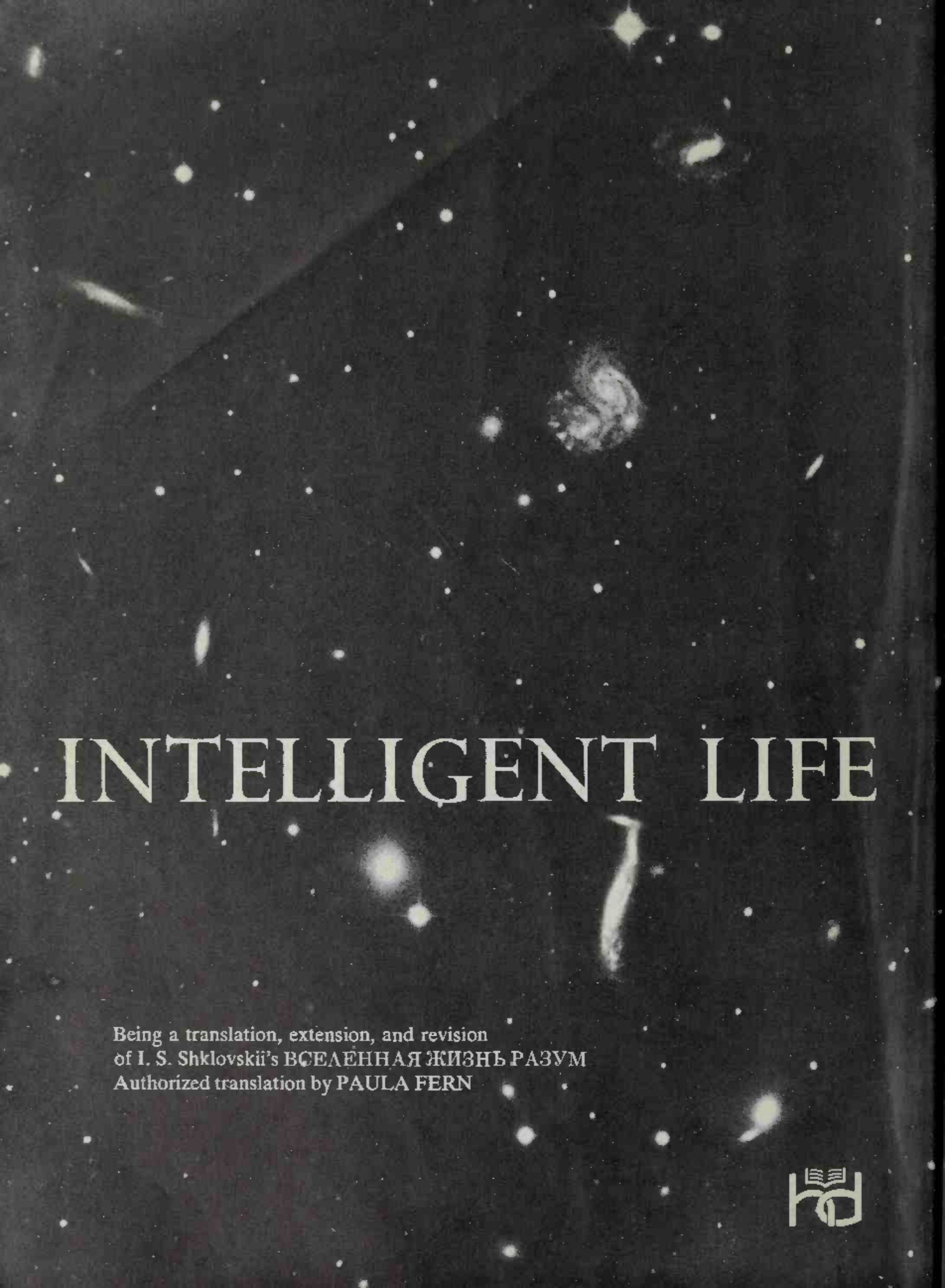
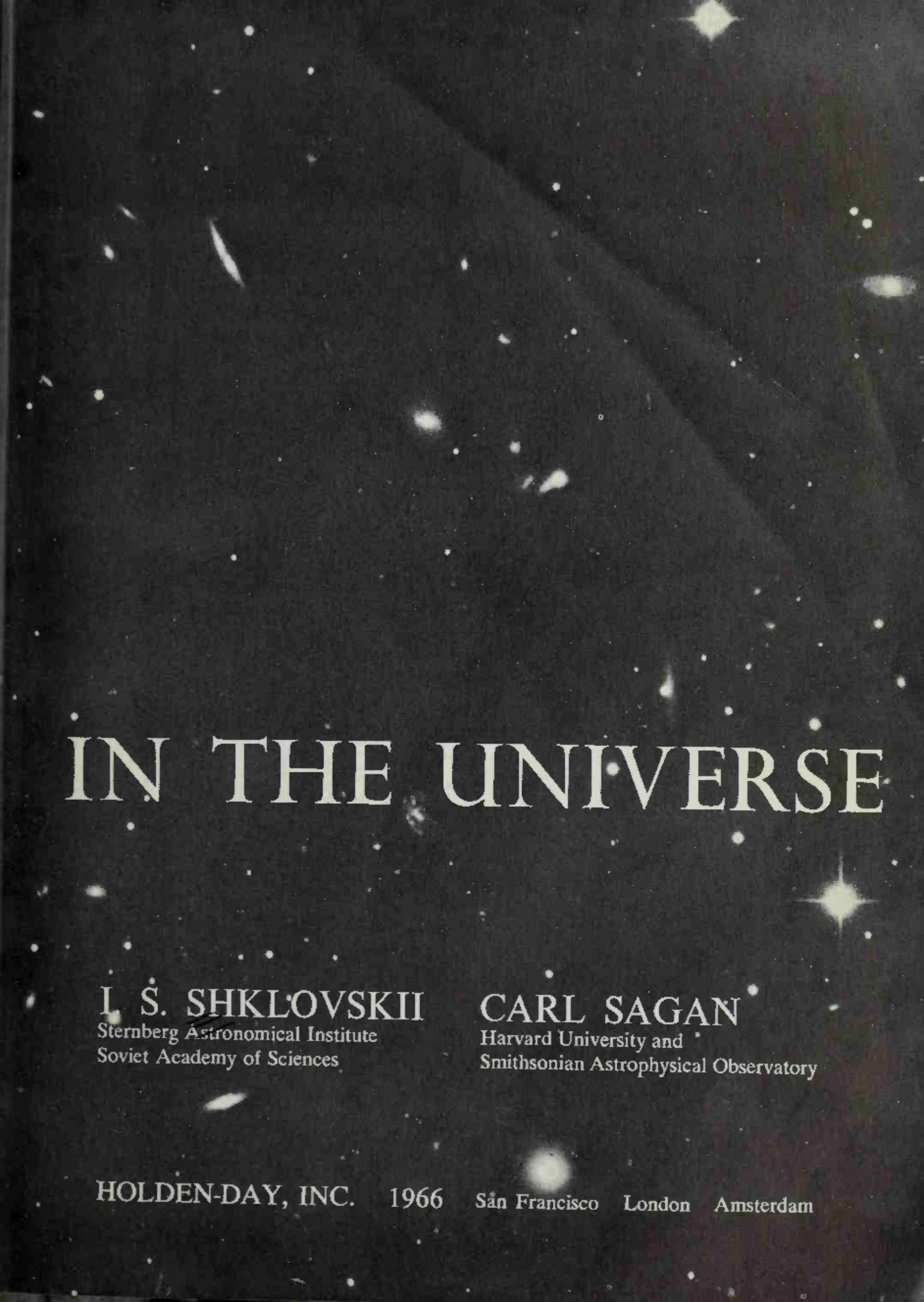
Intelligent Life in the Universe


INTELLIGENT LIFE
Being a translation, extension, and revision
of I. S. Shklovskii's ВСЕЛЁННАЯ ЖИЗНЬ РАЗУМ
Authorized translation by PAULA FERN



IN THE UNIVERSE

I. S. SHKLOVSKII
Sternberg Astronomical Institute
Soviet Academy of Sciences

CARL SAGAN
Harvard University and
Smithsonian Astrophysical Observatory

HOLDEN-DAY, INC. 1966 San Francisco London Amsterdam

Library of Congress Catalog Card Number: 64–18404
Printed in the United States of America

To the memory of John Burdon Sanderson Haldane, F.R.S., member of the National Academies of Sciences of the United States and of the Soviet Union, member of The Order of the Dolphins, and a local example of what this book is about.

Preface

"The prey runs to the hunter," wrote Iosef Shmuelovich Shklovskii in a letter to me in 1962. Knowing his wide-ranging interests, I had sent him a preprint of a paper of mine called "Direct Contact Among Galactic Civilizations by Relativistic Interstellar Spaceflight," a speculative exercise on a topic which I thought might interest him. Shklovskii wrote to tell me that he was completing a semipopular book, *Vselennaia, Zhizn, Razum* (in English, *Universe, Life, Mind*). It was being written for the fifth anniversary of the launching of the first Soviet artificial satellite on August 4, 1957. Shklovskii had been about to write a chapter on the possibility of interstellar spaceflight when my preprint arrived, just in time to be partially incorporated into the text. *Vselennaia, Zhizn, Razum* was published in Moscow in early 1963. Parts of it were also serialized in *Komsomolskaya Pravda*, and extracts published in the Soviet scientific journal *Priroda*. It has received enthusiastic acclaim in the Soviet Union and elsewhere, and is being translated into a variety of languages, including Chinese.

When I received a copy of the book, I was struck by its broad scope and novel insights. I wrote to Shklovskii, asking him if we might translate it into English. Shklovskii readily consented, and invited me to add additional material as I saw fit. As the translation proceeded, in the capable hands of Paula Fern, I found myself unable to resist the temptation to annotate the text, clarify concepts for the scientific layman, comment at length, and introduce new material. The delay in the publication of the English edition is attributable entirely to this cause. I have added about as much material as there was in the work initially, and the figures and captions are solely my responsibility. I hope that the book is, as a result, of somewhat wider appeal. Shklovskii also has made a large number of changes and additions which have been incorporated into the English edition.

The result is a peculiar kind of cooperative endeavor. I have sent much of the entirely new material to Shklovskii for his comment, and he has sent much new material to me for inclusion. Since he does not travel out of the Soviet Union and I have never traveled to the Soviet Union, we have been unable to discuss the present edition in person. "The probability of our meeting is unlikely to be smaller than the probability of a visit to the Earth by an extraterrestrial cosmonaut," he once wrote, in a puckish mood. Consequently, there are places in the present work where Shklovskii and I alternate sentences, or even occasionally insert clauses into each other's sentences. Shklovskii and I agree on almost all the substantive issues of this book, but to avoid the possibility of attribution to Shklovskii of a view which he does not hold, I have adopted the following strategem: Sentences or paragraphs which appeared either in the Russian edition of this book, or in additions provided by Shklovskii, are presented in ordinary type. Annotations, additions, and discussions of my devising are surrounded by the symbols ▽ and △, the first preceding and the second following my contribution. In those cases where Shklovskii uses the word "we," as in "we believe" or "we feel," this generally represents a sentiment shared by both of us.

As the reader might expect for a book written by two authors, one in the Soviet Union and one in the United States, there are occasional ideological differences. I have not tried to avoid these problems, but I also have not tried, in what is primarily a scientific work, to rebut each ideological assertion. When Shklovskii expresses his belief that lasting world peace is impossible while capitalism survives, or implies that lasers are being actively developed in the United States for their possible military applications alone, I have let the content of these statements stand, despite their political intent. I have occasionally interjected some remarks on related subjects, with which, perhaps, Shklovskii would disagree. I do not think the reader will be distressed by the occasional appearance of a dialogue.

The possibility of extraterrestrial life has caused some ideological embarrassment in the Soviet Union. There used to be, at Alma Ata, in the Kazakh Soviet Socialist Republic, an Institute of Astrobotany, some of whose members argued that the existence of extraterrestrial life was required by dialectical materialism, and implied strongly that the absence of life on Mars, or even on Jupiter, would be a clear disproof of the philosophical basis of Communism. This dangerous situation prompted an article in the September–October 1958 issue of the Soviet astronomical journal *Astronomicheskii Zhurnal,* called "Concerning the 'Philosophical Foundation' of One Question," by I. G. Perel', in which Perel' points out that both the materialist and the idealist philosophical schools seem to strongly support the likelihood of extraterrestrial life. He argues that dialectical materialism is a method, not a body of knowledge, much as Shklovskii does on p. 136 of this book; and in particular, that even if Mars or Jupiter is lifeless, dialectical materialism is not disproved. This debate has been echoed by other discussions in the United States, which, while on a different ideological basis, turn out to have very similar content.

The present work has ten more chapters than the original Russian edition. This is due almost entirely to new material. The over-all organization remains as in the Russian edition: a discussion first of the astronomical background, then of the nature of life and its possible occurrence in our solar system, and finally, a treatment of the possibility that advanced communicative technical civilizations exist on planets of other stars. A more detailed overview of the book appears in the introductory chapter by Shklovskii. I have added an introductory chapter on the psychological perils of the study of extraterrestrial life.

It is impossible for me to thank all those who have helped to shape my views on the topics of this book. I would, however, like to acknowledge my debt to the following people for specific discussion of relevant material or for reading and commenting on various parts of the present work in manuscript form: Dr. Elso Barghoorn, Dr. Geoffrey Burbidge, Dr. Frank Drake, Dr. Freeman Dyson, Dr. Owen Gingerich, Dr. J. B. S. Haldane, Dr. William Irvine, Dr. Luigi Jacchia, Dr. G. P. Kuiper, Dr. David Layzer, Dr. A. E. Lilley, Dr. Phillip Morrison, Dr. H. J. Muller, Dr. James B. Pollack, Dr. Lynn Sagan, Dr. Evry Schatzmann, Dr. Ellie Shneour, Dr. Charles H. Townes, and Dr. Andrew T. Young. I am also grateful to Dr. Leo Goldberg for editorial advice. None of the foregoing are, of course, responsible for any errors of fact or interpretation which may have crept into the manuscript.

In addition to credits referenced through the text the following acknowledgements are made: the jacket composite photograph is by courtesy of NASA and Mt. Wilson and Palomar Observatories. The frontispiece cartoon, from POGO is © by Walt Kelly and reprinted by kind permission of the Hall Syndicate, Inc. Cartoons, page 24, by Charles Schulz, © by United Features Syndicate and also published in *It's A Dog's Life, Charlie Brown* by Holt, Rinehart and Winston. Quotations from Loren Eiseley's *The Immense Journey* (1946) pages 25, 182, 356 and 432 are © by Random House, Inc., and reproduced through their permission. Mt. Wilson and Palomar Observatories through whose courtesy the photographs on page 7 and in many other places in the book are reproduced, are associated with the Carnegie Institution of Washington and the California Institute of Technology. Yerkes Observatory through whose courtesy the illustrations on page 41 and elsewhere are reproduced, is a unit of the University of Chicago. The diagram on page 80 is reproduced through the courtesy of Professor Chushiro Hayashi. The diagrams from the *Astrophysical Journal* on pages 81 and 84 are © in the years indicated by the University of Chicago. The diagram on page 86 also appeared in the volume "Stellar Populations," *Specola Vaticana Richerche Astronomiche,* volume 5, page 227. The figures from the *Astrophysical Journal* on pages 118 and 119 are © 1963 by the University of Chicago. The diagram on page 150 first appeared in *Sky and Telescope 26:* 3. The diagram on page 163 first appeared in *Physics Today 1:* 12–28, 1948. The illustration on page 180 is reproduced through the courtesy of Mr. Charles Addams and the *New Yorker* Magazine, Inc. The quotation on page 184 appears on page 1–9 of *The Feynman Lectures on Physics,* volume 1, by R. P. Feynman, R. Leighton and M. Sands, Addison-Wesley, 1963, through the kind permission of Professor Feynman and the publisher. The photographs on pages 338 and 339 appeared in *Science 138:* 1392–99 (1962), and are reproduced by permission. The upper illustration on page 354 is reproduced through the courtesy of Mr. Herbert Block; the lower through the courtesy of Mr. Charles Addams and the *New Yorker* Magazine, Inc. Parts of Chapters 29, 32, and 33, as well as Figure 32–1, are taken from an article by Carl Sagan in *Planetary and Space Science,* volume 11, p. 45 (1963) and are reproduced through the courtesy of Professor D. R. Bates, the editor. The illustrations on pages 458, 461 and 462 are reproduced through the courtesy of Macmillan and Co., Ltd., London. The quotations on pages 459 and 460 are reproduced from H. Frankfort *et al., Before Philosophy, The Intellectual Adventure of Ancient Man,* © 1946 by the University of Chicago Press.

Dr. Charles Federer and Dr. Joseph Ashbrook of *Sky and Telescope* have generously assisted in tracking down illustrative material.

Collecting the epigraphs for the chapters of this book has been a source of considerable pleasure for me. Some of these epigraphs have been suggested by others. The quotations from Pasternak, on page 26, and by Smelianov on page 272 appeared in the Russian edition. The quote from Herschel on page 110 was suggested by Dr. Owen Gingerich, as was the epigraph of Chapter 11. Dr. Sidney Coleman recommended the epigraph of Chapter 28. Dr. Robert B. Leighton called my attention to the Walt Kelly strip used as one of the frontispieces. The dedication of this book, while of my devising, was approved by Professor Shklovskii.

I would like, finally, to express my thanks to Mrs. Ruth Ellen Galper for intelligent commentary on and typing of what proved to be an unexpectedly long manuscript; and to Miss Jean Swift for reading and providing useful editorial comments on the entire work, and to the other members of the publisher's staff. Other assistance with the manuscript was generously provided by Mrs. Terry Brown, Mrs. Sondra Cohen, Mrs. Elinore Green, and Miss Kathryn Bloom.

Carl Sagan
Cambridge, Mass.
May 15, 1966

A note on mathematical and physical units

▽ In this book, we are concerned with quantities ranging from the immensely large to the minutely small. If we were to write out these sizes in the conventional way, using any unit of length we wish—miles or inches, for example—the pages of the book would be filled with large numbers of zeros, either before or after the decimal place. For any one of these numbers, such as 33 000 000 000 000 000, there is always the tedious problem of counting the number of zeros.

▽ To avoid these difficulties, mathematicians devised, some centuries ago, a very simple exponential notation for large and small numbers. The expression a^n means that we are to multiply the number a by itself n times. For example, $2^2 = 4$; $10^3 = 1000$; etc. a is called the base, and n, the exponent. A negative exponent denotes a reciprocal. Thus, $a^{-n} = 1/a^n$. In this way, $\frac{1}{8}$ can be written as 2^{-3}; $\frac{1}{10\,000}$, 10^{-4}; etc. A fractional exponent, such as $a^{1/2}$, indicates that we must multiply a by itself less than one time; therefore, we should obtain a number which is less than a. For example, an exponent of $\frac{1}{2}$ denotes the square root; an exponent of $\frac{1}{10}$ denotes the tenth root, etc. Therefore, $8^{1/3} = 2$; $10000^{1/4} = 10$; etc. Since $2^3 = 8$, and $10^4 = 10000$, these results can also be written as $(2^3)^{1/3} = 2^1 = 2$, and $(10^4)^{1/4} = 10^1 = 10$.

▽ With this notation, multiplication and division become very easy, provided we compare quantities with the same base. In multiplication, we add exponents; in division, we subtract exponents. $100 \times 1000 = 100000$; or, in our notation, $10^2 \times 10^3 = 10^5$. Such addition of exponents becomes very useful for extremely large numbers. For example, the product of 2 million and 4.5 billion may be written as

$$(2 \times 10^6) \times (4.5 \times 10^9) = 9 \times 10^{15}.$$

The expression 9×10^{15} is a much neater way of describing the number, which otherwise we might have to call "nine million trillion."

▽ Similarly, 7 billion divided by 2000 can be expressed as

$$\frac{7 \times 10^9}{2 \times 10^3} = \frac{7}{2} \times 10^6 = 3.5 \times 10^6.$$

Since division is equivalent to multiplication by reciprocals, we can write this as

$$(7 \times 10^9) \times (2^{-1}) \times (10^{-3}) = 3.5 \times 10^6.$$

▽ Our arithmetic is based upon the base 10, probably because we have ten fingers on our hands, and mathematics developed among people who counted on their fingers. The metric system of units, introduced in France at the time of the French Revolution, is also constructed on the base 10. For example, there are 10 millimeters to a centimeter; 1000 grams to a kilogram; and so forth. The United States, United Kingdom, and a few other countries are still stuck in the morass of what is called the English system of units, in which, for example, 12 inches make a foot; 16 ounces make a pound; and 5280 feet make a mile—as if the English system had developed among some species of animal with a strange and variable number of appendages. In this book we use primarily the metric system. Distances will be measured in centimeters, meters, kilometers, and so forth; masses will be measured in grams and their multiples; and time will be measured in seconds. A conversion from metric to English units will be made where the metric units are first introduced in the body of the text. The reader is encouraged to adopt the metric system of units in his thinking about the contents of this book.

▽ The exponential system of notation can also be used for the physical units. For example, a light year is defined as the distance which light travels in one year. In metric units, light travels 3×10^{10} cm in each second (or per second). This is the velocity of light. We may also write this as 3×10^{10} cm/sec, or 3×10^{10} cm sec^{-1}, where, as with numbers, the exponent -1 indicates a reciprocal. To determine the number of centimeters that light travels in a year from the number of centimeters that it travels in a second, we must find the number of seconds in a year. There are 60 seconds in a minute, 60 minutes in an hour, 24 hours in a day, about 365 days in a year. Thus, a light year may be written as

$$(3 \times 10^{10}\ \text{cm sec}^{-1}) \times (6 \times 10\ \text{sec min}^{-1}) \times (6 \times 10\ \text{min hr}^{-1}) \\ \times (2.4 \times 10\ \text{hr day}^{-1}) \times (3.65 \times 10^{2}\ \text{days yr}^{-1}).$$

If we perform the multiplication and cancel units—(sec^{-1}) $\times$ (sec) $= 1$, etc.—we find that a light year is about 10^{18} cm, a very large distance.

▽ As another problem, consider the density of the Sun. The volume of a sphere is $\frac{4}{3}\pi R^3$, where R is its radius. The density is simply the mass, M, divided by the volume. Therefore, the density of the Sun is $M/(\frac{4}{3}\pi R^3)$, or $3(4\pi)^{-1}MR^{-3}$. The mass of the Sun is 2×10^{33} gm; its radius is 7×10^{10} cm. If you substitute these values for M and R in the above equation and find that the density of the Sun is about 1.4 gm cm^{-3}, you will have no difficulty with any of the mathematics in this book. (Since the density of water is about 1 gm cm^{-3}, we have derived the somewhat surprising result that the Sun is only slightly more dense than water.) △

Table of contents

III: INTELLIGENT LIFE IN THE UNIVERSE

A star cloud in the region of the galactic center. There are approximately a million stars in this photograph. According to the estimates of Chapter 29, a planet of one of these stars holds a technical civilization vastly in advance of our own. (Courtesy of Mount Wilson and Palomar Observatories.)

© by Walt Kelly

There is one
race of men, one race of gods; both have breath
of life from a single mother. But sundered power
holds us divided, so that the one is nothing, while for the
other the brazen sky is established
their sure citadel forever. Yet we have some likeness in great
intelligence, or strength, to the immortals,
though we know not what the day will bring, what course
after nightfall
destiny has written that we must run to the end.

Pindar, *Sixth Nemean Ode*

1

Perspectives

Numberless are the world's wonders, but none
More wonderful than man; the stormgrey sea
Yields to his prows, the huge crests bear him high . . .
The lion on the hill, the wild horse windy-maned,
Resign to him; and his blunt yoke has broken
The sultry shoulders of the mountain bull.
Words also, and thought as rapid as air,
He fashions to his good use . . .

Sophocles, *Antigone*

My surprise reached a climax, however, when I found that [Sherlock Holmes] was ignorant of the Copernican Theory, and of the composition of the Solar System. That any civilized human being in the nineteenth century should not be aware that the Earth travelled round the sun appeared to me to be such an extraordinary fact that I could hardly realize it.

"You appear to be astonished," he said, smiling at my expression of surprise. "Now that I do know it, I shall do my best to forget it . . ."

"But the Solar System!" I protested.

"What the deuce is it to me?" he interrupted impatiently: "you say that we go round the sun. If we went round the moon it would not make a pennyworth of difference to me . . ."

Sir Arthur Conan Doyle, *A Study in Scarlet*

Since the dawn of history, man has speculated about the possibility that intelligent life may exist on other worlds beyond the Earth. This idea probably originated from the often unsuccessful attempts of primitive religions to give meaning to those aspects of the environment which had no simple explanations. In the ancient Vedda culture of Ceylon, the belief in the migration of the soul after death was linked with the concept of a plurality of habitable worlds. The dead souls were believed to migrate to the Sun, the Moon, and the stars before attaining the state of Nirvana.

As astronomy developed, the concept of the existence of life on other worlds began to acquire some scientific bases. Most of the early Greek philosophers, both the materialists and the idealists, thought that our Earth was not the sole dwelling place of intelligent life. Considering the limitations of science at that time, these early philosophers displayed great originality and ingenuity. Thales of Miletus, the founder of the Ionian school of philosophy, taught that the stars and the Earth were made of the same material. Anaximander asserted that worlds are created and destroyed. Anaxagoras, one of the first proponents of the heliocentric theory, believed the moon to be inhabited. He also maintained that invisible "seeds of life," from which all living things originated, were dispersed throughout the universe.

In later eras, similar concepts of "panspermia" (ubiquitous life) were propounded by various scientists and philosophers. This idea was incorporated into Christianity soon after its inception.

The Epicurean school of materialist philosophy taught that many habitable worlds, similar to our Earth, existed in space. The Epicurean, Metrodoros, maintained: "To consider the Earth the only populated world in infinite space is as absurd as to assert that in an entire field sown with millet only one grain will grow." It is of interest that the proponents of this doctrine considered that not only the planets, but also other heavenly bodies in the vast reaches of space, were inhabited.

The Roman philosopher, Titus Lucretius Carus, was an ardent exponent of the concept of the plurality of worlds. In his famous poem, *On the Nature of Things,* he wrote: "Nature is not unique to the visible world; we must have faith that in other regions of space there exist other earths, inhabited by other peoples and other animals." Curiously enough, Lucretius did not understand the true nature of the stars, but conceived of them as luminous terrestrial vapors; therefore, his inhabited worlds were located on the periphery of the visible universe.

For fifteen hundred years after the birth of Jesus of Nazareth, Christian cosmology, influenced by the theories of Ptolemy, taught that the Earth was the center of the universe. The concept of life on other worlds seemed to be incompatible with this philosophy. The extrication of cosmology from the Ptolemaic system began when the gifted Polish astronomer, Nicolaus Copernicus, placed man in his proper position in the solar system, downgrading the status of Earth to that of one planet among the many revolving about the Sun.

▽ Copernicus' achievement lay in his precise explanation, with a modest investment of hypothesis, of the motions of the planets. The Ptolemaic hypothesis that the Sun, Moon, and planets, embedded in crystal spheres, circled the Earth, encountered more and more difficulty with observations of the changing planetary, lunar, and solar motions as the centuries passed. A set of special motions, called epicycles, was a characteristic feature of Ptolemaic hypothesis. At the time of its

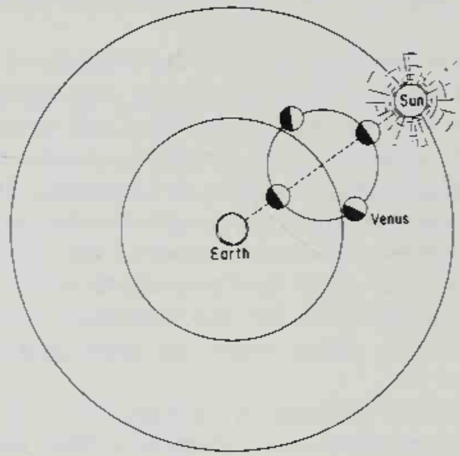


FIGURE 1–1. *Schematic illustration of the relative positions of the Earth, the Sun, and Venus in the Ptolemaic cosmology. To explain the motions of Venus, it was necessary to imagine the planet revolving about a point on a straight line between the Earth and the Sun. Note that the center of the bright side of Venus (towards the Sun) can never be seen by an observer on the Earth.*

invention, the heliocentric hypothesis of Copernicus merely explained the motions in a simpler way; at a later time, with Galileo's discovery that the planet Venus exhibits phases like those of the moon, the Copernican hypothesis was, in its gross features, demonstrated correct, and the Ptolemaic views overthrown. If any modern refutation of the Ptolemaic cosmology is needed, it is provided by space vehicles. The flights of Luna III, Mariner II, and Mariner IV were not accompanied by the tinkle of broken crystal spheres.

▽ The distinction between the Ptolemaic and the Copernican cosmologies is an interesting example of model-building, or hypothesis construction in science. Both the Ptolemaic and the Copernican views explained the motions of the planets. The heliocentric view of Copernicus was a simpler hypothesis. This in itself is not a demonstration of its validity. Nature may, after all, be complex. But if each view explains the planetary motions equally well, we certainly cannot be criticized if we think in terms of the simpler model. The Ptolemaic and Copernican pictures differed in another respect, however. According to Ptolemy, the Sun circled about the Earth, and inside the sphere of the Sun lay the sphere of Venus and Mercury. With such a geometry, it would be impossible for us ever to see the entire bright side of Venus [see Fig. 1–1]. According to Copernicus, however, both Venus and the Earth circled the Sun. Since Venus was sometimes beyond the Earth and the Sun, it would be possible for us to see its bright side [see Fig. 1–2]. Thus, when Galileo turned his telescope to Venus, and saw that its disk underwent phases from a "full Venus," corresponding to our full moon, to a "new Venus" (the dark side of Venus), corresponding to our new moon, it was clear that the Copernican hypothesis was vindicated. It does not follow that the Copernican view is completely valid in every respect; it is merely a model which conforms, with our desired degree of precision, to all the observations.

▽ In later years, Johannes Kepler showed that the paths of the planets about the sun were not circles, but ellipses. A prediction of the observed planetary motions on the basis of a law of gravitational interaction tested for the Moon was the

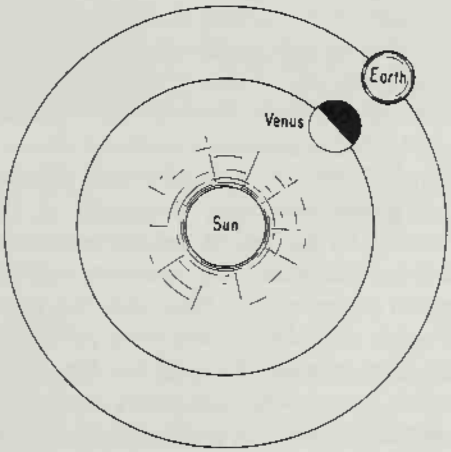


FIGURE 1–2. *Schematic illustration of the relative positions of the Earth, the Sun, and Venus in the Copernican cosmology. Note that the center of the bright side of Venus (towards the Sun) can sometimes be seen by an observer on the Earth.*

crowning achievement of Sir Isaac Newton, completing the goal of explaining in great detail the motions of the Sun, Moon, and planets from simple and largely testable starting points. △

The first telescopic observations by Galileo opened a new era in astronomy and dealt a mighty blow to the ideas of many of his contemporaries. It became evident that the planets were similar to the Earth in many respects. Galileo wrote in his *Sidereus Nuncius:*

> I have been led to the opinion and conviction that the surface of the Moon is not smooth, uniform, and precisely spherical as a great number of philosophers believe it (and the other heavenly bodies) to be, but is uneven, rough, and full of cavities and prominences, being not unlike the face of the Earth, relieved by chains of mountains and deep valleys.

This evoked the following questions: If mountains and valleys exist on the Moon, then might not cities inhabited by intelligent beings also exist there? And is our Sun the only star accompanied by planets?

These bold ideas were advanced by the Italian philosopher, Giordano Bruno, who wrote: "Innumerable suns exist; innumerable earths revolve about these suns in a manner similar to the way the seven planets revolve around our sun. Living beings inhabit these worlds." The Roman Catholic clergy of his time denounced Bruno for his radical views. He was tried by a tribunal of the Inquisition and burned at the stake in Rome on February 17, 1600. Until the end of the seventeenth century, the Church violently opposed the new heliocentric theory. But in time, the Church adapted its philosophy to the new scientific concepts. Many present-day theologians accept the premise that intelligent beings may exist on other planets and do not find this view inconsistent with the fundamental tenets of their religions.

During the second half of the seventeenth century and into the eighteenth century, a number of scientists, philosophers, and writers (notably, Cyrano de Bergerac, Christianus Huygens, Bernard de Fontenelle, and Voltaire) published works dealing with life on other planets. Some of these publications were eloquent; some, especially those of Voltaire, contained profound ideas; but at the same time they were purely speculative. Brilliant scientists and philosophers such as Kant, Laplace, Herschel, and Lomonosov were advocates of the hypothesis of the plurality of habitable worlds. By the end of the eighteenth century, this hypothesis had gained almost universal acceptance by scientists and intellectuals. However, some were cautious about adopting the view that life existed on *every* planet. William Whewell, an English philosopher, in a book published in 1853, stated that perhaps not all planets were suitable habitats for life. He conjectured that the larger planets of the solar system were composed of "water, gases, and vapor" which would make them unfit for life. "In proportion to their distance from the sun, the inner planets would have large amounts of hot water on their surfaces." Whewell also argued against life existing on the Moon, and his view gradually gained acceptance.

Belief in the existence of extraterrestrial life continued to spread during the eighteenth and early nineteenth centuries. William Herschel, an eminent English astronomer, believed the sun to be inhabited. He thought that sun spots [Fig. 1–3] were apertures in a brilliant shell around the sun enabling us to see into its interior. Hypothetical solar beings inhabited this interior and could admire the stars through the openings in their roof. Even Sir Isaac Newton believed that the sun was inhabited.

In the latter half of the nineteenth century, a book written by Camille

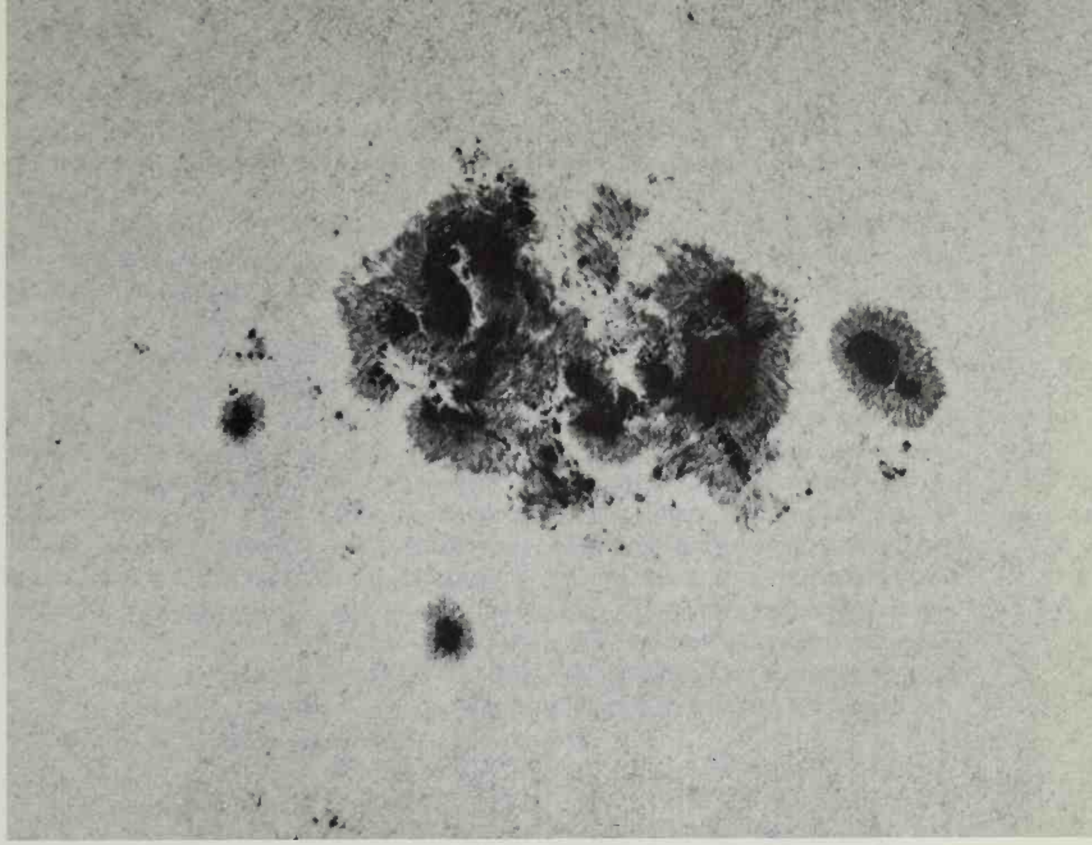


FIGURE 1–3. *Close-up of a sun-spot complex. Against the fine granulation, like grains of rice, of the visible disk of the Sun appear such dark sun-spots. They are much cooler than their surroundings, and are the sites of intense magnetic fields. Some early astronomers thought that sun-spots were windows to a cooler, more habitable region beneath the fiery exterior of the Sun. (Courtesy of Mt. Wilson and Palomar Observatories.)*

Flammarion, *On the Plurality of Habitable Worlds,* attained great popularity in France, and (through extensive translation) in many other countries. In this, as well as other publications, Flammarion contended that the planets were created purposefully for the formation of life. During a 20-year period, he published some 30 works, written in a florid style, which made a deep impression upon his contemporaries. Flammarion appealed more to the emotions of the reader than to his logic, but the appeal was often successful. Even today, reading his works is an evocative experience.

During the late nineteenth and early twentieth centuries various modifications of the panspermia hypothesis received wide circulation. According to this hypothesis, life in the universe exists eternally; living organisms never arise from nonliving matter, but are transmitted from one planet to another. At the turn of the century, the Swedish chemist, Svante Arrhenius, conjectured that microorganisms—spores or bacteria, probably adhering to small specks of dust—are propelled by the pressure of star light from one planet to another. If, by chance, they should land on some planet where conditions for life are favorable, these spores were thought to germinate and initiate the local evolution of life.

Although such transmission of life from planet to planet within a single planetary system cannot be completely discounted, the propagation of panspermia from one planetary system to another is today considered highly unlikely (see Chapter 15). The assumption that life is eternal seems inconsistent with current evidence on the evolution of stars and galaxies. This evidence, based on numerous observations, implies that the Galaxy at one time consisted purely of hydrogen or hydrogen and helium. Only as the Galaxy evolved were the heavier elements, necessary for the origin of life, formed (see Chapter 8). Finally, some calculations of the "red shift" of light from distant galaxies apparently indicate that 10 to 20 billion years ago the conditions in the universe were such that the existence of life would have been highly improbable (see Chapter 10). These observations seem to imply that life originates *de novo* in separate regions of the universe at such stages of the evolution of matter when conditions become favorable. The panspermia hypothesis therefore seems untenable as a fundamental concept.

Konstantin E. Tsiolkovskii, the Russian founder of astronautics, was an ardent believer in the plurality of worlds. He wrote,

> Is it possible that Europe is inhabited and other parts of the world are not? Is it possible for one island to be inhabited and other islands to be uninhabited? . . .
>
> All the phases of the development of life may be found on the various planets . . . Did man exist several thousand years ago and will he be extinct in several million years? . . . This entire process may be found on other planets . . .

While the first quotation seems to echo the view of the ancient philosophers, the second contains a new and important idea. Previous writers had depicted the civilizations on other planets as being socially and technologically similar to those of Earth. Tsiolkovskii, on the other hand, thought that extraterrestrial civilizations might exist at various developmental levels. We should note, however, that his point of view could not then (and cannot even now) be confirmed by direct evidence.

The development of the hypothesis of the plurality of worlds has often been linked to cosmogonic hypotheses—hypotheses concerning the creation or origin of the universe. The hypothesis of the English astronomer Sir James Jeans, which prevailed during the first third of the twentieth century, assumed that the planetary system of our sun was formed as a result of a rare cosmic cataclysm (perhaps the glancing collision of two stars). Life in the universe was an infrequent phenomenon; in our Galaxy (consisting of about 150 billion stars), it was considered highly improbable that more than one other star had a planetary system resembling ours. The failure of Jeans' hypothesis to explain the masses, motions, and composition of the planets, and the rapid development of astrophysics, has led to the present conclusions that there are a vast number of planetary systems within our Galaxy;

our solar system is the rule, not the exception, in a universe of stars. However, this theory, too, has not as yet been conclusively demonstrated [see Chapters 11–13].

Concepts about stellar and planetary cosmogony have had a considerable influence on the study of the origin of life. The age of a star and the time interval during which its luminosity is fairly constant (a condition necessary for the support of life on any accompanying planets) can now be determined. Stellar cosmogony also enables us to predict the future of our sun and hence the fate of life on Earth. Recent achievements in astrophysics have permitted a new scientific approach to the problem of the plurality of habitable worlds.

Today, the question is being approached from an entirely different direction—through molecular biology. It is now apparent that the origin of life can be explained, to a large extent, by studies in the field of chemistry. We are beginning to comprehend by what means and under what environmental circumstances those specific complex organic reactions leading to the origin of life can proceed. In recent years, chemists have made great strides in this direction. The outstanding advances in genetics and the clarification of the significance of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) have permitted a new understanding of the basis of life, and its beginnings. Now for the first time, problems of the origin of life have been subject to laboratory experimentation.

The concept of the plurality of habitable worlds entered a new phase with the first artificial Earth satellite hurled into orbit by the U.S.S.R. on October 4, 1957. The triumphal orbital flights of the Soviet cosmonauts Y. A. Gagarin, G. Titov, A. G. Nikolaev, V F. Byokovsky, P. R. Popovich, and V. Tereshkova, and the American astronauts, John Glenn, M. Scott Carpenter, L. Gordon Cooper, Walter Schirra, and their successors, were in one sense the culmination, and in another sense the bare beginnings, of a series of splendid successes in space technology. Today, many people, both scholars and laymen, are suddenly aware that they inhabit a small planet surrounded by boundless space. Astronomy is now taught in the schools, and students have a vague idea of the relative positions of the Earth and the other heavenly bodies. Some people may still be drawn to a kind of practical geocentrism; but the revolution in our awareness of our surroundings should not be underestimated. It *is* a revolution, one which marks the beginning of a new era in the history of mankind—an era of direct investigation and conquest of the cosmos.

The problem of life on other planets is no longer abstract. It has acquired practical meaning. Experimental methods are being designed for the direct investigation of our solar system. Special devices for the detection of extraterrestrial life will be landed on the surfaces of the planets and may provide a definitive answer to this age-old question. Astronauts will land on the Moon, Mars, and perhaps even on our mysterious and inhospitable neighbor, Venus. Then, at last, man will be able to seek and study extraterrestrial life by conventional biological methods.

There is enormous public interest in the possibility of extraterrestrial life. The primary purpose of this book is to acquaint the reader interested in life in the universe with our *current* state of knowledge. The word "current" is emphasized, because rapid progress is being made in the study of this problem. Other works devoted to this subject deal mainly with the question of life on other planets in our solar system. This book, however, includes, in addition to a study of our own solar system, a discussion of the possibility of life in other planetary systems, and of the problems of establishing contact with civilizations separated from us by interstellar distances.

The book is divided into three parts: Part I contains astronomical information necessary for an understanding of contemporary theories on the evolution of galaxies, stars, and planetary systems. Part II deals with the conditions necessary for the origin of life on a planet. We survey there the question of whether the conditions required for life exist on Mars, Venus, and other planets of the solar system and what direct evidence exists for life beyond the Earth. The contemporary variants of the panspermia hypothesis are also critically assessed. Part III contains an analysis of the possibility of intelligent life in other parts of the universe. Special attention is devoted to the problem of establishing contact with civilizations which may exist on the planets of distant stars.

The first two parts deal primarily with concrete results. In the third section, we are concerned with essentially speculative arguments; we have not as yet established contact with interstellar civilizations, nor do we know if we ever shall. It does not follow, however, that investigations of extraterrestrial intelligence are scientifically superfluous, or that they should be relegated to the realm of fantasy and science fiction. We have attempted to analyze rigorously the latest achievements of science and technology which may, in the future, give us an answer to this problem. The last section also illustrates the strength of man's intellect at its present level of development. Man, through his own activity, has already given real significance to and has in certain respects already changed the cosmos. What can we not expect in a few centuries?

2

Extraterrestrial life as a psychological projective test

"The Ethiopians make their gods black and snub-nosed; the Thracians say theirs have blue eyes and red hair. . . . Yes, and if oxen and horses or lions had hands, and could paint with their hands, and produce works of art as men do, horses would paint the forms of the gods like horses, and oxen like oxen, and make their bodies in the image of their several kinds."

Xenophanes

"If God did not exist, man would be obliged to invent him."

Voltaire

". . . are you so stupid as to think that just because we're alone here, there's nobody else in the room? Do you consider us so boring or so repulsive that of all the millions of beings, imaginary or otherwise, who are prowling around in space looking for a little company, there is not one who might possibly enjoy spending a moment with us? On the contrary, my dear—my house is full of guests . . ."

Jean Giraudoux, *The Madwoman of Chaillot*

▽ The possibility of life beyond the Earth evokes today strong and partisan emotions. There are some who want very much to believe that extraterrestrial life—particularly the intelligent variety—is common throughout the universe; and there are those who are committed to the view that extraterrestrial life is impossible, or so rare as to have neither practical nor philosophical interest. It seems to me appropriate that in this book more than passing attention be paid to such psychological predispositions.

One spring some years ago, the Department of Astronomy received a telephone call from the local District Attorney's office. There was in progress the criminal trial of a gentleman whom I shall call Helmut Winckler, a Nebraskan who claimed to have had personal contacts with inhabitants of the planet Saturn. The state desired an expert witness. With wry comments and professional asides, the message was conveyed to me. I agreed to serve as a surprise witness for the prosecution and was presented with a sheaf of publications written by or about the defendant, concerning his extraordinary adventures.

The following is my recollection of the statements made in these publications and subsequently verified by the defendant under oath:

Winckler was a salesman of agricultural implements in Nebraska; he was of German ancestry, but born in the United States. He had few intellectual pretensions, but had at least a grammar school education. Winckler was a trifle chubby, of florid complexion and robust countenance, and wore steel-rimmed glasses. He looked considerably younger than his sixty-odd years, was soft-spoken and polite on the witness stand. His accent was that of the rural midwest.

Winckler testified that one day while motoring along a back road in Nebraska, he had the remarkable good fortune to overtake a parked flying saucer. Naturally he stopped, as anyone would. To his surprise, he observed emerging from the saucer, several men and women of entirely human appearance, dressed in flowing robes and speaking mellifluously. The language which they were speaking so mellifluously was Hochdeutsch. Coincidentally, Winckler understood Hochdeutsch. The saucerians were delighted. Winckler was delighted.

The saucerians explained that they were from the planet Saturn, and had, for reasons of their own, selected Winckler as an "intermediary." They were to impart to him information of great value for the inhabitants of the planet Earth. It seems that the international political situation on Earth had reached serious proportions, a fact which concerned the Saturnians, an old, wise, and sympathetic race. They were here to save us from ourselves. Winckler never revealed why *he* had been selected. It appears that it was not because he knew Hochdeutsch. The Saturnians speak all human tongues.

Winckler accompanied the Saturnians into their saucer. His publications contain diagrams of the interior of the saucers, replete with esoteric and unintelligible descriptions of their method of propulsion. The interiors resemble those of the 1958 Buick.

The group then departed for the Arctic Circle, taking Winckler with them. On a subsequent saucer foray which Winckler made with the Saturnians, he was taken beneath the Bering Straits to inspect the Soviet Union's suboceanic missile emplacements. You may not have heard of them, but according to Winckler the government of the United States knows all about them.

In another of Winckler's flights, this time over the North Pole, the Saturnians were heard to remark that only a few months ago the axis of the Earth was tilting a dangerous six degrees. Winckler paled, but the Saturnians assured him that they had repaired the misalignment in short order.

When the court recessed, I queried Winckler about this delinquent behavior of the Earth's axis of rotation. I explained that even a very much smaller tilt would have been discovered immediately by astronomers who, each night, point their telescopes by assuming that they know precisely where the celestial poles are. Winckler's response was that he could hardly be held responsible for statements made by inhabitants of the planet Saturn. He was merely relaying information. But he left the distinct impression that the Saturnians knew quite a bit more about the subject than do bystanders in courtrooms.

On another expedition, the Saturnians took Winckler to that Mecca of the occult, the Great Pyramid of Gizeh in Egypt. They mingled with a group of tourists being guided *through* the interior of the pyramid. (I have a vivid mental image of this procession: Egyptian guide, two middle-aged ladies from Dubuque, some assorted French and German tourists, six Saturnians in flowing robes, and, bringing up the rear, Helmut Winckler in levis.) At a certain intersection of pathways, the tourists went in one direction, and Winckler and the Saturnians in the other. They were confronted with a blank wall. Appropriate pressures were applied to appropriate bricks, and the wall slid open, revealing a chamber within. The party entered, and the stone door slid silently shut behind them. In the room were (1) a small, one-man flying saucer, quite dusty with age; (2) a large and equally ancient wooden cross perhaps ten feet high; and (3) a toroid of thorns about eight inches in diameter. The Saturnians offhandedly explained that one of their number had attempted a mission to Earth some two thousand years ago. He had met with somewhat qualified success.

In extracting these marvels from Winckler, the Assistant District Attorney first displayed shock, then disbelief, and finally, righteous indignation. He shook his head and peered upwards—awaiting the thunderbolt which doubtless would terminate the proceedings. The courtroom was hushed. The jury was awed. Winckler was cool. From the tone of his voice, he might have been describing a reaper sale in Lincoln.

