The universe is big. We often say that we live in a small world, but the 25,000-mile trek around the planet Earth is still a longer trip than most of us have ever attempted. The farthest location ever reached by humans is the moon, about 240,000 miles from the earth. While the lunar landing was a spectacular achievement, we would have to travel 400 times farther if we wished to extend our exploration to the sun. This glowing sphere is so large that if a map of the earth were drawn to cover the sun’s surface, the entire area of our planet would fit comfortably within the outline of the Dominican Republic. The sun is not unique, but is one of many stars. A journey to our nearest stellar neighbor, a three-star system called Alpha Centauri, would carry our astronauts a hundred million times farther than a trip to the moon, a distance so great that even light requires four years to traverse it. If the astronauts looked homeward from Alpha Centauri, the separation between the earth and moon would look no bigger than a thumbtack viewed from 400 miles away.

The sun and Alpha Centauri are lazily circling in a spiral galaxy of stars, making four revolutions every billion years. The galaxy, known as the Milky Way, has a diameter about 20,000 times larger than the distance from earth to Alpha Centauri. The sun would be hard to find on a complete map of the galaxy, since the sun is only one of the hundred billion stars that make up the swirling disk. To see how large this number is, we can imagine trying to count the stars. At one star every second, day and night, we and our descendants would be rattling off integers for over three thousand years before the last star of the Milky Way would be tallied. If we stacked a hundred billion pieces of paper, the height of the pile would reach 6000 miles into space.
Even with its hundred billion stars, however, our galaxy is only a puny dot in the realm of the cosmos. Within the range of our telescopes there are a hundred billion galaxies, each one comparable in size to the Milky Way. When our descendants finished counting the stars of our own galaxy, they could begin the three-thousand-year task of counting the multitude of galaxies that are scattered across the vast expanse of the visible universe.

While there are many questions about the universe that cosmologists would like to answer, probably the most fascinating is the most fundamental question of all: Where did all this come from? Almost every human civilization in history has offered an answer to this question in the context of mythology or religion, but until recently the question had been thought to be outside the scope of science. Although the generally accepted big bang theory holds that the observable universe emerged from an explosion some ten to twenty billion years ago, the theory nonetheless assumes that all the matter in the universe was present from the start. The form of the matter may have been different, but it was all there. The classic big bang theory describes the aftermath of the bang, but makes no attempt to describe what "banged," how it "banged," or what caused it to "bang." Nothing can be created from nothing, we were always taught, so there was no hope for a scientific explanation for the actual origin of the matter in the universe.

The difficulty in constructing a scientific theory for the origin of matter stems from a set of rules, called conservation principles, that trace their origin to the very roots of science itself.

"Being is ungenerable and imperishable," wrote Parmenides in about 500 B.C., in a passage that helped to create the philosophical approach that we call science today. This basic idea, that things which exist continue to exist, became a cornerstone of the concept of natural order. Objects would not appear and disappear unpredictably, but instead would evolve continuously according to principles of nature. This notion of continuity in existence became more concrete a century later in the work of Leucippus and Democritus, who advanced the theory that all matter is composed of eternal, indivisible atoms which move through an otherwise empty space. These ideas are reflected strongly in Lucretius' De Rerum Natura (On the Nature of Things), written during the first century B.C., which includes the statement that "Nothing can be created from nothing." Lucretius went on to explain that "Material objects are of two kinds, atoms and compounds of atoms. The atoms themselves cannot be swamped by any force, for they are preserved indefinitely by their absolute solidity."

While the fundamental idea of continuity in existence can be traced to the ancient Greeks, it was not until much later that this line of inquiry evolved into the conservation laws of modern science. An important step occurred in the eighteenth century, with the work of Antoine-Laurent Lavoisier and his contemporaries. In 1772 Lavoisier discovered that when sulphur or phosphorus are burned, the products weigh more than the original material. While this might have been taken as proof that matter is not conserved, Lavoisier instead traced the added matter to a component of the air, which he named oxygen. When the oxygen was taken into account, the initial and final masses were equal. In Lavoisier's words, "nothing is created in the operations either of art or of nature, and it can be taken as an axiom that in every operation an equal quantity of matter exists both before and after the operation." These ideas were soon solidified in the work of the English scientist John Dalton, who in 1803 reintroduced the atomic theory of Leucippus and Democritus, and shortly thereafter produced the first table of atomic weights. By this time the idea of conservation of mass was firmly embedded in the developing scientific tradition.

