SPACETIME & GRAVITY: The GENERAL THEORY of RELATIVITY

We now come to one of the most extraordinary developments in the history of science - the picture of gravitation, spacetime, and matter embodied in the General Theory of Relativity (GR). This theory was revolutionary in several different ways when first proposed, and involved a fundamental change how we understand space, time, and fields. It was also almost entirely the work of one person, viz. Albert Einstein. No other scientist since Newton had wrought such a fundamental change in our understanding of the world; and even more than Newton, the way in which Einstein came to his ideas has had a deep and lasting influence on the way in which physics is done. It took Einstein nearly 8 years to find the final and correct form of the theory, after he arrived at his 'Principle of Equivalence'. During this time he changed not only the fundamental ideas of physics, but also how they were expressed; and the whole style of theoretical research, the relationship between theory and experiment, and the role of theory, were transformed by him. Physics has become much more mathematical since then, and it has become commonplace to use purely theoretical arguments, largely divorced from experiment, to advance new ideas.

In what follows little attempt is made to discuss the history of this field. Instead I concentrate on explaining the main ideas in simple language. Part A discusses the new ideas about geometry that were crucial to the theory - ideas about curved space and the way in which spacetime itself became a physical entity, essentially a new kind of field in its own right. In part B we look at how this theory has been tested, and at some of its astrophysical consequences. Finally, in a second document, we look at one of the most spectacular and fundamental predictions of General Relativity, that of 'black holes'. From being a theoretical abberation when they were first conjectured in 1938, these have become a central part of physics - we now know that 'supermassive black holes' play a key role in the evolution of the universe and of galaxies.

In the following, only those topics are discussed that can be understood using classical GR, without quantum mechanics (this of course misses out much of modern astrophysics).

A. SPACETIME as a FIELD: CURVED SPACE

We begin by looking again at the idea of a geometry. Recall that the foundations of geometry were first laid down by the Greeks, in the highly sophisticated framework of the Euclidean axiomatic method. The big outstanding question left over from the Greek period was whether or not one could construct Euclidean geometry from 4 basic axioms, or whether Euclid's famous "5th axiom", or "axiom of parallels", was required (see the notes on Greek mathematics). Then, in the early 19th century, the startling answer was provided by Gauss, Bolyai, and Lobachevsky: that the 5th axiom *was* necessary in order to define Euclidean geometry, *but* that one could also construct alternative geometric systems by dropping the axiom of parallels, and replacing it with some other axiom. These 3-dimensional geometries were extremely counter-intuitive, and not so easy to understand without a sophisticated training in mathematics.

Later work of Riemann made it much clearer what was going on here, and opened the door to a very general understanding of geometry. In what follows I will first explain the general ideas for 2-dimensional geometries, where we can easily visualize what is going on, and then go to higher geometries.

A.1: 2-DIMENSIONAL GEOMETRIES

The key to the approach of Riemann is to realize that one can define a geometry 'from inside' by just looking at how distances, angles, etc., are measured. Thus we can always define a *straight line* in some geometry by one of various means. For example, if we 'look along it', it should look straight (this means that all points on the line will be in the 'same direction', as seen from some point on the line or some point on an extension of it). Alternatively, if we take 2 points, then a straight line can be defined as the line which describes the path of minimum distance between them. And so on. These various definitions are all equivalent. In the same way, a circle can be defined as the set of all points which are at the same distance from some other point (this last point being the 'centre' of the circle, and the distance in question being the 'radius' of the circle). A 'right angle' can be defined as the angle between 2 perpendicular lines A ad B that cross - and here 'perpendicular' means that each of the 2 angles between the 2 lines is the same. We can define angles by taking them to be fractions of a circle, so that, eg., the angle between 2 lines is measured by measuring how far around some circle , centred on the junction between the lines, we have to go to get from one line to the other. One can then measure the area of some geometric figure by using some basic measure of

area (like a tiny square patch) and seeing how many times it fits into the figure.

Riemann's key insight was that a geometry, defined by these measures, can *vary* around some space - but that one can always define the geometry *locally* (ie., in some small neighbourhood of a point P) simply by defining distances for very small (actually infinitesimal) line segments, then angles between them by using very small circles at points, etc.; areas of some shape are then defined by using infinitesimal squares to fill in the shape, volumes by using an infinitesimal volume (ie., a sphere), and so on.

To see what we are talking about, let's first look at Fig. 1, which show various figures in Euclidean geometries.

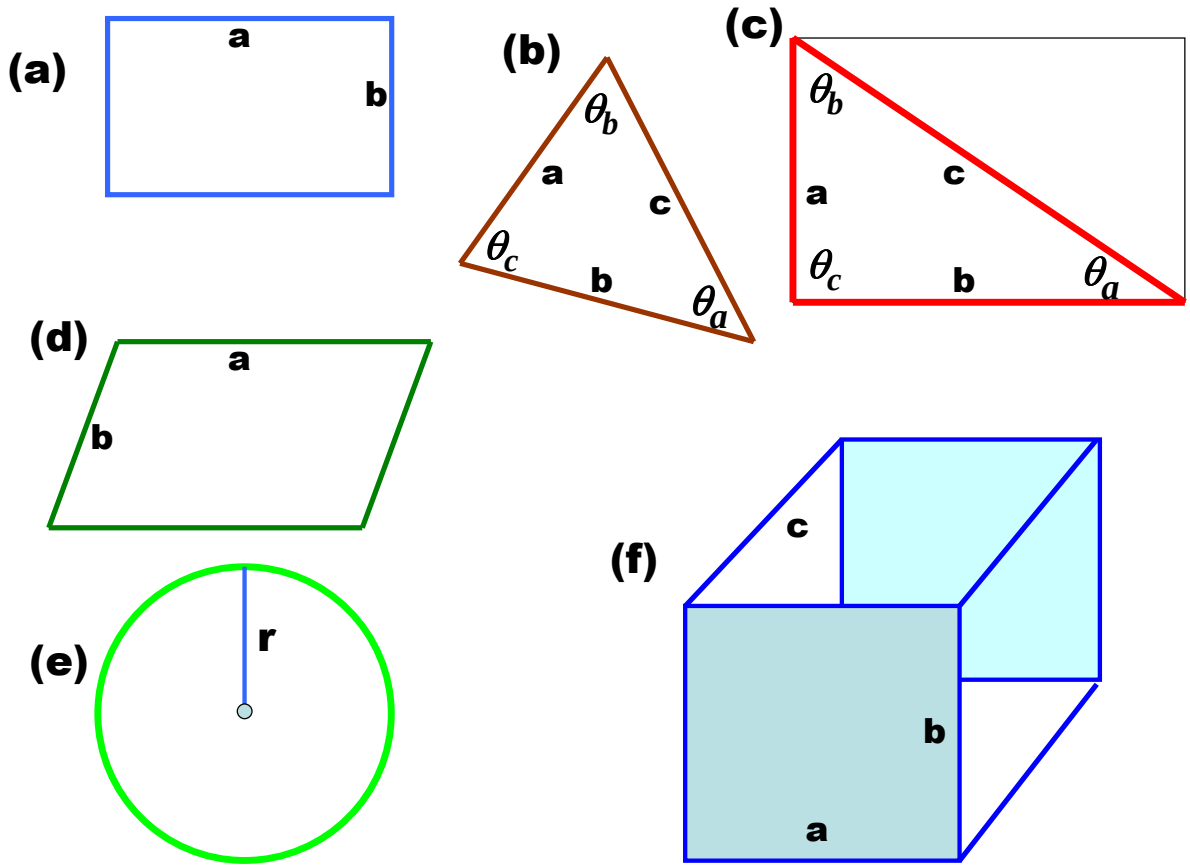

FIG. 1: Various geometric figures in Euclidean space. In (a) we see a 2-d rectangle of sides a and b. In (b) we see a triangle of sides A,b, and c, with angles θ_a, θ_b, and θ_c at the various vertices (ie., corners). In (c) this triangle is a right-angled triangle, ie., $\theta_c = 90^o$. In (d) we have a parallelogram, ie., opposite sides are parallel, but the angles are not right angles. In (e) we have a circle of radius r. Finally, in (f) we have a 3-d rectangular parallelepiped, with sides a,b, and c; this is the 3-d generalization of a rectangle, in which all opposite sides (planes) are parallel (co-planar).

All the objects in this figure are simple 2-d or 3-d objects; since you are familiar with them I won't go on too much, but merely make several remarks:

First, note that all internal angles in the rectangle (a) are right angles, and so the sum of the internal angles is $\Theta = 360^o$. If we look at (d) we see that $\Theta = 360^o$ here as well, even though the internal angles are not right angles (increasing one angle from 90^o decreases its neighbour by the same amount), and in fact this is true of any 4-sided plane figure. Likewise the sum $\Theta = \theta_a + \theta_b + \theta_c$ for the triangle in (b) is $\Theta = 180^o$, as for any triangle, including the right-angled triangle in (c). The length of the sides in this right-angled triangle satisfies the theorem of Pythagoras, ie., we have $a^2 + b^2 = c^2$. Finally, notice that the rectangular parallellepiped, or 'solid rectangle' in (f) is bounded by a set of 6 rectangles.

Now consider the question of *areas/volumes* in these figures. The area A of the rectangle in (a) is just $A = ab$, and the area of the right-triangle in (c) is easily seen from this to be $A = ab/2$ (we get the right triangle by dividing the rectangle into 2 equal parts). The area of the circle is πr^2 (this can be taken as a definition of π). And the volume of the solid rectangle in (f) is just $V = abc$.

Now, says Riemann, suppose that rather than looking at the *finite* objects in Fig. 1, we look instead at *infinitesimal* geometric objects - noting that we can always build up a finite object simply by adding together an enormous number of infinitesimal objects (eg., we can make a circle by adding together a huge number of infinitesimal pieces of it, as we saw with Archimedes). In particular, let's look at how we can define distances and angles on an infinitesimal scale.

We start by considering a 2-dimensional flat plane, into which we project a 2-d 'rectangular coordinate grid', which allows us to define and measure position in the plane (see Fig 2). If we lived in the plane, and wanted to make such a grid, we would simply set up a system of straight lines perpendicular to each other, separated from their neighbours by some well-defined interval of distance. In this flat plane, superposed on the grid, we then inscribe an infinitesimal right triangle having sides dx and dy, and dr. Here I am simply using the standard notation of calculus, according to which an infinitesimal change in some quantity Q is called dQ. The triangle has been set up so that the long side (the hypotenuse), of length $d\mathbf{r}$, takes us from a point P at position $\mathbf{r}$, to another point P' at position $\mathbf{r} + d\mathbf{r}$.