One of the principal pastimes of the Saturnians, while cavorting about the Earth, was their remote geological survey. They possessed instruments capable of determining, from quite remarkable altitudes, the location and distribution of mineral-bearing ores. They discovered untapped veins of gold, pristine platinum-bearing rocks, and uranium ores unknown to man. The locations of these finds were carefully kept from Winckler. One day, however, the Saturnians pointed out to Winckler the existence of a quartz mine—while flying over southern California. You may think that quartz is rather uninteresting, compared with gold, platinum, and uranium. But this was a rather special kind of quartz. It cured cancer.

Soon after landing, Winckler was selling quartz stock. I have the distinct impression that he sold half-interests in the mine—several dozen half-interests. Adding to his subsequent embarrassment was the fact that the mine was already owned by another gentleman and was actively producing quartz. It also transpired that by a perfectly remarkable coincidence, Winckler had visited this very mine several years earlier.

But none of these activities directly precipitated Winckler's indictment. His difficulties are traceable to his promotion of the healing properties of quartz among elderly and wealthy widows. Many ladies had lent him sizable sums to advance his venture. In the last months before his arrest, it was his practice to speak before large gatherings of flying saucer enthusiasts—some but by no means all of whom were also elderly, wealthy widows. They paid his transportation and living expenses, invited him to their homes, and accorded him the customary respect due prophets of new religions. The complaint leading to Winckler's arrest arose from a combination of money borrowed and not repaid, and affection promised and not delivered.

To give you the flavor of the courtroom proceedings, here is an approximation of the dialogue which preceded my appearance on the stand. After establishing that Winckler was married, back in Nebraska, the Assistant District Attorney pursued the following line of inquiry:

ASSISTANT DISTRICT ATTORNEY: Now, Mr. Winckler, you have several times told this court that it was not your custom to use words or gestures of affection in your dealings with Mrs. Brewster.

WINCKLER: Yeah, except when I would call her Lovey and Dear, like I do in my work.

ASSISTANT DISTRICT ATTORNEY: But you never expressed to Mrs. Brewster any deep feelings of love and affection?

WINCKLER: Yeah, that's right.

ASSISTANT DISTRICT ATTORNEY (hands clasped behind his back, eyes fixed on a point on the ceiling, and slowly pacing): Now, Mr. Winckler, did you, for example, ever convey to Mrs. Brewster sentiments such as these . . . ?

At this point, the Assistant District Attorney recited from memory some twenty lines of verse, dactylic trimeter in rhymed couplets, expressing very tragic sentiments.

WINCKLER: I never said nothing like that in my life.

ASSISTANT DISTRICT ATTORNEY: Your Honor, I would like to submit in evidence People's Exhibit Number 14.

The Assistant District Attorney passed a piece of paper to the judge. The judge read it and handed it to the Court Clerk. The clerk wrote in a large book and returned the paper to the Assistant District Attorney.

ASSISTANT DISTRICT ATTORNEY: Now, Mr. Winckler, would you care to examine the signature on the bottom of this card?

WINCKLER (polishing glasses, scrutinizing the paper): It *looks* like my signature.

ASSISTANT DISTRICT ATTORNEY: Mr. Winckler, do you now recall having sent this greeting card to Mrs. Brewster?

WINCKLER: Yeah, I guess so. But, see, it was this way. A guy come to my door a couple years ago. He was a disabled veteran and he was selling greeting cards. He sold me 200 cards, all kinds. I had these cards in a big box, and when it was her birthday, I just took one out and sent it. I didn't even read it.

ASSISTANT DISTRICT ATTORNEY: It's a wonder you didn't pick out a funeral condolence by mistake.

On the afternoon of the same day, I was called to the witness stand. I had been told that as a surprise witness, giving testimony in a specialized field, I would be subjected to no serious cross-examination. In this, the prosecution was mistaken.

The Assistant District Attorney inquired of my name and affiliation and established my academic credentials. He then invited me to discuss the likelihood of human beings inhabiting the planet Saturn.

I described the operation of the spectroscope and explained how it gives information on the chemical composition of distant objects. Astronomical spectroscopy of Saturn, I noted, shows its atmosphere to contain no oxygen, and large amounts of methane and ammonia, compounds which are poisonous to human beings.

I then outlined the way in which the temperatures of planets are measured with a thermocouple at the focus of a large telescope. The temperatures of those parts of Saturn which are accessible to our telescopes are several hundred degrees below zero Fahrenheit. Finally, I described how the surface gravity of a planet could be determined from its mass and radius, and mentioned that since the surface gravity on Saturn is some 17 percent greater than on Earth, any beings which evolved there would probably be squatter than we.

I summarized my testimony by saying that while these observations by no means exclude the possibility of some sort of life on Saturn, they provide quite convincing evidence that there are no human beings there. Indeed, I concluded, it would be most remarkable if four and a half billion years of independent biological evolution on the two planets had produced identical end results, even if their environments were not so dissimilar.

The defense attorney was then asked if he wished to cross-examine. Indicating that he might have a question or two, he approached the witness stand slowly, his voice gradually increasing in volume:

"Dr. Sagan, I don't mean to be disrespectful, but isn't it a fact that four or five hundred years ago, university scientists like yourself were maintaining that the Earth was . . . *flat?*"

The Assistant District Attorney leaped to his feet.

"Your Honor, I object!"

The judge inquired on what grounds he objected. Surely, on grounds of irrelevance, I thought. But no.

"Hearsay evidence."

The objection was sustained, and the defense attorney continued. The riposte and parry had not been lost on the judge, who was smiling faintly. But the jury maintained its glazed and somewhat haggard expression.

The defense attorney was puzzled by one part of my testimony. He understood, he thought, how astronomical spectroscopy worked, and how it was checked by laboratory comparison with the gas in question—for example, ammonia. But hadn't I been keeping an assumption from the jury? Weren't we tacitly assuming that the same physical laws apply on Saturn as on Earth?

Suddenly, in a proceeding for fraud in a criminal court, we had plunged into one of the basic questions of the philosophy of science. I explained that there are large numbers of spectral lines which are indicative of the presence of a compound, and that many of them exist both on Saturn and in the laboratory. The possibility of such a random coincidence would seem to be very small. I went on to describe how Newton had demonstrated that the same physical laws accounting for the gravitational attraction of objects on the Earth are responsible for the motion of the Moon. Glancing over to the jury, however, I had the distinct impression that the seed of doubt had been planted. I could imagine them thinking: After all, maybe the physical laws *are* different on Saturn. How does anyone know?

The defense then inquired about the temperature determination. What level in the atmosphere of Saturn did the temperatures which I quoted refer to? High in the atmosphere. How high? Well, possibly 10,000 miles above the surface —if any. Might the temperatures down at the surface—if any—be considerably warmer? Indeed they could, I replied; in fact, I had made a similar suggestion concerning the planet Jupiter. What temperature for the planet Earth would be derived by an extraterrestrial observer if he could only look at the top of our clouds? About $-60°$ or $-70°$ Fahrenheit. But we all know that that isn't the average temperature on the surface of the Earth, don't we? He had scored.

And now, doesn't the spectroscopic determination of composition also refer only to the very high atmosphere? And mightn't the chemical composition of the lower atmosphere near the surface be considerably different? In particular, mightn't it contain molecular oxygen, so that beings there could breathe as we do?

I replied in terms of chemical equilibrium. There is such an overabundance of hydrogen in the upper atmosphere of Saturn, I said, that it would instantly react with any of the oxygen around. Some estimate can be made of the abundance of hydrogen in the lower atmosphere; the body of Saturn is believed to be at least in part metallic hydrogen. I thought it highly unlikely that free molecular oxygen existed at the surface of Saturn. The defense attorney replied, "But these are indirect arguments, aren't they? You don't really *know* there's no oxygen on Saturn." I could only agree that the evidence, while convincing, was indirect. But astronomy is based on indirect evidence.

"Now, Dr. Sagan," he continued, "I have heard it said that fossil plants have been found in the arctic regions of the Earth, and that these fossil plants were of a tropical variety. Is this true? How could there be tropical plants at the pole?" I felt that I had completely lost the direction of the discussion. I explained the evidence for polar wandering garnered from studies of paleomagnetism. The present arctic regions of the Earth were once sub-tropical because the poles were once at a different geographical location.

"Now what I really want to know, Dr. Sagan, is this: At the time when the present poles of the Earth were in the tropics, were the present tropics at the poles? Or was it tropical all over?"

At this point the Assistant District Attorney somewhat resignedly got to his feet. "Your Honor, I must confess that I fail to see the relevance of this line of inquiry."

"I find that I must agree with the Assistant District Attorney," said the judge. "Perhaps the counsel for the defense could enlighten us."

"Well, Your Honor, it's not really relevant, but it's such an interesting topic that I thought, while we had Dr. Sagan on the stand, it might be fun to pursue it. I have no further questions."

I stepped down from the stand and resumed my seat in the audience. One local newspaper, the following day, printed in full the sequence of rhymed couplets that Winckler had sent to the unfortunate Mrs. Brewster, and then quoted me as saying that the temperatures on Saturn were several hundred degrees below zero Fahrenheit, and that this was too cold for love on any world.

The remaining topic of substantial interest—which was arranged by considerable side-of-mouth conversation between representatives of the prosecution and defense—was a motion picture film which Winckler had produced. It was shown in court, to the great delight of the jurors. It showed the landing of the Saturnians and their discussions and adventures with Winckler. Unfortunately, Winckler was unable to procure any Saturnians for the movie, and so had hired actors who were indistinguishable from Saturnians. In addition, Winckler had promised various leading roles in the movie to some of his lady friends, but at least in the case of Mrs. Brewster, this promise of stardom did not materialize.

Winckler was subsequently found guilty of fraud and sentenced to prison, despite the admirable efforts of his attorney. In my discussions with Winckler during recesses, I was unable to decide to what extent his escapades with the Saturnians were a conscious fraud, and to what extent he genuinely believed his account. But it was clear that many others found Winckler's adventures in ringing consonance with what they believed—or would like to believe.

Winckler's experience underlines the existence of an unfulfilled need in contemporary society. Almost any other of the many accounts of alleged contacts of human beings with the crews of flying saucers—accounts which regale the flying saucer societies—follow the same pattern and stress the same points. The extraterrestrials are human, with few even minor physical differences from local cosmetic standards. (I know of no case of Negro saucerians, or Oriental saucerians, reported in the United States; but there are very few flying saucer contact reports made in this country by Negroes or by Orientals.) The saucerians are wise and gentle and loving; concerned for our safety during this epoch of continuing international tensions, yet for some reason unwilling to intervene in force. They have long ago solved international disputes on their home planets. They have great gifts in the humanities—this is, of course, still an appropriate subject for them—but also immense technical abilities. In short, the saucerians are all-powerful, all-knowing, and concerned with the plight of mankind as a parent would be for his children. Yet they do not direct the course of the major events of the day, presumably on the grounds that mankind must work out its own destiny. I cannot help but conclude that the flying saucer societies represent a thinly disguised religion, and that the saucerians are the deities of the cult.

As science has progressed during the last few centuries, areas which were originally the exclusive province of religion have been increasingly preempted by science. We no longer hold that the Earth is stationary, or that it is at the center of the universe; nor that the world was made even approximately on October 23, 4004 B.C.; nor that it was made in seven days; nor that different species had separate creations; nor that the origin of the solar system and the origin of life are forever beyond the ken of man. Rather, the laboratory synthesis of life from materials which were abundant in the early environment of the Earth seems no more than a decade off. One result of these encroachments by science has been that there seems less and less for God to do. If he creates some hydrogen at the beginning of the universe, and establishes the physical laws, he can then retire. He is a *roi fainéant*. If God did not directly make life or man, it is hard to believe that he will intervene in our everyday lives, or answer our supplications.

Yet the temptation to believe in an omnipotent, omniscient, and loving God is especially great today. The pace of world events is out of the hands of the ordinary individual. We have no assurance that tomorrow will not find the world a radioactive pyre. Our destiny appears to be at the mercy of forces we little understand and only perilously control. If only there existed a god who was concerned with our plight, who would give some assurance of our survival; but who was explicable within the framework of contemporary science. . . . The saucer myths represent a neat compromise between the need to believe in a traditional paternal God and the contemporary pressures to accept the pronouncements of science.

While the saucerian contact cult is viable and widespread—at least, in the United States—it represents only a small fraction of the total number of saucer enthusiasts. There are large numbers of people who have, in all good conscience, observed unknown objects in the skies which they have called "unidentified flying objects"—UFO's—and which they believe to be of intelligent extraterrestrial origin. The diversity of these reports is as striking as the observations themselves.

UFO's have been described variously as rapidly moving or hovering; disk-shaped, cigar-shaped, or ball-shaped; moving silently or noisily; with fiery exhaust, with no exhaust whatever; accompanied by flashing lights, or uniformly glowing with a silvery cast. It is immediately clear that all UFO's do not share a common origin. Indeed, the use of a generic term such as "UFO's" or "flying saucers" has served to confuse the issue by implying a common origin.

As detailed by the American astronomer Donald H. Menzel of Harvard College Observatory, confirmed identifications of UFO's have been made with the following: unconventional aircraft; aircraft under uncommon weather conditions; aircraft with unusual external light patterns; meteorological and other high-altitude balloons; artificial earth satellites; flocks of birds; reflections of searchlights or headlights off clouds; reflection of sunlight from shiny surfaces; luminescent organisms, including one case of a firefly lodged between two adjacent panes of glass in an airplane cockpit window; optical mirages and looming; lenticular cloud formations; ball lightning; sundogs; meteors, including the green fireballs; planets, especially Venus; bright stars; and the Aurora Borealis. Radar detection of unidentified flying objects has also occurred occasionally. Many of these sightings have been explained in terms of radar reflection off temperature inversion layers in the atmosphere, and other sources of radar "angels."

Considering the difficulty in tracing out the visual and radar sightings—the visual sightings are often made by individuals with little experience in observing the skies—it is remarkable that all but a few percent of the reported UFO's have been identified as naturally occurring, if sometimes unusual, phenomena. It is remarkable that the professional astronomers, who are familiar with the skies and regularly scan them with sophisticated instrumentation, have never, to the best of my knowledge, photographed anything like the classical UFO. For example, in the Harvard Meteor Project, performed in New Mexico during the period 1954–1958, extensive photographic observations were made by Super-Schmidt cameras with a 60° field of view. In all a surface area of 7000 km^2 was observed to 80 km altitude for a total period of some 3000 hours. Visual and photographic observations were good down to magnitude +4. (The magnitude scale is defined in footnote 2 of the following chapter; a magnitude of +4 is close to the faintest object visible with the naked eye.) These observations by professional astronomers were made in a locale and period characterized by extensive reports of unidentified flying objects. No unexplained objects were detected, despite the fact that rapidly moving objects were being sought in a study of meteors. Similar negative results have been obtained by large numbers of astronomers, and help to explain the general skepticism of the astronomical community towards flying saucer reports. There is no way to exclude the very occasional presence of unidentified objects in our skies, but the run-of-the-mill flying saucer observations (made in the United States on the average of about once a day) are certainly common astronomical objects and atmospheric phenomena—and perhaps some not so common—which have been misinterpreted by the observer.

Repeated sightings of UFO's, and the persistence of the United States Air

Force and members of the responsible scientific community in explaining the sightings away have suggested to some that a conspiracy exists to conceal from the public the true nature of the UFO's. But precisely because people desire so intensely that unidentified flying objects be of benign, intelligent, and extraterrestrial origin, honesty requires that in evaluating the observations, we accept only the most rigorous logic and the most convincing evidence.

There is also the opposite danger. Public interest in flying saucers, contact reports, and extraterrestrial life in general has proved a frequent source of embarrassment to many scientists, whose statements tend to be distorted, exaggerated, and otherwise perturbed by the bright light of popular concern. There is then a tendency to reject out of hand the possibility of extraterrestrial intelligence as baseless, improbable, or unscientific. There are also covert Ptolemaicists who find the prospects of extraterrestrial intelligence threatening.

A typical example of this other projective extreme can be found in the circumstances attending the first release of scientific results from the United States spacecraft Mariner IV, which encountered Mars on Bastille Day, 1965. Among the early announcements was the finding that Mars has no detectable magnetic field. The conclusion drawn (by no means a secure one, incidentally) was that Mars lacked mountains and volcanoes and could be considered geologically dead. Some segments of the press then reported that by failing to measure the Martian magnetic field, scientists had proved Mars lifeless—truly a marvel of twentieth century thought. The confusion between these two senses of the word "dead" was never noted or retracted, to the best of my knowledge, by the press.

The magnetometer results then set the stage for the popular interpretation of the Mariner IV photographs. First, despite the fact that a similar experiment directed at the Earth would be incapable of detecting life on our planet (see Chapter 18), since no life could be *seen* on Mars, the news media deduced a lifeless planet. Second, since no sign of recent bodies of water could be found on Mars—as expected—it was concluded that there was no life on Mars. Finally, the existence of craters on the Martian surface implied to many that Mars is lifeless. The syllogism seemed to go, "There are craters on the Moon. There is no life on the Moon. There are craters on Mars. Thus, there is no life on Mars."

Newspapers, magazines, television, and press releases are still replete with descriptions of how the "widely held" view of a lush, vegetated and canal-crossed Mars has now been abandoned because of the decisive findings of Mariner IV, and replaced by a lifeless, cratered Moon-like world. Even the often reliable New York *Times* displayed an editorial entitled "The Dead Planet," listing the supposed new findings about life on Mars. As we shall see in Chapters 19 and 20, these scientific conclusions drawn by public relations officers and by the news media do justice neither to the painstaking efforts of ground-based astronomy, nor to the exciting and significant findings of Mariner IV. This spacecraft was not designed to search for life on Mars. As the experimenters were careful to point out, the mission neither demonstrated nor precluded the possibility of life on Mars.

Why then were the communications media so quick to deduce a lifeless Mars?

I believe a partial answer can be found in the responses made to the Mariner IV findings by political leaders, by Mr. Billy Graham, and by other American divines—often sure barometers of common attitudes. They were unmistakably *relieved.* Finding life beyond the Earth—particularly intelligent life, although this is highly unlikely on Mars—wrenches at our secret hope that Man is the pinnacle of creation, a contention which no other species on our planet can now challenge. Even simple forms of extraterrestrial life may have abilities and adaptations denied to us. The discovery of life on some other world will, among many things, be for us a humbling experience.

The question of extraterrestrial life—and even more so, the question of extraterrestrial intelligence—is then many things to many men. In assessing evidence for extraterrestrial life, and in evaluating statistical estimates of the likelihood of extraterrestrial intelligence, we may be at the mercy of our prejudices. At the present time, there is no unambiguous evidence for even simple varieties of extraterrestrial life, although the situation may change in the coming years. There are unconscious factors operating, in the present arguments of both proponents and opponents of extraterrestrial life.

I think Shklovskii and I can be described as cautious optimists on this question. In many places in the present book, we have made speculations, but I hope that we have labelled our speculations as such, and given the reader enough information to evaluate the basis for our speculations. In Part III, where we extrapolate from contemporary terrestrial technology to future extraterrestrial technologies, perhaps we have not been cautious enough, but I rather suspect that the opposite is the case. It is chastening to read nineteenth-century prognostications of the events of the middle twentieth century. Even their most grandiose extrapolations have proved a pale echo of our realities. It strained Jules Verne's imagination to picture giant passenger balloons transporting people through the air over thousands of miles in a period of only a week. He could not imagine contemporary jet transports, which cover the same distance in hours.

Whether we have been too optimistic or not optimistic enough, only the future will tell. △


I AM ALWAYS IMPRESSED BY THE CONSTANCY OF THE STARS..
IT GIVES ME A FEELING OF SECURITY TO LOOK UP, AND KNOW THAT THE STAR I SEE WILL ALWAYS BE THERE, AND WILL...

I

THE UNIVERSE

In a universe whose size is beyond human imagining, where our world floats like a dust mote in the void of night, men have grown inconceivably lonely. We scan the time scale and the mechanisms of life itself for portents and signs of the invisible. As the only thinking mammals on the planet—perhaps the only thinking animals in the entire sidereal universe—the burden of consciousness has grown heavy upon us. We watch the stars, but the signs are uncertain. We uncover the bones of the past and seek for our origins. There is a path there, but it appears to wander. The vagaries of the road may have a meaning, however; it is thus we torture ourselves.

Loren Eiseley, *The Immense Journey* (1946)

3

The size and structure of the universe

What a wonderful and amazing Scheme have we here of the magnificent Vastness of the Universe! So many Suns, so many Earths . . . !

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

And with an awful, dreadful list
Towards other galaxies unknown
Ponderously turns the Milky Way . . .

Boris Pasternak

▽ The great seventeenth-century religious philosopher and mathematician, Blaise Pascal, expressed a fear of the great spaces between the stars, and if some men have become more daring since his time, others have not become less afraid. △

The scale of space and time which we customarily observe on Earth, and which is part of our daily lives, is insignificant when compared with cosmic space and time. When, for the first time, we become aware of the vastness of our universe, we are awed and humbled.

But if astronomers spent all their time speculating about the immensity of the cosmos and the prodigious intervals of time necessary for the evolution of the stars, their accomplishments would be few. The primary concerns of the astronomer when studying the cosmos are physical and mathematical interpretations of observations, predictions of future observations, and the development and refinement of his instruments for such interpretation and prediction. To help him in the conceptualization of his problem, the astronomer may, consciously or unconsciously, visualize a small model representing the cosmic system under investigation. Using this method, he can arrive at an understanding of the relative dimensions of the system and an understanding of the time intervals involved.

I have spent a great deal of time in the study of the solar corona and the Galaxy. ▽ The solar corona is an extended halo of glowing gas which surrounds the Sun and which can be seen during a solar eclipse. The Galaxy is a system of stars, called the Milky Way, in which our Sun resides. The Galaxy is surrounded by a halo of gas called the Galactic corona [Fig. 3–1]. The sizes of the solar and Galactic coronae are, of course, very different. △ I have always visualized each of them as irregular, approximately spherical bodies with somewhat the same dimensions—about 10 centimeters (4 inches) across. Why 10 centimeters? This figure is entirely arbitrary; it is convenient and easily visualized. I have sketched the outlines of the objects of my reflections in my notebook, attempting to preserve the apparent scales of the phenomena. I am, of course, quite aware that the dimensions of the Galactic corona are at least 100 billion times greater than those of the solar corona. However, I could ignore this fact, since the absolute size was not important for an understanding of the problem at hand. When the actual dimensions of an object are of special significance, I use formal mathematics.

Until very recently the dimensions of the Earth were thought to be immense. Four and a half centuries ago it took Ferdinand Magellan and his men almost three years to circumnavigate the globe. Less than a hundred years ago, Phileas Fogg, the courageous hero of Jules Verne's novel, using the latest scientific achievements of his day, traveled *Around the World in 80 Days*. In 1961, our planet's first space travelers, Gagarin and Titov, flew around the globe in 89 *minutes* in the cosmic ships "Vostok." ▽ Thus, the apparent size of the Earth has shrunk as vehicles of increasingly greater speed have been constructed. At the same time, △ the thoughts of men have almost involuntarily turned to the vast reaches of space in which our tiny planet is nearly lost.

There are nine known planets in our solar system. The Earth is situated relatively close to the Sun, although both Mercury and Venus are closer. The mean distance between the Sun and its most distant planet, Pluto, is forty times greater than the distance between the Earth and the Sun. At the present time we do not know if there are any planets further from the Sun than Pluto. We can only speculate that, if such planets do exist, they are relatively small in size and so have escaped detection.

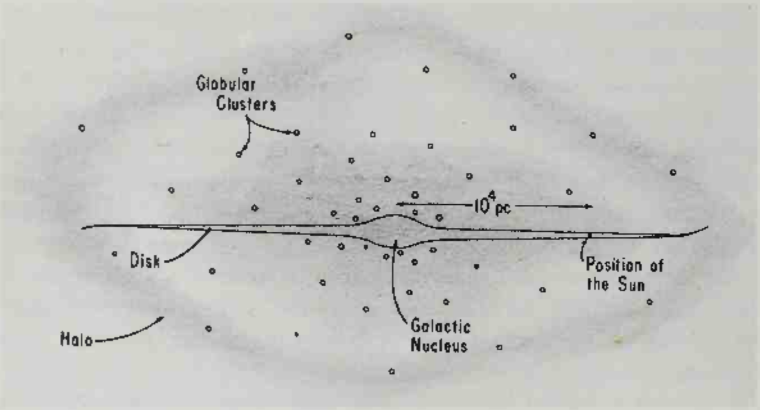


FIGURE 3–1. *Schematic illustration of our Galaxy. The Sun lies in a spiral arm in the disk, or Galactic plane, where the density of stars is relatively low compared with the nucleus. When we look above or below the plane of the Galaxy, we see only the stars in our local neighborhood, but when we look in along the disk, towards the Galactic nucleus, we see a broad, diffuse band of stars along the night sky, which we call the Milky Way. It is our Galaxy seen edge-on. Surrounding the Galaxy is a spherical cloud of hot gas called the halo, which is sparsely populated by stars, most of them in globular clusters.*

The diameter of the solar system is approximately 50 to 100 astronomical units, or 10 billion km.[1] This is an extremely large figure in our scale of distance, about a million times greater than the diameter of the Earth.

We can better visualize the relative sizes in our solar system by imagining a scale model: Let the Sun be represented by a billiard ball with a diameter of 7 cm. On this scale, Mercury, the planet closest to the Sun, would be at a distance of 280 cm, the Earth at a distance of 760 cm, Jupiter, the largest planet, approximately 40

[1] △An astronomical unit, abbreviated A.U., is the average distance from the Earth to the Sun, approximately 150 million km, or 93 million miles.

A kilometer (km) is about 0.62 miles
A meter (m) is 10^{-3} kilometers, or about 39 inches
A centimeter (cm) $= 10^{-2}$ m $= 10^{-5}$ km, or about 0.4 inches
A millimeter (mm) $= 0.1$ cm, or about 0.04 inches △

m, and Pluto, the most distant planet, would be nearly 300 m from the billiard ball. The diameter of the Earth would be slightly greater than 0.5 mm; the lunar diameter would be approximately 0.1 mm, and its orbit around the Earth would have a diameter of about 4 cm. The nearest star beyond the Sun, Proxima Centauri, would be at 2000 km, or about 1200 miles away—so far that the huge planetary distances shown on our scale would seem insignificant by comparison.

The kilometer, the centimeter, the mile, and all other units of measurement were adopted because of the practical requirements of man on Earth. They are obviously inadequate for gauging cosmic distances. The "light year" is used as a unit of measure for interstellar and intergalactic distances in science fiction—and sometimes in scientific literature. One light year is the distance traveled by light in one year, moving at a velocity of 300,000 km per second, ▽ or 186,000 miles per second. Since there are about 3×10^7 seconds in a year, a light year is about 2×10^5 miles per second $\times 3 \times 10^7$ seconds $= 6 \times 10^{12}$ miles, or six trillion miles. △

The special unit of measure usually employed in scientific literature is the "parsec," defined as the distance from which the radius of the Earth's orbit subtends an angle of one second of arc. ▽ Apparent distances in the sky, as seen from Earth, are often expressed in angular measure. There are 360 degrees in a circle; each one is a degree of arc. Thus, a large imaginary circle, drawn above us in the sky and passing directly overhead, would encompass 180° from horizon to horizon. Each degree of arc contains 60 minutes of arc, and each minute of arc contains 60 seconds of arc. A second of arc is a very small unit of angular measure. △ A one-kopeck coin would subtend an angle of one second of arc at a distance of 3 km. ▽ Or, for those readers with no kopecks handy, a second is the angle which a U.S. quarter dollar makes at a distance of $3\frac{1}{4}$ miles. △ The full moon subtends an angle of half a degree. ▽ In order for the radius of the Earth's orbit—150,000,000 km—to subtend one second of arc, it must be at a great distance. That distance, 1 parsec, is 3.26 light years. △

There are no known stars within one parsec of our solar system. Even Proxima Centauri, our nearest neighboring star, is 1.3 parsec (abbreviated "pc") away. In the scale used in our previous analogy to depict our solar system, we found the distance to the nearest star to be 2000 km. Our Sun and its planets are thoroughly isolated from the surrounding stellar systems.

The Sun is a modest member of an enormous collection of stars and dust which we call our "Galaxy" (from the Greek word *gala,* meaning milk). This mass of stars which, on a moonless night, resembles a broad swath of light crossing the sky, is called the Milky Way. It is estimated that there are more than 100 billion stars of various types and ages in the Milky Way. They are found, for the most part, within a gigantic disk with a diameter of approximately 100,000 light years, and with a thickness of some 1500 light years.

The Galaxy has an extremely complex structure. As a preliminary description, let us say that its shape is that of a flattened disk, or a large, rotating wheel. ▽ Figure 3–1 gives a schematic and idealized view of our Galaxy, as seen from a

FIGURE 3–2. *NGC 5364 in the constellation Virgo, a typical spiral galaxy seen face-on. (Courtesy of Mt. Wilson and Palomar Observatories.)*

million light years away. The thick, lens-shaped central region is called the *disk,* and contains the spiral arms which originate near the center and wind outward. They would be prominent if the Milky Way could be viewed from above or below, as in photographs of other galaxies [see Fig. 3–2]. The lens-shaped disk is surrounded by a roughly spherical halo, or galactic corona, which is composed most noticeably of hot gas. △

Stellar density in the Milky Way is not uniform. The brightest region, which has the densest aggregation of stars, is the Galactic nucleus, where, according to the latest data, there are approximately 2000 stars per cubic parsec. This is much greater than the average star density in the neighborhood of our own Sun. The stars also tend to form groups and clusters. The Pleiades is an example of a cluster visible to the naked eye [Fig. 3–3].

Certain types of stars are found in greater numbers in some parts of the Galaxy than in others. For example, giant, hot young stars occur mainly among the spiral arms near the Galactic plane. Older stars, of relatively small mass, appear more numerous in the Galactic center. The so-called globular clusters [Fig. 3–4] are found mostly in the center of the Galaxy, but out of the Galactic plane.

Our Sun is located well away from the center of the disk, in the Galactic suburbs. It lies about 30 light years from the Galactic plane, a distance which is

FIGURE 3–3. *The Pleiades, a nearby open, or "galactic" star cluster. The four luminescent spikes emanating from many of the bright stars in this picture are not intrinsic to the stars, but are due to diffraction effects in the reflecting telescope used to take this picture. (Courtesy of Mt. Wilson and Palomar Observatories.)*

FIGURE 3–4. *The globular star cluster M13 in the constellation Hercules. Some globular clusters are swarms of stars containing tens of thousands of individual members. Note that the density of stars in the center of M13 is so great that we are unable to resolve individual members. (Courtesy of Mt. Wilson and Palomar Observatories.)*

relatively small compared to the total thickness of the stellar disk. The distance from the Sun to the Galactic center is approximately 33,000 light years—10,000 pc.

The stars move in a very complicated way within the Galaxy. Primarily, they participate in the rotation of the Galaxy about its axis, which is perpendicular to the Galactic plane. This motion is different from the rotation of a solid body since different regions have different angular velocities. The Sun and the nearby stars in the solar neighborhood, a region several hundred light years across, rotate at about 250 km per second. A simple rule might be kept in mind: A velocity of 1 pc per million years is approximately equal to a velocity of 1 km per second. Other portions of the Galaxy may rotate at greater or lesser velocities. Our Sun takes some 200 million years to make a complete revolution. Since we estimate that our solar system has existed for about 5 billion years—from its birth from a cloud of gas and dust to its present state—we conclude that it has made some 25 revolutions about the axis of the Galaxy ▽ (5×10^9 yrs/2×10^8 yrs $= 25$). △ We can therefore say that the age of the Sun is 25 Galactic years.

In addition to this motion about the Galactic center, the stars have their own peculiar chaotic motions. These velocities are considerably less—about 10 to 50 km per second—although different types of stars move at different speeds. The hot, massive stars have the smallest velocities (6 to 8 km per second); stars similar to our Sun have a velocity of about 20 km per second. The smaller the velocity, the more time a given star spends in the vicinity of the Galactic plane.

▽ These stellar motions are determined in a variety of ways. For example, we can compare photographic plates taken of the heavens many years apart and see the changes which have occurred in the relative positions of the stars. These peculiar motions are recorded in angular measure—e.g., seconds of arc per century of observation. To convert angular velocities to real velocities—e.g., kilometers per second—it is necessary to know the actual distance of the star from the Earth.

▽ The earliest, and still the most fundamental, astronomical distance determination involves triangulation, the same method used by surveyors to determine the distance to an inaccessible point. The astronomer observes the star of interest from two different, widely separated points, and notes the apparent motion of the star against a background of more distant objects. This effect can be demonstrated easily by holding a pencil a foot away from you and alternately opening and closing each eye. The pencil is seen to move relative to the background. The further away the pencil is from you, the less it seems to move as you wink your eyes. If your eyes were much further apart, the pencil could be seen to move as you wink, even when it was quite far away. Similarly, in astronomical observations, the greater the baseline between the two observations, the larger the measurable distances.

▽ At first these observations were made at observatories in different parts of the world; then, observations were made at the same observatory, but six months apart, so the much larger baseline of the diameter of the Earth's orbit could be used. Since the Sun's own peculiar motion in relation to the neighboring stars is known, observations made many years apart can be utilized to give an even larger baseline, the distance the Sun has moved through its local neighborhood in the course of years.

▽ Another method used for determining velocities is the Doppler effect. For sound waves, the Doppler effect is familiar to us in the change of pitch of an approaching or receding automobile horn. An analogous Doppler effect exists for light waves where the frequency (or color) of the light changes according to the motion of the light source. A star moving away from us becomes redder; a star moving toward us becomes bluer. Astronomical spectroscopy provides a very precise method for measuring even minute changes in the frequency, or color, of light. Such observations of the Doppler effect are of particular significance in the study of the apparent recession of the galaxies from us [see Chapter 10]. △

On the scale model used earlier in this chapter, in which the Earth has a diameter of about 0.05 cm, the dimensions of our Galaxy would be approximately 60 million km. It is obvious that this scale cannot be used to illustrate the vast distances of the other galaxies in the universe. Another scale must therefore be set up to conceptualize galactic distances.

Imagine the Earth's orbit to be the size of the orbit of the electron in a hydrogen atom. ▽ In the simplest picture of this lightest atom, hydrogen is composed of a central proton, which has a positive electric charge, about which orbits an electron having a negative electric charge. The opposite sign of these charges—one positive, and the other negative—provides the electrical force holding the hydrogen atom together, since oppositely charged particles attract one another. △ The radius of this electron orbit is 0.53×10^{-8} cm. The nearest star would then be approximately 0.014 mm from the nucleus of the atom, the center of the Galaxy about 10 cm, and the diameter of our stellar system about 35 cm. The diameter of our Sun would be submicroscopic—about 4.6×10^{-11} cm.

It has already been stated that the stars are vast distances apart; for all practical purposes, they are isolated from each other. Hence, they almost never collide, although the motions of each of them is determined by the total gravitation of all the stars in the Galaxy. If we consider the Milky Way as a closed region filled with gas, with individual stars playing the role of molecules, this gas would be found to be extraordinarily rarefied. The average distance between the stars is ▽ 10^{19} cm. The diameter of the Sun is about 10^{11} cm. Thus, the average relative distance between the stars is 10^{19} cm/10^{11} cm = 10^{8}, or △ almost 100 million times greater than the average diameter of the stars. Under ordinary conditions the average distance between molecules of air is only several tens of times greater than the dimensions of the molecules. The air would have to be made at least 10^{19} times less dense in order to attain the same relative degree of rarity as the stars in our Galaxy. ▽ A gas so rarefied would have only about one atom in every cubic centimeter. This is accidentally the average density of matter in interstellar space. So, by an interesting coincidence, the distances between the stars in interstellar space, relative to their diameters, are just about the same as the distances between the atoms and molecules in interstellar space, relative to *their* diameters.

Interstellar space is as empty as a cubical building, 60 miles long, 60 miles wide, and 60 miles high, containing a single grain of sand. △

However, in the center region of the Galaxy, where the star density is relatively greater, collisions do occur from time to time, one collision every million years or so. During the history of our Galaxy, which is thought to be at least 10 billion years old, a collision between stars has almost never occurred in normal regions of the Galaxy [see Chapter 12].

For several decades, astronomers have been studying other galaxies which resemble our own in varying degrees. This field of investigation is called "extragalactic astronomy." During the past two decades, great strides have been made in understanding the configuration of the metagalaxy (the system of galaxies external to our own). Although the general structure of the metagalaxy is becoming increasingly clear, many questions still remain unanswered. The vast distances separating us from these galaxies create problems which can be solved only by employing more powerful instruments of observation combined with more intensive theoretical investigation.

The galaxies closest to us are the Magellanic Clouds, so named because the explorer Magellan sighted them on his famous voyage around the globe. They are seen clearly in the evening sky of the Southern hemisphere as two large patches of light with almost the same surface brightness as the Milky Way. The distance to the Magellanic Clouds is only about 200 thousand light years—about two diameters of our Galaxy away. Another nearby galaxy is the Great Nebula in the constellation Andromeda [Fig. 3–5]. It is seen by the naked eye as a faint, luminous speck of the fifth magnitude.[2] Actually, this vast stellar system is almost three times larger than our own Galaxy, both in numbers or stars and in total mass. The Andromeda galaxy, called M31 by astronomers because it is number 31 in the catalogue of the eighteenth-century French astronomer Messier, is about 1.8 million light years away, or almost twenty times the diameter of our Galaxy. The Great Nebula has a clearly defined spiral structure and characteristics similar to those of our Galaxy. A small satellite galaxy with an elliptical form can be seen to one side of M31.

▽ In Figure 3–6 is seen a detailed photograph of the region of the galactic nucleus of M31. The bright white dots are foreground stars in our own Galaxy. The photograph does not resolve the individual stars in the nucleus of M31. Several dark lanes of gas and dust may be seen. In Figure 3–7, however, in a photograph of the periphery of M31, resolution into individual stars is accomplished. This is also true in Figure 3–8, a photograph of the companion galaxy NGC 205 [cf. Fig. 3–5]. Three other fairly typical spiral galaxies—one seen edge-on, the others approximately face-on—are presented in Figures 3–9, 3–10, and 3–11. These

[2] The amount of radiation from the stars is measured by stellar magnitude. The smaller the magnitude, the brighter the star. If a star is one magnitude *less* than another star, it is 2.512 times *brighter*. A difference of five magnitudes corresponds to a brightness ratio of 100. Stars weaker than the sixth magnitude cannot be seen with the naked eye. The brightest stars have negative magnitudes; e.g., Sirius has a magnitude of −1.6.

FIGURE 3–5. *The nearest spiral galaxy, M31, the Great Nebula in the constellation Andromeda. The numerous bright points which cover the picture are foreground stars in the solar neighborhood, within our own Galaxy. Also shown are two smaller companion galaxies, NGC 205, further from M31, and NGC 221, closer to M31. The dark dust lanes in M31 are readily visible. (Courtesy of Mt. Wilson and Palomar Observatories.)*

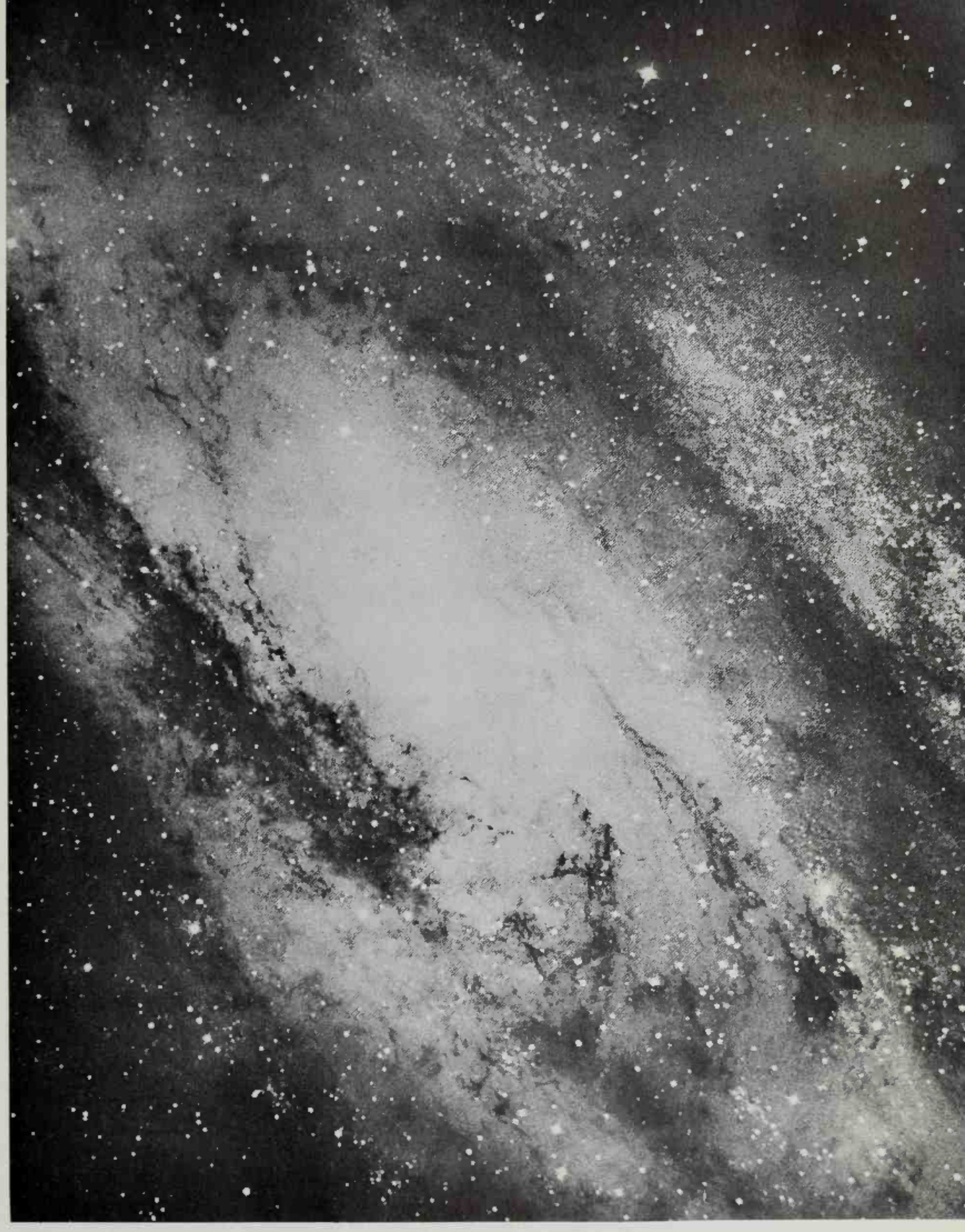


FIGURE 3–6. *Region of the galactic nucleus of the great galaxy M31. (Courtesy of Mt. Wilson and Palomar Observatories.)*

FIGURE 3–7. *The periphery of the great galaxy M31, showing resolution into individual stars. (Courtesy of Mt. Wilson and Palomar Observatories.)*

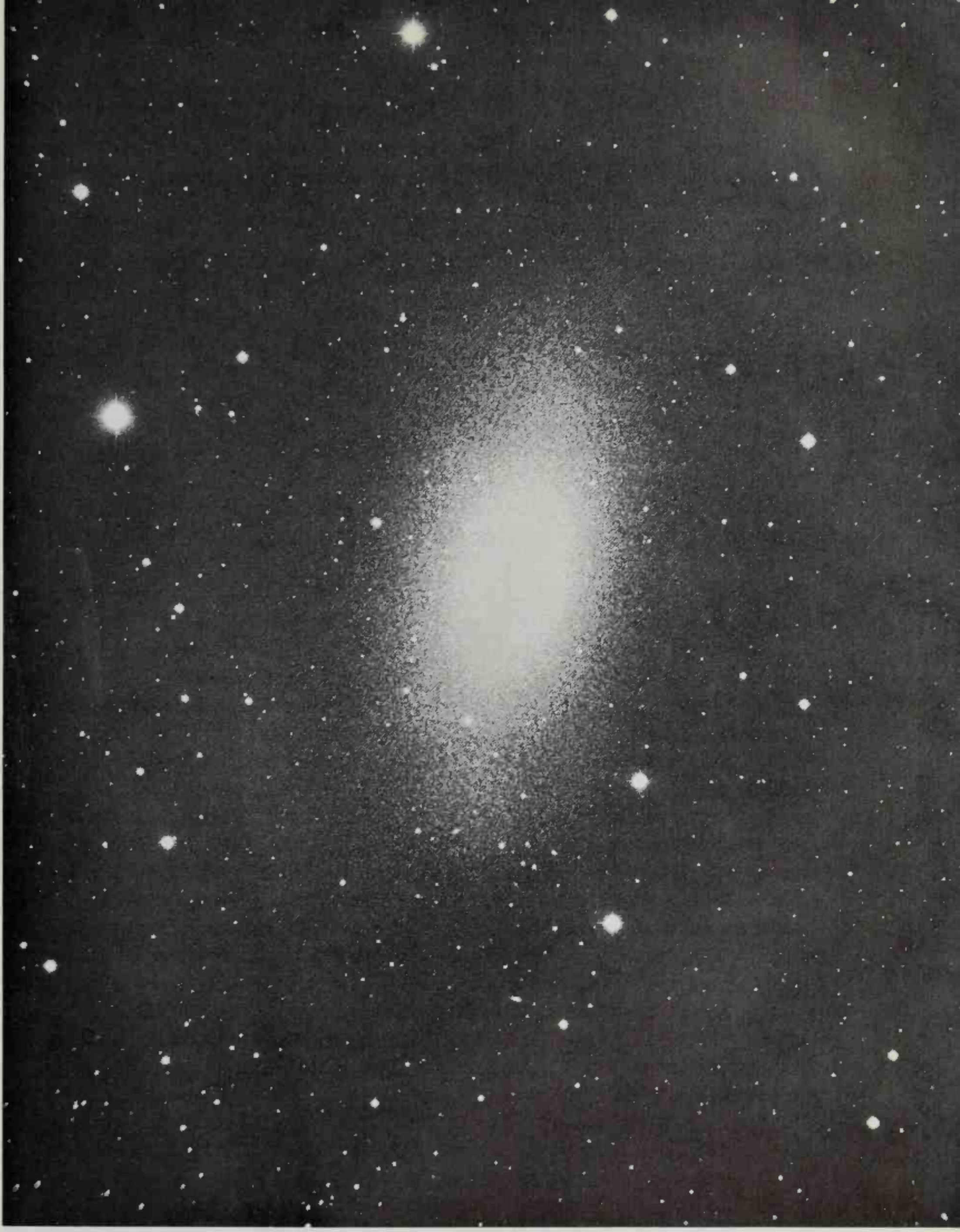


FIGURE 3–8. *The elliptical galaxy NGC 205, a companion to the great galaxy in Andromeda, M31. Note the resolution into individual stars in the outer portions of this galaxy. (Courtesy of Mt. Wilson and Palomar Observatories.)*

FIGURE 3–9. *The spiral galaxy M104 in the constellation Virgo. This galaxy is oriented so that we see it edge-on. Note the prominent dust lanes in the disk, and the luminous galactic nucleus. The object with diffraction spikes in the lower right is a foreground star in our own Galaxy. (Courtesy of Mt. Wilson and Palomar Observatories.)*

galaxies are too distant for us to resolve their individual stars. The bright nodules seen in such galaxies are either globular clusters or extensive glowing regions of hydrogen gas. Note the dark dust lanes in these photographs. △

The galaxies differ greatly from each other both in size and in shape. Besides the spiral systems, subdivided into groups a, b, and c, according to the development of the spiral structure, there are "elliptical" galaxies—for example, the small satellite galaxy of M31 mentioned above—and irregular galaxies such as the Magellanic Clouds.

A great many of these stellar systems can be observed with large telescopes. Only about 250 of them are brighter than the twelfth magnitude. There are, however, at least 50,000 galaxies brighter than the fifteenth magnitude. The faintest systems which can be photographed with the 200-inch reflecting telescope at Mt. Palomar have a magnitude of +24. Stellar systems of this magnitude may be billions of light years away. The light producing the image of a very remote galaxy on the photographic plate may have left the galaxy when the first cells were forming in the primitive oceans of the Earth.

▽ Even the light from our nearest neighboring spiral galaxy in Andromeda

FIGURE 3–10. *NGC 1300, a barred spiral galaxy in the constellation Eridanus, photographed at the McDonald Observatory. (Courtesy of Yerkes Observatory, University of Chicago.)*

(M31), which formed the image in Figure 3–6, left M31 during the Pleistocene epoch, when Megatherium, an eight-foot-tall ground sloth, roamed what is now the southwestern United States and when, somewhere east of the Atlantic, the first tool-using hominids were gradually emerging through slow evolutionary processes. The entire subject of extragalactic astronomy is based on light which left these remote galaxies in prehistoric times when man had not yet trod the Earth.

▽ Astronomers are not restricted to observations at visible wavelengths, where the eye is sensitive. Many other wavelengths exist, undetected by the unaided eye. Every object at any temperature above absolute zero (−273°C or −456°F) radiates at all the wavelengths of the electromagnetic spectrum. This is called "thermal" emission, because it depends on no special emission mechanism—simply the heat of the object.

▽ In many respects, light has wave properties. The distance from crest to crest in light waves—as in water waves—is called the wavelength. The number of crests which pass a fixed point in a given time (say, one second) is the frequency of the wave, and can be measured in crests or cycles per second (cps). With a little thought we can convince ourselves that the wavelength, λ, and the frequency, ν, are related to the velocity of light, c, by $\lambda\nu = c$.

