Cosmology

Cosmology, field of study that brings together the natural sciences, particularly astronomy and physics, in a joint effort to understand the physical universe as a unified whole.

If one looks up on a clear night, one will see that the sky is full of stars. During the summer months in the Northern Hemisphere, a faint band of light stretches from horizon to horizon, a swath of pale white cutting across a background of deepest black. For the early Egyptians, this was the heavenly Nile, flowing through the land of the dead ruled by Osiris. The ancient Greeks likened it to a river of milk. Astronomers now know that the band is actually composed of countless stars in a flattened disk seen edge on. The stars are so close to one another along the line of sight that the unaided eye has difficulty discerning the individual members. Through a large telescope, astronomers find myriads of like systems sprinkled throughout the depths of space. They call such vast collections of stars galaxies, after the Greek word for milk, and call the local galaxy to which the Sun belongs the Milky Way Galaxy or simply the Galaxy.

The Sun is a star around which Earth and the other planets revolve, and by extension every visible star in the sky is a sun in its own right. Some stars are intrinsically brighter than the Sun; others, fainter. Much less light is received from the stars than from the Sun because the stars are all much farther away. Indeed, they appear densely packed in the Milky Way only because there are so many of them. The actual separations of the stars are enormous, so large that it is conventional to measure their distances in units of how far light can travel in a given amount of time. The speed of light (in a vacuum) equals 3 × 1010 cm/sec (centimetres per second); at such a speed, it is possible to circle the Earth seven times in a single second. Thus in terrestrial terms the Sun, which lies 500 light-seconds from the Earth, is very far away; however, even the next closest star, Proxima Centauri, at a distance of 4.3 light-years (4.1 × 1018 cm), is 270,000 times farther yet. The stars that lie on the opposite side of the Milky Way from the Sun have distances that are on the order of 100,000 light-years, which is the typical diameter of a large spiral galaxy.

If the kingdom of the stars seems vast, the realm of the galaxies is larger still. The nearest galaxies to the Milky Way system are the Large and Small Magellanic Clouds, two irregular satellites of the Galaxy visible to the naked eye in the Southern Hemisphere. The Magellanic Clouds are relatively small (containing roughly 109 stars) compared to the Galaxy (with some 1011 stars), and they lie at a distance of about 200,000 light-years. The nearest large galaxy comparable to the Galaxy is the Andromeda Galaxy (also called M31 because it was the 31st entry in a catalog of astronomical objects compiled by the French astronomer Charles Messier in 1781), and it lies at a distance of about 2,000,000 light-years. The Magellanic Clouds, the Andromeda Galaxy, and the Milky Way system all are part of an aggregation of two dozen or so neighbouring galaxies known as the Local Group. The Galaxy and M31 are the largest members of this group.



The Andromeda Galaxy, also known as the Andromeda Nebula or M31. It is the closest spiral galaxy to Earth, at a distance of 2.48 million light-years.

The Galaxy and M31 are both spiral galaxies, and they are among the brighter and more massive of all spiral galaxies. The most luminous and brightest galaxies, however, are not spirals but rather supergiant ellipticals (also called cD galaxies by astronomers for historical reasons that are not particularly illuminating). Elliptical galaxies have roundish shapes rather than the flattened distributions that characterize spiral galaxies, and they tend to occur in rich clusters (those containing thousands of members) rather than in the loose groups favoured by spirals. The brightest member galaxies of rich clusters have been detected at distances exceeding several thousand million light-years from the Earth. The branch of learning that deals with phenomena at the scale of many millions of light-years is called cosmology—a term derived from combining two Greek words, kosmos, meaning “order,” “harmony,” and “the world,” and logos, signifying “word” or “discourse.” Cosmology is, in effect, the study of the universe at large.

**The Cosmological Expansion**

When the universe is viewed in the large, a dramatic new feature, not present on small scales, emerges—namely, the cosmological expansion. On cosmological scales, galaxies (or, at least, clusters of galaxies) appear to be racing away from one another with the apparent velocity of recession being linearly proportional to the distance of the object. This relation is known as the Hubble law (after its discoverer, the American astronomer Edwin Powell Hubble). Interpreted in the simplest fashion, the Hubble law implies that 13.8 billion years ago all of the matter in the universe was closely packed together in an incredibly dense state and that everything then exploded in a “big bang,” the signature of the explosion being written eventually in the galaxies of stars that formed out of the expanding debris of matter. Strong scientific support for this interpretation of a big bang origin of the universe comes from the detection by radio telescopes of a steady and uniform background of microwave radiation. The cosmic microwave background is believed to be a ghostly remnant of the fierce light of the primeval fireball reduced by cosmic expansion to a shadow of its former splendour but still pervading every corner of the known universe.

According to the evolutionary, or big bang, theory of the universe, the universe is expanding while the total energy and matter it contains remain constant. Therefore, as the universe expands, the density of its energy and matter must become progressively thinner. At left is a two-dimensional representation of the universe as it appears now, with galaxies occupying a typical section of space. At right, billions of years later the same amount of matter will fill a larger volume of space.

According to the evolutionary, or big bang, theory of the universe, the universe is expanding while the total energy and matter it contains remain constant. Therefore, as the universe expands, the density of its energy and matter must become progressively thinner. At left is a two-dimensional representation of the universe as it appears now, with galaxies occupying a typical section of space. At right, billions of years later the same amount of matter will fill a larger volume of space.

The simple (and most common) interpretation of the Hubble law as a recession of the galaxies over time through space, however, contains a misleading notion. In a sense, as will be made more precise later in the article, the expansion of the universe represents not so much a fundamental motion of galaxies within a framework of absolute time and absolute space, but an expansion of time and space themselves. On cosmological scales, the use of light-travel times to measure distances assumes a special significance because the lengths become so vast that even light, traveling at the fastest speed attainable by any physical entity, takes a significant fraction of the age of the universe (13.8 billion years old) to travel from an object to an observer. Thus, when astronomers measure objects at cosmological distances from the Local Group, they are seeing the objects as they existed during a time when the universe was much younger than it is today. Under these circumstances, Albert Einstein taught in his theory of general relativity that the gravitational field of everything in the universe so warps space and time as to require a very careful reevaluation of quantities whose seemingly elementary natures are normally taken for granted.

**The Nature Of Space And Time**

**Finite or infinite?**

An issue that arises when one contemplates the universe at large is whether space and time are infinite or finite. After many centuries of thought by some of the best minds, humanity has still not arrived at conclusive answers to these questions. Aristotle’s answer was that the material universe must be spatially finite, for if stars extended to infinity, they could not perform a complete rotation around Earth in 24 hours. Space must then itself also be finite because it is merely a receptacle for material bodies. On the other hand, the heavens must be temporally infinite, without beginning or end, since they are imperishable and cannot be created or destroyed.

Except for the infinity of time, these views came to be accepted religious teachings in Europe before the period of modern science. The most notable person to publicly express doubts about restricted space was the Italian philosopher-mathematician Giordano Bruno, who asked the obvious question that, if there is a boundary or edge to space, what is on the other side? For his advocacy of an infinity of suns and earths, he was burned at the stake in 1600.

