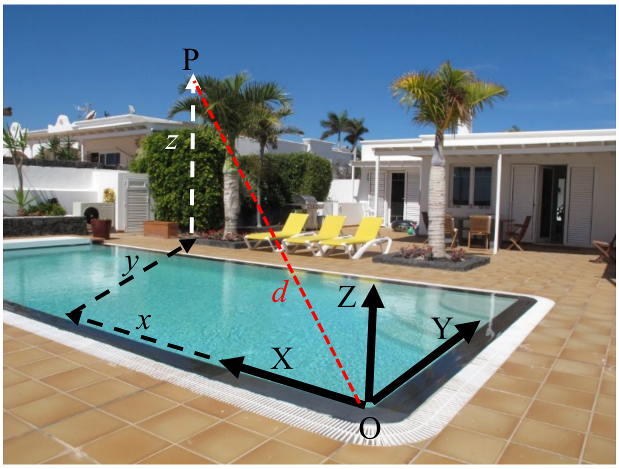
Space and Space-time

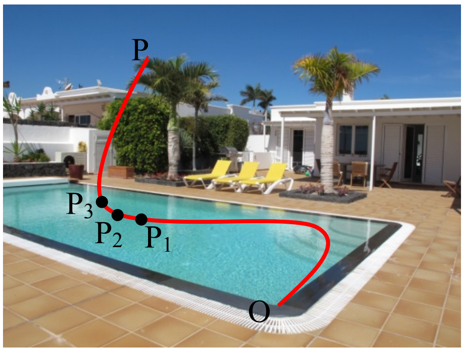
Is space-time really any different from space? Yes it is! I plan to creep up on it. First I'll talk about *ordinary space*, then the *space-time* that Minkowski introduced (Minkowski space) and finally *curved space-time* that comes into General Relativity. The point of this piece is to discuss some of the ideas and language used when talking about Special and General Relativity. I'm writing this in a villa in Lanzarote, hence the first illustration. The University view that scientists sit in their cells, I mean offices, in front of a desktop computer and think deeply about nature is fairly wide of the mark. Yes, we work things out in our offices but they aren't the most creative of places. I'll get on with what I'm planning to say.



Measuring distance in one dimension is easy and non-controversial. One starts somewhere (*the origin*) and measures along to the point of interest. In this way coordinates are measured. To locate a point in 3 dimensions, its *coordinates* are measured along axes in the 3 dimensions. To avoid any complications, my axes will be at right-angles to each other so that they are all independent. In the picture, the origin is at the corner of the swimming pool of our villa. The near sides are the *X* and *Y* axes and the *Z* axis is for height. The coordinates of a point P near the top of the palm tree are (x,y,z) in an obvious notation.

The relevant question is to ask what is the *distance*, d, from the origin O to P? Ordinary space is *Euclidean*, meaning that the distance rule is that $d^2 = x^2 + y^2 + z^2$. Clearly the distance d is given by the square root of d^2. This relationship defines the *metric* of ordinary space, basically how distances are measured. The *straight line* from O to P is the shortest distance from O to P. Is this not making a meal of the obvious? Well, it's always good to start from the obvious.

I haven't quite finished with the obvious. If you want to measure **from P** a short distance ds whose coordinates are dx, dy, dz (mathematically, these symbols stand for infinitesimally small distances but we're not going to worry about the finer points here) then the metric relationship is that $ds^2 = dx^2 + dy^2 + dz^2$. The special feature of Euclidean space is that this relationship doesn't include any mention of where the point P is located. It works everywhere. We'll find that the *metric* in curved space-time is different in that it has terms that depend on where you are.



The metric underlies how distance is measured along any path. For instance, how far has a bird flown that takes off from O and flies to P along the curved path shown? (The birds were very fond of the edge of our pool, treating it like a bird bath). On any bit of the path you can take a point like P_1 and find the short coordinates to the next point P_2 and use the metric to work out the short (ideally infinitesimal) distance to P_2. Then take the next step from P_2 to P_3, and so on.

Adding up all the short steps gives the total path length. It may sound a bit tedious but that's actually what you have to do to find the length of any path.

Well, maybe not for **any** path, for there's one simple case, that of the shortest path between O and P. That's just the straight line between O and P and because the metric in Euclidean space is as simple as they get then adding up all the bits of the path gives the very simple answer for the value of d mentioned earlier.