FIGURE 3-11. *The spiral galaxy NGC 7331 in the constellation Pegasus. Other galaxies may also be seen in this photograph. (Courtesy of Mt. Wilson and Palomar Observatories.)*

▽ Despite its general wave nature, light must also be considered composed of discrete packets of energy called quanta, or photons. The energy of a photon is proportional to its frequency, ν. Thus, the higher frequency (shorter wavelength) photons are more energetic, and penetrate further into matter.

▽ The visible spectrum lies at wavelengths between 4×10^{-5} and 7×10^{-5} cm. Another unit of length used in discussions of light is the Ångstrom (abbreviated Å): $1\ \text{Å} = 10^{-8}$ cm. Thus, visible light lies between 4×10^{-5} cm $\times 10^{8}$ Å/cm = 4000 Å and 7×10^{-5} cm $\times 10^{8}$ Å/cm = 7000 Å, corresponding to deep violet and deep red, respectively. Wavelengths below 4000 Å are in the ultraviolet region of the spectrum; below about 100 Å, in the x-ray region; and below about 1 Å, in the gamma-ray region. Since the shorter wavelengths penetrate deeper into matter, x-rays are used in medical diagnosis. Wavelengths longer than 7000 Å are in the infrared region. When matter absorbs infrared light, its constituent atoms are induced to vibrate, a phenomenon which we call heat in a solid. For this reason, infrared radiation is also called heat radiation. Its wavelength is commonly expressed in units of microns (μ): $1\mu = 10^{-4}$ cm $= 10^{4}$ Å. Wavelengths longer than 1 mm are in the radio region of the spectrum. △

Among the galaxies there are some which radiate excessively large amounts of energy at radio frequencies. These are called "radio galaxies." The flux of radio radiation may many times exceed the flux of visible radiation. The classic example of such a galaxy is Cygnus A. In terms of visible radiation, this galaxy appears as two insignificant specks of light of the seventeenth magnitude [Fig. 3–12]. Actually, however, the absolute output of visible light is very great—about ten times that of our own Galaxy. It appears faint to us because it is some 600 million light years distant. The flux of radio radiation reaching us from Cygnus A in meter wavelengths is so great that it exceeds the flux of radio radiation from our Sun during a period of low sun-spot activity. But then, the Sun is very close—only 8 light minutes—as compared to 600 million light years for Cygnus A. The flux of the radiation is inversely proportional to the square of the distance.

In visible light, the spectra of the majority of the galaxies are similar to the spectrum of the Sun. ▽ Such a spectrum can be seen when sunlight, or any other kind of light, passes through a triangular glass prism. The different colors constituting sunlight travel at different speeds through the glass and are therefore bent different amounts in passing through it. The white light of the Sun is then spread out on the other side of the prism into a broad, rainbow-hued pattern of color. This spectrum can be either viewed with the naked eye or photographed. Figure 3–13 shows some examples of stellar spectra. Discrete dark absorption lines can be seen superposed on the background. Each line, at its own characteristic frequency, can be considered due to the absorption of sunlight (arising from deeper levels in the star) by atoms at cooler, higher parts of the stellar atmosphere.

▽ Each chemical element has its own set of frequencies where it characteristically absorbs radiation. Strong absorption lines in stellar spectra are often caused by such elements as hydrogen, helium, sodium, calcium, and potassium. The darkest of these absorption lines do not necessarily correspond to the most abundant elements, since some atoms have stronger absorption lines than others. Hydrogen and helium are principal constituents of almost all stars, while sodium, calcium, and potassium are present in relatively minor quantities. But the lines of calcium can nevertheless be stronger than the lines of helium. Comparison with laboratory measurements can allow for this effect. Therefore, the chemical composition of distant stars can be determined from the light which they radiate to us and the interpretations provided by modern physics.

▽ The spectra of galaxies also show dark absorption features, an understandable circumstance, since the visible radiation of these galaxies consists of the cumulative radiation of billions of stars more or less similar to our Sun and to the stars in the solar neighborhood. Information can thus be garnered about the composite chemical make-up of galaxies millions of light years away, an extraordinary achievement of the human mind. △

By measuring the displacement of the wavelengths in the spectrum of a light source relative to a laboratory standard, using the Doppler effect mentioned earlier, we can determine whether the light source is approaching us or moving away from us. If the light source is approaching, the wavelengths are decreased, and the spectral lines are displaced toward the blue end of the spectrum. If the light source

FIGURE 3–12. *In the center of this photograph is the Cygnus A radio source as it appears at optical frequencies. At radio frequencies, it is one of the brightest objects in the heavens. At visual frequencies, except for its double nature, it is a very unextraordinary object.* (*Courtesy of Mt. Wilson and Palomar Observatories.*)

is receding, the wavelengths are increased, and the lines shift toward the red end of the spectrum.

A very important discovery was made about the spectra of galaxies several decades ago ▽ by the American astronomer V. M. Slipher, of Lowell Observatory. △ The spectral lines of all the galaxies—except those very close to us—undergo a shift towards the red end of the spectrum. This phenomenon is called the "red shift," and as later found by the American astronomer Edwin Hubble of Mt. Wilson Observatory this shift increases with the increasing distances of the galaxies. The simplest explanation is that all galaxies are receding from us, and that the velocity of this "expansion" increases with the distance. The greater the distance, the faster

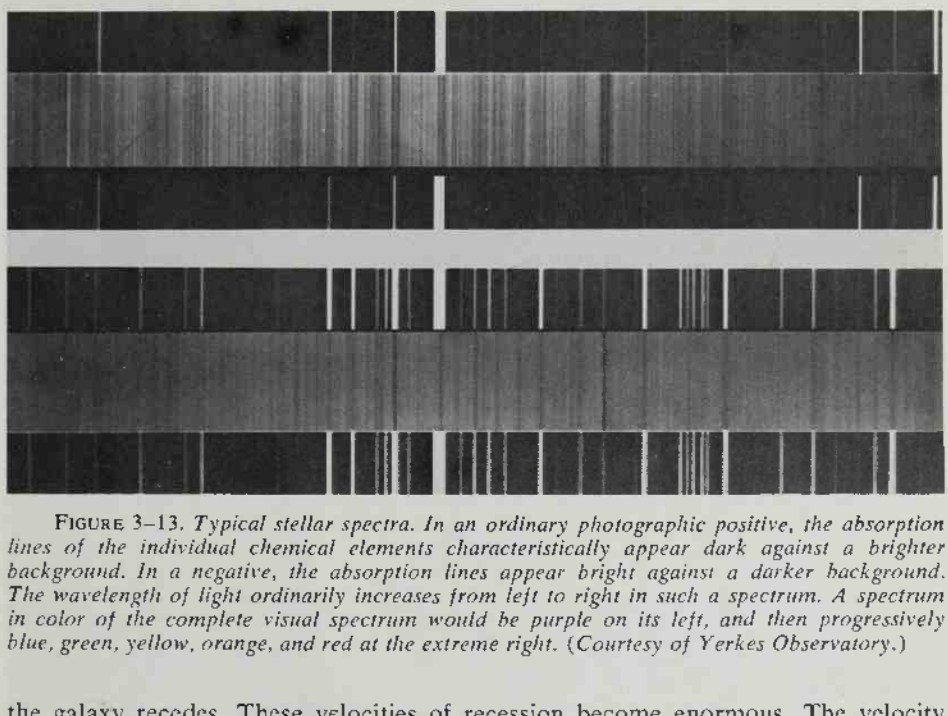


FIGURE 3–13. *Typical stellar spectra. In an ordinary photographic positive, the absorption lines of the individual chemical elements characteristically appear dark against a brighter background. In a negative, the absorption lines appear bright against a darker background. The wavelength of light ordinarily increases from left to right in such a spectrum. A spectrum in color of the complete visual spectrum would be purple on its left, and then progressively blue, green, yellow, orange, and red at the extreme right. (Courtesy of Yerkes Observatory.)*

the galaxy recedes. These velocities of recession become enormous. The velocity of the receding radio galaxy Cygnus A is almost 16,000 km/sec. A very weak radio galaxy, 3C–295, has a very great velocity of recession. Visually, it is of the twentieth magnitude. In 1961, its spectrum was obtained [see Fig. 3–14], and it appears that the ultraviolet spectral lines produced by (ionized) oxygen are displaced to the orange region of the spectrum. Therefore, by simple calculation, we find that the velocity of recession is 138,000 km/sec, or almost half the speed of light. This radio galaxy is five billion light years from us. ▽ More recently, even more distant objects have been detected by similar techniques. △ Astronomers are now investigating light which started its long journey through space when the Sun and the planets were forming.

FIGURE 3–14. *At the top is a photograph of the object 3C 295 in the constellation Bootes. It is some 5 billion light years away, and a source of cosmic radio noise. The lower half of the picture shows negatives of three sets of spectra. The upper and lower spectra are obtained in the laboratory to establish wave-length standards. The middle spectrum is taken of 3C 295 through the terrestrial atmosphere. Most of the lines in the middle spectrum are due to atoms and molecules in the terrestrial atmosphere. The arrow, however, points to a spectral feature which is not present in spectra of the sky, and which is due to 3C 295. This is a line which has been very greatly red-shifted, and provides the best grounds for attributing immense distances to 3C 395. Because of the faintness of such features, great care must be taken in obtaining and interpreting such spectra.* (*Courtesy of Mt. Wilson and Palomar Observatories.*)

FIGURE 3–15. *A cluster of galaxies in the constellation Hercules. Spiral, elliptical, and irregular galaxies may be seen in various inclinations to our line of sight. Some of the galaxies are connected by luminescent bridges. The spiked objects, and many of the small, perfectly circular objects, are foreground stars in our own Galaxy. (Courtesy of Mt. Wilson and Palomar Observatories.)*

In addition to the over-all expansion of the universe, individual galaxies have their own irregular, disordered motions. These velocities are usually several hundred kilometers per second. Since the velocity of expansion increases by about 100 km/sec for each million parsecs, the irregular velocities exceed the velocity of recession for those galaxies which are within a million parsecs of the Milky Way. Therefore, their red shift cannot be detected. Some of the nearest galaxies are in fact approaching us.

The galaxies are not distributed uniformly in metagalactic space; they form separate groups and clusters of galaxies. Galaxies are found tens of times less often outside such clusters than inside them. A group of seventeen galaxies, including our own, make up the so-called "local group." The local group, in turn, is part of a larger cluster, the center of which is in that part of the sky where the constellation Virgo is located. This large cluster is thought to contain several thousand galaxies. ▽ Figure 3–15 shows a cluster of galaxies in the constellation Hercules. Elliptical, spiral, and irregular galaxies may all be seen. Some clusters have as many as a thousand galaxies. △

Let us now consider the differences between clusters of stars and clusters of galaxies. The number of stars in a stellar cluster is much greater than the number of galaxies in a cluster of galaxies. The distances between the stars in a cluster are very large compared with the size of the stars themselves. The mean distances between the galaxies in a cluster of galaxies are only several times greater than the dimensions of the galaxies. If we think of all the galaxies in the sky as a gas in which the individual galaxies are represented by molecules, this medium would be very viscous and unlike an ordinary gas. (In many problems of metagalactic astronomy, it is often useful to treat the metagalactic medium as a continuum, possessing characteristics such as viscosity, electrical conductivity, etc.)

Now, let us turn back to our second model, in which the Earth's orbit is reduced to the dimensions of the first electron orbit of the hydrogen atom. On this scale, the distance to the Andromeda galaxy would be somewhat greater than 6 meters, the distance to the central part of the local system of galaxies in the constellation Virgo, about 120 meters, the distance to the radio galaxy Cygnus A, 2.5 km, and the distance to the radio galaxy 3C–295, about 25 km.

With this background, we have some conception of the scale and general structural characteristics of the universe as it exists today. The old picture of a static universe must be replaced by one of a dynamic universe filled with evolving cosmic objects. Everything is changing; stars appear, grow, and die; clouds of gas and dust and galaxies develop; everything is in motion. ▽ The universe is filled with planets, stars, galaxies, which, on the immense cosmic time scale, are evanescent, ephemeral entities, forming, flickering briefly, and then fading, lost forever in the infinite recesses of space and time. △

4

Fundamental properties of the stars

I must be of the same Opinion with all the greatest Philosophers of our Age, that the Sun is of the same Nature with the fix'd Stars.

. . . May not every One of these Stars or Suns have as great a Retinue as our Sun, of Planets, with their Moons, to wait upon them? Nay, there's a manifest Reason why they should. For if we imagine ourselves placed at an equal Distance from the Sun and fix'd Stars we should then perceive no difference between them. For, as for all the Planets that we now see attend the Sun, we should not have the least glimpse of them, either because their Light would be too weak to affect us, or that all the Orbs in which they move would make up one lucid point with the Sun.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

With the invaluable aid of the spectroscope, astronomers have amassed a great deal of information about such fundamental properties of the stars as their masses, radii, chemical composition, the total amount of energy each radiates per second (called the luminosity, and designated by the letter L), and the temperature of their surface layers.

A star's surface temperature determines both its color and its spectrum; at 3,000–4,000 degrees Kelvin,[1] the color is reddish; at 6,000–7,000°K, yellowish; very hot stars—at temperatures above 10,000–12,000°K—are blue or white.

▽ The relation between color and temperature in stellar atmospheres is very similar to that observed with everyday materials. For example, as a bar of iron is heated, its color passes progressively from red, to yellow, until, finally, it becomes white hot. Similarly, the color of the star can be used to determine the temperature of its atmosphere. △

A standard non-visual method for measuring the color of a star is based on the determination of its "color index," which is equal to the difference between the photographic magnitude and the apparent (visual) magnitude. The usual photographic plate is especially sensitive to blue light; the eye, to yellow and green light. ▽ (In fact, the wavelength of maximum sensitivity of the eye is very close to the wavelength at which the Sun emits the most energy. This is, of course, no coincidence: Human eyes have evolved in landscapes bathed with solar radiation, and there is clearly an evolutionary advantage in high sensitivity to low levels of light. If there are organisms with sight on other planets, we may expect a similar connection between ocular sensitivity and the color of the local sun.) △ Thus, the photographic and the visual magnitudes are not identical. The visual magnitude is determined by means of photographic plates which have a sensitivity resembling that of the eye. The color index of red stars can reach +1.5 stellar magnitudes and even higher; with blue stars, the index is negative. A specific type of spectrum accompanies each star's color index. The relatively cold red stars have spectra with absorption features characteristic of the neutral atoms of metals and of certain very simple compounds, for example, CN, CH, and others.

▽ A neutral atom is one which has no net electrical charge. In the hydrogen atom, discussed previously, the positive electrical charge of the nuclear proton is exactly balanced by the negative electrical charge of the electron, so that, from a distance, the atom is electrically neutral. Similarly, the next atom in order of complexity, helium, is electrically neutral because its nucleus contains two positive

[1] For stellar temperatures, the difference between Kelvin degrees (K°) and Centigrade degrees (C°) can be neglected in an introductory discussion. Those readers who like their degrees Fahrenheit can multiply the number of K° by 1.8 to get F°.

protons, and two neutral neutrons, while the nucleus is surrounded by two electrons. If somehow—for example, by the collision of another atom, or by the absorption of light—helium is made to lose an electron, it is said to be "ionized." It would then have a net positive charge, since there would be more positively charged protons in the nucleus than negatively charged electrons outside the nucleus. The absorption properties of ionized atoms are different from the absorption properties of the same atoms when neutral. As the temperature of a star increases, the relative number of ionized atoms of a given sort will also increase, because of a rise in both the number of collisions and the number of absorptions. In addition, as the temperature increases, simple chemical compounds such as CN are dissociated, torn apart by collisions with their more rapidly moving neighbors, and by bombardment by high energy photons. △

As the surface temperature increases, the molecular spectral absorption features vanish, many lines of the neutral atoms diminish in intensity, and the lines of ionized helium appear. Stars with a surface temperature of approximately 6,000°K have lines of ionized calcium located at the end of the visible and the beginning of the ultraviolet part of the spectrum. Our sun has such a spectrum. Stars with a surface temperature of approximately 10,000°K have intense lines of hydrogen. Very hot stars with a surface temperature exceeding 20,000°K have lines chiefly of neutral or ionized helium, and the continuous spectrum is very intense in the ultraviolet regions.

The spectral sequence of the bulk of the stars (their classification according to their spectra) is denoted by the following, essentially arbitrary, sequence of letters: O, B, A, F, G, K, M. ▽ The time-honored mnemonic device for remembering this sequence of letters is the immortal phrase, "*O*h *B*e *A* *F*ine *G*irl, *K*iss *M*e." △ Each letter indicates a spectral class, O being the hottest star and M being the coldest. The measurements are so sensitive that it is possible to divide each class into 10 subclasses, for example, B1, B2, B3, and so forth. A star with a spectrum B9 is nearer the spectral class A1 than B1.

▽ Figures 4–1 to 4–4 illustrate this change in the spectral properties with spectral class. Each horizontal strip corresponds to the spectrum of an individual star in our Galaxy. Each star shown belongs to the "main sequence," a category to which most stars belong. The spectra are taken from the Morgan-Keenan-Kellman atlas, a collection of spectra named after the compilers, American astronomers then at Yerkes Observatory, University of Chicago. The names of the stars are shown on the left in each figure, and their spectral types arranged sequentially on the right. Each spectrum is a "negative"; that is, for purposes of presentation, the absorption lines are shown as bright features on a dark background, instead of as dark lines on a bright background, their usual observed configuration. At the top and bottom, individual spectral lines are identified according to laboratory comparisons. The atom causing the absorption, its state of ionization, and the wavelength at which it absorbs are shown. For example, He I 4009 refers to neutral helium absorption at a wavelength of 4009 Å; the background would appear blue in the vicinity of this line, if the photographs of Figures 4–1 to 4–4

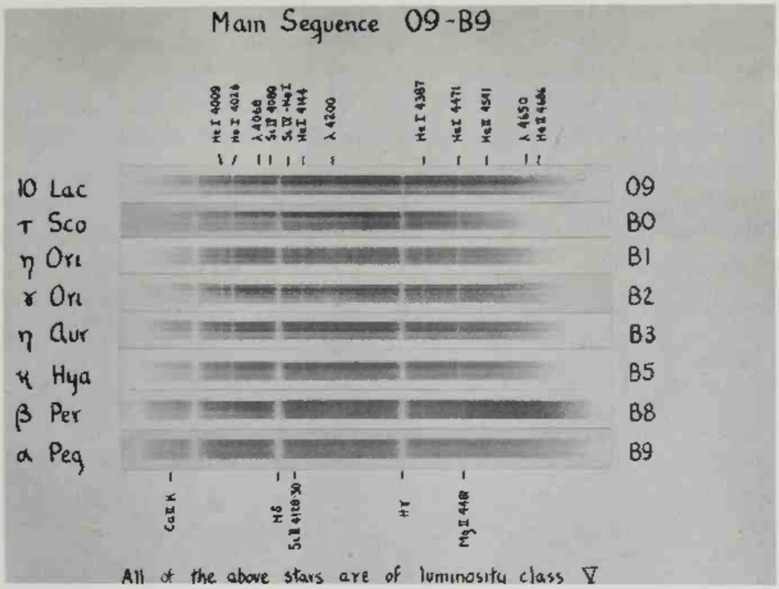


FIGURE 4–1. *Representative spectra of late O and early and late B stars of the main sequence. These spectra are reproduced from the Morgan-Keenan-Kellman Atlas. (Courtesy of Yerkes Observatory.)*

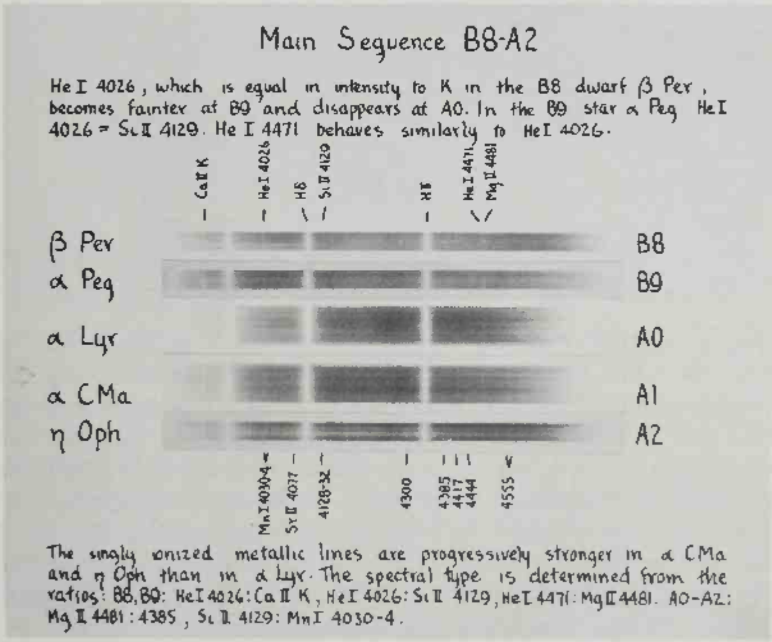


FIGURE 4–2. *Representative spectra of late B and early A stars of the main sequence. These spectra are reproduced from the Morgan-Keenan-Kellman Atlas. (Courtesy of Yerkes Observatory.)*

were in color. Helium II indicates singly ionized helium, that is, helium in which one electron has been lost. Si IV indicates silicon atoms from which three electrons have been lost. The names of the individual stars arrayed along the left-hand border of these figures also show a variety of nomenclature systems, the patrimony of a hoary astronomical classification convention in which each of many different workers compiles his own catalogue. △

▽ It had originally been intended by those who invented this stellar spectral type classification that the sequence of spectral types could be represented by letters in alphabetical sequence—A, B, C, D, etc. However, after the original assignment of letters to spectral types had been made, it was discovered that through errors in

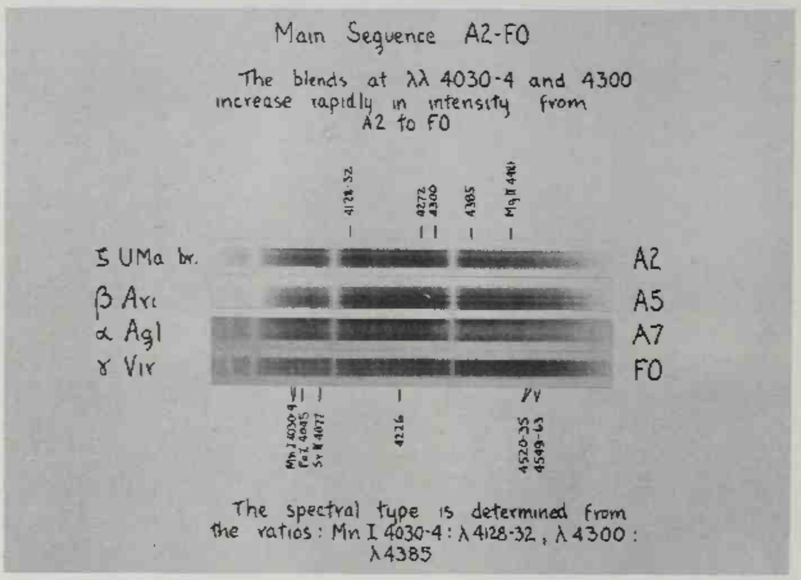


FIGURE 4–3. *Representative spectra of early and late A and early F stars of the main sequence. These spectra are reproduced from the Morgan-Keenan-Kellman Atlas. (Courtesy of Yerkes Observatory.)*

classification some letters had been assigned to nonexistent or insignificant spectral types while others had been assigned so that there was not a smooth transition between adjacent spectral types. A star of late O spectral type has a spectrum quite similar to a star of early B spectral type (see Figure 4–1). Thus, O had to be placed before B, B before A, and the entire astronomical alphabet almost randomly reassorted. It is an interesting commentary on human conservatism that this clerical error has been enshrined by repeated use, until there is now no hope of introducing another system. However, since the sequence of letters in the alphabet is also arbitrary, the astronomical nomenclature for spectral types is fundamentally not much more obscure than a true alphabetical system. △

The luminosity differs greatly from star to star and is often expressed in terms of the luminosity of the Sun. Our Sun has a luminosity of 4×10^{33} erg/sec.

▽ The basic metric system unit of mass is the gram. One gram equals approximately 0.035 ounce. An erg is the unit of energy expended in lifting 1 gm a distance of 10^{-3} cm on the Earth, obviously a very small quantity. It is very much in the range of interest of the flea. The output of a 100-watt bulb is 10^9 erg/sec. Thus, the Sun's output is the equivalent of four trillion trillion light bulbs, each putting out 100 watts. △

The vast majority of stars are "dwarfs" that are significantly less luminous than our Sun—as much as a thousand times less luminous. However the

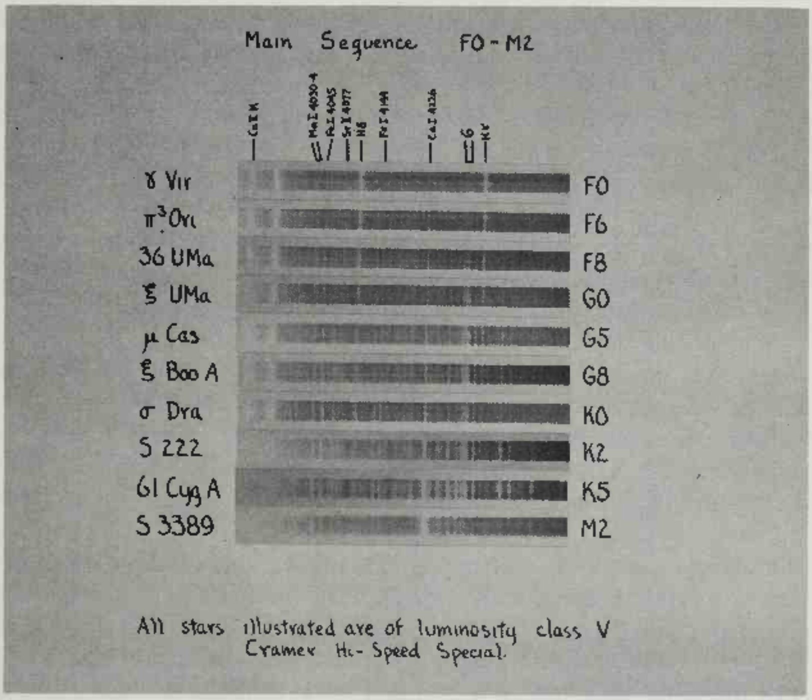


FIGURE 4–4. *Representative spectra of F, G, K, and early M stars of the main sequence. These spectra are reproduced from the Morgan-Keenan-Kellman Atlas. (Courtesy of Yerkes Observatory.)*

"supergiant" stars (relatively few in number) have luminosities which are from 10^4 to 10^5 times greater than that of our Sun.

▽ It is common, in astronomy, to refer to certain broad categories of stars both by their relative sizes and by their colors. The astronomical zoo is replete with "supergiants," "giants," "dwarfs," and "sub-dwarfs," but no individuals of ordinary stature, and a simple statement of solar evolution often sounds like an excursion into the world of the brothers Grimm. A typical star begins life auspiciously, as a bright yellow giant, and then metamorphoses, in early adolescence, into a yellow dwarf. After spending most of its life in this state, the yellow dwarf rapidly expands into a luminous red giant, jumps the Hertzsprung gap, and decays violently into a hot white dwarf. It ends its life, cooling inexorably, as a degenerate black dwarf. Few readers will recall the original title of this moderately depressing life history, but many will find it vaguely familiar. To understand the underlying causes of the varied careers of the stars, we must discuss further astronomical observations and their interpretations. △

The apparent magnitude of a star is a measure of its apparent brightness; that is, its brightness as it appears to us. The apparent magnitude is therefore dependent on both a star's intrinsic brightness and on its distance from us. Even a very brilliant star will appear inconspicuous if it is very distant. The ordinary bright stars, visible to the naked eye on an average night, have apparent magnitudes mostly between 1 and 4. (A star of the first magnitude is brighter than one of fourth magnitude.) △ Very bright stars have negative apparent magnitudes. Most stars have small positive magnitudes. The apparent magnitude of the Sun—much brighter than the stars, of course—is -26.8. However, if we moved the Sun to a distance of 10 parsecs (approximately 2 million times further than its actual distance), its apparent magnitude would be $+5$, and it would appear as a tiny point of light in the sky, barely visible to the naked eye. The faintest star which can be seen with the naked eye has a magnitude of $+6$.

If we take any star at the standard distance of 10 parsecs from the solar system, its magnitude is called "absolute." Stars of high intrinsic luminosity have negative absolute magnitudes—for example, -7 or -5; stars of low intrinsic luminosity have large positive magnitudes—for example, $+10$, $+12$, etc.

Stellar mass, in contrast to luminosity, varies within relatively narrow limits from star to star. The mass of our Sun is 2×10^{33} gm—more than 330,000 times the mass of the Earth. Few stars have masses as much as ten times greater or smaller than our Sun's.

Radii differ greatly from star to star. The dimensions of white dwarfs do not exceed those of the Earth. White dwarfs have an enormous average density, ranging to 10^4 and 10^5 gm/cm^3. ▽ By comparison, the density of water is only 1 gm/cm^3, and the density of an average rock is about 3 gm/cm^3. △ Other stars have such huge diameters that the orbit of Mars could be placed comfortably inside them. Such immense stars are sometimes called "bubbles." Since there is comparatively little variation in the masses of the stars, a star with a large radius will have a low average density. The density of the Sun is about 1.4 gm/cm^3, or slightly denser than water. In contrast, stellar "bubbles" are millions of times less dense than air.

Investigations during the last three decades have indicated that stars rotate about their axes. It is now clear that stars of different spectral classes rotate with different velocities. Chapter 13 will be devoted to this very important cosmogonical question.

Spectral analyses indicate that the chemical composition varies from star to star. The hot giant stars, concentrated in the Galactic plane, are relatively rich in heavy elements, for example, iron or silicon, while the stars in globular clusters [see

Fig. 3–4], quite far removed from the plane, have a heavy element content which is ten times smaller. This important fact is one starting point for contemporary theories on the evolution of stars and stellar systems.

The principal constituents of the stars are usually hydrogen and helium plasmas—ionized gas which is electrically neutral because the number of positive charges on the ions (e.g., He II) is just balanced by the number of free negative electrons, which are not bound by electrical forces to any atom. The remaining elements are present in the form of relatively insignificant impurities. ▽ The average

TABLE I. COSMIC ABUNDANCE OF THE ELEMENTS

Atom	*Relative atomic weight*	*Relative cosmic abundance, atoms*
Hydrogen	1.0	10,000,000.
Helium	4.0	1,400,000.
Lithium	6.9	0.003
Carbon	12.0	3,000.
Nitrogen	14.0	910.
Oxygen	16.0	6,800.
Neon	20.2	2,800.
Sodium	23.0	17.
Magnesium	24.3	290.
Aluminum	27.0	19.
Phosphorus	31.0	3.
Potassium	39.1	0.8
Argon	40.0	42.
Calcium	40.1	17.
Iron	55.8	80.

relative chemical composition of the outer layers of the stars is given in the accompanying table, which shows the relative abundance of other elements for every 10 million atoms of hydrogen. Also given is the weight of one atom of the elements listed, relative to hydrogen. The mass of a hydrogen atom is 1.66×10^{-24} gm. While the most abundant elements by far are hydrogen and helium, there is no obvious systematic dependence of abundance on atomic weight for the atoms displayed here. These abundance questions will arise again later, when we discuss the chemical reactions leading to the origin of life on Earth.

▽ Comparison of Table I with Figures 4–1 to 4–4 bears out the point that the most abundant atoms are not necessarily those most easily discerned by spectroscopy. △

Although the so-called heavy elements (those heavier than helium) are underabundant compared with hydrogen and helium, they play a very important role in the universe. The luminosity of a star depends on its opacity to the radiation generated in its interior. Many of the heavy elements tend to be quite opaque, so that small quantities of them may significantly influence the character of the light emitted from the stellar interior, and the subsequent evolution of the star.

The heavy elements have a decisive meaning for life in the universe. The role of carbon in the structure of living material is well known. Equally important for life on earth are nitrogen, oxygen, and phosphorus; and, for many life forms, iron, magnesium, sulfur, potassium, etc. Life is based upon the complex linkages of such

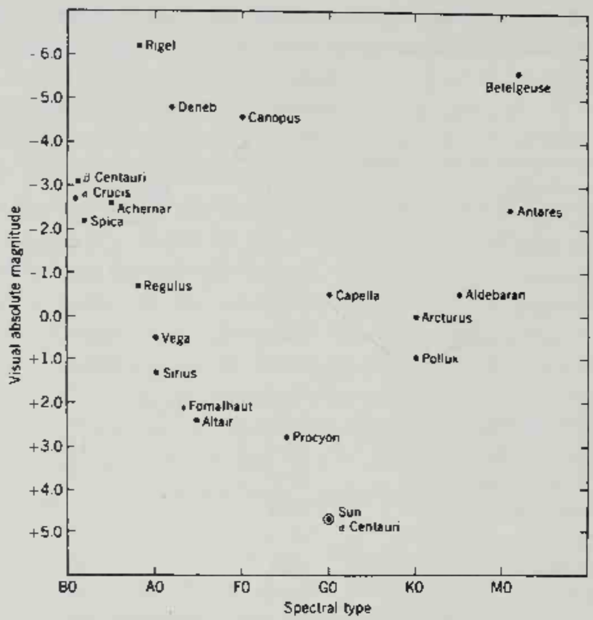


FIGURE 4–5. *The Hertzsprung-Russell diagram of the brightest stars in the sky taken from Bart J. Bok and Priscilla Bok,* The Milky Way, *third edition, Harvard University Press, Cambridge, Massachusetts, 1957. (Courtesy of Harvard University Press.)*

atoms. Thus, if there were no elements heavier than helium in the universe, there would be no life. Does this mean that stars which have a low content of heavy elements could not have habitable planets? We will consider this question later; here we merely stress the importance of the chemical composition of cosmic objects (stars, nebulae, planets) in assaying the possibilities of life in any particular region of space.

The following question can also be asked: Were the heavy elements always present? And if not, how were they formed? There is some evidence that in the distant past there were significantly smaller amounts of the heavier elements than there are today. Perhaps there were none at all, and the universe consisted solely of hydrogen and helium. The formation of these elements will be discussed in Chapter 8.

By spectroscopic methods astronomers have detected the presence of powerful magnetic fields in the atmospheres of certain stars. The intensity in individual cases can be as great as 10,000 gauss, that is, 20,000 times greater than the surface magnetic field of the Earth, which has a magnetic field strength of about half a gauss. We note that sun spots on our star have magnetic fields reaching an intensity of 3,000 to 4,000 gauss. Magnetic phenomena, as understood in recent years, play an important role in the physical processes which occur in the solar atmosphere. There is some basis for assuming that the same is true in other stellar atmospheres. At first glance it would appear that stellar magnetism is unrelated to the problem of the origin and development of life in the universe. However, the sequence of events which, when taken as a whole, leads to the origin of life is

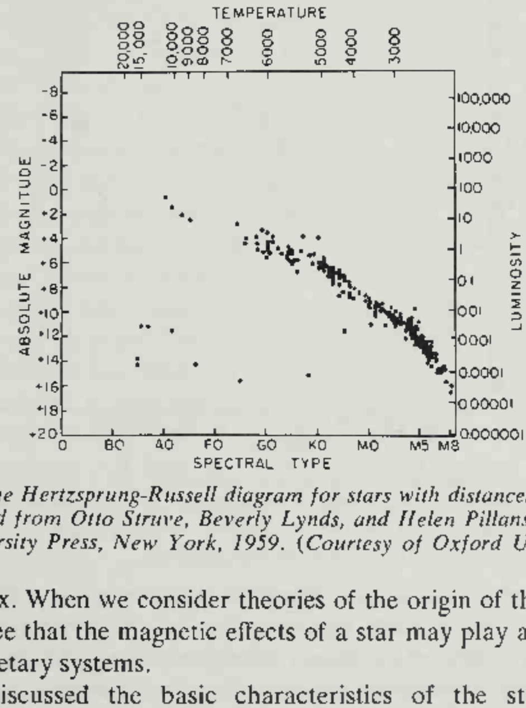


FIGURE 4–6. *The Hertzsprung-Russell diagram for stars with distances less than 10 pc from the Sun. Reproduced from Otto Struve, Beverly Lynds, and Helen Pillans,* Elementary Astronomy, *Oxford University Press, New York, 1959. (Courtesy of Oxford University Press.)*

unusually complex. When we consider theories of the origin of the planets in Chapter 13, we shall see that the magnetic effects of a star may play a critical role in the formation of planetary systems.

We have discussed the basic characteristics of the stars. Is there any connection among these properties? It appears that such a connection does indeed exist. It was independently discovered half a century ago by the Danish astronomer E. Hertzsprung and by the American astronomer Henry Norris Russell of Princeton University.

Figure 4–5 depicts a Hertzsprung–Russell diagram. The points represent the brightest stars in the night sky, the horizontal axis represents the spectral types of the stars, and the vertical axis represents the absolute magnitudes. We see that the majority of the stars are found within the limits of a relatively narrow band, going from the upper left-hand corner of the diagram to the lower right-hand corner. This is the so-called main sequence of the stars. ▽ There must be a fundamental reason that the stars are not strewn, more or less at random, throughout the Hertzsprung–Russell diagram. △ In the upper right-hand section we see some stars in a disordered array. Their spectral classes are G, K, and M, and their absolute magnitudes lie between +2 and −6. They are the "red giants," although there are yellow stars among them. Had we included stars which have smaller apparent luminosities we would have found in the lower left of the diagram a small number of stars with absolute magnitudes less than +10 and spectral classes within the limits B to F. These are very hot stars with low luminosity. But low luminosity with a high surface temperature can occur only when the radius of a star is quite small. Consequently, such stars are called "white dwarfs."

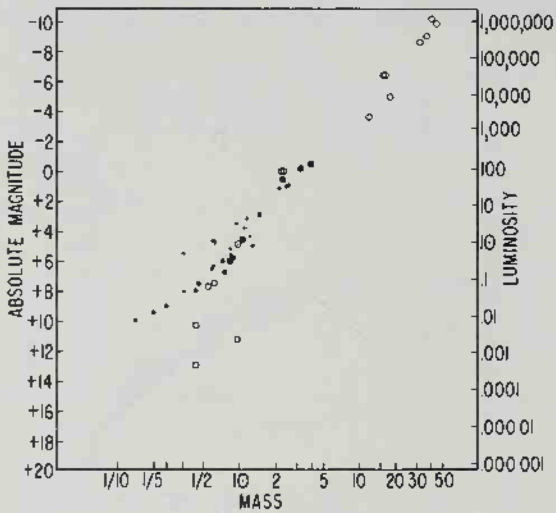


FIGURE 4–7. *The relation between mass and luminosity for a variety of stars. The squares which show deviations from the mass-luminosity relation of the bulk of the stars represent the white dwarfs. From Otto Struve,* Stellar Evolution, *Princeton University Press, Princeton, New Jersey, 1950. (Courtesy of Princeton University Press.)*

The number of points in the "spectrum-luminosity" or Hertzsprung–Russell diagram do not give an accurate representation of the relative number of stars in each spectral class within the Galaxy. The giant stars with high luminosity are represented in disproportionately large numbers because they can be seen from a very great distance. The dwarf stars are difficult to observe and are accordingly less equitably represented. We can obtain a more accurate idea of the relative numbers of stars in each spectral class if we consider only those which are found within 10 parsecs of our Sun (32.6 light years) [see Fig. 4–6]. Here we see that the lower right-hand side of the main sequence is very sharply defined, but there is an absence of giants. Within 10 parsecs, the overwhelming majority of stars are dimmer and colder than our Sun, ▽ a circumstance typical of other parts of the Galaxy as well △; these are the "red dwarfs," which lie in the lower right-hand part of the main sequence. Only eight stars in this diagram (of approximately 170 found within this area) are brighter than the Sun. Eight white dwarfs are represented. Since within a small radius of 10 parsecs we observe so many white dwarfs, we conclude that they are very numerous throughout the universe. Calculations show that there are at least several billion, and perhaps as many as ten billion white dwarfs in our Galaxy. There are approximately 150 billion stars of all types in the Milky Way. The number of white dwarfs is ten thousand times greater than the number of giants of high luminosity, which are represented in such great numbers in Figure 4–5. This example shows the important role in astronomy (as in other natural sciences) played by observational selection.

There are other categories of stars. In Figure 4–6 we see a number of stars situated a little lower than the main sequence. These are the "sub-dwarfs." Although there are relatively few sub-dwarfs near our Sun, they exist in vast numbers in the central regions of the Galaxy, and in the globular clusters. Rarely are sub-dwarfs found near the Galactic plane, but they are very numerous towards the Galactic center. They are apparently the most numerous type of star in the Galaxy. Sub-dwarfs differ from stars of the main sequence in their relatively low content of heavy elements. ▽ If an ordinary main sequence star were somehow to have its heavy element content suddenly decreased, its luminosity would increase, and it would move upward and to the left in the Hertzsprung–Russell diagram, entering the sub-dwarf region, but still lying below the main sequence. The low heavy element content causes less absorption of the radiation emitted from the depths of the star and therefore leads to higher luminosities. △

As we go along the main sequence from spectral classes O to M, the masses of the stars continuously decrease. For example, stars of type O have a mass which is several tens of times greater than the solar mass. Stars of class B have a mass approximately five times greater than that of the Sun, which is in spectral class G2. Most main sequence dwarfs are of spectral class M and have their masses approximately ten times less than the Sun's. Since mass and luminosity continuously change along the main sequence, there exists an empirical relationship between them [see Figure 4–7].

Soon after the "spectrum-luminosity" diagram was published, its intimate connection with the problem of the evolution of the stars was intuitively felt by astronomers. It was formerly believed that the stars evolved directly along the main sequence. According to these naïve concepts, the red giants were the first stars to be formed. As they condensed and shrank, their temperatures increased, and they entered the main sequence. Evolving along the main sequence, they became cooler and radiated less. The present terminology of astronomers still reflects these old concepts: the spectral classes O, B, A, and part of F are called the "early" types, and G, K, and M are called the "late" types. If the stars evolved directly along the main sequence, it would be necessary to conclude that they continuously lost a significant part of their original mass. Such concepts present insurmountable difficulties. The modern theory of stellar evolution, based on contemporary concepts of the source of stellar energy and on much observational material, was developed during the last decade. This theory, which successfully explains the "spectrum-luminosity" diagram, will be discussed in Chapter 6.

5

The interstellar medium

Nothing exists but atoms and the void.

Democritus

Since contemporary theories propose that the stars were formed from a condensation of the interstellar medium, we must investigate the properties and content of this material before discussing the evolution of the stars. This subject will also be important when we consider the problem of contacting life in other regions of the universe. The possibility of establishing communications between the civilizations of different planetary systems depends, to a certain extent, upon the characteristics of the matter which fills the intervening space.

The interstellar medium is composed largely of gas and dust. The mass ratio of gas to dust in a typical volume of space is approximately 100:1. Observations indicate that this material is distributed irregularly and unevenly throughout the Galaxy. There are clouds which are much denser than the medium in general; they appear to us as dark or luminous nebulae and are found mostly in the spiral arms. Individual clouds may have velocities of 6 to 8 km/sec. In other regions of the Galaxy, the density of the interstellar material is extremely low.

The dimensions of cosmic dust grains are from 10^{-4} to 10^{-5} cm, ▽ or about the same size as the wavelength of visible light △. These particles are responsible for the absorption of light in interstellar space that prevents us from observing objects in the Galactic plane at distances greater than 2,000 or 3,000 parsecs. Fortunately, this cosmic dust and the interstellar gas associated with it are concentrated near the Galactic plane in a layer only about 250 parsecs thick. Therefore, radiation emitted by objects at a sufficiently large angle to the plane is not significantly absorbed.

▽ Interstellar grains are built up by low-energy collisions among the atoms and molecules of the interstellar medium. Very energetic collisions vaporize the grains. It is believed that these energetic collisions occur just frequently enough to limit the size of the grains to the range of 10^{-4} to 10^{-5} cm. Because the composition of the grains should approximately parallel the composition of the interstellar gas, the grains should be composed primarily of molecules of the atoms carbon, nitrogen, oxygen, and hydrogen. By an interesting coincidence, the size range and chemical composition of interstellar grains are therefore quite close to the size and composition of terrestrial bacteria. The similarity in composition has an underlying cause in the cosmic abundance of the elements; the similarity in size is a coincidence only. △

The presence of interstellar gas was discovered in the early part of the twentieth century from the absorption lines of ionized calcium which occur in the spectra of remote hot stars, but which are really due to calcium in the intervening interstellar medium. The density of this gas is extremely low, approximately one atom per cubic centimeter, on the average, in regions near the Galactic plane. In air there are 2.7×10^{19} molecules per cubic centimeter. Even in the best vacuum which can be produced in the laboratory, the concentration of atoms is at least 10^{13} cm^{-3}. And yet we cannot consider interstellar space a vacuum. A vacuum is defined as a system in which the mean free path of the atoms or molecules (▽ that is, the average distance that the particles move between collisions △) exceeds the characteristic dimensions of the system. In interstellar space, the mean free path of the atoms is hundreds of times less than the distance between the stars. Therefore, we can rightly consider interstellar gas as a uniform, continuous medium, to which the laws of gas dynamics can be applied.

The chemical content of interstellar gas, revealed through spectroscopy, is similar to that of the external layers of stars on the main sequence. Atoms of hydrogen and helium predominate; metallic atoms are comparatively rare. The simplest molecular compounds (for example, CH, CN) are present in detectable amounts. It has been postulated that perhaps a significant part of the interstellar gas could be in the form of molecular hydrogen, H_2, but there are as yet no methods for determining the validity of this view.

The temperature of the interstellar gas depends upon its distance from a hot star. The ultraviolet radiation of hot stars of spectral class O5 will ionize nearly all the gas within a radius of approximately 100 parsecs. Such regions are called "H II" regions, and their temperatures can reach 10,000°K. (The temperature of a gas is defined by the velocities of the characteristic random motions of the particles.) Under these conditions, the medium emits radiation in the visible part of the spectrum at distinct frequencies, particularly at the frequencies characteristic of a red hydrogen line. When almost all of the interstellar material is far from the hot stars, the interstellar hydrogen is not ionized, and the temperature of the gas is as low as 100°K or lower. There are probably significant amounts of hydrogen molecules in such cold regions.

During the last decade, radio astronomy has proved very valuable in the study of interstellar gas. Investigations at a wavelength of 21 cm have been particularly fruitful. Why this wavelength? Several decades ago, it was theoretically predicted that neutral atoms of hydrogen under the conditions of interstellar space would radiate at a wavelength of 21 cm.

▽ At ultraviolet, visible, infrared, and radio wavelengths, atoms emit or absorb radiation because their electrons change energy. In the case of the simplest atom, hydrogen, there are many possible electron orbits. An electron in an orbit far from the nucleus has more energy than an electron in an orbit closer to the nucleus. When the electron's orbit changes from large to small, this difference in energy is emitted as a light photon. The lowest-energy orbit of the hydrogen atom is called the "ground" state. Actually, it consists of two different orbits of very slightly differing energy. The nucleus of the hydrogen atom is a proton, which has a small associated magnetic field. This field can be pictured as having a direction in space perpendicular to the plane of the electron orbit. In this simplest picture of the hydrogen atom, the electron, moving about the hydrogen nucleus, has a magnetic field associated with it, since moving charged particles produce magnetic fields. The magnetic field produced by the electron revolution about the proton can also be pictured as perpendicular to the plane of the electron orbit; but whether the electron field and the proton field are in the same direction or in exactly opposite directions depends on the direction in which the electron revolves around the proton, clockwise or counterclockwise. These two different directions of revolution correspond to the two slightly different energies which make up the ground state of the hydrogen atom. △ According to the laws of quantum physics, occasional spontaneous transitions take place from the higher ground energy level to the lower ground energy level. When this occurs, a photon of low energy is emitted. Its frequency is proportional to the difference between the two energy levels. Since the difference is very small, the frequency of the radiation will be low. The corresponding wavelength is 21 cm.

Calculations indicate that such transitions between levels of hydrogen atoms occur rarely; on the average, one transition per atom every 11 million years! For other lines in the visible spectrum, the transitions occur perhaps once every 100 millionth of a second.

Since the interstellar atoms have different velocities as seen from Earth, then, due to the Doppler effect, not all the radiation emitted by hydrogen will be at 21 cm. Those moving towards the observer will emit at wavelengths somewhat shorter than 21 cm; those moving away, at longer wavelengths. As a result, there will be a spread of wavelengths around 21 cm. Thus, by measuring the width of the 21 cm line, it is possible to determine the state of motion of the interstellar gas in the Galaxy, and investigate Galactic rotation and the disordered motions and temperatures of individual clouds of interstellar matter. The approximate number of hydrogen atoms in interstellar space has also been determined.

These methods have been used to study other galaxies, for example, the Andromeda Galaxy, M31 [Figure 3–5]. As the techniques of radio astronomy are improved, we shall be able to study the motions and rotations of very remote galaxies. The investigation of interstellar hydrogen at the 21 cm wavelength has inaugurated a new era in astronomy.