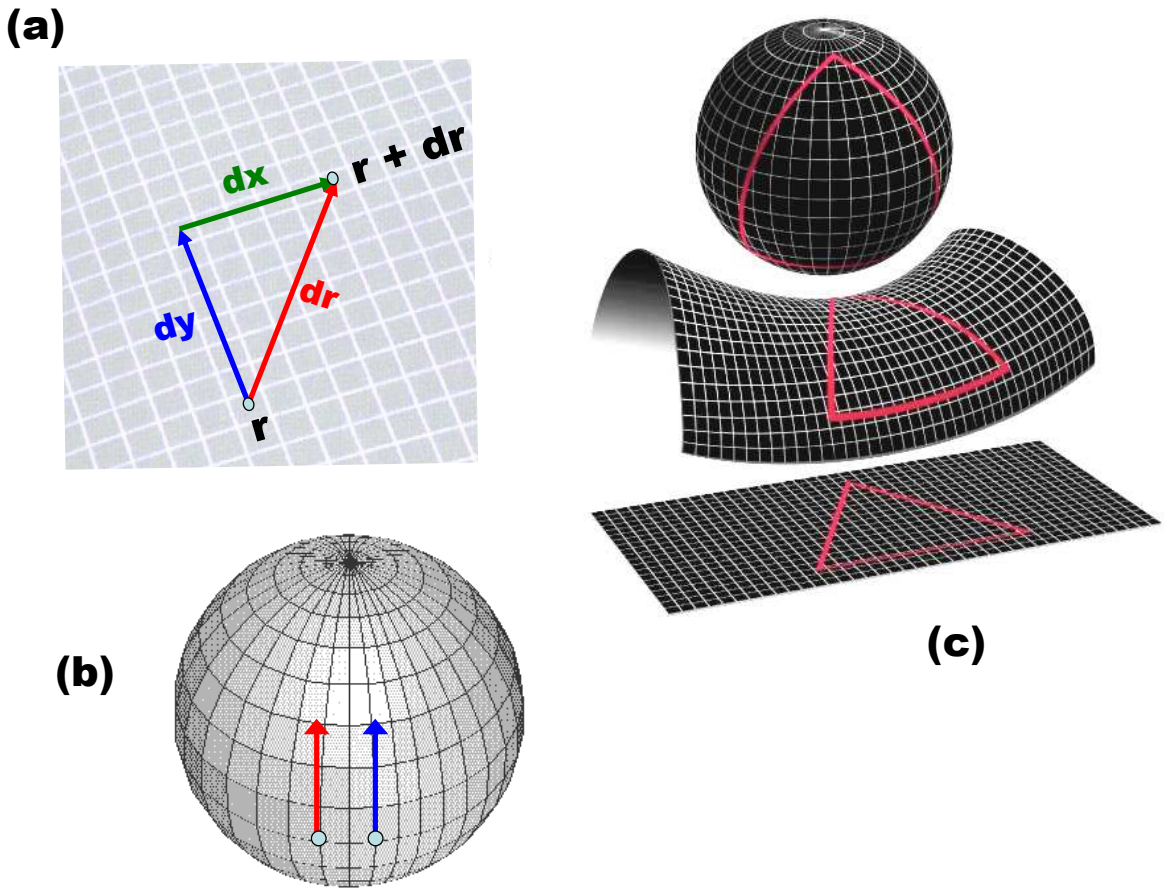

FIG. 2: Mapping out geometries in 2-dimensional spaces. In (a) we see how one can define a 'coordinate grid' in a 2-d flat space (ie., a plane), and define infinitesimal displacements dx and dy along the x and y directions; an arbitrary displacement $d\mathbf{r}$ from some point P to another nearby point P' can always be decomposed into 2 successive displacements dx and dy. In (b)we show how to set up a coordinate grid on the sphere, and in (c) we show 3 kinds of 2-d geometry, with zero curvature (the plane), positive curvature (the sphere) and negative curvature (the 'saddle'), and we see what happens when we draw a triangle in these geometries.

Now of course we know from Pythagoras that the distance $d\mathbf{r}$ is related to dx and dy by

$$dr^2 = dx^2 + dy^2 \qquad (Pythagoras) \tag{0.1}$$

and moreover, since the geometry is flat, we also know that when we make a *finite* triangle (not an infinitesimal one) of sides a and b, then Pythagoras's theorem is also valid for this - we just saw this above (cf. Fig 1).

But suppose we now look at what are clearly *curved* geometries, as in Figs. 2(b) and 2(c). We can still, if we wish, attempt to put down a coordinate grid on these surfaces, but now we will that things start to get more complicated.

Let's first try to set up a coordinate grid inside one of the surfaces - eg., the sphere in (b). We start by picking 2 nearby points, calling the displacement between them the x-axis, or "line of latitude", and then start projecting 2 lines perpendicular to this, along the y-axis (call these "lines of longitude" (if we actually lived in sucy a geometry, we would probably do this by setting up two parallel light beam emitters, both shining along their respective y-axes). If we do this projection only over very small distances, everything seems fine. But as we continue, we find something very disturbing - the two projected lines are no longer parallel, even though they were when we first started! In fact the lines of longitude (or the beams of light) start to converge - and eventually they will meet each other, once we have gone one-quarter way around the sphere. In fact the sphere is considered to be a 2-d geometry having 'positive curvature', so that parallel lines converge towards each other.

Now, try as we might, we cannot fit a rectangular (ie., Euclidean) coordinate grid onto this sphere - this is why cartographers are unable to produce maps of the earth on a flat sheet. Thus, *inside* this geometry, we find by exploring it, that it is not Euclidean. Of course we can always 'embed' this non-Euclidean 2-d geometry into a higher dimensional Euclidean geometry - this is precisely what we do when we make a 2-d spherical surface like a globe in our own 3-d Euclidean spatial world.

If we further explore 'from inside' some of these 2-d geometries, we will find that geometric objects like triangles do not behave as they should. This is shown in Fig. 2(c), where large triangles are drawn, by extending straight lines in different directions from different points until they meet. If we lived in these geometries, we would not notice anything odd except that the internal angles would not sum to 180^o (the sum would be more in sphere, and less in the saddle point geometry). However from outside we immediately see what is going on - the triangle on the sphere 'bulges outward' as seen in 3-d Euclidean space, and a triangle on the 'hyperbolic' saddle surface bulges inwards. The hyperbolic geometry actually has negative curvature - two initially parallel lines will actually start to diverge from each other.

Now a key question here is - how do we quantify or 'measure' the degree of curvature in some geometry, if we inside it? This was a key question addressed by Riemann along with many later mathematicians (notable, at least for their influence on Einstein, were Levi-Civita, Ricci, and Christoffel). The mathematics here is very complicated but we can give a flavour of it. Essentially what Riemann realized was that the curvature could be measured by considering just how far geometrical objects were distorted by the curvature - and that this could be done *entirely from inside the geometry*, ie., without ever having to be outside the geometry! To this it was necessary to define something called the 'intrinsic curvature' (or 'Riemann curvature') and show how the distortion in the shape of different objects depended on it. Some examples are shown in Fig. 3. In Fig. 3(a)-(c) we see 3 different ways in which this can be done. The first, in (a), shows that if we try to make a square in a curved surface (or indeed any rectangle in which all internal angles are right angles) it will not close on itself - how much it misses doing so is a measure of the curvature. Likewise in (b) we can look at a triangle. In this case, also seen in the last Figure, the curvature is measured by seeing how much the sum of the internal angles differs from 180^o. And in (c) we imagine drawing a circle inside the geometry (defined as usual by a line wherein all points are equidistant from the centre); this could be done by simply taking some constant length and moving it around the central point. In this case the area of the circle is known to be given by $A = \pi r^2$, where r s the radius of the circle. However in a curved space we actually find a different result. If the radius of the circle that we measure inside the geometry is called (r_{int} (where the subscript "int" stands for "intrinsic"), then we actually find that

$$\begin{aligned} A &> \pi r_{int}^2 \qquad (\textit{open geometry/negative curvature}) \\ A &< \pi r_{int}^2 \qquad (\textit{closed geometry/positive curvature}) \end{aligned} \tag{0.2}$$

so that if we lived inside a spherical 2d geometry, and measured the area of a circle of radius r_{int}, we would actually find that it was less than it should be.

Two points are of interest here. First, although I have not said anything precise or quantitative about how these deviations (in internal rectangle or triangle angles, or in the area of a circle) depend on the size of the objects concerend, it is fairly easy to understand that they are actually proportional to the size (ie., the area) of the objects, at least when the objects are small. Thus, eg., the deviation in the internal angles of the triangle from 180^o is proportional to the curvature, but it is also proportional to the area of the triangle. Thus, if we look at a very small triangle, the sum of the internal angles is only barely different from 180^o. Another way of saying this is that in the limit as the geometric object becomes infinitesimal, it becomes indistinguishable from the equivalent object in flat space. This should not surprise you. We know that the earth's surface is curved - but on a scale much smaller than the earth, we cannot tell - it looks flat, and ordinary geometry seems fine. But if we tried sending 2 roads parallel to each other along, eg., a north-south direction, separated by, eg., 1 km, we would begin to see after a while that they started to converge towards each other - and they would meet at both the north and south poles of the earth. Likewise, we would find that very large triangles did not look Euclidean, that the area A of very large circles was slightly less than πr^2, where r was their radius, and so on.

The second point is that of course of we could look from *outside* these geometries, embedding them is a higher-dimensional 3-d Euclidean space, we would see things quite differently. For example, we would see that the circle in Fig. 3(c) could be observed from outside to be a perfectly normal circle, with area A and a centre at a point C in the 3d space (marked in red in the Figure). This new "centre" is at a distance r_{ex}, in 3d Euclidean space, from all points on the circle (here the subscript "ex" stands for 'extrinsic"). Moreover we would find that $A = \pi r_{ex}^2$, just as we expect in Euclidean geometry. However we notice that the point C is *not* in the 2d spherical surface - it is in no way part of the 2d geometry at all. Thus, from the point of view of the 2d geometry *per se*, the point C is quite meaningless.

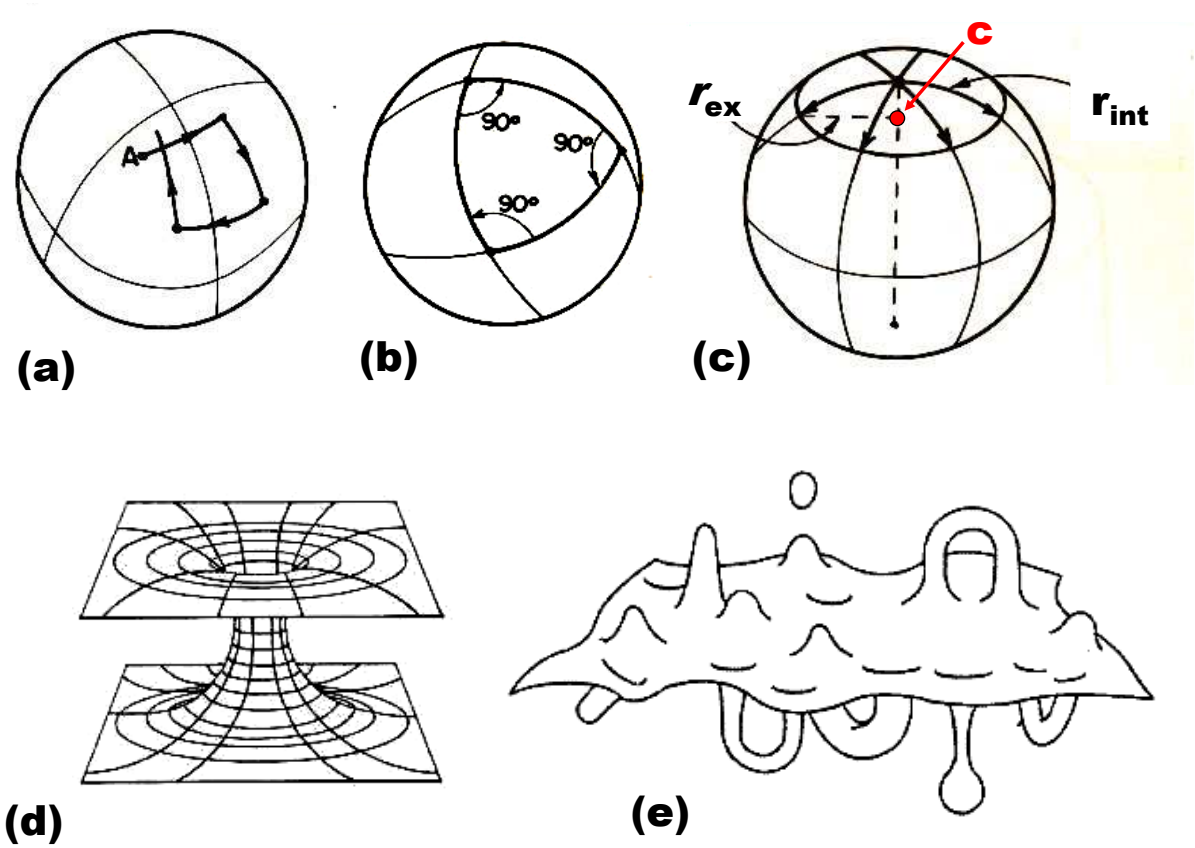

FIG. 3: Different kinds of curved 2-d geometries (which look like surfaces in 3-d). In (a)-(c) we look at spherical geometries; (a) shows what happens when we attempt to make a square, with 90^o internal angels and equal sides, (b) shows that a triangle with straight sides has internal angles summing to more than 180^o, and (c) shows how a circle whose radius r_{int} as seen from inside the geometry is different from the radius r_{ex} as seen in 3-d space, from outside. In (d) we see part of a 2-d geometry with a 'throat' or hole, and in (e) a geometry is shown with holes and even one part (a 'bubble') that is detached from the others.

