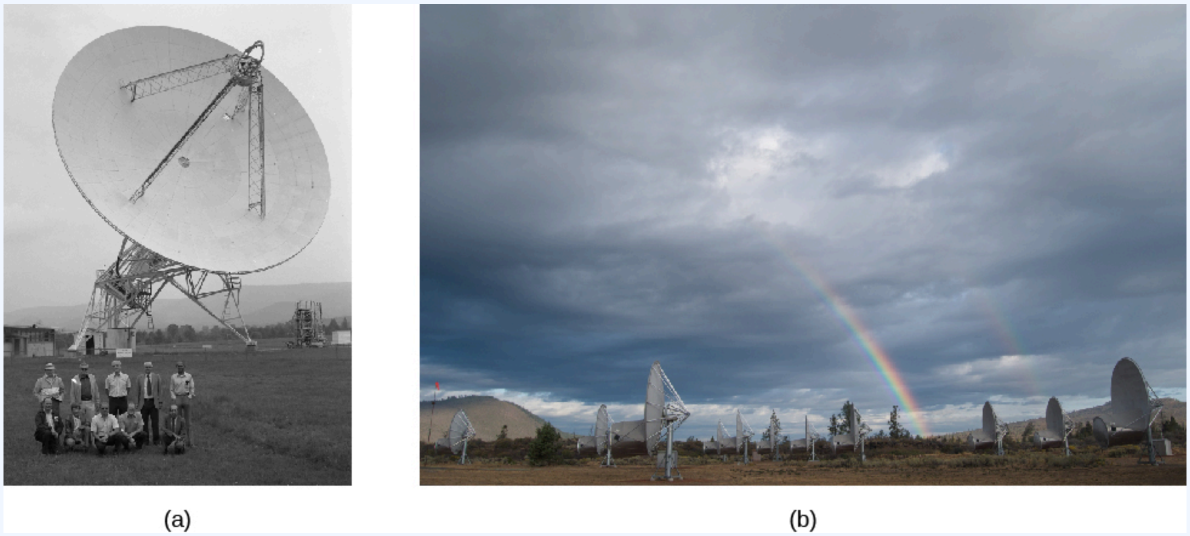


Figure 16.19. (a) This 25th anniversary photo shows some members of the Project Ozma team standing in front of the 85-foot radio telescope with which the 1960 search for extraterrestrial messages was performed. Frank Drake is in the back row, second from the right.
85 Foot Telescope and Ozma team, 1985 by NRAO, NRAO Image Use Policy (b) The Allen Telescope Array in California is made up of 42 small antennas linked together. This system allows simultaneous observations of multiple sources with millions of separate frequency channels. (credit b: modification of work by Colby Gutierrez-Kraybill Open Stax CC-BY)

Receivers are constantly improving, and the sensitivity of SETI programs—**SETI** stands for the search for extraterrestrial life—is advancing rapidly. Equally important, modern electronics and software allow simultaneous searches on millions of frequencies (channels). If we can thus cover a broad frequency range, the cosmic haystack problem of guessing the right frequency largely goes away. One powerful telescope array (funded with an initial contribution from Microsoft founder Paul Allen) that is built for SETI searches is the Allen Telescope in Northern California. Other radio telescopes being used for such searches include the giant Arecibo radio dish in Puerto Rico and the Green Bank Telescope in West Virginia, which is the largest steerable radio telescope in the world.

What kind of signals do we hope to pick up? We on Earth are inadvertently sending out a flood of radio signals, dominated by military radar systems. This is a kind of leakage signal, similar to the wasted light energy that is beamed upward by poorly designed streetlights and advertising signs. Could we detect a similar leakage of radio signals from another civilization? The answer is just barely, but only for the nearest stars. For the most part, therefore, current radio SETI searches are looking for beacons, assuming that civilizations might be intentionally drawing attention to themselves or perhaps sending a message to another world or outpost that lies in our direction. Our prospects for success depend on how often civilizations arise, how long they last, and how patient they are about broadcasting their locations to the cosmos.

Jill Tarter: Trying to Make Contact

1997 was quite a year for Jill Cornell Tarter (Figure 16.20), one of the world's leading scientists in the SETI field. The SETI Institute announced that she would be the recipient of its first endowed chair (the equivalent of an endowed research professorship) named in honour of Bernard Oliver. The National Science Foundation approved a proposal by a group of scientists and educators she headed to develop an innovative hands-on high school curriculum based on the ideas of cosmic evolution (the topics of this chapter). And, at roughly the same time, she was being besieged with requests for media interviews as news reports identified her as the model for Ellie Arroway, the protagonist of *Contact*, Carl Sagan's best-selling novel about SETI. The book had been made into a high-budget science fiction film, starring Jodie Foster, who had talked with Tarter before taking the role.

Jill Tarter

Figure 16.20. Jill Tarter (credit: Christian Schidlowski)
Jill Tarter by Christian Schidlowski , used under fair dealing. All rights reserved.

Tarter is quick to point out, "Carl Sagan wrote a book about a woman who does what I do, not about me." Still, as the only woman in such a senior position in the small field of SETI, she was the centre of a great deal of public attention. (However, colleagues and reporters pointed out that this was nothing compared to what would happen if her search for radio signals from other civilizations recorded a success.)

Being the only woman in a group is not a new situation to Tarter, who often found herself the only woman in her advanced science or math classes. Her father had encouraged her, both in her interest in science and her "tinkering." As an undergraduate at Cornell University, she majored in engineering physics. That training became key to putting together and maintaining the complex systems that automatically scan for signals from other civilizations.

Switching to astrophysics for her graduate studies, she wrote a PhD thesis that, among other topics, considered the formation of failed stars—those whose mass was not sufficient to ignite the nuclear reactions that power more massive stars like our own Sun. Tarter coined the term “brown dwarf” for these small, dim objects, and it has remained the name astronomers use ever since.

It was while she was still in graduate school that Stuart Bowyer, one of her professors at the University of California, Berkeley, asked her if she wanted to be involved in a small experiment to siphon off a bit of radiation from a radio telescope as astronomers used it year in and year out and see if there was any hint of an intelligently coded radio message buried in the radio noise. Her engineering and computer programming skills became essential to the project, and soon she was hooked on the search for life elsewhere.

Thus began an illustrious career working full time searching for extraterrestrial civilizations, leading Jill Tarter to receive many awards, including being elected fellow of the American Association for the Advancement of Science in 2002, the Adler Planetarium Women in Space Science Award in 2003, and a 2009 TED Prize, among others.

Example 16.1

The Drake Equation

At the first scientific meeting devoted to SETI, Frank Drake wrote an equation on the blackboard that took the difficult question of estimating the number of civilizations in the Galaxy and broke it down into a series of smaller, more manageable questions. Ever since then, both astronomers and students have used this **Drake equation** as a means of approaching the most challenging question: How likely is it that we are alone? Since this is at present an unanswerable question, astronomer Jill Tarter has called the Drake equation a “way of organizing our ignorance.” (See Figure 16.21 below)

Drake Equation

Figure 16.21. A plaque at the National Radio Astronomy Observatory commemorates the conference where the equation was first discussed.
National Radio Astronomy Observatory Green Bank by Nick Strobel, used under fair dealing. All rights reserved.

The form of the Drake equation is very simple. To estimate the number of communicating civilizations that currently exist in the Galaxy (we will define these terms more carefully in a moment), we multiply the rate of formation of such civilizations (number per year) by their average lifetime (in years). In symbols,

$$N = R_{\text{total}} \times L$$

To make this formula easier to use (and more interesting), however, Drake separated the rate of formation R_{total} into a series of probabilities:

$$R_{\text{total}} = R_{\text{star}} \times f_{\text{p}} \times f_{\text{e}} \times f_{\text{l}} \times f_{\text{i}} \times f_{\text{c}}$$

R_{star} is the rate of formation of stars like the Sun in our Galaxy, which is about 10 stars per year.