In 1610 the German astronomer Johannes Kepler provided a profound reason for believing that the number of stars in the universe had to be finite. If there were an infinity of stars, he argued, then the sky would be completely filled with them and night would not be dark! This point was rediscussed by the astronomers Edmond Halley of England and Jean-Philippe-Loys de Chéseaux of Switzerland in the 18th century, but it was not popularized as a paradox until Wilhelm Olbers of Germany took up the problem in the 19th century. The difficulty became potentially very real with American astronomer Edwin Hubble’s measurement of the enormous extent of the universe of galaxies with its large-scale homogeneity and isotropy. His discovery of the systematic recession of the galaxies provided an escape, however. At first people thought that the redshift effect alone would suffice to explain why the sky is dark at night—namely, that the light from the stars in distant galaxies would be redshifted to long wavelengths beyond the visible regime. The modern consensus is, however, that a finite age for the universe is a far more important effect. Even if the universe is spatially infinite, photons from very distant galaxies simply do not have the time to travel to Earth because of the finite speed of light. There is a spherical surface, the cosmic event horizon (13.8 billion light-years in radial distance from Earth at the current epoch), beyond which nothing can be seen even in principle; and the number (roughly 1010) of galaxies within this cosmic horizon, the observable universe, are too few to make the night sky bright.

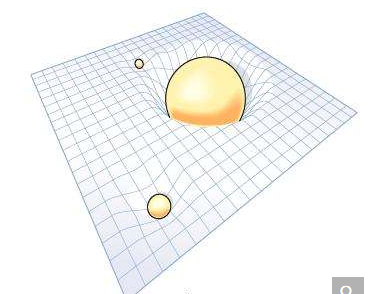
When one looks to great distances, one is seeing things as they were a long time ago, again because light takes a finite time to travel to Earth. Over such great spans, do the classical notions of Euclid concerning the properties of space necessarily continue to hold? The answer given by Einstein was: No, the gravitation of the mass contained in cosmologically large regions may warp one’s usual perceptions of space and time; in particular, the Euclidean postulate that parallel lines never cross need not be a correct description of the geometry of the actual universe. And in 1917 Einstein presented a mathematical model of the universe in which the total volume of space was finite yet had no boundary or edge. The model was based on his theory of general relativity that utilized a more generalized approach to geometry devised in the 19th century by the German mathematician Bernhard Riemann.

**Gravitation and the geometry of space-time**

The physical foundation of Einstein’s view of gravitation, general relativity, lies on two empirical findings that he elevated to the status of basic postulates. The first postulate is the relativity principle: local physics is governed by the theory of special relativity. The second postulate is the equivalence principle: there is no way for an observer to distinguish locally between gravity and acceleration. The motivation for the second postulate comes from Galileo’s observation that all objects—independent of mass, shape, colour, or any other property—accelerate at the same rate in a (uniform) gravitational field.

Einstein’s theory of special relativity, which he developed in 1905, had as its basic premises (1) the notion (also dating back to Galileo) that the laws of physics are the same for all inertial observers and (2) the constancy of the speed of light in a vacuum—namely, that the speed of light has the same value (3 × 1010 centimetres per second [cm/sec], or 2 × 105 miles per second [miles/sec]) for all inertial observers independent of their motion relative to the source of the light. Clearly, this second premise is incompatible with Euclidean and Newtonian precepts of absolute space and absolute time, resulting in a program that merged space and time into a single structure, with well-known consequences. The space-time structure of special relativity is often called “flat” because, among other things, the propagation of photons is easily represented on a flat sheet of graph paper with equal-sized squares. Let each tick on the vertical axis represent one light-year (9.46 × 1017 cm [5.88 × 1012 miles]) of distance in the direction of the flight of the photon, and each tick on the horizontal axis represent the passage of one year (3.16 × 107 seconds) of time. The propagation path of the photon is then a 45° line because it flies one light-year in one year (with respect to the space and time measurements of all inertial observers no matter how fast they move relative to the photon).

The principle of equivalence in general relativity allows the locally flat space-time structure of special relativity to be warped by gravitation, so that (in the cosmological case) the propagation of the photon over thousands of millions of light-years can no longer be plotted on a globally flat sheet of paper. To be sure, the curvature of the paper may not be apparent when only a small piece is examined, thereby giving the local impression that space-time is flat (i.e., satisfies special relativity). It is only when the graph paper is examined globally that one realizes it is curved (i.e., satisfies general relativity).



Curved space-timeThe four dimensional space-time continuum itself is distorted in the vicinity of any mass, with the amount of distortion depending on the mass and the distance from the mass. Thus, relativity accounts for Newton's inverse square law of gravity through geometry and thereby does away with the need for any mysterious “action at a distance.”

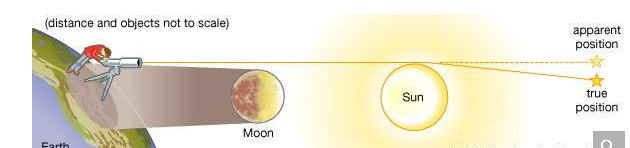
In Einstein’s 1917 model of the universe, the curvature occurs only in space, with the graph paper being rolled up into a cylinder on its side, a loop around the cylinder at constant time having a circumference of 2πR—the total spatial extent of the universe. Notice that the “radius of the universe” is measured in a “direction” perpendicular to the space-time surface of the graph paper. Since the ringed space axis corresponds to one of three dimensions of the actual world (any will do since all directions are equivalent in an isotropic model), the radius of the universe exists in a fourth spatial dimension (not time) which is not part of the real world. This fourth spatial dimension is a mathematical artifice introduced to represent diagrammatically the solution (in this case) of equations for curved three-dimensional space that need not refer to any dimensions other than the three physical ones. Photons traveling in a straight line in any physical direction have trajectories that go diagonally (at 45° angles to the space and time axes) from corner to corner of each little square cell of the space-time grid; thus, they describe helical paths on the cylindrical surface of the graph paper, making one turn after traveling a spatial distance 2πR. In other words, always flying dead ahead, photons would return to where they started from after going a finite distance without ever coming to an edge or boundary. The distance to the “other side” of the universe is therefore πR, and it would lie in any and every direction; space would be closed on itself.

Now, except by analogy with the closed two-dimensional surface of a sphere that is uniformly curved toward a centre in a third dimension lying nowhere on the two-dimensional surface, no three-dimensional creature can visualize a closed three-dimensional volume that is uniformly curved toward a centre in a fourth dimension lying nowhere in the three-dimensional volume. Nevertheless, three-dimensional creatures could discover the curvature of their three-dimensional world by performing surveying experiments of sufficient spatial scope. They could draw circles, for example, by tacking down one end of a string and tracing along a single plane the locus described by the other end when the string is always kept taut in between (a straight line) and walked around by a surveyor. In Einstein’s universe, if the string were short compared to the quantity R, the circumference of the circle divided by the length of the string (the circle’s radius) would nearly equal 2π = 6.2837853…, thereby fooling the three-dimensional creatures into thinking that Euclidean geometry gives a correct description of their world. However, the ratio of circumference to length of string would become less than 2π when the length of string became comparable to R. Indeed, if a string of length πR could be pulled taut to the antipode of a positively curved universe, the ratio would go to zero. In short, at the tacked-down end the string could be seen to sweep out a great arc in the sky from horizon to horizon and back again; yet, to make the string do this, the surveyor at the other end need only walk around a circle of vanishingly small circumference.

