THE RECORD-BREAKING BESTSELLER NOW IN PAPERBACK

A BRIEF HISTORY OF TIME

From the Big Bang to Black Holes

'This book marries a child's wonder to a genius's intellect. We journey into Hawking's universe, while marvelling at his mind'
Sunday Times

Introduction by Carl Sagan

STEPHEN HAWKING

FOREWARD

I didn't write a foreword to the original edition of A Brief History of Time. That was done by Carl Sagan. Instead, I wrote a short piece titled "Acknowledgments" in which I was advised to thank everyone. Some of the foundations that had given me support weren't too pleased to have been mentioned, however, because it led to a great increase in applications.

I don't think anyone, my publishers, my agent, or myself, expected the book to do anything like as well as it did. It was in the London Sunday Times best-seller list for 237 weeks, longer than any other book (apparently, the Bible and Shakespeare aren't counted). It has been translated into something like forty languages and has sold about one copy for every 750 men, women, and children in the world. As Nathan Myhrvold of Microsoft (a former post-doc of mine) remarked: I have sold more books on physics than Madonna has on sex.

The success of A Brief History indicates that there is widespread interest in the big questions like: Where did we come from? And why is the universe the way it is?

I have taken the opportunity to update the book and include new theoretical and observational results obtained since the book was first published (on April Fools' Day, 1988). I have included a new chapter on wormholes and time travel. Einstein's General Theory of Relativity seems to offer the possibility that we could create and maintain wormholes, little tubes that connect different regions of space-time. If so, we might be able to use them for rapid travel around the galaxy or travel back in time. Of course, we have not seen anyone from the future (or have we?) but I discuss a possible explanation for this.

I also describe the progress that has been made recently in finding "dualities" or correspondences between apparently different theories of physics. These correspondences are a strong indication that there is a complete unified theory of physics, but they also suggest that it may not be possible to express this theory in a single

fundamental formulation. Instead, we may have to use different reflections of the underlying theory in different situations. It might be like our being unable to represent the surface of the earth on a single map and having to use different maps in different regions. This would be a revolution in our view of the unification of the laws of science but it would not change the most important point: that the universe is governed by a set of rational laws that we can discover and understand.

On the observational side, by far the most important development has been the measurement of fluctuations in the cosmic microwave background radiation by COBE (the Cosmic Background Explorer satellite) and other collaborations. These fluctuations are the finger-prints of creation, tiny initial irregularities in the otherwise smooth and uniform early universe that later grew into galaxies, stars, and all the structures we see around us. Their form agrees with the predictions of the proposal that the universe has no boundaries or edges in the imaginary time direction; but further observations will be necessary to distinguish this proposal from other possible explanations for the fluctuations in the background. However, within a few years we should know whether we can believe that we live in a universe that is completely self-contained and without beginning or end.

Stephen Hawking

CHAPTER 1
OUR PICTURE OF THE UNIVERSE

A well-known scientist (some say it was Bertrand Russell) once gave a public lecture on astronomy. He described how the earth orbits around the sun and how the sun, in turn, orbits around the center of a vast collection of stars called our galaxy. At the end of the lecture, a little old lady at the back of the room got up and said: "What you have told us is rubbish. The world is really a flat plate supported on the back of a giant tortoise." The scientist gave a superior smile before replying, "What is the tortoise standing on." "You're very clever, young man, very clever," said the old lady. "But it's turtles all the way down!"

Most people would find the picture of our universe as an infinite tower of tortoises rather ridiculous, but why do we think we know better? What do we know about the universe, and how do we know it? Where did the universe come from, and where is it going? Did the universe have a beginning, and if so, what happened before then? What is the nature of time? Will it ever come to an end? Can we go back in time? Recent breakthroughs in physics, made possible in part by fantastic new technologies, suggest answers to some of these longstanding questions. Someday these answers may seem as obvious to us as the earth orbiting the sun - or perhaps as ridiculous as a tower of tortoises. Only time (whatever that may be) will tell.

As long ago as 340 BC the Greek philosopher Aristotle, in his book On the Heavens, was able to put forward two good arguments for believing that the earth was a round sphere rather than a Hat plate. First, he realized that eclipses of the moon were caused by the earth coming between the sun and the moon. The earth's shadow on the moon was always round, which would be true only if the earth was spherical. If the earth had been a flat disk, the shadow would have been elongated and elliptical, unless the eclipse always occurred at a time when the sun was directly under the center of the disk.

Second, the Greeks knew from their travels that the North Star appeared lower in the sky when viewed in the south than it did in more northerly regions. (Since the North Star lies over the North Pole, it appears to be directly above an observer at the North Pole, but to someone looking from the equator, it appears to lie just at the horizon. From the difference in the apparent position of the North Star in Egypt and Greece, Aristotle even quoted an estimate that the distance around the earth was 400,000 stadia. It is not known exactly what length a stadium was, but it may have been about 200 yards, which would make Aristotle's estimate about twice the currently accepted figure. The Greeks even had a third argument that the earth must be round, for why else does one first see the sails of a ship coming over the horizon, and only later see the hull?

Aristotle thought the earth was stationary and that the sun, the moon, the planets, and the stars moved in circular orbits about the earth. He believed this because he felt, for mystical reasons, that the earth was the center of the universe, and that circular motion was the most perfect. This idea was elaborated by Ptolemy in the second century AD into a complete cosmological model. The earth stood at the center, surrounded by eight spheres that carried the moon, the sun, the stars, and the five planets known at the time, Mercury, Venus, Mars, Jupiter, and Saturn (Fig. 1.1). The planets themselves moved on smaller circles attached to their respective spheres in order to account for their rather complicated observed paths in the sky. The outermost sphere carried the so-called fixed stars, which always stay in the same positions relative to each other but which rotate together across the sky. What lay beyond the last sphere was never made very clear, but it certainly was not part of mankind's observable universe.

Ptolemy's model provided a reasonably accurate system for predicting the positions of heavenly bodies in the sky. But in order to predict these positions correctly, Ptolemy had to make an assumption that the moon followed a path that sometimes brought

it twice as close to the earth as at other times. And that meant that the moon ought sometimes to appear twice as big as at other times! Ptolemy recognized this flaw, but nevertheless his model was generally, although not universally, accepted. It was adopted by the Christian church as the picture of the universe that was in accordance with Scripture, for it had the great advantage that it left lots of room outside the sphere of fixed stars for heaven and hell.

A simpler model, however, was proposed in 1514 by a Polish priest, Nicholas Copernicus. (At first, perhaps for fear of being branded a heretic by his church, Copernicus circulated his model anonymously.) His idea was that the sun was stationary at the center and that the earth and the planets moved in circular orbits around the sun. Nearly a century passed before this idea was taken seriously. Then two astronomers - the German, Johannes Kepler, and the Italian, Galileo Galilei - started publicly to support the Copernican theory, despite the fact that the orbits it predicted did not quite match the ones observed. The death blow to the Aristotelian/Ptolemaic theory came in 1609. In that year, Galileo started observing the night sky with a telescope, which had just been invented. When he looked at the planet Jupiter, Galileo found that it was accompanied by several small satellites or moons that orbited around it. This implied that everything did not have to orbit directly around the earth, as Aristotle and Ptolemy had thought. (It was, of course, still possible to believe that the earth was stationary at the center of the universe and that the moons of Jupiter moved on extremely complicated paths around the earth, giving the appearance that they orbited Jupiter. However, Copernicus's theory was much simpler.) At the same time, Johannes Kepler had modified Copernicus's theory, suggesting that the planets moved not in circles but in ellipses (an ellipse is an elongated circle). The predictions now finally matched the observations.

As far as Kepler was concerned, elliptical orbits were merely an ad hoc hypothesis, and a rather repugnant one at that, because ellipses were clearly less perfect than circles. Having discovered almost by

accident that elliptical orbits fit the observations well, he could not reconcile them with his idea that the planets were made to orbit the sun by magnetic forces. An explanation was provided only much later, in 1687, when Sir Isaac Newton published his Philosophiae Naturalis Principia Mathematica, probably the most important single work ever published in the physical sciences. In it Newton not only put forward a theory of how bodies move in space and time, but he also developed the complicated mathematics needed to analyze those motions. In addition, Newton postulated a law of universal gravitation according to which each body in the universe was attracted toward every other body by a force that was stronger the more massive the bodies and the closer they were to each other. It was this same force that caused objects to fall to the ground. (The story that Newton was inspired by an apple hitting his head is almost certainly apocryphal. All Newton himself ever said was that the idea of gravity came to him as he sat "in a contemplative mood" and "was occasioned by the fall of an apple.") Newton went on to show that, according to his law, gravity causes the moon to move in an elliptical orbit around the earth and causes the earth and the planets to follow elliptical paths around the sun.

The Copernican model got rid of Ptolemy's celestial spheres, and with them, the idea that the universe had a natural boundary. Since "fixed stars" did not appear to change their positions apart from a rotation across the sky caused by the earth spinning on its axis, it became natural to suppose that the fixed stars were objects like our sun but very much farther away.

Newton realized that, according to his theory of gravity, the stars should attract each other, so it seemed they could not remain essentially motionless. Would they not all fall together at some point? In a letter in 1691 to Richard Bentley, another leading thinker of his day, Newton argued that this would indeed happen if there were only a finite number of stars distributed over a finite region of space. But he reasoned that if, on the other hand, there were an infinite number of stars, distributed more or less uniformly

over infinite space, this would not happen, because there would not be any central point for them to fall to.

This argument is an instance of the pitfalls that you can encounter in talking about infinity. In an infinite universe, every point can be regarded as the center, because every point has an infinite number of stars on each side of it. The correct approach, it was realized only much later, is to consider the finite situation, in which the stars all fall in on each other, and then to ask how things change if one adds more stars roughly uniformly distributed outside this region. According to Newton's law, the extra stars would make no difference at all to the original ones on average, so the stars would fall in just as fast. We can add as many stars as we like, but they will still always collapse in on them-selves. We now know it is impossible to have an infinite static model of the universe in which gravity is always attractive.