Each of the other terms is a fraction or probability (less than or equal to 1.0), and the product of all these probabilities is itself the total probability that each star will have an intelligent, technological, communicating civilization that we might want to talk to. We have:

- f_p = the fraction of these stars with planets
- f_e = the fraction of the planetary systems that include habitable planets
- f_l = the fraction of habitable planets that actually support life
- f_i = the fraction of inhabited planets that develop advanced intelligence
- f_c = the fraction of these intelligent civilizations that develop science and the technology to build radio telescopes and transmitters

Each of these factors can be discussed and perhaps evaluated, but we must guess at many of the values. In particular, we don't know how to calculate the probability of something that happened once on Earth but has not been observed elsewhere—and these include the development of life, of intelligent life, and of technological life (the last three factors in the equation). One important advance in estimating the terms of the Drake equation comes from the recent discovery of exoplanets. When the Drake equation was first written, no one had any idea whether planets and planetary systems were common. Now we know they are—another example of the Copernican principle.

Solution

Even if we don't know the answers, we can make some guesses and calculate the resulting number *N*. Let's start with the optimism implicit in the Copernican principle and set the last three terms equal to 1.0. If *R* is 10 stars/year and if we measure the average lifetime of a technological civilization in years, the units of years cancel. If we also assume that f_p is 0.1, and f_e is 1.0, the equation becomes

$$N={R}_{\text{total}}\phantom{\rule{0.2em}{0ex}}\times\phantom{\rule{0.2em}{0ex}}L=L$$

Now we see the importance of the term L, the lifetime of a communicating civilization (measured in years). We have had this capability (to communicate at the distances of the stars) for only a few decades.

Exercise 16.1

Suppose we assume that this stage in our history lasts only one century.

Solution

With our optimistic assumptions about the other factors, L = 100 years and N = 100 such civilizations in the entire Galaxy. In that case, there are so few other civilizations like ours that we are unlikely to detect any signals in a SETI search. But suppose the average lifetime is a million years; in that case, there are a million such civilizations in the Galaxy, and some of them may be within range for radio communication.

The most important conclusion from this calculation is that even if we are extremely optimistic about the probabilities, the only way we can expect success from SETI is if other civilizations are much older (and hence probably much more advanced) than ours.

PBS has an interactive version of the Drake Equation

https://www.pbs.org/wgbh/nova/space/drake-equation.html

SETI outside the Radio Realm

For the reasons discussed above, most SETI programs search for signals at radio wavelengths. But in science, if there are other approaches to answering an unsolved question, we don't want to neglect them. So astronomers have been thinking about other ways we could pick up evidence for the existence of technologically advanced civilizations.

Recently, technology has allowed astronomers to expand the search into the domain of visible light. You might think that it would be hopeless to try to detect a flash of visible light from a planet given the brilliance of the star it orbits. This is why we usually cannot measure the reflected light of planets around other stars. The feeble light of the planet is simply swamped by the "big light" in the neighborhood. So another civilization would need a mighty strong beacon to compete with their star.

However, in recent years, human engineers have learned how to make flashes of light brighter than the Sun. The trick is to "turn on" the light for a very brief time, so that the costs are manageable. But ultra-bright, ultra-short laser pulses (operating for periods of a billionth of a second) can pack a lot of energy and can be coded to carry a message. We also have the technology to detect such short pulses—not with human senses, but with special detectors that can be "tuned" to hunt automatically for such short bursts of light from nearby stars.

Why would any civilization try to outshine its own star in this way? It turns out that the cost of sending an ultra-short laser pulse in the direction of a few promising stars can be less than the cost of sweeping a continuous radio message across the whole sky. Or perhaps they, too, have a special fondness for light messages because one of their senses evolved using light. Several programs are now experimenting with "optical SETI" searches, which can be done with only a modest telescope. (The term *optical* here means using visible light.)

If we let our imaginations expand, we might think of other possibilities. What if a truly advanced civilization should decide to (or need to) renovate its planetary system to maximize the area for life? It could do so by breaking apart some planets or moons and building a ring of solid material that surrounds or encloses the star and intercepts some or all of its light. This huge artificial ring or sphere might glow very brightly at infrared wavelengths, as the starlight it receives is eventually converted to heat and re-radiated into space. That infrared radiation could be detected by our instruments, and searches for such infrared sources are also underway (Figure 16.22).

Wide-Field Infrared Survey Explorer (WISE)

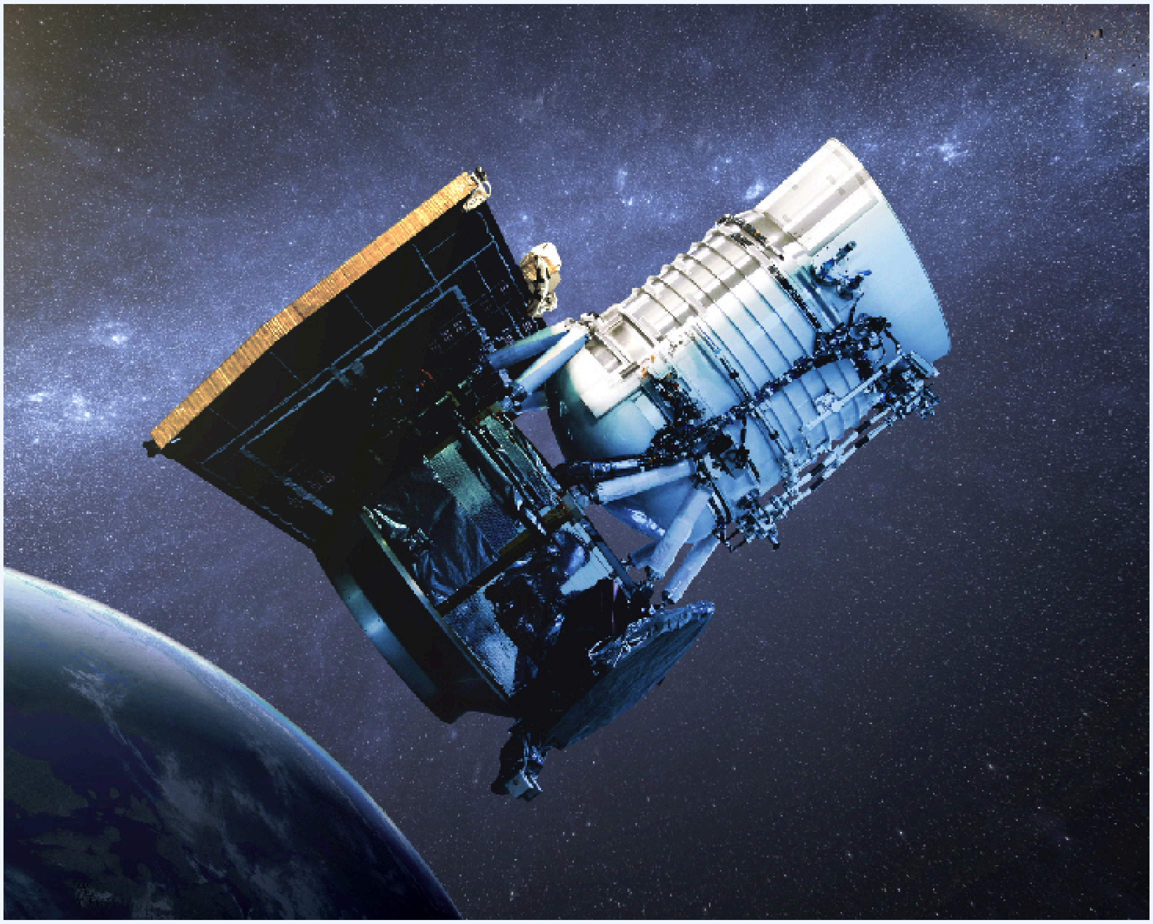


Figure 16.22. Astronomers have used this infrared satellite to search for infrared signatures of enormous construction projects by very advanced civilizations, but their first survey did not reveal any. (credit: modification of work by NASA/JPL-Caltech)
NEOWISE: Back to Hunt More Asteroids (Artist Concept) by NASA/JPL-Caltech, JPL Image Use Policy.