To understand why gravitation can curve space (or more generally, space-time) in such startling ways, consider the following thought experiment that was originally conceived by Einstein. Imagine an elevator in free space accelerating upward, from the viewpoint of a woman in inertial space, at a rate numerically equal to g, the gravitational field at the surface of Earth. Let this elevator have parallel windows on two sides, and let the woman shine a brief pulse of light toward the windows. She will see the photons enter close to the top of the near window and exit near the bottom of the far window because the elevator has accelerated upward in the interval it takes light to travel across the elevator. For her, photons travel in a straight line, and it is merely the acceleration of the elevator that has caused the windows and floor of the elevator to curve up to the flight path of the photons.

Let there now be a man standing inside the elevator. Because the floor of the elevator accelerates him upward at a rate g, he may—if he chooses to regard himself as stationary—think that he is standing still on the surface of Earth and is being pulled to the ground by its gravitational field g. Indeed, in accordance with the equivalence principle, without looking out the windows (the outside is not part of his local environment), he cannot perform any local experiment that would inform him otherwise. Let the woman shine her pulse of light. The man sees, just like the woman, that the photons enter near the top edge of one window and exit near the bottom of the other. And just like the woman, he knows that photons propagate in straight lines in free space. (By the relativity principle, they must agree on the laws of physics if they are both inertial observers.) However, since he actually sees the photons follow a curved path relative to himself, he concludes that they must be bent by the force of gravity. The woman tries to tell him there is no such force at work; he is not an inertial observer. Nonetheless, he has the solidity of Earth beneath him, so he insists on attributing his acceleration to the force of gravity. According to Einstein, they are both right. There is no need to distinguish locally between acceleration and gravity—the two are in some sense equivalent. But if that is the case, then it must be true that gravity—“real” gravity—can actually bend light. And indeed it can, as many experiments have shown since Einstein’s first discussion of the phenomenon.

It was the genius of Einstein to go even further. Rather than speak of the force of gravitation having bent the photons into a curved path, might it not be more fruitful to think of photons as always flying in straight lines—in the sense that a straight line is the shortest distance between two points—and that what really happens is that gravitation bends space-time? In other words, perhaps gravitation is curved space-time, and photons fly along the shortest paths possible in this curved space-time, thus giving the appearance of being bent by a “force” when one insists on thinking that space-time is flat. The utility of taking this approach is that it becomes automatic that all test bodies fall at the same rate under the “force” of gravitation, for they are merely producing their natural trajectories in a background space-time that is curved in a certain fashion independent of the test bodies. What was a minor miracle for Galileo and Newton becomes the most natural thing in the world for Einstein.



Experimental evidence for general relativityIn 1919, observation of a solar eclipse confirmed Einstein's prediction that light is bent in the presence of mass. This experimental support for his general theory of relativity garnered him instant worldwide acclaim.

To complete the program and to conform with Newton’s theory of gravitation in the limit of weak curvature (weak field), the source of space-time curvature would have to be ascribed to mass (and energy). The mathematical expression of these ideas constitutes Einstein’s theory of general relativity, one of the most beautiful artifacts of pure thought ever produced. The American physicist John Archibald Wheeler and his colleagues summarized Einstein’s view of the universe in these terms:

Curved spacetime tells mass-energy how to move;

mass-energy tells spacetime how to curve.

Contrast this with Newton’s view of the mechanics of the heavens:

Force tells mass how to accelerate;

mass tells gravity how to exert force.

Notice therefore that Einstein’s worldview is not merely a quantitative modification of Newton’s picture (which is also possible via an equivalent route using the methods of quantum field theory) but represents a qualitative change of perspective. And modern experiments have amply justified the fruitfulness of Einstein’s alternative interpretation of gravitation as geometry rather than as force. His theory would have undoubtedly delighted the Greeks.

**Relativistic Cosmologies**

**Einstein’s model**

To derive his 1917 cosmological model, Einstein made three assumptions that lay outside the scope of his equations. The first was to suppose that the universe is homogeneous and isotropic in the large (i.e., the same everywhere on average at any instant in time), an assumption that the English astrophysicist Edward A. Milne later elevated to an entire philosophical outlook by naming it the cosmological principle. Given the success of the Copernican revolution, this outlook is a natural one. Newton himself had it implicitly in mind when he took the initial state of the universe to be everywhere the same before it developed “ye Sun and Fixt stars.”

The second assumption was to suppose that this homogeneous and isotropic universe had a closed spatial geometry. As described above, the total volume of a three-dimensional space with uniform positive curvature would be finite but possess no edges or boundaries (to be consistent with the first assumption).

The third assumption made by Einstein was that the universe as a whole is static—i.e., its large-scale properties do not vary with time. This assumption, made before Hubble’s observational discovery of the expansion of the universe, was also natural; it was the simplest approach, as Aristotle had discovered, if one wishes to avoid a discussion of a creation event. Indeed, the philosophical attraction of the notion that the universe on average is not only homogeneous and isotropic in space but also constant in time was so appealing that a school of English cosmologists—Hermann Bondi, Fred Hoyle, and Thomas Gold—would call it the perfect cosmological principle and carry its implications in the 1950s to the ultimate refinement in the so-called steady-state theory.

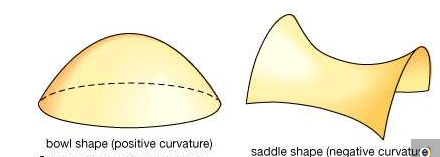
To his great chagrin Einstein found in 1917 that with his three adopted assumptions, his equations of general relativity—as originally written down—had no meaningful solutions. To obtain a solution, Einstein realized that he had to add to his equations an extra term, which came to be called the cosmological constant. If one speaks in Newtonian terms, the cosmological constant could be interpreted as a repulsive force of unknown origin that could exactly balance the attraction of gravitation of all the matter in Einstein’s closed universe and keep it from moving. The inclusion of such a term in a more general context, however, meant that the universe in the absence of any mass-energy (i.e., consisting of a vacuum) would not have a space-time structure that was flat (i.e., would not have satisfied the dictates of special relativity exactly). Einstein was prepared to make such a sacrifice only very reluctantly, and, when he later learned of Hubble’s discovery of the expansion of the universe and realized that he could have predicted it had he only had more faith in the original form of his equations, he regretted the introduction of the cosmological constant as the “biggest blunder” of his life. Ironically, observations of distant supernovas have shown the existence of dark energy, a repulsive force that is the dominant component of the universe.

**De Sitter’s model**

It was also in 1917 that the Dutch astronomer Willem de Sitter recognized that he could obtain a static cosmological model differing from Einstein’s simply by removing all matter. The solution remains stationary essentially because there is no matter to move about. If some test particles are reintroduced into the model, the cosmological term would propel them away from each other. Astronomers now began to wonder if this effect might not underlie the recession of the spiral galaxies.