There's another feature of the way things are done in ordinary life which is not quite so obvious but which also represents the simplest of circumstances. The distance found between two points doesn't depend on the speed of an observer making the measurements. Of course how a moving observer would measure distances fixed in another frame of reference is something to think about. For example how do I measure the length of a platform as I pass by it at speed in a train? If I were passing my scene in a bus, then I could determine the distance OP and if I measured correctly I would find it to be d. Likewise if I passed in a plane. Just as well, really, for the Ordnance Survey would be in trouble if the distance found between two points on the ground depended on the speed of their survey planes. In the language of Physics and Maths, the distance d is *invariant* with the speed of the observer.

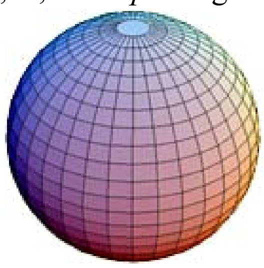
I'll add that in Newtonian physics, *time* is a completely different idea from space and the time interval found between two events is also invariant with the speed of the observer. We'll find that things will be different in space-time.

Curvature and why ordinary space is 'flat'

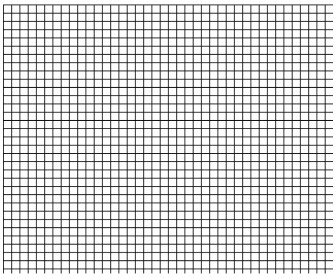
What I've done so far is to put obvious ideas, hopefully ones you pretty well knew in any case, into less obvious language. The point of doing so is that this language and the ideas behind it are concepts that are good for describing the less obvious ideas on which Relativity is built. Many elderly physicists in Einstein's day couldn't get it, but more than 100 years on last century's shock of the new has become this century's background knowledge.

Curvature is a concept that is central to General Relativity. We associate curvature with the bending of straight lines but the curvature Einstein needed is curvature of space and time and that's a harder idea to get hold of. We need to continue creeping up on it.

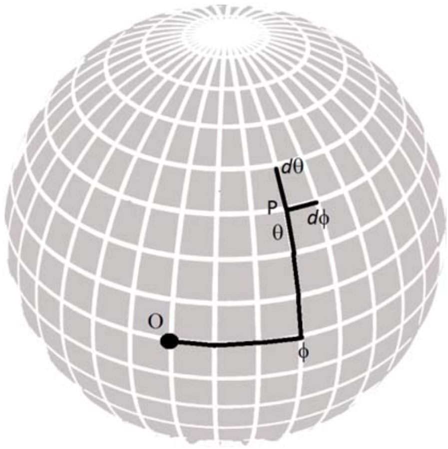
Sticking with ordinary space for the moment, a *surface* is defined by a single relationship between (x,y,z) coordinates. Thus $lx + my + nz = p$ defines a plane, with l, m, n and p being constants for a given plane. Twiddling with the values of l, m, n and p one can alter the orientation of the plane and how far it is from the origin. Taking another example, $x^2 + y^2 + z^2 = r^2$ describes a sphere centred on the origin with radius r. Surfaces have no thickness so they are really 2 dimensional objects. For example after setting up suitable axes, by using coordinates of latitude and longitude, any point on a sphere can be located by just 2 numbers. The two dimensional sphere, however, is *embedded* in 3 dimensions and we need this extra dimension to appreciate its curved nature.



[You might be stuck here by the assertion that a spherical surface is just two-dimensional. Surely a ball is a 3D object? Well, yes it is but a ball is bounded by a closed surface with an internal volume. A football encloses air, a lifebelt is toroidal and may enclose polystyrene, a pear is, well, pear-shaped and encloses sweet mushy stuff and seeds, etc. A cap shape is an example of a surface that's not closed. Whether closed or not, a surface is a mathematical idealisation that has no thickness. In this sense it is two-dimensional. Whether it encloses a volume or not doesn't change its dimension.]

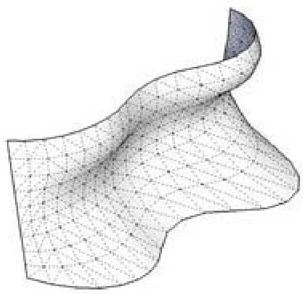


There is one readily visualisable difference between a plane surface and a spherical surface that in a way doesn't need an extra dimension to show itself. The plane surface can be covered with identical squares of unit length *dx* and *dy* defining a measuring grid. This makes it *flat*. Euclidean space can be covered with identical measuring cubes and this is why Euclidean space is defined as flat. It's not 'flat' as with a pancake, but flat because it has no curvature.