There are of course lots of other different kinds of 2d geometries that one can imagine - and in 3d space one can visualize them as well. In Figs. 3(d) and 3(e) we show two important examples. Fig. 3(d) shows a region of negative curvature called a 'throat' or 'wormhole' (negative because parallel lines diverge in the throat region). Not only is this interesting in itself, but we see that it can form a kind of 'bridge' between 2 other surfaces (think of a narrow 'neck' formed between 2 parts of a balloon that have been blown up). Then in Fig. 3(e) we show a very complicated geometry full of tunnels and bumps and troughs (as seen from outside); and there is even a closed surface or 'bubble' which is disconnected from the main surface. Now the key here is that from inside this geometry, all one would notice was that the curvature would vary wildly form one region to another, and that it would be 'multiply connected', ie., that one could perform multiple 'loop' trajectories (to see that these existed if one lived inside the geometry, one would need to map out the whole geometry - it would not be possible to see it just by measurements performed locally). If one lived inside the 'bubble' region of this geometry, there would be no possible way of getting to or even knowing about the other larger part of the geometry - unless for some region the 2 parts were able to join (in which case the initial join would be via a wormhole or 'neck' like that shown in (d)).

This is all that I wish to say here about 2d geometries, except to reiterate that the key insight of Riemann was to realize that one had to *define geometries from within*, and that one could in fact do so in a complete way. Of course one gains an extra perspective by looking at them from outside, but the key is that this is *not necessary*.

Now of course the interesting question is - what about geometries in higher dimensions? One can certainly define geometries in an arbitrary number of dimensions - but we cannot visualize them, simply because visualization of some 'shape' for us is a process of embedding that shape in the visual field of our own 3d spatial world. It is of course at this point that we really begin to need the methods and ideas developed by the mathematicians, which do not require visualization. It is interesting, for example, to ask what sort of observations we would have to make inside a 3d geometry to notice that it was not Euclidean. Clearly we would now have to look at 3-dimensional objects, such as the solid rectangle we saw in Fig. 1, or at spheres. If these were small, or infinitesimal, then we would notice nothing - they would look Euclidean. But what kinds of distortion would we see if they were larger, and the 3d space was not Euclidean?

The answer to this question is interesting - what you have to imagine is *what kinds of distortion are possible*. And in fact, you have already seen this in our discussion of EM fields, where we imagined the distortions of an 'aether', in analogy with those of a solid 'jelly'. We can have twist-like distortions, clockwise or counterclockwise around some line, as well as compressions/expansions around a point. Actually it turns out we can also have another kind of distortion, called a 'splay', in which the medium or the space 'spread out' as we move in some direction. Now all of these distortions will affect the geometry of 3-d objects, and it is clear that we would have to make precise measurements of angles, areas, and volumes on these objects in order to map out the 3d geometry (on spheres we would look at how the volume and circumference depended on radius). Obviously this would be more complicated than we saw for 2d geometries, but the general idea is exactly the same.

A.2: 4-D SPACETIME GEOMETRY & GRAVITY

With all the above notions of geometry in mind, we now proceed to the incredible picture that was found by Einstein. The first key point we have to take on board is that we are now dealing with the 4-dimensional geometry of spacetime. Until Einstein and Minkowski found that space and time formed a single 4-dimensional continuum (ie., the Theory of Special Relativity), nobody ever imagined that it would be necessary to go to 4 dimensions (some mathematicians had even speculated, starting with Gauss and Riemann, that 3-d space might be curved, but the idea of spacetime was never contemplated).

Einstein was led to the General Theory of Relativity by a number of different arguments. The historical development was rather tortuous, and here I will drastically simplify it, boiling things down to two key insights:

(i) The Principle of Equivalence: The first key part of Einstein's argument was formulated by him in Dec 1907, and can be understood very simply from Fig. 4(a). We imagine that we are in an elevator which is completely closed, so that the person inside it has no idea what is going on outside. They observe that they seem to have weight - they are pressed against the floor - and moreover if they let go of an apple, it falls, just as it would in a gravitational field. If we aimed a light beam inside at what we thought was a horizontal direction, it would bend downwards slightly - and we could interpret this as the beam 'falling' in the gravitational field. The question is - does this mean that the elevator really is sitting stationary on the ground, in a field? The answer is of course no - it could just as easily be accelerating upwards, as shown in the figure. We would see the apple falling in the same way, we would feel the same weight, and the light beam would bend in the same way. Another possibility is that is that the elevator could be travelling downwards but decelerating - this is the same as an acceleration upwards. Indeed, we could have some combination of gravitational field directed downwards and acceleration upwards.

In the same way, Einstein realized that if the person inside experienced no acceleration at all - so that everything in the elevator seemed weightless - this could mean that it was either (i) just floating or moving uniformly in empty space, with no gravitational fields around at all, or (ii) that the elevator was in free fall (and so accelerating) in a gravitational field (as would happen if, eg., the cables in an earthbound elevator broke). Any attempt to 'probe', from inside the elevator, what was the real situation would not be able to tell the difference - a light beam would behave the same way in both cases (it would be undeflected).

Einstein realized that for the local region of spacetime in the elevator, this apparent equivalence between acceleration and gravitational fields concealed a profound fact about Nature. He argued that the apparent equivalence was in fact real - that from a fundamental point of view these two were absolutely identical, and that they needed to be treated as such in a fundamental theory describing both gravitation and accelerated objects. This he called the "Principle of Equivalence".

Later, Einstein realized that this would only be true for small regions of spacetime - and that if we went to larger regions, we would be able to see the difference between gravitation and acceleration because the gravitational fields would vary in strength and direction over the larger region. Such a situation is shown in Figs. 4(d) and (e), where we show the effects of gravity in a large region near a massive body like the earth. In Fig. 4(d) we imagine a number

of large spherical 'fluid masses' arranged near the massive body; and the effect of the field is to distort their shapes. The reason for this is exactly the same as that causing tides on the earth, and so these distortions are called 'tidal distortions'. If we take one of the 'fluid drops', we notice that the attractive gravitational force on it from the massive body is slightly larger on the side of the droplet nearest the massive body, and so it is pulled more strongly towards the body than is the far side of the droplet. Thus the drop is distorted as shown. The forces acting on different parts of the drop are shown in Fig. 4(c), and in Fig. 4(e) we show these forces for each of the drop shown in Fig. 4(d). Notice that these drops have to be big, much bigger than the elevator - they have to notice the difference in gravitational fields at different points inside themselves. The bigger they get, the more they will be distorted.

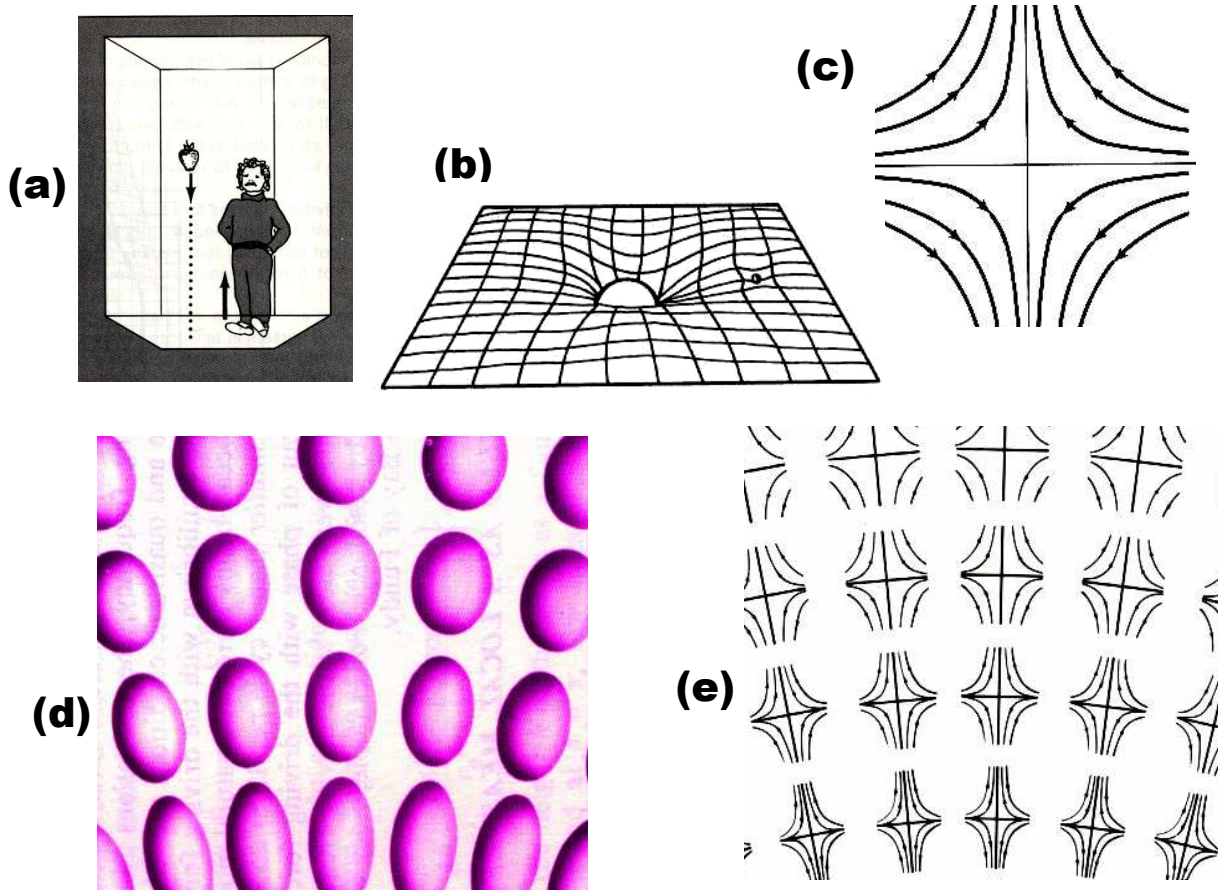

FIG. 4: The local effects of spacetime curvature. In (a) we show the famous 'thought experiment' of Einstein, which he used in 1908 to illustrate the Principle of Equivalence; a person in an elevator cannot tell if the elevator is either accelerating, or is stationary in a gravitational field. In (b) we see a 2-d cross-section of the spacetime curvature near the Earth-moon system, and (c) shows the forces involved in a local 'tidal 'distortion' of spacetime. In (d) we see tidal gravitational distortion, as probed by liquid spheres, near a massive body; and (e) shows the forces acting on these spheres.