**Friedmann-Lemaître models**

In 1922 Aleksandr A. Friedmann, a Russian meteorologist and mathematician, and in 1927 Georges Lemaître, a Belgian cleric, independently discovered solutions to Einstein’s equations that contained realistic amounts of matter. These evolutionary models correspond to big bang cosmologies. Friedmann and Lemaître adopted Einstein’s assumption of spatial homogeneity and isotropy (the cosmological principle). They rejected, however, his assumption of time independence and considered both positively curved spaces (“closed” universes) as well as negatively curved spaces (“open” universes). The difference between the approaches of Friedmann and Lemaître is that the former set the cosmological constant equal to zero, whereas the latter retained the possibility that it might have a nonzero value. To simplify the discussion, only the Friedmann models are considered here.



Intrinsic curvature of a surface.

The decision to abandon a static model meant that the Friedmann models evolve with time. As such, neighbouring pieces of matter have recessional (or contractional) phases when they separate from (or approach) one another with an apparent velocity that increases linearly with increasing distance. Friedmann’s models thus anticipated Hubble’s law before it had been formulated on an observational basis. It was Lemaître, however, who had the good fortune of deriving the results at the time when the recession of the galaxies was being recognized as a fundamental cosmological observation, and it was he who clarified the theoretical basis for the phenomenon.

The geometry of space in Friedmann’s closed models is similar to that of Einstein’s original model; however, there is a curvature to time as well as one to space. Unlike Einstein’s model, where time runs eternally at each spatial point on an uninterrupted horizontal line that extends infinitely into the past and future, there is a beginning and end to time in Friedmann’s version of a closed universe when material expands from or is recompressed to infinite densities. These instants are called the instants of the “big bang” and the “big squeeze,” respectively. The global space-time diagram for the middle half of the expansion-compression phases can be depicted as a barrel lying on its side. The space axis corresponds again to any one direction in the universe, and it wraps around the barrel. Through each spatial point runs a time axis that extends along the length of the barrel on its (space-time) surface. Because the barrel is curved in both space and time, the little squares in the grid of the curved sheet of graph paper marking the space-time surface are of nonuniform size, stretching to become bigger when the barrel broadens (universe expands) and shrinking to become smaller when the barrel narrows (universe contracts).

It should be remembered that only the surface of the barrel has physical significance; the dimension off the surface toward the axle of the barrel represents the fourth spatial dimension, which is not part of the real three-dimensional world. The space axis circles the barrel and closes upon itself after traversing a circumference equal to 2πR, where R, the radius of the universe (in the fourth dimension), is now a function of the time t. In a closed Friedmann model, R starts equal to zero at time t = 0 (not shown in barrel diagram), expands to a maximum value at time t = tm (the middle of the barrel), and recontracts to zero (not shown) at time t = 2tm, with the value of tm dependent on the total amount of mass that exists in the universe.

Imagine now that galaxies reside on equally spaced tick marks along the space axis. Each galaxy on average does not move spatially with respect to its tick mark in the spatial (ringed) direction but is carried forward horizontally by the march of time. The total number of galaxies on the spatial ring is conserved as time changes, and therefore their average spacing increases or decreases as the total circumference 2πR on the ring increases or decreases (during the expansion or contraction phases). Thus, without in a sense actually moving in the spatial direction, galaxies can be carried apart by the expansion of space itself. From this point of view, the recession of galaxies is not a “velocity” in the usual sense of the word. For example, in a closed Friedmann model, there could be galaxies that started, when R was small, very close to the Milky Way system on the opposite side of the universe. Now, 1010 years later, they are still on the opposite side of the universe but at a distance much greater than 1010 light-years away. They reached those distances without ever having had to move (relative to any local observer) at speeds faster than light—indeed, in a sense without having had to move at all. The separation rate of nearby galaxies can be thought of as a velocity without confusion in the sense of Hubble’s law, if one wants, but only if the inferred velocity is much less than the speed of light.

On the other hand, if the recession of the galaxies is not viewed in terms of a velocity, then the cosmological redshift cannot be viewed as a Doppler shift. How, then, does it arise? The answer is contained in the barrel diagram when one notices that, as the universe expands, each small cell in the space-time grid also expands. Consider the propagation of electromagnetic radiation whose wavelength initially spans exactly one cell length (for simplicity of discussion), so that its head lies at a vertex and its tail at one vertex back. Suppose an elliptical galaxy emits such a wave at some time t1. The head of the wave propagates from corner to corner on the little square grids that look locally flat, and the tail propagates from corner to corner one vertex back. At a later time t2, a spiral galaxy begins to intercept the head of the wave. At time t2, the tail is still one vertex back, and therefore the wave train, still containing one wavelength, now spans one current spatial grid spacing. In other words, the wavelength has grown in direct proportion to the linear expansion factor of the universe. Since the same conclusion would have held if n wavelengths had been involved instead of one, all electromagnetic radiation from a given object will show the same cosmological redshift if the universe (or, equivalently, the average spacing between galaxies) was smaller at the epoch of transmission than at the epoch of reception. Each wavelength will have been stretched in direct proportion to the expansion of the universe in between.

A nonzero peculiar velocity for an emitting galaxy with respect to its local cosmological frame can be taken into account by Doppler-shifting the emitted photons before applying the cosmological redshift factor; i.e., the observed redshift would be a product of two factors. When the observed redshift is large, one usually assumes that the dominant contribution is of cosmological origin. When this assumption is valid, the redshift is a monotonic function of both distance and time during the expansional phase of any cosmological model. Thus, astronomers often use the redshift z as a shorthand indicator of both distance and elapsed time. Following from this, the statement “object X lies at z = a” means that “object X lies at a distance associated with redshift a”; the statement “event Y occurred at redshift z = b” means that “event Y occurred a time ago associated with redshift b.”

The open Friedmann models differ from the closed models in both spatial and temporal behaviour. In an open universe the total volume of space and the number of galaxies contained in it are infinite. The three-dimensional spatial geometry is one of uniform negative curvature in the sense that, if circles are drawn with very large lengths of string, the ratio of circumferences to lengths of string are greater than 2π. The temporal history begins again with expansion from a big bang of infinite density, but now the expansion continues indefinitely, and the average density of matter and radiation in the universe would eventually become vanishingly small. Time in such a model has a beginning but no end.

**The Einstein–de Sitter universe**

In 1932 Einstein and de Sitter proposed that the cosmological constant should be set equal to zero, and they derived a homogeneous and isotropic model that provides the separating case between the closed and open Friedmann models; i.e., Einstein and de Sitter assumed that the spatial curvature of the universe is neither positive nor negative but rather zero. The spatial geometry of the Einstein–de Sitter universe is Euclidean (infinite total volume), but space-time is not globally flat (i.e., not exactly the space-time of special relativity). Time again commences with a big bang and the galaxies recede forever, but the recession rate (Hubble’s “constant”) asymptotically coasts to zero as time advances to infinity. Because the geometry of space and the gross evolutionary properties are uniquely defined in the Einstein–de Sitter model, many people with a philosophical bent long considered it the most fitting candidate to describe the actual universe.