It is an interesting reflection on the general climate of thought before the twentieth century that no one had suggested that the universe was expanding or contracting. It was generally accepted that either the universe had existed forever in an unchanging state, or that it had been created at a finite time in the past more or less as we observe it today. In part this may have been due to people's tendency to believe in eternal truths, as well as the comfort they found in the thought that even though they may grow old and die, the universe is eternal and unchanging.

Even those who realized that Newton's theory of gravity showed that the universe could not be static did not think to suggest that it might be expanding. Instead, they attempted to modify the theory by making the gravitational force repulsive at very large distances. This did not significantly affect their predictions of the motions of the planets, but it allowed an infinite distribution of stars to remain in equilibrium - with the attractive forces between nearby stars balanced by the repulsive forces from those that were farther away. However, we now believe such an equilibrium would be unstable: if the stars in some region got only slightly nearer each other, the

attractive forces between them would become stronger and dominate over the repulsive forces so that the stars would continue to fall toward each other. On the other hand, if the stars got a bit farther away from each other, the repulsive forces would dominate and drive them farther apart.

Another objection to an infinite static universe is normally ascribed to the German philosopher Heinrich Olbers, who wrote about this theory in 1823. In fact, various contemporaries of Newton had raised the problem, and the Olbers article was not even the first to contain plausible arguments against it. It was, however, the first to be widely noted. The difficulty is that in an infinite static universe nearly every line of sight would end on the surface of a star. Thus one would expect that the whole sky would be as bright as the sun, even at night. Olbers' counter-argument was that the light from distant stars would be dimmed by absorption by intervening matter. However, if that happened the intervening matter would eventually heat up until it glowed as brightly as the stars. The only way of avoiding the conclusion that the whole of the night sky should be as bright as the surface of the sun would be to assume that the stars had not been shining forever but had turned on at some finite time in the past. In that case the absorbing matter might not have heated up yet or the light from distant stars might not yet have reached us. And that brings us to the question of what could have caused the stars to have turned on in the first place.

The beginning of the universe had, of course, been discussed long before this. According to a number of early cosmologies and the Jewish/Christian/Muslim tradition, the universe started at a finite, and not very distant, time in the past. One argument for such a beginning was the feeling that it was necessary to have "First Cause" to explain the existence of the universe. (Within the universe, you always explained one event as being caused by some earlier event, but the existence of the universe itself could be explained in this way only if it had some beginning.) Another argument was put forward by St. Augustine in his book The City of

God. He pointed out that civilization is progressing and we remember who performed this deed or developed that technique. Thus man, and so also perhaps the universe, could not have been around all that long. St. Augustine accepted a date of about 5000 BC for the Creation of the universe according to the book of Genesis. (It is interesting that this is not so far from the end of the last Ice Age, about 10,000 BC, which is when archaeologists tell us that civilization really began.)

Aristotle, and most of the other Greek philosophers, on the other hand, did not like the idea of a creation because it smacked too much of divine intervention. They believed, therefore, that the human race and the world around it had existed, and would exist, forever. The ancients had already considered the argument about progress described above, and answered it by saying that there had been periodic floods or other disasters that repeatedly set the human race right back to the beginning of civilization.

The questions of whether the universe had a beginning in time and whether it is limited in space were later extensively examined by the philosopher Immanuel Kant in his monumental (and very obscure) work Critique of Pure Reason, published in 1781. He called these questions antinomies (that is, contradictions) of pure reason because he felt that there were equally compelling arguments for believing the thesis, that the universe had a beginning, and the antithesis, that it had existed forever. His argument for the thesis was that if the universe did not have a beginning, there would be an infinite period of time before any event, which he considered absurd. The argument for the antithesis was that if the universe had a beginning, there would be an infinite period of time before it, so why should the universe begin at any one particular time? In fact, his cases for both the thesis and the antithesis are really the same argument. They are both based on his unspoken assumption that time continues back forever, whether or not the universe had existed forever. As we shall see, the concept of time has no meaning before the beginning of the universe. This

was first pointed out by St. Augustine. When asked: "What did God do before he created the universe?" Augustine didn't reply: "He was preparing Hell for people who asked such questions." Instead, he said that time was a property of the universe that God created, and that time did not exist before the beginning of the universe.

When most people believed in an essentially static and unchanging universe, the question of whether or not it had a beginning was really one of metaphysics or theology. One could account for what was observed equally well on the theory that the universe had existed forever or on the theory that it was set in motion at some finite time in such a manner as to look as though it had existed forever. But in 1929, Edwin Hubble made the landmark observation that wherever you look, distant galaxies are moving rapidly away from us. In other words, the universe is expanding. This means that at earlier times objects would have been closer together. In fact, it seemed that there was a time, about ten or twenty thousand million years ago, when they were all at exactly the same place and when, therefore, the density of the universe was infinite. This discovery finally brought the question of the beginning of the universe into the realm of science.

Hubble's observations suggested that there was a time, called the big bang, when the universe was infinitesimally small and infinitely dense. Under such conditions all the laws of science, and therefore all ability to predict the future, would break down. If there were events earlier than this time, then they could not affect what happens at the present time. Their existence can be ignored because it would have no observational consequences. One may say that time had a beginning at the big bang, in the sense that earlier times simply would not be defined. It should be emphasized that this beginning in time is very different from those that had been considered previously. In an unchanging universe a beginning in time is something that has to be imposed by some being outside the universe; there is no physical necessity for a beginning. One

can imagine that God created the universe at literally any time in the past. On the other hand, if the universe is expanding, there may be physical reasons why there had to be a beginning. One could still imagine that God created the universe at the instant of the big bang, or even afterwards in just such a way as to make it look as though there had been a big bang, but it would be meaningless to suppose that it was created before the big bang. An expanding universe does not preclude a creator, but it does place limits on when he might have carried out his job!

In order to talk about the nature of the universe and to discuss questions such as whether it has a beginning or an end, you have to be clear about what a scientific theory is. I shall take the simpleminded view that a theory is just a model of the universe, or a restricted part of it, and a set of rules that relate quantities in the model to observations that we make. It exists only in our minds and does not have any other reality (whatever that might mean). A theory is a good theory if it satisfies two requirements. It must accurately describe a large class of observations on the basis of a model that contains only a few arbitrary elements, and it must make definite predictions about the results of future observations. For example, Aristotle believed Empedocles's theory that everything was made out of four elements, earth, air, fire, and water. This was simple enough, but did not make any definite predictions. On the other hand, Newton's theory of gravity was based on an even simpler model, in which bodies attracted each other with a force that was proportional to a quantity called their mass and inversely proportional to the square of the distance between them. Yet it predicts the motions of the sun, the moon, and the planets to a high degree of accuracy.

Any physical theory is always provisional, in the sense that it is only a hypothesis: you can never prove it. No matter how many times the results of experiments agree with some theory, you can never be sure that the next time the result will not contradict the theory. On the other hand, you can disprove a theory by finding

even a single observation that disagrees with the predictions of the theory. As philosopher of science Karl Popper has emphasized, a good theory is characterized by the fact that it makes a number of predictions that could in principle be disproved or falsified by observation. Each time new experiments are observed to agree with the predictions the theory survives, and our confidence in it is increased; but if ever a new observation is found to disagree, we have to abandon or modify the theory.

At least that is what is supposed to happen, but you can always question the competence of the person who carried out the observation.

In practice, what often happens is that a new theory is devised that is really an extension of the previous theory. For example, very accurate observations of the planet Mercury revealed a small difference between its motion and the predictions of Newton's theory of gravity. Einstein's general theory of relativity predicted a slightly different motion from Newton's theory. The fact that Einstein's predictions matched what was seen, while Newton's did not, was one of the crucial confirmations of the new theory. However, we still use Newton's theory for all practical purposes because the difference between its predictions and those of general relativity is very small in the situations that we normally deal with. (Newton's theory also has the great advantage that it is much simpler to work with than Einstein's!)

The eventual goal of science is to provide a single theory that describes the whole universe. However, the approach most scientists actually follow is to separate the problem into two parts. First, there are the laws that tell us how the universe changes with time. (If we know what the universe is like at any one time, these physical laws tell us how it will look at any later time.) Second, there is the question of the initial state of the universe. Some people feel that science should be concerned with only the first part; they regard the question of the initial situation as a matter for metaphysics or religion. They would say that God, being

omnipotent, could have started the universe off any way he wanted. That may be so, but in that case he also could have made it develop in a completely arbitrary way. Yet it appears that he chose to make it evolve in a very regular way according to certain laws. It therefore seems equally reasonable to suppose that there are also laws governing the initial state.

It turns out to be very difficult to devise a theory to describe the universe all in one go. Instead, we break the problem up into bits and invent a number of partial theories. Each of these partial theories describes and predicts a certain limited class of observations, neglecting the effects of other quantities, or representing them by simple sets of numbers. It may be that this approach is completely wrong. If every-thing in the universe depends on everything else in a fundamental way, it might be impossible to get close to a full solution by investigating parts of the problem in isolation. Nevertheless, it is certainly the way that we have made progress in the past. The classic example again is the Newtonian theory of gravity, which tells us that the gravitational force between two bodies depends only on one number associated with each body, its mass, but is otherwise independent of what the bodies are made of. Thus one does not need to have a theory of the structure and constitution of the sun and the planets in order to calculate their orbits.