Should We Transmit in Addition to Listening?

Our planet has some leakage of radio waves into space, from FM radio, television, military radars, and communication between Earth and our orbiting spacecraft. However, such leakage radiation is still quite weak, and therefore difficult to detect at the distances of the stars, at least with the radio technology we have. So at the present time our attempts to communicate with other civilizations that may be out there mostly involve trying to receive messages, but not sending any ourselves.

Some scientists, however, think that it is inconsistent to search for beacons from other civilizations without announcing our presence in a similar way. (We discussed earlier the problem that if every other civilization confined itself to listening, no one would ever get in touch.) So, should we be making regular attempts at sending easily decoded messages into space? Some scientists warn that our civilization is too immature and defenseless to announce ourselves at this early point in our development. The decision whether to transmit or not turns out to be an interesting reflection of how we feel about ourselves and our place in the universe.

Discussions of transmission raise the question of who should speak for planet Earth. Today, anyone and everyone can broadcast radio signals, and many businesses, religious groups, and governments do. It would be a modest step for the same organizations to use or build large radio telescopes and begin intentional transmissions that are much stronger than the signals that leak from Earth today. And if we intercept a signal from an alien civilization, then the issue arises whether to reply.

Who should make the decision about whether, when, and how humanity announces itself to the cosmos? Is there freedom of speech when it comes to sending radio messages to other civilizations? Do all the nations of Earth have to agree before we send a signal strong enough that it has a serious chance of being received at the distances of the stars? How our species reaches a decision about these kinds of questions may well be a test of whether or not there is intelligent life on Earth.

Conclusion

Whether or not we ultimately turn out to be the only intelligent species in our part of the Galaxy, our exploration of the cosmos will surely continue. An important part of that exploration will still be the search for biomarkers from inhabited planets that have not produced technological creatures that send out radio signals. After all, creatures like butterflies and dolphins may never build radio antennas, but we are happy to share our planet with them and would be delighted to find their counterparts on other worlds.

Whether or not life exists elsewhere is just one of the unsolved problems in astronomy that we have discussed in this book. A humble acknowledgment of how much we have left to learn about the universe is one of the fundamental hallmarks of science. This should not, however, prevent us from feeling exhilarated about how much we have already managed to discover, and feeling curious about what else we might find out in the years to come.

Our progress report on the ideas of astronomy ends here, but we hope that your interest in the universe does not. We hope you will keep up with developments in astronomy through media and online, or by going to an occasional public lecture by a local scientist. Who, after all, can even guess all the amazing things that future research projects will reveal about both the universe and our connection with it?

Attribution

16.6 KEY TERMS

Acidophiles: the most acid-tolerant organisms that are capable of living at pH values near zero—about ten million times more acidic than human blood. 16.3

Alkaliphiles: organisms that can grow at pH levels of about 13, which is comparable to the pH of household bleach and almost a million times more alkaline than human blood. 16.3

Amino acids: organic compounds that are the molecular building blocks of proteins. 16.3

Astrobiology: the multidisciplinary study of life in the universe: its origin, evolution, distribution, and fate; similar terms are exobiology and bioastronomy. 16.3

Biomarker: evidence of the presence of life, especially a global indication of life on a planet that could be detected remotely (such as an unusual atmospheric composition). 16.4

Continuously habitable zone: defined by the range of orbits that would remain within the habitable zone during the entire lifetime of the star system. 16.4

Copernican principle: the idea that there is nothing special about our place in the universe. 16.2

Deinococcus radiodurans: an organism that can tolerate ionizing radiation (such as that released by radioactive elements) a thousand times more intense than humans would be able to withstand. 16.3

DNA (deoxyribonucleic acid): a molecule that stores information about how to replicate a cell and its chemical and structural components. 16.3

Drake equation: a formula for estimating the number of intelligent, technological civilizations in our Galaxy, first suggested by Frank Drake. 16.5

Extremophile: an organism (usually a microbe) that tolerates or even thrives under conditions that most of the life around us would consider hostile, such as very high or low temperature or acidity. 16.3

Fermi paradox: question asked by physicist Enrico Fermi based on teh Copernican principle's suggestion that intelligent life like us might be common: "Where are they? If life and intelligence are common and have such tremendous capacity for growth, why is there not a network of galactic civilizations whose presence extends even into a "latecomer" planetary system like ours?" 16.2

Gene: the basic functional unit that carries the genetic (hereditary) material contained in a cell. 16.3

Habitable environment: an environment capable of hosting life. 16.3

Habitable zone: a region around a star where suitable conditions might exist for life. 16.4

Organic compound: a compound containing carbon, especially a complex carbon compound; not necessarily produced by life. 16.3

Organic molecule: a combination of carbon and other atoms—primarily hydrogen, oxygen, nitrogen, phosphorus, and sulfur—some of which serve as the basis for our biochemistry. 16.2

Oxidizing chemistry: a chemistry where molecules tend to accept electrons readily. 16.4

Photosynthesis: a complex sequence of chemical reactions through which some living things can use sunlight to manufacture products that store energy (such as carbohydrates), releasing oxygen as one by-product. 16.3

Protein: a key biological molecule that provides the structure and function of the body's tissues and organs, and essentially carries out the chemical work of the cell.

Psychrophiles: cold-adapted cells. 16.3

Reducing chemistry: a chemistry where molecules tend to give up electrons readily. 16.4

RNA (ribonucleic acid): a molecule that aids in the flow of genetic information from DNA to proteins. 16.3

SETI: the search for extraterrestrial intelligence; usually applied to searches for radio signals from other civilizations. 16.5

Stromatolites: solid, layered rock formations that are thought to be the fossils of oxygen-producing photosynthetic bacteria in rocks that are 3.5 billion years old. 16.3

Thermophile: an organism that can tolerate high temperatures. 16.3

Tholins: complex organic compounds formed when the Sun's ultraviolet light breaks apart and recombines the nitrogen and methane present in the Titan's upper atmosphere. 16.4

APPENDIX A: HOW TO STUDY FOR AN INTRODUCTORY ASTRONOMY CLASS

In this brief appendix, we want to give you some hints for the effective study of astronomy. These suggestions are based on ideas from good teachers and good students around the United States. Your professor will probably have other, more specific suggestions for doing well in your class.

Astronomy, the study of the universe beyond the borders of our planet, is one of the most exciting and rapidly changing branches of science. Even scientists from other fields often confess to having had a lifelong interest in astronomy, though they may now be doing something earthbound—like biology, chemistry, engineering, or writing software.

But some of the things that make astronomy so interesting also make it a challenge for the beginning student. The universe is a big place, full of objects and processes that do not have familiar counterparts here on Earth. Like a visitor to a new country, it will take you a while to feel familiar with the territory or the local customs. Astronomy, like other sciences, also has its own special vocabulary, some of which you will have to learn to communicate well with your professor and classmates.