▽ Recently, an additional interstellar radio absorption line has been discovered at a wavelength of 18 cm. This line is due to absorption by the molecular fragment OH, called the hydroxyl radical. That OH absorbs at 18 cm was first predicted by Shklovskii many years ago. It already appears that the distribution of OH in interstellar space is different from the distribution of H. As we shall see in Chapter 8, it is believed that oxygen, but not hydrogen, is synthesized in the deep interiors of hot stars. Thus, the difference in distribution of oxygen and hydrogen in interstellar space may provide some significant clues on the sites of element generation within the Galaxy. At the present time many aspects of the interstellar absorption and emission spectrum near 18 cm have been investigated, but they are not at all well understood. The spectral features seem to be localized preferentially near H II regions. The details of the spectrum have been very difficult to interpret, especially since the emission is strongly polarized and sometimes varies in strength over a period of months. Other interstellar lines besides H and OH have been predicted by Shklovskii and others. If they are discovered and their distributions mapped, it may be possible one day to draw maps of the relative abundances of various chemical elements through the Galaxy. △

Astronomers have obtained a number of indirect proofs of the presence of interstellar magnetic fields. These magnetic fields are associated with clouds of interstellar gas and move with them. Their intensity is approximately 10^{-5} gauss. The general direction of the lines of magnetic force coincides with the direction of the arms of the spiral structure of the Galaxy. We can say that the spiral arms themselves represent magnetic tubes of force of gigantic dimensions. If interstellar gas is found in a magnetic field, the 21 cm lines should be split into several components (which differ in polarization). Since the magnitude of the magnetic field is very small, the splitting will be slight. The width of the absorption line is also affected by the magnetic field. ▽ Confirmed observation of this magnetic splitting and broadening of interstellar radio lines should permit a more direct determination of interstellar magnetic field strengths. △

The primary cosmic rays which fill interstellar space are closely associated with the interstellar magnetic fields. Cosmic ray primaries are particles (protons, the nuclei of the heavier elements, and also electrons) which frequently have energies exceeding 1 erg per particle, and sometimes approach $10^6 - 10^7$ ergs/particle. ▽ (A hydrogen atom has a mass of 1.66×10^{-24} gm; for particles of such small mass to have such high energies, they must move exceedingly fast. The high energy cosmic ray primaries in fact have velocities very close to the velocity of light.) △ They move along the lines of force of the magnetic fields in spiral trajectories. Until recently, cosmic rays could be studied only in the immediate vicinity of the Earth's surface. Now, with radio astronomy, we can study cosmic radiation indirectly in the depths of the Milky Way and even beyond its limits, since the electrons of the cosmic rays radiate radio waves. Radio astronomy has placed the problem of the origin of cosmic radiation on a firm scientific foundation.

Until fairly recently, investigators working on the problem of the origin of life did not consider the question of ultra-high energy radiation. In my opinion, however, cosmic rays are an essential factor in evolution. The evolution of life on Earth might have been entirely different if the level of high energy radiation had been ten times greater than the present level. The rate of mutations would have increased greatly. Hence, a very important question arises: Does the level of cosmic radiation remain constant on all planets where life develops? (We are here concerned with periods of time covering many hundreds of millions of years.) Contemporary astrophysics and radio astronomy have answered this question, as we shall see in Chapter 7. ▽ This view of Shklovskii's is, in my opinion, not strongly supported by existing biological evidence. Some discussion of the causes of evolution and the role that radiation may have played in the development of life on Earth is made in Chapters 14 and 17. △

The mass of the interstellar gas in our Galaxy is approximately a billion times the mass of our Sun; yet it is no more than one percent of the total mass of the Galaxy. The remainder of the mass of the Galaxy is mostly in the form of stars.

In other galaxies, the relative content of the interstellar gas varies greatly. In elliptical galaxies [Figure 3–8] it is very small, approximately 10^{-4} or even less. In the irregular galaxies (for example, the Magellanic Clouds), interstellar gas makes up from 20 to 50 percent of the total mass of the galaxy. This circumstance is closely connected with the question of the evolution of the galaxies, a topic which we treat in Chapter 9.

6

The evolution of the stars

But what comes after? What passes when all Creation is destroyed, when the gods are dead, and the chosen warriors, and the races of men? . . . Will there be gods again; will there be any earth or heaven?

The Ragnarok

▽ On a clear night, we can look up into the sky and see myriads of twinkling stars. These tiny pinpoints of light seem to move across the sky from the eastern to the western horizon, as the Earth rotates from west to east; but their relative positions are unchanged from night to night. Therefore, we can recognize certain random star groupings which suggest mythological or other images to us and call them "constellations." We know from ancient manuscripts and tablets that the constellations had much the same form many thousands of years ago as they do today—although, as is also the case with non-cosmic projective tests, they evoked different images. However, if stars were forming, evolving, and dying on a timescale of a few thousand years, the constellations of ancient times would have been very different from the present ones. Hence, we may conclude that the characteristic stellar evolutionary timescale must be at least ten thousand years—and may be much greater.

▽ Our Sun is a typical star in many respects. Its mass, radius, luminosity, and chemical composition are not extraordinary. Some stars are more massive than our Sun, others less. Some are larger in radius, others smaller. The brightest visible stars have an intrinsic luminosity much greater than our Sun's. But the nearest stars are intrinsically less bright. There are stars with a higher proportion of heavy elements than the Sun's, and others which appear to be composed primarily of hydrogen.

▽ Since the Sun is typical in so many of its characteristics, its age may also be expected to be typical. How can the Sun's age be determined? If we understood why the Sun radiates energy into space at such a prodigious rate, we could estimate its total fuel supply. This would give us a rough idea of its age and anticipated lifetime. We are therefore led to another fundamental question: Why does the Sun shine?

▽ Combustion is a familiar source of energy. If a substance such as coal—composed primarily of the element carbon—is heated in the presence of the oxygen in the atmosphere, a chemical reaction follows producing the gas carbon dioxide, CO_2, from C and O_2. Carbon and oxygen have such an affinity for each other at elevated temperatures that the reaction is violent, and much more heat is generated than is used in initiating the reaction. This released energy is observed as fire. Let us make the naïve but instructive hypothesis that the Sun shines because it is on fire. Let us assume the Sun's composition to be half coal and half oxygen, ignoring, for the moment, the spectroscopic evidence that the Sun is composed primarily of hydrogen, and contains little carbon and oxygen. The formation of one gram of carbon dioxide by the reaction $C + O_2 \rightarrow CO_2$ yields about 3.4×10^{11} ergs. Therefore, if all 2×10^{33} grams of the Sun were burnt, 3.4×10^{11} erg

$gm^{-1} \times 2 \times 10^{33}$ gms $= 6.8 \times 10^{44}$ ergs would be released. The Sun is radiating to space 4×10^{33} ergs sec^{-1}. This is the value of the solar luminosity. Therefore, the length of time the Sun could be radiating to space with our hypothetical combustion energy source at its present luminosity is $(6.8 \times 10^{44}$ ergs$)/(4 \times 10^{33}$ ergs $sec^{-1}) = 1.7 \times 10^{11}$ seconds. There are about 3.16×10^{7} seconds in a year; therefore, the lifetime of our burning Sun appears to be about 5400 years. If it were any older, it would be extinguished by now.

▽ Our conclusion is in reasonable agreement with the age of the Earth determined by Archbishop Ussher who, in the seventeenth century, added the lifetimes of various personages in the Bible, made appropriate interpolations and extrapolations, and concluded that the Earth was formed approximately 6000 years ago. We assume the Sun is on fire, and deduce that its lifetime is about 5400 years. This is an interesting example of the pitfalls of science. The methods are different, but the conclusions are roughly the same. (Presumably the Sun and the Earth are approximately coeval.) Some might be tempted to conclude that Biblical chronology and the coal-burning Sun hypothesis are both strengthened, and that an apocalyptic death of the Sun is imminent. However, there are other facts which are inconsistent with these views.

▽ For example, geologists find that the Earth is covered by sedimentary rock, layered down by the action of rivers and waterways. At the present sedimentation rate, tens of millions of years would be required to give the observed amount of sedimentary rock. Paleontologists find these sedimentary layers filled with fossils of organisms, now extinct, which once had world-wide distribution. Tens or hundreds of millions of years are necessary to explain the evolutionary origin of these creatures with present rates of evolution. The amount of salt in the ocean comes from alluvial erosion. From the present abundance of salt in the ocean and the present rate of erosion, it can be concluded that the salt took at least a hundred million years to accumulate.

▽ This kind of argument was common in written discussions of the ages of the Earth and the Sun late in the nineteenth century. Since that time, the discovery of radioactivity has placed the entire subject of the Earth's chronology on a firm basis. The isotopes of some elements, such as uranium, spontaneously and unpredictably emit charged particles from their nuclei and then weigh less; they have been transmuted to a different atom of lower atomic weight. When a uranium isotope completes its sequence of radioactive decays, it becomes a particular isotope of lead. The lead is stable, and decays no more. The characteristic time for, say, half a given lump of uranium to turn into lead can be determined; the time is unaffected by the local temperature, pressure, or other conditions. Therefore, by measuring the amounts of uranium and lead isotopes in a given sample of rock, we can derive the time elapsed since the piece of rock first formed, as well as its original chemical composition. In this way, it has been possible to conclude that the Earth assumed its present form some 4.5×10^{9} years ago. Analyses of meteorites—small chunks of stone and iron from the asteroid belt—show they were formed in the same epoch as Earth. Since it seems unlikely that the Earth or the asteroids were formed much before the Sun, we can conclude that the Sun is at least 4.5×10^9 years old. And, because the Sun is a typical star, the characteristic ages of many stars are therefore several billions of years.

▽ But what is the energy source which makes the Sun shine? We have seen that combusion is far too feeble. At the turn of the century, other explanations were proposed. Some thought that solar energy was supplied by the collision of large numbers of meteors with the Sun; others suggested that the Sun was slowly contracting, and that the slow increase of density in the solar interior was responsible for the observed solar luminosity. But the lifetimes computed with these assumptions were too small by factors of 100 or more. Obviously, some other energy source existed, but its nature could only be dimly surmised. In 1926, the British astrophysicist Sir Arthur Stanley Eddington mused, "Does energy issue freely from matter at 40,000,000° as steam issues from water at 100°?"

▽ It is curious that the same discovery of radioactivity which led to an accurate determination of the age of the Earth also resulted in an understanding of the solar luminosity. From the mass and composition of the Sun, it is possible to compute the pressure in its interior, since this pressure is determined by the weight of the overlying material. The gases near the Sun's center are found to have temperatures of ten million degrees or more. Advances made in nuclear physics in the 1930's proved that at such temperatures atoms collide with each other so vigorously that enormous quantities of energy are released, in an atomic analogy to the way heat is released during the molecular combustion of carbon. Eddington's question was answered affirmatively. Since the Sun is composed primarily of hydrogen, the thermonuclear reactions which occur in the solar interior involve the jamming together of four hydrogen nuclei, or protons, and the formation of one helium nucleus. The rate of such reactions is strongly dependent upon the temperature. The liberated energy slowly emerges from the interior of the Sun, and is finally transformed, near the surface, into radiation that is emitted into space. This is the only source of the Sun's power today.

▽ The energy released is about 6×10^{18} ergs for each gram of hydrogen converted into helium. Thus, the process is about ten million times more efficient than the burning of coal. This is also approximately the same factor of increased efficiency—if that is the word—of thermonuclear weapons over ordinary explosives such as TNT. A nuclear device weighing about a ton may provide an explosive energy of ten megatons—that is, of ten million tons of TNT. This common nomenclature for nuclear weapon firepower testifies to the 10^7 greater efficiency of thermonuclear processes over chemical processes.

▽ The same factor of 10^7 is the increase in the age of the Sun which we obtain by using thermonuclear rather than chemical energy sources. Instead of 5×10^3 years for the lifetime of the Sun, we obtain $5 \times 10^3 \times 10^7 = 5 \times 10^{10}$ years, or, comfortably, about ten times the age of the Earth. Thus, if the Sun should spend all of its hydrogen nuclear capital, it could shine with its present luminosity for another 45 billion years or so ($5 \times 10^{10} - 5 \times 10^9 = 4.5 \times 10^{10}$). However, there is a limit to the amount of hydrogen which the Sun can convert into helium before other complications occur. Recent calculations estimate that the Sun can be expected to shine at its present luminosity, on the main sequence, for about another 8 billion years. What happens after that, and the resulting consequences for the Earth will be discussed presently.

▽ It was once believed that all the stars in the heavens were formed at about the same time, several billions of years ago. But there are now a number of lines of evidence indicating that stars are being formed continuously, by condensation of the interstellar gas and dust. Even today, the mysterious processes of stellar origins are occurring in regions of the Galaxy difficult for us to observe, and by mechanisms we only incompletely understand.

▽ The problem of stellar evolution can be likened to the processes of human development. Imagine, in keeping with the ultimate subject matter of this book, that you are an intelligent extraterrestrial being—probably not a Martian—landed for the first time on Earth. You would most probably, quickly pass over other living creatures, from viruses to baleen whales, and come to focus your attention on human beings as the most important life form on the planet. Looking at a random sample of human beings, you would observe two sexes, a fairly wide range of colors, various physiognomic peculiarities, and a continuum of sizes, from about 50 cm to about 200 cm. The 50 cm human beings would be very rare. There would be a continuous distribution of sizes, with a peak around 170 cm. Occasionally, you would come upon locales—say, schoolyards—where there was a concentration of human beings in the 120 cm size range. You would be faced with the problem of explaining the origin of human beings. Is each size, sex, and color immutable? Do the little ones become big ones? Do the big ones become little ones? Do changes in sex or color occur? Frequently, or rarely? If you were only on the Earth for a period much shorter than the characteristic timescale for human development—say you were on the Earth for a week—any conclusion regarding human growth would have to be inferential, and not directly observed. You would also have to eliminate effects of observational selection, because the newly formed human beings would be the most difficult to observe. And even if you had formulated the correct hypothesis about human development, the methods of origin of individual human beings might still be obscure; in fact, the correct explanation might seem, on superficial grounds, highly unlikely. Our understanding of the life cycle of the stars is not dissimilar.

▽ We have been observing stars in detail for only a few hundred years. We have seen the lifetime of a star such as the Sun to be about 10^{10} years. Therefore, we have observed typical stars for only 10^{-8} of their life cycle. The comparable situation in the observation of human beings (characteristic lifetimes, 60 years) would be observations for $60 \times 10^{-8} = 6 \times 10^{-7}$ years $= 6 \times 10^{-7}$ years $\times 3 \times 10^{7}$ seconds year^{-1} = 18 seconds. It would be a very intelligent extraterrestrial indeed who could understand the human life cycle after 18 seconds of observation. The depth of our understanding of stellar evolution is therefore especially remarkable.

▽ We know that all stars, like all people, are not the same age, and that, as a star passes through its life cycle, it takes on the characteristics of many superficially different kinds of stars. Some stars are relatively young objects; others are much older than the Sun. The young stars are generally the very bright ones. Stars with large intrinsic luminosities burn their nuclear fuel very rapidly, and have corresponding short lifetimes. △

The groups of young stars are concentrated in the spiral arms of the Galaxy, those twisting tubes of magnetic force which contain most of the interstellar gas in the Milky Way. We recall that the Galactic magnetic field strengths are very weak, not exceeding 10^{-5} gauss. Magnetic forces, then, would not be strong enough to affect the motions of such dense and massive objects as the stars. Thus, we conclude that young stars are found only within the spiral arms, not because the Galactic magnetic field holds them there, but because they have only recently been formed out of the interstellar gas concentrated there. The older stars are found in great numbers in the Galactic nucleus, and in the halo of the Galaxy. Here, the gas density is very low. Some of these older stars have traveled far from their birthplaces over timescales of 10^8 or 10^9 years: Thus the distribution of old stars and young stars in the Galaxy is an important argument supporting the view that stars are formed from the interstellar medium.

▽ Let us now consider the evolution of the stars. Since the earliest stages of stellar evolution are still unknown, we will concern ourselves with the current theoretical picture of the evolution of interstellar gas masses. It is postulated that, under certain specific conditions, a cloud of interstellar gas and dust begins to condense. Perhaps this occurs by gravitational forces, each part of the interstellar cloud attracting all the other parts. The cloud may fragment into smaller clouds; the smaller clouds, into still smaller ones, until eventually a cloud of approximately stellar mass is formed. We may picture it as a relatively dense, opaque gas sphere. △

Strictly speaking, this sphere is not yet a star, since the temperatures in its central regions are not high enough for the occurrence of thermonuclear reactions. At these low temperatures, the gas pressures inside the sphere are not great enough to overcome the forces of gravitational attraction between individual grains and atoms. Hence, the sphere continues to contract. It is believed that protostars in this opaque, distended stage, can be seen, in gaseous nebulae, as small, dark splotches, called globules [see Figures 6–1 and 6–2]. There are reasons to believe that protostars are formed collectively. Subsequently, these groups of protostars evolve into stellar associations or clusters. It is highly probable that during this early stage of evolution aggregates of smaller masses are formed around the stars, and that these are gradually transformed into planets [see Chapters 11–13].

As a protostar contracts, its gravitational potential energy is being converted into heat and light. It takes tremendous amounts of energy to heat a stellar mass from temperatures near absolute zero to tens of millions of degrees. The remainder of the potential energy released during contraction is radiated into surrounding space. Since the dimensions of the contracting gas sphere are very large, the amount of energy radiated to space by each unit of surface area—say, one square

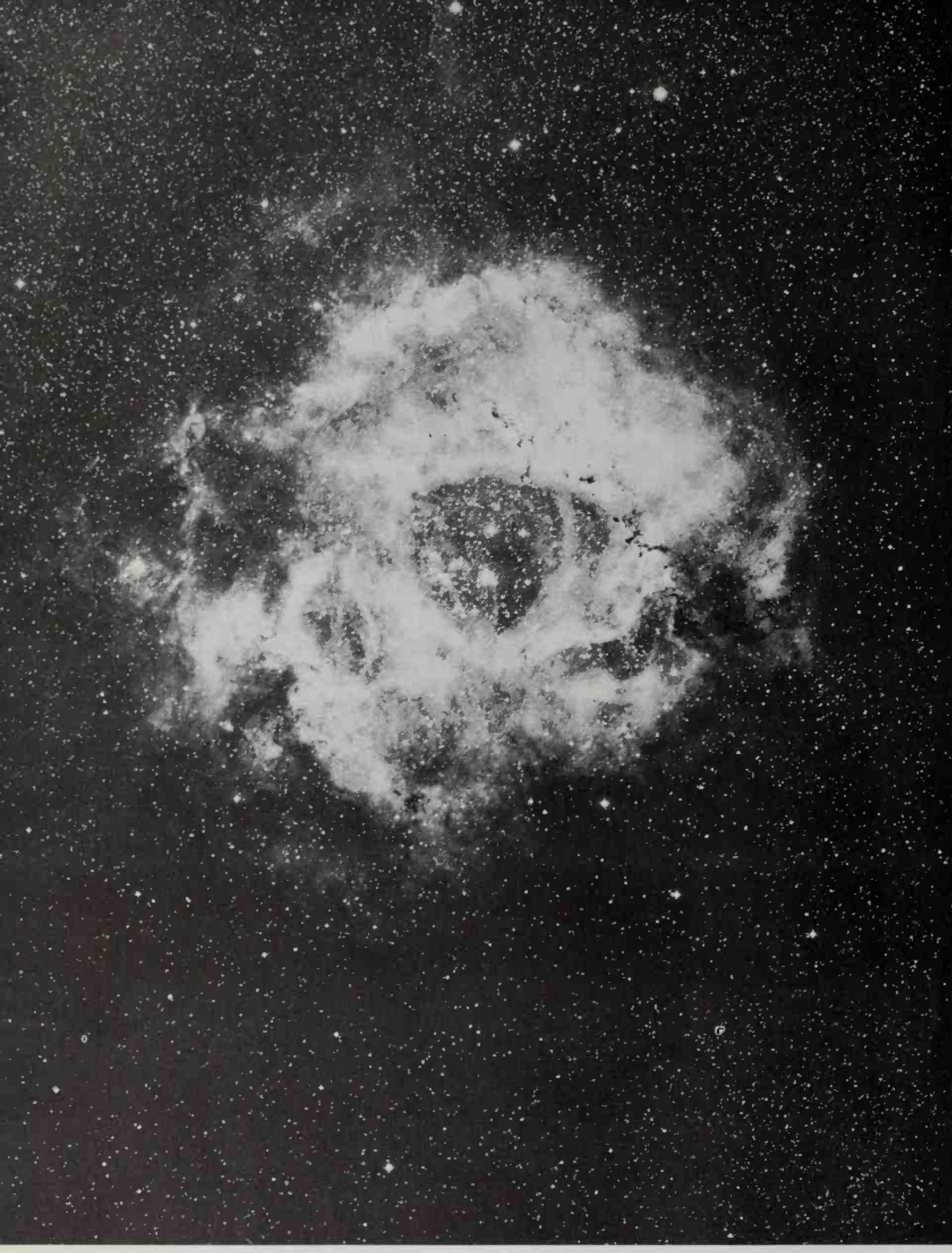


FIGURE 6–1. *The gas and dust nebula NGC 2237 in the constellation Monoceros. The dark splotches are thought to be great concentrations of absorbing dust. The small black globules may be cold stars in the earliest stages of formation. (Courtesy of Mt. Wilson and Palomar Observatories.)*

FIGURE 6–2. *An enlarged portion of the nebula NGC 2237, showing the absorbing clouds and globules more distinctly. (Courtesy of Mt. Wilson and Palomar Observatories.)*

centimeter—will be very small. The Stefan-Boltzmann law of physics shows that the amount of energy radiated by unit surface area is proportional to the fourth power of the temperature. Thus, if the temperature is doubled, the radiation per unit surface area is increased by $2^4 = 16$ times. The surface layers of such protostars may therefore be relatively cool, while the luminosity may be almost the same as for older evolved stars of similar mass, the larger surface area compensating for the lower temperatures. Such protostars, with low temperatures, but not necessarily low luminosities, therefore lie to the right of the main sequence in the spectrum-luminosity, or Hertzsprung-Russell, diagram described in Chapter 4; that is, they fall into the region of the red giants or red dwarfs, depending on the mass of the protostar.

As time passes, the protostar continues to contract. Its dimensions become smaller, and both its interior and surface temperatures increase. ▽ This earliest contraction phase of the protostar occurs relatively rapidly, on the cosmic timescale. The contraction time depends only on the initial density of the gas cloud. If the original gas cloud had a density of 10^4 atoms in every cubic centimeter, the collapse time would be about 5×10^5 years. If the original density is greater, the collapse time is smaller, because the attractive gravitational forces leading to collapse become more efficient as the density is increased. The interior temperatures of the protostar reach about 100,000°K at this point. Since this is not high enough to initiate thermonuclear reactions, a star has not yet formed. But the temperature is large enough to ionize the hydrogen and helium which are the predominant constituents of the protostar; that is, the temperatures become high enough so that electrons are stripped off atoms of hydrogen and helium by violent collisions. These ionized atoms are much more efficient absorbers of the radiation generated in the protostellar interior than are their neutral counterparts having a full complement of electrons. The increase in stellar opacity, in turn, raises the internal temperatures. Radiation which previously escaped to space is now trapped in the interior and causes it to heat still further. At this critical temperature, the atoms and ions in the interior of the star are moving fast enough to exert an upward pressure approximately balancing the weight of the overlying material. In this approximate equilibrium situation, the rate of contraction of the star declines.

▽ The star now goes into a stage of convection, in which there is mass exchange between the interior and the exterior. Having increased in luminosity during the collapse phase, the star now decreases in luminosity and approaches the main sequence, as shown in Figure 6–3. Here, we see a temperature-luminosity diagram such as was described in Chapter 4. The oblique, solid line represents the main sequence.

▽ The vertical axis represents the luminosity, in units of the solar luminosity, $L_\odot$. For example, $L/L_\odot = 10$ represents a luminosity ten times greater than the Sun's. The horizontal axis represents the temperature of the outer layers of the star, those which radiate directly to space. The temperatures are expressed in thousands of degrees Kelvin. The present position of the Sun is located at the point $L/L_\odot = 1$, and a temperature of 6000°K, approximately. We see that the contraction phase in the early evolution of a star of solar mass begins at the very

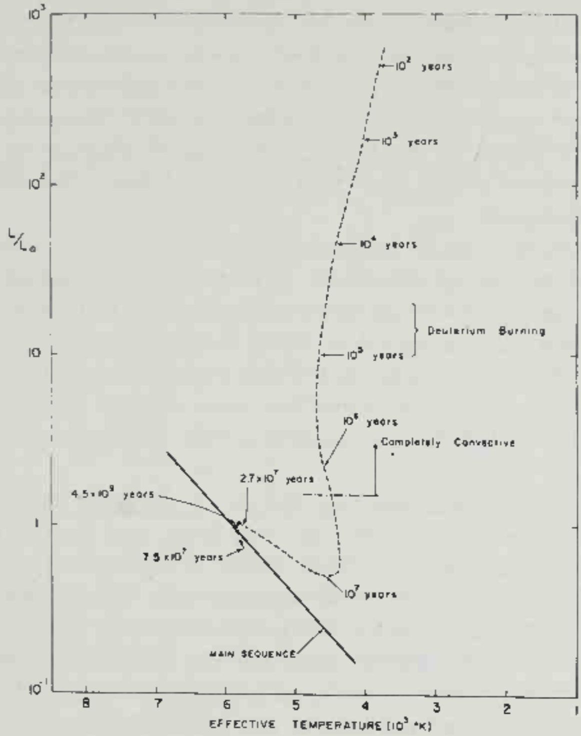


FIGURE 6–3. *A theoretical diagram of the evolutionary track of a star of solar mass, on its way toward the main sequence. The evolutionary track is shown as the dashed line beginning at very high luminosities and descending towards the inclined solid line which represents the main sequence. The point marked 4.5×10^9 years represents the present position of the Sun. (Courtesy of Dr. A. G. W. Cameron and Dr. G. Ezer, Institute for Space Studies, New York.)*

high luminosity of $L/L_\odot$ somewhat less than 1000. We see, however, that it is declining in luminosity very rapidly, and in a few million years it has approximately the present solar luminosity. In something like 100 million years, it has traversed the hook, shown by the dashed line, and entered onto the main sequence. △ The temperature in the stellar interior has now become sufficient for thermonuclear reactions to begin. The pressure of the gas in the interior regions balances the forces of gravitational attraction, and the gaseous sphere stops contracting. The protostar has become a star.

We may, in fact, be observing stars in the stage of vertical contraction towards the main sequence. A species called the T Tauri stars are found embedded in dark nebulae. ▽ Their luminosity changes, sometimes erratically, with time. There is some evidence that they are losing mass, and they are found in the appropriate part of the temperature-luminosity diagram. △

After the star terminates its contraction phase and enters the main sequence, its position in the temperature-luminosity diagram changes very little, over long periods of time. ▽ For example, our Sun has moved up and to the left along the main sequence only a very small amount (corresponding to an increase in luminosity of half a magnitude, or by about 20%) during the past 5×10^9 years. △ On the main sequence, the radiation is maintained by thermonuclear reactions in the interior and by the accompanying conversion of hydrogen into helium. Thus, the main sequence represents not an evolutionary track but the geometric locus of points in the Hertzsprung–Russell diagram on which stars of various masses radiate stably over long periods of time, due to thermonuclear reactions in their interiors.

All stars of equal mass do not occupy the same position in the Hertzsprung–Russell diagram, because of differences in chemical compositions. If a protostar has a relatively small amount of heavy elements, the star arrives on the main sequence at a lower position than if its heavy element content were high. We have already mentioned this theoretical result which explains the sequence of subdwarfs, having a heavy element content about ten times smaller than that of the main sequence stars, and lying below the main sequence appropriate to the majority of stars.

A star's initial mass determines its lifetime on the main sequence. If the mass is great, the interior temperatures are also very large, and the star becomes a very powerful source of radiation. But as a result, it rapidly depletes its supply of hydrogen fuel. Thus, for example, main sequence stars with masses 20 or 30 times greater than that of the Sun—the hot blue giants of spectral class O—remain on the main sequence only a few million years. On the other hand, stars with masses close to that of our Sun reside on the main sequence for 10 or 15 billion years. In Table II, we see a computed estimate of the main sequence lifetimes of stars in various spectral classes. The values of the masses, radii, and luminosities of these stars are also indicated in solar units—a mass of 10 means a mass ten times that of the Sun, or $10 \times 2 \times 10^{33}$ grams $= 2 \times 10^{34}$ grams, and so forth.

According to present estimates, the Galaxy is approximately 10 to 20 billion years old. In Table II, we see that the calculated period of time that stars later than K0 spend on the main sequence is much greater than the Galaxy's age. Therefore we conclude that none of these stars have left the main sequence.

The burning of hydrogen—its conversion into helium by thermonuclear reactions—occurs in stellar interiors, because both the high temperatures required to initiate such reactions and the convection necessary for supplying fresh hydrogen for future reactions are found only in this region. Since there is a finite amount of hydrogen in the core, the nuclear fuel will, sooner or later (depending on the star's mass) become entirely depleted. All the hydrogen in the star is not available for nuclear reactions, and eventually a star is left with a hot core composed almost entirely of the thermonuclear reaction product, helium.

What happens to a star when all, or almost all, of the hydrogen in its core is exhausted? The generation of nuclear energy in the central regions must cease. At this juncture the temperatures and the pressures will not be great enough to oppose the gravitational forces which originally contracted the star. The core will then begin to contract while—▽ because of the high temperatures of the interior △—the surface layers expand. The internal temperatures will then increase, while

TABLE II. PROPERTIES OF MAIN SEQUENCE STARS

Stellar spectral type	*Mass in units of the Sun*	*Radius in units of the Sun*	*Luminosity in units of the Sun*	*Residence time on the main sequence in years*
B0	17.0	9.0	30,000	8×10^6
B5	6.3	4.2	1,000	8×10^7
A0	3.2	2.8	100	4×10^8
A5	1.9	1.5	12	2×10^9
F0	1.5	1.25	4.8	4×10^9
F5	1.3	1.24	2.7	6×10^9
G0	1.02	1.02	1.2	1.1×10^{10}
G2 (the Sun)	1.00	1.00	1.0	1.3×10^{10}
G5	0.91	0.92	0.72	1.7×10^{10}
K0	0.74	0.74	0.35	2.8×10^{10}
K5	0.54	0.54	0.10	7.0×10^{10}

the surface temperatures decline. ▽ The increase in surface area of the star more than compensates for the decline in surface temperature; therefore, after the hydrogen in its core becomes depleted, the luminosity of the star will increase. If the star increases in luminosity and decreases in temperature, it must move off the main sequence upward and to the right in the Hertzsprung–Russell diagram. The star has now become a red giant. △

Meanwhile, back in the interior, a very dense, hot region is formed within the core, consisting of helium and small amounts of heavier elements. Nuclear reactions will not occur in this hot region, because the hydrogen there is exhausted. Such reactions *will* take place in a relatively thin layer on the periphery of the nucleus. As the star becomes a red giant, its luminosity is maintained by a thin shell of hydrogen "burning," which separates the helium-rich core from the hydrogen-rich envelope. If the heavy element content is smaller, the red giant will have a higher luminosity.

▽ Figure 6–4 shows a temperature-luminosity diagram giving the theoretically computed evolutionary tracks of stars of various masses. The figure is not as complicated as it appears at first sight. The vertical axis is the logarithm of the stellar luminosity, in units of the sun—that is, 0 indicates a luminosity $10^0 = 1$ times that of the sun, or just the solar luminosity; 2 denotes $10^2 = 100$ times the solar luminosity; -2 indicates 10^{-2}, or 1/100 the solar luminosity, and so forth. The horizontal axis gives the logarithm of the effective surface temperature; hence a value of 3.0 means a temperature of $10^3 = 1000$ degrees; 4.0 indicates a temperature of $10^4 = 10{,}000°$; 5.0 indicates a temperature of $10^5 = 100{,}000°$, etc. The evolutionary tracks off the main sequence are shown for stars having 0.7 of a solar mass, 4 solar masses, and 15.6 solar masses. These stars leave the main sequence at the positions marked by H; that is, they are burning hydrogen in a shell source. We see that a star of 0.7 solar mass leaves the main sequence almost

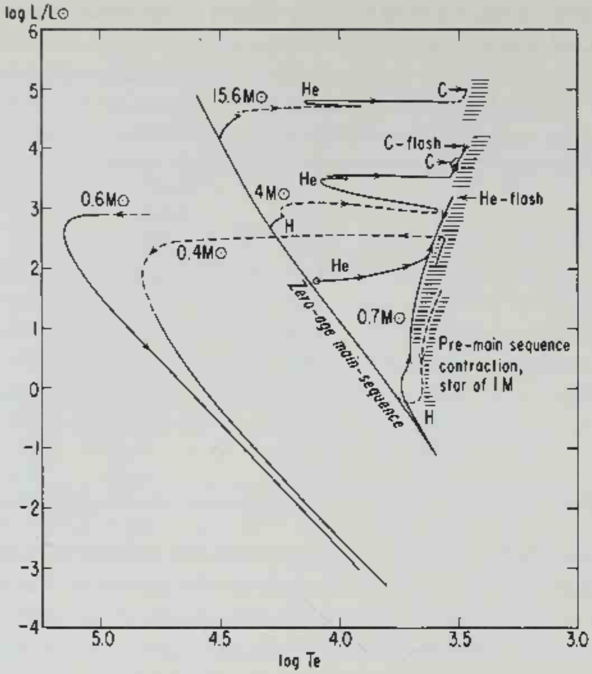


FIGURE 6–4. *Theoretically computed evolutionary tracks on the Hertzsprung-Russell diagram for stars of various masses. After C. Hayashi, R. Hoshi, and D. Sugimoto,* Progress of Theoretical Physics, *Suppl. 22, Kyoto (1962).*

vertically, while stars of larger solar masses leave more horizontally. As a result, the evolutionary tracks tend to "focus" the stars into the same red giant region of the diagram. △ The subsequent evolutionary history of these red giants, as indicated on the diagram, will be discussed presently.

It is important to compare the observed spectrum-luminosity diagrams for individual star clusters with the results of calculations such as those exhibited in Figure 6–4. Clusters of stars—for example, the Pleiades—are chosen because we can assume that such stellar aggregations are approximately coeval. ▽ If they had not been formed at about the same time, it would be difficult to understand their physical association. The faster-moving members of such stellar associations would escape; the cluster also tends to be disrupted by external gravitational perturbations. △ By comparing the spectrum-luminosity diagrams of older clusters with those of younger ones, it is possible to confirm the theoretical calculations of stellar evolution, and even to deduce the ages of individual star clusters. In Figures 6–5 and 6–6, we see spectrum-luminosity diagrams for two different stellar clusters. The horizontal axis is given in terms of color index (B – V), a quantity closely related to the spectral type and temperature of the star, and defined at the beginning of Chapter 4. In Figure 6–5, each dot represents the color index and luminosity of an individual star in the galactic star cluster NGC 2254. (▽ NGC is

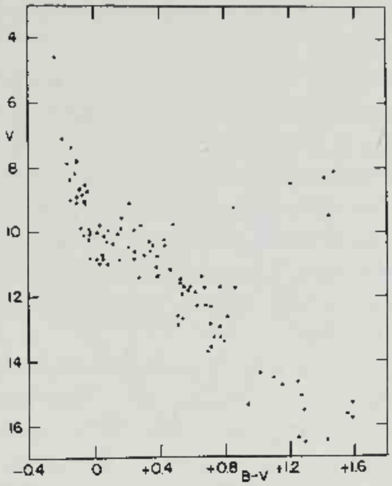


FIGURE 6–5. *Hertzsprung-Russell diagram of the very young galactic star cluster NGC 2254. Reproduced, with permission, from M. Walker,* Astrophysical Journal, *Suppl. 23 (1956).*

an abbreviation for New General Catalogue, a fairly indiscriminate collection of galaxies, gaseous nebulae, and star clusters. It *was* new, when first published, in 1888, but the name now seems less appropriate as the Catalogue has become musty with age. △) The diagram for NGC 2254 shows a heavy concentration of hot, massive stars situated in the upper left of the main sequence. (The color index 0.2 corresponds to a surface temperature of 20,000°K, that is, to a spectrum of

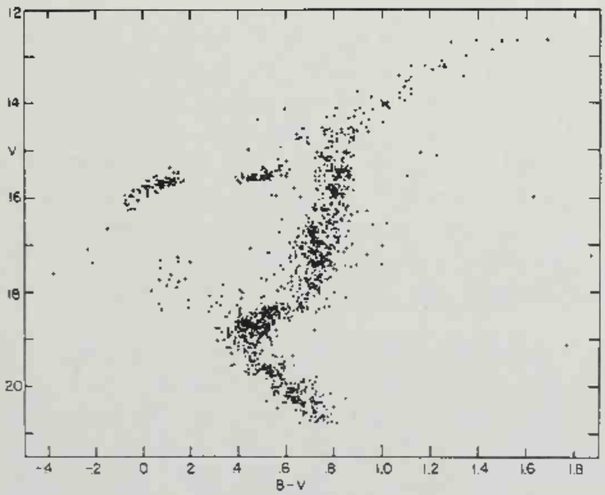


FIGURE 6–6. *Hertzsprung-Russell diagram of the highly evolved globular cluster M3. Reproduced, with permission, from H. L. Johnson and A. R. Sandage,* Astrophysical Journal 124, *379 (1956).*

type B.) ▽ The presence of such hot stars immediately tells us that NGC 2254 is a relatively young formation. △

The globular cluster M3 is an old object. (▽ It is called M3 after an entirely separate, and partially overlapping, catalogue compiled by a French gentleman-astronomer named Messier. △) The diagram of M3 contains almost no stars in the upper left segment of the main sequence. On the other hand, the red giant branch, extending to the right of the main sequence, is very populated. We note that in the diagram for NGC 2254, there are very few red giants. Thus, we conclude that the older clusters, such as M3, have a large number of stars which have already left the main sequence, while the young clusters, such as NGC 2254, contain only a few stars which have evolved this far.

In Figure 6–6, for M3, we see an almost horizontal branch of stars, gently sloping from the upper right to the middle left. There is no analogous branch in Figure 6–5, for NGC 2254. Is it possible that highly evolved stars travel through this horizontal branch? After the temperatures of the dense, contracting helium core of the red giant reach 1.0 to 1.5×10^8 °K, a new kind of thermonuclear reaction sets in. At these temperatures, helium nuclei are jammed together to form carbon nuclei, and additional radiation is released. As soon as this helium burning begins, the contraction of the core ceases. The temperature of the surface layers increases, and the star moves to the left in the spectrum-luminosity diagram. ▽ This feature can be seen in the theoretical evolutionary tracks [Figure 6–4], for example, for a star of 4 solar masses. Detailed computations show even more complicated tracks, and a single star may traverse sections of the horizontal branch many times during its lifetime. The sudden onset of helium burning is known as the "helium flash." The star of 0.7 solar masses, in Figure 6–4, can be observed to reverse its upward motion in the Hertzsprung–Russell diagram very rapidly, after the helium flash.

▽ After much of the helium in the core is exhausted, the situation is somewhat analogous to that after hydrogen depletion. We have carbon burning only near the center, and a shell source of helium burning around the carbon core of the star. In the star of 4 solar masses [Figure 6–4], the onset of carbon burning occurs very rapidly, and is described as a "carbon flash." The main products of carbon burning are oxygen, neon, and magnesium. In general, we can see that there will be a continual progression of core contractions, increases in interior temperatures, and the synthesis of more massive elements in the stellar interior, as the star wends its complicated path through the horizontal branch of the Hertzsprung–Russell diagram.

▽ Figure 6–7 shows the combined color-luminosity diagram for eleven clusters, one of which, M3, is a globular cluster. We see that the main sequence of the different clusters are bent to the right and upwards. The horizontal axis as before, is the color index, which is related to the spectrum and to the temperature of the star, the temperature increasing to the left. The vertical axis is the visual magnitude of the star, which is proportional to the logarithm of the stellar luminosity. These units are the same as for the previous diagrams, in which individual stars were shown for the clusters M3 and NGC 2254. The only

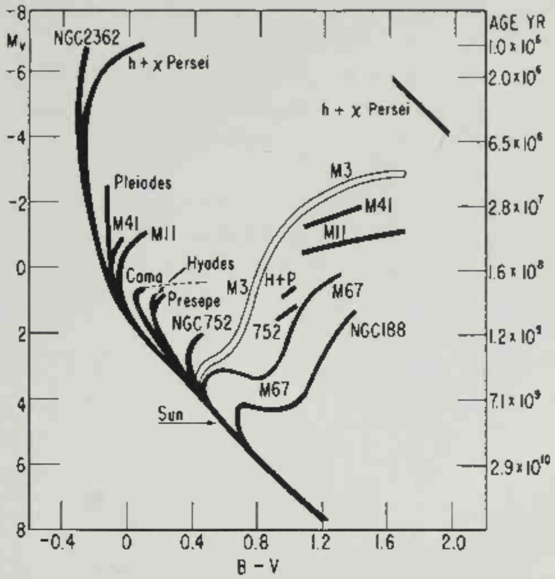


FIGURE 6–7. *Combined Hertzsprung-Russell diagrams for a variety of star clusters. (After A. R. Sandage.)*

difference between the figures is that in Figure 6–7 the absolute visual magnitude is used, to facilitate the comparison of different clusters. In Figures 6–5 and 6–6 only the apparent visual magnitudes have been displayed.

▽ We have seen that the most luminous stars on the main sequence—those toward the top of the Hertzsprung–Russell diagram—burn their nuclear fuel most quickly and are therefore the first stars in a given cluster to turn off toward the right in the spectrum-luminosity diagram and become red giants. With the passage of time, stars further down the main sequence will move off it into the red giant region. Therefore a simple examination of the turnoff point in the main sequence of any star cluster provides an estimate of its age. The ages of the clusters can also be plotted along the vertical axis on the right-hand side of Figure 6–7. We see, for example, that the cluster h and χ Persei has an age of a few million years, while M3 has an age of perhaps 6 billion years, and the galactic cluster M67 is even older. The Hertzsprung–Russell diagram of M67, showing the individual stars, is given in Figure 6–8. We see that the turnoff is at about apparent visual magnitude 12.5. The existence of the giant branch and the horizontal branch testify to its age. The age of the oldest cluster in the Galaxy, determined from its turnoff point, sets a lower limit to the age of the Galaxy. The Galaxy must certainly be older than the star clusters contained within it. In this way, an estimate of 10 to 20 billion years has been obtained for the age of the Galaxy. If the individual stars in globular or galactic clusters of other galaxies could be resolved and their magnitudes measured, we would be able to make some estimates of the lifetimes of other galaxies. Unfortunately, the other galaxies are too far away to permit such measurements. △

This advancement in our understanding of the constitution and evolution of the stars has been one of the great achievements of astronomy during the second half of the twentieth century. It would not have been possible without investigations in the field of nuclear physics, which have led to a detailed comprehension of the nuclear reactions which take place in stellar interiors, or without the aid of high-speed electronic computers.

Let us consider the further evolution of the stars, the stage that comes after red giants. Helium, carbon, and similar burnings in the interiors of the stars cannot proceed indefinitely. What happens when all the nuclear fuel sources are exhausted?

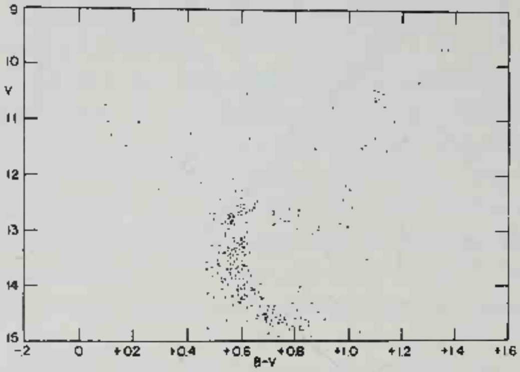


FIGURE 6–8. *Hertzsprung-Russell diagram for the very old galactic star cluster M67. Reproduced, with permission, from H. L. Johnson and A. R. Sandage,* Astrophysical Journal 121, *616 (1955).*

Direct observations and a number of theoretical considerations suggest that in the next stage of the stellar life cycle, a significant fraction of the mass of the star is shed. The outer layers may become separated from the star, and move further and further away, to form a planetary nebula, such as that shown in Figure 6–9.

The intense ultraviolet radiation of the central star, the "nucleus" of the planetary nebula, will ionize the neutral atoms in the nebula and cause them to fluoresce. In some tens of thousands of years, the nebula will dissipate, and only the small, hot, dense central star will remain. Gradually cooling, this star eventually becomes a white dwarf. Thus, the white dwarfs grow in the interiors of the red giants and appear after the external layers of the red giants are ejected. ▽ It is conjectured that in some cases, the shedding of the stellar envelope may occur, not by the formation of a planetary nebula, but by gradual mass ejection. Some examples are known of red giants which are gradually losing matter to space. △

That white dwarfs, of immense density, are a final stage in stellar evolution has been supported by direct observations. The older stellar clusters, for example the Hyades and Praesepe, contain many white dwarfs, while the younger clusters, for example the Pleiades, contain few.

As the white dwarfs gradually cool, they radiate less and less, changing into invisible "black" dwarfs. These stars are dead and cold, but have an enormous

FIGURE 6–9. *The planetary nebula NGC 7293 in the constellation Aquarius. Note the fine lines radiating away from the central star, suggesting that this planetary nebula was formed in an explosion of titanic proportions.* (*Courtesy of Mt. Wilson and Palomar Observatories.*)

density, millions of times greater than water. Their dimensions are less than the dimensions of the Earth, but their masses are comparable with that of the Sun. The cooling process goes on for many billions of years. ▽ Such final evolutionary tracks in the Hertzsprung–Russell diagram can be seen in Figure 6–4, in the lower left of the diagram. The stars move downward and to the right, that is, towards lower temperatures and lower luminosities. The light from the star becomes feebler and feebler, and ultimately disappears. The star has died. △

We must emphasize again that the rate of stellar evolution is determined by the initial composition and initial stellar mass. Since our Galaxy has been in existence approximately 10 to 20 billion years, only those stars with masses exceeding a certain critical value will have gone through the entire evolutionary sequence towards the black dwarf. This critical mass seems to be only 10 or 20 percent greater than the mass of our Sun. ▽ That is, a star of 1.2 solar masses, formed 10 billion years ago, would now be well into the white dwarf evolutionary stage. △

Our Sun was formed about 5 billion years ago. It is believed that the Galaxy at that time was similar, in its basic characteristics, to its present form. For at least 4.5 billion years, our Sun has resided on the main sequence, stably radiating energy due to the thermonuclear reactions in its interior. How long will this stability continue? ▽ The evolution of the Sun can be seen in Figure 6–10. The vertical axis shows the luminosity and the radius of the Sun, in terms of their present values. The horizontal axis shows the age of the Sun in units of 10^{17} seconds. We recall that one year is about 3×10^7 seconds, so 10^{17} seconds is about 3×10^9 years. The two curves for the change in the luminosity and radius of the Sun as time progresses intersect at about 1.5×10^{17} seconds, or about 4.5×10^9 years, approximately the age of the Sun. We see that from the Sun's origin to the present, its luminosity has increased somewhat, while its radius has hardly increased at all. As time goes on, the luminosity of the Sun will begin to increase sharply. In another 6 billion years, when the age of the Sun is 3.5×10^{17} seconds, the luminosity will be increasing very rapidly indeed, and the radius will also begin to increase. The Sun will be well on its way to becoming a red giant. Several calculations have been performed, giving results similar to those of Figure 6–10. △

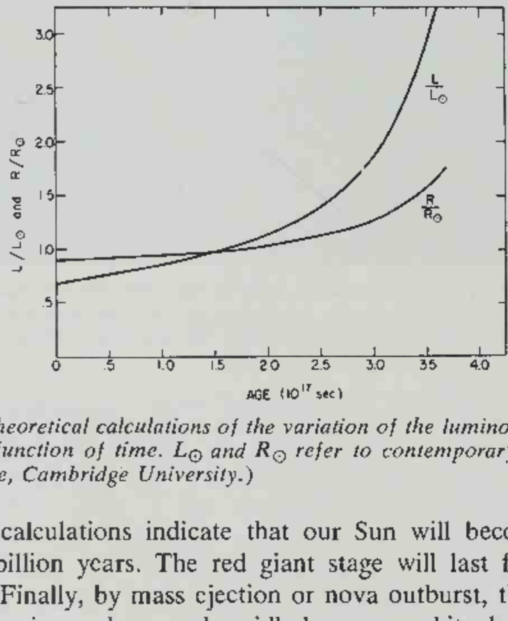


FIGURE 6–10. *Theoretical calculations of the variation of the luminosity, L, and the radius, R, of the Sun as a function of time. $L_\odot$ and $R_\odot$ refer to contemporary values. (Courtesy of Professor Fred Hoyle, Cambridge University.)*

The most recent calculations indicate that our Sun will become a red giant in approximately 8 billion years. The red giant stage will last for several hundred millions of years. Finally, by mass ejection or nova outburst, the gigantic Sun will discard its atmospheric envelope, and rapidly become a white dwarf.

▽ One result of the evolution of our Sun through the red giant phase will very likely be the reduction of our Earth to a bleak, charred cinder. The increase in luminosity of the Sun will cause the surface temperatures of the Earth and the other planets to increase. As the swollen, distended red Sun increases in size, the oceans of the Earth will boil away. The top of the terrestrial atmosphere will become exceedingly hot, and our atmosphere will evaporate away to space. Eventually the Sun will engulf the orbits of Mercury and Venus; its surface will approach the orbit of the Earth. Whether life of any sort will survive until that remote epoch, and whether terrestrial intelligence could conceivably be equal to the supreme challenges of those times, are questions worth pondering. △

7

Supernovae

If the radiance of a thousand suns were to burst into the sky that would be the splendor of the Mighty One.

Bhagavad Gita

In the previous chapter we discussed the evolution of a normal star from its origin as a condensing cloud of gas and dust to its old age as a super-dense, cold, black dwarf. All stars, however, do not pass through these normal stages of development. Certain stars, at definite periods during their evolution, explode, creating a brilliant display of cosmic pyrotechnics, called supernovae.