Today scientists describe the universe in terms of two basic partial theories - the general theory of relativity and quantum mechanics. They are the great intellectual achievements of the first half of this century. The general theory of relativity describes the force of gravity and the large-scale structure of the universe, that is, the structure on scales from only a few miles to as large as a million million million million (1 with twenty-four zeros after it) miles, the size of the observable universe. Quantum mechanics, on the other hand, deals with phenomena on extremely small scales, such as a millionth of a millionth of an inch. Unfortunately, however, these two theories are known to be inconsistent with each other - they

cannot both be correct. One of the major endeavors in physics today, and the major theme of this book, is the search for a new theory that will incorporate them both - a quantum theory of gravity. We do not yet have such a theory, and we may still be a long way from having one, but we do already know many of the properties that it must have. And we shall see, in later chapters, that we already know a fair amount about the predications a quantum theory of gravity must make.

Now, if you believe that the universe is not arbitrary, but is governed by definite laws, you ultimately have to combine the partial theories into a complete unified theory that will describe everything in the universe. But there is a fundamental paradox in the search for such a complete unified theory. The ideas about scientific theories outlined above assume we are rational beings who are free to observe the universe as we want and to draw logical deductions from what we see.

In such a scheme it is reasonable to suppose that we might progress ever closer toward the laws that govern our universe. Yet if there really is a complete unified theory, it would also presumably determine our actions. And so the theory itself would determine the outcome of our search for it! And why should it determine that we come to the right conclusions from the evidence? Might it not equally well determine that we draw the wrong conclusion.? Or no conclusion at all?

The only answer that I can give to this problem is based on Darwin's principle of natural selection. The idea is that in any population of self-reproducing organisms, there will be variations in the genetic material and upbringing that different individuals have. These differences will mean that some individuals are better able than others to draw the right conclusions about the world around them and to act accordingly. These individuals will be more likely to survive and reproduce and so their pattern of behavior and thought will come to dominate. It has certainly been true in the past that what we call intelligence and scientific discovery have

conveyed a survival advantage. It is not so clear that this is still the case: our scientific discoveries may well destroy us all, and even if they don't, a complete unified theory may not make much difference to our chances of survival. However, provided the universe has evolved in a regular way, we might expect that the reasoning abilities that natural selection has given us would be valid also in our search for a complete unified theory, and so would not lead us to the wrong conclusions.

Because the partial theories that we already have are sufficient to make accurate predictions in all but the most extreme situations, the search for the ultimate theory of the universe seems difficult to justify on practical grounds. (It is worth noting, though, that similar arguments could have been used against both relativity and quantum mechanics, and these theories have given us both nuclear energy and the microelectronics revolution!) The discovery of a complete unified theory, therefore, may not aid the survival of our species. It may not even affect our life-style. But ever since the dawn of civilization, people have not been content to see events as unconnected and inexplicable. They have craved an understanding of the underlying order in the world. Today we still yearn to know why we are here and where we came from. Humanity's deepest desire for knowledge is justification enough for our continuing quest. And our goal is nothing less than a complete description of the universe we live in.

CHAPTER 2
Space and Time

Our present ideas about the motion of bodies date back to Galileo and Newton. Before them people believed Aristotle, who said that the natural state of a body was to be at rest and that it moved only if driven by a force or impulse. It followed that a heavy body

should fall faster than a light one, because it would have a greater pull toward the earth.

The Aristotelian tradition also held that one could work out all the laws that govern the universe by pure thought: it was not necessary to check by observation. So no one until Galileo bothered to see whether bodies of different weight did in fact fall at different speeds. It is said that Galileo demonstrated that Aristotle's belief was false by dropping weights from the leaning tower of Pisa. The story is almost certainly untrue, but Galileo did do something equivalent: he rolled balls of different weights down a smooth slope. The situation is similar to that of heavy bodies falling vertically, but it is easier to observe because the Speeds are smaller. Galileo's measurements indicated that each body increased its speed at the same rate, no matter what its weight. For example, if you let go of a ball on a slope that drops by one meter for every ten meters you go along, the ball will be traveling down the slope at a speed of about one meter per second after one second, two meters per second after two seconds, and so on, however heavy the ball. Of course a lead weight would fall faster than a feather, but that is only because a feather is slowed down by air resistance. If one drops two bodies that don't have much air resistance, such as two different lead weights, they fall at the same rate. On the moon, where there is no air to slow things down, the astronaut David R. Scott performed the feather and lead weight experiment and found that indeed they did hit the ground at the same time.

Galileo's measurements were used by Newton as the basis of his laws of motion. In Galileo's experiments, as a body rolled down the slope it was always acted on by the same force (its weight), and the effect was to make it constantly speed up. This showed that the real effect of a force is always to change the speed of a body, rather than just to set it moving, as was previously thought. It also meant that when-ever a body is not acted on by any force, it will keep on moving in a straight line at the same speed. This idea was first

stated explicitly in Newton's Principia Mathematica, published in 1687, and is known as Newton's first law. What happens to a body when a force does act on it is given by Newton's second law. This states that the body will accelerate, or change its speed, at a rate that is proportional to the force. (For example, the acceleration is twice as great if the force is twice as great.) The acceleration is also smaller the greater the mass (or quantity of matter) of the body. (The same force acting on a body of twice the mass will produce half the acceleration.) A familiar example is provided by a car: the more powerful the engine, the greater the acceleration, but the heavier the car, the smaller the acceleration for the same engine. In addition to his laws of motion, Newton discovered a law to describe the force of gravity, which states that every body attracts every other body with a force that is proportional to the mass of each body. Thus the force between two bodies would be twice as strong if one of the bodies (say, body A) had its mass doubled. This is what you might expect because one could think of the new body A as being made of two bodies with the original mass. Each would attract body B with the original force. Thus the total force between A and B would be twice the original force. And if, say, one of the bodies had twice the mass, and the other had three times the mass, then the force would be six times as strong. One can now see why all bodies fall at the same rate: a body of twice the weight will have twice the force of gravity pulling it down, but it will also have twice the mass. According to Newton's second law, these two effects will exactly cancel each other, so the acceleration will be the same in all cases.

Newton's law of gravity also tells us that the farther apart the bodies, the smaller the force. Newton's law of gravity says that the gravitational attraction of a star is exactly one quarter that of a similar star at half the distance. This law predicts the orbits of the earth, the moon, and the planets with great accuracy. If the law were that the gravitational attraction of a star went down faster or increased more rapidly with distance, the orbits of the planets

would not be elliptical, they would either spiral in to the sun or escape from the sun.

The big difference between the ideas of Aristotle and those of Galileo and Newton is that Aristotle believed in a preferred state of rest, which any body would take up if it were not driven by some force Or impulse. In particular, he thought that the earth was at rest. But it follows from Newton's laws that there is no unique standard of rest. One could equally well say that body A was at rest and body B was moving at constant speed with respect to body A, or that body B was at rest and body A was moving. For example, if one sets aside for a moment the rotation of the earth and its orbit round the sun, one could say that the earth was at rest and that a train on it was traveling north at ninety miles per hour or that the train was at rest and the earth was moving south at ninety miles per hour. If one carried out experiments with moving bodies on the train, all Newton's laws would still hold. For instance, playing Ping-Pong on the train, one would find that the ball obeyed Newton's laws just like a ball on a table by the track. So there is no way to tell whether it is the train or the earth that is moving.

The lack of an absolute standard of rest meant that one could not determine whether two events that took place at different times occurred in the same position in space. For example, suppose our Ping-Pong ball on the train bounces straight up and down, hitting the table twice on the same spot one second apart. To someone on the track, the two bounces would seem to take place about forty meters apart, because the train would have traveled that far down the track between the bounces. The nonexistence of absolute rest therefore meant that one could not give an event an absolute position in space, as Aristotle had believed. The positions of events and the distances between them would be different for a person on the train and one on the track, and there would be no reason to prefer one person's position to the other's.

Newton was very worried by this lack of absolute position, or absolute space, as it was called, because it did not accord with his

idea of an absolute God. In fact, he refused to accept lack of absolute space, even though it was implied by his laws. He was severely criticized for this irrational belief by many people, most notably by Bishop Berkeley, a philosopher who believed that all material objects and space and time are an illusion. When the famous Dr. Johnson was told of Berkeley's opinion, he cried, "I refute it thus!" and stubbed his toe on a large stone.

Both Aristotle and Newton believed in absolute time. That is, they believed that one could unambiguously measure the interval of time between two events, and that this time would be the same whoever measured it, provided they used a good clock. Time was completely separate from and independent of space. This is what most people would take to be the commonsense view. However, we have had to change our ideas about space and time. Although our apparently commonsense notions work well when dealing with things like apples, or planets that travel comparatively slowly, they don't work at all for things moving at or near the speed of light.

The fact that light travels at a finite, but very high, speed was first discovered in 1676 by the Danish astronomer Ole Christensen Roemer. He observed that the times at which the moons of Jupiter appeared to pass behind Jupiter were not evenly spaced, as one would expect if the moons went round Jupiter at a constant rate. As the earth and Jupiter orbit around the sun, the distance between them varies. Roemer noticed that eclipses of Jupiter's moons appeared later the farther we were from Jupiter. He argued that this was because the light from the moons took longer to reach us when we were farther away. His measurements of the variations in the distance of the earth from Jupiter were,

¿ however, not very accurate, and so his value for the speed of light was 140,000 miles per second, compared to the modern value of 186,000 miles per second. Nevertheless, Roemer's achievement, in not only proving that light travels at a finite speed, but also in measuring that speed, was remarkable - coming as it did eleven years before Newton's publication of Principia Mathematica. A

proper theory of the propagation of light didn't come until 1865, when the British physicist James Clerk Maxwell succeeded in unifying the partial theories that up to then had been used to describe the forces of electricity and magnetism. Maxwell's equations predicted that there could be wavelike disturbances in the combined electromagnetic field, and that these would travel at a fixed speed, like ripples on a pond. If the wavelength of these waves (the distance between one wave crest and the next) is a meter or more, they are what we now call radio waves. Shorter wavelengths are known as microwaves (a few centimeters) or infrared (more than a ten-thousandth of a centimeter). Visible light has a wavelength of between only forty and eighty millionths of a centimeter. Even shorter wavelengths are known as ultraviolet, X rays, and gamma rays.