Still, hundreds of thousands of non-science majors take an introductory astronomy course every year, and surveys show that students from a wide range of backgrounds have succeeded in (and even enjoyed) these classes. Astronomy is for everyone, not just those who are "science oriented."

So, here are some suggestions to help you increase your chances of doing well in your astronomy class.

1. The best advice we can give you is to be sure to leave enough time in your schedule to study the material in this class regularly. It sounds obvious, but it is not very easy to catch up with a subject like astronomy by trying to do everything just before an exam. (As astronomers like to put it, you can't learn the whole universe in one night!) Try to put aside some part of each day, or every other day, when you can have uninterrupted time for reading and studying astronomy.
2. In class, put your phone away and focus on the class activities. If you have to use a laptop or tablet in class, make a pact with yourself that you will *not* check email, get on social media, or play games during class. A number of careful studies of student behaviour and grades have shown that students are not as good at such multi-tasking as they think they are, and that students who do *not* use screens during class get significantly better grades in the end.
3. Try to take careful notes during class. Many students start college without good note-taking habits. If you are not a good note-taker, try to get some help. Many colleges and universities have student learning centres that offer short courses, workbooks, tutors, or videos on developing good study habits. Good note-taking skills will also be useful for many jobs or activities you are likely get involved with after

college.

4. Try to read each assignment in the textbook twice, once before it is discussed in class, and once afterwards. You can highlight text or make notes in your copy of this electronic book: just move your cursor over some text and a box with highlight colours and room for notes will become visible.
5. Form a small astronomy study group with people in your class. Get together with them regularly and discuss what you have been learning. Also, focus on the topics that may be giving group members trouble. Make up sample exam questions and make sure everyone in the group can answer them confidently. If you have always studied alone, you may at first resist this idea, but don't be too hasty to say no. Study groups are a very effective way to digest a large amount of new information.
6. Before each exam, create a concise outline of the main ideas discussed in class and presented in your text. Compare your outline with those of other students as a check on your own study habits.
7. If your professor suggests doing web-based sample quizzes, or looking at online apps, animations, or study guides, take advantage of these resources to enhance your studying.
8. At the end of each chapter in this textbook you, will find four kinds of questions. The Collaborative Group Activities are designed to encourage you follow up on the material in the chapter as a group, rather than individually. Review Questions help you see if you have learned the material in the chapter. Thought Questions test deeper understanding by asking you to apply your knowledge to new situations. And Figuring for Yourself exercises test and extend some of the mathematical examples in the chapter. (Not all professors will use the math sections; if they don't, you may not have homework from this section.)
9. If you find a topic in the text or in class especially difficult or interesting, talk to your professor or teaching assistant. Many students are scared to show their ignorance in front of their teacher, but we can assure you that most professors and TA's *like* it when students come to office hours and show that they care enough about the course to ask for help.
10. Don't stay up all night before a test and then expect your mind to respond well. For the same reason, don't eat a big meal just before a test, since we all get a little sleepy and don't think as clearly after a big meal. Take many deep breaths and try to relax during the test itself.
11. Don't be too hard on yourself! If astronomy is new to you, many of the ideas and terms in this book may be unfamiliar. Astronomy is like any new language: it may take a while to become a good conversationalist. Practice as much as you can, but also realize that it is natural to feel overwhelmed by the vastness of the universe and the variety of things that are going on in it.

Attribution

"A | How to Study for an Introductory Astronomy Class" from Astronomy 2e by Andrew Fraknoi, David

APPENDIX B: ASTRONOMY WEBSITES, IMAGES, AND APPS

Throughout the textbook, we suggest useful resources for students on the specific topics in a given chapter. Here, we offer some websites for exploring astronomy in general, plus good sites for viewing and downloading the best astronomy images, and guides to astronomical apps for smartphones and tablets. This is not an exhaustive listing, but merely a series of suggestions to whet the appetite of those wanting to go beyond the textbook.

Websites for Exploring Astronomy in General

Astronomical Organizations

Amateur Astronomy Clubs. In most large cities and a number of rural areas, there are *amateur astronomy clubs*, where those interested in the hobby of astronomy gather to observe the sky, share telescopes, hear speakers, and help educate the public about the night sky. To find an astronomy club near you, you can try the following sites:

- Night Sky Network club finder.
- *Sky & Telescope* Magazine astronomy clubs and organizations
- *Astronomy* Magazine club finder
- Astronomical League astronomy clubs and societies
- Go-Astronomy club search

American Astronomical Society: Composed mainly of professional astronomers. They have an active education office and various materials for students and the public on the education pages of their website.

Astronomical League: The league is the umbrella organization of American astronomy clubs. They offer a newsletter, national observing programs, and support for how to form and support a club.

Astronomical Society of the Pacific: Founded in 1889, this international society is devoted to astronomy education and outreach. They have programs, publications, and materials for families, teachers, amateur astronomers, museum guides, and anyone interested in astronomy.

European Space Agency (ESA): Information on European space missions with an excellent gallery of images.

International Astronomical Union (IAU): International organization for professional astronomers; see the menu choice “IAU for the Public” for information on naming astronomical objects and other topics of interest to students.

International Dark-Sky Association: Dedicated to combating light pollution, the encroachment of stray light that wastes energy and washes out the glories of the night sky.

NASA: NASA has a wide range of information on its many websites; the trick is to find what you need. Most space missions and NASA centres have their own sites.

Planetary Society: Founded by the late Carl Sagan and others, this group works to encourage planetary exploration and the search for life elsewhere. While much of their work is advocacy, they have some educational outreach too.

Royal Astronomical Society of Canada: Unites professional and amateur astronomers around Canada; has 28 centres with local activities, plus national magazines and meetings.

Some Astronomical Publications Students Can Read

Astronomy Now: A colourful British monthly, with excellent articles about astronomy, the history of astronomy, and stargazing.

Astronomy: Has the largest circulation of any magazine devoted to the universe and is designed especially for astronomy hobbyists and armchair astronomers.

Free Astronomy: A new web-based publication, with European roots.

Scientific American: Offers one astronomy article about every other issue. These articles, a number of which are reproduced on their website, are at a slightly higher level, but—often being written by the astronomers who have done the work—are authoritative and current.

Sky & Telescope: An older and somewhat higher-level magazine for astronomy hobbyists. Many noted astronomers write for this publication.

Sky News: A Canadian publication, featuring both astronomy and stargazing information. It also lists Canadian events for hobbyists.

StarDate: Magazine that accompanies the brief radio program, with a useful website for beginners.

Sites that Cover Astronomy News

Portal to the Universe: A site that gathers online astronomy and space news items, blogs, and pictures.

Science@NASA news stories and newscasts: Well-written stories with, of course, a NASA focus.

Space.com: A commercial site, but with wide coverage of space and astronomy news.

Universe Today: Another commercial site, with good articles by science journalists, but a lot of ads.

Sites for Answering Astronomical Questions

Ask an Astrobiologist: On this site from the National Astrobiology Institute at NASA, astronomer David Morrison answered questions about the search for life on other planets, the origin of life on Earth, and many other topics.

Ask an Astronomer at Lick Observatory: Graduate students and staff members at this California observatory answered selected astronomy questions, particularly from high school students.

Ask an Astrophysicist: Questions and answers at NASA's Laboratory for High-Energy Astrophysics focus on X-ray and gamma-ray astronomy, and such objects as black holes, quasars, and supernovae.

