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. Figure 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.

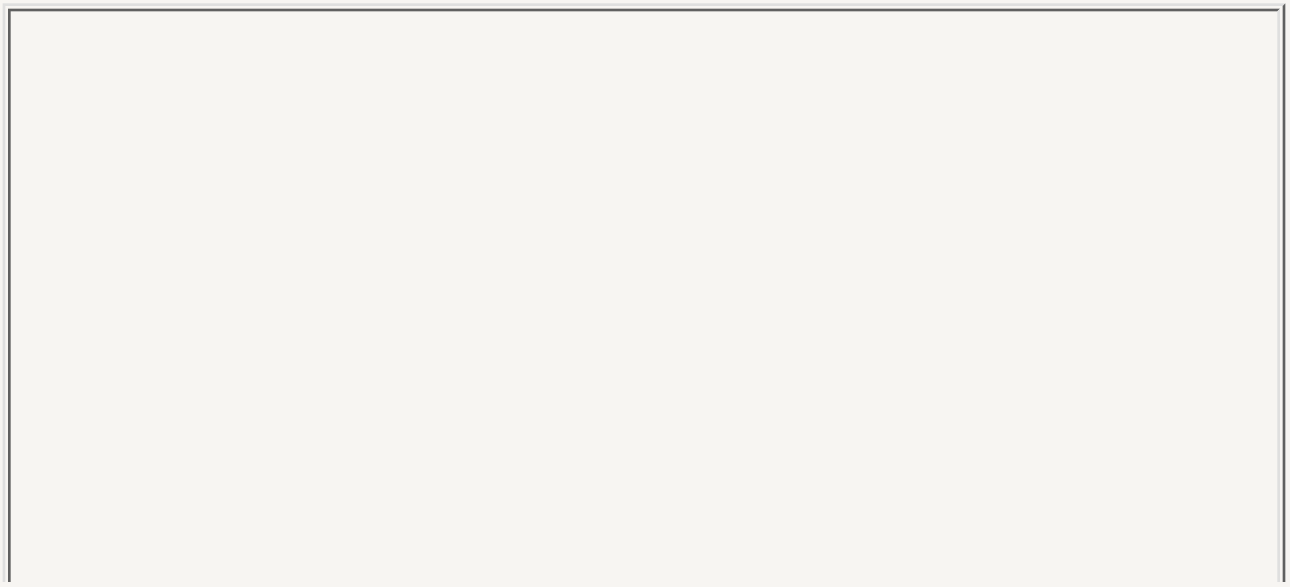

Figure 3:1

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.) Moreover, 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 wavelength 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.

In 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 changing 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, because 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 something 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.

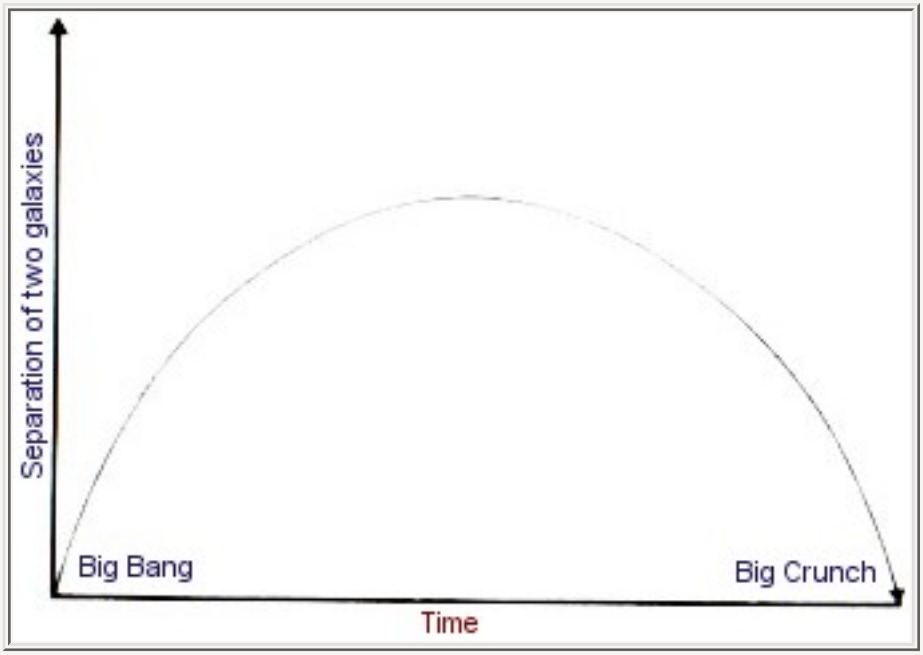

Figure 3:2

Figure 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.