The other great conservation principle developed during this period was the conservation of energy. This principle is more subtle, and it understandably took longer to develop. Unlike mass, the definition of energy—the conserved quantity—is far from obvious. The word itself is derived from the Greek energia, constructed from the roots en (in) and ergon (work). Roughly speaking, energy is the capacity of any system to do work.

The concept of energy is intimately linked to the study of motion, the main subject of Isaac Newton's legendary Philosophiae Naturalis Principia Mathematica (Mathematical Principles of Natural Philosophy), published in 1687. The laws of motion described by Newton lead directly to the conservation of energy, provided that forces such as friction are excluded from consideration. Newton himself was clearly aware of what we now call the conservation of energy, but he apparently did not appreciate the deep importance that this idea would later acquire. It was the German philosopher Gottfried Wilhelm Leibniz (1646–1716), a contemporary of Newton's, who first realized that energy was sufficiently important to merit a name. He defined the term Vis Viva (Living Force) to measure what is now called kinetic energy, the energy of motion. The kinetic energy of a moving object is equal to half its mass times the square of its velocity. (If the mass is measured in grams and the velocity is measured in centimeters per second, the resulting energy is given in a unit called an erg. An erg is a small amount of energy—a 150-pound person walking at 2 miles per hour has a kinetic energy of 272 million ergs. The conversion to other units of energy is described in Appendix D.)

In most circumstances, however, kinetic energy itself is not conserved, as one can see by imagining a baseball that is held stationary at arm's length and then released. The kinetic energy starts from zero and increases as the
ball gains speed on its downward plunge. Where does this energy come from? The answer can be found by a careful examination of the motion. To eliminate the complication of air friction, consider a weight that is dropped in a vacuum. If the mass is one gram, the kinetic energy is 980 ergs after one centimeter of fall, twice this amount after two centimeters of fall, three times this amount after three centimeters, etc. Since the kinetic energy is determined by the height of the ball, the height must be related to a new form of energy. It is called gravitational potential energy, or just gravitational energy for short. For each centimeter of fall the gravitational energy decreases by 980 ergs, just as the kinetic energy increases by 980 ergs, so the total energy is constant. (If the mass is more or less than one gram, the gravitational energy varies in proportion to the mass.) As long as the gravitational energy is properly included, then the total energy is conserved.

The skeptical reader might be wondering whether the conservation of energy is really a principle of nature, or just a tautological proclamation of obstinate physicists. Suppose, for comparison, that somebody claimed that a child's height was really conserved. Like the conservation of energy, this conservation principle could be made to work by inventing a new form of "height." One could define a "potential height" for the child, which decreases by one inch for each inch the child grows. In this way the "potential height" plus the ruler height will always have the same value. There is an important difference, however, between the conservation of energy and the conservation of a child's height. Both principles are technically correct, but the former can be used to make predictions while the latter cannot. If a ball is thrown into the air, then at any time during its flight one could determine the total energy by measuring both the speed and the height. At any later time one might measure only the speed, and the conservation of energy would allow the height to be calculated. Alternatively, one could measure only the height, and the speed could be calculated. By contrast, the potential height of the child is not directly measurable, and so the child-height conservation principle leads to no predictions whatever.

If one deals with the motion of planets instead of the motion of baseballs, then the force of gravity cannot be treated as constant, as it was in the previous example. The gravitational force between two planets becomes weaker as they are separated, decreasing in inverse proportion to the square of the distance between them. That is, if the distance is doubled, then the gravitational force that each exerts on the other is decreased by a factor of four. In this case the formula for the gravitational energy is less apparent, but it was discovered in 1847 by the German physiologist and physicist, Hermann von Helmholtz [1]. Helmholtz showed that, in the absence of friction, the sum of the kinetic and gravitational potential energies does not change with time.