Maxwell's theory predicted that radio or light waves should travel at a certain fixed speed. But Newton's theory had got rid of the idea of absolute rest, so if light was supposed to travel at a fixed speed, one would have to say what that fixed speed was to be measured relative to.

It was therefore suggested that there was a substance called the "ether" that was present everywhere, even in "empty" space. Light waves should travel through the ether as sound waves travel through air, and their speed should therefore be relative to the ether. Different observers, moving relative to the ether, would see light coming toward them at different speeds, but light's speed relative to the ether would remain fixed. In particular, as the earth was moving through the ether on its orbit round the sun, the speed of light measured in the direction of the earth's motion through the ether (when we were moving toward the source of the light) should be higher than the speed of light at right angles to that motion (when we ar not moving toward the source). In 1887Albert Michelson (who later became the first American to receive the Nobel Prize for physics) and Edward Morley carried out a very careful experiment at the Case School of Applied Science in

Cleveland. They compared the speed of light in the direction of the earth's motion with that at right angles to the earth's motion. To their great surprise, they found they were exactly the same!

Between 1887 and 1905 there were several attempts, most notably by the Dutch physicist Hendrik Lorentz, to explain the result of the Michelson-Morley experiment in terms of objects contracting and clocks slowing down when they moved through the ether. However, in a famous paper in 1905, a hitherto unknown clerk in the Swiss patent office, Albert Einstein, pointed out that the whole idea of an ether was unnecessary, providing one was willing to abandon the idea of absolute time. A similar point was made a few weeks later by a leading French mathematician, Henri Poincare. Einstein's arguments were closer to physics than those of Poincare, who regarded this problem as mathematical. Einstein is usually given the credit for the new theory, but Poincare is remembered by having his name attached to an important part of it.

The fundamental postulate of the theory of relativity, as it was called, was that the laws of science should be the same for all freely moving observers, no matter what their speed. This was true for Newton's laws of motion, but now the idea was extended to include Maxwell's theory and the speed of light: all observers should measure the same speed of light, no matter how fast they are moving. This simple idea has some remarkable consequences. Perhaps the best known are the equivalence of mass and energy, summed up in Einstein's famous equation E=mc2 (where E is energy, m is mass, and c is the speed of light), and the law that nothing may travel faster than the speed of light. Because of the equivalence of energy and mass, the energy which an object has due to its motion will add to its mass. In other words, it will make it harder to increase its speed. This effect is only really significant for objects moving at speeds close to the speed of light. For example, at 10 percent of the speed of light an object's mass is only 0.5 percent more than normal, while at 90 percent of the speed of light it would be more than twice its normal mass. As an object

approaches the speed of light, its mass rises ever more quickly, so it takes more and more energy to speed it up further. It can in fact never reach the speed of light, because by then its mass would have become infinite, and by the equivalence of mass and energy, it would have taken an infinite amount of energy to get it there. For this reason, any normal object is forever confined by relativity to move at speeds slower than the speed of light. Only light, or other waves that have no intrinsic mass, can move at the speed of light.

An equally remarkable consequence of relativity is the way it has revolutionized our ideas of space and time. In Newton's theory, if a pulse of light is sent from one place to another, different observers would agree on the time that the journey took (since time is absolute), but will not always agree on how far the light traveled (since space is not absolute). Since the speed of the light is just the distance it has traveled divided by the time it has taken, different observers would measure different speeds for the light. In relativity, on the other hand, all observers must agree on how fast light travels. They still, however, do not agree on the distance the light has traveled, so they must therefore now also disagree over the time it has taken. (The time taken is the distance the light has traveled - which the observers do not agree on - divided by the light's speed - which they do agree on.) In other words, the theory of relativity put an end to the idea of absolute time! It appeared that each observer must have his own measure of time, as recorded by a clock carried with him, and that identical clocks carried by different observers would not necessarily agree.

Each observer could use radar to say where and when an event took place by sending out a pulse of light or radio waves. Part of the pulse is reflected back at the event and the observer measures the time at which he receives the echo. The time of the event is then said to be the time halfway between when the pulse was sent and the time when the reflection was received back: the distance of the event is half the time taken for this round trip, multiplied by the speed of light. (An event, in this sense, is something that takes

place at a single point in space, at a specified point in time.) This idea is shown in Fig. 2.1, which is an example of a space-time diagram. Using this procedure, observers who are moving relative to each other will assign different times and positions to the same event. No particular observer's measurements are any more correct than any other observer's, but all the measurements are related. Any observer can work out precisely what time and position any other observer will assign to an event, provided he knows the other observer's relative velocity.

Nowadays we use just this method to measure distances precisely, because we can measure time more accurately than length. In effect, the meter is defined to be the distance traveled by light in 0.000000003335640952 second, as measured by a cesium clock. (The reason for that particular number is that it corresponds to the historical definition of the meter - in terms of two marks on a particular platinum bar kept in Paris.) Equally, we can use a more convenient, new unit of length called a light-second. This is simply defined as the distance that light travels in one second. In the theory of relativity, we now define distance in terms of time and the speed of light, so it follows automatically that every observer will measure light to have the same speed (by definition, 1 meter per 0.000000003335640952 second). There is no need to introduce the idea of an ether, whose presence anyway cannot be detected, as the Michelson-Morley experiment showed. The theory of relativity does, however, force us to change fundamentally our ideas of space and time. We must accept that time is not completely separate from and independent of space, but is combined with it to form an object called space-time.

It is a matter of common experience that one can describe the position of a point in space by three numbers, or coordinates. For instance, one can say that a point in a room is seven feet from one wall, three feet from another, and five feet above the floor. Or one could specify that a point was at a certain latitude and longitude and a certain height above sea level. One is free to use any three

suitable coordinates, although they have only a limited range of validity. One would not specify the position of the moon in terms of miles north and miles west of Piccadilly Circus and feet above sea level. Instead, one might de-scribe it in terms of distance from the sun, distance from the plane of the orbits of the planets, and the angle between the line joining the moon to the sun and the line joining the sun to a nearby star such as Alpha Centauri. Even these coordinates would not be of much use in describing the position of the sun in our galaxy or the position of our galaxy in the local group of galaxies. In fact, one may describe the whole universe in terms of a collection of overlapping patches. In each patch, one can use a different set of three coordinates to specify the position of a point.

An event is something that happens at a particular point in space and at a particular time. So one can specify it by four numbers or coordinates. Again, the choice of coordinates is arbitrary; one can use any three well-defined spatial coordinates and any measure of time. In relativity, there is no real distinction between the space and time coordinates, just as there is no real difference between any two space coordinates. One could choose a new set of coordinates in which, say, the first space coordinate was a combination of the old first and second space coordinates. For instance, instead of measuring the position of a point on the earth in miles north of Piccadilly and miles west of Piccadilly, one could use miles northeast of Piccadilly, and miles north-west of Piccadilly. Similarly, in relativity, one could use a new time coordinate that was the old time (in seconds) plus the distance (in light-seconds) north of Piccadilly.

It is often helpful to think of the four coordinates of an event as specifying its position in a four-dimensional space called space-time. It is impossible to imagine a four-dimensional space. I personally find it hard enough to visualize three-dimensional space! However, it is easy to draw diagrams of two-dimensional spaces, such as the surface of the earth. (The surface of the earth is

two-dimensional because the position of a point can be specified by two coordinates, latitude and longitude.) I shall generally use diagrams in which time increases upward and one of the spatial dimensions is shown horizontally. The other two spatial dimensions are ignored or, sometimes, one of them is indicated by perspective. (These are called space-time diagrams, like Fig. 2.1.) For example, in Fig. 2.2 time is measured upward in years and the distance along the line from the sun to Alpha Centauri is measured horizontally in miles. The paths of the sun and of Alpha Centauri through space-time are shown as the vertical lines on the left and right of the diagram. A ray of light from the sun follows the diagonal line, and takes four years to get from the sun to Alpha Centauri.

As we have seen, Maxwell's equations predicted that the speed of light should be the same whatever the speed of the source, and this has been confirmed by accurate measurements. It follows from this that if a pulse of light is emitted at a particular time at a particular point in space, then as time goes on it will spread out as a sphere of light whose size and position are independent of the speed of the source. After one millionth of a second the light will have spread out to form a sphere with a radius of 300 meters; after two millionths of a second, the radius will be 600 meters; and so on. It will be like the ripples that spread out on the surface of a pond when a stone is thrown in. The ripples spread out as a circle that gets bigger as time goes on. If one stacks snapshots of the ripples at different times one above the other, the expanding circle of ripples will mark out a cone whose tip is at the place and time at which the stone hit the water (Fig. 2.3). Similarly, the light spreading out from an event forms a (three-dimensional) cone in (the four-dimensional) space-time. This cone is called the future light cone of the event. In the same way we can draw another cone, called the past light cone, which is the set of events from which a pulse of light is able to reach the given event (Fig. 2.4).

Given an event P, one can divide the other events in the universe into three classes. Those events that can be reached from the event P by a particle or wave traveling at or below the speed of light are said to be in the future of P. They will lie within or on the expanding sphere of light emitted from the event P. Thus they will lie within or on the future light cone of P in the space-time diagram. Only events in the future of P can be affected by what happens at P because nothing can travel faster than light.

Similarly, the past of P can be defined as the set of all events from which it is possible to reach the event P traveling at or below the speed of light. It is thus the set of events that can affect what happens at P. The events that do not lie in the future or past of P are said to lie in the elsewhere of P (Fig. 2.5). What happens at such events can neither affect nor be affected by what happens at P. For example, if the sun were to cease to shine at this very moment, it would not affect things on earth at the present time because they would be in the elsewhere of the event when the sun went out (Fig. 2.6). We would know about it only after eight minutes, the time it takes light to reach us from the sun. Only then would events on earth lie in the future light cone of the event at which the sun went out. Similarly, we do not know what is happening at the moment farther away in the universe: the light that we see from distant galaxies left them millions of years ago, and in the case of the most distant object that we have seen, the light left some eight thousand million years ago. Thus, when we look at the universe, we are seeing it as it was in the past.