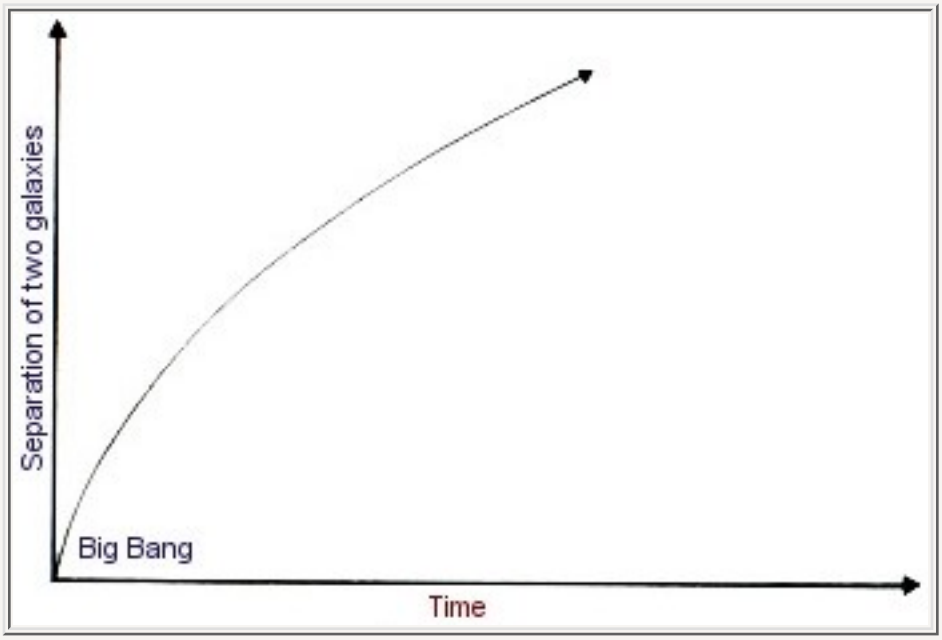

Figure 3:3

Figure 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.

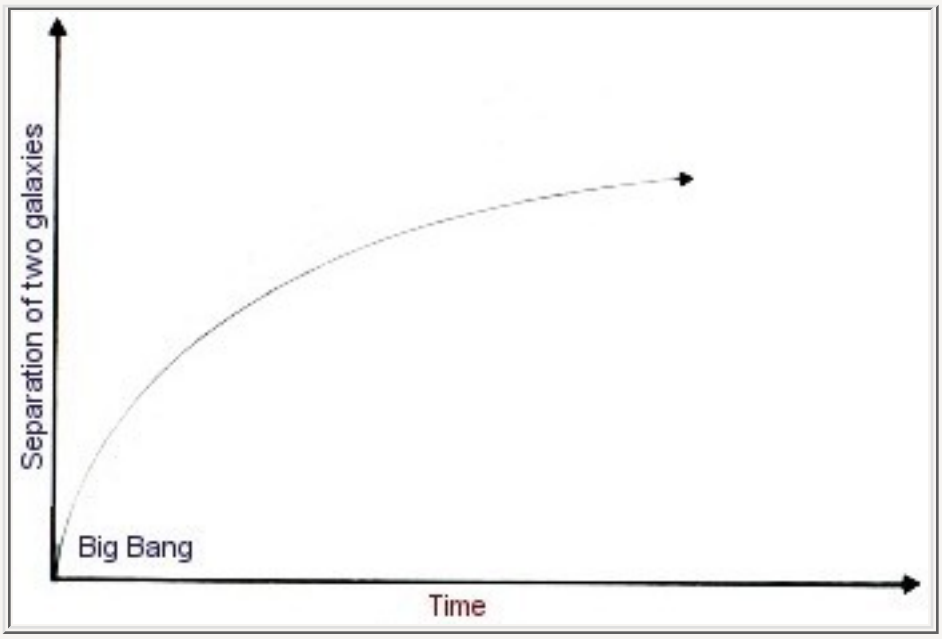

Figure 3:4

In this case the separation, shown in Figure 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.

PREVIOUS NEXT