In the real world, however, there is always at least a small amount of friction, so the sum of the kinetic and potential energies is never quite conserved. A pendulum with friction will gradually come to a halt, so it appears that energy is lost. This is no doubt one of the reasons why the conservation of energy was not given much importance in the early development of mechanics by Newton and his followers. However, once again it is possible to make the conservation principle work, by realizing that there is another form of energy that has to be defined and included in the bookkeeping. The new form of energy is heat, and it was Benjamin Thompson, one of the more colorful scientists of the eighteenth century,* who first showed that mechanical energy can be converted to heat energy with a fixed conversion rate.

While supervising a military arsenal in Munich, Thompson became impressed by the surprising amount of heat that was produced when a brass cannon was bored by a horse-turned machine. He performed an experiment, which he reported to the Royal Society of London in 1798, in which a metal cylinder enclosed in a box containing about two gallons of water was bored until the water became hot. He measured the temperature of the water as it rose with time, until the water started to boil after about two and a half hours. According to Thompson's report to the Royal Society,

It would be difficult to describe the surprise and astonishment expressed in the countenances of the bystanders, on seeing so large a quantity of cold water heated, and actually made to boil, without any fire.

These observations constituted the first measurement of the conversion rate of mechanical energy to heat. Thompson's measurements were improved in 1843 by the English physicist James Prescott Joule, who also measured the conversion of electrical energy into heat.† By the end of the 1840s, the law

---

* Benjamin Thompson was born in Woburn, Massachusetts, in 1753. An unmitigated opportunist, at the age of 19 he married an extremely wealthy 33-year-old widow. He remained loyal to the British crown at the outbreak of the American Revolution, and for a short period served as a British spy. His espionage career had all the flair of an Ian Fleming novel, including the use of secret messages written in invisible ink, and a brief affair with the wife of a leading revolutionary newspaper publisher. In March 1776 he abandoned his own wife and daughter and fled to London, where his credentials as an important loyalist from Massachusetts earned him the position of undersecretary of state for the colonies. He was later knighted by King George III, and then with the Crown's permission served as an advisor to the government of Bavaria. There he performed his famous experiments on heat, and he also brought James Watt's steam engine into common use. He produced inventions including an improved fireplace, a double boiler, a kitchen range, and a drip coffeemaker. In 1791 he became Count von Rumford of the Holy Roman Empire, taking the name from Rumford, Massachusetts (now Concord, New Hampshire), where he had lived for a short time [2].

† Joule's name is used for one of the standard units of energy, equal to 10 million ergs. The well-known unit of electrical power, the watt, is equal to one joule per second.
of conservation of energy had been enunciated clearly by Joule, Helmholtz, and also the German physicist Julius Robert von Mayer. If all forms of energy are included, then the total amount of energy cannot be changed by any physical process.

Thus, by the middle of the nineteenth century the two most basic conservation laws—the conservation of mass and the conservation of energy—were firmly established. In terms of Newton’s laws of motion, it was now understood that fundamental forces, such as gravity, always conserve energy. Those forces which fail to conserve energy, such as friction, do so only because the description is incomplete. As the motion of a pendulum dissipates due to friction, the energy is not really lost. Instead it shows up as the kinetic energy of the random motion of atoms—in the pendulum and in the air surrounding the pendulum—energy which we recognize as heat. If Newton’s laws were applied not to the pendulum as a whole, but instead to the individual atoms in the pendulum and in the air, then energy would always be conserved.

No one in the middle of the nineteenth century, however, could have anticipated the dramatic leap that was to be made by Albert Einstein in 1905. In that year Einstein published his two papers on the subject that we now call special relativity. The first laid out the foundations of the subject, while the second was a short follow-up, only three pages long. In this second paper Einstein derived the now-famous formula, $E = mc^2$. Here $E$ denotes the total energy of any body, $m$ denotes its mass, and $c^2$ denotes the square of the speed of light. According to this formula, mass and energy are equivalent, in the sense that they measure exactly the same thing. The only difference is the factor of $c^2$, which really just means that mass and energy refer to different systems of units. We get a bigger number if we measure a distance in inches than if we measured it in miles, but we understand that they are merely two ways to express the same thing. Similarly, the mass/energy of an object is a single property that we can equally well choose to measure in grams or in ergs.