If one neglects gravitational effects, as Einstein and Poincare did in 1905, one has what is called the special theory of relativity. For every event in space-time we may construct a light cone (the set of all possible paths of light in space-time emitted at that event), and since the speed of light is the same at every event and in every direction, all the light cones will be identical and will all point in the same direction. The theory also tells us that nothing can travel faster than light. This means that the path of any object through

space and time must be represented by a line that lies within the light cone at each event on it (Fig. 2.7). The special theory of relativity was very successful in explaining that the speed of light appears the same to all observers (as shown by the Michelson-Morley experiment) and in describing what happens when things move at speeds close to the speed of light. However, it was inconsistent with the Newtonian theory of gravity, which said that objects attracted each other with a force that depended on the distance between them. This meant that if one moved one of the objects, the force on the other one would change instantaneously. Or in other gravitational effects should travel with infinite velocity, instead of at or below the speed of light, as the special theory of relativity required. Einstein made a number of unsuccessful attempts between 1908 and 1914 to find a theory of gravity that was consistent with special relativity. Finally, in 1915, he proposed what we now call the general theory of relativity.

Einstein made the revolutionary suggestion that gravity is not a force like other forces, but is a consequence of the fact that space-time is not flat, as had been previously assumed: it is curved, or "warped," by the distribution of mass and energy in it. Bodies like the earth are not made to move on curved orbits by a force called gravity; instead, they follow the nearest thing to a straight path in a curved space, which is called a geodesic. A geodesic is the shortest (or longest) path between two nearby points. For example, the surface of the earth is a two-dimensional curved space. A geodesic on the earth is called a great circle, and is the shortest route between two points (Fig. 2.8). As the geodesic is the shortest path between any two airports, this is the route an airline navigator will tell the pilot to fly along. In general relativity, bodies always follow straight lines in four-dimensional space-time, but they nevertheless appear to us to move along curved paths in our three-dimensional space. (This is rather like watching an airplane flying over hilly ground. Although it follows a straight line in three-dimensional

space, its shadow follows a curved path on the two-dimensional ground.)

The mass of the sun curves space-time in such a way that although the earth follows a straight path in four-dimensional space-time, it appears to us to move along a circular orbit in three-dimensional space.

fact, the orbits of the planets predicted by general relativity are almost exactly the same as those predicted by the Newtonian theory of gravity. However, in the case of Mercury, which, being the nearest planet to the sun, feels the strongest gravitational effects, and has a rather elongated orbit, general relativity predicts that the long axis of the ellipse should rotate about the sun at a rate of about one degree in ten thousand years. Small though this effect is, it had been noticed before 1915 and served as one of the first confirmations of Einstein's theory. In recent years the even smaller deviations of the orbits of the other planets from the Newtonian predictions have been measured by radar and found to agree with the predictions of general relativity.

Light rays too must follow geodesics in space-time. Again, the fact that space is curved means that light no longer appears to travel in straight lines in space. So general relativity predicts that light should be bent by gravitational fields. For example, the theory predicts that the light cones of points near the sun would be slightly bent inward, on account of the mass of the sun. This means that light from a distant star that happened to pass near the sun would be deflected through a small angle, causing the star to appear in a different position to an observer on the earth (Fig. 2.9). Of course, if the light from the star always passed close to the sun, we would not be able to tell whether the light was being deflected or if instead the star was really where we see it. However, as the earth orbits around the sun, different stars appear to pass behind the sun and have their light deflected. They therefore change their apparent position relative to other stars. It is normally very difficult to see this effect, because the light from the sun makes it impossible to

observe stars that appear near to the sun the sky. However, it is possible to do so during an eclipse of the sun, when the sun's light is blocked out by the moon. Einstein's prediction of light deflection could not be tested immediately in 1915, because the First World War was in progress, and it was not until 1919 that a British expedition, observing an eclipse from West Africa, showed that light was indeed deflected by the sun, just as predicted by the theory. This proof of a German theory by British scientists was hailed as a great act of reconciliation between the two countries after the war. It is ionic, therefore, that later examination of the photographs taken on that expedition showed the errors were as great as the effect they were trying to measure. Their measurement had been sheer luck, or a case of knowing the result they wanted to get, not an uncommon occurrence in science. The light deflection has, however, been accurately confirmed by a number of later observations.

Another prediction of general relativity is that time should appear to slower near a massive body like the earth. This is because there is a relation between the energy of light and its frequency (that is, the number of waves of light per second): the greater the energy, the higher frequency. As light travels upward in the earth's gravitational field, it loses energy, and so its frequency goes down. (This means that the length of time between one wave crest and the next goes up.) To someone high up, it would appear that everything down below was making longer to happen. This prediction was tested in 1962, using a pair of very accurate clocks mounted at the top and bottom of a water tower. The clock at the bottom, which was nearer the earth, was found to run slower, in exact agreement with general relativity. The difference in the speed of clocks at different heights above the earth is now of considerable practical importance, with the advent of very accurate navigation systems based on signals from satellites. If one ignored the predictions of general relativity, the position that one calculated would be wrong by several miles!

Newton's laws of motion put an end to the idea of absolute position in space. The theory of relativity gets rid of absolute time. Consider a pair of twins. Suppose that one twin goes to live on the top of a mountain while the other stays at sea level. The first twin would age faster than the second. Thus, if they met again, one would be older than the other. In this case, the difference in ages would be very small, but it would be much larger if one of the twins went for a long trip in a spaceship at nearly the speed of light. When he returned, he would be much younger than the one who stayed on earth. This is known as the twins paradox, but it is a paradox only if one has the idea of absolute time at the back of one's mind. In the theory of relativity there is no unique absolute time, but instead each individual has his own personal measure of time that depends on where he is and how he is moving.

Before 1915, space and time were thought of as a fixed arena in which events took place, but which was not affected by what happened in it. This was true even of the special theory of relativity. Bodies moved, forces attracted and repelled, but time and space simply continued, unaffected. It was natural to think that space and time went on forever.

The situation, however, is quite different in the general theory of relativity. Space and time are now dynamic quantities: when a body moves, or a force acts, it affects the curvature of space and time - and in turn the structure of space-time affects the way in which bodies move and forces act. Space and time not only affect but also are affected by everything that happens in the universe. Just as one cannot talk about events in the universe without the notions of space and time, so in general relativity it became meaningless to talk about space and time outside the limits of the universe.

In the following decades this new understanding of space and time was to revolutionize our view of the universe. The old idea of an essentially unchanging universe that could have existed, and could continue to exist, forever was replaced by the notion of a dynamic,

expanding universe that seemed to have begun a finite time ago, and that might end at a finite time in the future. That revolution forms the subject of the next chapter. And years later, it was also to be the starting point for my work in theoretical physics. Roger Penrose and I showed that Einstein's general theory of relativity implied that the universe must have a beginning and, possibly, an end.

CHAPTER 3
THE EXPANDING UNIVERSE

If one looks at the sky on a clear, moonless night, the brightest objects one sees are likely to be the planets Venus, Mars, Jupiter, and Saturn. There will also be a very large number of stars, which are just like our own sun but much farther from us. Some of these fixed stars do, in fact, appear to change very slightly their positions relative to each other as earth orbits around the sun: they are not really fixed at all! This is because they are comparatively near to us. As the earth goes round the sun, we see them from different positions against the background of more distant stars. This is fortunate, because it enables us to measure directly the distance of these stars from us: the nearer they are, the more they appear to move. The nearest star, called Proxima Centauri, is found to be about four light-years away (the light from it takes about four years to reach earth), or about twenty-three million million miles. Most of the other stars that are visible to the naked eye lie within a few hundred light-years of us. Our sun, for comparison, is a mere light-minutes away! The visible stars appear spread all over the night sky, but are particularly concentrated in one band, which we call the Milky Way. As long ago as 1750, some astronomers were suggesting that the appearance of the Milky Way could be explained if most of the visible stars lie in a single disklike configuration, one example of what we now call a spiral galaxy.

Only a few decades later, the astronomer Sir William Herschel confirmed this idea by painstakingly cataloging the positions and distances of vast numbers of stars. Even so, the idea gained complete acceptance only early this century.

Our modern picture of the universe dates back to only 1924, when the American astronomer Edwin Hubble demonstrated that ours was not the only galaxy. There were in fact many others, with vast tracts of empty space between them. In order to prove this, he needed to determine the distances to these other galaxies, which are so far away that, unlike nearby stars, they really do appear fixed. Hubble was forced, therefore, to use indirect methods to measure the distances. Now, the apparent brightness of a star depends on two factors: how much light it radiates (its luminosity), and how far it is from us. For nearby stars, we can measure their apparent brightness and their distance, and so we can work out their luminosity. Conversely, if we knew the luminosity of stars in other galaxies, we could work out their distance by measuring their apparent brightness. Hubble noted that certain types of stars always have the same luminosity when they are near enough for us to measure; therefore, he argued, if we found such stars in another galaxy, we could assume that they had the same luminosity - and so calculate the distance to that galaxy. If we could do this for a number of stars in the same galaxy, and our calculations always gave the same distance, we could be fairly confident of our estimate.

In this way, Edwin Hubble worked out the distances to nine different galaxies. We now know that our galaxy is only one of some hundred thousand million that can be seen using modern telescopes, each galaxy itself containing some hundred thousand million stars. Fig. 3.1 shows a picture of one spiral galaxy that is similar to what we think ours must look like to someone living in another galaxy. We live in a galaxy that is about one hundred thousand light-years across and is slowly rotating; the stars in its spiral arms orbit around its center about once every several

hundred million years. Our sun is just an ordinary, average-sized, yellow star, near the inner edge of one of the spiral arms. We have certainly come a long way since Aristotle and Ptolemy, when thought that the earth was the center of the universe!