There is no cataclysm of individual stars which is larger or more magnificent than the supernova. After the explosion, the stellar luminosity may increase 100 million times; for a short period, one supernova may radiate more light than a billion stars. Cases are known where the brightness of a supernova surpasses the brightness of the entire galaxy which contains it.

▽ The spectra of supernovae show that, compared with ordinary stars, they contain a relatively small amount of hydrogen and a relatively large amount of helium, iron, and other heavy elements. Because it is thought that older evolved stars have transmuted their hydrogen into heavier elements, the spectra support the hypothesis that supernovae are one cause, more violent than most, of the death of a star. △

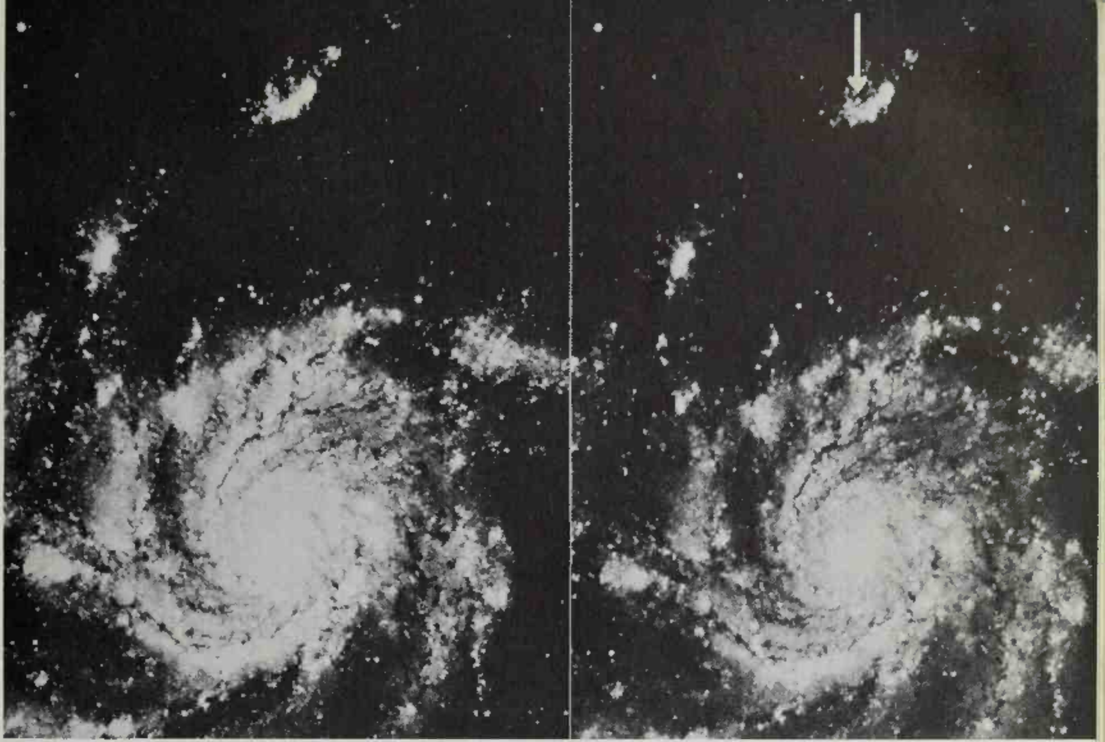
Supernovae occur infrequently. In large stellar systems such as the Milky Way, there is only about one explosion each century. As a result, astronomers are much more likely to observe such phenomena in the other galaxies. If we systematically search several hundred galaxies during a period of one year, it is highly probable that we will discover at least one supernova. This is a more expedient observational technique than waiting for an explosion to occur in our own Galaxy.

▽ The appearance of a supernova in the spiral galaxy NGC 4725 may be seen in Figure 7–1. The top photograph was taken on May 10, 1940, when the supernova exceeded in brightness all other regions of the spiral arms of this galaxy, except for the galactic nucleus; the bottom photograph was taken on January 2, 1941, when no supernova was evident. Figure 7–2 shows a less spectacular but more common variety of supernova; this occurred in the nearer galaxy M101. Here too, we have a "before" and "after" sequence, with the arrow indicating the supernova. △

Despite the infrequency of supernovae in the Milky Way, a number have been recorded in historical times. On July 4 (▽ sic! △), 1054, a "guest-star" appeared in the sky; it was duly reported by Chinese scholars. This star was so bright that it could be seen during the daylight hours. It surpassed Venus in luminosity; only the Moon and Sun were brighter. For several months this star was visible to the naked eye; then it gradually faded from view.

In compiling his catalogue of nebulae, Messier placed as first an object of unusual form which subsequently became known as the "Crab Nebula," or, more

FIGURE 7–1. *Top: Photograph of the galaxy NGC 4725 in the constellation Coma Berenices, taken on 10 May, 1940. A supernova explosion is indicated by the straight line. Bottom: The same galaxy photographed on 2 January, 1941. The great decline in intensity of the supernova is evident. (Courtesy of Mt. Wilson and Palomar Observatories.)*



June 9, 1950 Feb. 7, 1951

FIGURE 7–2. *Two photographs of the type Sc galaxy NGC 5457 in the constellation Ursa Major. The supernova is observed in one of the extensive spiral arms of this galaxy, which is also known as Messier 101. (Courtesy of Mt. Wilson and Palomar Observatories.)*

familiarly, "The Crab." Figure 7–3 shows a photograph of the Crab taken through a filter which passed only red light. Systematic observations indicate that the Crab Nebula is slowly expanding, as if it were unraveling in the sky. Since it is more than 1000 parsecs from us, the very fact that we can detect this increase in dimensions means that its velocity of expansion is enormous. ▽ Since we can measure its apparent rate of expansion in angular units, and since we know its distance, we can compute its true velocity of expansion. △ It has been calculated that this velocity approaches 1000 km/sec; that is, more than 100 times the velocity of an artificial Earth satellite. In contrast, the velocity of the motion of normal gaseous nebulae is usually no greater than 20 to 30 km/sec. Only an explosion of titanic proportions could have caused the observed expansion.

The Crab is located in that region of the sky where the strange stellar "guest" was observed in 1054 A.D. The velocity of expansion indicates that approximately 900 years ago the whole cloud was contained within a very small volume. It can be concluded that this nebula is indeed the residue of the gigantic cosmic cataclysm observed in China during the time of the Sung Dynasty.

The Crab Nebula has played a particularly important role in astrophysics during the last decade. As one of the nearest supernova remnants, it is more easily

FIGURE 7–3. *Photograph of Messier 1, the Crab Nebula, in the constellation Taurus, taken in red light. These turbulent expanding gases are the remnants of a supernova which exploded in our own Galaxy in A.D. 1054. (Courtesy of Mt. Wilson and Palomar Observatories.)*

investigated than others. The remains of similar stellar explosions which have briefly flared from time to time in our Galaxy are scattered throughout the sky. All, with a few exceptions, are older than the Crab. The clouds of Figure 7–4, a filamentary nebula of supernova origin in the constellation Cygnus, are estimated to be several tens of thousands of years old.

How can we distinguish normal gas clouds, the diffuse nebulae such as M 20, the Trifid Nebula [shown in Figure 7–5], from a supernova? In 1949 it was discovered that the Crab Nebula is a very powerful source of radio radiation. ▽ There are several possible sources of radio radiation. Might there be a titanic radio broadcasting station in the Crab Nebula? This seems unlikely for several reasons. The signals from the Crab Nebula are not modulated and occasionally intelligible, as the emission from radio broadcasting stations sometimes is; nor is the emission confined to one frequency, or "station." Instead, the radiation is spread over a wide range of radio frequencies. It sounds very much like "static." Any hot object will emit electromagnetic radiation at all frequencies—gamma rays, x-rays, ultraviolet light, visible light, infrared radiation, and radio waves. However, in thermal radiation, the intensity at radio frequencies should be inversely proportional to the square of the wavelength. Instead, the intensity of radiation from the

FIGURE 7–4. *A filamentary nebula in the constellation Cygnus, showing a region of strong radio emission. Photographed in red light. (Courtesy of Mt. Wilson and Palomar Observatories.)*

Crab is found to be more nearly the same at many frequencies. Hence, the emission must be non-thermal.

▽ Shklovskii was able to demonstrate that the non-thermal radio emission of the Crab Nebula can be explained as synchrotron radiation. This is radiation such as is observed in an electron synchrotron, a successor to the cyclotron. In the synchrotron, strong magnetic fields cause electrons to move at very high velocities, approaching the velocity of light. They are constrained to spiral along the magnetic lines of force. The acceleration of the electrons causes them to emit light; by adjusting the acceleration, the electrons can be made to emit light of any frequency desired, from the visible to the radio region. If we postulate appropriate magnetic fields in the Crab Nebula, the radio emission can be explained in a similar way. △ The radio radiation of the entire Galaxy can be explained in the same way. Normal nebulae such as M20 do not exhibit this intense radio emission.

▽ If the Crab Nebula is an explosion relic, with an envelope expanding at 1000 km every second, might there not have been much faster-moving particles which long ago escaped? If the radio emission is due to synchrotron radiation, then some of these fast-moving charged particles must also sometimes escape from the Crab. It is possible that supernovae are a source of cosmic rays [see Chapter 5]? △

FIGURE 7–5. *Messier 20, the Triffid Nebula, in the constellation Sagittarius. This is an excellent example of a cosmic projective test. The object to the right is the diffraction pattern of a nearby star. (Courtesy of Mt. Wilson and Palomar Observatories.)*

FIGURE 7–6. *A photograph in red light of the region of the intense radio source Cassiopeia A. Some faint wisps of nebular material can be seen. (Courtesy of Mt. Wilson and Palomar Observatories.)*

The total amount of cosmic rays emitted by the Crab Nebula can be estimated by using the theory of the synchrotron radiation of fast-moving electrons, the measured flux of radio radiation, and the known distance and dimensions of the Crab. Then, taking into consideration the estimated frequency of occurrence of supernovae in our Galaxy, we find that the total cosmic radiation emitted by supernovae is sufficient to account for the cosmic ray intensity observed on Earth. Thus, the evidence seems to indicate that supernovae are the main source of cosmic rays in our Galaxy. In addition, these explosions enrich interstellar space with the heavier elements. This is very significant both for the evolution of the stars and for the Galaxy as a whole, as we shall see in the next chapter.

The Crab possesses another remarkable characteristic. As I pointed out in 1953, its visual radiation—at least 95 percent of it—is also produced by high energy electrons as synchrotron emission. The energy of the electrons which radiate in visual wavelengths is one hundred times greater than the energy of the electrons which radiate in radio wavelengths. Their energy reaches 0.1 to 1 erg/electron. On the basis of my then new explanation of the optical radiation of the Crab Nebula, I predicted that this radiation would be polarized. Soviet and American observations have since confirmed this deduction. Subsequently, synchrotron optical radiation was detected from other objects, mainly the radio galaxies.

All supernova remnants, without exception, are powerful sources of radio radiation. One nebula found in the constellation Cassiopeia has a meter wavelength radiation flux which is ten times that of the Crab Nebula, although it is almost three times as far away. However, this supernova remnant is a very weak source of optical radiation [see Figure 7–6]. It has been calculated that the Cassiopeia explosion occurred approximately 300 years ago. The exploding star was not noticed at the time because it was embedded in dense clouds of interstellar dust.

The amount of radiation now emitted by supernovae which occurred even 10,000 years ago differs greatly from that now emitted by more recent explosions. The nebula in Cygnus [Figure 7–4] is a radio source ten times less powerful than the Crab.

The last observed supernova in the Milky Way—recorded by Johannes Kepler—took place in 1604, before the telescope had been invented, or spectral analysis discovered. Recent data concerning the course and mechanism of these explosions have been acquired solely from observations of other stellar systems.

These data indicate that supernovae fall into two categories, Types I and II. Supernovae of Type I are older stars, with masses only slightly greater than that of our Sun; the radiation from their explosion is very great, although the mass of the gaseous cloud does not exceed several tenths of the mass of the Sun; and they are found in the elliptical and spiral galaxies. Such supernovae have a certain characteristic time for their brightness to decline, after the explosion occurs. From an examination of the Chinese records of 1054, we can conclude that the Crab Nebula was a Type I supernova.

Type II supernovae take place only in the spiral galaxies. They are initially massive, hot, young stars, usually occurring in the spiral arms, where the process of star formation is localized. A number of stars of spectral class O probably end their existence spectacularly, in explosions of this type. The mass of the gases expelled exceeds by several times the mass of our Sun. The material therefore requires a significantly longer time for its dispersal than does the less massive supernovae of Type I. The powerful radio source in the constellation Cassiopeia [Figure 7–4] is a remnant of a Type II supernova.

There are several important hypotheses attempting to explain the causes of these titanic stellar explosions. In all probability, they are due to the catastrophic, sudden release of gravitational potential energy attending the collapse of the internal layers of a star. It has been postulated that the interior, remaining after the outburst, would be an object more dense than the white dwarfs.

There is no generally accepted hypothesis to enable us to predict when a supernova will occur. The question of whether our Sun will become a supernova is of some interest for the present generation of human beings on Earth and for all future generations. Such an explosion would entirely vaporize all the planets, with the possible exceptions of Jupiter and Saturn. However, there is little cause for worry. We can quite definitely state that, because of its small mass, the Sun will never become a supernova.

▽ It is possible that some day we will be able to determine which stars are about to become supernovae. All modern theories of supernovae require very high temperatures at the center of the star—hundreds of millions of degrees or more. At these high temperatures, electrons and positrons (positively charged electrons) are colliding with each other at fantastic rates. Such collisions often result in the complete conversion of matter (the electron-positron pair), into energy (e.g. gamma rays). But it also happens that electron-positron interaction produces much less familiar particles, a neutrino-antineutrino pair. A neutrino is an elementary particle, in some respects like the photon. It has no mass and travels at the speed of light. The reason it is so unfamiliar is that the neutrino effortlessly passes through matter. The bulk of neutrinos pass through the planet Earth as easily as light through window glass. Neutrinos have been discovered only by very patient searching for their rare interaction with matter. At a temperature of several hundreds of millions of degrees, a star should give off more energy as neutrinos than as photons. Neutrinos pass through the overlying bulk of the stars into space. When neutrino telescopes become feasible, it will be possible to, in a sense, peer directly into the innermost cores of the red giants. Stars gradually evolving into supernovae will be detectable as sources of gradually increasing amounts of neutrinos, and we should be able to determine long in advance any potential supernovae in our stellar neighborhood. △

As previously stated, supernovae occur infrequently. But our Galaxy has existed for so long that there have probably been a fair number of these events since the formation of our solar system. Could a supernova have exploded relatively close to our planet during the Earth's history? To answer this question, we shall make the following simple calculations:

Let us assume that a supernova of Type II exploded somewhere in the Galaxy 100 years ago. An explosion of this type occurs only in a thin region near the Galactic plane, within a thickness, *d*, of some 100 parsecs. The Galactic orbit of the Sun is now (and has always been) within this thickness. Consider a spherical region of radius R which surrounds the Sun. Its volume will be $\frac{4}{3}\pi R.^3$ If r is the characteristic size of the spiral arms of our Galaxy, and d the thickness of the region of the Galactic plane in which a supernova of Type II can occur, then the volume of the disc in which such supernovae can occur is $\pi r^2 d$ [see the sketch in Figure 7–7]. The ratio of the volumes of these two regions, of the sphere to the disc, will be $\frac{4}{3}\pi R^3/\pi r^2 d$. The ratio of these volumes is also the probability that, when a chance supernova explosion occurs somewhere within the Galaxy, the Sun will be at a distance R or less from the explosion. ▽ We see from Figure 7–7 that R must always be less than or equal to d. Therefore, since $\frac{4}{3}\pi R^3$ is always less than $\pi r^2 d$, the probability that the Sun will be near any particular explosion is less than 1, as, of course, it should be, since a probability of 1 indicates a certainty of occurrence. △

If one supernova occurs, on the average, every T years, then a "nearby" explosion will occur once every

$$t = \frac{\pi r^2 d}{\frac{4}{3}\pi R^3} T = \frac{3r^2 d}{4R^3} T \text{ years.}$$

Let us now put numbers in our algebraic equation. Assuming that $r = 10{,}000$ parsecs, $d = 100$ parsecs, $R = 10$ parsecs, and $T = 100$ years, we find that $t = 750$ million years.

Thus, during the 4.5-billion-year history of the Earth, the Sun has several times been closer than 10 parsecs to a supernova explosion. If we believe our estimate for t to be reliable, then this has happened $(4.5 \times 10^9)/(7.5 \times 10^8) = 6$ times. It is possible that t may be smaller, because the Galactic orbit of the Sun will occasionally take it into regions where Type II supernovae occur more frequently. ▽ It might also be larger, if our estimate for the average period for recurrence of

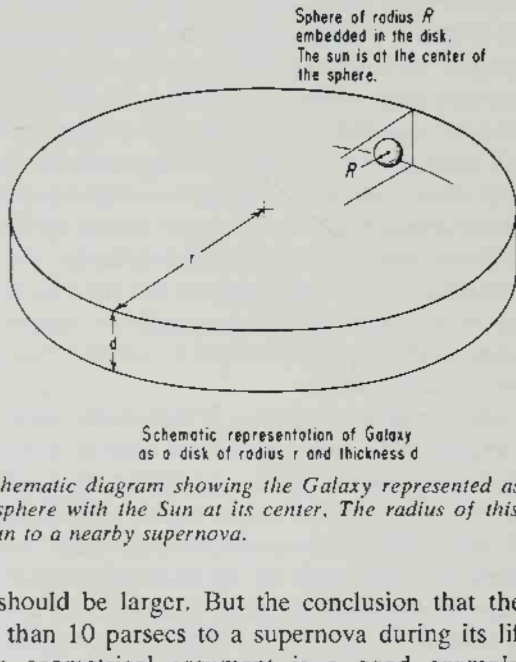


FIGURE 7–7. *Schematic diagram showing the Galaxy represented as a disk. Embedded in the disk is a small sphere with the Sun at its center. The radius of this sphere represents the distance from the Sun to a nearby supernova.*

a supernova, T, should be larger. But the conclusion that the Earth has several times been closer than 10 parsecs to a supernova during its lifetime seems rather firm. This simple geometrical argument is a good example of the power of elementary mathematical reasoning in physics and astrophysics. △

How would a nearby supernova explosion affect the Earth? If intelligent life were present, an unusually bright star would be seen in the heavens. It would be a million times brighter than Sirius (the brightest star in the sky); but 10,000 times less brilliant than the Sun. At night, the star would illuminate the countryside.

The flow of radiation in the ultraviolet region of the spectrum would be tens of times greater than the Sun's. Although this would give rise to significant ionization in the upper layers of the Earth's atmosphere, it would have no catastrophic biological consequences with our present oxygen atmosphere. ▽ The ozone in our atmosphere would absorb essentially all the ultraviolet radiation before it reached the Earth's surface. But in the Earth's earlier history [see Chapter 16], such an ultraviolet intensity increase may have had a more profound significance for living systems. △

A supernova occurring near the Earth would shine in the heavens for several months and then gradually fade. Around the star would form a nebula which, expanding rapidly at a velocity of several thousand km/sec, would cover a significant part of the sky in a few hundred years. Although the night sky would glow at the wavelengths of light characteristic for such nebulae, the fluorescence would be weak, hardly visible to the naked eye. In a thousand years, because of the gradual retarding influence of the interstellar medium, the velocity of expansion would decline. The expanding gas would reach our solar system in approximately 10,000 years. Then, during several tens of thousands of years, the Sun and its planets would be embedded inside a "radio nebula" of supernova origin.

What would happen on Earth? First, the density of primary cosmic rays would increase greatly, since radio nebulae are a source of such high-energy particles. However, cosmic rays are distributed irregularly within a radio nebula, so during one time interval—perhaps several centuries long—the intensity of cosmic radiation would be 100 times greater than at other intervals.

Such an increase in the flux of primary cosmic rays could conceivably have a serious effect on living organisms. The evolution of life on Earth is regulated by natural selection. Of the range of physical types available for a given organism, only a certain fraction is, by chance, best adapted to the environment, and reproduces its kind. ▽ The remainder eventually perish, because, for example, of competition or predation. The variety of types available for natural selection to act upon is determined by the mutation rate—that is, by the frequency of occurrence of inheritable biological changes.

▽ Mutations are caused by a variety of factors: the natural radioactivity of the soil, the waters, and the air; the cosmic ray flux; and a sizable remainder of causes, largely unknown, possibly random chemical changes in the hereditary material. All of the foregoing contribute to the "spontaneous" mutation rate and are regarded as "spontaneous" mutations. To some extent, the word "spontaneous" is a cover for our ignorance of the ultimate causes of such naturally occurring, inheritable changes. An increase in the background radiation intensity will cause an increase in the mutation rate. The majority of mutations are random, and hence deleterious. The genetic material is a finely tuned molecular instrument; a mutation is no more likely to improve its functioning than a watch is likely to work better after having been dropped from a tall building. The possibility exists, but it is unlikely. On the other hand, mutations provide the raw material on which natural selection acts. If there were no mutations, there would be no genetic inventory of possible adaptations to future environmental changes. If the mutation rate were extremely high, any characteristic selected would soon be mutated away. Thus, there is a most appropriate mutation rate for any organism. In fact, organisms exert control on their own mutation rates. There are specialized regions of the genetic material which can increase or decrease the general mutation rate, and even regions which control the mutation rate of specific characteristics. △

As a result, the biological response to an increase in the background radiation varies from organism to organism. Forms with a short reproductive cycle often require an increase of 100 to 1000 times in the background radiation for the mutation rate to double. Many long-lived forms, on the other hand, require an increase of only 3 to 10 times, for a doubling of the mutation rate.

At the present time, the average background ionizing radiation near the surface of the Earth is 0.12 roentgens per year. ▽ A roentgen is an arbitrary unit of radiation dose. △ Two-thirds of this background radiation comes from terrestrial sources, mainly from the radioactivity of the crust of the Earth. About 0.04 roentgens per year is due to primary cosmic radiation.

If the cosmic ray intensity were increased 30 times, the average radiation dose near the Earth's surface would increase by about 10 times—an increase which might have serious genetic consequences for long-lived organisms. Those organisms which are highly specialized in narrowly circumscribed environmental niches would be particularly vulnerable. For such forms, prolonged exposure to this increased dose for a period of tens of thousands of years could conceivably be catastrophic.

V. I. Krasovskii and I have suggested that the extinction of the dinosaurs at the end of the Cretaceous period in the history of the Earth, about one hundred million years ago, was caused by such an increase in the cosmic ray background. We postulated that in the epoch, the Sun was embedded in a radio nebula a distance of five to ten parsecs from a recently-exploded supernova. If the background cosmic ray intensity had increased by factors of tens or hundreds, then the increase in the dinosaur mutation rate may have caused their extinction. The lifetimes of such enormous beasts were almost certainly several centuries.

▽ One difficulty with this interesting suggestion is that it predicts the extinction of dinosaurs everywhere on the Earth during a relatively short period of time. The paleontological evidence, however, indicates that the timescale for the extinction of the dinosaurs was about 10^7 years. This is longer than the period of 10^4 years which the supernova hypothesis would suggest, unless a very long time were occupied in the reassortment by mating of deleterious recessive genes.

▽ There are in fact too many causes proposed for the extinction of the dinosaurs—not too few. One author has pointed out that the changing climatic conditions on the Earth in the middle Cretaceous period evidently eliminated a fern similar to contemporary plants which have laxative properties. In his view, the dinosaurs died of constipation.

▽ Nevertheless, there have undoubtedly been some biological effects of supernova explosions in the vicinity of the Earth during geological time, although perhaps not so spectacular as the extinction of the dinosaurs. △ At any rate, a prolonged increase in the background high-energy radiation dose would not necessarily be fatal to all living organisms. Perhaps such an exposure would be favorable for the evolution of certain life forms and the origin of some life-related substances during the early history of the Earth.

There is one other curious circumstance which may be related to supernovae. For a decade, an unexplained detail has remained in our picture of the distribution in the sky of cosmic radio noise. The intensity of the radio emission usually tends to concentrate towards the nucleus of the Milky Way, and in the Galactic plane containing the spiral arms. However, this rule does not apply to an intense tongue of radio emission extending across the sky almost perpendicular to the Milky Way. The tongue begins in a region removed from the Galactic center by approximately 30 degrees, and extends almost to the north Galactic pole, on the axis perpendicular to the Galactic plane. In Figure 7–8 we see a schematic diagram of

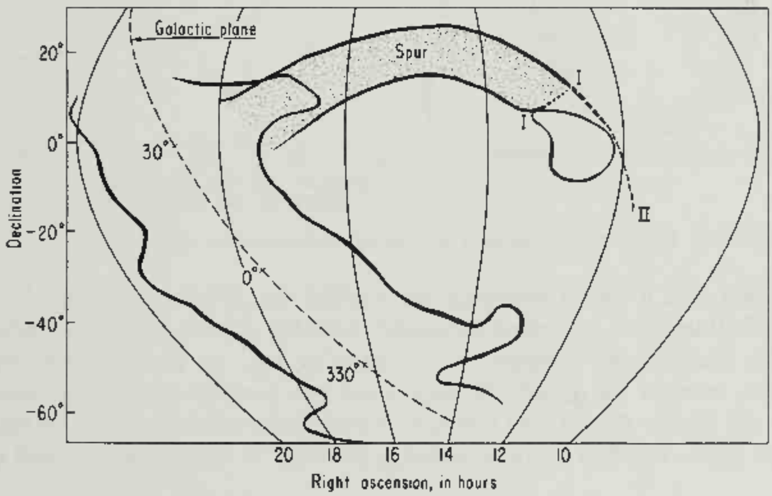


FIGURE 7–8. *Schematic diagram showing isophotes of the radio emission of the Milky Way galaxy. The region extending downward from the upper left corner is approximately in the Galactic plane. The "spur" is the shaded region extending out of the plane.*

the heavens, showing curves joining the regions of the sky where the radio brightness is the same. Such curves, called "isophotes," give a graphic representation of the distribution of the intensity of radio radiation in the sky. In this sketch, we can clearly see the concentration of intensity toward the plane of the Milky Way at Galactic latitude 0°. At the same time, it is apparent that to the left of the Galactic center the isophotes of radio radiation abruptly climb upwards. This is the unexplained tongue, or "spur."

A hypothesis of the English radio astronomer Hanbury Brown and his colleagues concerning the nature of this anomaly deserves special attention. They believe that it may be the radio envelope of a supernova which exploded very close to our solar system several tens of thousands of years ago. Since this envelope is at a distance of 30 to 40 parsecs and its linear dimensions cover 30 to 40 parsecs, it must occupy a vast part of the sky. This is shown in the diagram of Figure 7–9. However, Brown's hypothesis runs into difficulty. There are no optical traces of a supernova in this part of the heavens. Recently, in the southern part of the sky, another radio tongue was detected. The presence of two supernova remnants, both of which exploded near our Sun, during the last few tens of thousands of years seems highly improbable and suggests that these radio features have another

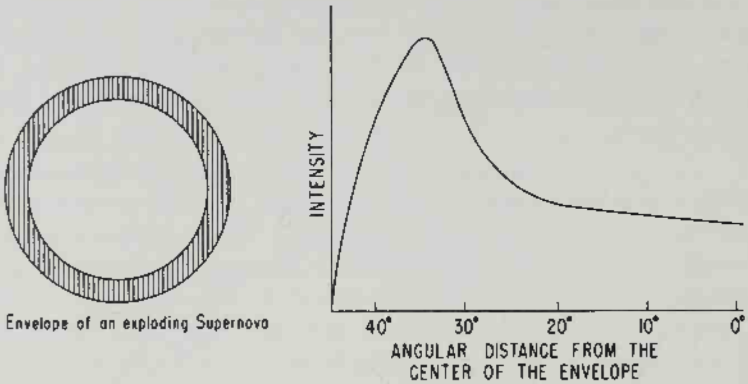


FIGURE 7–9. *Diagram illustrating a supernova explosion near the Sun.*

explanation. But if future investigations confirm the Brown hypothesis in spite of these difficulties, then in several thousand years the density of cosmic radiation in the solar system may increase by a factor of ten, as the radiation from the supernova reaches the Earth. Perhaps even the present cosmic ray intensity is abnormally higher than it was typically during the evolution of life on the Earth. We hope that a solution to this interesting problem in cosmic physics will soon be found.

8

The origin of the elements

[Atoms] move in the void and catching each other up jostle together, and some recoil in any direction that may chance, and others become entangled with one another in various degrees according to the symmetry of their shapes and sizes and positions and order, and they remain together and thus the coming into being of composite things is effected.

Simplicius (sixth century A.D.)

I believe a leaf of grass is no less than the journeywork of the stars.

Walt Whitman, *Leaves of Grass*

▽ What is the origin of matter? Were the chemical elements made together, in the same epoch, or have they evolved through time, one from another? Fifty years ago, this question would have been scientifically meaningless. The origin of the elements! Yet today, we think we understand the basic processes involved. There has been a debate between proponents of evolutionary origin and proponents of independent creation which evokes some of the flavor of the debate which followed the publication, in 1859, of Charles Darwin's *Origin of Species*. But it now appears that the origin of the elements occurs mainly in the remote interiors of red giant stars. Not only can the observed cosmic abundance of the elements be explained by this hypothesis, but also, direct evidence exists for element-building in stars. The most striking example of such direct evidence is the discovery of the element technetium in the atmospheres of giant stars.

▽ Technetium is a very unstable element. Given a lump of the longest-lived kind of technetium, half of it would have decayed to other elements in about 200,000 years. If it had been produced in the early history of the solar system, there would now be none of it on the surface of the Earth, since its lifetime is so short. This expectation is confirmed. In fact, it is just the absence of naturally-occurring technetium on the Earth, and the fact that it can only be produced synthetically, in nuclear accelerators, which has led to its name. Of course 200,000 years is much shorter than the lifetime of the star. Thus, technetium must either be made near the surface of giant stars, or made in the interior and carried up towards the surface in times not much greater than 200,000 years.

▽ Technetium is only one of the hundred or so elements which are known. Each element can be characterized by its atomic number, which is simply the number of electrons surrounding the atomic nucleus [see Chapter 4]. The sum of the number of neutrons and protons in the atomic nucleus is called the atomic weight and is written as a superscript to the right of the chemical symbol; e.g., He^4. Since the atom is electrically neutral, the number of protons in the nucleus must equal the number of surrounding electrons. Therefore, the difference between the atomic weight and the atomic number equals the number of neutrons in the nucleus.

▽ Since the chemical properties are determined only by the number of electrons, a statement of the atomic number is equivalent to a specification of the element. Thus, atomic number 1 indicates hydrogen; atomic number 6 means carbon, etc. By changing the number of neutrons, but leaving the numbers of protons and electrons fixed, we produce different *isotopes* of the same chemical element. Thus, we find C^{12}, the most abundant naturally occurring form of carbon which contains six protons and six neutrons. There is also, however, a C^{13}, which contains seven neutrons, and a C^{14}, which contains eight neutrons. C^{13} and C^{14} are naturally radioactive—that is, they have a tendency to decay spontaneously to another isotope, either of carbon or of some other element, in times short, compared with the age of the solar system. The spontaneous decay of C^{14} is widely used in radioactive dating of organic matter. From analyses of the chemical material in the crust of the Earth and in meteorites, and from astronomical spectroscopy, it has been possible to determine the cosmic distribution of a great many of the isotopes of the known chemical elements. A table of the resulting cosmic distribution of the most stable isotopes of some familiar elements was given as Table I of Chapter 4, where we saw that hydrogen and helium were by far the most abundant elements in the universe. In Chapter 6, on stellar evolution, we have seen that reactions such as

$$
\begin{array}{lll}
4H^1 & \longrightarrow He^4 & + \text{ energy} \\
3He^4 & \longrightarrow C^{12} & + \text{ energy} \\
C^{12} & + He^4 \rightarrow O^{16} & + \text{ energy} \\
O^{16} & + He^4 \rightarrow Ne^{20} & + \text{ energy} \\
Ne^{20} & + He^4 \rightarrow Mg^{24} & + \text{ energy} \\
Mg^{24} & + He^4 \rightarrow Si^{28} & + \text{ energy} \\
Si^{28} & + He^4 \rightarrow S^{32} & + \text{ energy} \\
S^{32} & + He^4 \rightarrow Ar^{36} & + \text{ energy} \\
Ar^{36} & + He^4 \rightarrow Ca^{40} & + \text{ energy}
\end{array}
$$

are the successive sources of stellar energy for more and more highly evolved stars. These successive syntheses of elements of atomic weights which are multiples of four account for the high cosmic abundances of these familiar elements. The helium nuclei, symbolized by He^4, are also called alpha particles. The temperatures in the stellar interiors are, of course, so hot that all the atoms are ionized. Successive reactions with alpha particles construct elements of higher and higher atomic weight.

▽ Can this process continue indefinitely? The answer is no. After the formation of Fe^{56}, the most abundant isotope of iron, successive reactions with alpha particles produce elements which are naturally unstable, and spontaneously decay back into iron and other elements. Some other processes are needed to account for the synthesis of elements above atomic weight 56, and the interstitial elements whose atomic weights are not multiples of four. Nevertheless, the general form of the cosmic abundance of the elements can be understood from the four-proton reaction and the successive alpha processes alone.

▽ The general shape of the cosmic distribution of the elements shows a decline in abundance with increasing atomic weight. This is entirely expected, from the alpha processes, because the higher mass number elements have to wait their turn to be synthesized; the lower mass number isotopes must be made first. One of the principal exceptions to the smooth decline in cosmic abundance with increasing atomic weight is the case of the elements near iron. Because Fe^{56} is the highest mass number isotope which can be made by alpha processes, there tends to be a piling up of nuclear products near this isotope. △

We have noted in Chapter 6 that near the end of their evolutionary lifetimes, red giants discard, in one of a variety of ways, their outer gaseous envelopes, which then gradually diffuse into interstellar space. Thus, during the evolution of a star, some fraction of its mass returns into the interstellar medium from which it arose. Subsequent generations of stars, forming anew from the interstellar medium, will be composed in part from the debris of previous stellar generations. Since a major fraction of the mass of a star is not so ejected into interstellar space, but eventually ends up in the interior of a black dwarf, it is clear that, as time passes, the amount of matter in the interstellar medium declines.

▽ The sequence of nuclear reactions described above is appropriate for a star initially formed only of hydrogen. Alpha processes dominate its subsequent chemical evolution, and when it passes through the red giant stage, it will eject into the interstellar medium such newly formed isotopes as C^{12}, O^{16}, etc. Thus, if the galaxy had originally been composed only of hydrogen, the chemical composition of the interstellar medium would, in time, become gradually enriched with heavier elements. If, now, a second generation star is formed in the interstellar medium, it will have, in addition to hydrogen, smaller amounts of helium, C^{12}, O^{16}, and so forth. The presence of these isotopes, even in small amounts, during the hydrogen-burning phase on the main sequence, leads to nuclear reactions which are impossible in a star formed of pure hydrogen. Protons are added to C^{12}, and, after some intermediate reactions, successively produce C^{13}, C^{14}, and N^{15}. Ne^{20} successively forms Ne^{21}, Ne^{22}, and Na^{23}. In a similar way, the bulk of the interstitial isotopes can be synthesized.

▽ The synthesis of elements of higher mass number than iron is believed to occur by neutron capture. Neutrons are produced in stellar interiors by such reactions as:

$$Ne^{21} + He^{4} \rightarrow Mg^{24} + \text{one neutron.}$$

Successive neutron capture is capable of synthesizing heavy elements up to Bi^{209}. Neutron capture by bismuth and heavier elements ordinarily does not lead to the synthesis of more complex elements, because the newly-synthesized elements are radioactive and spontaneously decay. But if the supply of neutrons were very great, the construction of the heavy elements could occur before they had a chance to decay spontaneously. Such a high neutron flux is believed to be available only during a supernova explosion. The existence of the elements gold and uranium on the Earth provides strong evidence that the material from which the Earth was formed once passed through a supernova. The influence of these two supernova products, gold and uranium, on the recent history of mankind is striking. Possibly, there are other planets in the Galaxy, formed in regions of the Galaxy where supernova explosions are few. Are their inhabitants happier for having no gold and uranium? △

If it is true that the elements beyond bismuth are constructed only in supernovae, then the abundance of such elements must tell us something about the rate at which supernovae occur. We saw in Chapter 7 that supernovae of Type II occur in young, massive stars. The rate of formation of such stars is strongly dependent on the density of the interstellar medium. There are some reasons for believing that this rate is proportional to the cube of the density. Thus, in the early history of the Milky Way, when the interstellar gas density was considerably greater than at present, and the rate of star formation was much higher, supernovae must have exploded at a much greater frequency than occurs today. Calculations suggest that when our Galaxy was less than one billion years old, the frequency of supernova outbursts was approximately 100 times greater than today. ▽ This corresponds to an interstellar density only a little less than 5 times greater than the present value, since the cube root of 100 is 4.7. The present rate of supernova explosion in our Galaxy is about one per century, and therefore, about 10^7 per billion years. If the rate in the first billion years was 100 times greater than this, then △ we can conclude that since the birth of the Milky Way, approximately one billion supernovae have occurred. This number completely accounts for the observed content of elements heavier than bismuth in the Galaxy. ▽ Note, however, that this computation implicitly assumes that the mechanism of supernova outburst is not dependent upon the presence of heavy elements. △

The oldest stars in our Galaxy are the sub-dwarfs, and those stars in globular clusters which have a mass of less than 1.2 times that of the Sun. More nearly than any other stars, we expect them to preserve, at least in their outer layers, the original distribution of the elements characteristic of the medium from which they were formed. And indeed, it is found that the content of heavy elements in such old stars is some tens of times less than that in the Sun. The fact that the stars of the main sequence are much richer in heavy elements than are the sub-dwarfs can thus be explained by the continuous enrichment of the interstellar medium by ejected stellar material.

▽ From spectroscopic observations of the Sun, it has become clear that our Sun is not a first generation star, and probably not even a second generation star—that is, many of the atoms which constitute the Sun have, in the past, been in the insides of other stars, stars now long since decayed into white dwarfs. And the atoms inside us—they too were cooked in the interiors of stars. Our bones are made of calcium formed by alpha processes in some red giant, billions of years ago. The same is true of the iron in our bloodstreams, the carbon, the nitrogen, and the oxygen which are constituents of all our tissues. Only the hydrogen which in our bodies is chemically bound to C, N, and O has any chance of having avoided stellar cookery. It is the most ancient of the elements, and if it was formed at all, it was formed eons ago, on the grandest imaginable scale, and by processes which even today, with our understanding of the synthesis of all the other elements, we cannot even dimly guess at.

▽ The atoms which, by spectroscopy, signal to us their presence in distant stars, are the same as their congeners here. The iron atom which we see in zeta Ursae

Majoris [Figure 4–3] is indistinguishable from the iron in the girders of a modern building. It is one of the triumphs of astronomical spectroscopy that we now know the universe to be constructed of just the same kinds of atoms as are present here on Earth.

▽ In the middle of the nineteenth century, the French philosopher Auguste Comte was seeking an example of a kind of knowledge which human beings would never achieve. Unfortunate man! He chose the chemical composition of the stars. Less than half a century later, astronomical spectroscopy was in full flourish. Negative prognostications are risky. Not only do we now know the chemical composition of the stars; we also understand how the elements were made. The atoms in zeta Ursae Majoris and in steel girders were made in precisely the same way. They have simply been put to different uses. △

9

The evolution of the galaxies

We may also draw a very important additional conclusion from the gradual dissolution of the Milky Way: for the state into which the incessant action of the clustering power has brought it at present, is a kind of chronometer that may be used to measure the time of its past and future existence; and although we do not know the rate of going of this mysterious chronometer, it is nevertheless certain, that since the breaking up of the parts of the Milky Way affords a proof that it cannot last forever, it equally bears witness that its past duration cannot be admitted to be infinite.

William Herschel, *The Construction of the Heavens* (1811)

▽ When large telescopes and long exposure times are used to gather light from very distant sources out of the plane of the Milky Way, a point is eventually reached when one can see more distant galaxies than foreground stars. There may be more galaxies within reach of the 200-inch telescope on Mount Palomar than stars in our own Galaxy; in photographing objects fainter than about the eighteenth magnitude, we can obtain a view of the universe like that of Figure 3–15. Here, we see spiral galaxies with various degrees of winding of the spiral arms; we see them face-on, edge-on, and at all intermediate inclinations. We see the fuzzy, almost featureless elliptical galaxies, and a variety of misshapen, irregular galaxies. The situation is the same in every observable direction. (Due to heavy concentration of stars and dust in our own Galaxy, we cannot see galaxies on the other side of the Galactic nucleus.)

▽ The galaxies, however, are not uniformly distributed. There are occasional close groupings, such as Stefan's Quintet seen in Figure 9–1, in which clouds of gas connect individual galaxies. There are also looser associations of galaxies; our Galaxy, M31, M33, the Magellanic Clouds, and a handful of ellipticals and irregulars form an association of nearby galaxies called the "local group."

▽ Beyond the local group are other clusters of galaxies, some much richer than the local group. The nearest of these is the Virgo cluster, approximately 40 million light years distant. Beyond 60 million light years, the number of galaxies unexpectedly declines, and there are no rich clusters for much greater distances. This has led us to suspect that clusters of galaxies may in turn be grouped into vast superclusters. Is it possible that the superclusters in turn are aggregated into clusters of clusters of clusters? Is the universe arranged into an infinite regression of such clusterings, forming a grand and endless hierarchy?

▽ Since the chance juxtaposition of galaxies in an association of this sort is very unlikely, there must be some physical connection, probably gravitational, among the members of a given cluster of galaxies. In addition, since it is very difficult for galaxies to come together by random encounters, the existence of clusters of galaxies provides some clue to the origin of galaxies. Are galaxies, like stars, formed collectively, in great associations which only subsequently dissipate? △

It is postulated that some 10 or 20 billion years ago, a vast but diffuse cloud of gas existed at a very high temperature. The chemical composition of this cloud differed substantially from the present composition of the interstellar gas. It is possible that hydrogen was the only element present, perhaps as only its constituent fundamental particles, protons and electrons. ▽ The production of all the other elements would, in this view, have occurred at later times, in the interiors of

FIGURE 9–1. *A group of galaxies in the constellation Serpens, exhibiting gaseous connecting bridges. This group is known as Stefan's Quintet. (Courtesy of Mt. Wilson and Palomar Observatories.)*

stars. △ It is believed that the various regions of the cloud gravitationally attracted one another, resulting in a contraction of the cloud and an increase in its density, just as is postulated for the origin of stars. As the density increased, collisions among protons and electrons and the resulting emission of radiation would have occurred increasingly often. As the cloud cooled, the internal gas pressure was unable to support the overlying weight of material, and contraction proceeded at an even greater tempo.

Some mathematical calculations suggest that in such a condensation the cloud would inevitably fragment into smaller masses which many astronomers believe are the forerunners of clusters of galaxies. The potential energy of the original cloud, liberated during contraction, was transformed into the kinetic energy of motion of the individual gaseous fragments. Further contraction of the individual fragments led to a secondary decomposition into even smaller gas masses, each with high random velocities. These secondary masses are believed to have been the nascent forms of individual galaxies. Each of these newly formed protogalaxies, contracting again under the influence of their gravity, fragmented into still smaller masses, which later became the globular star clusters. ▽ The fourth level fragmentation of the globular clusters led, it is believed, to the individual stars of the first generation

—that is, to those composed entirely, or almost entirely, of hydrogen. One of the many remaining questions in this hypothetical hierarchy of condensations is this: Why do the forces which led to the preceding fragmentations not also fragment individual stars into smaller units? Why are the stars stable end products in the fragmentation hierarchy? △

During this remote era of fragmentations, the relative velocities of the individual condensations were quite high, and the protogalaxies must have had an approximately spherical shape. Evidence supporting this contention can be found in the spatial distribution of stars of the first or second generations, and in the spatial distributions of the very old globular clusters. These objects form an almost spherical system about the center of our Galaxy, their numbers increasing towards the Galactic center. Their numbers do not increase towards the Galactic plane. ▽ These old stars and globular clusters must have been formed at a time when the Galaxy had not yet condensed towards the present plane of the Milky Way. As the gas and dust then contracted towards this plane, those stars which had been formed earlier were left behind. Figure 9–2 outlines the early shape of the Milky Way.

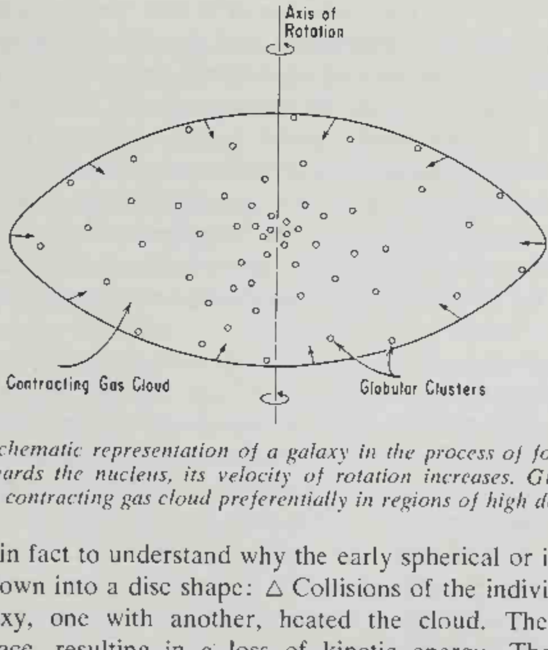


FIGURE 9–2. *Schematic representation of a galaxy in the process of formation. As the gas cloud collapses towards the nucleus, its velocity of rotation increases. Globular star clusters condense out of the contracting gas cloud preferentially in regions of high density.*

▽ It is easy in fact to understand why the early spherical or irregularly shaped galaxies settled down into a disc shape: △ Collisions of the individual gas particles in the protogalaxy, one with another, heated the cloud. The hot cloud then radiated into space, resulting in a loss of kinetic energy. The motions of the relatively cold particles were now dominated by the gravitational forces and settled down towards the Galactic plane. Consequently, stars that were formed at a later time were highly concentrated toward the disc.

▽ A protogalaxy that was rotating even very slightly when condensation began would have been rotating much faster as contraction proceeded, because of the conservation of angular momentum. The situation can be easily illustrated by taking a brick in each hand, and sitting on a piano stool with arms outstretched. Have someone spin you rapidly, then draw in your arms. The result is striking. Along the axis of rotation (with the piano stool, the legs are along the axis of rotation) contraction is easy, because there is no rotational force opposing it. Contraction in the plane perpendicular to the axis of rotation, however, is very difficult, since the centrifugal forces of rotation counteract the gravitational forces of contraction. Therefore, in the case of a protogalaxy, the net effect of gravitational attraction and rotational forces is a highly flattened rapidly rotating system, where material has contracted along the axis of rotation, but not perpendicular to it. The same circumstance explains why the planets in the solar system lie in a plane, why the satellite systems of the individual planets lie in a plane, and why the sun and all the planets are slightly flattened along their axes of rotation. △

We have outlined the present hypothesis on the origins of galaxies. ▽ All astronomers agree that many aspects of it still remain incomplete. Our present picture can be tested only by examining other galaxies. △ As noted in Chapter 3, there are large variations among different categories of galaxies. How may these differences be explained? ▽ Is there an evolution of galactic types similar to the evolution of stellar types? The character of a star, we remember, is determined primarily by its initial mass, composition, and present age. In the opinion of many astronomers, △ the characteristics of a galaxy are determined by its initial mass, its present age, and its initial velocity of rotation—that is, the rate of rotation of the protogalaxy. For example, if the initial galactic mass is relatively small, the average density of interstellar gas will be rather low, and star formation will proceed slowly. Indeed, the formation of the first generation stars in appreciable numbers may be delayed several billion years. The Magellanic Clouds may be examples of such a galaxy. These irregular galaxies have prominent hot, massive, young stars which, as shown by spectroscopic investigations, contain only small amounts of the heavy elements.

If the initial mass of the protogalaxy is large, but its rotational velocity relatively small, star formation may occur very rapidly. The interstellar medium should quickly condense into stars, and the density of the interstellar gas should soon become very low. Further star formation will occur in such galaxies only towards the galactic nucleus, where residual interstellar gas may be concentrated. Eventually, star formation ceases altogether, and the galaxies should be characterized by little gas and dust, and highly evolved stars. Elliptical galaxies [see, e.g., Figure 3–8] show such properties. If the protogalaxy is massive, and also rotates relatively rapidly, there is some expectation that spiral arms will form, and a spiral galaxy similar to our own evolve.

▽ The preceding is one scheme of galactic evolution, in which irregular galaxies, elliptical galaxies, and spiral galaxies have independent origins, and do not evolve one into another. The presentation of another scheme, such as the following by the American astronomer Allan Sandage of Mt. Wilson and Palomar Observatories, illustrates the range of uncertainty in our present knowledge of the evolution of the galaxies. Sandage points out that there is a close correspondence between the presence of dust in galaxies and the presence of the very bright young O and B stars. Thus, where dust is present, star formation must be occurring today. Such young stars are present in the irregular galaxies (Irr), the ordinary spiral galaxies (Sc) of type c, and the barred spiral galaxies (SBc) of type c. These and other spiral galaxies are illustrated in Figures 9–3 and 9–4. Moreover, whenever the spiral arms are tightly wound about the galactic nucleus, the bright blue stars cannot be seen, and dust is rare. Such galaxies are the Sa and Sb galaxies. It is reasonable to conclude that the dust in such galaxies has been depleted in previous processes of stellar formation, and that the origin of stars occurs infrequently, if at all, at the present time. If, then, the tight winding of the spiral arms is an index of a highly evolved galaxy, we may imagine the arms as winding up about the galactic nucleus as the galaxy turns. Galaxies of types Sa, SBa, S0, Sb, and all the elliptical galaxies, show no bright young stars, and an almost entire absence of dust. Finally, the integrated spectra of the billions of stars in such galaxies show characteristics of very highly evolved stars, such as the red giants. White dwarfs may exist in large numbers, but because of their low luminosity their contribution to integrated spectra of galaxies is probably negligible.