Now, let's summarize these observations. In a small region of spacetime (ie., roughly speaking, an 'infinitesimal' region), there is no way to distinguish an accelerated frame of reference from one in a gravitational field, ie., there is no way for any 'local probe' to distinguish acceleration from gravitation - and in particular, a weightless state could either be a result of free fall in a gravitational field, or uniform unaccelerated motion in a completely empty region devoid of gravitational fields. On the other hand, if we consider a larger region of spacetime, we can tell if there is a gravitational field, as opposed to just acceleration, because the *variation* in the field at different points in spacetime will show up.

(ii) Spacetime Curvature and Mass-Energy: Eventually Einstein realized that the way to understand these observations is in precise analogy with what we saw for geometries in section A.1. Recall that we found there that if we are inside a geometry, there is no way to distinguish it from a flat Euclidean geometry if we only make local measurements, at an infinitesimal scale. However if we go to a larger scale, we will see departures from the results of

Euclidean geometry, which become more obvious as we increase the size of the geometric objects we examine. Imagine now that we are in a 4-dimensional spacetime geometry. The observations we have to make inside this geometry, to find out what it is like, are naturally more complicated than those one would make inside a 2d surface - but Einstein realized that one way to distinguish a 'flat' 4d spacetime geometry from a curved one was precisely by looking for tidal effects. These tidal effects are one aspect of the generalization to 4 dimensions that one can make of the 'stretching', 'twisting', and 'splay' distortions that one can have in a 3d medium, or in a 3d geometry. Naturally, because we are now dealing with 4 dimensions, we have to imagine even more complicated kinds of distortion, which will be less intuitive now that we are in a higher number of dimensions than our own - moreover, one of these dimensions is time, so that the distortions will also involve what we see as *motion*.

Nevertheless the point was clear. Gravitation and gravitational fields had to somehow be thought of as a kind of distortion of the 4d spacetime geometry itself - indistinguishable from ordinary acceleration in infinitesimal regions of spacetime, but having quite different effects for larger regions. In other words, gravity had to be somehow linked to the *curvature of spacetime*.

It is primarily because of the complexities of geometry in higher dimensions that Einstein took a few years to finally formulate GR. In the period from roughly Aug 1912 to the publication of a paper in 1913 along with his friend Grossmann, Einstein was engaged in a furious orgy of learning about what is now called 'differential geometry', ie., the mathematics of curved spaces of arbitrary dimensions. From this point until the final success in Nov 1915, he spent enormous effort in trying to use this new 'mathematical technology' to translate his physical ideas into a precise form. The realization by Einstein in 1912 that he needed to learn these new mathematical methods was crucial - as often happens in creative endeavours of many different kinds, new ideas require new techniques for their expression.

Finally, however, success came. This involved 2 key steps, to answer 2 key questions. The first question was a technical one - how to properly describe this curvature mathematically. The second question was more physical, and could be expressed simply as: what was the *source* of this curvature? In answering this latter question, Einstein made the link to the kind of field theory that Maxwell had invented to describe EM fields. Recall that in EM theory, the source of the *distortions* of the underlying EM field that we call electric and magnetic fields is just *charge*, either static or moving. Charge both creates these distortions, and then is acted upon by them (ie., charge feels forces from them). So what was the source of the spacetime curvature (ie., the distortion of spacetime) that we experience as a gravitational field?

The obvious answer to this question is mass - this is what we would guess from Newton's law of gravitation. But Einstein already had his special theory of relativity, and he knew that the proper way to think of mass was as a manifestation of the more general concept of 'stress energy', described by a mathematical object called the 'stress-energy tensor'. Without going into any details, the important thing to grasp here is that the source of spacetime curvature was going to have to be a combination of mass and energy, which we will call here 'mass-energy'. Thus not only could a simple massive object like the earth cause spacetime curvature, but so could, eg., an electric field (which, as a distortion of the EM field, involves energy, which can be used to move charge around).

You might be curious to see the equation that Einstein finally wrote down - the so-called "general field equation of General Relativity". It involves the 'stress-energy tensor', written as $T_{\mu\nu}(x)$, which describes the concentration or density of 'mass-energy' at some point x in the 4d spacetime; and it involves a quantity called $G_{\mu\nu}(x)$, which is related to the spacetime curvature at the same point in spacetime. The equation then takes the simple form

$$G_{\mu\nu}(x) = -\kappa T_{\mu\nu}(x) \tag{0.3}$$

where κ is just a constant (it is in fact $\kappa = 8\pi G/c^4$, where G is the gravitational constant of Newton, and c is the velocity of light). This equation simply says that the the *distortion of spacetime* (the left-hand side) is proportional, at any point in spacetime, to the mass-energy at that point (the right hand side). The reason for the complicated tensor form (each tensor is in fact a collection of 16 different numbers, each of which depends on the position x in spacetime), is simply that geometry is much more complicated in 4 dimensions than it is in two.

Nevertheless we can get the basic idea of all of this without worrying about all these technical complications. To do so it is convenient to compare with the EM field - the comparison gives us a better appreciation of both. Let us do this, point by point:

(i) In both theories, we have "*fields*" (which you can think of as 'aethers'). in electromagnetism, we have the EM field, and in GR, we have the spacetime field. These fields themselves are quite inaccessible to us - we are only able to experience *distortions* of them. These distortions are caused by "sources", or "charges". The source of distortion for an EM field is just electric charge and its motion; the source for the spacetime field is mass-energy. The distortions of the EM field are what we call magnetic and electric fields; the distortions of the spacetime field are what we call gravitational fields. The distortions caused by a source at some point extend over some finite range - thus, the electric field caused by a point charge extend out from the charge, decreasing in strength as one moves away; and likewise the gravitational field extends out from a mass-energy source. And these field distortions have energy associated with them.

(ii) The distortions of the fields, and they energy these distortions carry, cause physical effects - this is how we know about them. In particular, they react back on the sources themselves. In the case of the EM field, the magnetic and electric fields cause forces on charges and currents - these cause the charges to move. The distortion of spacetime that we call the gravitational field causes forces on mass-energy, making the mass-energy move. We can visualize these forces by imagining 'test objects' in the fields - these test objects are themselves just miniature sources, whose effect on the underlying field is too small to appreciably change it. In the case of EM fields, we do this by putting down infinitesimal charges and currents. In the case of the spacetime field, we put down test 'mass-energies'. These latter can take various forms - we have already seen that one sort of test object for gravity is just a small mass, and another is a liquid object which is free to distort in shape; and we will see others.

(iii) Because the distortions of the field, caused by sources, then act back on sources, interactions between 2 different sources are automatically generated - the distortion caused by one source at one point will cause a distortion of the field which spreads out from that source, and which can then affect another source at some distance away - and vice-versa. Thus in any field theory, interactions are generated between sources via the field - the field acts as the invisible medium or 'aether' through which interactions are mediated.

(iv) Fields can carry waves. These arise because any change in a field distortion with time will make the local energy where this change occurs different from elsewhere, and the attempt by the field to re-establish things will cause this change to propagate away at some velocity. This change is most easily accomplished by moving a source/charge - if we do so in an oscillatory fashion, then we get a regular wavelike pattern emitted from the source, which can of course be picked up (ie., 'received') by another source/charge. Thus oscillating electric charges generate EM waves, which will be detected by other charges; and oscillating 'mass-energies' will cause 'gravitational waves' which will be detectable by other mass-energies.

In Fig. 4(b), we see some of this summed up in a single picture. The spacetime distortion in the vicinity of the earth-moon system is shown, in schematic form (ie., as a stretching of a 2d surface), and greatly exaggerated (the field strengths of the earth and moon are actually extremely small). We see how the stronger field of the earth (which has a mass 81 times that of the moon, and a surface gravity 6 times stronger) causes a much larger distortion, and the moon's distortion is just a minor perturbation on this - in this sense, the moon is not a bad 'test object' to probe the earth's field. Moreover, the moon is clearly going to be affected by this field, which is pulling it in towards the earth. So will any other object - we can imagine the path of a light beam being similarly distorted as it moves near the earth and attempts to follow the distorted spacetime field. The analogy with the pictures you have seen of electric fields will be very clear.

There are however differences between the spacetime and EM fields. The most profoundly interesting one is this. In the case of EM fields, we see that electric charge generates electric and magnetic fields. We think of the charge and the field distortion as 2 quite separate entities - and although we create field distortion with a charge, we certainly cannot create charge with a field distortion - the charges are just given to us, as a part of nature. The situation is entirely different with gravitation. This is because the distortion of spacetime caused by mass-energy itself carries energy! And this energy can then create a *further* distortion of spacetime - and so on. Thus we say that not only is mass-energy a source for gravitational fields, but gravitational fields themselves can act as a source for gravitational fields - ie., they can 'create themselves'. Or, to put it another way, spacetime curvature can engender itself. This remarkable fact means that gravity can 'feed on itself' and 'self-amplify'; as we will see, this is what ultimately leads to black holes.

There can be little doubt that Einstein understood, right from the very beginning, how profound was the change he had wrought in our understanding of the universe - although even he was to be surprised by the consequences of his theory within the next few years. However, in the first few decades after Einstein discovered GR, very few physicists paid much professional attention to it. This was partly because it was strange and rather hard to understand, even for professional physicists; but more importantly, because it was apparently irrelevant to most of the physical phenomena of interest to physicists at that time. There was a reason for this - it is simply that gravity is a weak force, and to see any interesting effects in gravity that cannot be understood in simple Newtonian terms requires huge sources of mass-energy, far beyond anything in the immediate vicinity of the earth. Nevertheless, in spite of the apparent irrelevance to any earthly concerns, there were key predictions that came out of the theory, right at the very beginning, and their subsequent experimental verification passed quickly into legend. That this happened was primarily a result of historical accident, because of the way in which these confirmations were announced to the public; they soon made Einstein one of the most famous people in the world (this development is described in section B below).

The effect on Einstein himself, of finally finishing his theory, was initially one of elation, but also relief - it had been a long and hard road, lasting for 8 years. The work had begun when he was in Prague, and had continued throughout a move to Zurich, and then to Berlin; during this time his first marriage finished, he was separated from his children, and, nearer the end of this period, the First World War was being fought (a war which engaged him politically, against the prevailing regime in Germany). In early 1917 he fell seriously ill, in part because of the immense strain

he had undergone - he was not to recover properly for several years. He had managed to pull off what many feel is the greatest single intellectual achievement in human history, almost single-handedly, and unsurprisingly, this left its mark on him both physically and psychologically, in ways that are hard to imagine. As he remarked in 1933:

" *The years of searching in the dark for a truth that one feels but is incapable of expressing, the intense desire and the alternations between feelings of confidence and despair, until finally one breaks through to clarity and understanding; these are only known to him who has himself experienced them*"

A Einstein, in " *The Origins of the General Theory of Relativity*" (Glasgow, 1933)

However, his efforts were soon to be repaid, both in the form of experimental confirmation of the theory, and in terms of public recognition, as we will now see.

B. EXPERIMENTAL TESTS & COSMOLOGICAL IMPLICATIONS

As noted above, the reaction of the scientific community to GR was for a long time very muted, in spite of its successes, which we will be discussing here. However, the disinterest in GR amongst professional physicists and astronomers began to lift in the 1960's, years after Einstein had died (in 1955). This was for two reasons. The first was that observational discoveries, and the theory surrounding them, made it clear that many astrophysical phenomena were quite inexplicable without the use of GR - or else that, like the Big Bang or neutron stars, they were almost inevitable theoretical consequences of GR. The second reason was that a better understanding of the mathematical structure of GR led Penrose to the conclusion that 'singular solutions' of Einstein's equations (later to be called 'black holes') were not some weird pathology of the equations, but rather an unavoidable consequence of them. More recent astrophysical work (both theoretical and observational) has made it clear that black holes play a key role in the internal dynamics and evolution of galaxies, and that they also likely control much of the long-term evolution of the entire universe. Thus now, in the 21st century, GR has become a central feature of astrophysics, an essential tool for understanding the universe.