Thus, there are no longer separate conservation laws for mass and energy. With the advent of special relativity, the two conservation laws meld into one—the conservation of mass/energy. For brevity, the combined conservation principle is often called simply the “conservation of energy.” It follows that any particle—an electron, proton, or neutron, for example—has a significant energy even when it is at rest. This energy is called the rest energy of the particle.

Since $c$ is such a large number, relativity implies that a small amount of mass is equivalent to a very large amount of energy. If $m$ is measured in grams and $c$ is measured in centimeters per second ($c = 2.998 \times 10^{10}$ centimeters per
second*), then the energy $E = mc^2$ is obtained in ergs.\(^1\) If one pound of matter could be converted entirely to energy, the result would be eleven billion kilowatt-hours—enough energy to supply the annual electric power consumption of a million average Americans, two million Europeans or Japanese, or eight million Latin Americans. In an article for *Science Illustrated* in 1946, Einstein himself gave a nontechnical explanation of how such a large amount of energy could have gone undetected for so long:

But if every gram of material contains this tremendous energy, why did it go so long unnoticed? The answer is simple enough: so long as none of the energy is given off externally, it cannot be observed. It is as though a man who is fabulously rich should never spend or give away a cent; no one could tell how rich he was.

In a nuclear reactor only about one tenth of one percent of the mass of fissionable uranium is converted to energy, but the energy output is nonetheless enormous.

Probably the most dramatic challenge in this century to the conservation of energy was the discovery in 1914 by the British physicist, James Chadwick, that energy appears to be lost in a process called beta decay. In this process a radioactive nucleus decays by emitting an electron, which was found to have a range of possible energies extending up to the value predicted by energy conservation. Physicists searched for other ways in which the energy might escape, but by the late 1920s all reasonable possibilities were ruled out [3].

Physicists were puzzled, but in 1931 the Austrian-born physicist Wolfgang Pauli suggested what he called a “desperate solution.” He proposed that beta decay produces a new type of particle that interacts so weakly that it escaped all efforts at detection. Two years later this suggestion was elaborated by the Italian physicist Enrico Fermi, who developed a detailed theory of the decay process. Fermi called the unseen particle the “neutrino,” a word constructed from Italian roots which mean “little neutral one.” (Curiously, Fermi’s now-classic paper on beta decay was rejected by the prominent British journal *Nature*) Since neutrinos interact with normal matter only very rarely, their detection remained a formidable challenge that required incredible ingenuity and patience. It was not until June 14, 1956, a year and half after Fermi’s death, that Pauli received a telegram informing him that his prediction had at last been experimentally confirmed. Twenty-five years after its prediction, the neutrino had been detected by F. L.

---

\[^*\] Note on scientific notation (for the reader who is not familiar with this way of writing numbers): \(10^n\) means a 1 followed by \(n\) zeros, or \(1,000,000,000\). Here 10 is called the exponent. For example, with a negative exponent, \(10^{-n}\) means a decimal point, \(n\) zeros, and then a 1, or .0001.

\[^1\] The conversion to other units of energy is described in Appendix D.

---

Cowan, Jr. and Frederick Reines, working at the Savannah River reactor in South Carolina. (In 1995, after Cowan’s death, Reines was awarded the Nobel Prize in physics for this work.) Nonetheless, in spite of this long gap between the prediction and the observation, physicists had never seriously doubted Pauli’s proposal. Their confidence in the conservation of energy was so strong that they were willing to accept for a quarter of a century, the existence of a particle that was totally unseen.

Finally then, we come to the key question: Given the present understanding of conservation laws, is there any hope for a scientific description of the creation of the universe? If the conservation laws imply that “nothing can be created from nothing,” as Lucretius put it, then how could the universe have come into being? We would be forced to believe either that the universe is eternal, or that it was created by some force that worked outside the restrictions of physical laws. If, however, we seek to understand how the universe might have been created within the context of physical laws, then some loophole has to be found in the age-old dictum of Lucretius.