Stars are so far away that they appear to us to be just pinpoints of light. We cannot see their size or shape. So how can we tell different types of stars apart? For the vast majority of stars, there is only one characteristic feature that we can observe - the color of their light. Newton discovered that if light from the sun passes through a triangular-shaped piece of glass, called a prism, it breaks up into its component colors (its spectrum) as in a rainbow. By focusing a telescope on an individual star or galaxy, one can similarly observe the spectrum of the light from that star or galaxy. Different stars have different spectra, but the relative brightness of the different colors is always exactly what one would expect to find in the light emitted by an object that is glowing red hot. (In fact, the light emitted by any opaque object that is glowing red hot has a characteristic spectrum that depends only on its temperature - a thermal spectrum. This means that we can tell a star's temperature from the spectrum of its light.) More-over, we find that certain very specific colors are missing from stars' spectra, and these missing colors may vary from star to star. Since we know that each chemical element absorbs a characteristic set of very specific colors, by matching these to those that are missing from a star's spectrum, we can determine exactly which elements are present in the star's atmosphere.

In the 1920s, when astronomers began to look at the spectra of stars in other galaxies, they found something most peculiar: there were the same characteristic sets of missing colors as for stars in our own galaxy, but they were all shifted by the same relative amount toward the red end of the spectrum. To understand the implications of this, we must first understand the Doppler effect. As we have seen, visible light consists of fluctuations, or waves, in the electromagnetic field. The wavelength (or distance from one

wave crest to the next) of light is extremely small, ranging from four to seven ten-millionths of a meter. The different wavelengths of light are what the human eye sees as different colors, with the longest wavelengths appearing at the red end of the spectrum and the shortest wavelengths at the blue end. Now imagine a source of light at a constant distance from us, such as a star, emitting waves of light at a constant wavelength. Obviously the wave-length of the waves we receive will be the same as the wavelength at which they are emitted (the gravitational field of the galaxy will not be large enough to have a significant effect). Suppose now that the source starts moving toward us. When the source emits the next wave crest it will be nearer to us, so the distance between wave crests will be smaller than when the star was stationary. This means that the wavelength of the waves we receive is shorter than when the star was stationary. Correspondingly, if the source is moving away from us, the wavelength of the waves we receive will be longer. In the case of light, therefore, means that stars moving away from us will have their spectra shifted toward the red end of the spectrum (red-shifted) and those moving toward us will have their spectra blue-shifted. This relationship between wavelength and speed, which is called the Doppler effect, is an everyday experience. Listen to a car passing on the road: as the car is approaching, its engine sounds at a higher pitch (corresponding to a shorter wavelength and higher frequency of sound waves), and when it passes and goes away, it sounds at a lower pitch. The behavior of light or radio waves is similar. Indeed, the police make use of the Doppler effect to measure the speed of cars by measuring the wavelength of pulses of radio waves reflected off them.

ln the years following his proof of the existence of other galaxies, Rubble spent his time cataloging their distances and observing their spectra. At that time most people expected the galaxies to be moving around quite randomly, and so expected to find as many blue-shifted spectra as red-shifted ones. It was quite a surprise, therefore, to find that most galaxies appeared red-shifted: nearly all

were moving away from us! More surprising still was the finding that Hubble published in 1929: even the size of a galaxy's red shift is not random, but is directly proportional to the galaxy's distance from us. Or, in other words, the farther a galaxy is, the faster it is moving away! And that meant that the universe could not be static, as everyone previously had thought, is in fact expanding; the distance between the different galaxies is
g all the time.

The discovery that the universe is expanding was one of the great intellectual revolutions of the twentieth century. With hindsight, it is easy wonder why no one had thought of it before. Newton, and others should have realized that a static universe would soon start to contract under the influence of gravity. But suppose instead that the universe is expanding. If it was expanding fairly slowly, the force of gravity would cause it eventually to stop expanding and then to start contracting. However, if it was expanding at more than a certain critical rate, gravity would never be strong enough to stop it, and the universe would continue to expand forever. This is a bit like what happens when one fires a rocket upward from the surface of the earth. If it has a fairly low speed, gravity will eventually stop the rocket and it will start falling back. On the other hand, if the rocket has more than a certain critical speed (about seven miles per second), gravity will not be strong enough to pull it back, so it will keep going away from the earth forever. This behavior of the universe could have been predicted from Newton's theory of gravity at any time in the nineteenth, the eighteenth, or even the late seventeenth century. Yet so strong was the belief in a static universe that it persisted into the early twentieth century. Even Einstein, when he formulated the general theory of relativity in 1915, was so sure that the universe had to be static that he modified his theory to make this possible, introducing a so-called cosmological constant into his equations. Einstein introduced a new "antigravity" force, which, unlike other forces, did not come from any particular source but was built into the very fabric of

space-time. He claimed that space-time had an inbuilt tendency to expand, and this could be made to balance exactly the attraction of all the matter in the universe, so that a static universe would result. Only one man, it seems, was willing to take general relativity at face value, and while Einstein and other physicists were looking for ways of avoiding general relativity's prediction of a nonstatic universe, the Russian physicist and mathematician Alexander Friedmann instead set about explaining it.

Friedmann made two very simple assumptions about the universe: that the universe looks identical in whichever direction we look, and that this would also be true if we were observing the universe from anywhere else. From these two ideas alone, Friedmann showed that we should not expect the universe to be static. In fact, in 1922, several years before Edwin Hubble's discovery, Friedmann predicted exactly what Hubble found!

The assumption that the universe looks the same in every direction is clearly not true in reality. For example, as we have seen, the other stars in our galaxy form a distinct band of light across the night sky, called the Milky Way. But if we look at distant galaxies, there seems to be more or less the same number of them. So the universe does seem to be roughly the same in every direction, provided one views it on a large scale compared to the distance between galaxies, and ignores the differences on small scales. For a long time, this was sufficient justification for Friedmann's assumption - as a rough approximation to the real universe. But more recently a lucky accident uncovered the fact that Friedmann's assumption is in fact a remarkably accurate description of our universe.

In 1965 two American physicists at the Bell Telephone Laboratories in New Jersey, Arno Penzias and Robert Wilson, were testing a very sensitive microwave detector. (Microwaves are just like light waves, but with a wavelength of around a centimeter.) Penzias and Wilson were worried when they found that their detector was picking up more noise than it ought to. The noise did

not appear to be coming from any particular direction. First they discovered bird droppings in their detector and checked for other possible malfunctions, but soon ruled these out. They knew that any noise from within the atmosphere would be stronger when the detector was not pointing straight up than when it was, because light rays travel through much more atmosphere when received from near the horizon than when received from directly overhead. The extra noise was the same whichever direction the detector was pointed, so it must come from outside the atmosphere. It was also the same day and night and throughout the year, even though the earth was rotating on its axis and orbiting around the sun. This showed that the radiation must come from beyond the Solar System, and even from beyond the galaxy, as otherwise it would vary as the movement of earth pointed the detector in different directions.

In fact, we know that the radiation must have traveled to us across most of the observable universe, and since it appears to be the same in different directions, the universe must also be the same in every direction, if only on a large scale. We now know that whichever direction we look, this noise never varies by more than a tiny fraction: so Penzias and Wilson had unwittingly stumbled across a remarkably accurate confirmation of Friedmann's first assumption. However, be-cause the universe is not exactly the same in every direction, but only on average on a large scale, the microwaves cannot be exactly the same in every direction either. There have to be slight variations between different directions. These were first detected in 1992 by the Cosmic Background Explorer satellite, or COBE, at a level of about one part in a hundred thousand. Small though these variations are, they are very important, as will be explained in Chapter 8.

At roughly the same time as Penzias and Wilson were investigating noise in their detector, two American physicists at nearby Princeton University, Bob Dicke and Jim Peebles, were also taking an interest in microwaves. They were working on a suggestion,

made by George Gamow (once a student of Alexander Friedmann), that the early universe should have been very hot and dense, glowing white hot. Dicke and Peebles argued that we should still be able to see the glow of the early universe, because light from very distant parts of it would only just be reaching us now. However, the expansion of the universe meant that this light should be so greatly red-shifted that it would appear to us now as microwave radiation. Dicke and Peebles were preparing to look for this radiation when Penzias and Wilson heard about their work and realized that they had already found it. For this, Penzias and Wilson were awarded the Nobel Prize in 1978 (which seems a bit hard on Dicke and Peebles, not to mention Gamow!).

Now at first sight, all this evidence that the universe looks the same whichever direction we look in might seem to suggest there is some-thing special about our place in the universe. In particular, it might seem that if we observe all other galaxies to be moving away from us, then we must be at the center of the universe. There is, however, an alternate explanation: the universe might look the same in every direction as seen from any other galaxy too. This, as we have seen, was Friedmann's second assumption. We have no scientific evidence for, or against, this assumption. We believe it only on grounds of modesty: it would be most remarkable if the universe looked the same in every direction around us, but not around other points in the universe! In Friedmann's model, all the galaxies are moving directly away from each other. The situation is rather like a balloon with a number of spots painted on it being steadily blown up. As the balloon expands, the distance between any two spots increases, but there is no spot that can be said to be the center of the expansion. Moreover, the farther apart the spots are, the faster they will be moving apart. Similarly, in Friedmann's model the speed at which any two galaxies are moving apart is proportional to the distance between them. So it predicted that the red shift of a galaxy should be directly proportional to its distance from us, exactly as Hubble found. Despite the success of his model

and his prediction of Hubble's observations, Friedmann's work remained largely unknown in the West until similar models were discovered in 1935 by the American physicist Howard Robertson and the British mathematician Arthur Walker, in response to Hubble's discovery of the uniform expansion of the universe.