In this section we will (i) concentrate on the tests of GR, and also discuss one or two other consequences of GR for ordinary astrophysical phenomena; and (ii) look at one of the most surprising consequences of General Relativity, the application of the theory to the entire universe. These are of course not the only areas of astrophysics where GR is important. However a proper discussion of the role of GR in the physics of stars, as well as to the early universe, will have to wait until we have met quantum mechanics. The discussion of black holes, another area where GR really comes into its own, is in the next section (section C).

B.1: TESTS of GENERAL RELATIVITY

There are many different experimental and observational tests of General Relativity that have been performed since 1915. A survey of these is both interesting and informative - it gives us some idea of how the theory works. Tests can be divided into '*weak field*' tests, which look at small departures from Newtonian dynamics when the gravitational fields are not strong, and strong field tests, which are necessarily more difficult because the gravitational fields required can only exist around extreme objects like black holes or neutron stars. we begin with weak field effects.

B.1(a) Precession of Mercury's Orbit: On Nov 18th 1915, just one week short of producing the final equations for his general theory, Einstein announced the results of his calculations of 2 physical phenomena that were linked to his theory. The first of these concerned the orbital motion of the planet Mercury, and the second the deflection of the light of distant stars by the sun. Both of these results were destined to play a key role in the subsequent developments of the subject, and illustrate the enormous importance of precise theoretical predictions in physics. Neither of them required the final equations, and both calculations were fully justified when the final equations were produced a week later (ie.. it was found that the final correct field equations gave the same answer).

It had already been known by then for nearly 60 years (announced in fact by Le Verrier in Sept 1859) that there was an anomaly in Mercury's motion - the eccentric orbit of the planet was not stationary in space, but precessed around the sun more than it was supposed to, by some 43" of arc every century. This situation is depicted in Fig. 5(a), although the rate of extra precession is greatly exaggerated in the Figure (in one century Mercury completes over 400 revolutions around the sun, and yet 47" of arc is less than 1/2000th of a full circle; ie., the angular precession of the orbit is just over one millionth of a full circle each revolution). Nevertheless by this time the measurements of planetary dynamics were so accurate, and the calculations of these dynamics so precise, that this discrepancy was

a real problem. To have such a precession in Newtonian mechanics, one would require some extra force to 'drag' Mercury's orbit in the way - the obvious explanation was another planet. Certainly Venus could do such a thing, and there would be a smaller effects from the Earth and other planets - but all of these effects could be calculated and they were not enough to explain the full precession - hence the discrepancy (in fact one finds, in observations, a total precession of 574.1" $\pm$ 0.4" of arc per century, of which all but 43.1" $\pm$ 0.5" can be accounted for by the gravitational effects of the other planets). Thus there was no obvious explanation for the extra precession, at least in terms of Newtonian theory - it seemed to require either an extra unknown planet to exist closer to the sun than Mercury, or a change of perhaps 10% in the mass of the planet Venus. Neither of these possibilities seemed very plausible.

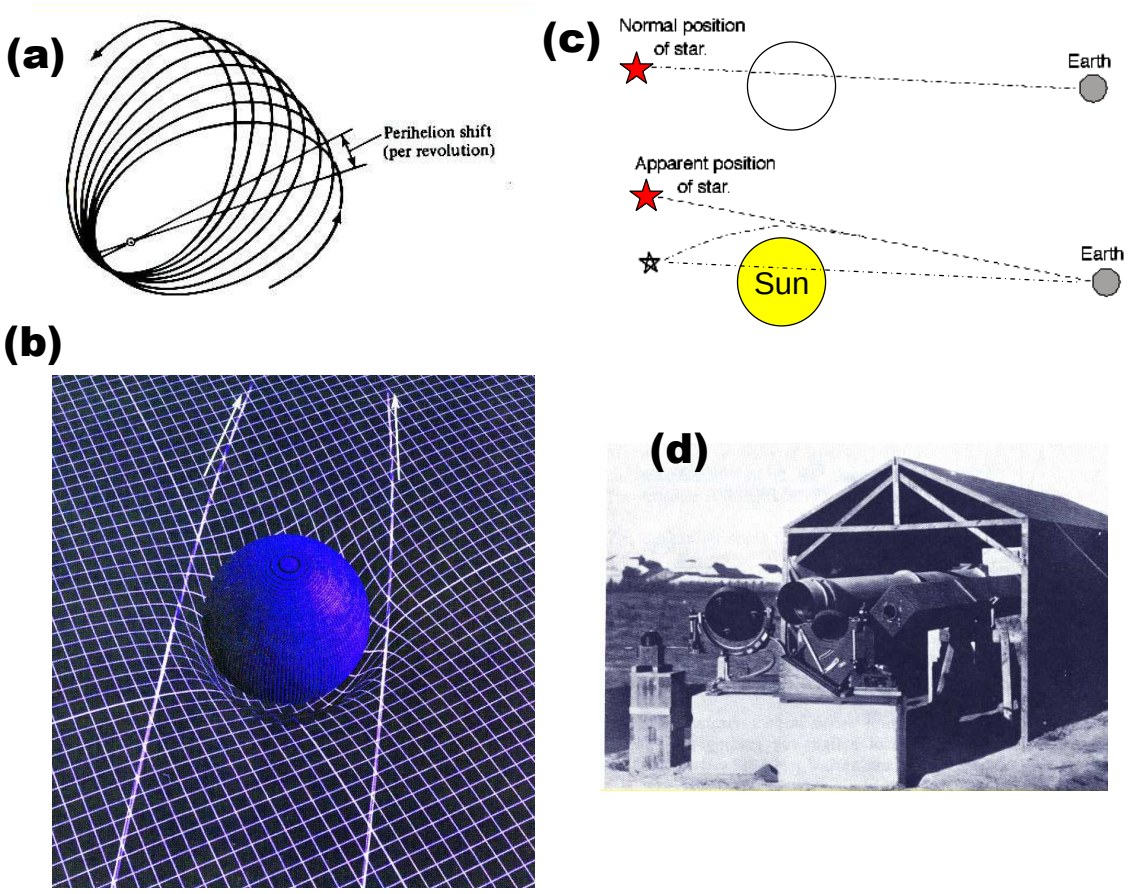

FIG. 5: Early tests of the General Theory of Relativity. In (a) we show schematically the precession of the orbit of Mercury (greatly exaggerated). In (b) the bending of two light beams around the sun is shown (again exaggerated); and in (c) we show how this leads to a shift in the apparent position of a star behind the sun. In (d) one of the telescopes used in the 1919 solar eclipse expedition is shown.

Einstein had already decided that an explanation of this precession could come from his theory of gravitation, well before he had the final theory - and indeed he had even tried to calculate the precession for an early version of the theory (and obtained the wrong answer, ie., an answer which did not agree with astronomical observations). Thus when he performed his calculation in November 1915, he was well aware of how important the answer might be. The result came out to be exactly right (in fact, Einstein calculated an extra precession from GR of 43.03" per century). As Einstein remarked in his paper: "the theory explains quantitatively the secular motion of the orbit of Mercury, discovered by Le Verrier, ...without the need for any special hypothesis." The psychological impact on Einstein was enormous - he felt that he had established a kind of direct contact with Nature, in a way that he would never forget. A few months later, he wrote to Ehrenfest that "for a few days, I was beside myself with joyous excitement"; and apparently the emotion was such that he experienced palpitations of the heart. He later remarked to de Haas that he felt that something had snapped inside him, once he compared his calculated result with the observations.

In fact the precession of orbits turns out to be a general feature of GR. The reason is simple - even though the relativistic effects are very small for a planet like Mercury orbiting around the sun, they result in a very slight

correction to the standard Newtonian force law for gravitational attraction. Recall that in Newtonian theory, there is an attractive gravitational force which varies with the distance r between 2 masses like $F(r) \sim 1/r^2$. However in GR, there is a slight correction to this in weak fields, arising essentially because the energy in the gravitational field also exercises a force on the orbiting planet. It is a simple fact, already realized by Newton, that orbits only form stable simple figures like ellipses or circles if the attractive force goes like $1/r^2$; otherwise they will change with time. In fact, as we will see below, ultimately they will decay with time, because of the emission of gravitational waves.

B.1(b) Gravitational Bending of Light by the Sun: The second calculation of weak field effects that Einstein gave in his Nov 18th 1915 paper was that of the bending of light paths by a large massive object. That this should happen was not surprising - one expects light to "fall" towards a massive object, in a gravitational field, even in Newtonian theory (and indeed, Newton had made just such a prediction). However, as with the orbital motion of planets, everything was in the details, and in particular, in the predicted angle of bending of the light beam. In fact, it is easy to show in Newtonian theory that a light ray passing close the the edge of the sun, as seen from earth, ought to be deflected by the tiny angle of 0.87" of arc (about 1/2500 of the apparent diameter of the sun or moon in the sky, as seen from earth). However, by Nov. 1915 Einstein knew for sure that the angular deflection predicted by GR would be twice this much, ie., 1.74" of arc. The physics of this is illustrated in Fig. 5(b); light rays passing on either side of a massive body feel the strong curvature of spacetime around the body, particularly as the graze the edge of the body, and their pats are deflected.

To test such a prediction experimentally was clearly going to be very hard - and yet the sun was the only really massive body around, so there was no real choice. The problem was not just the smallness of the deflection - far worse was the enormous brightness of the sun. In fact the only hope lay in waiting for a solar eclipse by the moon - by blocking out all but the feeble light from the corona, the eclipse makes it possible to see stars in the daytime, right up to the edge of the sun. Already eclipses were being planned for 1916 and 1917, but the war intervened and made these efforts impossible. Finally, in 1919, an expedition was planned jointly by the British Royal Society and the Royal Astronomical Society, under the stimulus of Sir Arthur Eddington, Britain's foremost astronomer, and Sir Frank Dyson, the Astronomer Royal. Two teams were sent out to intercept the eclipse path of the moon's shadow, which would begin in Brazil and end in equatorial Africa. One of these, under Eddington's supervision, went to the island of Principe, just off the coast of Africa, and one to Sobral in Brazil, under the supervision of A Crommelin from the Greenwich observatory. That the British astronomical community should make such a strenuous effort, shortly after the end of the first World War, was a remarkable proof of both the importance attached to the theory, and to the internationalization of science.

The eclipse took place on May 29, 1919, and both expeditions were successful - the weather at Sobral was perfect, and despite some cloud, the Principe expedition was also able to get results. The basic idea is shown in Fig. 5(c). If we compare the apparent position of a star as seen from earth (given by the direction of the light as it arrives at the earth) we see that the deflection by the sun causes these positions to move slightly *outwards* from the sun (ie., away from the edge), as compared to what one would see if the sun were not there. Thus the key is to make really accurate observations of the position of the stars before the sun arrives, and then to make the same observations during the eclipse - picking stars that are close to the edge of the sun during the eclipse. To make such measurements, the astronomers were helped enormously by a technique that had not been available to, eg., Le Verrier in 1859 when measuring the motion of Mercury - this was the use of photographic plates, which had been invented in the interim. Not only did this allow in principle more accurate measurements, but one could also take multiple and very quick exposures, and then make the careful measurements at one's leisure later on, after the plates had been developed. By this time the techniques of astronomical photography were already quite well developed, and all that was required was to be able to set up a stable observing platform and properly calibrate the instruments before the eclipse arrived (cf. Fig. 5(d)).