Ask an Infrared Astronomer: A site from the California Institute of Technology, with an archive focusing on infrared (heat-ray) astronomy and the discoveries it makes about cool objects in the universe. No longer taking new questions.

Ask the Astronomer: This site, run by astronomer Sten Odenwald, is no longer active, but lists 3001 answers to questions asked in the mid-1990s. They are nicely organized by topic.

Ask the Experts at PhysLink: Lots of physics questions answered, with some astronomy as well, at this physics education site. Most answers are by physics teachers, not astronomers. Still taking new questions.

Ask the Space Scientist: An archive of questions about the Sun and its interactions with Earth, answered by astronomer Sten Odenwald. Not accepting new questions.

Curious about Astronomy?: An ask-an-astronomer site run by graduate students and professors of astronomy at Cornell University. Has searchable archives and is still answering new questions.

Miscellaneous Sites of Interest

A Guide to Careers in Astronomy: From the American Astronomical Society.

Astronomical Pseudo-Science: A Skeptic's Resource List: Readings and websites that analyze such claims as astrology, UFOs, moon-landing denial, creationism, human faces on other worlds, astronomical disasters, and more.

Astronomy for Beginners: A page to find resources for getting into amateur astronomy.

Black Lives in Astronomy.

Contributions of Women to Astronomy.

Music Inspired by Astronomy.

Science Fiction Stories with Good Astronomy and Physics.

Astronomy Calendar: Entitled "This Day in Astronomical History," this list provides 158 astronomical anniversaries or events that you can celebrate.

Unheard Voices: The Astronomy of Many Cultures: A guide to resources about the astronomy of native, African, Asian, and other non-Western groups.

Videos To Go with Each Chapter of this Textbook: Brief Astronomy Videos.

Selected Websites for Viewing and Downloading Astronomical Images

The Top Image Sites

Astronomy Picture of the Day: Two space scientists scour the internet and feature one interesting astronomy image each day.

European Southern Observatory Photo Gallery: Magnificent colour images from ESO's largest telescopes. See the topical menu at the top.

Hubble Space Telescope Images: Click on the menu of categories or search for any object the telescope may have photographed.

James Webb Space Telescope Images: Growing library of infrared images from JWST; use the menu of categories or search for an object you may have seen on the news.

National Optical Astronomy Observatories Image Gallery: Growing archive of images from the many telescopes that are at the United States' National Observatories.

Planetary Photojournal: Features thousands of images from NASA's extensive set of planetary exploration missions with a good search menu. Does not include most of the missions from other countries.

The World at Night: Dramatic night-sky images by professional photographers who are amateur astronomers. Note that while many of the astronomy sites allow free use of their images, these are copyrighted by photographers who make their living selling them.

Other Useful General Galleries

Anglo-Australian Observatory: Great copyrighted colour images by leading astro-photographer David Malin and others.

Canada-France-Hawaii Telescope: Remarkable colour images from a major telescope on top of the Maunakea peak in Hawaii.

European Space Agency Gallery: Access images from such missions as Mars Express, Rosetta, and Herschel.

Gemini Observatory Images: Images from a pair of large telescopes in the northern and the southern hemispheres.

Isaac Newton Group of Telescopes Image Gallery: Beautiful images from the Herschel, Newton, and Kapteyn telescopes on La Palma.

National Radio Astronomy Observatory Image Gallery: Organized by topic, the images show objects and processes that give off radio waves.

James Webb Space Telescope Images: Here you can find the infrared images from the new large space telescope.

Our Infrared World Gallery: Images from a variety of infrared astronomy telescopes and missions. See also their "Cool Cosmos" site for the public.

Some Galleries on Specific Subjects

Astronaut Photography of Earth.

Chandra X-Ray Observatory Images.

NASA Human Spaceflight Gallery or Astronaut images.

Robert Gendler: One of the amateur astro-photographers who comes closest to being professional.

Sloan Digital Sky Survey Images.

Solar Dynamics Observatory Gallery: Sun images.

Spitzer Infrared Telescope Images.

Astronomy Apps for Smartphones and Tablets

The Best Astronomy Apps for IOS and Android

The Best Stargazing Apps for Looking at the Night Sky.

The 19 Best Astronomy Apps for Stargazing.

The 10 Best Astronomy Apps for Enjoying the Night Sky.

The Best Astronomy Apps for Smartphones, Tablets, and Computers.

Attribution

APPENDIX C: PHYSICAL AND ORBITAL DATA FOR THE PLANETS

Physical Data for the Major Planets

Table C1

Major Planet	Mean Diameter (km)	Mean Diameter (Earth = 1)	Mass (Earth = 1)	Mean Density (g/cm^3)	Rotation Period (d)	Inclination of Equator to Orbit (°)	Surface Gravity (Earth = 1[g])	Velocity of Escape (km/s)
Mercury	4879	0.38	0.055	5.43	58.	0.0	0.38	4.3
Venus	12,104	0.95	0.815	5.24	−243.	177	0.90	10.4
Earth	12,756	1.00	1.00	5.51	1.000	23.4	1.00	11.2
Mars	6779	0.53	0.11	3.93	1.026	25.2	0.38	5.0
Jupiter	140,000	10.9	318	1.33	0.414	3.1	2.53	60.
Saturn	117,000	9.13	95.2	0.69	0.440	26.7	1.07	36.
Uranus	50,700	3.98	14.5	1.27	−0.718	97.9	0.89	21.
Neptune	49,200	3.86	17.2	1.64	0.671	29.6	1.14	23.

Physical Data for Well-Studied Dwarf Planets

Table C2

Well-Studied Dwarf Planet	Diameter (km)	Diameter (Earth = 1)	Mass (Earth = 1)	Mean Density (g/cm^3)	Rotation Period (d)	Inclination of Equator to Orbit (°)	Surface Gravity (Earth = 1[g])	Velocity of Escape (km/s)
Ceres	950	0.07	0.0002	2.2	0.378	3	0.03	0.5
Pluto	2470	0.18	0.0024	1.9	−6.387	122	0.06	1.3
Haumea	1700	0.13	0.0007	3	0.163	—	—	0.8
Makemake	1400	0.11	0.0005	2	0.321	—	—	0.8
Eris	2326	0.18	0.0028	2.5	1.25[1]	—	—	1.4

Orbital Data for the Major Planets

Table C3

Major Planet	Semimajor Axis (AU)	Semimajor Axis (10^6 km)	Sidereal Period (y)	Sidereal Period (d)	Mean Orbital Speed (km/s)	Orbital Eccentricity	Inclination of Orbit to Ecliptic (°)
Mercury	0.39	58	0.24	88.0	47.9	0.206	7.0
Venus	0.72	108	0.6	224.7	35.0	0.007	3.4
Earth	1.00	149	1.00	365.2	29.8	0.017	0.0
Mars	1.52	228	1.88	687.0	24.1	0.093	1.9
Jupiter	5.20	778	11.86	—	13.1	0.048	1.3
Saturn	9.54	1427	29.46	—	9.6	0.056	2.5
Uranus	19.19	2871	84.01	—	6.8	0.046	0.8
Neptune	30.06	4497	164.82	—	5.4	0.010	1.8

Orbital Data for Well-Studied Dwarf Planets

Table C4

Well-Studied Dwarf Planet	Semimajor Axis (AU)	Semimajor Axis (10^6 km)	Sidereal Period (y)	Mean Orbital Speed (km/s)	Orbital Eccentricity	Inclination of Orbit to Ecliptic (°)
Ceres	2.77	414.0	4.6	18	0.08	11
Pluto	39.5	5915	248.6	4.7	0.25	17
Haumea	43.1	6452	283.3	4.5	0.19	28
Makemake	45.8	6850	309.9	4.4	0.16	29
Eris	68.0	10,120	560.9	3.4	0.44	44

---

1. This measurement is quite uncertain.

Attribution

APPENDIX D: SELECTED MOONS OF THE PLANETS

Nearly two hundred moons are now known in the solar system and more are being discovered on a regular basis. Of the major planets, only Mercury and Venus do not have moons. In addition to moons of the planets, there are many moons of asteroids. In this appendix, we list only the largest and most interesting objects that orbit each planet (including dwarf planets). The number given for each planet is discoveries through 2015. For further information see Our Solar System and List of Natural Satellites.