Although Friedmann found only one, there are in fact three different kinds of models that obey Friedmann's two fundamental assumptions. In the first kind (which Friedmann found) the universe is expanding sufficiently slowly that the gravitational attraction between the different galaxies causes the expansion to slow down and eventually to stop. The galaxies then start to move toward each other and the universe contracts. Fig. 3.2 shows how the distance between two neighboring galaxies changes as time increases. It starts at zero, increases to a maximum, and then decreases to zero again. In the second kind of solution, the universe is expanding so rapidly that the gravitational attraction can never stop it, though it does slow it down a bit. Fig. 3.3 Shows the Separation between neighboring galaxies in this model. It starts at zero and eventually the galaxies are moving apart at a steady speed. Finally, there is a third kind of solution, in which the universe is expanding only just fast enough to avoid recollapse. In this case the separation, shown in Fig. 3.4, also starts at zero and increases forever. However, the speed at which the galaxies are moving apart gets smaller and smaller, although it never quite reaches zero.

A remarkable feature of the first kind of Friedmann model is that in it the universe is not infinite in space, but neither does space have any boundary. Gravity is so strong that space is bent round onto itself, making it rather like the surface of the earth. If one keeps traveling in a certain direction on the surface of the earth, one never comes up against an impassable barrier or falls over the edge, but eventually comes back to where one started.

In the first kind of Friedmann model, space is just like this, but with three dimensions instead of two for the earth's surface. The fourth dimension, time, is also finite in extent, but it is like a line with two ends or boundaries, a beginning and an end. We shall see later that when one combines general relativity with the uncertainty principle of quantum mechanics, it is possible for both space and time to be finite without any edges or boundaries.

The idea that one could go right round the universe and end up where one started makes good science fiction, but it doesn't have much practical significance, because it can be shown that the universe would recollapse to zero size before one could get round. You would need to travel faster than light in order to end up where you started before the universe came to an end - and that is not allowed!

In the first kind of Friedmann model, which expands and recollapses, space is bent in on itself, like the surface of the earth. It is therefore finite in extent. In the second kind of model, which expands forever, space is bent the other way, like the surface of a saddle. So in this case space is infinite. Finally, in the third kind of Friedmann model, with just the critical rate of expansion, space is flat (and therefore is also infinite).

But which Friedmann model describes our universe? Will the universe eventually stop expanding and start contracting, or will it expand forever? To answer this question we need to know the present rate of expansion of the universe and its present average density. If the density is less than a certain critical value, determined by the rate of expansion, the gravitational attraction will be too weak to halt the expansion. If the density is greater than the critical value, gravity will stop the expansion at some time in the future and cause the universe to recollapse.

We can determine the present rate of expansion by measuring the velocities at which other galaxies are moving away from us, using the Doppler effect. This can be done very accurately. However, the distances to the galaxies are not very well known because we can

only measure them indirectly. So all we know is that the universe is expanding by between 5 percent and 10 percent every thousand million years. However, our uncertainty about the present average density of the universe is even greater. If we add up the masses of all the stars that we can see in our galaxy and other galaxies, the total is less than one hundredth of the amount required to halt the expansion of the universe, even for the lowest estimate of the rate of expansion. Our galaxy and other galaxies, however, must contain a large amount of "dark matter" that we cannot see directly, but which we know must be there because of the influence of its gravitational attraction on the orbits of stars in the galaxies. Moreover, most galaxies are found in clusters, and we can similarly infer the presence of yet more dark matter in between the galaxies in these clusters by its effect on the motion of the galaxies. When we add up all this dark matter, we still get only about one tenth of the amount required to halt the expansion. However, we cannot exclude the possibility that there might be some other form of matter, distributed almost uniformly throughout the universe, that we have not yet detected and that might still raise the average density of the universe up to the critical value needed to halt the expansion. The present evidence therefore suggests that the universe will probably expand forever, but all we can really be sure of is that even if the universe is going to recollapse, it won't do so for at least another ten thousand million years, since it has already been expanding for at least that long. This should not unduly worry us: by that time, unless we have colonized beyond the Solar System, mankind will long since have died out, extinguished along with our sun!

All of the Friedmann solutions have the feature that at some time in the past (between ten and twenty thousand million years ago) the distance between neighboring galaxies must have been zero. At that time, which we call the big bang, the density of the universe and the curvature of space-time would have been infinite. Because mathematics cannot really handle infinite numbers, this means that

the general theory of relativity (on which Friedmann's solutions are based) predicts that there is a point in the universe where the theory itself breaks down. Such a point is an example of what mathematicians call a singularity. In fact, all our theories of science are formulated on the assumption that space-time is smooth and nearly fiat, so they break down at the big bang singularity, where the curvature of space-time is infinite. This means that even if there were events before the big bang, one could not use them to determine what would happen afterward, because predictability would break down at the big bang.

Correspondingly, if, as is the case, we know only what has happened since the big bang, we could not determine what happened beforehand. As far as we are concerned, events before the big bang can have no consequences, so they should not form part of a scientific model of the universe. We should therefore cut them out of the model and say that time had a beginning at the big bang.

Many people do not like the idea that time has a beginning, probably because it smacks of divine intervention. (The Catholic Church, on the other hand, seized on the big bang model and in 1951officially pronounced it to be in accordance with the Bible.) There were therefore a number of attempts to avoid the conclusion that there had been a big bang. The proposal that gained widest support was called the steady state theory. It was suggested in 1948 by two refugees from Nazi-occupied Austria, Hermann Bondi and Thomas Gold, together with a Briton, Fred Hoyle, who had worked with them on the development of radar during the war. The idea was that as the galaxies moved away from each other, new galaxies were continually forming in the gaps in between, from new matter that was being continually created. The universe would therefore look roughly the same at all times as well as at all points of space. The steady state theory required a modification of general relativity to allow for the continual creation of matter, but the rate that was involved was so low (about one particle per cubic kilometer per

year) that it was not in conflict with experiment. The theory was a good scientific theory, in the sense described in Chapter 1: it was simple and it made definite predictions that could be tested by observation. One of these predictions was that the number of galaxies or similar objects in any given volume of space should be the same wherever and whenever we look in the universe. In the late 1950s and early 1960s a survey of sources of radio waves from outer space was carried out at Cambridge by a group of astronomers led by Martin Ryle (who had also worked with Bondi, Gold, and Hoyle on radar during the war). The Cambridge group showed that most of these radio sources must lie outside our galaxy (indeed many of them could be identified with other galaxies) and also that there were many more weak sources than strong ones. They interpreted the weak sources as being the more distant ones, and the stronger ones as being nearer. Then there appeared to be less common sources per unit volume of space for the nearby sources than for the distant ones. This could mean that we are at the center of a great region in the universe in which the sources are fewer than elsewhere. Alternatively, it could mean that the sources were more numerous in the past, at the time that the radio waves left on their journey to us, than they are now. Either explanation contradicted the predictions of the steady state theory. Moreover, the discovery of the microwave radiation by Penzias and Wilson in 1965 also indicated that the universe must have been much denser in the past. The steady state theory therefore had to be abandoned.

Another attempt to avoid the conclusion that there must have been a big bang, and therefore a beginning of time, was made by two Russian scientists, Evgenii Lifshitz and Isaac Khalatnikov, in 1963. They suggested that the big bang might be a peculiarity of Friedmann's models alone, which after all were only approximations to the real universe. Perhaps, of all the models that were roughly like the real universe, only Friedmann's would contain a big bang singularity. In Friedmann's models, the galaxies are all moving directly away from each other - so it is not

surprising that at some time in the past they were all at the same place. In the real universe, however, the galaxies are not just moving directly away from each other - they also have small sideways velocities. So in reality they need never have been all at exactly the same place, only very close together. Perhaps then the current expanding universe resulted not from a big bang singularity, but from an earlier contracting phase; as the universe had collapsed the particles in it might not have all collided, but had flown past and then away from each other, producing the present expansion of the the universe that were roughly like Friedmann's models but took account of the irregularities and random velocities of galaxies in the real universe. They showed that such models could start with a big bang, even though the galaxies were no longer always moving directly away from each other, but they claimed that this was still only possible in certain exceptional models in which the galaxies were all moving in just the right way. They argued that since there seemed to be infinitely more Friedmann-like models without a big bang singularity than there were with one, we should conclude that there had not in reality been a big bang. They later realized, however, that there was a much more general class of Friedmann-like models that did have singularities, and in which the galaxies did not have to be moving any special way. They therefore withdrew their claim in 1970.

The work of Lifshitz and Khalatnikov was valuable because it showed that the universe could have had a singularity, a big bang, if the general theory of relativity was correct. However, it did not resolve the crucial question: Does general relativity predict that our universe should have had a big bang, a beginning of time? The answer to this carne out of a completely different approach introduced by a British mathematician and physicist, Roger Penrose, in 1965. Using the way light cones behave in general relativity, together with the fact that gravity is always attractive, he showed that a star collapsing under its own gravity is trapped in a region whose surface eventually shrinks to zero size. And, since

the surface of the region shrinks to zero, so too must its volume. All the matter in the star will be compressed into a region of zero volume, so the density of matter and the curvature of space-time become infinite. In other words, one has a singularity contained within a region of space-time known as a black hole.

At first sight, Penrose's result applied only to stars; it didn't have anything to say about the question of whether the entire universe had a big bang singularity in its past. However, at the time that Penrose produced his theorem, I was a research student desperately looking for a problem with which to complete my Ph.D. thesis. Two years before, I had been diagnosed as suffering from ALS, commonly known as Lou Gehrig's disease, or motor neuron disease, and given to understand that I had only one or two more years to live. In these circumstances there had not seemed much point in working on my Ph.D.- I did not expect to survive that long. Yet two years had gone by and I was not that much worse. In fact, things were going rather well for me and I had gotten engaged to a very nice girl, Jane Wilde. But in order to get married, I needed a job, and in order to get a job, I needed a Ph.D.