▽ The following picture of galactic evolution is suggested from these data: Protogalaxies are chaotic configurations of gas and dust, contracting from the intergalactic medium. As time progresses, there is a great initial flurry of star formation, and the galaxy is recognizable as an irregular. The galaxy then contracts towards its median plane and, by some processes not currently understood, forms open, loosely trailing spiral arms, in which dust is concentrated and star formation preferentially occurs. It may be that magnetic fields control the production of spiral arms, but the details of this process are, at the present time, very obscure. As successive generations of stars deplete the interstellar gas and dust, the density of the interstellar medium declines, the number of highly evolved stars increases, and the spiral arms wind closer and closer to the galactic nucleus. The galaxy evolves from Sc to Sb to Sa (or through the corresponding sequence for a barred spiral). Eventually, essentially all the gas and dust has been utilized in star formation, the spiral arms have entirely wound about the galactic nucleus, and the galaxy is characterized by old, evolved stars having disordered motions, and no longer restricted to the galactic plane. An S0, or elliptical galaxy, is produced. △

In still other discussions of galactic evolution the existence of spiral arms is associated with the presence of a galactic, magnetic field. The Soviet astrophysicist N. S. Kardashev, of the Sternberg Astronomical Institute, has proposed that the magnetic field of any galaxy has a metagalactic origin. As intergalactic gas contracted in the development of a protogalaxy the intergalactic magnetic field must also have been compressed, increased in strength, and then twisted by the rotation of the galaxy. Developing these ideas, the Australian astrophysicist, J. H. Piddington, has pointed out that the strength of a galactic magnetic field formed in this way depends on the angle between orientation of the metagalactic field and the axis of

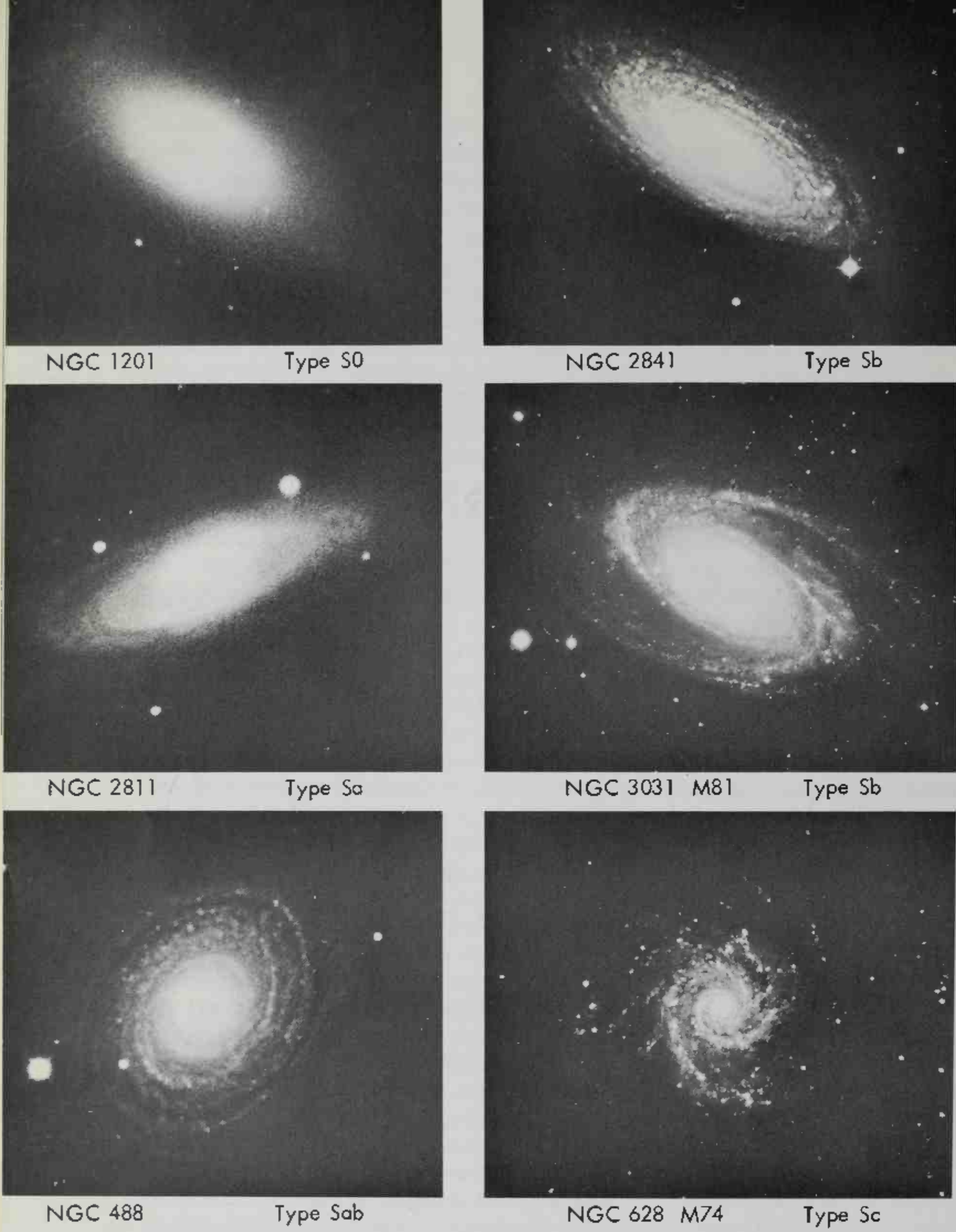
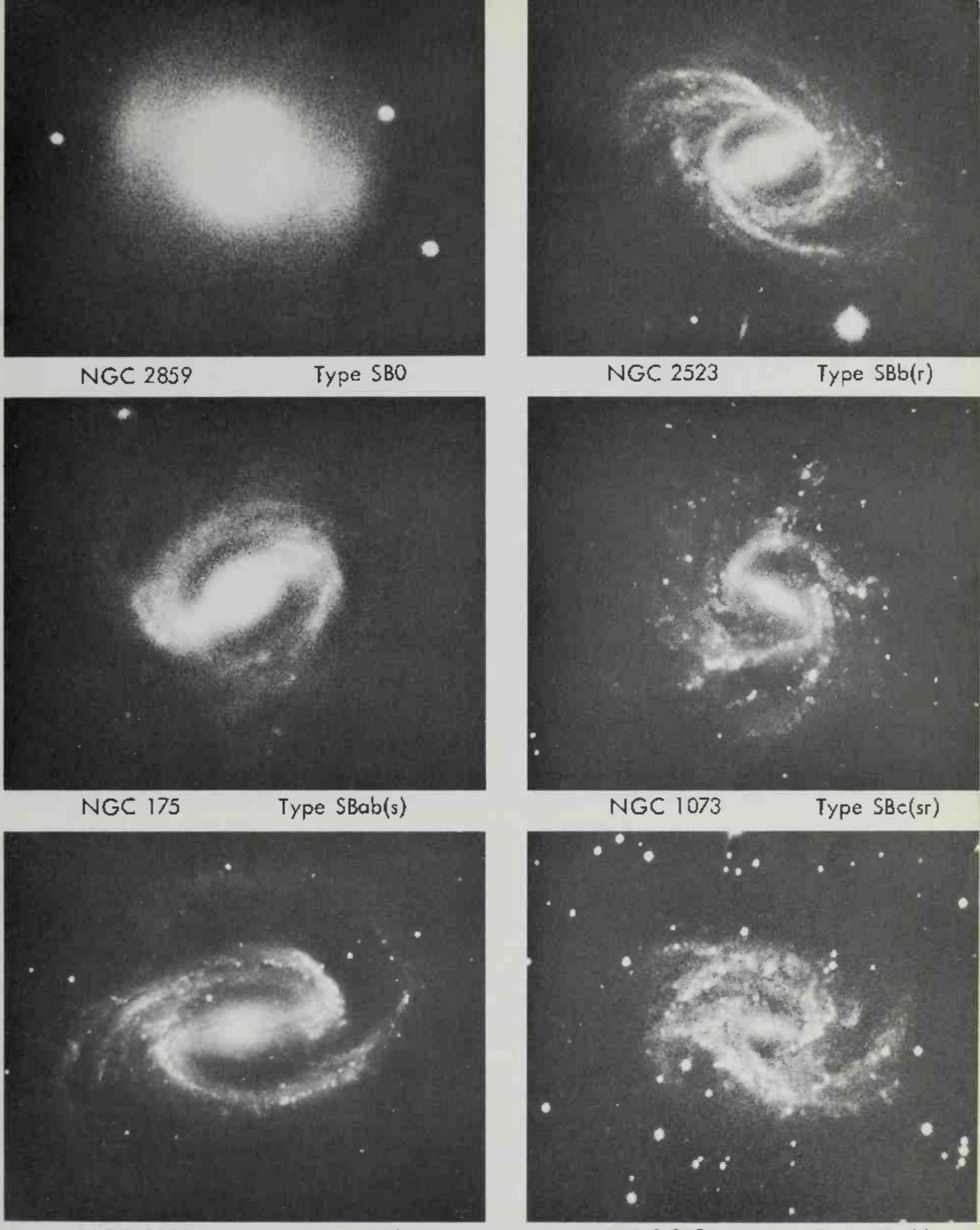


FIGURE 9–3. *Some representative normal spiral galaxies. (Courtesy of Mt. Wilson and Palomar Observatories.)*



NGC 2859 Type SB0

NGC 2523 Type SBb(r)

NGC 175 Type SBab(s)

NGC 1073 Type SBc(sr)

NGC 1300 SBb(s)

NGC 2525 Type SBc(s)

FIGURE 9–4. *Some representative barred spiral galaxies. The "bar" is the broad band of stars and gas which connects the nucleus of the galaxy with the spiral arms. Compare with Figure 9–3, which exhibits no bars.* (*Courtesy of Mt. Wilson and Palomar Observatories.*)

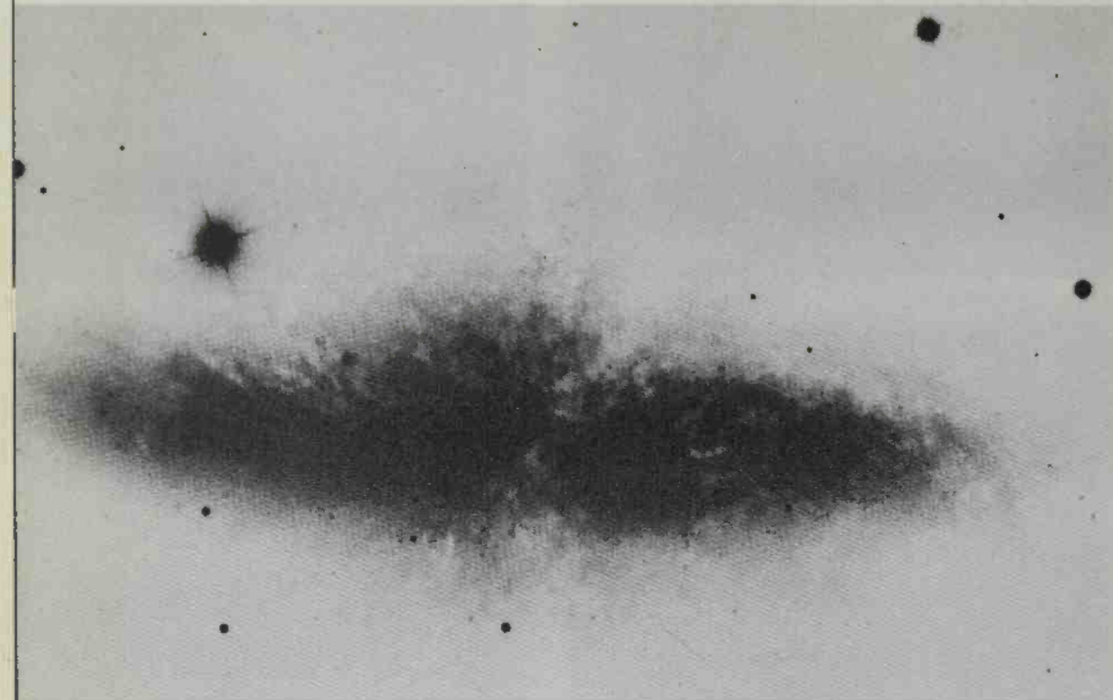


FIGURE 9–5a. *A photograph taken in blue light of the galaxy M82. This photograph is designed to bring out the overall distribution of stars, gas, and dust in this galaxy.* [*Reproduced, with permission, from C. R. Lynds and A. R. Sandage,* Astrophysical Journal *137, 1005 (1963).*]

rotation of the galaxy. For example, if this angle is small, the galactic magnetic field will be relatively weak. Even more recently, the Soviet astrophysicist S. B. Pikelner, also of the Sternberg Institute, has extended these views, and worked out an orderly theory which attempts to explain the multiple forms of the galaxies, and in particular their spiral structures.

Much more work, both observational and theoretical, is needed to resolve the conflicts among these hypotheses and to determine the validity of their common features. In addition to questions about the relationships among different galactic types, we do not fully understand such matters as the origin of spiral arms or the reason for the difference between normal and barred spirals. The important problem of the nature of the galactic nucleus is only beginning to be investigated, and the promising technique of analyzing the composite spectrum of an entire galaxy in terms of the spectra of its constituent stars is still in its infancy.

▽ There is, however, an increasing body of evidence which shows that, whatever the process, the evolution of galaxies is not a smooth one. Violent events occur in the nuclei of galaxies. We can sometimes detect such events by seeing the actual displacement of material. For example, in the galaxy M82, there is a jet of outward-moving interstellar matter leaving the galactic nucleus. The jet can be

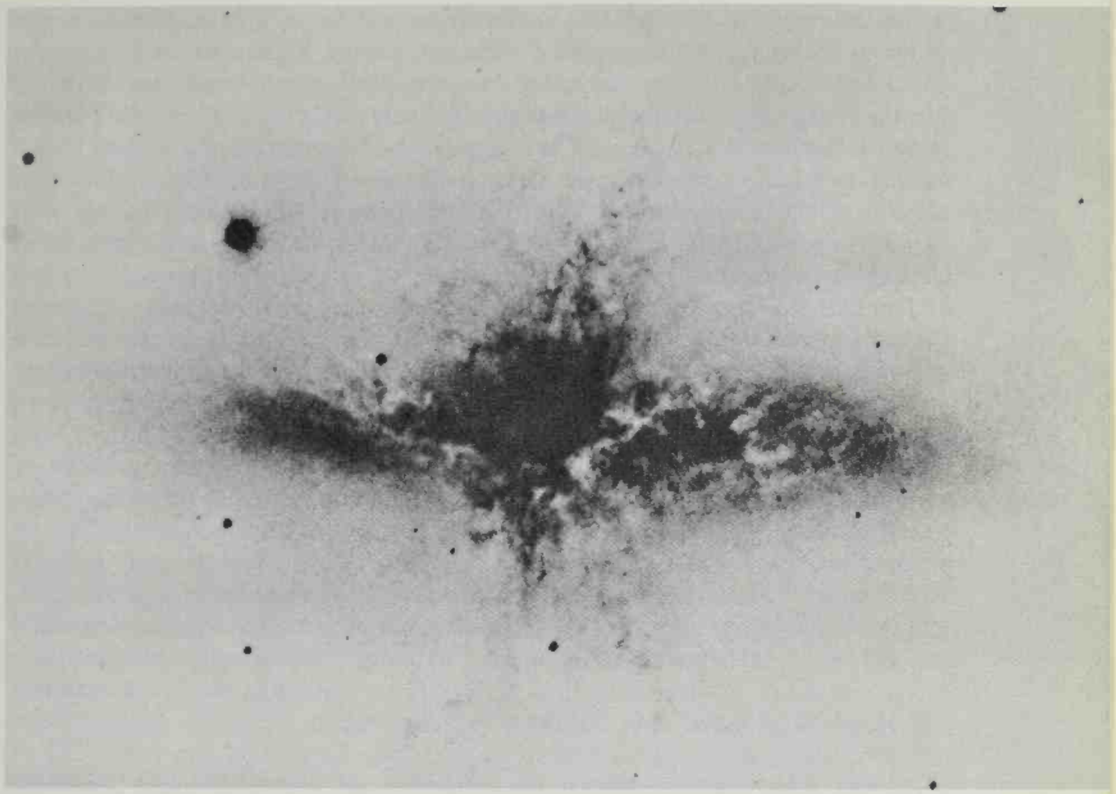


FIGURE 9–5b. *Photograph of M82, the same galaxy shown in Figure 9–5a, but here, observed in the light of hydrogen emission. The enormous quantity of material coursing upwards from the plane of the galaxy indicates the existence of a great explosion in M82. The turbulence throughout this galaxy, and the fainter plumes in other directions, suggest previous explosions. [Reproduced, with permission, from C. R. Lynds and A. R. Sandage,* Astrophysical Journal *137, 1005 (1963).*]

observed in photographs of M82 taken in light emitted by hydrogen atoms in that galaxy [see Figure 9–5]. In the case of our own Galaxy, there is evidence from radio observations in the 21-cm line of neutral hydrogen that large amounts of gas are flowing out from the center of our Galaxy in the Galactic plane. The rate of outflow is so great that, had it been continuing for the last 10^{10} years, no gas would be left in the Galactic nucleus. Thus, either the gas in the nucleus of the Galaxy is being replenished from some source, such as the Galactic halo, or the outflow of gas is a temporary occurrence.

▽ Many galaxies are radio sources. The radio emission appears to be due to synchrotron radiation, in which energetic charged particles are constrained by relatively strong magnetic fields to move in restricted regions of space. The most energetic particles produce the most energetic, highest frequency radiation; but by the same token, they lose their energy fastest. If there is no source of replenishment for these high energy particles, the lifetime of effective radio frequency radiation of a galactic radio source can be computed. One can also compute the time for the material violently leaving a galactic nucleus to be dissipated. Both of these timescales turn out to be short, compared with the lifetime of a galaxy, and are characteristically between 10^4 and 10^7 years. There is reason to suspect that the events are correlated—that is, that the violent events leading to the expulsion of large amounts of matter from a galactic nucleus also provide the high energy particles which lead to synchrotron radiation.

▽ Now, within 300 megaparsecs (3×10^8 pc) of the Sun there are about 10^6 large galaxies more or less like our own, not including the dwarf galaxies. Of these million galaxies, some 300 are known radio sources. Let us set the average lifetime for radio emission by a radio source at 3×10^6 years. If the average lifetime of a galaxy is 10^{10} years, then an average outburst lasts approximately $3 \times 10^6/10^{10} = 3 \times 10^{-4}$ of the lifetime of the galaxy. Likewise, the fraction of nearby galaxies which are currently radio sources is $300/10^6 = 3 \times 10^{-4}$. Therefore, we can conclude that every galaxy is likely to be a source of radio emission at least once during its lifetime. These violent events are observed in elliptical, spiral, and irregular galaxies. We do not know whether they play a fundamental role, or are only a passing incident, in galactic evolution. △

In recent years, observations have been made of remarkable objects which may be radio galaxies in the initial stages of formation. Through the telescope, they resemble ordinary stars, ▽ and, for many years, were thought to be just that. A typical object of this sort might be of the seventeenth magnitude—entirely unpresuming. Since they are also intense radio sources, these objects have been called "quasi-stellar" radio sources or quasars. The name itself is a confession of ignorance. When we better understand what they are, it is very likely that there will be a change in nomenclature. △

In their visible spectra, intense lines due to hydrogen, oxygen, magnesium, and other elements are observed. These lines are strongly displaced toward the red end of the spectrum. This red-shift is almost certainly due to the Doppler effect; ▽ all other attempts to explain this wavelength displacement have met with apparently insuperable difficulties. △ Their velocities can be computed from the Doppler effect. These velocities are very great, and all in a direction away from our Galaxy. From these immense velocities of recession and Hubble's law [see Chapter 10], it is concluded that the quasi-stellar radio sources are many hundreds of millions of parsecs away. One of the most distant at the time of writing, 3C-9, may be about 3 billion parsecs distant. By scaling their apparent brightness by these immense distances, using the inverse square law, it is possible to reach the remarkable conclusion that their luminosity is about 100 times greater than the luminosity of the most intense galaxy otherwise known. It seems possible that the optical radiation, as well as the radio radiation, is produced by synchrotron emission.

The most astonishing characteristic of the quasi-stellar radio sources, however, is that their radiation at visible frequencies varies with time. Individual stars of a galaxy are known to be variable, ▽ but the variation in the light output of an entire galaxy over a period of years, or even months, is entirely unprecedented. Indeed, it seems at first impossible. How could the light output of millions of stars be synchronized? The fastest velocity at which information can be transmitted is the velocity of light. If a galactic nucleus is 3,000 light years across, it takes at least 3,000 years for information to travel from one end of the nucleus to the other. Only if the actual size of the quasi-stellar radio source is less than a light year across does there seem to be any hope of understanding the luminosity variations, if the quasars are the nuclei of distant galaxies. △

If the quasi-stellar radio sources are indeed the initial stages of galactic outbursts, then the timescale of this early phase could scarcely exceed several thousand years. The radio galaxies 3C-295 and Cygnus A are older than this. Particles traveling close to the speed of light, which arose in their initial outbursts, have left the limits of these galaxies and formed two symmetrical clouds about them, which are detected by their radio emission. 3C-295 is several tens of thousands of years old; Cygnus A, several hundred thousand years old; and such radio galaxies as Centaurus A, in still later stages of development.

What could cause these explosions? The Anglo-American astrophysicist Geoffrey Burbidge believes that a chain reaction of supernovae could be the initiating factor [see Chapter 7]. Other investigators, for example, the British astronomer Fred Hoyle of Cambridge University and I, believe that the explosion of an immense, unstable starlike body, with a mass at least millions of times that of the Sun has occurred.

▽ Could an object of this mass be stable? It was believed—at least until recently—that any star with a mass exceeding about 100 times the mass of the Sun would be unstable. In fact, the British astronomer Sir Arthur Stanley Eddington once suggested that the hypothetical inhabitants of a cloud-bound planet—say, Venus—could deduce the existence and mass range of stars partly from this fact, using only physics, and no astronomy. Consider objects of successively larger masses. A one-gram object just sits there. As the masses increase, so, of course, do the sizes of the objects. Eventually, their masses become so large that the effects of gravity become important, and their interiors become compressed, because of the high internal pressures. Eventually, still increasing the mass, we reach and surpass characteristic planetary masses. Soon, due to the release of gravitational potential energy, as we construct larger and larger masses, the surface temperatures of our objects become so great that they glow, and slowly radiate energy into space. By the time the masses reach about 10^{32} gm, Eddington reasoned, these objects would be bright enough to see (in the absence of clouds), and the stars would "turn on." We now know that at slightly larger masses, thermonuclear reactions in the interior would begin, and we certainly would have a star.

▽ As the mass of the star increases towards 10^{35} gm, a new phenomenon enters the picture: radiation pressure. In reflecting off a wall, or any other object, light exerts a pressure. Ordinarily, due to the small intensity of the light, the pressure exerted is very small and is not a common feature of our everyday lives.

As the intensity of the radiation increases, the pressure becomes more and more important. In a star with a mass of about 10^{35} gm, the radiation pressure exerted by the hot interior on the cooler exterior becomes enormous, and stars of much greater mass may be blown apart by radiation pressure. In addition to demonstrating the existence of stars to sophisticated inhabitants of cloud-bound planets, this argument also makes it seem unlikely that stars with masses many millions of times that of our Sun can exist.

▽ Nevertheless, the quasars must have some energy source. The quasar 3C-273 emits a total energy supply at all frequencies of about 10^{54} ergs per year. If we take its lifetime to be that of other radio sources—say, 10^5 years—then its total energy output during its lifetime is about 10^{59} ergs. There are only two general types of energy sources which might be capable of supplying this energy: nuclear and gravitational. The Sun, which is burning its nuclear fuel with an efficiency of about one percent, has a luminosity of 4×10^{33} ergs per second. Therefore, in its lifetime of 10^{10} years it will emit approximately 4×10^{33} erg $\text{sec}^{-1} \times 10^{10}$ years $\times\ 3 \times 10^7$ sec $\text{yr}^{-1} = 10^{51}$ ergs, approximately. A quasar, working on thermonuclear energy as does the Sun, would require a mass of 10^{59} ergs/10^{51} ergs $\text{sun}^{-1} = 10^8$ suns. If a quasar were capable of complete conversion of mass to energy, by the equation $E = mc^2$, its efficiency would be 100%, and its total mass would be only 10^6 suns.

▽ On the other hand, if the energy is supplied by the gravitational collapse of a large cloud, an even larger total mass of about 10^9 suns is required. Thus, whatever the energy source, the quasar must be extremely massive; yet because of the variation in the light output of the quasar, it seems unlikely that its total dimensions can be much larger than a few light months. The quasar must be an extremely massive object in a very small volume, that is, an object with an extraordinarily high density. Because of the difficulties which we have already alluded to in understanding the stability of such an object, some astronomers feel that quasars are not intrinsically starlike and do not run on nuclear energy, but rather, run on the energy of a vast gravitational collapse. The mass required, about 10^9 suns, is less than one percent of the total mass of a typical galaxy. A reasonable place for such a collapse to occur is in the nucleus of a galaxy, where, indeed, violent events have been directly observed [see Fig. 9–5b].

▽ Perhaps, during the formation of a galaxy, much of the matter which does not condense out into dust and stars falls into the galactic nucleus. Because the total mass of the collapsing matter is so large, the pressure of the underlying material—as in the case of red giant evolution—is insufficient to support the exterior layers, and the mass continues to collapse, converting more and more gravitational potential energy into kinetic energy. Eventually, so much kinetic energy is produced that the collapse rebounds, implosion becomes explosion, and we call it a quasar. After a quasar explosion, other material in the galaxy may continue to fall into the galactic nucleus, and a series of collapses and explosions of gradually diminishing amplitude may ensue. In the particular case of the galaxy

M82 [Fig. 9–5], there is some evidence for more than one explosion. A quasar would then be merely a more violent example of the kind of explosions which have previously been observed in the nearby galaxies.

▽ The study of quasars has generated another interesting speculation. As we shall see in the next chapter, when the density of matter is very great, a light ray passing near it is deviated from a rectilinear path. As the density becomes greater, so does the curvature of the path of the light ray. When the density reaches a certain critical value, different for different masses, the light ray has curved to such an extent that it might be considered to be in a circular orbit about the dense mass. At such a density, radiation emitted by the object is unable to escape to space. Instead, it is gravitationally bound to the object. For a given mass, the radius of an object which exhibits this behavior is called the Schwarzschild radius, after the German astronomer Karl Schwarzschild, of the University of Göttingen.

▽ Now, an object smaller than the Schwarzschild radius cannot be seen, but its gravitational influence can be felt. It has been postulated that the quasars are objects of immense density which are oscillating between two radii, one larger and the other smaller than the Schwarzschild radius. Since there is no way of communicating with or receiving information from an object with a radius smaller than the Schwarzschild radius, such an oscillating object might be described as periodically entering and leaving our universe. However, it has since been shown by the Indian-American astronomer Subrahmanyan Chandrasekhar of the University of Chicago that other instabilities are experienced by a collapsing object long before the Schwarzschild radius is reached. These instabilities tend to reverse the collapse and produce an explosion. Quasars may be exotic objects, but they are not so exotic that they are smaller than the Schwarzschild radius.

▽ A slightly less exotic source of energy for the quasars has been suggested by the American astrophysicist Lyman Spitzer, of Princeton University, by Thomas Gold, and by the Dutch-American physicist, L. Woltjer, of Columbia University, among others: they invoke frequent and violent collision of stars in the nuclei of galaxies. For example, at the center of the galaxy M31 there is a nucleus with a luminosity as bright as 10^8 suns; yet, its diameter is less than 5 pc. The average star density in this and other galactic nuclei is at least 10^5 to 10^6 stars per cubic parsec. The corresponding average distance between stars in such galactic nuclei is a few hundredths of a parsec or a few thousand astronomical units. If there are occasional galactic nuclei where the star density is substantially greater, and the average distance between stars is tens of astronomical units, then collisions among the stars will occur frequently enough to provide the observed energy outputs and deduced lifetimes of the quasars. In such galactic nuclei the average distance between the stars would be comparable to the distance between the Earth and Saturn. In the version of this theory due to Spitzer, a few percent of the mass of each star is driven off in each stellar collision. This gas cools, falls towards the center of the galactic nucleus and there condenses into further stars which again lose mass in collision. This stellar collision model of quasars invokes no new physical principles, but it does require star densities in galactic nuclei which have not yet been observed.

▽ The study of the quasars is just beginning, but there is little doubt that they represent a very significant episode in galactic evolution. △

The distribution of radio galaxies may not be irrelevant to the question of life in the universe. For example, in the giant stellar system Cygnus A, the level of hard, ionizing radiation is hundreds of thousands of times more intense than on the surface of the Earth. It is unlikely that higher life forms, when exposed to such radiation intensities, could exist for long. If radio galaxies such as Cygnus A are old, it is possible that galactic explosions in the nuclei of old stellar systems can destroy the life which evolved before the outburst. ▽ Events occurring in the galaxy's interior can destroy life on millions of planets throughout the entire galaxy. △ If a smaller explosion of this type occurred in the nucleus of our own Galaxy at some past time, it apparently did not constitute an insuperable hazard to the development of life on Earth. ▽ However, no one can predict the likelihood of an explosion in the center of our Galaxy at some future time. △

10

Cosmology

Who knows for certain? Who shall here declare it?
Whence was it born, whence came creation?
The gods are later than this world's formation;
Who then can know the origins of the world?

None knows whence creation arose;
And whether he has or has not made it;
He who surveys it from the lofty skies,
Only he knows—or perhaps he knows not.

The Rig Veda, X. 129

Human thought knows no bounds. Scientists have studied the origin of elements, stars, and galaxies, as we have discussed in the preceding chapters. But what is the origin of the initial cloud of gas from which the galaxies were formed?

In this chapter we come face to face with the most momentous question of contemporary natural science: the cosmological problem. Cosmology is defined as the study of the structure and development of the universe as a whole. It goes to the heart of profound problems, both scientific and philosophical. ▽ Is the universe finite or infinite? Is it eternal, or did it have a finite beginning in time? If it was created at an instant in time, how was this accomplished? If the universe is infinitely old, is there any sense in which it has a purpose? Are the physical laws fixed, or do they alter with time? What determines the physical laws? Does the universe have the same appearance at all places and times? What is its geometry? Why do the galaxies seem to be flying apart, one from another? Is there an over-all irreversible conversion of hydrogen to the heavy elements? The cosmologists grapple with many of these problems; eventually they may be solved.

▽ The sky is dark at night. This seemingly trivial observation has profound cosmological consequences. Why should it be dark at night? Because it isn't bright. Why isn't it bright? Because there are insufficient stars close enough together to make the night sky appear bright. Consider the following geometrical argument. In Figure 10–1 we show our Galaxy, in the center of the figure, surrounded by two imaginary spherical shells of radii R_1 and R_2. The thickness, s, of each shell is much less than either R_1 or R_2. Consider the innermost shell. The inside surface has an area approximately equal to $4\pi R_1^2$. The outer surface of the innermost shell also has an area approximately equal to $4\pi R_1^2$, since the thickness is negligible. Therefore the volume of the spherical shell is $4\pi R_1^2 s$.

▽ Now, suppose that space is uniformly filled with galaxies. We represent the space density of galaxies by N, the number of galaxies in a unit volume of space—say, in a cubic megaparsec. A megaparsec is a million parsecs; a cubic megaparsec is, of course, a cube a million parsecs on a side. Let us call the average absolute luminosity per galaxy L. Thus, N, the number of galaxies per unit volume, times L, the luminosity per galaxy, is NL, the luminosity emitted by all the galaxies within a unit volume of space. The absolute luminosity of the innermost spherical shell is therefore $NL4\pi R_1^2 s$. By the same type of reasoning, the absolute luminosity of the outer shell, of radius R_2, is $NL4\pi R_2^2 s$, since we have assumed the thickness of the two shells, s, to be the same. Since R_2 is greater than R_1, the volume of the outer shell is larger than the volume of the inner shell. There are more galaxies in the outer shell, and therefore the outer shell has a greater intrinsic luminosity. But

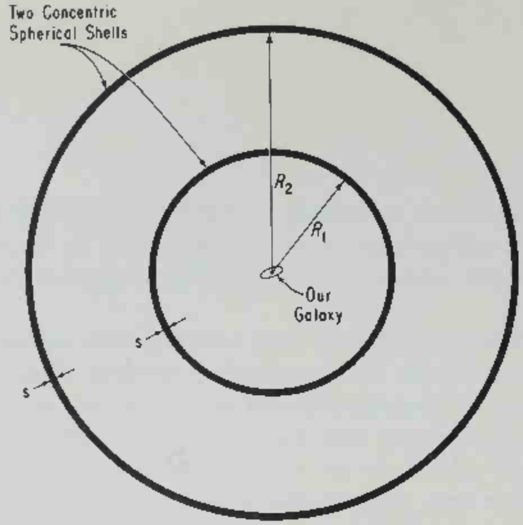


FIGURE 10–1. *Two concentric, spherical shells of equal thickness, seen in cross-section, surrounding our Galaxy. The distances from our Galaxy to these imaginary shells should be considered to be millions of parsecs, at least.*

the outer shell is further away! The light from a distant object is attenuated by the inverse square of its distance. Thus, while a distant spherical shell has a greater number of galaxies, proportional to R^2, than a nearby shell, the luminosity of each of those galaxies is reduced by the same factor R^2. Hence any spherical shell anywhere in the universe contributes the same *apparent* luminosity, as seen from the Earth.

▽ The argument is valid, of course, only insofar as our assumptions are valid; in particular, if the distribution of galaxies throughout the universe is constant, and if the luminosity per galaxy does not systematically change with distance from the Earth. If each spherical *shell* contributes an apparent luminosity $\mathcal{L}$, ten spherical shells will contribute an apparent luminosity $10\mathcal{L}$, and so forth. Thus, if the universe is infinite, so that there are an infinite number of shells, the light of the night sky should be infinitely bright. However, observation does not confirm this prediction. This contradiction between prediction and observation is known as Olber's paradox.

▽ Actually, we should not expect the night sky to be infinitely bright, even if the universe is infinite in extent, because as we successively add more and more spherical shells, a point would eventually be reached at which nearby galaxies block the light from more distant galaxies. The largest amount of light that could reach the Earth would occur when, in whatever direction we looked, our line of sight intercepted a star. We would see a sky uniformly illuminated by an immense galactic population, with no spaces between stars. The sky would be radiating to the Earth at the surface temperature of an average star—say, 5,000°K. As an equivalent, we can imagine the Earth (and our solar system and Galaxy) being placed inside an enormous oven at a temperature of 5,000°. The temperature of the Earth would soon also rise to 5,000°, and life would then be impossible. Thus, the existence of life on Earth must be due in part to the solution of Olber's paradox.

▽ Let us explore some attempted solutions to Olber's paradox. First, there is dust. Could there be enough dust and other absorbing material in intergalactic space for all the light from the most distant objects to be absorbed? This is equivalent to saying that the luminosity per galaxy, L, in the above expression is not the same for each spherical shell. But this does not solve Olber's paradox, since the dust grains would be heated by starlight, and as hot objects they would re-emit radiation. At equilibrium, they would be at the same temperatures as the stars.

▽ Alternatively, we might imagine that the space density of galaxies, N, declines with distance. As cited in Chapter 9, the galaxies are apparently distributed in a hierarchical manner; that is, the average density of galaxies in the universe does seem to decrease with distance from the Earth, if we take large enough distances. In detail, however, the hierarchical solution is not in accord with the observations.

▽ Perhaps we could add only a finite number of spherical shells and then stop. If we stop adding spherical shells soon enough, then there should be places in the sky as seen from Earth where no galaxies are visible. Thus, Olber's paradox can be solved if we postulate that the amount of matter in the universe is finite. We shall explore this possibility in more detail later in this chapter.

▽ Finally, we may solve Olber's paradox by assuming, instead of an end to space, a beginning of time. We recall that light travels with a finite propagation velocity, $c = 300{,}000$ km sec^{-1}. Thus, as we look to the most distant galaxies, we are seeing epochs further and further back into the past. If the universe began at a finite moment in time, eventually we would see at great distances to a point in space corresponding to the time of the origin of the universe. Beyond that, of course, there would be no galaxies. This is a possible solution of Olber's paradox, but to many people it is an uncomfortable one. An apocryphal story relating to the life of St. Augustine is relevant to this issue. Augustine was delivering an address on some of the same topics we are discussing in this chapter, but, of course, in the conceptual framework of his time—just as this chapter is in the conceptual framework of ours. A member of the audience objected: "Now see here, Augustine. You have told us that in the beginning God created heaven and Earth. You have also told us that God is immortal, with no beginning and no end. What, then, was God doing before he created heaven and Earth?" Augustine's riposte was: "He was creating Hell, for people who ask such questions."

▽ It is important to remember what infinity is; it is not merely a large number. There are not an infinite number of grains of sand on the beach, or an infinite number of hands of bridge. These are large but finite numbers. To illustrate how large numbers can be without coming even close to infinity, the 8-year-old nephew of the American mathematician, Edward Kasner, has called the number 10^{100} a *googol*. This can be written as a 1, followed by one hundred zeros, and is larger than the number of elementary particles in the known universe, that is, particles out to a distance of a few billion light years. A much larger number, still nowhere near infinity, is a *googolplex,* which is $10^{\text{googol}} = 10^{10^{100}}$. If it were a smaller number, a googolplex is so large that simply writing down its number of zeros in ordinary decimal notation would occupy much more than a human lifetime. Yet a googolplex is not infinity. However, in this chapter, we are concerned with whether the universe is *infinitely* large, whether it contains an *infinite* amount of matter, whether it is *infinitely* old, and whether it has an *infinite* future life expectancy.

▽ As another example of large numbers, let us calculate the number of elementary particles—protons and electrons—in the observable universe. We have mentioned that our Sun has a mass of about 2×10^{33} gm. A hydrogen atom has a mass of 1.66×10^{-24} gm. If the Sun is made only of hydrogen—not a bad approximation—it contains 10^{57} hydrogen atoms. The number of electrons is about equal. There are about 10^{11} stars in our Galaxy. Therefore, the number of protons and electrons in our Galaxy is about 2×10^{68}. There are at least 10^{9} other galaxies within range of the 200-inch telescope at Mt. Palomar, giving 2×10^{77} elementary particles. If we make a generous allowance for the amount of interstellar and intergalactic matter, and for the possibility of undiscovered galaxies out to a distance of some 10 billion light years, we find that the number of elementary particles in the observable universe is not more than about 10^{80}—a hundred million trillion times smaller than a googol.

▽ With 10^{11} stars in our Galaxy and 10^{9} other galaxies, there are at least 10^{20} stars in the universe. Most of them, as we shall see in subsequent chapters, may be accompanied by solar systems. If there are 10^{20} solar systems in the universe, and the universe is 10^{10} years old—and if, further, solar systems have formed roughly uniformly in time—then one solar system is formed every 10^{-10} yr $= 3 \times 10^{-3}$ seconds. On the average, a million solar systems are formed in the universe each hour. △

▽ We will return to Olber's paradox presently. But first, we must consider the question of the over-all geometry of the universe. The simplest and most natural assumption is that the universe is three-dimensional and Euclidean; that is, the position of an object can be specified by three coordinates, and the familiar axioms of Euclidean geometry apply to the measurement of distance. While simplicity is of heuristic value, nothing compels the universe to be simple. However, what is difficult for one generation of scientists is simple for the next.

▽ In the first and second decades of this century, Albert Einstein proposed that we live in a four-dimensional universe in which time (or rather, the velocity of light multiplied by time, to make the units consistent) is on an equal footing with the ordinary spatial coordinates. Instead of talking about points in space, we must talk about events in a four-dimensional space-time continuum. This, at least, seems quite reasonable. Of course, we cannot picture four physical dimensions—length, width, height, and something else at right angles to the other three; but mathematically, four dimensions can be dealt with almost as simply as three. If the four-dimensional space-time continuum were flat, or Euclidean, calculations in it would be especially easy. For example, if the side of a square is of length a, the area of the square is a^2. A cube with a side of length a has a volume a^3. The corresponding four-dimensional object, all of whose sides have length a, is called a *tesseract*. It has an interior "capacity" of a^4. (There is, of course, no word in everyday parlance for the four-dimensional equivalent of the three-dimensional volume.)

▽ By comparing the area of a circle, πr^2, with the volume of a sphere, $\frac{4}{3}\pi r^3$, we immediately see that geometrical excursions into non-Euclidean four-space may be complicated. Einstein proposed that in the presence of matter space-time becomes curved, and non-Euclidean geometry must apply to the motion of material objects and of light. In the General Theory of Relativity, Einstein made specific numerical predictions of, for example, the deflection of starlight on passing near the Sun (easily visible, of course, only during a total solar eclipse) and of anomalies in the motion of Mercury, the planet nearest the Sun.

▽ These brilliant predictions have been partially confirmed by observation, and most physicists agree that space-time is curved. However, the observational difficulties in performing these checks, and the fact that some of them do not test the full General Theory of Relativity, permits some room for skepticism still. A new means of testing General Relativity has been suggested by the American physicist Irwin I. Shapiro, of the Massachusetts Institute of Technology. He recommends that radar signals be transmitted to Venus or Mercury when these planets are on the other side of the Sun from the Earth. Then, the radar pulse must pass near the Sun to reach the planet. According to General Relativity, the path of the radar pulse will be deflected towards the Sun, and Shapiro concludes that the resulting delay in the echo should be about 2×10^{-4} seconds. Such an extraordinarily short time can be measured by existing equipment, and this critical test of General Relativity will probably be performed in the near future.

▽ Having successfully predicted the behavior of space-time in the vicinity of a single material object such as the Sun, Einstein went on to consider the behavior of space-time in the universe as a whole. Assuming an approximately uniform distribution of galaxies, Einstein deduced, in his initial cosmological foray, a closed universe of "positive" curvature. Its three-dimensional analog would be a sphere. Consider a two-dimensional organism walking on the inside of a hollow sphere. He would find that while there were no obstacles in his path, the amount of two-dimensional space available for his peregrinations was limited. The reason, of course, is that the two-dimensional area of the sphere, $4\pi r^2$ (where r is the radius of the sphere), is cleverly rolled through a third dimension, so that there are no edges. In an analogous manner, Einstein imagined that the three ordinary physical dimensions are curved, so that if we were able to travel over immense distances, we might set out in what we think is a straight line, and without ever turning back or meeting a boundary, find ourselves returned to the point of origin. It would be possible to circumnavigate a closed universe.

▽ In the 1920's, when the universe was thought to be closed, and the radius of curvature modest by present standards, some astronomers turned their telescope in the exact opposite direction in the sky from our nearest spiral galaxy, M31 [see Figure 3–5]. There, sure enough, was another spiral galaxy of similar form, and the exciting possibility arose that this was M31 seen from the other side. In fact, beyond M31 there should be another spiral galaxy, our own. If light traveled infinitely fast, and if telescopes had infinite resolving power, astronomers might photograph the backs of their heads by pointing the telescope out into space. Perhaps fortunately, such observations are impossible.

▽ Einstein also found, with his early cosmological models, that unless he made some other assumption, static universes were impossible. It seemed that a closed universe tended to contract, due to the gravitational attraction of the matter contained within it. Models in which the universe expanded were also found. Since at that time no one believed the universe itself to undergo any net motion, Einstein introduced, to circumvent this difficulty, a new force of cosmical repulsion. It was so weak at small distances that no one could measure it, and so large at great distances that it propped up the universe, making it stable against contraction.

▽ But almost as soon as these model universes were formulated, astronomical observations made them obsolete. In the early 1920's, the American astronomer Edwin Hubble at Mt. Wilson Observatory was able to deduce the distance of these galaxies from the Earth. They turned out not to be solar systems of our own Galaxy in the process of formation, as had previously been thought, but independent galaxies like our own, or, as they were called in those days, "island universes." (The existence of more than one universe is impossible, by definition, so we now call them galaxies.) By successively using the so-called Cepheid variables, bright O and B stars, associations of bright O and B stars, and all the stars in the galaxy together, as absolute luminosity indicators, it was possible to compare absolute and apparent luminosities, and derive the distances of even fairly remote galaxies. At about the same time, V. M. Slipher discovered that the spectral lines of the galaxies which lie beyond our local group were shifted towards the red. Interrelating these observations, Hubble discovered that the red-shift and the distance of a given galaxy from the Earth were correlated; the further the galaxy, the greater the red-shift. The only interpretation of the red-shift which has stood the test of time is the Doppler effect, discussed in Chapter 3. It must then follow that except for the nearby galaxies, whose random motions obscure the effect, the further a galaxy is from us, the faster it moves away from us. This can be expressed by the equation $V = Hr$, where V is the velocity of recession, r is the distance to the galaxy, and H is the proportionality constant known as Hubble's constant, after the discoverer of this linear relation. Current estimates for Hubble's constant lie between 75 and 100 km sec^{-1} per megaparsec—that is, for every megaparsec in the distance of a galaxy, we must add an increment of about 100 km sec^{-1} to its recessional velocity.

▽ But what does this mean? Are all the galaxies fleeing from us? Why should they be fleeing *us?* Are we at the center of the universe? The bulk of the astronomical evidence shows that there is nothing unique about our particular corner of the universe. First men discovered that the Earth was not at the center of the universe; instead of the Sun moving about the Earth, it was the other way around. Then it was found, from the distribution of the globular clusters, that our Sun was not at the center of the Galaxy, but was rather in an obscure position near the rim. Now, are we to find that our Galaxy, one of at least 10^9 others, happens to be in the center of the universe?

▽ Not if the universe as a whole is expanding. Let us use the following analogy. Imagine the universe is an unbaked raisin cake. (Worse analogies have been made.) Each raisin represents a galaxy. The cake is placed in an oven and, after a while, it rises. The volume of the cake has increased—that is, the "universe" has expanded—but, in addition, there has been an increase in distance from any raisin to any other raisin. If we were to stand on a raisin and view the other raisins, it would appear that all the other raisins were receding from us, and that the most distant raisins were moving away at the greatest speed. The same view would be obtained regardless of which raisin we stood on. In the same way, if our universe is expanding, something similar to Hubble's law should be seen by astronomers on any of the more than 10^9 galaxies.

▽ Very well then, the universe is expanding, and there is nothing special about our position in it. But it must follow that in the past the galaxies were closer together. △ If the velocity of expansion increases at a rate of 75 or 100 km sec^{-1} with each million parsecs, then by extrapolating into the past, we come to the following remarkable conclusion: Approximately 12 billion years ago, ▽ roughly, $1/H$, △ the entire universe was concentrated into an extremely small volume. Some scientists believe that, at that time, the density was in excess of 10^{14} or 10^{15} gm cm^{-3}; that is, the density of the universe exceeded the density of the atomic nucleus. The universe, in other words, was one gigantic, super-dense, nuclear "drop." For some reason, the drop became unstable and exploded. The results are now observed as the expanding universe. ▽ This cosmogony is known as the evolutionary, or "big bang" hypothesis.

▽ But it is not necessary to conclude that the universe was *created* when the galaxies were close together. Other models of the universe exist. At about the same time that the evidence for Hubble's law was accumulating, a new solution was found to the equations of general relativistic cosmology. It was discovered that open, hyperbolic universes and closed, pulsating universes were both possible. An open, hyperbolic universe is one in which the universe begins after taking an infinity to contract, then expands towards infinity, taking an infinite amount of time to reach it. Or, the universe could have been created at any moment in these infinities. In this view, the universe has a finite age, but an infinite future. The infinities here are true mathematical infinities, and not simply large numbers like googolplexes. If the hyperbolic model is correct, then at the present moment, the universe is on the expansion leg [see Figure 10–2].

▽ In the pulsating universe, the universe is always closed, in the same sense that a sphere is, but its radius of curvature varies periodically with time, as in

Figure 10–3. Here, the universe is infinitely old, and has an infinite future life expectancy. We are situated on one of the upward rises of the oscillation, as shown in Figure 10–3.

▽ Both models are consistent with general relativity. The distinction between them lies mainly in the value of the average density of the universe. If we were to take all the stellar and interstellar matter which we are certain exists in the universe and spread it uniformly in a sphere of radius some ten billion light years, we would have a density of matter of about 10^{-30} gm cm^{-3}. If this is the mean density of matter in the universe, the various parts of the universe do not exert a sufficient mutual gravitational attraction to counterbalance the high expansion velocities towards the periphery, and the universe will be hyperbolic. We are using words like "periphery" somewhat loosely, but having spent most of our lives as three-dimensional beings, we have no ready stores of four-dimensional experience to draw upon.

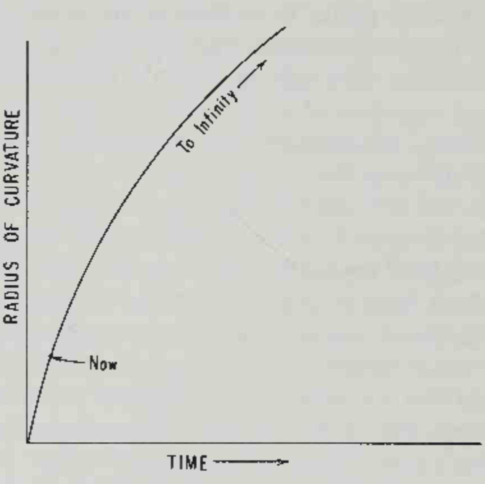


FIGURE 10–2. *Schematic representation of the change of the radius of curvature in an open, hyperbolic universe with time.*

▽ If, on the other hand, there exists some tens of times more intergalactic matter, at present undiscovered, than catalogued matter, then the mean density of the universe would be about 10^{-29} gm cm^{-3}. At these densities, the gravitational self-attraction of the universe is sufficient to resist the expansion, and the pulsating model will more nearly describe the universe as a whole.