A sphere (of radius r) can be covered with unit shapes for latitude and longitude whose sides always point due north or due east but they aren't all identical in size or shape. Take a point P on a sphere whose longitude is the angle ϕ and whose latitude is the angle θ. Measuring a small distance *ds* from P we must account for the fact that 1° of longitude is shorter that it is at the origin point O. The metric is then described by $ds^2 = r^2 d\theta^2 + r^2\cos^2\theta d\phi^2$. Curvature therefore makes itself apparent by changes in the metric used for measurement. Because of the dependence of the metric on where you are, the sphere is *curved* without any reference to an extra dimension.

The metric, remember, is the idea behind distance. A straight line is the shortest distance in the 'space' being described. **Distance on a surface is always measured over the surface**. On a sphere, the shortest distance between two points is the shorter arc of the great circle between the two points. From a viewpoint in the embedding dimension a great circle looks curved but on the sphere it is effectively a straight line. The geometry of lines on a sphere seems odd. You can't draw another straight line (great circle) parallel to a given straight line; the internal angles of a triangle no longer add up to 180°, and so on. The metric on a general curved surface can be complicated but it will tell you how to measure distance everywhere on the surface. This is a kind of warning that curved space-time when we get around to it isn't going to be intuitive.



So the world is round, to use the everyday term, but it sits in a flat (Euclidean) space as far as our usual descriptions go.

Minkowski space

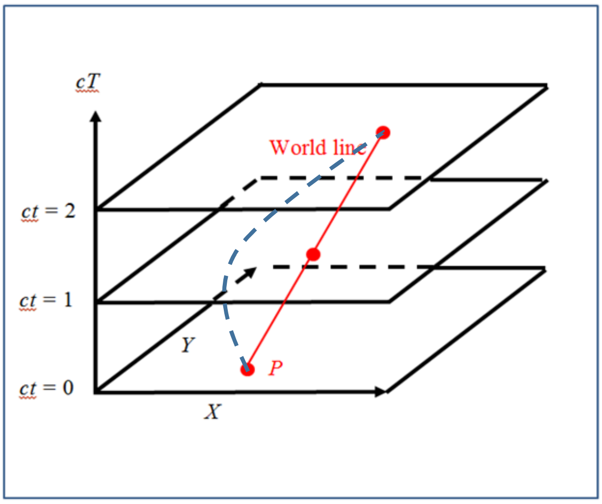
Hermann Minkowski taught Albert Einstein in one of his maths classes when Einstein was a student at ETH Zurich. Einstein was a less than star pupil so by the time Minkowski had moved to Göttingen in 1902 he had probably forgotten about Herr Einstein. He would surely have been surprised to realise that it was the same Albert Einstein who published two papers in Annalen der Physik in 1905 setting out the fundamentals of what would become 'Special Relativity'. By 1908 Minkowski had shown that Special Relativity could be elegantly presented in theoretical terms by introducing the concept of *space-time*, often now called *Minkowski space* in the context of Special Relativity. In fact it was Minkowski's work that made theoreticians take Einstein very seriously, for in 1908 Einstein was still an obscure patent clerk in Bern.

Space-time is not just a matter of adding a time dimension (T) conceptually at right angles to however many space dimensions are relevant in a given situation. Oh, no. Minkowski space is intrinsically different because it has a different metric from Euclidean space. Let's go to the hard bit first. The word 'space' in 'Minkowski space' is best understood as 'realm'. The realm of Minkowski space has one axis with coordinate ct and the other axes with coordinates x, y and z. That's 4 coordinates, hence Minkowski space is 4 dimensional. As usual, c is the speed of light in vacuum and ct is a distance, so the 4th 'dimension' is another distance. The speed of light is the constant chosen to multiply time by so that the transformations between one frame of reference and another moving at a constant velocity are correctly given by the equations in Special Relativity. That's a digression. We'll just accept that ct is the right choice.

A point in Minkowski space is called *an event* since it involves both place and time. The idea of distance is so closely associated with ordinary space that the equivalent in Minkowski space is called the *interval between events*, sometimes written s. An interval is a line in Minkowski space. Now intervals can be either *time-like* or *space-like*. As you might guess, time-like intervals are nearer the time axis. The interval, though, is determined by the *Minkowski metric* $s^2 = (c^2t^2 - x^2 - y^2 - z^2)$. In fact for time-like intervals you can find a frame of reference moving at less than the speed of light in which the two events (at either end of the interval) take place at the same spot. In this frame the interval represents just a time, usually called the *proper time* and given the symbol τ (Greek tau). So you'll see the metric written $c^2\tau^2 = (c^2t^2 - x^2 - y^2 - z^2)$. The consequences of having such a different metric from Euclidean space aren't obvious. [For intervals that are space-like, it is convenient to use the metric $s^2 = (x^2 + y^2 + z^2 - c^2t^2)$. In this case one can find a frame of reference where the events take place at the same time and the interval becomes the *proper distance*].