In 1965 I read about Penrose's theorem that any body undergoing gravitational collapse must eventually form a singularity. I soon realized that if one reversed the direction of time in Penrose's theorem, so that the collapse became an expansion, the conditions of his theorem would still hold, provided the universe were roughly like a Friedmann model on large scales at the present time. Penrose's theorem had shown that any collapsing star must end in a singularity; the time-reversed argument showed that any Friedmann-like expanding universe must have begun with a singularity. For technical reasons, Penrose's theorem required that the universe be infinite in space. So I could in fact, use it to prove that there should be a singularity only if the universe was expanding fast enough to avoid collapsing again (since only those Friedmann models were infinite in space).

During the next few years I developed new mathematical techniques to remove this and other technical conditions from the theorems that proved that singularities must occur. The final result was a joint paper by Penrose and myself in 1970, which at last proved that there must have been a big bang singularity provided only that general relativity is correct and the universe contains as much matter as we observe. There was a lot of opposition to our work, partly from the Russians because of their Marxist belief in scientific determinism, and partly from people who felt that the whole idea of singularities was repugnant and spoiled the beauty of Einstein's theory. However, one cannot really argue with a mathematical theorem. So in the end our work became generally accepted and nowadays nearly everyone assumes that the universe started with a big bang singularity. It is perhaps ironic that, having changed my mind, I am now trying to convince other physicists that there was in fact no singularity at the beginning of the universe - as we shall see later, it can disappear once quantum effects are taken into account.

We have seen in this chapter how, in less than half a century, man's view of the universe formed over millennia has been transformed. Hubble's discovery that the universe was expanding, and the realization of the insignificance of our own planet in the vastness of the universe, were just the starting point. As experimental and theoretical evidence mounted, it became more and more clear that the universe must have had a beginning in time, until in 1970 this was finally proved by Penrose and myself, on the basis of Einstein's general theory of relativity. That proof showed that general relativity is only an incomplete theory: it cannot tell us how the universe started off, because it predicts that all physical theories, including itself, break down at the beginning of the universe. However, general relativity claims to be only a partial theory, so what the singularity theorems really show is that there must have been a time in the very early universe when the universe was so small that one could no longer ignore the small-scale effects

of the other great partial theory of the twentieth century, quantum mechanics. At the start of the 1970s, then, we were forced to turn our search for an understanding of the universe from our theory of the extraordinarily vast to our theory of the extraordinarily tiny. That theory, quantum mechanics, will be described next, before we turn to the efforts to combine the two partial theories into a single quantum theory of gravity.

CHAPTER 4
THE UNCERTAINTY PRINCIPLE

The success of scientific theories, particularly Newton's theory of gravity, led the French scientist the Marquis de Laplace at the beginning of the nineteenth century to argue that the universe was completely deterministic. Laplace suggested that there should be a set of scientific laws that would allow us to predict everything that would happen in the universe, if only we knew the complete state of the universe at one time. For example, if we knew the positions and speeds of the sun and the planets at one time, then we could use Newton's laws to calculate the state of the Solar System at any other time. Determinism seems fairly obvious in this case, but Laplace went further to assume that there were similar laws governing everything else, including human behavior.

The doctrine of scientific determinism was strongly resisted by many people, who felt that it infringed God's freedom to intervene in the world, but it remained the standard assumption of science until the early years of this century. One of the first indications that this belief would have to be abandoned came when calculations by the British scientists Lord Rayleigh and Sir James Jeans suggested that a hot object, or body, such as a star, must radiate energy at an infinite rate. According to the laws we believed at the time, a hot body ought to give off electromagnetic waves (such as radio waves, visible light, or X rays) equally at all frequencies. For example, a hot body should radiate the same amount of energy in

waves with frequencies between one and two million million waves a second as in waves with frequencies between two and three million million waves a second. Now since the number of waves a second is unlimited, this would mean that the total energy radiated would be infinite.

In order to avoid this obviously ridiculous result, the German scientist Max Planck suggested in 1900 that light, X rays, and other waves could not be emitted at an arbitrary rate, but only in certain packets that he called quanta. Moreover, each quantum had a certain amount of energy that was greater the higher the frequency of the waves, so at a high enough frequency the emission of a single quantum would require more energy than was available. Thus the radiation at high frequencies would be reduced, and so the rate at which the body lost energy would be finite.

The quantum hypothesis explained the observed rate of emission of radiation from hot bodies very well, but its implications for determinism were not realized until 1926, when another German scientist, Werner Heisenberg, formulated his famous uncertainty principle. In order to predict the future position and velocity of a particle, one has to be able to measure its present position and velocity accurately. The obvious way to do this is to shine light on the particle. Some of the waves of light will be scattered by the particle and this will indicate its position. However, one will not be able to determine the position of the particle more accurately than the distance between the wave crests of light, so one needs to use light of a short wavelength in order to measure the position of the particle precisely. Now, by Planck's quantum hypothesis, one cannot use an arbitrarily small amount of light; one has to use at least one quantum. This quantum will disturb the particle and change its velocity in a way that cannot be predicted. moreover, the more accurately one measures the position, the shorter the wavelength of the light that one needs and hence the higher the energy of a single quantum. So the velocity of the particle will be disturbed by a larger amount. In other words, the more accurately

you try to measure the position of the particle, the less accurately you can measure its speed, and vice versa. Heisenberg showed that the uncertainty in the position of the particle times the uncertainty in its velocity times the mass of the particle can never be smaller than a certain quantity, which is known as Planck's constant. Moreover, this limit does not depend on the way in which one tries to measure the position or velocity of the particle, or on the type of particle: Heisenberg's uncertainty principle is a fundamental, inescapable property of the world.

The uncertainty principle had profound implications for the way in which we view the world. Even after more than seventy years they have not been fully appreciated by many philosophers, and are still the subject of much controversy. The uncertainty principle signaled an end to Laplace's dream of a theory of science, a model of the universe that would be completely deterministic: one certainly cannot predict future events exactly if one cannot even measure the present state of the universe precisely! We could still imagine that there is a set of laws that determine events completely for some supernatural being, who could observe the present state of the universe without disturbing it. However, such models of the universe are not of much interest to us ordinary mortals. It seems better to employ the principle of economy known as Occam's razor and cut out all the features of the theory that cannot be observed. This approach led Heisenberg, Erwin Schrodinger, and Paul Dirac in the 1920s to reformulate mechanics into a new theory called quantum mechanics, based on the uncertainty principle. In this theory particles no longer had separate, well-defined positions and velocities that could not be observed, Instead, they had a quantum state, which was a combination of position and velocity.

In general, quantum mechanics does not predict a single definite result for an observation. Instead, it predicts a number of different possible outcomes and tells us how likely each of these is. That is to say, if one made the same measurement on a large number of similar systems, each of which started off in the same way, one

would find that the result of the measurement would be A in a certain number of cases, B in a different number, and so on. One could predict the approximate number of times that the result would be A or B, but one could not predict the specific result of an individual measurement. Quantum mechanics therefore introduces an unavoidable element of unpredictability or randomness into science. Einstein objected to this very strongly, despite the important role he had played in the development of these ideas. Einstein was awarded the Nobel Prize for his contribution to quantum theory. Nevertheless, Einstein never accepted that the universe was governed by chance; his feelings were summed up in his famous statement "God does not play dice." Most other scientists, however, were willing to accept quantum mechanics because it agreed perfectly with experiment. Indeed, it has been an outstandingly successful theory and underlies nearly all of modern science and technology. It governs the behavior of transistors and integrated circuits, which are the essential components of electronic devices such as televisions and computers, and is also the basis of modern chemistry and biology. The only areas of physical science into which quantum mechanics has not yet been properly incorporated are gravity and the large-scale structure of the universe.

Although light is made up of waves, Planck's quantum hypothesis tells us that in some ways it behaves as if it were composed of particles: it can be emitted or absorbed only in packets, or quanta. Equally, Heisenberg's uncertainty principle implies that particles behave in some respects like waves: they do not have a definite position but are "smeared out" with a certain probability distribution. The theory of quantum mechanics is based on an entirely new type of mathematics that no longer describes the real world in terms of particles and waves; it is only the observations of the world that may be described in those

terms. There is thus a duality between waves and particles in quantum mechanics: for some purposes it is helpful to think of

particles as waves and for other purposes it is better to think of waves as particles. An important consequence of this is that one can observe what is called interference between two sets of waves or particles. That is to say, the crests of one set of waves may coincide with the troughs of the other set. The two sets of waves then cancel each other out rather than adding up to a stronger wave as one might expect (Fig. 4.1). A familiar example of interference in the case of light is the colors that are often seen in soap bubbles. These are caused by reflection of light from the two sides of the thin film of water forming the bubble. White light consists of light waves of all different wavelengths, or colors, For certain wavelengths the crests of the waves reflected from one side of the soap film coincide with the troughs reflected from the other side. The colors corresponding to these wavelengths are absent from the reflected light, which therefore appears to be colored. Interference can also occur for particles, because of the duality introduced by quantum mechanics. A famous example is the so-called two-slit experiment (Fig. 4.2). Consider a partition with two narrow parallel slits in it. On one side of the partition one places a source of fight of a particular color (that is, of a particular wavelength). Most of the light will hit the partition, but a small amount will go through the slits. Now suppose one places a screen on the far side of the partition from the light. Any point on the screen will receive waves from the two slits. However, in general, the distance the light has to travel from the source to the screen via the two slits will be different. This will mean that the waves from the slits will not be in phase with each other when they arrive at the screen: in some places the waves will cancel each other out, and in others they will reinforce each other. The result is a characteristic pattern of light and dark fringes.

The remarkable thing is that one gets exactly the same kind of fringes if one replaces the source of light by a source of particles such as electrons with a definite speed (this means that the corresponding waves have a definite length). It seems the more