▽ Before we go on to discuss the possibility of observational tests of these two models, let us examine some of their implications. First, both of them admit the possibility of an infinitely old universe. Many people will not grant this possibility. They are willing to admit a universe which has an infinite life expectancy, but they feel much more comfortable in a universe which was formed at a finite point in time. In the pulsating model, there is no method of determining which oscillation we are presently immersed in, because at the cusps of the curves there is a definite possibility that all matter of the previous cycle, no matter how complex or elaborate its architecture, is ground down into elementary particles, and then—like the legendary phoenix—emerges from its own ashes. The pulsating universe also bears some curious resemblances to Hindu cosmology, particularly since the period per cycle is of the order of some tens of billions of years in both systems. It certainly appears possible, although there is no compelling scientific reason for the belief, that the universe began at an arbitrary point in the pulsation sequence—say, the point marked "*P*" in Figure 10–3.

∇ In the Soviet Union, the pulsating model encounters difficulties of another sort. In the Russian edition of the present work, Professor Shklovskii writes: "The simple repetition of cycles in essence excludes the development of the universe as a whole; it therefore seems philosophically inadmissible. Further, if the universe at some time exploded and began to expand, would it not be simpler to believe that this process occurred just once?"

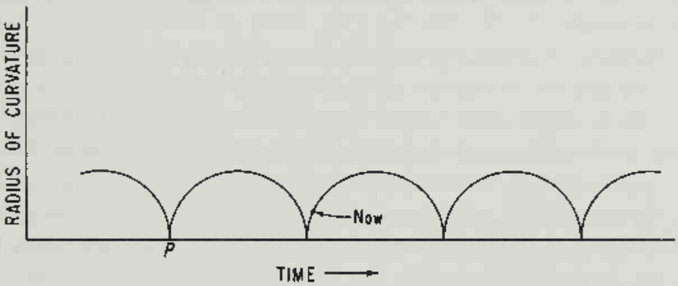


FIGURE 10–3. *Schematic representation of the change of the radius of curvature of a closed, pulsating universe with time.*

∇ In Marxist philosophy, the potentiality for self-improvement is axiomatic at all levels of study, and for all subject matters. In biology, for example, the development of Lysenkoism in the Soviet Union was in part a simple development of this precept of dialectical materialism. In modern versions of the theory of natural selection, organisms evolve one from another because the inherited characteristics of some are better fit to the environment than others; yet, the raw material for natural selection to operate upon is provided by mutation, which is essentially a random event. There is no purpose in the evolutionary process, and in an evolutionary sense organisms do not "learn" by experiences acquired during their lifetimes. Except for a few higher animals, and except for mutation, each organism starts out its life without benefit of the learning experiences of its forebears. But such a view is in apparent contradiction to the self-improvement precept of Marxist philosophy. The view that inheritable acquired characteristics are developed in particular organisms, when subjected to rigorous environments, was a position strongly maintained by Trofim Lysenko, a Soviet agronomist much in favor during the Stalin regime. This hypothesis, totally unsupported by careful experiment, has in recent years become unfashionable in the Soviet Union.

▽ An analogous involvement of dialectical materialism with cosmology is a recurrent theme in Soviet scientific literature, and is criticized in the following words by Shklovskii in the Russian edition of the present work: "The opinion that the theory of a finite universe is incompatible with the philosophy of dialectical materialism is in error. It is nonsense to connect this philosophy with any concrete property or characteristic of the universe. Dialectical materialism states that among the basic attributes of the universe are its objective reality and its knowability. Therefore, the laws of nature do not depend on the preconceived opinions of various individuals who do not comprehend the underlying spirit of dialectical materialism."

▽ A universe which grinds to ashes 50 to 100 billion years of galactic, stellar, planetary, biological, and cultural evolution is an uncomfortable one to live in, and not only for Marxists. Some religions, such as the Hindu, find no difficulties in the notion of a cyclical cosmos, but the pointlessness of such a universe is awesome. Fortunately, the determination of whether the universe is hyperbolic, pulsating, or in some other configuration is not restricted to philosophical skirmishes. There are always observations. △

Observations are the basic practical criteria for the evaluation of hypotheses in astronomy, as in any other science. It is of great interest to examine any observed phenomena which might corroborate the big-bang hypothesis. If we extrapolate back into the past the present recessional velocities of the galaxies, then, as we have seen, approximately 12 billion years ago conditions were certainly so unusual that no stars could have existed. If we could show that there are no stars with ages older than about 12 billion years, we would have a powerful argument in favor of the big-bang hypothesis.

As described in Chapter 6, studies of the main sequence turnoff point of certain globular clusters indicate that their constituent stars may be as much as 20 billion years old. However, these conclusions must be qualified. Purely theoretical considerations enter into the estimation of the ages of the stars from the turnoff point on the main sequence. Although these calculations are based on well-known principles of nuclear physics and seem to be valid for some stars, many uncertainties remain. Furthermore, the estimation of the age of the expanding universe itself is not entirely reliable. During the past decades, there have been several radical revisions of the intergalactic distance scale, and it is quite possible that errors of at least several tens of percents still exist. Thus, we can neither verify nor disprove the big-bang hypothesis on the basis of presently available information on the ages of globular clusters.

▽ Whether the universe is open and hyperbolic, or closed and pulsating, or satisfies some other cosmological model, can, in principle, be determined by other observations. One method is to search for deviations from Hubble's law. We recall that Hubble's law states that the velocity of recession of a galaxy is directly proportional to its distance. As we look to more and more distant galaxies, we are, of course, seeing them as they were in past epochs. If the rate of expansion of the universe is now accelerating, we should expect the observations of the most distant nebulae to show smaller velocities than those predicted by Hubble's law. Alternatively, if the velocity of expansion is now slowing down, we should expect the most distant galaxies, seen as they were billions of years ago, to be receding faster than predicted by the Hubble law. If very reliable estimates of the distance and velocity of the most remote galaxies could be obtained, then not only could changes in the rate of expansion of the universe be determined, but also a choice could be made among the various cosmological models. Unfortunately, the observations are very difficult to perform, and at the present time many more-or-less questionable assumptions must enter into the derivation of the distance scale. The distance to the furthest galaxies is determined by comparing their apparent magnitude with an assumed absolute magnitude typical of bright nearby galaxies. Such a calculation implicitly neglects the effects of galactic evolution, because a galaxy several billion light years away is being seen when it was several billion years younger. If such galaxies have ages of the order of 10 billion years, then we would be seeing them, say, when their ages were between 5 and 8 billion years, but not when they were less than 1 billion years old, when the most rapid evolution of stellar populations is believed to occur [see Chapter 6]. However, it is apparent that as we look to immensely distant galaxies, the effect of galactic aging will become increasingly important. Unfortunately, it is not even clear what effect differences in galactic evolution will have on the determination of the distance scale. Very early in the history of a galaxy, there should be few stars, and therefore it should appear less bright than a typical nearby mature galaxy. At a later time—perhaps during the first billion years of a galaxy's lifetime—a flurry of star formation is expected, and the average luminosity of the galaxy may be much higher than the local average. Because of such complications as aging and non-uniform rates of star formation, at the present time no unambiguous determination can be made by this method about whether the universe is open or closed, finite or infinite. Some astronomers hold that the evidence most nearly suggests a closed and finite universe, such as the oscillating model; others believe that the evidence is most consistent with an open and infinite universe, such as the hyperbolic model previously discussed.

The measurement of the angular distance between components of binary radio galaxies is another method of determining whether the universe is open or closed. It has recently been established that such duality among radio galaxies is very widespread. ▽ Figure 10–4 is a photograph of an extraordinary galaxy, NGC 5128 in the constellation Centaurus. It was once thought to be two galaxies, a spiral seen edge-on, and an elliptical, in collision. It now appears possible that this object is one more galaxy caught in the act of an explosion in its nucleus. This galaxy is also known to be an intense source of radio emission, and is more commonly known as Centaurus A. The radio emission comes predominantly from the periphery of the optical image of Figure 10–4. It is divided into two main components, one above, and the other below the dark and disrupted lanes of dust in the plane of the galaxy. △ The distance between the components of radio galaxies is characteristically about 100,000 pc. If this figure is constant throughout the

FIGURE 10–4. *The extraordinary galaxy NGC 5128 in the constellation Centaurus. As a source of intense radio emission, it is known as Centaurus A. (Courtesy of Mt. Wilson and Palomar Observatories.)*

universe, then in Euclidean space the apparent angle between the components will decrease as our distance from the radio galaxies increases. However, if space is closed, then, as we look to increasingly remote galaxies, this angle should decrease only to a definite limit. With a further increase in distance, the angle will begin to grow. No observations of this nature have as yet been carried out. They would have to be very numerous, in order to draw statistically significant conclusions from such difficult and delicate measurements.

▽ An alternative method for deciding between open and closed universes is to determine whether there exists in the universe some tens of times more matter than has been detected to date. But such an observational determination must await a theoretical prediction of the nature, temperature, and distribution of such matter. △ Weak magnetic fields are probably connected with the intergalactic medium. Their intensity could be tens or hundreds of times weaker than the intensity of interstellar magnetic fields. There is reason to believe that the intergalactic fields will rotate the polarization vector of radio radiation from the galaxies by an angle which increases with the distance to the source. (Recently, it was found that the radiation from the radio galaxies is indeed linearly polarized; the polarization is a few percent. These observations are consistent with the hypothesis that radio galaxies emit primarily because of synchrotron emission.) Although this may be a potential method for determining the density of intergalactic matter, it must be noted that this, too, is a very complex and delicate task.

▽ One further cosmological model should be mentioned, which differs in principle from the preceding evolutionary models. This is the "steady state" universe of the British astronomers Herman Bondi, Thomas Gold, and Fred Hoyle. The evolutionary models explicitly assume what is called the "cosmological principle," which states that the universe is so constructed that no observation made by any hypothetical observer anywhere in the universe can tell him where he is, in an absolute sense. The cosmological principle then predicts that the universe should look more-or-less the same to observers situated anywhere in space. An observer 10 billion light years away from us does not see fewer galaxies ahead of him than behind him because of the curvature of space; just as, for example, a two-dimensional inhabitant of the surface of a sphere finds no edge to his universe, even though its area is finite.

▽ In one formulation of the steady state theory, a "perfect cosmological principle" is introduced. It is assumed that the universe has the same over-all appearance not only to observers in all places, but also to observers at all times. That is, an observer cannot tell, in an absolute sense, what epoch he lives in by any observations made of his surroundings. Yet the galaxies are receding, and we would imagine that the average density of matter in any volume of space should decrease with time, so that an observational determination of the density of matter should give us an estimate of what cosmic time it is. To avoid this contradiction, proponents of the steady state hypothesis have postulated that matter is continually being created from nothing throughout space, at a rate which just exactly compensates the depletion of matter in a given volume by the recession of the galaxies. The hypothesis that matter is continually created at a low rate—so low that no present instruments could possibly detect it on Earth—certainly seems no more absurd than the hypothesis that all the matter in the universe was created from nothing, some 10 or 20 billion years ago. But there are other observational consequences of the steady state hypothesis.

▽ In the steady state models, the universe is infinitely old. Moreover, at no time is highly evolved matter—synthesized in the interiors of many generations of red giants—ground down again into a nuclear pulp, as occurs, for example, at the cusps of pulsation in some evolutionary models [see, for example, Figures 10–2 and 10–3]. Thus, *somewhere* in the universe, there should be immensely old—indeed, infinitely old—galaxies. It is difficult to predict what such galaxies would be like. Almost all the hydrogen should be depleted; the predominant nuclear species may have high atomic weights; and because all the stars are white dwarfs or black dwarfs, its intrinsic luminosity should be very low. We know of no such galaxies, but because of their low luminosity, they would be difficult to see unless, by chance, they were situated nearby. Intermediate stages in galactic evolution should also be seen. If they are being seen, we do not now know it.

▽ In the steady state theory, the reciprocal of the Hubble constant is *not* approximately the time since the galaxies were all very close together. Therefore, in the steady state theory, it would be possible to have in our Galaxy stellar ages which exceed the reciprocal of the Hubble constant. Such stars could not exist, of course, in the evolutionary cosmologies. There is some weak evidence that such very old stars may indeed exist in globular clusters of our Galaxy, as mentioned on page 82. Unfortunately, as we have seen, the uncertainties in both observation and theory do not permit a distinction between evolutionary and steady state cosmologies on these grounds.

▽ However, recent observations in radio astronomy now permit a tentative decision to be made between evolutionary and steady state cosmologies. △ Radio telescopes enable us to study radio galaxies which are separated from us by such vast distances that relativistic effects enter the interpretation of our observations. It has been found that the spatial density of radio galaxies which are several billions of light years distant from us is significantly greater than the spatial density of the relatively nearby radio galaxies. ("Nearby" is, of course, a purely relative term. For these purposes, a nearby galaxy is one which is as close as one billion light years.) This implies that in earlier epochs, the percentage of radio galaxies was significantly greater than at present. ▽ In itself, this circumstance is a contradiction of the perfect cosmological principle and is inconsistent with at least that formulation of steady state hypothesis. △

A possible cause for this non-uniform distribution of radio sources would be the presence of greater amounts of intergalactic gas in earlier times. The influx of this gas into the central regions of galaxies would have been much greater, leading to more explosions in galactic nuclei [see Chapter 9].

▽ An alternative explanation of the non-uniform distribution of radio galaxies at very large distances can be made in terms of the evolutionary cosmologies, in which we expect the galaxies to have been closer together in remote epochs than they are today. Since the distant galaxies are seen as they were in remote epochs, we would indeed expect the space density of radio galaxies to increase with distance.

▽ Finally, the deviations from Hubble's law shown by the motion of very remote galaxies also seem inconsistent with the steady state hypothesis. The evidence taken as a whole, then, appears to support the evolutionary cosmologies, although it cannot yet make a firm choice among them.

▽ We may now return to Olber's paradox. One reason that the sky is dark at night may be the absence of stars more than 20 billion years old. Then, no contribution to the light of the night sky could be made by objects more than 20 billion light years away. Another contemporary explanation of Olber's paradox arises from the recession of the galaxies. As a galaxy moves away from us, the Doppler effect decreases the energy of the light it emits towards Earth. This decrease is seen as a reddening of the light. As their speeds approach the velocity of light, the energy from more and more distant galaxies is increasingly diminished, so that ultimately a photon emitted in the visible region may be received on the Earth in the infrared, or indeed, in the microwave or radio region. Thus, as we look to distant galaxies, receding at velocities closer and closer to the velocity of light, the energy which we receive from them progressively declines, because of the Doppler effect, but also, of course, because of the law of the inverse square. No matter how sophisticated a radiation detection system we construct, we can always imagine a galaxy traveling so close to the speed of light that its energy output cannot be detected on Earth. Therefore, there is an effective cut-off to the distance of the furthest galaxy we can see. This cut-off is approximately 10 or 20 billion light years away.

▽ It may be asked if this explanation does not contradict the conservation of energy. Hypothetical inhabitants of a very distant, rapidly-receding galaxy measure a sizable radiation output directed towards the Earth. When a photon of visible light leaves that galaxy, the inhabitants measure it as a photon of visible light. Yet when that photon arrives at Earth, we find it reddened in frequency, and attribute to it a much smaller energy. Yet ultimately, our description and theirs of the energy output of the galaxy should tally. The solution to this conundrum is provided if we imagine some other observers, equally distant from the receding galaxy, but situated on the other side of it. These observers, if situated at rest with respect to us, will see the galaxy in question rushing towards them. The light emitted from the galaxy will now be shifted towards the violet, and the new set of observers will attribute to each photon more energy than is measured on the emitting galaxy. Thus, the deficit of radiation emitted backwards from the rapidly-moving galaxy is compensated by an enhancement of the energy of radiation emitted forwards. When the energy bookkeeping is completed, the energy remains conserved.

▽ In fact, however, there are no distant observers stationary with respect to us. The universe is expanding, and the cosmological principle requires that all observers arrive at something like the Hubble law. The energy degradation which the red-shift represents provides some of the energy for continued expansion of the universe, and no observers should now be witnessing a violet-shift.

▽ This consideration of the violet-shift is of some relevance when we consider that, in the pulsating universe, the universe would have been contracting some 20 billion years ago. △ This must have led to a significant increase in the short wave, "hard" radiation, due to the violet-shift of the radiation emitted by the converging galaxies. Eventually, the violet-shift would be so large that the likelihood of the origin and development of life in those times would be remote.

Thus, if the hypothesis of the pulsating universe is valid, and the epoch of the red-shift is ultimately to be replaced by the epoch of the violet-shift, we must come to the following conclusion: The origin and evolution of life in suitable regions of the universe are more probable during the times of the red-shift than during the times of the violet-shift ▽ because of the gradually increasing flux of ultraviolet light, x-rays, and gamma rays. △ Throughout the long period of expansion, life could negotiate the evolutionary journey from the lowest to the highest forms. But with the advent of the contracting phase, life would become increasingly difficult, and would eventually disappear, only to arise and develop again in the following epoch of expansion.

▽ While we do not yet know whether our universe must experience a contraction phase, we are on the verge of finding out: Does the universe expand forever, or are we trapped in a vast cycle of cosmic deaths and rebirths? △

11

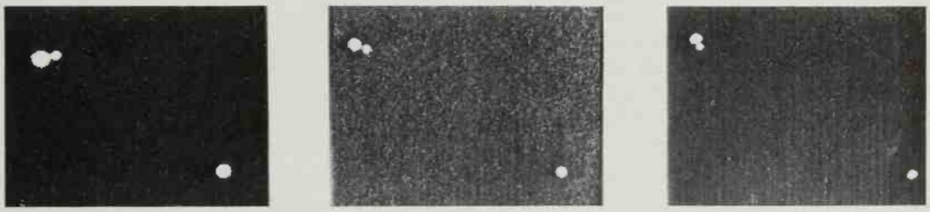
Multiple star systems

Canst thou bind the chains of the Pleiades, or loose the bands of Orion?

Job, 38:31

▽ To pursue our inquiry into the nature and distribution of life in the universe, we must now contract our field of view from the grand cosmological panorama to the seemingly commonplace study of the individual stars. If the stars move in isolated splendor, unaccompanied by planetary systems, we can expect few occurrences of life in the reaches of cosmic space. In the next three chapters, we consider the observational and the theoretical evidence that beyond our solar system other planets circle other stars. △

In Chapters 3 and 4, we considered certain fundamental characteristics of the stars, their diameter, luminosity, color, age, and evolution. Here, we consider still another property—their multiplicity. Many stars ▽ (at least 30 percent; perhaps more than 50 percent) △ form double, triple, or multiple systems in which the individual stars revolve about each other. But the periods of revolution may vary from several hours to thousands of years, depending on their masses and separation. Figures 11–1 and 11–2 show the successive relative positions of the



1908 1915 1920

Figure 11–1. *Photograph of the visual binary star xi Ursae Majoris, in 1908, 1915, and 1920. The relative motion of the dimmer secondary about the brighter primary is evident. (Courtesy of Yerkes Observatory.)*

individual stars of the double star system xi Ursae Majoris. In many cases, the components of a multiple star system are so close together that they cannot be resolved into separate stars. ▽ The system then appears to the eye or the photographic plate as a single star. △ In such cases, the multiplicity can often be confirmed by spectral observations. Because of the orbital motions of the stars about each other, their velocities along the line of sight are unequal and time-variable. For example, at a given moment in their mutual orbital motion, one star may be approaching us, the other receding. ▽ Half an orbital period later, approach and recession will be reversed. △ These motions lead, through the Doppler effect, to a small wavelength displacement of the spectral lines of one star with respect to the corresponding lines of the other. Because of the orbital motions, the velocities along the line of sight will vary periodically, and hence the displacements of the spectral lines will also vary periodically. By systematically observing such spectral line displacements as they vary with time, not only can the multiplicity of the system be deduced, but the basic characteristics of the orbit can be reliably calculated, and some information about the masses of the individual components derived. Such close double stars are called *spectroscopic binaries*.

▽ The components of a multiple star system revolve about each other, usually in a common orbital plane, similar to that in which we find the Sun and the planets of our solar system. The orientation of these planes is more or less random, so that some stars will revolve about each other in the plane of the Galaxy; others in a plane perpendicular to that of the Galaxy; and most, at some intermediate angle. △ Thus, only rarely does the plane of the orbit of a double or multiple star system form a small angle with the line of sight. If it does, however, it is possible to

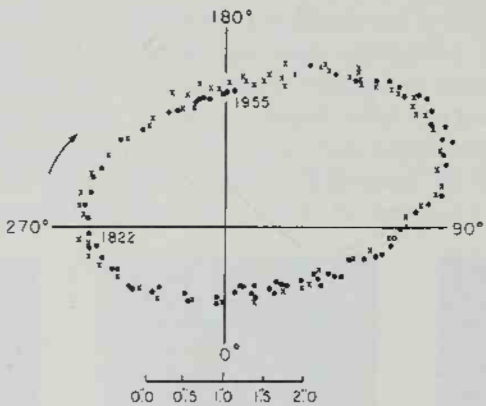


FIGURE 11–2. *The dots, circles, and crosses represent the apparent motion of the secondary of xi Ursae Majoris around the primary, in observations made between 1822 and 1955. [Reproduced from O. Struve, B. Lynds, and H. Pillans,* Elementary Astronomy *(Oxford University Press, New York, 1959). Courtesy of Oxford University Press.]*

observe an eclipse of one star by another. Since neither component can be seen separately, even by the most powerful telescopes, we observe only a periodic variation in the brightness of the double star system. At the beginning of the eclipse, the brightness declines; at its end, the star recovers its usual luminosity. (Often an eclipse lasts for several hours.) By plotting the brightness of the star against time (the so-called "brightness curve"), we can determine not only the basic parameters of the orbit, but also the diameters of the stars, and even information on the decline of the brightness of the stellar disks from their centers to their edges, ▽ or *limbs*. Since even the nearest star appears as a point of light to the most powerful telescope, such determinations of "limb-darkening" are all the more remarkable. △ A diagram of the orbit of the eclipsing variable star Algol, and the brightness curve which corresponds to it, are shown in Fig. 11–3. ▽ The Algol system, in the constellation Perseus, undergoes an eclipse approximately once every three days. When the bright component is partially eclipsed by the dimmer companion, the over-all brightness of Algol falls by more than 50 percent. This variation in luminosity is readily detectable by the naked eye and is the reason that Algol was known in ancient times as "The Demon Star." △

In both spectroscopic binaries and eclipsing binaries, the component stars are situated very close to one another. In fact, sometimes their surfaces are in physical contact. ▽ Tidal forces then draw material from one star towards the other, in complex and beautiful patterns. The material being exchanged from star to star is itself luminous. From a hypothetical planet circling such a stellar system (and such planets are possible), an observer would see two suns in his sky, perhaps of different sizes, luminosities and colors, with a brilliant and luminous ribbon of light seemingly binding them together. △ Often such multiple star systems are immersed in extensive rarefied gaseous envelopes. Diagrams of two of these close pairs are shown in Figures 11–4 and 11–5. Unfortunately, these fascinating processes are not observed visually, even through the largest telescopes, and our understanding of them is acquired only from an analysis of the spectra and brightness of the stars.

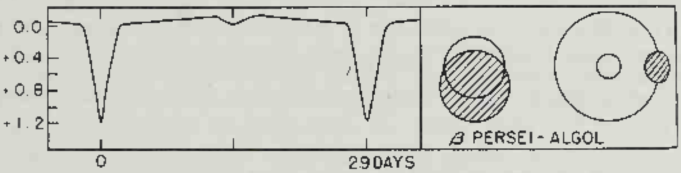


FIGURE 11–3. *Schematic representation of the light curve of an eclipsing binary star. On the left: The intensity and apparent magnitudes as plotted against time. The two major intensity minima correspond to the two configurations per orbit in which the stars eclipse each other. On the right: The relative configuration when the brighter star is eclipsed by the darker star, and the corresponding orbits. [Reproduced from O. Struve, B. Lynds, and H. Pillans,* Elementary Astronomy *(Oxford University Press, New York, 1959). Courtesy of Oxford University Press.]*

We know from Kepler's laws that the short periods of rotation belong to the very closest stars. The shortest known period, approximately 80 minutes long, belongs to the eclipsing variable WZ Sagittae.

▽ WZ Sagittae is the forty-eighth variable star discovered in the constellation Sagitta (which means "arrow," and which is to be distinguished from the nearby constellation Sagittarius, which means "archer"). The astronomical notation for variable stars in a given constellation is a tragic testimony to the deficiencies of the short-term view, and it will be profitable to tarry a moment and consider the vagaries of variable star notation. If a star which has already been named is discovered to be variable, it retains its original name. Thus, delta Cephei and Algol pose no problems. The first star in a given constellation which is recognized as a variable, and which has no proper, or Greek-letter name, is designated with the capital letter R; thus, R Sagittae. Subsequent discoveries are designated S, T, . . . , Z. This stratagem works as long as there are no more than nine new variable stars discovered per constellation. However, to the astronomers' surprise, more than nine such stars per constellation *were* discovered. Having exhausted the

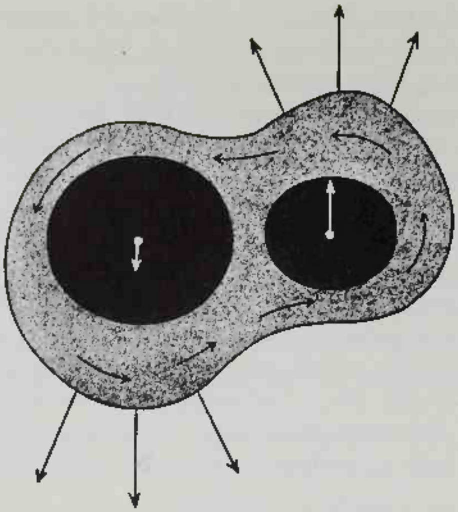


FIGURE 11–4. *Schematic representation of a close double star system, sharing a common gaseous envelope which circulates between them. The individual stars are distorted from a spherical configuration because of the mutual gravitational interaction and their rotation. [Reproduced from O. Struve,* Stellar Evolution *(Princeton University Press, Princeton, 1950). Courtesy of Princeton University Press.]*

letters R through Z, it was decided to use a double letter notation, beginning with RR, RS, RT, . . . , RZ; SS, ST, SU, . . . , SZ; and so on, until ZZ was reached, providing for the first 54 new variables discovered. But there are many stars in the skies, and as astronomical activities continued, variables beyond the ZZ's were discovered.

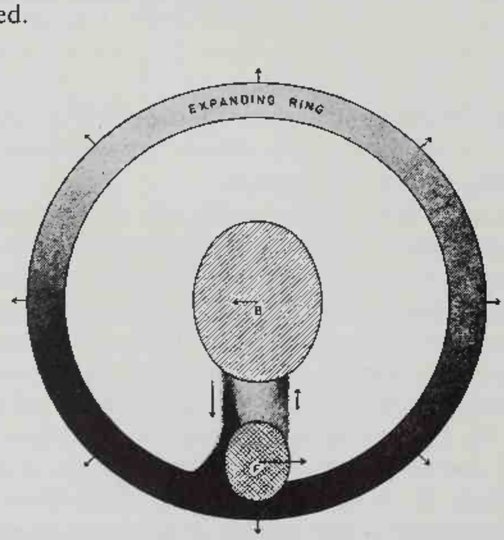


FIGURE 11–5. *Another example of an hypothesized circulating gaseous stream connecting close binary stars. [Reproduced from O. Struve,* Stellar Evolution *(Princeton University Press, Princeton, 1950). Courtesy of Princeton University Press.]*

▽ The next procedure was to begin with AA, and continue with AB, AC, . . . , AZ; BB, BC, BD, . . . , BZ, until QZ. Because of transcription problems, the letter J is omitted. This sequence stops at QZ, because the next logical doublet which has not previously been used would be RA. But since doublets beginning with R have been used earlier in the sequence, the designations RA, RB, . . . , were rejected.

▽ This elaborate tomfoolery accounts for the first 334 variable stars in a given constellation. A typical large constellation—Sagittarius, for example—may have 1700 identified variable stars. The letter notation obdurately assumed that the number of variable stars to be discovered was small. In fact, the number is very large, and the 1958 edition of the Soviet General Catalogue of Variable Stars lists 14,711 known variables in the Galaxy. To designate these, astronomers have finally taken to using the letter V to indicate a variable star, and follow it with a number, beginning with 335, to indicate its order in the discovery list; thus, V 678 Centauri. How much more sensible it would have been, had the first such variable star been designated V1, and the sequence continued *seriatum!* It is still possible, of course, to revise the notation, calling R, S, T, . . . 1, 2, 3, . . . , etc., but the letter notation has jelled, and there seems little hope for a more rational system. Such notational anachronisms, while they need not concern us further here, do not aid the popularization of astronomy. △

It now seems very possible that all novae occur in close binary systems. During a nova explosion, the luminosity of the star increases greatly in a brief period of time, although the luminosity is still a thousand times less than that of a supernova. The mass of material expelled during each nova explosion is approximately 10^{-4} to 10^{-5} the mass of our Sun. Apparently the presence of a close stellar neighbor interferes with the normal evolution of a star, particularly when it enters the red giant stage. Then, as its radius greatly increases, instabilities arise which lead to repeated—sometimes periodic—explosions.

Often, the masses of the components of a binary star system are very similar. But it sometimes happens that the mass of one component is ten or more times the mass of the other. Their relative luminosities can vary widely. For example, the faint companion of Sirius is a white dwarf. ▽ It was the first white dwarf ever discovered, and its high mass for its small radius was first derived from its orbital motion. △

Certain components of multiple star systems have such a small size that their luminosity is insignificant. These components are impossible to observe visually, even through a large telescope. But if such underluminous components are at a great distance from the primary (the brighter and more massive star), their existence and properties can be deduced. The classic example of such a system is 61Cygni, ▽ which was investigated by the American astronomer K. A. Strand and the Soviet astronomer A. N. Deutsch. The primary has its own random, so-called "proper" motion against the background of a relatively fixed field of the stars. If the primary has a massive, underluminous companion, it will have, in addition to its proper motion, a smaller, periodic motion about the center of mass of the double star system. The resulting total motion, over a period of many years, is a wavy line such as that of Figure 11–6. Observations of such periodic motions require great patience, precision, and dedication. The further away such a perturbed primary is from us, the more difficult it is to see the periodic variations in the proper motion. It is therefore not surprising that of the several invisible companions found in this way all belong to primaries which are among the nearest stars.

▽ To obtain some idea of the frequency of multiple star systems and of dark companions, consider the 12 star systems closest to the Sun. They are listed in Table III, with their spectral type and their distance from the Sun, measured in light years. The multiplicity of names arises from the astronomer's insistence on using a

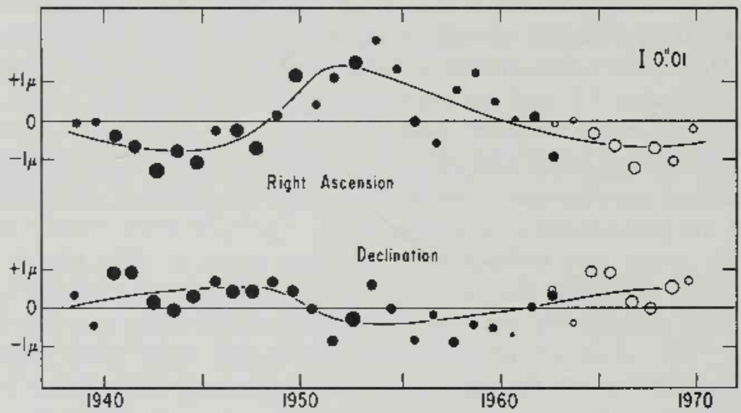


FIGURE 11–6. *The relative proper motion of Barnard's Star, as deduced by van de Kamp. Right ascension and declination are two common coordinates which are at right angles to each other, used in astronomy. The error bar in the upper right-hand corner illustrates a characteristic deviation of the actual observed points from this curve. While such observations are extremely difficult to perform, there seems little doubt about the reality of these curves. (Courtesy of Dr. Peter van de Kamp, Sproul Observatory.)*

multiplicity of catalogues. The symbol e in the spectral type column indicates stars which show emission lines; the symbol wd denotes a white dwarf. We note that the bulk of the nearest stars are M stars, of low luminosity. Of these dozen systems, at least four are multiple star systems, on the basis of visual and spectrographic evidence alone. One of them, alpha Centauri, is a triple system. The low-mass component, alpha Centauri C, is in orbit about the other two components. Since it is therefore sometimes slightly closer to the Sun than alpha Centauri A and B, it is the closest known star to the Sun, and is, accordingly, sometimes called Proxima Centauri.

▽ In addition, at least three of these twelve systems have dark companions with masses about one percent of the Sun's mass, or less. The first such companion, discovered by K. A. Strand in 1943, is associated with the 61 Cygni system, and has a mass about 0.8 percent that of the Sun. A companion having about one percent of the solar mass was discovered in 1960 by the American astronomer, Sarah Lee Lippincott, of Sproul Observatory, for the star Lalande 21185. In 1964, Peter van de Kamp, also of Sproul Observatory, found an exceedingly interesting companion to Barnard's Star, the second nearest system, only six light years away. Barnard's Star is a faint red dwarf, discovered in 1916 and named after its discoverer; it has the largest proper motion of any known star. The results of 25 years' careful observations by van de Kamp are shown in Figure

TABLE III

THE NEAREST STARS

Number	*Star system*	*Component*	*Spectral type*	*Distance from Sun (Light years)*
0	The Sun		G 0	0
1	alpha Centauri	A	G 0	4.3
		B	K 5	4.3
		C	M 5e	4.3
2	Barnard's Star		M 5	6.0
3	Wolf 359		M 6e	7.7
4	Luyten 726–8	A	M 6e	7.9
		B	M 6e	7.9
5	Lalande 21185		M 2	8.2
6	Sirius	A	A 0	8.7
		B	wd	8.7
7	Ross 154		M 5e	9.3
8	Ross 248		M 6e	10.3
9	epsilon Eridani		K 2	10.8
10	Ross 128		M 5	10.9
11	61 Cygni	A	K 6	11.1
		B	M 0	11.1
12	Luyten 789–6		M 6	11.2

11–6. The dots are the observational points, and the solid line is the periodic curve which fits them. The dark companion inferred from these observations orbits Barnard's Star once every 24 years, and has a mass about 0.15 percent of the Sun. This is a mass only 50 percent greater than that of Jupiter, which orbits our Sun once every 11.9 years. The companions of Lalande 21185 and 61 Cygni are sufficiently massive that they may be shining dimly by their own light, and therefore may be classified as very subluminous stars, rather than as very massive planets. But the companion of Barnard's Star is almost certainly a planet. This as yet unnamed world is the first planet to be discovered since the American astronomer Clyde Tombaugh found Pluto in 1930. But it has the additional and unique distinction of being the first planetary companion, discovered with a fair degree of confidence, of another sun. Van de Kamp's planet may not be the only companion to Barnard's Star, but less massive planets would be much more difficult to detect. Similarly, there may be further planetary companions to Lalande 21185, to 61 Cygni, and indeed, to all the other stars listed in Table III. The more distant the star, the more difficult it is to detect such low-mass planets.

▽ The sample of stars presented in Table III is a very small fraction of the total number of stars in our Galaxy, but it is the only sample in which we can make even a preliminary search for dark companions. We will consider it typical of similar regions elsewhere in the Galaxy. Of these thirteen systems, including the Sun, at least two have planets—the Sun and Barnard's Star. But, from our vantage point, these are two of the three closest systems. These data suggest that at least 10 percent, and perhaps more than 50 percent of the stars are accompanied by planetary systems. △

The difference between giant planets and dark companion stars is not absolute. Both consist mainly of hydrogen and helium. They are primarily gaseous spheres, held together by the force of gravity. If the mass of Jupiter were five times larger, the temperature in its central regions would increase, and it would begin to radiate, although feebly, in the visible part of the spectrum. Jupiter would then become a very dim red dwarf, with a surface temperature of approximately 1000 to 2000°K.

▽ To have a better idea of the distribution of planets throughout nearby space, we should like to extend the search for companions of Jovian mass and smaller to other star systems. This problem will be discussed in theoretical terms in the following two chapters. △ Here, we consider possible future astronomical techniques which may be used to extend our knowledge of planetary companions and nearby stars. We limit our discussion to the detection of large planets because if we cannot observe *them*, searches for the smaller, and perhaps more interesting, planets would be fruitless.

Let us assume that at a distance of 10 parsecs—about 33 light years—there is a star similar to the Sun. A large planet rotates about our hypothetical star at a distance of 5.2 astronomical units (A.U.), the same distance that Jupiter is from our Sun. Assume further that this planet has the same size and mass as Jupiter, and also that ▽ (lucky for us) △ this planet is in almost exactly the same orbital plane as is the Earth.

In principle, there are three methods which can be used to detect our planet. The first is by the periodic perturbations in the star's proper motion, ▽ just the technique used to discover van de Kamp's planet. △ The period of this perturbation would be equal to the period of revolution of the planet—in our case, 11.9 years. This periodic motion is again due to the fact that the star, influenced by the gravitational field of the planet, moves along an elliptical orbit about the center of gravity of the star-planet system. This orbital motion is superposed on the stellar proper motion. Since the mass of the star is about 1000 times the mass of the planet, the center of gravity of the system lies close to the star. △ Thus, the amplitude of the periodic motion would be very small.

Calculations carried out by the Russian-American astronomer Otto Struve indicate that the amplitude of this wave for a star 10 pc distant, superposed on the proper motion, would be smaller than 0.0005 seconds of arc per year. This is an extremely small angle [see Chapter 3], and lies far beyond the limits of accuracy of present astronomical techniques. However, if the mass were 10 to 20 times greater than the mass of Jupiter, such oscillations could be detected—although with some difficulty.

Another method of detecting the presence of such planets is by spectroscopy. The perturbations of the planetary companion produce periodic variations in the velocity of the star along the line of sight. It is easy to see that sometimes the star will be pulled slightly towards us, and at other times slightly away from us. The oscillation period of the radial velocities will again be equal to the period of revolution of the planet, but this effect is also very small. Otto Struve showed that the periodic variations in the radial velocity ▽ (that is, the velocity towards us or away from us) △ would not exceed about 10 meters per second. This is roughly 10^{-3} of the total radial velocity of an average star due to its proper motion. Through the Doppler effect, velocities of 10 meters per second correspond to a displacement of a spectral line by approximately 0.0001 Å. ▽ (We recall that 1 Å = 10^{-8} cm, and that the wavelength of visible light falls between 4000 and 7000 Å.) △ Such small variations in wavelength cannot be measured at present, especially if we consider that spectral lines are not infinitely narrow, but have a finite width of the order of 0.1 Å.

A third method of detecting planetary systems about nearby stars is the photometric method—that is, the systematic measurement of the brightness of a star. We have assumed that the orbit of our hypothetical planet is in the orbital plane of the Earth ▽—a circumstance which will in fact occur only fortuitously. △ Hence, periodically each 11.9 years, the planet will be projected onto the disk of the star as it passes in front of it. Similar phenomena are observed in our solar system —for example, when the planets Venus and Mercury transit the disk of the Sun. Since the planet is dark, and not self-luminous, when it passes in front of the star, the brightness of the star (or more exactly, the flux of radiation from it) will appear slightly diminished, as measured on Earth. The phenomenon is quite analogous to that already described for eclipsing binaries [see Figure 11–3].

Calculations indicate that if a planet the size of Jupiter passed in front of a star similar to the Sun, the star's luminosity would decrease by about 0.01 stellar magnitudes. Such a small variation in the radiation flux can be recorded by existing photomultipliers. We must recall, however, that we are considering a very idealized case, in which such an eclipse actually occurs, as seen from the Earth. An inclination of the orbital plane of the hypothetical planet by only half a second of arc would cause the planet not to pass in front of the star at all, as seen from the Earth. ▽ Thus, while the method works in principle, it works in practice only if we are very lucky, because the planes of planetary orbits should have a random distribution. Only if we observe a large number of stars will we be fortunate enough to detect a planetary companion by such a method. △

These three methods for the discovery of extrasolar planetary systems are summarized in Figure 11–7.

▽ Many of the difficulties in implementing these methods are due to our unimaginative choice of the Earth as an observing station. △ The large telescopes on Earth cannot operate close to their theoretical capabilities because of atmospheric turbulence, or what the astronomer calls "poor seeing." Even a brilliant point source of light—for our purposes, a star—is smeared into a disk with a

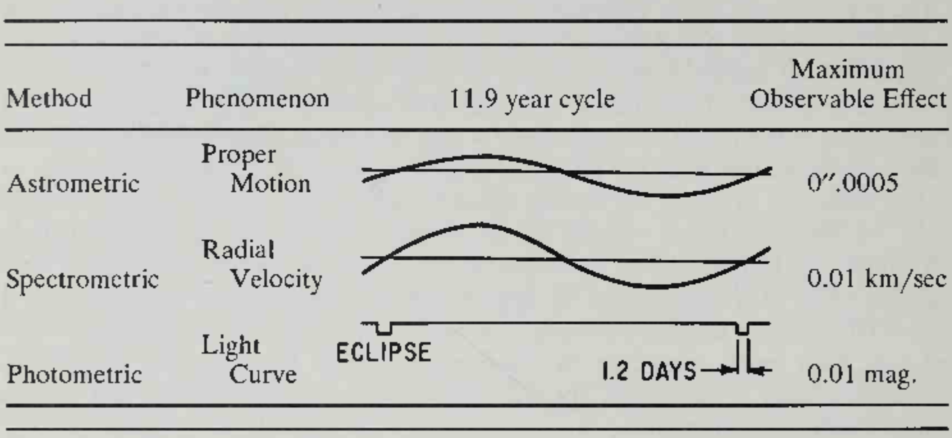


FIGURE 11–7. *Schematic representation of the astrometric, spectrometric, and photometric methods of detecting extrasolar planetary systems about stars 10 parsecs distant.*

diameter of between 0.5 and 2 seconds of arc. If a planet were 1 A.U. from its primary, and the primary star itself were distant from us about 10 parsecs, then the angular distance between the planet and the star would be about 0.1 second of arc. Thus, no telescope on Earth, of whatever size, could separate the image of the planet from the smeared image of the star.

In the near future, there may nevertheless be a possibility of directly observing large planets associated with nearby stars. The required instrument is a large telescope in orbit about the Earth. ▽ Such Orbiting Astronomical Observatories are under intensive development by the National Aeronautics and Space Administration in the United States, although their primary scientific objectives are studies of stars and the interstellar medium in the presently inaccessible ultraviolet wavelengths. The Orbiting Astronomical Observatory of the United States space program is an unmanned satellite. Similar instruments may be under development in the Soviet Union, perhaps in conjunction with manned spaceflight. △

Although an orbiting observatory will be well above atmospheric turbulence, it will still not be possible to observe or photograph objects which have indefinitely small angular separations. A limit to our ability to resolve nearby objects is imposed by the wave nature of light itself. Due to diffraction in the objective lens or mirror of a telescope, the focal plane image of any star is a system of rings of definite width. The diffraction-limited resolution of any telescope is approximately equal to the ratio of the wavelength of the light to the diameter of the objective. For example, with blue light and an objective lens or mirror of diameter 1 meter (39 inches), two close stars will be distinguished if they are more than 0.1 second of arc apart. ▽ If they are any closer, their separate images will fuse, and an indistinct diffraction smear will result. △

The use of a special instrument called an interferometer may, however, permit us to measure light sources separted by so small an angular distance as 0.01 second of arc. ▽ If both the planet and the star had large and comparable intensities, interferometric observations could be carried out with telescopes on the Earth. But since the brightness of the planet is negligible, compared with the brightness of the primary star, light scattering and turbulence in the Earth's atmosphere would preclude such observations. △

Let us consider a Jovian planet 1 A.U. away from its star, which is of solar type, and about 10 parsecs from us. The apparent magnitude of this planet is assumed to be approximately +24. Even such a weak light source could probably be detected from an orbiting space observatory, using current astronomical techniques. However, it seems unlikely that the critical observations could be performed automatically, at least in the near future. Thus, an astronomer-astronaut is, ▽ in the context of this problem, △ required for the orbiting astronomical observatory. Perhaps some day a large and stationary observatory will be established on the Moon. It will then be possible to carry out extensive observations on many significant astronomical problems which are currently difficult or impossible to pursue because of obscuration by the Earth's atmosphere.

Although our direct evidence for extrasolar planetary systems is at the present time limited, it is entirely reasonable to regard stellar multiplicity and planetary systems as different aspects of the same phenomenon. According to the investigations of the Dutch-American astronomer G. P. Kuiper of the University of Arizona, the average distance between the components of a binary star system is approximately 20 A.U. This is close to the dimensions of our own solar system.

How do multiple star systems arise? In the past, theories have been proposed which attempted to explain the formation of binary stars by the separation of a single star into two components. The cause of this stellar fission was supposedly the rapid rotation of the star. Because of centrifugal forces, the surface of a rapidly-rotating star would cease to be spherical. Mathematical calculations indicate that under certain ideal conditions, a rapidly-rotating body takes on a characteristic pear-shaped form; as it rotates still faster, an instability is established and the star separates into two components, ▽ each of which is rotating more slowly than the parent star. △ However, this hypothesis does not adequately account for the observations. We therefore discount this model of binary star formation.

An alternative hypothesis, proposed, for example, by O. Y. Schmidt, of the Soviet Union, postulates entrapment; that is, under certain conditions, two stars moving independently in space are imagined to become gravitationally coupled during a random encounter, and form a double star system. Although such a process is possible mathematically (for example, in the accidental encounter of three independent stars), the probability of such an event is exceedingly small. In addition, the entrapment model is contradicted by observations. For example, it cannot explain why quadruple star systems are always found in the systematic two-by-two pattern of Figure 11–8.

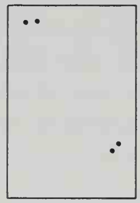


FIGURE 11–8. *The invariable relative positions of a quadruple star system. In such systems, two binaries are always observed orbiting each other.*

The bulk of relevant astronomical information garnered during the last two decades indicates that the stars in multiple star systems are coeval, formed simultaneously out of the gas and dust of the interstellar medium. From such condensations, entire groups of stars, associations and clusters, are formed [see Chapter 6]. Multiple star systems seem to be formed in the same way. In many cases, the components of a multiple star system seem to be of the same age. Often, we see a system in which both components are hot stars of spectral class O, or are early B stars. ▽ In Table III, we find that the components of Luyten 726–8 are both of spectral type M6e; 61 Cygni A and B have similar spectral types. △ According to contemporary concepts of the evolution of stars, such components are of similar mass and were formed at the same time from the primordial nebula. They are now at the same stage of evolution. Sometimes, one component will be a hot, main sequence giant, and its companion will be a red, relatively cold supergiant. We can conclude that the masses of the two stars were initially different, and that the supergiant represents the more rapid evolution of the more massive component [see Chapter 6].

In recent years, an extraordinary occurrence relevant to binary star systems has been discovered. Massive, hot young stars have, as a rule, a relatively small random velocity, generally less than 10 km sec^{-1}, ▽ corresponding to a small proper motion. △ Such stars are concentrated towards the Galactic plane, ▽ since they have been formed recently, and travel slowly. △ The exceptions to this rule are of interest. A small number of hot, massive stars move at unusually high random velocities, sometimes approaching 100 km sec^{-1}. Such stars may have escaped from stellar associations. In Figure 11–9, the points depict three such rapidly moving hot stars. The straight dotted lines indicate the directions of their motion through the sky. These lines almost intersect in the region of the constellation Orion, where there is a large association of hot stars. Since their velocities and

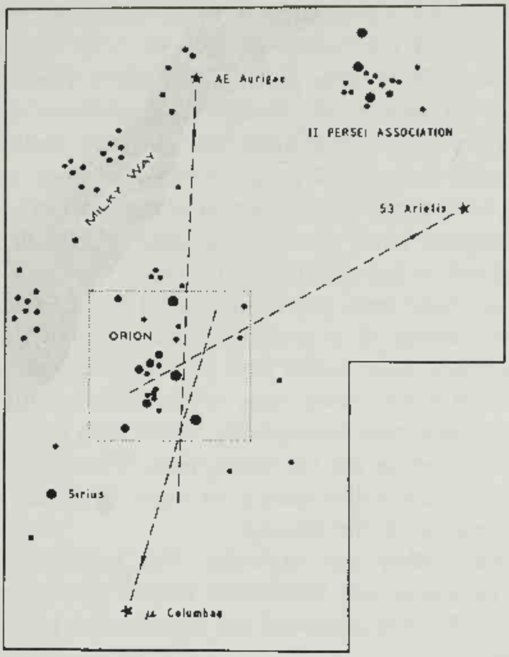


FIGURE 11–9. *Computed paths of the three best-known runaway stars. Note that they diverge from a comparatively small area in the constellation Orion. (Courtesy of Sky and Telescope, Cambridge, Massachusetts.)*

their distance to the Orion association are known, it is possible to compute backwards in time, and establish that these runaway stars left the Orion association some 2 to 5 million years ago.

Why are such stars ejected from their clusters? The Dutch astronomer Adriaan Blaauw has noted that these runaways are always single stars, an unusual circumstance, since multiplicity is particularly widespread among early type stars. According to Blaauw's hypothesis, the runaway stars were previously components of binary systems. The second components are postulated to be extremely massive hot stars of spectral class O, which became supernovae of Type II [see Chapter 7].