I should really have said that the metric was $ds^2 = c^2dt^2 - dx^2 - dy^2 - dz^2$ but you can see that there is no need when measuring distance to know where and when you are. So the metric may be odd but Minkowski space is *flat* in the extended meaning of the word. That's a relief. What does the odd metric imply?

Clearly it's not possible to draw diagrams of what's going on in all 4 dimensions of Minkowski space. About the best that can be done is to use only 2 spatial dimensions, with say X and Y axes, and the third axis as cT. How points move around in the XY plane is represented up the time line. For example at ct = 0 a point P is at its starting point in the XY plane;



at $ct = 1$ it has moved a bit and its new x,y coordinates can be plotted on a slice in Minkowski space parallel to the starting plane at unit distance up the cT axis; likewise for $ct = 2$, etc. The track of a point in Minkowski space is given the grandiose title of a *world line*.

This seems fine in terms of symbols but if you think about using everyday units – metres for distance, seconds for time and metres per second for speed – then intervals, s, are measured in metres. Given that $c = 3\times10^8$ m s^{-1}, there is an issue. The world lines of things moving around at ordinary speeds are lines hugging the cT axis. The distinctive features of Minkowski space will only become relevant for objects moving very quickly, in fact at least a few percent of the speed of light. 'ct ' is a distance and $ct = 1$ metre when $t = 3.33$ ns (nanoseconds). So if we are thinking spatial distances in metres, then how far has the point moved in a few nanoseconds? If it's less than a few centimetres, then forget about Minkowski space. Equivalently, if time in seconds is a convenient unit then the light-second is the convenient distance unit. A light second is about 300,000 km, almost here to the Moon, a long distance for a unit of length.

What passes for a straight line in space-time is called a *geodesic*. But what exactly is a 'straight line' when the interval between two events is measured by the Minkowski metric? A straight line in space-time is the world line taken by a point travelling at constant speed between the two events. For example, if a point P starts at the origin and travels with speed v_x along the X axis, v_y along the Y axis and v_z along the Z axis then the coordinates of P with increasing time will be (v_xt, v_yt, v_zt) and the interval between the origin and the moving point will be given by $s^2 = (c^2t^2 - v^2t^2)$, since the velocity of the point is given by $v^2 = v_x^2 + v_y^2 + v_z^2$. Hence the interval $s = \sqrt{}(c^2 - v^2)t$. So the interval increases linearly with time. I've drawn a straight world line in red in the previous diagram.

Since light always travels at a constant speed in a vacuum (the speed c, whatever the frame of reference), then in all frames of reference light travels along geodesics. Diagrams can be deceptive, though, for we 'see' the world line as a length but this isn't the interval between points on the world line. The interval is s. If the world line represented a ray of light, then the interval along any length is zero, as can be seen by setting $v = c$ in the expression in the previous paragraph. [For this reason, books on Relativity describe a light ray as following a 'null geodesic'].

For a curved world line between two points, like the blue dashed line in the diagram, the interval along the line is calculated as the sum of all the tiny intervals between successive points, just as the path distance of the bird flying was calculated at the foot of the first page. One of the properties of Minkowski space that gives rise to 'weird' features of Special Relativity is that the interval along any two world lines between two events is the same. The equality of intervals along different paths is the feature behind the famous 'twin paradox' of Special Relativity. In the frame of reference of the twin who stays at home his/her world line is a straight line up the ct axis. Since there is no motion in space, the interval is simply $c\tau$. For the twin who goes away the interval has both time and space components and since the total interval is the same when they meet again then the time component must be less than τ. This just sounds like a piece of maths but the reality is that the elapsed time in the departing twin's life is less than the stay-at-home twin. The more general conclusion is that the straight world line (the geodesic) has the greatest *proper time*. This concept can be made the definition of a geodesic that will work in curved space-time too. The geodesic is the path that maximises the proper time over all neighbouring paths. That's maybe getting a bit abstract but if you look at General Relativity texts this is how a geodesic is usually found.