**Bound and unbound universes and the closure density**

The different separation behaviours of galaxies at large timescales in the Friedmann closed and open models and the Einstein–de Sitter model allow a different classification scheme than one based on the global structure of space-time. The alternative way of looking at things is in terms of gravitationally bound and unbound systems: closed models where galaxies initially separate but later come back together again represent bound universes; open models where galaxies continue to separate forever represent unbound universes; the Einstein–de Sitter model where galaxies separate forever but slow to a halt at infinite time represents the critical case.

How the relative size of the universe changes with time in four different models. The red line shows a universe devoid of matter, with constant expansion. Pink shows a collapsing universe, with six times the critical density of matter. Green shows a model favoured until 1998, with exactly the critical density and a universe 100 percent matter. Blue shows the currently favoured scenario, with exactly the critical density, of which 27 percent is visible and dark matter and 73 percent is dark energy.

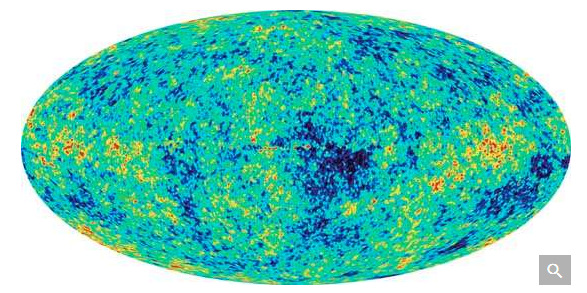
How the relative size of the universe changes with time in four different models. The red line shows a universe devoid of matter, with constant expansion. Pink shows a collapsing universe, with six times the critical density of matter. Green shows a model favoured until 1998, with exactly the critical density and a universe 100 percent matter. Blue shows the currently favoured scenario, with exactly the critical density, of which 27 percent is visible and dark matter and 73 percent is dark energy.

The advantage of this alternative view is that it focuses attention on local quantities where it is possible to think in the simpler terms of Newtonian physics—attractive forces, for example. In this picture it is intuitively clear that the feature that should distinguish whether or not gravity is capable of bringing a given expansion rate to a halt depends on the amount of mass (per unit volume) present. This is indeed the case; the Newtonian and relativistic formalisms give the same criterion for the critical, or closure, density (in mass equivalent of matter and radiation) that separates closed or bound universes from open or unbound ones. If Hubble’s constant at the present epoch is denoted as H0, then the closure density (corresponding to an Einstein–de Sitter model) equals 3H02/8πG, where G is the universal gravitational constant in both Newton’s and Einstein’s theories of gravity. The numerical value of Hubble’s constant H0 is 22 kilometres per second per million light-years; the closure density then equals 10−29 gram per cubic centimetre, the equivalent of about six hydrogen atoms on average per cubic metre of cosmic space. If the actual cosmic average is greater than this value, the universe is bound (closed) and, though currently expanding, will end in a crush of unimaginable proportion. If it is less, the universe is unbound (open) and will expand forever. The result is intuitively plausible since the smaller the mass density, the smaller the role for gravitation, so the more the universe will approach free expansion (assuming that the cosmological constant is zero).

The mass in galaxies observed directly, when averaged over cosmological distances, is estimated to be only a few percent of the amount required to close the universe. The amount contained in the radiation field (most of which is in the cosmic microwave background) contributes negligibly to the total at present. If this were all, the universe would be open and unbound. However, the dark matter that has been deduced from various dynamic arguments is about 23 percent of the universe, and dark energy supplies the remaining amount, bringing the total average mass density up to 100 percent of the closure density.

**The Hot Big Bang**

Given the measured radiation temperature of 2.735 kelvins (K), the energy density of the cosmic microwave background can be shown to be about 1,000 times smaller than the average rest-energy density of ordinary matter in the universe. Thus, the current universe is matter-dominated. If one goes back in time to redshift z, the average number densities of particles and photons were both bigger by the same factor (1 + z)3 because the universe was more compressed by this factor, and the ratio of these two numbers would have maintained its current value of about one hydrogen nucleus, or proton, for every 109 photons. The wavelength of each photon, however, was shorter by the factor 1 + z in the past than it is now; therefore, the energy density of radiation increases faster by one factor of 1 + z than the rest-energy density of matter. Thus, the radiation energy density becomes comparable to the energy density of ordinary matter at a redshift of about 1,000. At redshifts larger than 10,000, radiation would have dominated even over the dark matter of the universe. Between these two values radiation would have decoupled from matter when hydrogen recombined. It is not possible to use photons to observe redshifts larger than about 1,090, because the cosmic plasma at temperatures above 4,000 K is essentially opaque before recombination. One can think of the spherical surface as an inverted “photosphere” of the observable universe. This spherical surface of last scattering probably has slight ripples in it that account for the slight anisotropies observed in the cosmic microwave background today. In any case, the earliest stages of the universe’s history—for example, when temperatures were 109 K and higher—cannot be examined by light received through any telescope. Clues must be sought by comparing the matter content with theoretical calculations.



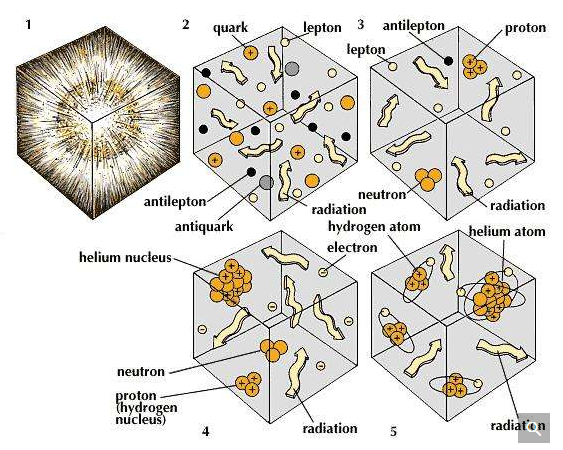
Wilkinson Microwave Anisotropy Probe

A full-sky map produced by the Wilkinson Microwave Anisotropy Probe (WMAP) showing cosmic background radiation, a very uniform glow of microwaves emitted by the infant universe more than 13 billion years ago. Colour differences indicate tiny fluctuations in the intensity of the radiation, a result of tiny variations in the density of matter in the early universe. According to inflation theory, these irregularities were the "seeds" that became the galaxies. WMAP's data support the big bang and inflation models.

For this purpose, fortunately, the cosmological evolution of model universes is especially simple and amenable to computation at redshifts much larger than 10,000 (or temperatures substantially above 30,000 K) because the physical properties of the dominant component, photons, then are completely known. In a radiation-dominated early universe, for example, the radiation temperature T is very precisely known as a function of the age of the universe, the time t after the big bang.

**Primordial nucleosynthesis**

According to the considerations outlined above, at a time t less than 10-4 seconds, the creation of matter-antimatter pairs would have been in thermodynamic equilibrium with the ambient radiation field at a temperature T of about 1012 K. Nevertheless, there was a slight excess of matter particles (e.g., protons) compared to antimatter particles (e.g., antiprotons) of roughly a few parts in 109. This is known because, as the universe aged and expanded, the radiation temperature would have dropped and each antiproton and each antineutron would have annihilated with a proton and a neutron to yield two gamma rays; and later each antielectron would have done the same with an electron to give two more gamma rays. After annihilation, however, the ratio of the number of remaining protons to photons would be conserved in the subsequent expansion to the present day. Since that ratio is known to be one part in 109, it is easy to work out that the original matter-antimatter asymmetry must have been a few parts per 109.