Carefully encouraged by Eddington, excitement over the results had reached a fever pitch by the time a public joint meeting of the Royal Society and the Royal Astronomical Society was organized in Nov 6th, 1919, in the meeting hall of the Royal Society (with participants seated under the large portrait of Sir Isaac Newton). The reports of Crommelin and Eddington were given, with Eddington finally announcing that "the observed bending of light was found in Sobral to be 1.98" $\pm$ 0.30", and in Principe to be 1.61" $\pm$ 0.30"; in close concordance with the theoretical predictions of Einstein, as opposed to that which would follow from the principles of Newton". In the subsequent discussion, the president of the Royal Society, Sir William Thomson, remarked that "this is the most important result obtained in connection with the theory of gravitation since Newton's day, and that it was fitting that it should be announced at a meeting of the Society so closely connected with him...."; and that "the audience had just listened to one of the most momentous, if not the most momentous, pronouncements of human thought". Nevertheless, Thomson remarked that "nobody had yet succeeded in stating in clear language what the theory of Einstein really was"; nevertheless he felt that our conceptions of the fabric of the universe must be fundamentally altered".

To understand the public and media reaction that ensued, one must remember that only a few months beforehand,

the "War to end all wars" had just finished, with the annihilation of a large part of the youth of countries like Britain, France, and Germany (as well as substantial numbers from the USA and from Commonwealth countries like Canada, Australia, and New Zealand). Thus, pubic resentment in the UK against Germany was still extremely high. And yet here was the British scientific establishment, apparently willingly overthrowing the iconic status of their greatest scientist (indeed, of the most famous scientist of all time). The very next day, the London Times published an article detailing the proceedings, fairly accurately as it turned out, followed by another article the following day, with descriptions of discussions that had taken place about the expedition results, the previous day, in the British Parliament (one can only imagine the day that an equivalent discussion will ever take place in a Canadian parliament!). Very quickly, interest spread to the international press; curiously, it took some time for the German press to react (although when it did, some of the writings were very congratulatory). Perhaps the most interesting reaction was from the New York Times, which in the following month published nearly a dozen articles and editorials devoted to Einstein and relativity theory, as well as the larger implications of work which was apparently so important and yet well-nigh impossible for most people to understand. It was the New York Times which first spread the idea that "only 12 people in the world were capable of understanding the theory", and which, probably more than any other newspaper, began the process of turning Einstein into a living legend, by emphasizing both the heroic and mysterious qualities of the man and his work, and at the same time his apparently accessible and rather engaging personality (which was far more complex than it seemed at first).

As we will see below the whole idea of gravitational bending of light has recently acquired a new importance, in the phenomenon of gravitational lensing (something quite unobservable without modern telescopic techniques).

B.1(c) Gravitational redshift: Well before he found the final GR field equations, Einstein already knew and had discussed in print the possibility of a test of the theory using a gravitational version of the well-known 'Doppler effect'. Most of you will be familiar with this effect from its auditory version, viz., in the sound coming from moving bodies. For example, if you have ever heard the whistle of a passing train, you will notice that the pitch drops as the train passes - it is higher when the train is approaching than when it is receding. One way to explain this is to say that when the train is approaching, it is catching up with the waves it is emitting, and so 'cramming them together', with a shorter wavelength and higher frequency. On the other hand, when it is receding, the waves are being 'stretched out', and the frequency is lowered. The popular name given to the reduction in frequency when the source of the waves is receding is the "red shift", so called because when it refers to visible light, longer wavelengths are at the red end of the spectrum. Likewise, one refers to a 'blue shifted' signal from an approaching object.

When this idea is applied to the emission of a light wave from a massive object, one finds that the light is also red shifted. A proper discussion of this effect in the GR theory is a little subtle, but the general idea is not too surprising - if a light wave is climbing out a gravitational potential well around, eg., a massive star, we might expect it to lose energy (and lowering the frequency of an EM wave lowers its energy). In fact a proper relativistic discussion has to discuss the 'time dilation' effect occurring in a gravitational field, but we will not go into these details here.

It turned out to be quite hard, in the early days of relativity, to do accurate measurements of this effect, and in fact proper tests had to wait until the 1960's. The obvious way to do tests, by looking at the shift in spectral lines of massive dense stars, was frustrated somewhat by the high temperature of most stars, which blurs the spectral lines. The first really accurate tests were either earthbound (light was allowed to fall down a tower, and extremely accurate measurements were possible using something called the "Mossbauer effect", whose details are unimportant here), or else done by bouncing radar signals off either the surface of Venus, or off reflectors that had been left by robot probes on Mars (in trhese experiments, the redshift involved actually came from the sun's gravitational field). These effects were necessarily very small, but they also agree witht he GR predictions.

However in more recent years it has been possible to see much larger gravitational redshifts, from extremely massive objects in deep space. These come from the 'accretion discs' around black holes, and we will discuss them in section C. The observation of these redshifts does not really test GR (since we are not doing controlled experiments); here the redshift is being used as a tool to probe regions of spacetime close to the event horizon of the black hole.

B.1(d) Gravitational waves and the Binary Pulsar: In the previous sub-section we mentioned that spacetime can support wavelike oscillations called 'gravitational waves', and that these can be excited by moving massive objects (and likewise detected by the same). To this day, gravity waves have yet to be detected by any earthbound measuring system, despite experimental efforts going back over more than 50 years. The basic problem is that any reasonably large waves are being emitted from very far away, so that their amplitude on reaching earth is very small; and then to have any hope of detecting them, a very large and extremely sensitive detector is required.

The general idea behind current gravity wave detector designs is shown in Fig. 6. In Fig. 6(a) we show again the tidal forces that one would see on a sphere suspended above a massive object (stretching it vertically, and compressing it laterally). Now imagine that there is a gravity wave coming towards you (out of the page). It turns out that it will have the same sort of effect on the spacetime as it passes you. Suppose we begin looking, at time $t = 0$, at the spacetime distortion at the very beginning of one cycle of the wave (call it the 'wavecrest'). In Fig. 6(b) we show

the way in which the spacetime distortion will change as this this wave passes. Initially, at time $t = 0$, a sphere will be distorted as shown for $\omega t = 0$ (ie., to the shape of a vertically oriented egg). However, as time passes, the stress pattern of the wave will change, until at a time half-way between the passage of one crest and the passage of the next crest (ie., when we are in the 'trough') the pattern of stresses will have reversed, so the it is being compressed in a vertical direction, and stretched horizontally - this is the situation shown for $\omega t = \pi$ in the figure (NB: in this notation, $\omega = 2\pi f$, where f is the frequency of the wave in oscillations per second; thus $\omega t = \pi$ corresponds to $ft = 1/2$, meaning that half the wave has passed).

From this discussion we see that a gravitational wave is not a compressional 'scalar' wave like a sound wave (where we see a medium alternately compress and expand as a wavecrest and then a trough pass). Nor is it a 'vector' wave like a light wave (where a field oscillates back and forth in some direction, or rotates around and round, as the wave moves past). It is in fact a 'tensor wave', in which we have a pattern of stretching and compressing taking place in the plane perpendicular to the direction the wave is moving.

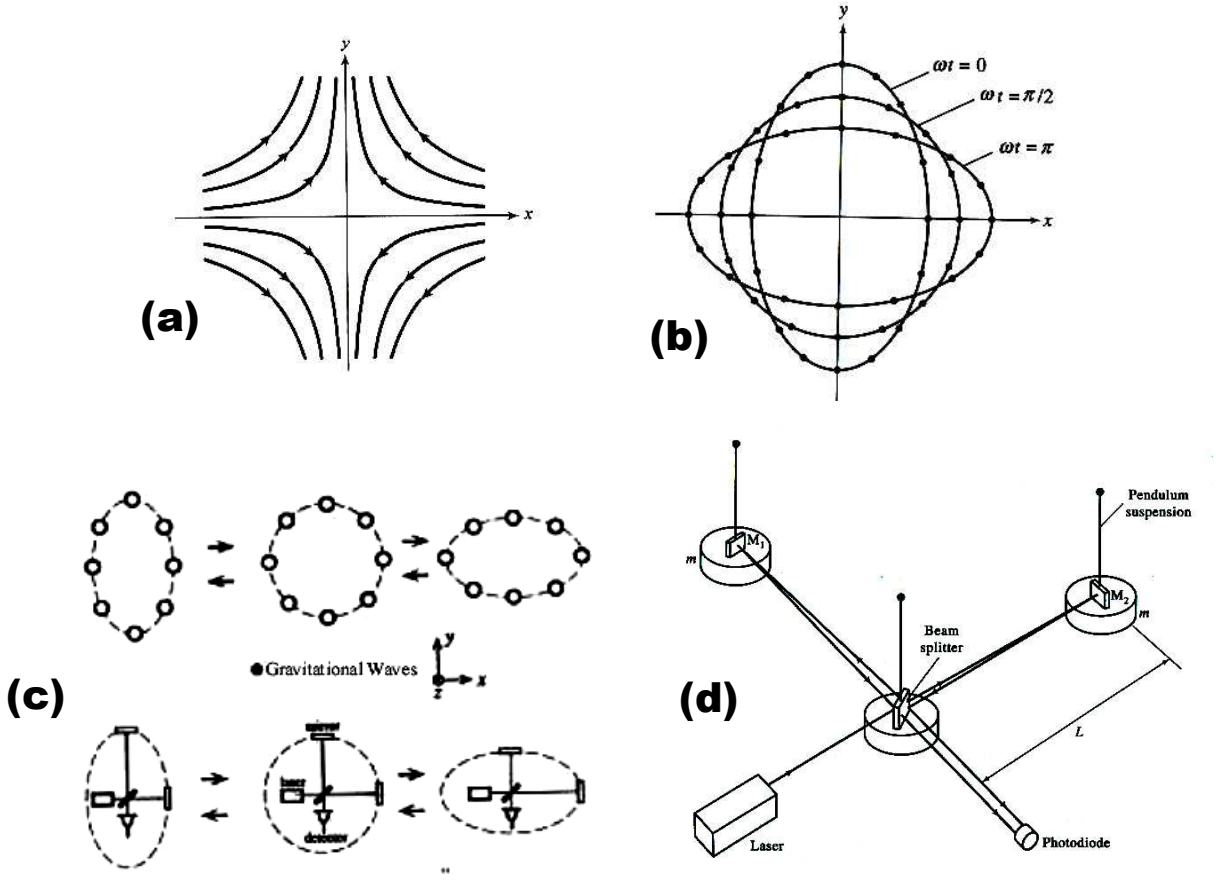

FIG. 6: Gravitational Waves. In (a) we see the direction of tidal forces acting on an object (compare Fig. 4(c); and in (b) we show how a gravitational wave moving past a spherical object will distort it with time: if the initial distortion squashes it into a 'vertical eggshape' (when $\omega\tau = 0$), then a half-period later (when $\omega\tau = \pi$) it will be squashed horizontally. In (c) we see how a ring of detectors will then be distorted by a passing wave coming out of the page, and how one might measure this distortion using laser detectors; the LIGO detector is shown schematically in (d).

In Fig. 6(c) we see how we might actually try and observe the passage of such a tensor wave distortion of spacetime. Imagine, as shown in the top of this figure, a circular ring of detectors all set up in the xy-plane (ie., the plane of the paper), perpendicular to the direction the wave is traveling in (the wave is again traveling straight towards you, out of the plane of the paper). As the wave passes, this ring will be distorted back and forth between the vertical and horizontal elliptical shapes shown. A schematic design for a detector is shown just below this (and in blown-up form in Fig. 6(d)). One imagines a laser set up to emit a light beam moving horizontally to the right - this beam then pases through what is called a "beam splitter", which allows half of the light wave to pass straight through and continue on to a mirror on the right (this mirror is called "M2" in Fig. 6(d)). The other half of the light wave is

reflected vertically upwards by the beam splitter to another mirror (called "M1" in Fig. 6(d)).