If the creation of the universe is to be described by physical laws that embody the conservation of energy, then the universe must have the same energy as whatever it was created from. If the universe was created from nothing, then the total energy must be zero. But the universe is clearly filled with energy: the Earth, the Sun, the Milky Way, and the hundred billion galaxies that make up the observable universe clearly contain an unfathomable amount of mass/energy. How, then, is there any hope that the creation of the universe might be described scientifically?

An answer to this question arises from an extraordinary feature of one particular form of energy—gravitational potential energy, which was discussed earlier in the chapter. To understand this feature, however, we must delve just a bit deeper into the description of gravity.

According to Newton, any two objects exert an attractive gravitational force on each other. To find the total gravitational force that pulls on a given object, one computes the sum of the forces exerted on it by every other object.* In principle one has to include the forces exerted by every other object in the universe, but in practice only nearby objects are significant. Newton’s description of gravity is often called an *action-at-a-distance* formulation, since gravity is interpreted as a force that one object exerts on a distant object.

* Forces are added by a procedure called vector addition, which takes into account the directions as well as the strengths of the forces. If two forces point in opposite directions, for example, their effects tend to cancel rather than to reinforce each other.
Most modern physicists, however, think about gravity using an alternative formulation in which forces at a distance are avoided. While maintaining the underlying content of Newton's law, this newer formulation replaces the action-at-a-distance with the notion of a gravitational field. In some contexts this phrase is familiar—an interplanetary space probe, for example, can be said to escape the gravitational field of the earth. More precisely, the physicist uses the concept of a gravitational field to characterize the effect of gravity at each point in space. Whether or not any object is actually located at a given point, one can always ask what force would be experienced by a one-gram mass, if the mass were placed at that point. The answer to this question, usually expressed in a unit of force called a dyne,\textsuperscript{*} defines the gravitational field. The force on a one-gram mass near the surface of the earth is about 980 dynes, so the gravitational field is 980 dynes per gram. At an altitude of 50,000 feet the same one-gram mass would experience a force of only 975 dynes, so the gravitational field at this altitude is 975 dynes per gram.\textsuperscript{1}

The gravitational field is often depicted by drawing gravitational field lines in space, as in Figure 1.2. The field lines point in the direction of the gravitational field, and the strength of the field can be indicated by how closely the lines are spaced. Figure 1.2 shows the gravitational field lines for a sphere: a collection of evenly spaced lines, each pointing directly toward the center of the sphere. The spacing between the lines is closest near the surface of the sphere, indicating that the field is strongest in this region.

In this formulation each particle produces a gravitational field in the space around it, similar to the illustration in Figure 1.2. The field in turn exerts a force on distant particles.

In the discussion of gravitational potential energy earlier in this chapter, we saw that energy can be stored by lifting a weight in the gravitational field of the earth. We know that the energy is not lost, even though it has disappeared from view, because it can be recovered by allowing the weight to fall. This argument, however, tells us nothing about where the energy is stored. Using the field formulation, however, the question has a simple answer: The energy is stored in the gravitational field. In any region of space with a gravitational field, there is a density of energy proportional to the square of the field. By using a beautiful technique first published in 1828 by George

\textsuperscript{*} A dyne is the force necessary to accelerate a mass of one gram at an acceleration of one centimeter per second per second.

\textsuperscript{1} The reader might wonder whether the one-gram mass remains a one-gram mass at 50,000 feet, since its weight—the gravitational pull toward the earth—decreases by about half of a percent. By the standard definition used in physics, mass is a measure of the inertia of an object. The mass of an object is in principle determined by observing the acceleration that is produced by applying a known force. By this definition a one-gram mass remains a one-gram mass, so far as we know, no matter where in the universe it is transported.

\textbf{Figure 1.2} The gravitational field of a sphere. The gravitational field at a specified point measures the force that would be experienced by a mass if it were located at that point. A gravitational field can be depicted by gravitational field lines, as illustrated in the diagram. The lines are shown here as thick rods to make it easier to see their direction.