Immediately after the big bang (1), the universe was filled with a dense “soup” of subatomic particles (2), called quarks and leptons (such as electrons), and their antiparticle equivalents. By 0.01 second after the big bang (3), some of the quarks had united to form neutrons and protons. (After another 2 seconds, the only leptons remaining were electrons; the antiparticles had been annihilated.) After 3.5 minutes (4), hydrogen and helium nuclei had formed. After a million years (5), the universe was populated with hydrogen and helium atoms, the raw material of stars and galaxies. The initial radiation from the big bang had grown less energetic.

In any case, after proton-antiproton and neutron-antineutron annihilation but before electron-antielectron annihilation, it is possible to calculate that for every excess neutron there were about five excess protons in thermodynamic equilibrium with one another through neutrino and antineutrino interactions at a temperature of about 1010 K. When the universe reached an age of a few seconds, the temperature would have dropped significantly below 1010 K, and electron-antielectron annihilation would have occurred, liberating the neutrinos and antineutrinos to stream freely through the universe. With no neutrino-antineutrino reactions to replenish their supply, the neutrons would have started to decay with a half-life of 10.6 minutes to protons and electrons (and antineutrinos). However, at an age of 1.5 minutes, well before neutron decay went to completion, the temperature would have dropped to 109 K, low enough to allow neutrons to be captured by protons to form a nucleus of heavy hydrogen, or deuterium. (Before that time, the reaction could still have taken place, but the deuterium nucleus would immediately have broken up under the prevailing high temperatures.) Once deuterium had formed, a very fast chain of reactions set in, quickly assembling most of the neutrons and deuterium nuclei with protons to yield helium nuclei. If the decay of neutrons is ignored, an original mix of 10 protons and two neutrons (one neutron for every five protons) would have assembled into one helium nucleus (two protons plus two neutrons), leaving more than eight protons (eight hydrogen nuclei). This amounts to a helium-mass fraction of 4/12 = 1/3—i.e., 33 percent. A more sophisticated calculation that takes into account the concurrent decay of neutrons and other complications yields a helium-mass fraction in the neighbourhood of 25 percent and a hydrogen-mass fraction of 75 percent, which are close to the deduced primordial values from astronomical observations. This agreement provides one of the primary successes of hot big bang theory.

**The deuterium abundance**

Not all of the deuterium formed by the capture of neutrons by protons would be further reacted to produce helium. A small residual can be expected to remain, the exact fraction depending sensitively on the density of ordinary matter existing in the universe when the universe was a few minutes old. The problem can be turned around: given measured values of the deuterium abundance (corrected for various effects), what density of ordinary matter needs to be present at a temperature of 109 K so that the nuclear reaction calculations will reproduce the measured deuterium abundance? The answer is known, and this density of ordinary matter can be expanded by simple scaling relations from a radiation temperature of 109 K to one of 2.735 K. This yields a predicted present density of ordinary matter and can be compared with the density inferred to exist in galaxies when averaged over large regions. The two numbers are within a factor of a few of each other. In other words, the deuterium calculation implies much of the ordinary matter in the universe has already been seen in observable galaxies. Ordinary matter cannot be the hidden mass of the universe

**The Very Early Universe**

**Inhomogeneous nucleosynthesis**

One possible modification concerns models of so-called inhomogeneous nucleosynthesis. The idea is that in the very early universe (the first microsecond) the subnuclear particles that later made up the protons and neutrons existed in a free state as a quark-gluon plasma. As the universe expanded and cooled, this quark-gluon plasma would undergo a phase transition and become confined to protons and neutrons (three quarks each). In laboratory experiments of similar phase transitions—for example, the solidification of a liquid into a solid—involving two or more substances, the final state may contain a very uneven distribution of the constituent substances, a fact exploited by industry to purify certain materials. Some astrophysicists have proposed that a similar partial separation of neutrons and protons may have occurred in the very early universe. Local pockets where protons abounded may have few neutrons and vice versa for where neutrons abounded. Nuclear reactions may then have occurred much less efficiently per proton and neutron nucleus than accounted for by standard calculations, and the average density of matter may be correspondingly increased—perhaps even to the point where ordinary matter can close the present-day universe. Unfortunately, calculations carried out under the inhomogeneous hypothesis seem to indicate that conditions leading to the correct proportions of deuterium and helium-4 produce too much primordial lithium-7 to be compatible with measurements of the atmospheric compositions of the oldest stars.

**Matter-antimatter asymmetry**

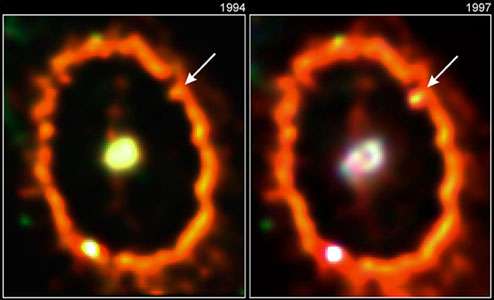
A curious number that appeared in the above discussion was the few parts in 109 asymmetry initially between matter and antimatter (or equivalently, the ratio 10−9 of protons to photons in the present universe). What is the origin of such a number—so close to zero yet not exactly zero?

At one time the question posed above would have been considered beyond the ken of physics, because the net “baryon” number (for present purposes, protons and neutrons minus antiprotons and antineutrons) was thought to be a conserved quantity. Therefore, once it exists, it always exists, into the indefinite past and future. Developments in particle physics during the 1970s, however, suggested that the net baryon number may in fact undergo alteration. It is certainly very nearly maintained at the relatively low energies accessible in terrestrial experiments, but it may not be conserved at the almost arbitrarily high energies with which particles may have been endowed in the very early universe.

An analogy can be made with the chemical elements. In the 19th century most chemists believed the elements to be strictly conserved quantities; although oxygen and hydrogen atoms can be combined to form water molecules, the original oxygen and hydrogen atoms can always be recovered by chemical or physical means. However, in the 20th century with the discovery and elucidation of nuclear forces, chemists came to realize that the elements are conserved if they are subjected only to chemical forces (basically electromagnetic in origin); they can be transmuted by the introduction of nuclear forces, which enter characteristically only when much higher energies per particle are available than in chemical reactions.

In a similar manner it turns out that at very high energies new forces of nature may enter to transmute the net baryon number. One hint that such a transmutation may be possible lies in the remarkable fact that a proton and an electron seem at first sight to be completely different entities, yet they have, as far as one can tell to very high experimental precision, exactly equal but opposite electric charges. Is this a fantastic coincidence, or does it represent a deep physical connection? A connection would obviously exist if it can be shown, for example, that a proton is capable of decaying into a positron (an antielectron) plus electrically neutral particles. Should this be possible, the proton would necessarily have the same charge as the positron, for charge is exactly conserved in all reactions. In turn, the positron would necessarily have the opposite charge of the electron, as it is its antiparticle. Indeed, in some sense the proton (a baryon) can even be said to be merely the “excited” version of an antielectron (an “antilepton”).