[Another feature of Minkowski space is that the *interval between 2 events is invariant for observers travelling at different velocities*. This is like the invariance of distance with the speed of observers mentioned on page 2. It is the basis of time dilation in Special Relativity, the phenomena that "moving clocks run slow". The underlying reason is that the interval between two consecutive ticks of a clock has different spatial components in the two different frames of reference and, since the interval itself is invariant, the time component must also be different. Very odd, when first met. Some of Einstein's older contemporaries never accepted it.]

Curved space-time

OK, I need to get back to the concept of curvature. I've said that curvature shows up in the metric used to describe a 'space' but I haven't said how it's measured.

I've hinted that all surfaces exhibit *curvature*, perhaps with the trivial exception of planes, which have zero curvature. Intuitively at least, a sphere is equally curved everywhere. The side of a cylindrical can, though, has curvature in one direction (directly towards the interior) but no curvature parallel to the axis of the cylinder. So curvature at a point on a surface is going to depend on the direction it is measured in. If you imagine a very short straight line drawn on a curved surface at a point then if that line is part of a circle of radius R in the embedding dimension then the curvature is determined by $1/R$ and of course must be labelled with the direction of the line. The smaller the radius, the greater the curvature. In some circumstances $1/R^2$ is used as a measure of curvature.

You can see one problem with this picture of 'curvature'. It needs the idea of an embedding dimension to picture it. How does a bead sliding on a one-dimensional wire so that it only records distances along the wire know if the wire is curved? Any curvature is easily appreciated if the wire is seen in two dimensions as a line on a plane. The bead, though, has no concept of 'plane'. The curvature of a two-dimensional surface like a sphere is readily visualisable if we have a third dimension in which to see it. So we can easily see the difference in curvature properties between a ball, a tin can and a lifebelt visualised as three-dimensional objects, thought the defining surfaces are only two-dimensional. This idea of curvature runs into trouble when the curvature itself is more than two dimensional. Fortunately, it turns out you don't need the idea of an embedding dimension to measure curvature. Mathematicians were onto this issue long before you or I were born. Bernhard Riemann (1826 – 1866) realised what to do but his general discussion is still too advanced for many Honours Physics courses so the details won't feature here. At least we can see what the problem is and know that there is a solution.

We don't need a '5 dimensional space' to talk about a 4-dimensional Minkowski space-time that is curved. That's fortunate. Remember that a flat space-time can be measured with a grid of little unit measuring boxes that are the same everywhere. In curved space-time the metric depends on where and when you are. I should add that very far from matter, curved space-time flattens out into Minkowski space-time. This is mathematics more than physics, for by the time you get far enough from one star for gravitational effects to be negligible you run into the next star.

The fundamental equation of General Relativity is about curvature, so that's why I'm focusing on it. It was gravity that got Einstein developing the concepts of General Relativity but apart from the constant G the concept of gravity doesn't appear in General Relativity. Matter in general introduces curvature into space-time and objects within the curved space-time are acted on by whatever other forces are around, such as electricity and magnetism, **but not gravity**. Gravity is eliminated, or rather its effects are built into the curvature. Clearly curvature is a big issue.

To make the point even more clearly, astronauts in the International Space Station experience weightlessness. Ask Einstein 'why?'. They are weightless, he might say if he were around, because they don't experience a force of gravity. They are travelling at a uniform speed in a straight line (a geodesic) in curved space-time, curved by the mass of the Earth. Gravity doesn't exist. Curved space-time does. Einstein's concept isn't a simplification of the normal basic physics of orbiting bodies – indeed in detail it's a lot more complicated because his 'curvature' in space-time has at least 10 components. Einstein's treatment, though, turns out to be more accurate than the basic physics using Newtonian gravity and Newtonian equations of motion, so much so that even a century on it has passed all the experimental tests that have been put to it.

Minkowski space is just about how things are measured in differently moving frames of reference and nothing interacts with space-time. Apart from ideas like observers and rulers and clocks and particles and such like, Minkowski space is essentially empty, and empty space-time is 'flat'. General Relativity describes how space-time is altered by the presence of matter. You'll gather by now that in curved space-time the metric is not as simple as that in Euclidean space or in Minkowski space. The simplest case of all in General Relativity (apart from empty space) is a point mass at the origin. Karl Schwarzschild obtained the metric from Einstein's equations (without the cosmological constant term Λ) in 1915, before Einstein had written his definitive paper. I was going to quote this but there have been enough equations already. The crucial point for us is that it contains the term $(1-r_s/r)$ where r is the distance from the origin and r_s is the Schwarzschild radius, or the distance to the event horizon. So the metric depends on how far you are from the mass. This is what makes curved space-time curved.