---

Green, a British baker and self-educated mathematician, one can show that all of the energy that is stored in lifting a weight can be attributed to the increased energy in the gravitational field. Thus, gravitational potential energy is really just the energy of the gravitational field. The energy of the gravitational field is therefore part of the energy of the universe, and must be included in the bookkeeping of the energy conservation law.

Now, we can return to the key question: How is there any hope that the creation of the universe might be described by physical laws consistent with
energy conservation? Answer: the energy stored in the gravitational field is represented by a negative number! That is, the energy stored in a gravitational field is actually less than zero. If negative numbers are allowed, then zero can be obtained by adding a positive number to a negative number of equal magnitude; for example, 7 plus -7 equals zero. It is therefore conceivable that the total energy of the universe is zero. The immense energy that we observe in the form of matter can be canceled by a negative contribution of equal magnitude, coming from the gravitational field. There is no limit to the magnitude of energy in the gravitational field, and hence no limit to the amount of matter/energy that it can cancel.

For the reader interested in learning why the energy of a gravitational field is negative, an argument is presented in Appendix A. A specific example is used to show that no energy is needed to produce a gravitational field, but instead energy is released when a gravitational field is created.

Given this peculiar property of gravity, a scientific description of the creation of the universe is not precluded by the conservation of energy. Other conservation laws also need to be considered, in particular the conservation of a quantity called baryon number, which will be discussed in Chapter 6. But the conclusion will not be changed: The universe could have evolved from absolutely nothing in a manner consistent with all known conservation laws. While no detailed scientific theory of creation is known, the possibility of developing such a theory now appears open.

In the late 1960s, a young assistant professor at Columbia University named Edward P. Tryon attended a seminar given by Dennis Sciama, a noted British cosmologist. During a pause in the lecture, Tryon threw out the suggestion that "maybe the universe is a vacuum fluctuation." Tryon intended the suggestion seriously, and was disappointed when his senior colleagues took it as a clever joke and broke into laughter. It was, after all, presumably the first scientific idea about where the universe came from.

By a vacuum fluctuation, Tryon was referring to the very complicated picture of the vacuum, or empty space, that emerges from relativistic quantum theory. The hallmark of quantum theory, developed to describe the behavior of atoms, is the probabilistic nature of its predictions. It is impossible, even in principle, to predict the behavior of any one atom, although it is possible to predict the average properties of a large collection of atoms. The vacuum, like any physical system, is subject to these quantum uncertainties. Roughly speaking, anything can happen in the vacuum, although the probability for a digital watch to materialize is absurdly small. Tryon was advancing the outlandish proposal that the entire universe materialized in this fashion!

Tryon put the idea of a vacuum fluctuation universe out of his mind for a while, but returned to it several years later while he was preparing a popular review of cosmology. In 1973 Tryon published an article in the journal Nature, with the title "Is the Universe a Vacuum Fluctuation?" [4]. He had understood the crucial point: the vast cosmos that we see around us could have originated as a vacuum fluctuation—essentially from nothing at all—because the large positive energy of the masses in the universe can be counterbalanced by a corresponding amount of negative energy in the form of the gravitational field. "In my model," Tryon wrote, "I assume that our universe did indeed appear from nowhere about 10^{10} years ago. Contrary to popular belief, such an event need not have violated any of the conventional laws of physics."

A weak point of Tryon's paper was its failure to explain why the universe had become so large. While the scale of vacuum fluctuations is typically subatomic, Tryon was asking us to believe that all the matter in the universe appeared in a single vacuum fluctuation. He pointed out that the laws of physics place no strict limit on the magnitude of vacuum fluctuations, but he
did not estimate the probability of such an unusually large fluctuation. "In answer to the question of why it happened," he wrote, "I offer the modest proposal that our Universe is simply one of those things which happen from time to time." Although the creation of a universe might be very unlikely, Tryon emphasized that no one had counted the failed attempts. Nonetheless, the immensity of the observed universe remained a striking feature for which Tryon's proposal had no explanation. For a number of years Tryon's work was largely ignored, as most other physicists apparently believed that any universe produced from a quantum fluctuation would, with overwhelming probability, be much smaller than the one that we observe.