Table D1. Selected Moons of the Planets

Planet (moons)	Satellite Name	Discovery	Semimajor Axis (km × 1000)	Period (d)	Diameter
Earth (1)	Moon	—	384	27.32	3476
Mars (2)	Phobos	Hall (1877)	9.4	0.32	23
	Deimos	Hall (1877)	23.5	1.26	13
Jupiter (79)	Amalthea	Barnard (1892)	181	0.50	200
	Thebe	Voyager (1979)	222	0.67	90
	Io	Galileo (1610)	422	1.77	3630
	Europa	Galileo (1610)	671	3.55	3138
	Ganymede	Galileo (1610)	1070	7.16	5262
	Callisto	Galileo (1610)	1883	16.69	4800
	Himalia	Perrine (1904)	11,460	251	170
Saturn (82)	Pan	Voyager (1985)	133.6	0.58	20
	Atlas	Voyager (1980)	137.7	0.60	40
	Prometheus	Voyager (1980)	139.4	0.61	80
	Pandora	Voyager (1980)	141.7	0.63	100
	Janus	Dollfus (1966)	151.4	0.69	190
	Epimetheus	Fountain, Larson (1980)	151.4	0.69	120
	Mimas	Herschel (1789)	186	0.94	394
	Enceladus	Herschel (1789)	238	1.37	502
	Tethys	Cassini (1684)	295	1.89	1048
	Dione	Cassini (1684)	377	2.74	1120
	Rhea	Cassini (1672)	527	4.52	1530
	Titan	Huygens (1655)	1222	15.95	5150
	Hyperion	Bond, Lassell (1848)	1481	21.3	270
	Iapetus	Cassini (1671)	3561	79.3	1435
	Phoebe	Pickering (1898)	12,950	550 (R)[1]	220
Uranus (27)	Puck	Voyager (1985)	86.0	0.76	170
	Miranda	Kuiper (1948)	130	1.41	485
	Ariel	Lassell (1851)	191	2.52	1160

Planet (moons)	**Satellite Name**	**Discovery**	**Semimajor Axis (km × 1000)**	**Period (d)**	**Diameter**
	Umbriel	Lassell (1851)	266	4.14	1190
	Titania	Herschel (1787)	436	8.71	1610
	Oberon	Herschel (1787)	583	13.5	1550
Neptune (14)	Despina	Voyager (1989)	53	0.33	150
	Galatea	Voyager (1989)	62	0.40	150
	Larissa	Reitsema, et al (1981)	74	0.55	194
	Proteus	Voyager (1989)	118	1.12	420
	Triton	Lassell (1846)	355	5.88 (R)[2]	2720
	Nereid	Kuiper (1949)	5511	360	340
Pluto (5)	Charon	Christy (1978)	19.7	6.39	1200
	Styx	Showalter et al (2012)	42	20	20
	Nix	Weaver et al (2005)	48	24	46
	Kerberos	Showalter et al (2011)	58	24	28
	Hydra	Weaver et al (2005)	65	38	61
Eris (1)	Dysnomea	Brown et al (2005)	38	16	684
Makemake (1)	(MK2)	Parker et al (2016)	—	—	160
Haumea (2)	Hi’iaka	Brown et al (2005)	50	49	400
	Namaka	Brown et al (2005)	39	35	200

Attribution

---

1. R stands for retrograde rotation (backward from the direction that most objects in the solar system revolve and rotate).
2. R stands for retrograde rotation (backward from the direction that most objects in the solar system revolve and rotate).

APPENDIX E: FUTURE TOTAL ECLIPSES

Future Total Solar Eclipses

We also include eclipses that are *annular*—where the Moon is directly in front of the Sun, but doesn't fully cover it—leaving a ring of light around the dark Moon's edges)**.**

Table E1. Future Total Solar Eclipses.

Date	**Type of Eclipse**	**Location on Earth**[1]
April 20, 2023	Total[2]	Mostly in Indian and Pacific oceans, Indonesia
October 14, 2023	Annular	OR, NV, UT, NM, TX, C America, Colombia, Brazil
April 8, 2024	Total	N Mexico, U.S. (TX to ME), SE Canada and oceans on either side
October 2, 2024	Annular	S Chile, S Argentina, and oceans on either side
February 17, 2026	Annular	Only in Antarctica
August 12, 2026	Total	Greenland, Iceland, Spain
February 6, 2027	Annular	S Pacific, Argentina, Chile, Uruguay, S Atlantic
August 2, 2027	Total	Spain, Morocco, Egypt, Saudi Arabia, Yemen, Arabian Sea
January 26, 2028	Annular	Ecuador, Peru, Brazil, North Atlantic Ocean, Portugal, Spain
July 22, 2028	Total	Indian Ocean, Australia, New Zealand, South Pacific Ocean

---

[1] Remember that a total or annular eclipse is only visible on a narrow track. The same eclipse will be partial over a much larger area, but partial eclipses are not as spectacular as total ones.

[2] This is a so-called hybrid eclipse, which is total in some places and annular in others.

Future Total Lunar Eclipses

Table E2. Future Total Lunar Eclipses

Date	Location on Earth
November 8, 2022	Asia, Australia, Pacific Ocean, N America, S America
March 14, 2025	Pacific Ocean, N America, S America, Atlantic Ocean, W Europe, W Africa
September 7, 2025	Europe, Africa, Asia, Australia, Indian Ocean
March 3, 2026	E Asia, Australia, Pacific Ocean, N America, C America
June 26, 2029	E North America, S America, Atlantic Ocean, W Europe, W Africa
December 20, 2029	E North America, E South America, Atlantic Ocean, Europe, Africa, Asia