The beam-splitter having separated these 2 beams, they are both now reflected back towards the beam-splitter by the 2 mirrors, where they recombine. In fact, the beam coming back from mirror M2 is now reflected downwards by the beam-splitter, and the beam coming back down from mirror M1 simply passes through the beam-splitter, so that they are now reunited and moving downwards towards their final destination, the 'photodiode detector'.

Now the key to this experiment is that light is a wave, so that the final beam moving down towards the detector is produced by just summing the 2 waves that have come back in from the mirrors. Thus we will get wave interference between the two waves when they recombine. The question then is - do these 2 waves add constructively or destructively? This is the key to the whole set-up, because the answer depends on haw far each beam has traveled on its path out from the beam splitter and back. Suppose we set it up so that when the system is in its quiescent 'circular' state, undistorted by any gravity wave, the path lengths are exactly equal. In this case the 2 waves must add constructively, since they have gone through the same distance. However now suppose a gravity wave is passing, so that one path is longer than another. In this case they will not necessarily add constructively - in fact, if one path is longer than another by half a wavelength, they will always be completely out of phase, and add destructively, canceling each other out.

Thus we have a very sensitive detector of spacetime distortions, which can in principle see the passage of a gravitational wave (which will simply cause it to oscillate back and forth between the vertical and horizontal elliptical shapes). The key now is to make this detector as big as possible. The most sophisticated such detector is called the "LIGO" detector (LIGO = Light Interferometer Gravitational-wave Observatory). In LIGO, each of the arms leading out to the mirrors is roughly 4 km long. The light passes along completely evacuated tubes - enormous care has to be taken to get rid of or compensate for unwanted distortions of the tube lengths, caused by, eg., thermal changes, or by tiny perturbations of the earth's crust (faraway earthquakes, vibrations from traffic, etc.). Thus this is a very difficult experiment. And, as noted above, neither it nor any other detector as yet observed any gravitational waves.

And yet, paradoxically, we actually have very good *indirect* evidence for the existence of such waves. This comes from the remarkable 'binary pulsar' PSR B1913+16, discovered and investigated in a beautiful series of observations by the team of Hulse and Taylor, starting in 1974. The binary pulsar is actually a pair of neutron stars, each having a mass 1.4 times that of the sun, orbiting around each other. Neutron stars (which have to be understood using quantum mechanics) are the core remnants of a supernova explosion, in which a massive star explodes catastrophically at the end of its life. Incredibly, in the binary pulsar, we have a system which was formerly a *pair* of massive supergiants, both of which then exploded and yet still remained together afterwards. The neutron star has the density of nuclear matter, some 10^{15} times that of ordinary matter (thus, a piece of neutron star matter the size of a sugar lump would have roughly the mass of Grouse mountain). Thus it is very small - the neutron stars in PSR B1913+16 have a diameter of only 10 km! They emit mostly very high-energy X-rays, but are also enough visible light and radio waves to be seen from earth. The name 'pulsar' comes because the radiation emitted from the neutron star appears to come in regular pulses - this is simply because the neutron star is spinning very rapidly, and we see the radiation being spewed out from the poles passing quickly across us like some giant searchlight.

Now the key to the importance of the binary pulsar is that these two pulsars are orbiting around each other very closely - they are roughly 1.5 million km apart, 100 times closer than the earth is to the sun, and their orbital period is only 7.75 hours. Thus we have 2 massive objects, causing a very significant distortion of spacetime around them, and this distortion is engaged in a fairly rapid (by astrophysical standards) regular oscillatory motion - we expect gravity waves to be emitted by the pair.

Now of course at the distance we are from this pulsar pair, we have little hope of directly detecting these waves. What we can measure, however, is the orbital period of the pair, very accurately indeed (since the pulses are strongly Doppler shifted by their orbital motion, allowing us to see exactly where the stars are in their orbits). Now, remarkably, the emission of gravitational waves by the pair is sufficiently strong that we expect it to suck a measurable amount of energy from the system as time goes by - and the inevitable result of this is that the orbits will slowly decay - very slowly but inexorably, the 2 stars are spiralling in towards each other (and indeed eventually they will fall into each other in a very dramatic collapse). As they do this, the orbital period must decrease - the orbits become smaller, and they move ever faster around each other. And we can see this slow change in the orbital periods - the difference between the number of revolutions actually accomplished, and what we would see if no gravity waves were being emitted - gradually accumulates until it is measurably large.

In Fig. 7 we see how this works. Fig. 7(a) shows, schematically, a snapshot of the pattern of spacetime distortion being emitted from the system as time goes on, with the wavecrest forming an outgoing spiral shape moving at the velocity of light. In Fig. 7(b) we see the pair of stars orbiting each other (with their diameters enormously exaggerated); the 'streamers' moving out from the stars are the tails of radiation emerging along the direction of the poles of the stars (which have very strong magnetic fields, so that radiation is funneled along the magnetic poles). Then in Fig. 7(c) we see the output of the observations - results for the accumulated change in the orbital period over a timescale of 30 years (this is a long-term research project!). The predictions from GR for the accumulated change

in period of this system are shown by the blue curve, and the observational points in red. The great advantage of this system is that the test becomes ever more stringent as time goes on - at the time of writing (2012) the results have verified the predictions of GR to better than 0.25 % accuracy.

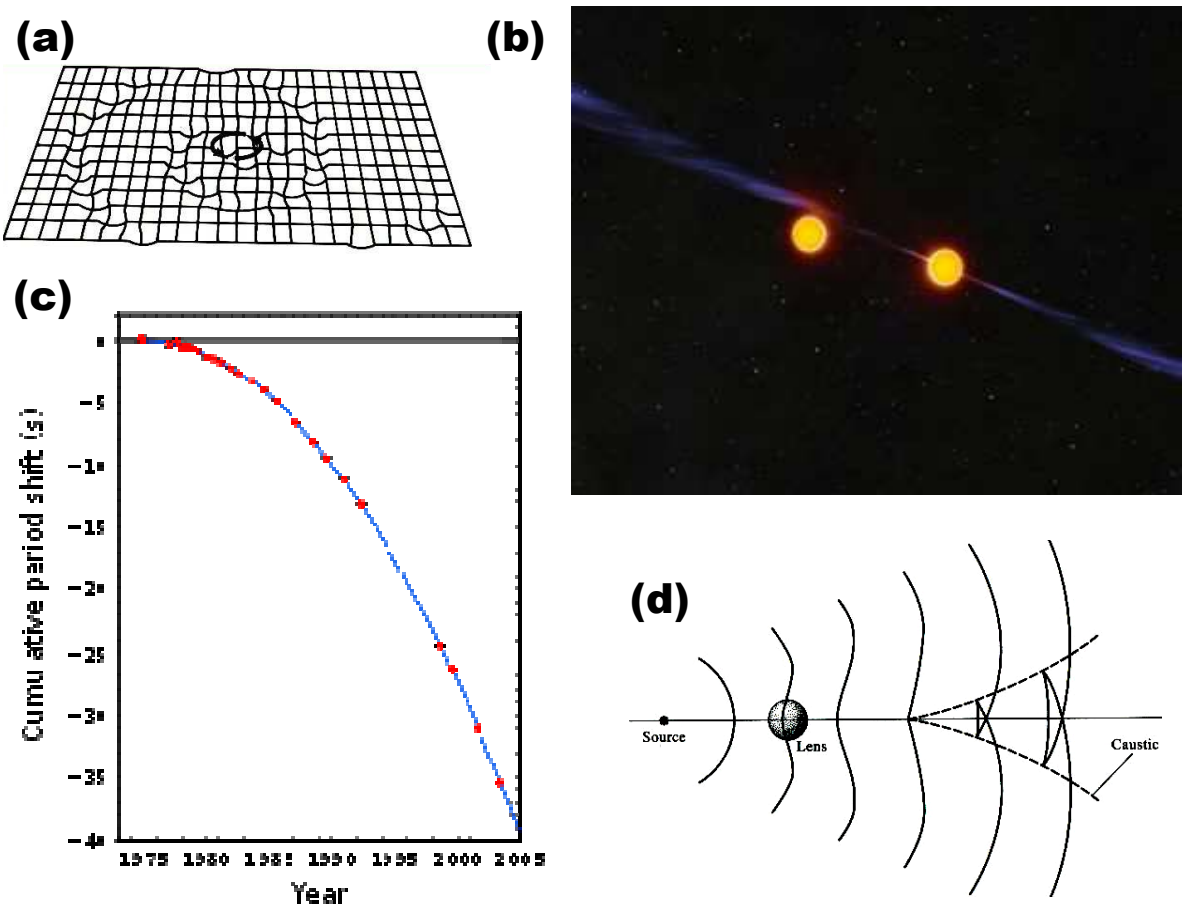

FIG. 7: The binary pulsar PSR B1913+16. In (a) we see a snapshot of the spacetime distortion caused by the orbiting pair of pulsars, which are shown in (b) (with the pulsar sizes greatly exaggerated). In (c) we see the accumulated observational results from 1974-2005; the GR prediction is in blue, and the observations with red points. In (d) the effect of a large mass on a passing gravity wave is shown - the mass attracts the energy in the wave, so that the different parts of the wave start to move towards each other, eventually focussing and forming a caustic pattern.

In recent years more thought has been given to the way in which gravitational waves will interact with a massive body. It was already shown a very long time ago by Penrose that because a gravity wave contains energy, such a wave will eventually 'self-focus' - the different parts of a wavefront will weakly attract each other, and slowly come together. Thus a beam of gravity waves would be unstable - it would slowly collapse inwards as it travels along. Now such a process can be enormously accelerated by having the wave pass a large mass - the mass itself will attract the wave towards itself, behind the wavefront inwards, just as happens with light. Thus a large mass can act as a kind of lens, focussing not only light but also focussing gravity waves. If we treat the gravity wave in a purely optical fashion (ie., not worrying about wave interference, etc.) then we see that eventually all parts of a wide beam must intersect, and form a 'caustic' pattern of intensity (if you have ever looked at the pattern of light at the bottom of a swimming pool you will have seen such a pattern - in this case produced because the irregular pattern of the water surface refracts the light in different directions, bringing it to a focus in some directions). In these caustic regions the gravity wave intensity can be very high indeed. Since the beams are in fact beams of waves, then in reality we must think of them in wave-theoretic terms - indeed, the will diffract around a mass, which will behave a little like a "2-slit" system for the gravity waves, not only focussing them but causing an interference pattern where they focus.

It is clear that in some parts of the universe we may expect fairly high-intensity beams of gravitational radiation, produced by this kind of lensing effect - the consequences of this have yet to be seriously explored.

B.1(e) Gravitational Lensing: As we just saw, a large mass can act as a kind of lens for both gravitational radiation and for light. Curiously, the first person to discuss this theoretically this was Einstein himself, in response to an inquiry he had received in a letter from R Mandl in 1936. The basic idea was quite simple - suppose that between us and some light emitter there is some very massive object, so that light from the more distant object has to pass right by and around (and perhaps even through) the nearer massive object. Clearly lensing will take place, and we will see a focussed image of the distant object (which might otherwise be so faint as to remain invisible).

It is not so easy to find the right alignment so for this sort of thing, and moreover we require really massive objects to do the job. Only in 1979 was a first identification made, this time of the lensed image of a quasar (a quasar is an extremely distant and highly luminous galaxy, powered by a supermassive black hole at its centre). The 'double quasar' QSO 0957 +561 was shown to be in reality a single quasar - the double image being produced by an intervening galaxy, must closer to us but almost invisible because it is much less luminous than the quasar. Other such candidates were soon found - in Fig. 8(a) we see the famous 'Einstein cross' quasar. The 5 images are really all images from a single distant quasar, refocussed by a closer galaxy which in this case is almost exactly in front of the distant quasar.

