A Brief History of Time

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 everything 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.

A Brief History of Time

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 describe 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 northwest 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