Consider now what would happen if the more massive component in a binary star system suddenly vanished, completely disappearing because of the supernova explosion. Gravitational forces could not hold the remaining star in its pre-explosion orbit, and it would fly off along a tangent to its original orbit, but with a velocity equal to its former orbital velocity. This "slingshot effect"—it is difficult to call it anything else—▽ is analogous to the situation in which a string, tied to a stone and whirled about the head, is suddenly cut. △

Now, the mass of the exploded star does not actually vanish without a trace; the supernova remnant is an expanding nebula which has a mass approximately equal to that of the original star [see Chapter 7]. If this nebula were inside the orbit of the surviving star, the gravitational forces would not be greatly altered, and the companion would not run away. If, however, the star were far inside the nebula, the nebula would have little gravitational effect. For the slingshot effect to work, it is necessary for most of the supernova remnant to leave the orbit of the surviving star during a time significantly less than the period of revolution. A binary star system with components quite far removed from one another—say, 10 to 20 A.U.—would have periods of revolution of the order of several years, and the condition for the slingshot effect would be satisfied. For sufficiently massive stars, the orbital velocities will be approximately 100 km sec^{-1}.

▽ The events we have been describing would be of considerable interest for the hypothetical inhabitants of a possible planet circling the runaway star. In earlier, more placid times, there would have been two brilliant suns in the sky, one near, the other further away. Some days would have no nights, because a star would be in the sky above each hemisphere. When both stars are above the same hemisphere, night must fall in the far hemisphere. The night would be extraordinary, because the stars in a stellar cluster, or association, are much more densely packed than in our region of the Galaxy.

▽ Suddenly, the farther star explodes. The biological consequences of a nearby supernova explosion are formidable indeed [see Chapter 7]. Unless an advanced technical civilization inhabited our hypothetical planet, all its inhabitants would be incinerated.

▽ For the purposes of this narrative, let us assume that our observers survive the supernova explosion. Shortly afterwards, the remaining star and its companion planets embark on an extraordinary interstellar voyage, moving about a parsec every 10,000 years. In less than a million years, the star would be well clear of the stellar association in which it was formed, and the nights, more frequent now, would exhibit a much more mundane celestial panoply.

▽ We must emphasize that the foregoing narrative is unrealistic in several respects. It is most unlikely for an advanced civilization to exist on such a planet, because the lifetime of the hot and massive local sun can itself only be a few million years old [see Chapter 6], not nearly enough time for the origin of life and the evolution of a technical civilization. We will return to the question of which stars are likely to hold indigenous planetary civilizations in Chapter 24. △

The presence of white dwarfs in multiple star systems—for example, in the system of Sirius (see Table III)—is explained by the fact that the more massive component has nearly completed its evolutionary history. However, it would be difficult to imagine a binary system which contained a hot, massive star of spectral class O, and also a red giant with a mass one and a half or two times that of the Sun. To leave the main sequence and become a red giant, a star of this mass would require 2 to 4 billion years (see Table I). A hot star of spectral type O, on the other hand, cannot remain on the main sequence for more than about 10 million years. Fortunately, binary systems similar to the one just described are entirely unknown.

Thus, the facts seem to indicate that the components of multiple star systems are formed simultaneously. If we can establish that the formation of planetary systems does not differ fundamentally from the formation of multiple star systems, then we shall be able to conclude that planets are coeval with their primary stars. The origin of planetary systems is the subject of the next two chapters.

12

Historical views on the origin of the solar system

But indeed all the whole Story of Comets and Planets, and the Production of the World, is founded upon such poor and trifling Grounds, that I have often wonder'd how an ingenious Man could spend all that pains in making such Fancies hang together. For my part, I shall be very well contented, and shall count I have done a great Matter, if I can but come to any Knowledge of the Nature of Things, as they now are, never troubling my self about their Beginning, or how they were made, knowing that to be out of the reach of human Knowledge, or even Conjecture.

Christianus Huygens, *New Conjectures Concerning the Planetary Worlds, Their Inhabitants and Productions* (c. 1670)

Upon a slight conjecture I have ventured on a dangerous journey, and I already behold the foothills of new lands. Those who have the courage to continue the search will set foot upon them.

Immanuel Kant, *Allgemeine Naturgeschichte und Theorie des Himmels* (1755)

From earliest times, the question of the origin and evolution of the Earth and the other planets in our solar system has challenged the keenest minds. Philosophers and scientists of the caliber of Kant and Laplace have wrestled with this problem; yet it still remains largely unresolved.

During the last decade, the theory of stellar evolution, sketched in Chapter 6, has gained widespread scientific acceptance. At first glance, it seems strange that astronomers should know more about the distant stars, which are in many cases difficult to observe, than about the relatively nearby planets. But the numbers of stars which astronomers can observe are vast, and stars are known that represent each stage of stellar evolution. We have been able to establish empirically, that the rate of evolution of a star depends on certain initial conditions—for example, on its mass. ▽ Given such observational hints, the theoretician's task has been simplified enormously. But no similar body of information on planetary evolution is currently available. △ If we should ever succeed in obtaining observations of numerous planetary systems in various stages of their development, then questions of planetary evolution may also be solved empirically.

Does it follow that we can say absolutely nothing about the origin of our solar system except that it was somehow formed, no later than 5 billion years ago? Such a stolid point of view is as reprehensible as unbounded speculation. Progress in stellar cosmogony has provided us with significant clues for planetary cosmogony. In addition to the observational evidence of the preceding chapter, there are today valid scientific arguments supporting the contention that many stars have planetary systems. ▽ We briefly consider in this chapter some earlier views on the origin of the solar system, and then proceed, in Chapter 13, to more current views.

▽ One of the earliest attempts to explain the origin of the world in scientific terms—insofar as science was, in those days, understood—was provided by Lucretius, in his "De Rerum Natura." Lucretius toyed with the idea that the universe was infinitely old, and that matter had always been in it. But he imagined a time before the origin of *things;* there was matter, but not in an organized form. The matter, naturally, was falling. Why? Because, as everyone knows, if you let go of an object, it falls. This was, of course, long before any connection was seen between the motion of falling bodies and the mass of the Earth. It was the nature of objects to fall; and in the beginning, matter fell. The motion was uniform, however, so one bit of matter could not jostle its neighbor. No forces were postulated between the particles. At this rate, the fall would continue forever, and the world would never begin.

▽ Lucretius therefore postulated the "Swerve." It was just the tiniest Swerve, barely adequate to produce collisions among the particles. Once there were collisions, and the colliding particles adhered one to another, Lucretius could imagine the origin of the Earth through a random interaction of material particles. The origin of the First Swerve was never explained, to say nothing of the fact that, in Lucretius' view, we must still be falling today.

▽ In the light of modern knowledge, Lucretius' naïveté was very great, but he was hypothesizing in an information vacuum. He stressed *random* interactions because the laws of physics were almost entirely unknown. One key ingredient which Lucretius missed was the possibility of forces acting between objects not in material contact. Only after the theory of gravitation was developed by Newton could planetary cosmogony enter a stage of profitable speculation.

▽ Among the earliest Newtonian cosmogonies are those of Kant and Laplace. △ It will be apparent, in Chapter 13, that the essential premise of their remarkable hypothesis has withstood the test of time, and we can find their basic idea in all contemporary theories of planetary origins. The nebular hypothesis was first expressed by the German philosopher Immanuel Kant, and several decades later was formulated independently by the French mathematician Pierre Simon, Marquis de Laplace. Although it is called the Kant–Laplace hypothesis, these two brilliant scholars did not concur on a number of important questions. Kant, for example, assumed that first the Sun and then the planets were formed from an initially cold, dusty nebula. Laplace, on the other hand, believed that the planets formed somewhat before the Sun, from an initially hot, gaseous nebula in a state of rapid rotation. Despite these differences, both agreed on the central point of what is now called the Kant–Laplace nebular hypothesis, that the solar system arose from the condensation of a primordial nebula, a cloud of gas and dust.

In their view, the Sun and the planets were formed (roughly) at the same time, from the identical nebula. Contracting under the influence of gravitational forces, the cloud began to rotate more and more rapidly, owing to the conservation of angular momentum. As a result of the large centrifugal forces generated by this rapid rotation, matter was thrown out from the periphery of the contracting nebula, forming an equatorial belt of gas and dust. ▽ As the contraction continued, the nebula shed successive rings of matter. The rings were thought to have condensed at a later time, to form the planets. This view of the early development of the contracting nebula was considered supported by the observations, then being made for the first time, of spiral nebulae [see Figures 9–3 and 9–4]. We now know, of course, that these objects are in fact distant galaxies, not nearby protostars. △ The fundamental premises of the Kant–Laplace hypothesis are sketched in Figure 12–1.

In the middle of the nineteenth century, ▽ primarily through the work of the celebrated Scottish physicist James Clerk Maxwell, △ it became clear that the nebular hypothesis contained a fundamental difficulty. Our planetary system, which consists of the Sun, nine planets of assorted sizes, ▽ and miscellaneous debris, ▽ possesses one quite remarkable peculiarity: a singular distribution of angular momentum.

Angular momentum is one of the most important characteristics of every

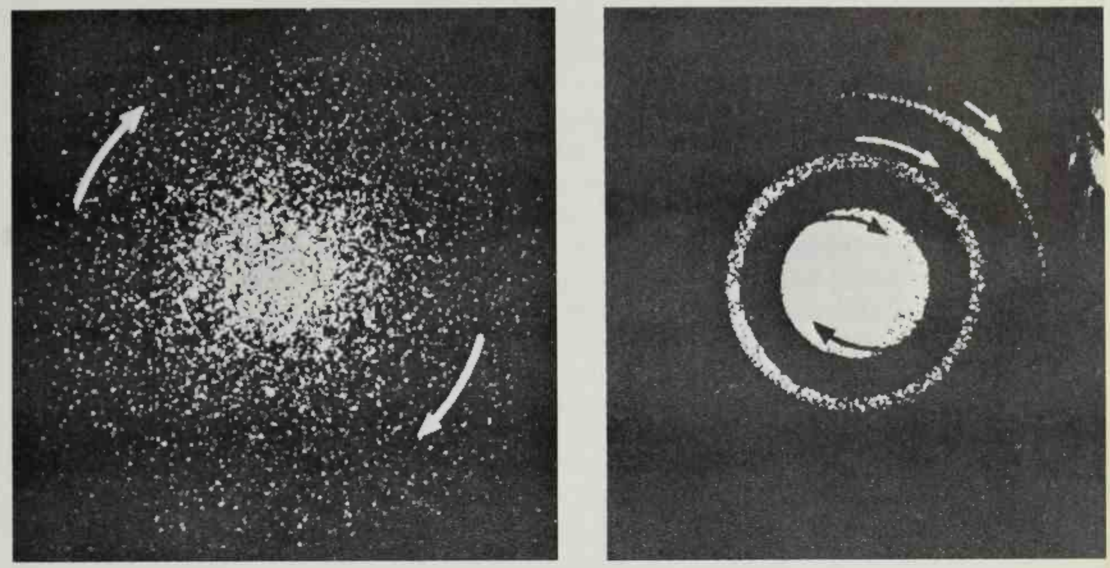


FIGURE 12–1. *Schematic representation of Laplace's nebular hypothesis. On the left is a rotating nebula of hot gas. On the right we see the nebula having contracted, increased its rotational velocity, and shed successive rings of gas. The rings were thought to condense into the planets, the dense central parts of the nebula forming the Sun.* (*Courtesy of Dr. Thornton Page, Wesleyan University.*)

mechanical system which is isolated from its environment. The Sun and its planetary accompaniment can be considered such an isolated system. Angular momentum in such a system is a reserve of rotational inertia, the tendency of a rotating body to resist braking. This rotational inertia arises from the orbital motion of the planets about the Sun, and from the rotation of the Sun and the planets on their respective axes. ▽ But the contribution from planetary rotation is negligible, compared to the contribution from planetary revolution and from solar rotation. The important point about angular momentum is that, like energy, it is a conserved quantity. In an isolated system, just as the total energy remains constant, so the total angular momentum remains constant. Since we can compute the angular momentum of the solar system at the present time, we can obtain some idea of the angular momentum of the solar system at the time of its formation—providing, again, that it is proper to consider the solar system as an isolated system. △

The orbital angular momentum of a planet is measured with respect to the center of gravity of the system, a point very close to the center of the Sun. It is defined as the product of the mass of the planet, m; its velocity of revolution about the Sun, v; and its distance, r, to the center of rotation—that is, to the Sun. In the case of a rotating, spherical, solid body, of uniform density the angular momentum with respect to an axis passing through its center is equal to $0.4\ MVR$, where M is the mass of the body, V is its equatorial velocity, and R is its radius. ▽ Note that we are explicitly distinguishing revolution from rotation by lower case and capital letters, respectively. △

The total mass of all the planets is only about $\frac{1}{700}$ of the solar mass. On the other hand, the distance of the planets from the Sun is much greater than the radius of the Sun, and many of the planets have a velocity of revolution about the Sun which is greater than the velocity of rotation of the Sun itself. For example, the orbital velocity of the Earth is about 30 km sec^{-1}, while the velocity of rotation of the Sun at its equator is only about 2 km sec^{-1}. When we take these numbers into consideration, we find that 98 percent of the angular momentum of the solar system derives from the orbital motions of the planets, and only 2 percent from the rotation of the Sun on its axis. In Figure 12–2 we see the distribution of angular momentum among the Sun and planets.

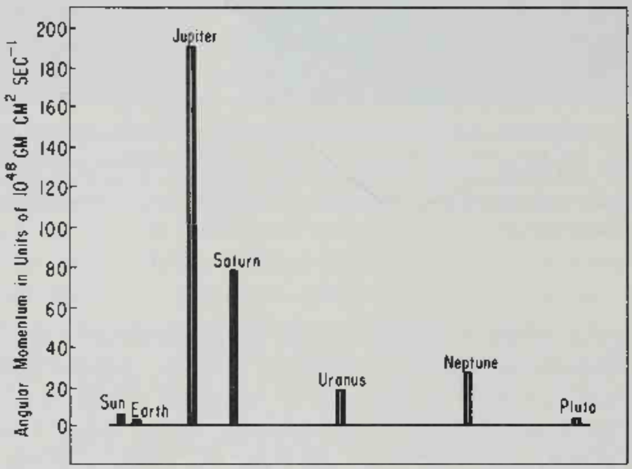


FIGURE 12–2. *Schematic representation of the distribution of angular momentum in the solar system. Despite its much larger mass, the Sun contains only a few percent of the total angular momentum of the solar system.*

In physics, masses are commonly expressed in grams, distances in centimeters, and time in seconds. ▽ Thus, the units of angular momentum, a product of mass, velocity, and distance, will be gm cm^2 sec^{-1}. △ Let us call the angular momentum of Jupiter, l. The mass of Jupiter is equal to $m = 2 \times 10^{30}$ gm (about 10^{-3} the mass of the Sun). The distance of Jupiter from the Sun, r, is 7.8×10^{13} cm (or 5.2 A.U.). The orbital velocity of Jupiter, v, is 1.3×10^{6} cm sec^{-1} (13 km sec^{-1}). Therefore, the angular momentum of revolution is $l = mvr = 2 \times 10^{50}$ gm cm^2 sec^{-1}. ▽ On the other hand, the mass of the Sun is 2×10^{33} gm; its equatorial velocity is about 2×10^{5} cm sec^{-1}; and its equatorial radius is about 7×10^{10} cm. The Sun is not quite a solid body; but for our purposes it will be sufficiently accurate to set the angular momentum of the Sun equal to 0.4 MVR. You can then easily convince yourself that the angular momentum of the rotating Sun is only about

1×10^{49} gm cm^2 sec^{-1}. Thus, Jupiter, having only 10^{-3} the mass of the Sun, has 20 times more angular momentum than the Sun. The actual figure, based on more exact calculations, is about 50 times more. △ From Figure 12–2, it is apparent that the terrestrial planets—Mercury, Venus, Earth, and Mars—have a combined total angular momentum which is some 380 times less than that of Jupiter. Thus, the lion's share of the angular momentum of the solar system is concentrated in the orbital motions of the giant planets Jupiter and Saturn.

This circumstance is entirely incomprehensible from the standpoint of the Kant–Laplace nebular hypothesis. In their view, the angular velocities of a given ring and of the contracting nebula should have been almost identical. When the ring separated from the nebula, it had approximately the same angular momentum per unit mass as the portion which continued to contract and form the Sun. But since the mass of the material destined to form the Sun was much greater than the mass of the protoplanetary ring, the angular momentum eventually residing in the Sun should have been much greater than that in the planets, if angular momentum were conserved. The Kant–Laplace hypothesis lacks any mechanism for the transfer of angular momentum from the protosun to the ring. ▽ Thus this otherwise promising theory floundered on the angular momentum difficulty. △

Other views replaced the nebular hypothesis. ▽ We will discuss here only one alternative—the collision hypothesis in the form enunciated by the English astronomer Sir James Jeans. It has both historical and philosophical interest. △ Jeans' hypothesis received wide acceptance during the first third of the present century, ▽ and is still treated warmly in obscure encyclopedias and some rural school textbooks in the United States. △ This theory is in all respects the complete antithesis of the nebular hypothesis. According to Kant and Laplace, the formation of the planetary system is a common process in stellar evolution. Jeans pictured the origin of our solar system as an exceedingly rare, perhaps even a unique, event.

Jeans believed that the original material from which the planets were formed was ejected by the Sun after the Sun had already reached its present form. The ejection was caused by the near collision of a passing star. The gravitational tidal forces of the interloper star drew out a filament of material from the surface layers of the Sun. The filament remained within the gravitational influence of the Sun as the interloper passed. After a period of time, the ejected material condensed and became the planets.

What can we say about this hypothesis? First of all, the initial event is exceedingly unlikely. As we mentioned in Chapter 3, the collisions of stars, and even their close mutual approach, are very rare phenomena in our Galaxy. Our Sun moves at a velocity of about 22 km sec^{-1} with respect to the nearest stars. Our closest neighbor—Proxima Centauri—is at a distance of 4.2 light years. It would take the Sun approximately 100,000 years, moving at the indicated velocity, to traverse this distance; ▽ and the Sun is, of course, not moving towards Proxima Centauri at all. △ It is not difficult to show that during its five-billion-year history, the Sun had only one chance in 10 billion of colliding with, or closely approaching, any other star. Similar probabilities apply to any other star. Since there are approximately 150 billion stars in our Galaxy, the total number of such close approaches throughout the entire Galaxy during the past 5 billion years would be approximately *ten*. (Near the center of the Galaxy, where the stellar density is tens of thousands of times greater than near the Sun, the probability of stellar collision would naturally be higher. ▽ But the bulk of the stars in the Milky Way are not in the Galactic nuclear region. △) Hence, if the hypothesis of Jeans were valid, the number of planetary systems formed in our Galaxy during 10 billion years could literally be counted on your fingers. ▽ (Actually, you might have to take off your shoes, and perhaps borrow someone else's toes, because *both* stars in the near collision would presumably have protoplanetary filaments expelled during the encounter. But the essential point is that Jeans' hypothesis makes our planets and the dark companions of nearby stars exceedingly, almost absurdly, improbable.) △

The number of further difficulties with Jeans' hypothesis is so great that we can consider such views completely refuted. It contains the same fatal inadequacy as the Kant–Laplace hypothesis: it does not explain why most of the angular momentum of the solar system is concentrated in the orbital motions of the planets. Mathematical calculations carried out by the Soviet astrophysicist N. N. Paruskii show that for all cases within the framework of Jeans' hypothesis, the planets would have very small, low-angular-momentum orbits.

Second, it does not follow that the stream of hot gas thrown off by the Sun would necessarily condense into planets. ▽ The reason is essentially this: since the putative filament initially had solar composition, it must have been composed primarily of hydrogen and helium (see Table I, Chapter 4). But hydrogen and helium are not even approximately the primary constituents of the terrestrial planets, and even in the outer planets like Jupiter and Saturn there seems to be less than the solar proportions of hydrogen and helium. This means that if the planets evolved from such a filament there must subsequently have been an escape of the lighter gases. If we add the cosmic proportions of hydrogen and helium to the present masses of the planets, we find that the filament must initially have had a mass of about 0.01 to 0.1 the mass of the Sun. The filament, then, could not have come from the ejection of *surface* material from the Sun; instead, it must have come from the solar interior. But the temperatures in the solar interior are so great—about 10^6 °K—that △ the filamentary material would not have condensed into planets. It would rather have been dissipated into interstellar space, as Lyman Spitzer, now at Princeton University, and other well-known astrophysicists, have shown. These difficulties, obvious by the late 1930's, provide an adequate basis for the rejection of the collision hypothesis.

▽ This last critique of Jeans' hypothesis is of considerable relevance to contemporary hypotheses. If the planets were formed initially from material of solar or cosmic composition, then the initial planetary mass must have been about 0.01 to 0.1 the present mass of the Sun. Since the total mass of the planets is now approximately 0.001 the mass of the Sun, we must imagine some means of dissipating most of this material. But for any condensation to have occurred, the planets must have been formed at relatively low temperatures, certainly temperatures below those of the interior of the Sun. We will return to these points later, in Chapters 13 and 16. △

The rejection of Jeans' views has led to a return, with various modifications, of the classical Kant–Laplace nebular hypothesis. At the time this hypothesis was initially formulated, astronomers knew virtually nothing about the nature of the stars and the galaxies; now, due to outstanding achievements in astrophysics, a vast amount of factual material has been collected. The Kant–Laplace hypothesis was originally based only on Newtonian mechanics, since at that time only mechanics had been mathematically well developed. Contemporary planetary cosmogony draws greatly on achievements in other areas of physics, particularly in magnetohydrodynamics. We shall see that magnetohydrodynamic considerations appear to eliminate the angular momentum difficulty in the nebular hypothesis.

13

Stellar rotation and the origin of the solar system

If then, Socrates, amid the many opinions about the gods and the generation of the universe, we are not able to give notions which are altogether and in every respect exact and consistent with one another, do not be surprised. Enough if we adduce probabilities . . .

Plato, *The Timaeus*

Who can number the clouds by wisdom
Or who can pour out the bottles of heaven,
When the dust runneth into a mass,
And the clods cleave fast together?

Job, 38:37

Before considering contemporary hypotheses on the origin of planetary systems, we must first discuss an important property of the stars—they rotate. In 1877, the English astronomer Captain W. de W. Abney, who is now all but forgotten, proposed that spectrographic observations could be used to determine how fast the stars rotate. If a star is rotating rapidly enough about its axis, and if the axis makes a large enough angle with the line of sight, then part of the surface of the star will be moving away from the observer, and part of it will be moving towards the observer. ▽ The side that moves away from the observer will be red-shifted by the Doppler effect; the side which moves towards the observer will be blue-shifted. These red- and blue-shifts apply equally to all the lines in the stellar spectrum. As a result, since the lines are shifted both to shorter and to longer wavelengths all spectral lines observed in the stellar atmosphere should be broadened. The more rapidly the star is rotating, the broader will be the lines. The only exception is the relatively rare case of a rapidly-rotating star which we see pole-on. If the axis of rotation is in the line of sight, no part of the star will be moving towards us or away from us, and therefore, despite the rapid rotation, there will be no Doppler broadening of the line from our vantage point. △

Abney's brilliant idea was not appreciated at the time, because astronomical spectroscopy was then in an embryonic state. Also, even early observations indicated that there could be both broad and narrow lines in the spectrum of the same star. Several decades passed before astronomers could distinguish among the several possible sources of line-broadening in stellar spectra. We now know that there can be a variety of causes for line-broadening ▽ —for example, strong electric or magnetic fields, local turbulence, or high temperatures and pressures— △ that have nothing to do with stellar rotation. In particular, we now know that lines belonging to the more abundant chemical elements can, under the physical conditions in stellar atmospheres, exhibit line-broadening independent of the rotation of the stars.

In 1928, the Russian-American astronomer Otto Struve and the Soviet astronomer G. A. Shajn solved this problem. Figure 13–1 shows schematically spectra of three hot stars, J Hercules, eta Ursae Majoris, and HR 2142. The three most intense lines which we observe in these spectra arise from transitions in neutral hydrogen and in neutral helium. A comparison of the upper and middle spectra shows that the hydrogen lines ($H\gamma$) appear about equally broad in both, while the helium lines are noticeably broader and more diffuse in the middle than in the upper spectrum. In the bottom spectrogram, all the lines are very broad and diffuse, to such an extent that they are almost invisible; ▽ they must be examined closely on the original photographic plate in order to determine their widths. △

Interpretation of these spectra is simple. In the upper spectrogram, the rotational velocity relative to the line of sight is almost zero; that is, either the star is rotating very slowly, or, fortuitously, its axis is almost along the line of sight. The width of the hydrogen line can be explained by reasons having nothing to do with the stellar rotation. ▽ In the middle spectrogram, the hydrogen line is not noticeably broader, but it is already so broad that it is difficult for it to be influenced by stellar rotation. The helium lines, however, △ indicate a rotational velocity of approximately 210 km sec^{-1}. In the bottom spectrum, the rotational velocity is so great—approximately 450 km sec^{-1}—that all the lines, even the hydrogen line Hγ, are greatly broadened and appear washed out.

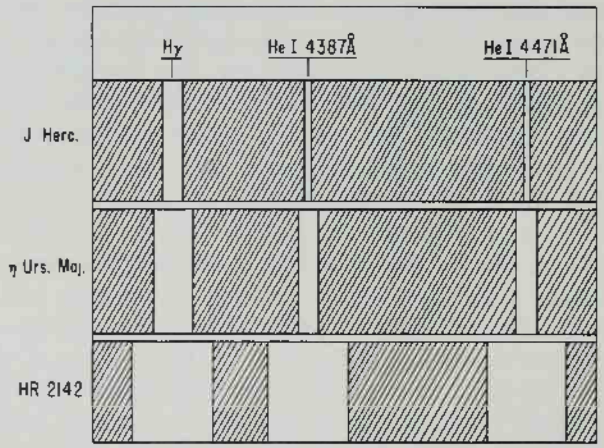


FIGURE 13–1. *Line-broadening in the spectrum of three stars, showing the effects of increasing equatorial rotational velocity. (Schematic drawing after photographic spectra of Struve and Shajn.)*

At the present time, the rotational velocities of a large number of stars are under investigation. The data in hand indicate that the rotational velocities differ greatly from star to star, and— ▽ a point first appreciated by Otto Struve △ — there is some relation between rotational velocity and spectral type. The hot, massive stars rotate very rapidly; the yellow and red main sequence dwarfs barely rotate at all. The velocity of the Sun at its equator is only about 2 km sec^{-1}. In Table IV, data on the velocities of rotation of stars of various spectral types is presented. Careful spectroscopic observations have shown, as is reflected in Table IV, that near spectral class F2, the velocity of rotation abruptly and characteristically decreases. The surface temperature of an F2 star is about 7000°K, ▽ about 1000 K° hotter than that of the Sun. △ What is the origin of this F2 discontinuity? ▽ Why do stellar rotational velocities decrease at all, as we go towards later spectral types?

▽ In itself, there is nothing strange about a variation of rotational velocity with spectral type. △ Other fundamental characteristics of the stars—for example, their luminosity and their surface temperature—vary uniformly from early type stars to late type stars. ▽ A tentative answer to the problem of the F2 discontinuity has been presented by Struve, and can be most easily grasped in the following terms: △ Let us imagine all the planets of the solar system somehow incorporated into the Sun. Since, in an isolated system, angular momentum must be conserved, such an imaginary fusion of the planets and the Sun would cause the latter to rotate much more rapidly. ▽ Since the planets represent only a small fraction of the total mass of the Sun, this fusion would change the mass of the Sun by a negligible amount.

Table IV

Relation Between Rotational Velocity and Spectral Type

Stellar equatorial rotational velocity (in km sec^{-1})	*Percent of stars of a given spectral type with rotational velocities in the ranges given at left*					
	Oe, Be	O,B	A	F0–F2	F5–F8	G, K, M
0–50	0	21	22	30	80	100
50–100	0	51	24	50	20	0
100–150	0	20	22	15	0	0
150–200	1	6	22	4	0	0
200–250	3	2	9	1	0	0
250–300	18	0	1	0	0	0
300–500	78	0	0	0	0	0

On the other hand, the planets contain so much angular momentum that △ the Sun would, under these circumstances, be forced to rotate some 50 times faster than it does at present—since its angular momentum would have to increase from its present 2 percent to the entire 100 percent of the total angular momentum of the solar system. The Sun would then be rotating with an equatorial velocity of about 100 km sec^{-1}. But this is exactly a typical velocity for stars which are more massive and hotter than spectral type F2. Thus, it is at least a reasonable hypothesis that the velocity of the Sun is small because somehow angular momentum has been transferred from the Sun to the planets.

▽ What, then, of all the other stars with spectral type later than F2? If our hypothesis is correct, we expect that they are rotating slowly because they too have transferred angular momentum to their planetary systems. Consider the implications of this result. We are in fact concluding that almost all stars of spectral type later than F2 are accompanied by planetary systems. But 93 percent of the main sequence stars in the sky are later than F2. The bulk of the remaining stars are giants, supergiants, white dwarfs, and dark companions. All but possibly the dark companions have gone through a main sequence evolutionary phase. Thus, from the widths of lines on spectral plates, we have reached the remarkable conclusion that the Galaxy is populated with planets! △ Since this conclusion is obviously of very great importance, we must convince ourselves that stars cannot lose angular momentum by other means—for example, by transferring it to the interstellar gas. In other words, we must prove that a rotating star with no planets is an isolated system, that it satisfies the law of conservation of angular momentum. How can we demonstrate this?

Let us consider first the case of the red giants. A great many of the giants are characterized by relatively rapid equatorial rotation. For example, the star xi Geminorum is of spectral type F5, and has a rotational velocity of 73 km sec^{-1}. This typical red giant (actually, it should be called a yellow giant) is a very old star; according to contemporary theories of stellar evolution [see Chapter 6], when it was on the main sequence, its spectral type was A [see the evolutionary tracks of Figure 6–3]. ▽ We can compare the present radius of xi Geminorum with the radius of a typical main sequence A star. △ When xi Geminorum was a main sequence star, its radius must have been about one-half its present radius. If it has conserved angular momentum during its evolution to the red giant stage, its equatorial velocity as a main sequence A star must have been 146 km sec^{-1}. We see from Table IV that this is a typical rotational velocity for main sequence stars of spectral type A. Thus, we can conclude that the angular momentum of xi Geminorum has been conserved during its evolution off the main sequence.

Now let us consider another example. ▽ There is a class of stars which resemble their prototype, T Tauri, the third variable star discovered in the constellation Taurus, the Bull. △ T Tauri stars belong to spectral type G, and have fairly high rotational velocities, up to 100 km sec^{-1}. They are believed to be very young—still in the stage of gravitational contraction [see Chapter 6]—and have not yet entered into the main sequence. Such stars are characteristically found to the right and a little above the main sequence, in the spectrum-luminosity diagram [see Figure 6–2]. As time progresses, their radii decrease. It appears that several millon years hence, when their contraction ends, T Tauri stars enter the main sequence at about spectral type A. In the course of such contraction, their radii decrease by a factor of approximately 2. Therefore, if they conserve angular momentum during contraction, their main sequence equatorial rotational velocity should be about 200 km sec^{-1}. We again see from Table IV that this is a typical equatorial rotational velocity for main sequence stars of spectral type A. Thus, in both early and late stages of stellar evolution, angular momentum seems to be conserved.

Nevertheless, we must emphasize that the results of these examples are based on contemporary theories of stellar evolution. The consistencies which we have found cannot serve as a *rigorous* proof of the conservation of angular momentum in stellar evolution. In fact, we shall see presently that there is at least a possibility that angular momentum may be lost independently of the existence of a planetary system.

The Swedish physicist Hannes Alfvén of the Royal Institute of Technology, Stockholm, first considered the problem of angular momentum transfer from stars to planets. He showed that a magnetic field could provide the mechanism for angular momentum transfer. Alfvén's view is incorporated in the theory of the origin of the solar system propounded by the English astrophysicist Fred Hoyle of Cambridge University. We believe that Hoyle's hypothesis is the most promising of those currently proposed, and we therefore will consider its basic tenets in some detail.

Following a now classic tradition, Hoyle pictures the planets formed from a cold nebula of gas and dust. Initially, the nebular density was very low. Individual regions of the cloud moved at varying velocities with respect to each other. By analogy with gaseous nebulae, Hoyle assumes that these velocities were about 1 km sec^{-1}.

As a result of such motions, the initial nebula must have possessed a small, but definite angular momentum. Moreover, the cloud must have had interstellar dimensions, being several light years across. Thus, if, in the process of condensation from interstellar to solar dimensions, angular momentum were conserved, the final equatorial velocity of the newly-formed star would have been almost equal to the speed of light. Such is not the case, as we have seen. Therefore, we must assume that more than 99 percent of the initial angular momentum was lost by the nebula during star formation. Such an angular momentum loss can, according to Hoyle, be explained by interstellar magnetic fields. ▽ The cloud and the interstellar medium initially share the same magnetic field, and we can imagine magnetic lines of force joining the contracting nebula with the material of the interstellar medium. As the nebula rotates more rapidly, because of its contraction, the lines of force cause an increased rotation of the interstellar medium external to the nebula, △ and thereby initiate an angular momentum transfer from the nebula to the surrounding interstellar medium. The lines of force act as bent strings. But, for a reason which we shall not discuss here, such transfer of angular momentum can occur only while the density of the nebula is low. When the density of the nebula ▽ increasing because of contraction △ reaches a certain value, angular momentum transfer from the cloud to the interstellar medium ceases.

▽ (The collective motion of charged particles in magnetic fields is a new branch of physics called magnetohydrodynamics. The hypothesis that solar rotation was braked because of angular momentum transfer along magnetic lines of force is an example of the application of magnetohydrodynamics to astronomy. Since magnetic fields are widespread in the universe, and since ionized atoms influenced by magnetic fields are commonplace, magnetohydrodynamics also has many other astronomical applications.) △

This theory has wide implications. As calculations carried out by Hoyle indicate, the remaining angular momentum, if it were concentrated only in the condensing star, would correspond to an equatorial rotational velocity of several hundred kilometers per second. As we have seen, relatively hot stars have rotational velocities of just this order. The theory does not yet account for the slow rotation of the relatively colder stars such as the Sun. We must assume that such stars lose angular momentum after contraction of the initial nebula to relatively small dimensions—say, to the dimensions of the solar system.

Therefore, two main problems remain with us: Why do stars of spectral type later than F2 lose nearly all of their rotational angular momentum? Why is this process of angular momentum loss inoperative in hotter stars?

To approach these questions, we note that as a nebula contracts, it rotates more and more rapidly about its axis. Let us call the contracting nebula a protostar. With the initial rotational velocities assumed by Hoyle, by the time a protostar of mass equal to that of our Sun contracts to a radius of 40 solar radii (about 0.2 A.U.), it will, according to Hoyle, be rotating so rapidly that the

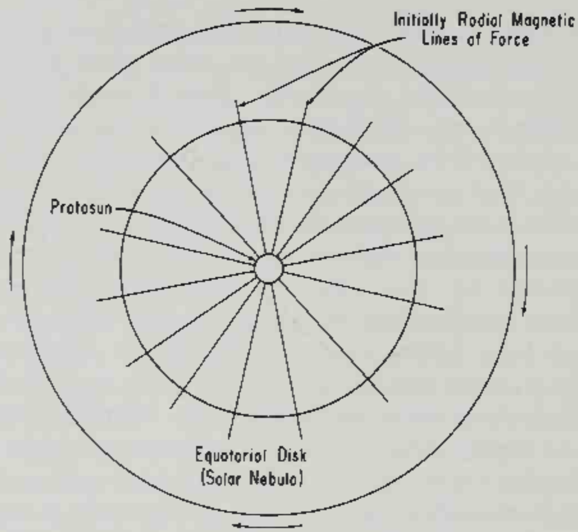


FIGURE 13–2. *An early stage in the evolution of the solar system, according to Hoyle. The contracting Sun has shed an equatorial disk of gas and dust which is connected to the Sun by magnetic lines of force.*

centrifugal force at the equator will just balance the force of gravity. Therefore a state of instability arises, and material is shed by the protostar to form an equatorial disk.

So far, the theory corresponds to the classical view of Laplace. The central problem in the Laplace hypothesis, we recall, concerns the transfer of angular momentum from the protostar to the disk which later condenses into the planets. Today we expect to find magnetic fields in condensing protostars. When the equatorial disk of gas is separated from the protostar, magnetic lines of force are likely to exist, connecting it to the protostar [Fig. 13–2]. The contracting protostar soon rotates more rapidly than the disk, and the lines of force connecting them, initially straight lines, now become twisted [Fig. 13–3]. As a result, the rotation of the protostar is effectively braked, and the disk is slowly forced outward from the protostar. In time, the width of the disk will increase due to internal friction, and part of this material should condense into the planets. In this way, the planets are the repository for the major fraction of the initial nebular angular momentum lost by the protostar.

Why does this process occur only in protostars destined to have main sequence spectral type later than F2? ▽ The protostar and the disk are, we recall, connected by magnetic lines of force which are initially radial, as in Figure 13–2. As the protostar continues to contract and rotate more rapidly, the magnetic lines of force tend to be wound around its periphery [Fig. 13–3]. The angular momentum is transferred through the lines of force to the disk, which also rotates more rapidly,

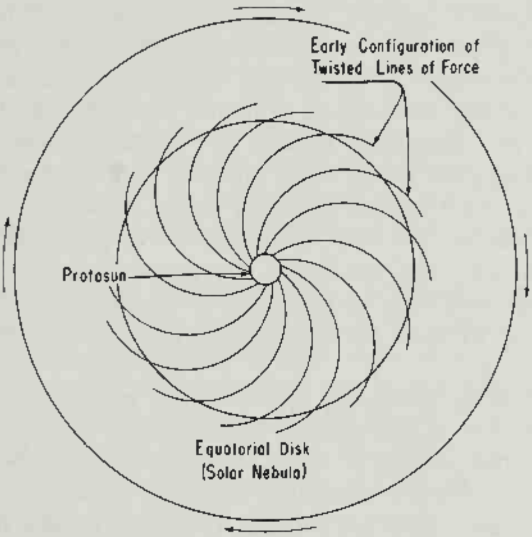


FIGURE 13–3. *A somewhat later stage in the early evolution of the solar system, according to Hoyle. Angular momentum is being transferred from the rapidly rotating protosun to the equatorial disk by the magnetic lines of force.*

thus preventing very high rotational velocities in the protostar. It is clear that the more massive the disk, the more difficult it will be for the disk to be accelerated by the winding magnetic lines of force. △ The mass of the disk is, however, not very large. Therefore, there will not be very much winding of the magnetic lines of force about the disk. These twisted lines of force are attached to the outer layers of the protostar, which are characterized by turbulent and disordered mass motion.

These outer, turbulent layers are believed to originate in the following way: The temperatures in stellar interiors are, of course, much greater than those of the radiating surfaces of stars. Hydrogen which is ionized in the hot interior becomes a neutral gas in the outer layers of the star. ▽ The transition from the ionized region to the neutral region is fairly abrupt, and leads to mechanical instability and the establishment of a turbulent zone composed primarily of neutral hydrogen in the outer reaches of stellar atmospheres. △ With a deep hydrogen convection zone, such as is found in cool stars, the magnetic lines of force, acting as if they were glued to the moving gas, can become embedded quite deeply in a protostellar atmosphere. However, in a hot protostar, hydrogen will remain ionized until quite close to the surface, and the hydrogen convection zone will be small or absent. In this case, the lines of force of the magnetic field will not penetrate deeply into the star but will wind themselves about the superficial layers. ▽ The mass of the shallow hydrogen convection zone of a hot protostar will be very small, and the winding lines of force will, according to Hoyle, be unable to transfer significant angular momentum to the more massive protoplanetary disk. △

Hoyle thus explains the abrupt decline in stellar rotational velocities near spectral type F2 in terms of the dependence of hydrogen ionization on temperature. In stars of spectral type F0, where the temperature of the surface layers is only about 1000° greater than for stars of spectral type F2, the convection zone begins so close to the surface of the star that magnetic lines of force are only slightly submerged in the stellar atmosphere. ▽ In this case, the angular rotation of the protostar will not be magnetically braked, and a main sequence star of relatively high rotational velocity will result. Note, however, that a protostar destined to be a main sequence star of F0 spectral type may shed a protoplanetary disk during the process of its formation. We have not excluded the possibility that such a disk formed around hot stars also condenses into planets. Hoyle's arguments support the view that all late type stars have transferred angular momentum to protoplanetary disks, but for all we know the condensation of such disks into planets may not be easy.

▽ Current thinking about the condensation processes is focused on two mechanisms. The first, a gravitational process, has been suggested by G. P. Kuiper. Here, the local density of the nebula is imagined to be so large that the mutual gravitational attraction of contiguous matter leads to further condensation, until objects of planetary mass are formed. The initial density of the nebula must be large enough that the tidal forces of the Sun do not pull the early condensations apart. The familiar ocean tides on Earth occur in part because the Sun and Moon pull more strongly on the oceans—slightly closer to them—than on the underlying land. The mutual gravitational attraction between the oceans and the Earth is, however, much greater than the dissipative tidal forces, so the oceans do not fly away into space each day. But unless the density of the solar nebula were rather large, condensations would not hold together gravitationally, and, in the absence of other influences, tidal instability would prevent the formation of the planets. This argument is originally due to the Scottish physicist James Clerk Maxwell.

▽ The second mechanism, suggested by the American cosmochemist Harold C. Urey of the University of California and others, is based on weak chemical bonds between colliding bits of matter. In effect, the condensates in the primitive nebula are imagined to be sticky. The view is not so very different from that of Lucretius.

▽ It is clear that if the nebula had very high temperatures, gases would boil off, condensates would evaporate, and the continued formation of planets would be difficult. It is generally believed that condensation occurred at low temperatures, perhaps as low as a few tens of degrees above absolute zero. Yet, as the Sun contracted towards the main sequence, its luminosity was very high [see Figure 6–2]. This apparent inconsistency between theories of solar and of planetary origins can possibly be resolved if we imagine a significant absorption of solar radiation by dust in the inner parts of the solar nebula, so that very little solar heating of the bulk of the nebula occurred; or alternatively that most of the planetary condensation was complete before the luminosity of the protosun became very great.

▽ One possible difficulty with the gravitational condensation mechanism is that the nebular densities required lead to very large nebular masses. The condensing protoplanets must then have been much more massive than they are today, and some process of dissipation of the bulk of their mass must be imagined. If there was an intense flux of charged particles emitted by the sun in those early times, such escape may have taken place. In the chemical accretion mechanism, much lower densities are adequate, and the dissipation difficulty is avoided. ▽ The problem of the escape of the light gases during the early history of the solar system is a pivotal one. Regardless of the temperatures at the top of the protoplanetary atmospheres, dissipation encounters very serious problems which were first pointed out by Shklovskii. △

The chemical content of the planets is clearly different from that of the Sun. If we added hydrogen and helium to the masses of the planets until these elements were present in cosmic proportions, the total mass of all the planets would be roughly ten times as great as it is at present—that is, the planets would comprise a few percent of the solar mass, a mass comparable to that of the invisible companion stars discussed in Chapter 11. Conceivably, the low-mass components of binary star systems were formed by processes similar to those which led to the formation of the planets. The difference may be that when the planets were formed the excess hydrogen and helium evaporated into interstellar space, while when the dark companion stars were formed the lighter elements remained gravitationally bound.

We now consider a more serious objection to Hoyle's theory. The existence of stellar magnetic fields raises the possibility that angular momentum may be lost even if planets are not formed. It is known that the Sun emits streams of charged particles ▽ —the so-called solar wind, which arises from the solar atmosphere and pervades interplanetary space. △ Individual clouds of hot, ionized gas are ejected from the vicinity of sunspots [Figure 1–3] at velocities of several hundred, or even several thousand, kilometers per second. This ionized material is an excellent conductor of electricity. It therefore moves along the magnetic lines of force of the solar magnetic field. At great distances from the Sun these magnetic lines of force have an almost radial direction, ▽ like pins in a pincushion. △ Moving along the lines of force, the ionized clouds of the solar wind can be ejected some tens of solar radii from the surface of the Sun.

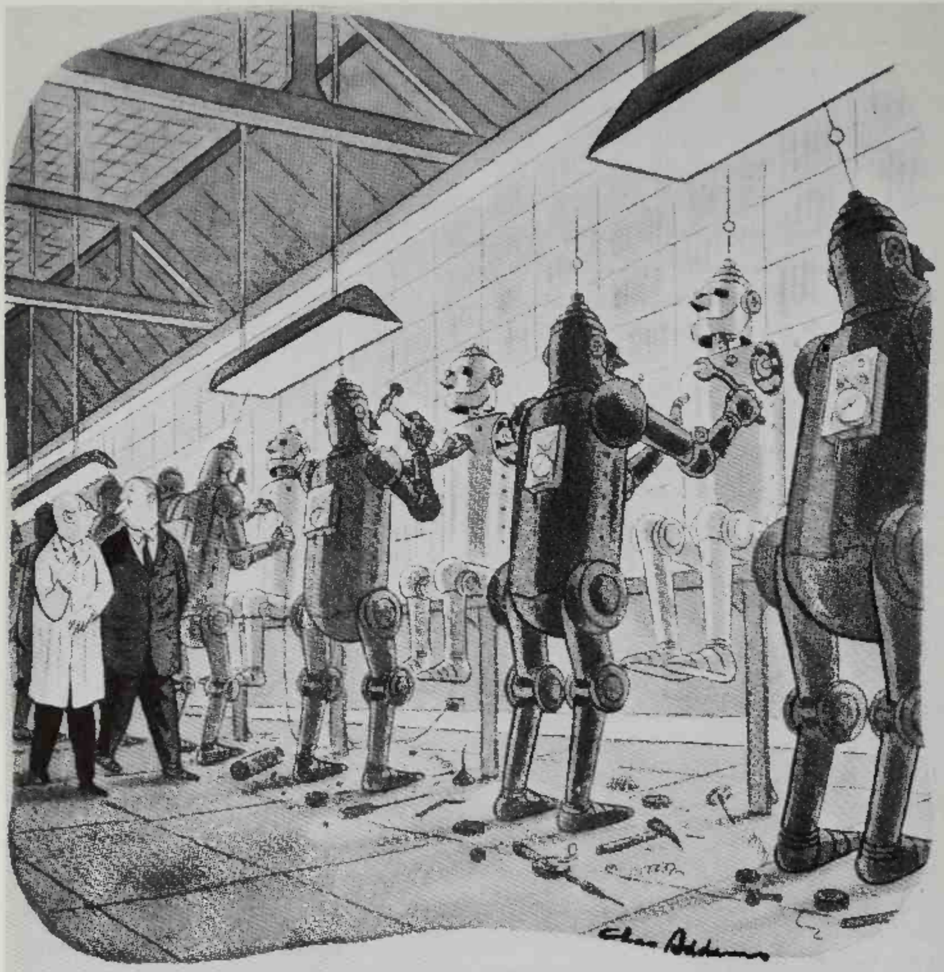
The magnetic lines of force rotate about the axis of the Sun with the same speed of rotation as do the surface layers. We can imagine the lines of force as a rigid wire framework fastened to a rotating sphere. Thus, the clouds of gas ejected from the Sun will increase their angular momentum as they move outward along the lines of force, ▽ because, we recall, the angular momentum of revolution is proportional to the distance from the Sun. △ If the clouds can break loose from the lines of force at great distances, where the magnetic fields are weak, then significant quantities of angular momentum could be lost into interstellar space. ▽ Similar ideas have been expressed by the French astrophysicist Evry Schatzmann of the Paris Observatory. △ Assume, for example, that gas clouds escape characteristically at a distance of 30 solar radii from the surface of the Sun. Then, in order to lose nearly all its initial angular momentum, the Sun need eject only about 0.1% of its mass. Such a relatively small mass loss during billions of years of solar evolution is entirely possible. At the present time, the rotational velocity of the Sun is not being braked by the loss of angular momentum through the ejection of ionized gas clouds, because the mass loss due to the solar wind is too small. But is it possible that, perhaps in the past, the mass loss was greater.

Thus, while the slow rotation of stars of spectral type later than F2 strongly suggests that planetary systems are associated with these stars, the evidence is not conclusive. There is an alternative and still admissible hypothesis—the loss of angular momentum to the interstellar medium—which does not connect planetary rotation with stellar rotation.

▽ In addition to the hypotheses we have mentioned, other views of planetary cosmogony have been propounded in recent years. △ For example, the Soviet scientist O. Y. Schmidt did not believe that the Sun ever possessed a cloud of gas and dust which later formed the planets. ▽ Schmidt pictured the Sun as capturing an interstellar cloud of gas, dust, and larger, accreted objects some time after its formation. △ However, the capture process is highly improbable. In addition, investigations of the last decade indicate, as we have seen, that the processes of star formation and of planet formation are closely interrelated.

Recently, the English astronomer W. H. McCrea of Royal Holloway College has proposed a cosmogonic hypothesis of a purely mechanical character, which does not take electromagnetic phenomena into account. Although McCrea's hypothesis explains why the angular momentum must be concentrated in the orbital motions of the planets, it does not explain the abrupt decline in stellar rotational velocity near spectral type F2. ▽ The advantage of the magnetic braking hypothesis is that it explains these two observations which otherwise are unconnected. △

Before the problem of the origin of planetary systems is definitively solved, much more work in theoretical physics and observational astronomy must be performed. But a beginning has been made, and the contours of a well structured theory have emerged.



"Sometimes I ask myself, 'Where will it ever end?'"