But cannot the small give rise to a larger universe?

It is the dream of every young scientist to be caught up in an important discovery, a scientific revolution which changes the way that people think about some fundamental problem. The problem of the origin of matter is about as fundamental as any that one can imagine. Prompted by an unlikely series of events beginning in 1978, I became embroiled in the writing of a new chapter about this subject, a chapter which involved many other physicists and which culminated in the development of the inflationary universe theory. This theory is a new twist on the big bang theory, proposing a novel picture of how the universe behaved for the first minuscule fraction of a second of its existence.

The central feature of the theory is a brief period of extraordinarily rapid expansion, or inflation, which lasted for a time interval perhaps as short as $10^{-30}$ seconds. During this period the universe expanded by at least a factor of $10^{25}$, and perhaps a great deal more. After the stupendous growth spurt of inflation the description merges smoothly with the standard big bang theory, which for several decades has been the generally accepted picture of cosmic evolution.

Working within the general framework of accepted laws of physics, the inflationary theory can explain how the universe might have evolved from an initial seed as small as Tryon's vacuum fluctuations. Inflation provides a natural mechanism for tapping the unlimited reservoir of energy that can be extracted from the gravitational field—energy that can evolve to become the galaxies, stars, planets, and human beings that populate the universe today. While the standard big bang theory assumes that all the matter in the universe was present in some form since the beginning, the inflationary theory shows how all the mass could evolve from an initial seed weighing only about an ounce, with a diameter more than a billion times smaller than a proton.

The status of the theory of cosmic inflation today can be described by the same words used by Tryon in 1973 to describe his theory: "My model is admittedly speculative, and is still in an early stage of development... I am encouraged to believe that the origin and properties of our Universe may be explicable within the framework of conventional science, along the lines indicated here."

While the final verdict on inflationary cosmology is not yet in, the basic outline of the theory seems very persuasive. The theory not only accounts for the vast amount of matter in the universe, but it also offers plausible explanations for a number of features of the universe that otherwise remain unexplained. These features will be described later in the book. There is no doubt that inflation has caught the attention of the scientific community, since there are now over 3000 publications on it in the scientific literature.

If inflation is correct, then the inflationary mechanism is responsible for the creation of essentially all the matter and energy in the universe. The theory also implies that the observed universe is only a minute fraction of the entire universe, and it strongly suggests that there are perhaps an infinite number of other universes that are completely disconnected from our own.

Most important of all, the question of the origin of the matter in the universe is no longer thought to be beyond the range of science. After two thousand years of scientific research, it now seems likely that Lucretius was wrong. Conceivably, everything can be created from nothing. And "everything" might include a lot more than what we can see. In the context of inflationary cosmology, it is fair to say that the universe is the ultimate free lunch.
Since the negative energy of a gravitational field is crucial to the notion of a zero-energy universe, it is a subject worth examining carefully. In this appendix I will explain how the properties of gravity can be used to show that the energy of a gravitational field is unambiguously negative. The argument will be described in the context of Newton's theory of gravity, although the same conclusion can be reached using Einstein's theory of general relativity.

A simple way to demonstrate the sign of gravitational energy is to imagine a thin spherical shell of mass, as shown in Figure A.1(a). The shell will create a gravitational field, which at each point in space provides a measure of the force that would be experienced by a mass if it were located at that point. The gravitational field can be calculated by using Newton's methods. Newton first considers an ideal point mass—a mass concentration that is so small that it can be treated as if all the mass were located at a single point in space. For this case the gravitational field points directly toward the mass, with a strength that is described by the inverse square law—that is, the strength decreases as the square of the distance from the point mass. For a more complicated object such as the shell of Figure A.1(a), the gravitational field is in principle determined by mentally dividing the object into an infinite number of point masses, each with a mass that is infinitely small. A schematic illustration of this division into point masses is shown in part (b) of the figure. For each point mass one uses the simple inverse square law, and then one has to add the infinite number of contributions to obtain the final answer. The technique for handling such
an infinite number of quantities is the main subject of what is called integral calculus, which was largely developed by Newton himself.