Additional Resources

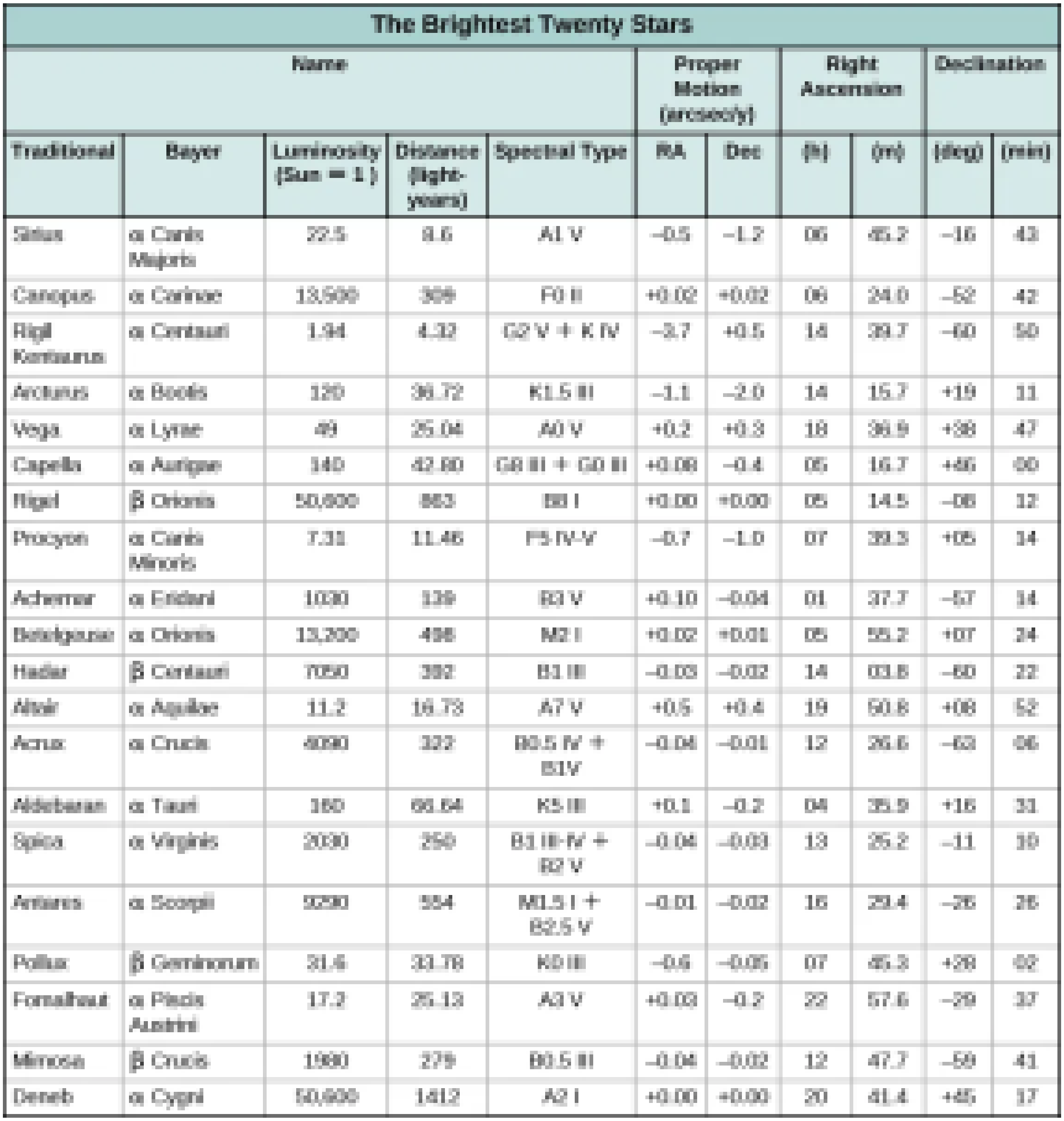
For more information and detailed maps about eclipses, see these resources.

- American Astronomical Society Eclipse Site
- NASA's Eclipse Site
- Mr. Eclipse site for beginners by Dr. Fred Espenak
- Eclipse Weather and Maps by Meteorologist Jay Anderson
- Eclipse Maps by Michael Zeiler
- Eclipse Information and Maps by Xavier Jubier

Attribution

APPENDIX F: THE BRIGHTEST TWENTY STARS

Note: These are the stars that *appear* the brightest visually, as seen from our vantage point on Earth. They are not necessarily the stars that are intrinsically the most luminous.



The Brightest Twenty Stars

Name					Proper Motion (arcsec/y)		Right Ascension		Declination	
Traditional	Bayer	Luminosity (Sun = 1)	Distance (light-years)	Spectral Type	RA	Dec	(h)	(m)	(deg)	(min)
Sirius	α Canis Majoris	22.5	8.6	A1 V	–0.5	–1.2	06	45.2	–16	43
Canopus	α Carinae	13,600	309	F0 II	+0.02	+0.02	06	24.0	–52	42
Rigil Kentaurus	α Centauri	1.94	4.32	G2 V + K IV	–3.7	+0.5	14	39.7	–60	50
Arcturus	α Bootis	120	36.72	K1.5 III	–1.1	–2.0	14	15.7	+19	11
Vega	α Lyrae	49	25.04	A0 V	+0.2	+0.3	18	36.9	+38	47
Capella	α Aurigae	140	42.80	G8 III + G0 III	+0.08	–0.4	05	16.7	+46	00
Rigel	β Orionis	50,600	863	B8 I	+0.00	+0.00	05	14.5	–08	12
Procyon	α Canis Minoris	7.31	11.46	F5 IV-V	–0.7	–1.0	07	39.3	+05	14
Achernar	α Eridani	1030	139	B3 V	+0.10	–0.04	01	37.7	–57	14
Betelgeuse	α Orionis	13,200	498	M2 I	+0.02	+0.01	05	55.2	+07	24
Hadar	β Centauri	7050	392	B1 III	–0.03	–0.02	14	03.8	–60	22
Altair	α Aquilae	11.2	16.73	A7 V	+0.5	+0.4	19	50.8	+08	52
Acrux	α Crucis	4090	322	B0.5 IV + B1V	–0.04	–0.01	12	26.6	–63	06
Aldebaran	α Tauri	160	66.64	K5 III	+0.1	–0.2	04	35.9	+16	31
Spica	α Virginis	2030	250	B1 III-IV + B2 V	–0.04	–0.03	13	25.2	–11	10
Antares	α Scorpii	9290	554	M1.5 I + B2.5 V	–0.01	–0.02	16	29.4	–26	26
Pollux	β Geminorum	31.6	33.78	K0 III	–0.6	–0.05	07	45.3	+28	02
Fomalhaut	α Piscis Austrini	17.2	25.13	A3 V	+0.03	–0.2	22	57.6	–29	37
Mimosa	β Crucis	1990	279	B0.5 III	–0.04	–0.02	12	47.7	–59	41
Deneb	α Cygni	50,600	1412	A2 I	+0.00	+0.00	20	41.4	+45	17

Figure F1. The brightest stars typically have names from antiquity. Next to each star's ancient name, we have added a column with its name in the system originated by Bayer. The distances of the more remote stars are estimated from their spectral types and apparent brightnesses and are only approximate. The luminosities for those stars are approximate to the same degree. Right ascension and declination is given for Epoch 2000.0.

Attribution

APPENDIX G: THE CHEMICAL ELEMENTS

Table G1.

Element	Symbol	Atomic Number	Atomic Weight[1]	Percentage of Naturally Occurring Elements in the Universe
Hydrogen	H	1	1.008	75
Helium	He	2	4.003	23
Lithium	Li	3	6.94	6×10^{-7}
Beryllium	Be	4	9.012	1×10^{-7}
Boron	B	5	10.821	1×10^{-7}
Carbon	C	6	12.011	0.5
Nitrogen	N	7	14.007	0.1
Oxygen	O	8	15.999	1
Fluorine	F	9	18.998	4×10^{-5}
Neon	Ne	10	20.180	0.13
Sodium	Na	11	22.990	0.002
Magnesium	Mg	12	24.305	0.06
Aluminum	Al	13	26.982	0.005
Silicon	Si	14	28.085	0.07
Phosphorus	P	15	30.974	7×10^{-4}
Sulfur	S	16	32.06	0.05
Chlorine	Cl	17	35.45	1×10^{-4}
Argon	Ar	18	39.948	0.02
Potassium	K	19	39.098	3×10^{-4}
Calcium	Ca	20	40.078	0.007
Scandium	Sc	21	44.956	3×10^{-6}
Titanium	Ti	22	47.867	3×10^{-4}
Vanadium	V	23	50.942	3×10^{-4}
Chromium	Cr	24	51.996	0.0015
Manganese	Mn	25	54.938	8×10^{-4}
Iron	Fe	26	55.845	0.11
Cobalt	Co	27	58.933	3×10^{-4}

---

1. Where mean atomic weights have not been well determined, the atomic mass numbers of the most stable isotopes are given in parentheses.