Motivated by this line of reasoning, experimental physicists searched hard during the 1980s for evidence of proton decay. They found none and set a lower limit of 1032 years for the lifetime of the proton if it is unstable. This value is greater than what theoretical physicists had originally predicted on the basis of early unification schemes for the forces of nature. Later versions can accommodate the data and still allow the proton to be unstable. Despite the inconclusiveness of the proton-decay experiments, some of the apparatuses were eventually put to good astronomical use. They were converted to neutrino detectors and provided valuable information on the solar neutrino problem, as well as giving the first positive recordings of neutrinos from a supernova explosion (namely, supernova 1987A).



Supernova 1987A in the Large Magellanic Cloud.This picture shows the faint outer rings and bright inner ring characteristic of an hourglass nebula.

With respect to the cosmological problem of the matter-antimatter asymmetry, one theoretical approach is founded on the idea of a grand unified theory (GUT), which seeks to explain the electromagnetic, weak nuclear, and strong nuclear forces as a single grand force of nature. This approach suggests that an initial collection of very heavy particles, with zero baryon and lepton number, may decay into many lighter particles (baryons and leptons) with the desired average for the net baryon number (and net lepton number) of a few parts per 109. This event is supposed to have occurred at a time when the universe was perhaps 10−35 second old.

Another approach to explaining the asymmetry relies on the process of CP violation, or violation of the combined conservation laws associated with charge conjugation (C) and parity (P) by the weak force, which is responsible for reactions such as the radioactive decay of atomic nuclei. Charge conjugation implies that every charged particle has an oppositely charged antimatter counterpart, or antiparticle. Parity conservation means that left and right and up and down are indistinguishable in the sense that an atomic nucleus emits decay products up as often as down and left as often as right. With a series of debatable but plausible assumptions, it can be demonstrated that the observed imbalance or asymmetry in the matter-antimatter ratio may have been produced by the occurrence of CP violation in the first seconds after the big bang. CP violation is expected to be more prominent in the decay of particles known as B-mesons. In 2010, scientists at the Fermi National Acclerator Laboratory in Batavia, Ill., finally detected a slight preference for B-mesons to decay into muons rather than anti-muons.

**Superunification and the Planck era**

Why should a net baryon fraction initially of zero be more appealing aesthetically than 10−9? The underlying motivation here is perhaps the most ambitious undertaking ever attempted in the history of science—the attempt to explain the creation of truly everything from literally nothing. In other words, is the creation of the entire universe from a vacuum possible?

The evidence for such an event lies in another remarkable fact. It can be estimated that the total number of protons in the observable universe is an integer 80 digits long. No one of course knows all 80 digits, but for the argument about to be presented, it suffices only to know that they exist. The total number of electrons in the observable universe is also an integer 80 digits long. In all likelihood these two integers are equal, digit by digit—if not exactly, then very nearly so. This inference comes from the fact that, as far as astronomers can tell, the total electric charge in the universe is zero (otherwise electrostatic forces would overwhelm gravitational forces). Is this another coincidence, or does it represent a deeper connection? The apparent coincidence becomes trivial if the entire universe was created from a vacuum since a vacuum has by definition zero electric charge. It is a truism that one cannot get something for nothing. The interesting question is whether one can get everything for nothing. Clearly, this is a very speculative topic for scientific investigation, and the ultimate answer depends on a sophisticated interpretation of what “nothing” means.

The words “nothing,” “void,” and “vacuum” usually suggest uninteresting empty space. To modern quantum physicists, however, the vacuum has turned out to be rich with complex and unexpected behaviour. They envisage it as a state of minimum energy where quantum fluctuations, consistent with the uncertainty principle of the German physicist Werner Heisenberg, can lead to the temporary formation of particle-antiparticle pairs. In flat space-time, destruction follows closely upon creation (the pairs are said to be virtual) because there is no source of energy to give the pair permanent existence. All the known forces of nature acting between a particle and antiparticle are attractive and will pull the pair together to annihilate one another. In the expanding space-time of the very early universe, however, particles and antiparticles may separate and become part of the observable world. In other words, sharply curved space-time can give rise to the creation of real pairs with positive mass-energy, a fact first demonstrated in the context of black holes by the English astrophysicist Stephen W. Hawking.

Yet Einstein’s picture of gravitation is that the curvature of space-time itself is a consequence of mass-energy. Now, if curved space-time is needed to give birth to mass-energy and if mass-energy is needed to give birth to curved space-time, which came first, space-time or mass-energy? The suggestion that they both rose from something still more fundamental raises a new question: What is more fundamental than space-time and mass-energy? What can give rise to both mass-energy and space-time? No one knows the answer to this question, and perhaps some would argue that the answer is not to be sought within the boundaries of natural science.

Hawking and the American cosmologist James B. Hartle have proposed that it may be possible to avert a beginning to time by making it go imaginary (in the sense of the mathematics of complex numbers) instead of letting it suddenly appear or disappear. Beyond a certain point in their scheme, time may acquire the characteristic of another spatial dimension rather than refer to some sort of inner clock. Another proposal states that, when space and time approach small enough values (the Planck values; see below), quantum effects make it meaningless to ascribe any classical notions to their properties. The most promising approach to describe the situation comes from the theory of “superstrings.”

Superstrings represent one example of a class of attempts, generically classified as superunification theory, to explain the four known forces of nature—gravitational, electromagnetic, weak, and strong—on a single unifying basis. Common to all such schemes are the postulates that quantum mechanics and special relativity underlie the theoretical framework. Another common feature is supersymmetry, the notion that particles with half-integer values of the spin angular momentum (fermions) can be transformed into particles with integer spins (bosons).

The distinguishing feature of superstring theory is the postulate that elementary particles are not mere points in space but have linear extension. The characteristic linear dimension is given as a certain combination of the three most fundamental constants of nature: (1) Planck’s constant h (named after the German physicist Max Planck, the founder of quantum physics), (2) the speed of light c, and (3) the universal gravitational constant G. The combination, called the Planck length (Gh/c3)1/2, equals roughly 10−33 cm, far smaller than the distances to which elementary particles can be probed in particle accelerators on Earth.

The energies needed to smash particles to within a Planck length of each other were available to the universe at a time equal to the Planck length divided by the speed of light. This time, called the Planck time (Gh/c5)1/2, equals approximately 10−43 second. At the Planck time, the mass density of the universe is thought to approach the Planck density, c5/hG2, roughly 1093 grams per cubic centimetre. Contained within a Planck volume is a Planck mass (hc/G)1/2, roughly 10−5 gram. An object of such mass would be a quantum black hole, with an event horizon close to both its own Compton length (distance over which a particle is quantum mechanically “fuzzy”) and the size of the cosmic horizon at the Planck time. Under such extreme conditions, space-time cannot be treated as a classical continuum and must be given a quantum interpretation.