Once you have the metric you can work out the lengths of paths in general and geodesics in particular. It takes come careful maths but out come phenomena like the bending of light round stars, the precession of planetary orbits and the spiralling in of mass very close to a black hole. None of these are correctly predicted by Newtonian gravity.

I don't see how curved space time works

Ask the Newtonian astronaut on the International Space Station why he is 'weightless' and he might reply that there are two forces acting on him and everything else in the station. In his frame of reference one is gravity, directed towards the Earth, and the other is the force due to the acceleration of his frame of reference, which is in the opposite direction. The two exactly balance, leaving him 'weightless'.

In this context, the 'force due to the acceleration of the frame of reference' is called the 'centrifugal force'. Purists have labelled it a 'fictitious force' since it's not one of the four fundamental forces in nature. Believe me, in the accelerated frame of the space station it's just as real as any other force. Indeed it balances gravity. So called 'fictitious forces' are found in all accelerated frames of reference. The Coriolis force is another example.

Ask the Einsteinian astronaut why she is weightless and she may well say that's because she is in free-fall in space, she is in an inertial frame, there are no forces acting on her and she is travelling along a geodesic in space-time, or at least the projection of it into 3D space. What could be simpler?

If 'gravity' as such no longer exists, how do we explain falling off a log, or a wall or a roof or a cliff? I think the answer is 'slightly clumsily', which is why people still talk of Newtonian gravity in everyday life. My description is this.

Stand on a bathroom scales or other weighing machine and you'll read the size of the force acting on your feet that stops you falling through the floor. It's almost 900 Newtons in my case. What else is acting on you? Ignoring the slight effect of atmospheric buoyancy, then nothing else is visible. This force, then, will result in your acceleration relative to the local inertial frame, which as Einstein has argued so well is the frame of a free-falling object. This acceleration is represented by the symbol g. In an accelerated frame of reference there is an apparent force opposite to the direction of acceleration that is proportional to the mass m considered. This force is therefore mg in the downwards direction. If you are at rest, for example on the scales, then these two forces balance, so for me $mg \approx 900$ N.

If I'm balancing on a floating log that rolls over, then the force acting on my feet quickly reduces as the log rolls round while the force of mg due to my acceleration relative to the local inertial frame doesn't reduce so quickly. The net result is a downwards force, pitching me into the water. I think this description is a little clumsy but by replacing the concept of 'the apparent force arising from acceleration relative to the local inertial frame' by the word 'gravity' the explanation becomes much simpler. Talk of the balance of gravity and the upward force of the ground on your feet and every pedestrian in Union Street will understand the explanation. It's therefore not the concept of curved space-time as such that is needed to understand the everyday effects of gravity but the equivalence principle upon which Einstein built General Relativity – the equivalence of gravitational force and inertial force.

Dynamics in general relativity is mainly about changes in frames of reference. Suitable symbols need to be defined to keep track of every dependence and it's easy to get lost in the detail. It's here I bail out. At least I hope I've hinted that what's going on is understandable. Producing the right answer is tricky.

What about the universe as a whole?

Around every comet, asteroid, moon, planet and star there is enough curvature of space-time to allow bodies to orbit, and technically around smaller masses too. But neglecting these localised features, does the universe as a whole exhibit curved space-time? Matter curves space-time but the average density of matter in the observable universe is astonishingly small – a few atoms per cubic metre. There are, though, a lot of cubic metres in the universe! It turns out that the curvature of space as a whole is related to the average density of matter and energy in the universe and that this average can be deduced from an analysis of the microwave background radiation. Broadly speaking, there are 3 main contributions to the matter and energy in the universe: ‘star-stuff’, that’s the atoms in the periodic table, ‘dark matter’ and ‘dark energy’. The microwave background tells us the total and to within measurement accuracy it’s within a percent of the amount that makes space flat. If the density were higher, space would have a positive curvature and eventually collapse in on itself. So on a cosmological scale, space-time is curved but the spatial dimensions are flat.

If the concept of ‘inflation’ is true, then that’s not quite the end of the story. In the inflation scenario, the observable universe is only a very small part of the whole thing. The apparent flatness of space is analogous to the apparent flatness of the Earth as seen from a village. What the space-time geometry of the entire universe might be is unobservable and what is unobservable isn’t science. It’s a pity to end on the unknowable but any philosophy that claims to answer all questions is likely to be deluding itself.

JSR