Element	**Symbol**	**Atomic Number**	**Atomic Weight**	**Percentage of Naturally Occurring Elements in the Universe**
Nickel	Ni	28	58.693	0.006
Copper	Cu	29	63.546	6×10^{-6}
Zinc	Zn	30	65.38	3×10^{-5}
Gallium	Ga	31	69.723	1×10^{-6}
Germanium	Ge	32	72.630	2×10^{-5}
Arsenic	As	33	74.922	8×10^{-7}
Selenium	Se	34	78.971	3×10^{-6}
Bromine	Br	35	79.904	7×10^{-7}
Krypton	Kr	36	83.798	4×10^{-6}
Rubidium	Rb	37	85.468	1×10^{-6}
’Strontium	Sr	38	87.62	4×10^{-6}
Yttrium	Y	39	88.906	7×10^{-7}
Zirconium	Zr	40	91.224	5×10^{-6}
Niobium	Nb	41	92.906	2×10^{-7}
Molybdenum	Mo	42	95.95	5×10^{-7}
Technetium	Tc	43	(98)	—
Ruthenium	Ru	44	101.07	4×10^{-7}
Rhodium	Rh	45	102.906	6×10^{-8}
Palladium	Pd	46	106.42	2×10^{-7}
Silver	Ag	47	107.868	6×10^{-8}
Cadmium	Cd	48	112.414	2×10^{-7}
Indium	In	49	114.818	3×10^{-8}
Tin	Sn	50	118.710	4×10^{-7}
Antimony	Sb	51	121.760	4×10^{-8}
Tellurium	Te	52	127.60	9×10^{-7}
Iodine	I	53	126.904	1×10^{-7}
Xenon	Xe	54	131.293	1×10^{-6}

Element	**Symbol**	**Atomic Number**	**Atomic Weight**	**Percentage of Naturally Occurring Elements in the Universe**
Cesium	Cs	55	132.905	8×10^{-8}
Barium	Ba	56	137.327	1×10^{-6}
Lanthanum	La	57	138.905	2×10^{-7}
Cerium	Ce	58	140.116	1×10^{-6}
Praseodymium	Pr	59	140.907	2×10^{-7}
Neodymium	Nd	60	144.242	1×10^{-6}
Promethium	Pm	61	(145)	—
Samarium	Sm	62	150.36	5×10^{-7}
Europium	Eu	63	151.964	5×10^{-8}
Gadolinium	Gd	64	157.25	2×10^{-7}
Terbium	Tb	65	158.925	5×10^{-8}
Dysprosium	Dy	66	162.500	2×10^{-7}
Holmium	Ho	67	164.930	5×10^{-8}
Erbium	Er	68	167.259	2×10^{-7}
Thulium	Tm	69	168.934	1×10^{-8}
Ytterbium	Yb	70	173.054	2×10^{-7}
Lutetium	Lu	71	174.967	1×10^{-8}
Hafnium	Hf	72	178.49	7×10^{-8}
Tantalum	Ta	73	180.948	8×10^{-9}
Tungsten	W	74	183.84	5×10^{-8}
Rhenium	Re	75	186.207	2×10^{-8}
Osmium	Os	76	190.23	3×10^{-7}
Iridium	Ir	77	192.217	2×10^{-7}
Platinum	Pt	78	195.084	5×10^{-7}
Gold	Au	79	196.967	6×10^{-8}
Mercury	Hg	80	200.592	1×10^{-7}
Thallium	TI	81	204.38	5×10^{-8}

Element	**Symbol**	**Atomic Number**	**Atomic Weight**	**Percentage of Naturally Occurring Elements in the Universe**
Lead	Pb	82	207.2	1×10^{-6}
Bismuth	Bi	83	208.980	7×10^{-8}
Polonium	Po	84	(209)	—
Astatine	At	85	(210)	—
Radon	Rn	86	(222)	—
Francium	Fr	87	(223)	—
Radium	Ra	88	(226)	—
Actinium	Ac	89	(227)	—
Thorium	Th	90	232.038	4×10^{-8}
Protactinium	Pa	91	231.036	—
Uranium	U	92	238.029	2×10^{-8}
Neptunium	Np	93	(237)	—
Plutonium	Pu	94	(244)	—
Americium	Am	95	(243)	—
Curium	Cm	96	(247)	—
Berkelium	Bk	97	(247)	—
Californium	Cf	98	(251)	—
Einsteinium	Es	99	(252)	—
Fermium	Fm	100	(257)	—
Mendelevium	Md	101	(258)	—
Nobelium	No	102	(259)	—
Lawrencium	Lr	103	(262)	—
Rutherfordium	Rf	104	(267)	—
Dubnium	Db	105	(268)	—
Seaborgium	Sg	106	(271)	—
Bohrium	Bh	107	(272)	—
Hassium	Hs	108	(270)	—
Meitnerium	Mt	109	(276)	—

Element	Symbol	Atomic Number	Atomic Weight	Percentage of Naturally Occurring Elements in the Universe
Darmstadtium	Ds	110	(281)	—
Roentgenium	Rg	111	(280)	—
Copernicium	Cn	112	(285)	—
Nihonium	Nh	113	(284)	—
Flerovium	Fl	114	(289)	—
Moskovium	Mc	115	(288)	—
Livermorium	Lv	116	(293)	—
Tennessine	Ts	117	(294)	—
Oganesson	Og	118	(294)	—

Attribution

"K | The Chemical Elements" from Astronomy 2e by Andrew Fraknoi, David Morrison, Sidney C. Wolff, © OpenStax – Rice University is licensed under a Creative Commons Attribution 4.0 International License, except where otherwise noted.

APPENDIX H: STAR CHART AND SKY EVENT RESOURCES

Star Charts

To obtain graphic charts of the sky over your head tonight, there are a number of free online resources.

- One of the easiest is at the Skymaps website: Here you can print out a free PDF version of the northern sky (for roughly latitude 40°, which is reasonable for much of the United States). Maps of the sky from the equator and the southern hemisphere can also be printed.
- Sky charts and other summaries of astronomical information can also be found at Heavens Above.
- A free, open-source computer application that shows the sky at any time from any place is called Stellarium.
- Appendix B provides a section of information for finding astronomy apps for cell phones and tablets. Many of these also provide star charts. If you have a smartphone, you can find a variety of inexpensive apps that allow you to simply hold your phone upward to see what is in the sky behind your phone.
- A *planisphere* is a sky chart that turns inside a round frame and can show you the night sky at your latitude on any date and time of the year. You can buy them at science supply and telescope stores or online. Or you can construct your own from templates at these two websites:
    - Dennis Schatz's AstroAdventures Star Finder
    - Uncle Al's Star Wheel

Calendars of Night Sky Events

The following resources offer calendars of night sky events.

- Sea and Sky (click on the fourth menu button)
- *Sky & Telescope's* This Week's Sky at a Glance
- *Astronomy* Magazine's The Sky This Week
- Dr. Fred Espenak's Year by Year Tables of Events
- Night Sky Network Sky Planner

Attribution

VERSIONING HISTORY

This page provides a record of edits and changes made to this book since its initial publication. Whenever edits or updates are made in the text, we provide a record and description of those changes here. If the change is minor, the version number increases by 0.1. If the edits involve a number of changes, the version number increases to the next full number.

The files posted alongside this book always reflect the most recent version.

Version	Date	Change	Affected Web Page
1.0	August 2023	First Publication	N/A