The result for the gravitational field of a spherical shell of mass was first calculated by Newton, and it is the sort of calculation that is likely to show up in any college-level course in physics. Newton found that, outside the shell, the gravitational field at any point is directed radially inward toward the center of the shell. This answer could have been anticipated from the symmetry of the problem: there is no other reasonable answer to the question “What direction could it possibly point?” The strength of the gravitational field outside the shell can be described with surprising simplicity: the gravitational field has exactly the same strength as it would if all of the mass were concentrated at the center point of the shell.

What is the gravitational field inside the spherical shell? Consider, for example, the force on a particle at position B, as shown in part (c). By symmetry the force will be along the horizontal broken line, because the upward force caused by attraction toward the mass in the upper half of the diagram will be canceled by an opposing downward attraction toward the mass in the lower half of the diagram. We still must decide, however, if the force will point to the right or to the left. There is a persuasive argument which says that the force should be to the right: since the matter to the right is much closer than the matter to the left, the inverse square law should mean that the attraction to the matter on the right should dominate. There is also, however, a persuasive argument which says that the force should be to the left: There is more matter on the left, so the attraction toward it should dominate.

Which of the arguments in the previous paragraph is correct? Newton showed that these two arguments are equally valid, and in fact the forces cancel out exactly for a particle placed at any point inside the spherical cavity.

Figure A.1 (facing page) The gravitational field of a thin spherical shell. Part (a) shows the thin shell. The shell is really a three-dimensional sphere, but the diagram shows only a two-dimensional slice through the center of the sphere. In part (b) the shell has been divided into a large number of point masses, for the purpose of calculating the gravitational field at the point A. In principle there should be an infinite number of mass points, but only a finite number can be shown. The arrows indicate the gravitational attraction toward the mass points labeled P and Q. The attraction toward P is larger, because it is closer. Part (c) shows a point B, inside the spherical cavity, at which the gravitational field is to be calculated. The mass to the right of the vertical broken line is closer than the mass to the left, but there is more mass to the left.
To proceed with the discussion of gravitational energy, we need to answer one more question: How will gravity affect the spherical shell itself? Each point on the spherical shell will be attracted gravitationally toward each of the other points on the shell, and the net effect is a force pulling each point toward the center of the sphere. If the matter from which the shell is constructed is soft and compressible, then gravity will cause the shell to contract.

The situation is illustrated in Figure A.2, where part (a) shows the thin shell and the gravitational field that it generates. Outside the shell the gravitational field points inward, and inside the gravitational field is zero. Now imagine what would happen if the shell were allowed to uniformly contract, keeping its spherical shape. One can imagine, for example, extracting energy by tying ropes to each piece of the shell, as is illustrated in part (b). These ropes can be used to drive electric generators as each piece is lowered to its new position. Part (c) shows the sphere after the new radius is attained. The dashed circle indicates the original radius of the shell, and outside of the dashed circle the gravitational field is identical to that in part (a). (Recall that the field outside is the same as if all the mass were concentrated at the center, so it does not depend on the radius of the shell.) Inside the shell in its new position, the gravitational field remains zero. However, in the shaded region between the original and new positions of the shell, a gravitational field now exists where no field had existed before. The net effect of this operation is to extract energy, and to create a new region of gravitational field. Thus, energy is released when a gravitational field is created. The energy contained in the shaded region must therefore decrease, just as the water level in a tank decreases if water is released. Since the region began with no gravitational field and hence no energy, the final energy must be negative. In most physical processes the exchange of gravitational energy is much smaller than the rest energy (mc²) of the particles involved, but cosmologically the total gravitational energy can be very significant.

---

Figure A.2 (facing page) Thought experiment to understand the energy of gravity. Part (a) shows a hollow spherical shell of mass, and the gravitational field lines that it produces. There is a force on each piece of the shell, pulling inward. Part (b) shows how energy can be extracted as the shell is allowed to uniformly contract. Each piece of the shell is tied by a rope to an electrical generator, producing power as the piece is “lowered” toward its final position. Part (c) shows the final configuration, which includes a gravitational field in the shaded region where no field existed before. Thus, the creation of the gravitational field is associated with the release of energy.