The latter is the goal of the superstring theory, which has as one of its features the curious notion that the four space-time dimensions (three space dimensions plus one time dimension) of the familiar world may be an illusion. Real space-time, in accordance with this picture, has 26 or 10 space-time dimensions, but all of these dimensions except the usual four are somehow compacted or curled up to a size comparable to the Planck scale. Thus has the existence of these other dimensions escaped detection. It is presumably only during the Planck era, when the usual four space-time dimensions acquire their natural Planck scales, that the existence of what is more fundamental than the usual ideas of mass-energy and space-time becomes fully revealed. Unfortunately, attempts to deduce anything more quantitative or physically illuminating from the theory have bogged down in the intractable mathematics of this difficult subject. At the present time superstring theory remains more of an enigma than a solution.

**Inflation**

One of the more enduring contributions of particle physics to cosmology is the prediction of inflation by the American physicist Alan Guth and others. The basic idea is that at high energies matter is better described by fields than by classical means. The contribution of a field to the energy density (and therefore the mass density) and the pressure of the vacuum state need not have been zero in the past, even if it is today. During the time of superunification (Planck era, 10−43 second) or grand unification (GUT era, 10−35 second), the lowest-energy state for this field may have corresponded to a “false vacuum,” with a combination of mass density and negative pressure that results gravitationally in a large repulsive force. In the context of Einstein’s theory of general relativity, the false vacuum may be thought of alternatively as contributing a cosmological constant about 10100 times larger than it can possibly be today. The corresponding repulsive force causes the universe to inflate exponentially, doubling its size roughly once every 10−43 or 10−35 second. After at least 85 doublings, the temperature, which started out at 1032 or 1028 K, would have dropped to very low values near absolute zero. At low temperatures the true vacuum state may have lower energy than the false vacuum state, in an analogous fashion to how solid ice has lower energy than liquid water. The supercooling of the universe may therefore have induced a rapid phase transition from the false vacuum state to the true vacuum state, in which the cosmological constant is essentially zero. The transition would have released the energy differential (akin to the “latent heat” released by water when it freezes), which reheats the universe to high temperatures. From this temperature bath and the gravitational energy of expansion would then have emerged the particles and antiparticles of noninflationary big bang cosmologies.

Cosmic inflation serves a number of useful purposes. First, the drastic stretching during inflation flattens any initial space curvature, and so the universe after inflation will look exceedingly like an Einstein–de Sitter universe. Second, inflation so dilutes the concentration of any magnetic monopoles appearing as “topological knots” during the GUT era that their cosmological density will drop to negligibly small and acceptable values. Finally, inflation provides a mechanism for understanding the overall isotropy of the cosmic microwave background because the matter and radiation of the entire observable universe were in good thermal contact (within the cosmic event horizon) before inflation and therefore acquired the same thermodynamic characteristics. Rapid inflation carried different portions outside their individual event horizons. When inflation ended and the universe reheated and resumed normal expansion, these different portions, through the natural passage of time, reappeared on our horizon. And through the observed isotropy of the cosmic microwave background, they are inferred still to have the same temperatures. Finally, slight anisotropies in the cosmic microwave background occurred because of quantum fluctuations in the mass density. The amplitudes of these small (adiabatic) fluctuations remained independent of comoving scale during the period of inflation. Afterward they grew gravitationally by a constant factor until the recombination era. Cosmic microwave photons seen from the last scattering surface should therefore exhibit a scale-invariant spectrum of fluctuations, which is exactly what the Cosmic Background Explorer satellite observed.

**How the polarization of light influenced the birth of the universe.**

As influential as inflation has been in guiding modern cosmological thought, it has not resolved all internal difficulties. The most serious concerns the problem of a “graceful exit.” Unless the effective potential describing the effects of the inflationary field during the GUT era corresponds to an extremely gently rounded hill (from whose top the universe rolls slowly in the transition from the false vacuum to the true vacuum), the exit to normal expansion will generate so much turbulence and inhomogeneity (via violent collisions of “domain walls” that separate bubbles of true vacuum from regions of false vacuum) as to make inexplicable the small observed amplitudes for the anisotropy of the cosmic microwave background radiation. Arranging a tiny enough slope for the effective potential requires a degree of fine-tuning that most cosmologists find philosophically objectionable.

**Steady State Theory And Other Alternative Cosmologies**

Big bang cosmology, augmented by the ideas of inflation, remains the theory of choice among nearly all astronomers, but, apart from the difficulties discussed above, no consensus has been reached concerning the origin in the cosmic gas of fluctuations thought to produce the observed galaxies, clusters, and superclusters. Most astronomers would interpret these shortcomings as indications of the incompleteness of the development of the theory, but it is conceivable that major modifications are needed.

An early problem encountered by big bang theorists was an apparent large discrepancy between the Hubble time and other indicators of cosmic age. This discrepancy was resolved by revision of Hubble’s original estimate for H0, which was about an order of magnitude too large owing to confusion between Population I and II variable stars and between H II regions and bright stars. However, the apparent difficulty motivated Bondi, Hoyle, and Gold to offer the alternative theory of steady state cosmology in 1948.

An overview of the 20th-century controversy concerning the origin of the universe: the big-bang model was disparaged by Fred Hoyle, but Edwin Hubble's findings supported it.

By that year, of course, the universe was known to be expanding; therefore, the only way to explain a constant (steady state) matter density was to postulate the continuous creation of matter to offset the attenuation caused by the cosmic expansion. This aspect was physically very unappealing to many people, who consciously or unconsciously preferred to have all creation completed in virtually one instant in the big bang. In the steady state theory the average age of matter in the universe is one-third the Hubble time, but any given galaxy could be older or younger than this mean value. Thus, the steady state theory had the virtue of making very specific predictions, and for this reason it was vulnerable to observational disproof.

The first blow was delivered by British astronomer Martin Ryle’s counts of extragalactic radio sources during the 1950s and ’60s. These counts involved the same methods discussed above for the star counts by Dutch astronomer Jacobus Kapteyn and the galaxy counts by Hubble except that radio telescopes were used. Ryle found more radio galaxies at large distances from Earth than can be explained under the assumption of a uniform spatial distribution no matter which cosmological model was assumed, including that of steady state. This seemed to imply that radio galaxies must evolve over time in the sense that there were more powerful sources in the past (and therefore observable at large distances) than there are at present. Such a situation contradicts a basic tenet of the steady state theory, which holds that all large-scale properties of the universe, including the population of any subclass of objects like radio galaxies, must be constant in time.

The second blow came in 1965 with the discovery of the cosmic microwave background radiation. Though it has few adherents today, the steady state theory is credited as having been a useful idea for the development of modern cosmological thought as it stimulated much work in the field.

At various times, other alternative theories have also been offered as challenges to the prevailing view of the origin of the universe in a hot big bang: the cold big bang theory (to account for galaxy formation), symmetric matter-antimatter cosmology (to avoid an asymmetry between matter and antimatter), variable G cosmology (to explain why the gravitational constant is so small), tired-light cosmology (to explain redshift), and the notion of shrinking atoms in a nonexpanding universe (to avoid the singularity of the big bang). The motivation behind these suggestions is, as indicated in the parenthetical comments, to remedy some perceived problem in the standard picture. Yet, in most cases, the cure offered is worse than the disease, and none of the mentioned alternatives has gained much of a following. The hot big bang theory has ascended to primacy because, unlike its many rivals, it attempts to address not isolated individual facts but a whole panoply of cosmological issues. And, although some sought-after results remain elusive, no glaring weakness has yet been uncovered.