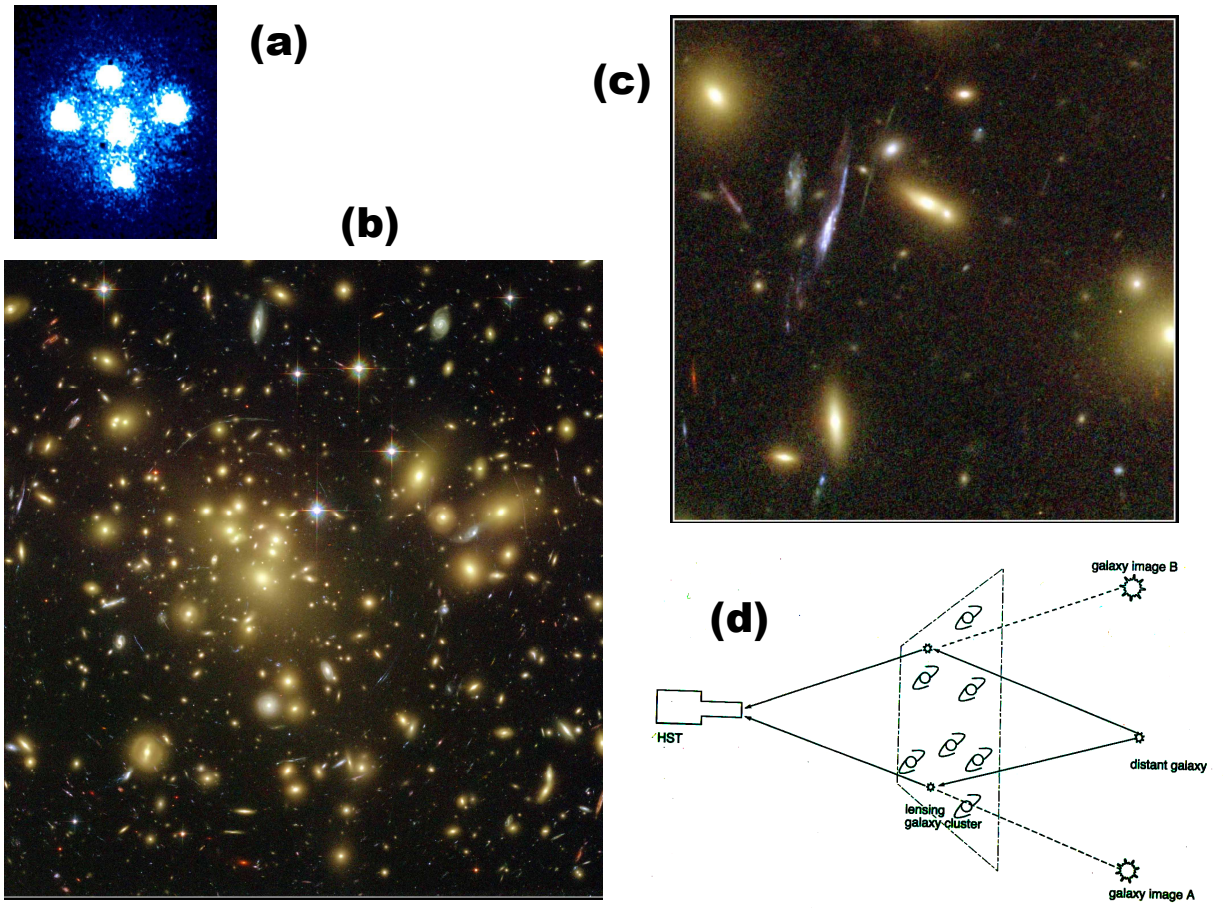

FIG. 8: Gravitational lensing of light. In (a) we see the 'Einstein cross' quasar. In (b) we see the massive (and very distant) Abell 1689 cluster of galaxies. By looking carefully one can just make out streaks of light apparently circulating around this cluster - these are highly focussed 'caustics' coming from the light of more distant galaxies . In (c) a blow-up is shown of a small part of the left-centre of the image in (b); we see several much bluer caustic images of a more distant galaxy crossing this image. In (d) the whole configuration is shown schematically (the light bending caused by the cluster is greatly exaggerated).

Far more dramatic than the Einstein cross are the images shown in Fig. 8(b) and (c). We are looking here at the central regions of a very large cluster of galaxies, the cluster Abell 1689, at a huge distance from us (2.2 billion light years); but the bluish streaks that we see in a kind of circulating pattern around this central region are actually images, in the shape of caustics, of much more distant galaxies. The most distant of these galaxies to be imaged in this way, the galaxy A1689-zD1, is a colossal 13.8 billion light years away, and without the lensing effect, it would be far too faint to be seen by even the most powerful telescopes on earth.

The basic idea of how this works is shown in Fig. 8(d). The cluster of galaxies forms a very large mass - large

galaxies like our own contain nearly 200 billion stars, with an average mass perhaps half that of the sun (the sun is rather an average star). In Abell 1689, there are at least 50 large galaxies, some considerably more massive than out own, and so spacetime is distorted over a large region. The distortion is not strong, because the average density is rather low, but the light from much further galaxies is bent slightly inwards by this distortion, and by the time it gets to earth, another 2.2 billion light years further on, it has been focussed.

It turns out that there is a huge amount of information in images like this. The reason is that the exact shape and position of the distorted images depends in a very precise way on the way in which the spacetime curvature is distributed in the intervening galactic cluster. It is as though one tried to refract light through a rather imperfect lens, full of regions of differing density - instead of getting a clear image of an object at the other side, you would get a peculiar image like the one we see. However, one can use this image to find out exactly what is the distribution of spacetime curvature in the galactic cluster (this involves advanced mathematical techniques, and computers to do the analysis). Since the distribution of curvature is just proportional to the distribution of mass-energy (this is what the equation (0.3) on page 8 is saying), then we can immediately deduce from this what is the distribution of mass-energy in the cluster. And this is where a really very important conclusion emerges - almost all of the spacetime curvature is not being caused by matter as we know it, but by something else, containing 2 components called dark energy and dark matter. Only 4% of the universe is actually in the form of matter, light, etc., ie., in the form of anything we can see - everything else is something else quite different. We will look properly at this in the next sub-section B.2.

We see that gravitational lensing has emerged into a very powerful tool, both to look at things like dark matter, and also for the investigation of the furthest reaches of the universe - the light we are receiving from A1689-zD1 was emitted only 700 million years after the Big Bang, ie., 13.8 billion years ago. Thus we can look very far back in time with lensing observations. Another crucial discovery that this has allowed us to make is that the expansion rate of the universe is slowly increasing. As it turns out, this discovery may have enormous consequences for our understanding of new physics - all this will be discussed immediately below.

B.2: GENERAL RELATIVITY and the UNIVERSE: COSMOLOGY

All of the observations discussed above are of individual phenomena and/or objects which exhibit general relativistic effects, usually quite small. But in fact the most important of all the consequences of GR was quite extraordinary. This was the discovery that GR made predictions for the behaviour of the *entire universe*. Such a situation was unprecedented in physics, and in subsequent years it changed the whole course of science. In what follows I will single out two of the most important and interesting developments, already mentioned above. The first is the idea that the universe is expanding from an initial "Big Bang". This at it turns out was a clear prediction of the theory, and has been confirmed in many different ways, most notably by the discovery of the 'microwave background'. The second is the existence of a completely unexpected source of spacetime curvature, the 'dark' component of the universe, which makes up 96% of the mass-energy of the universe, and yet is completely invisible except by its gravitational effects.

B.2(a) The Expanding Universe

Only a few months after Einstein published the final form of his theory, the German mathematician Karl Schwarzchild, at that time fighting on the Russian front, discovered the first solution to Einstein's equations; and within several years, a number of other solutions had been found, notably by de Sitter. What Einstein had not anticipated was how rapidly mathematical analysis, applied to some new set of equations, can generate all sorts of unexpected consequences, in the form of solutions to these equations - the theory began rapidly to take on a life of its own. In fact the Einstein equations turn out to be a very rich source of different possible solutions, and new ones are still being discovered now, nearly a century later.

The most important and extraordinary feature of these solutions was that many of them described *the universe as a whole*. Never before in science had such a situation presented itself, and this development essentially created the modern science of 'Cosmology'. Some of the predictions coming out from the equations were really extraordinary, and yet one of them - the expansion of the universe - was to be verified already in 1929. In what follows we look first at some of the historical development, and then at more recent developments, which have to some extent turned the subject upside down.

The initial period after Einstein discovered the GR equations was a little confused. In a remarkable *tour de force*, Schwarzchild found a solution to the equations, less than 2 months after they were published, which later turned out to be central to the development of the subject (and to our later understanding of black holes). But Schwarzchild died only a few weeks later, on the Russian front. Einstein was then left, for a short time, holding the ball, and his initial efforts to find solutions were motivated more by physical than mathematical ideas. He was attempting to flesh out an idea first discussed by Mach, now called "Mach's principle", to the effect that gravitational inertia (ie., the need for a force to accelerate a massive object, proportional to the inertial mass) must arise from surrounding matter (so that if the universe was empty, any "test object" put into the universe would have zero inertial mass). This took some time

- it is clear that Einstein was having some trouble dealing with the purely mathematical side of the problem.

However finally, in Feb 1917, he sent in a paper, which we can now see constitutes a huge milestone, for it is the genesis of modern cosmology. Einstein considered a universe having finite volume but no boundary - the 4-dimensional analogues of a 2-dimensional balloon. This of course was, from the point of view of any previous ideas about the universe, a very striking idea indeed, quite inconceivable without the ideas of non-Euclidean geometry we have already covered. Amongst other things, such a finite universe offered a solution to two very old problems. One was 'Olber's paradox - that if the universe was infinite and filled uniformly with stars, then no matter how low the mean density of the stars, the sky would be filled with light of an intensity equal to the typical brightness of the surface of a star (to see this, realize that no matter which direction you look, you will eventually see a star, if the universe is infinite). The second paradox is sometimes called 'Marshall's paradox, although it was clearly understood by Newton - if the universe is infinite and filled with matter at some density, the gravitational force will be infinite everywhere.

Einstein realized that if he was going to get a universe that produced a Mach's principle, as well as solving these paradoxes, he was going to have to change his equations a little bit - in fact he would have to add an extra term. Just so you know what we are talking about here, the revised form of his equations was:

$$G_{\mu\nu}(x) + \Lambda g_{\mu\nu}(x) \;=\; -\kappa T_{\mu\nu}(x) \tag{0.4}$$

so that we have the extra term $\Lambda g_{\mu\nu}$, now known as the 'cosmological term' (the tensor $g_{\mu\nu}(x)$ is known as the "metric tensor"; its precise definition is unimportant). The effect of this term is actually quite simple to understand - it causes an extra "*pressure*" in the universe, a sort of very weak 'anti-gravity'. Einstein's idea was very simple - the attraction mediated by gravity could be counterbalanced by this cosmological term, rendering the universe static; and at the same time the term would actually implement Mach's principle. At first he thought he had succeeded - it seemed that he new equations had no solutions unless the universe contained mass.

However things did not go according to his expectations. As further solutions started to be found (notably by de Sitter and Friedmann, and then later by Lemaitre), both to the original equations and now to his modified equations, it became clear that

(i) Both equations had a whole variety of solutions, describing either closed or open universes, and in which matter was clearly possessed of inertia, even when there was no matter in them (thus contradicting Mach's principle); and

(ii) Much more serious, In GR it was essentially impossible to maintain a static universe - it was unstable, and the general tendency was for it to expand from a point-like singularity. Under some circumstances it would reach a maximum volume, and then recontract - this solution was called an "oscillating" solution, since it was imagined that the universe might then 'rebound', to set off on a further expansion cycle. But the most likely solution seemed to be that of an expanding universe.

Thus Einstein was left with a rather mystifying situation - the universe did not want to remain static, in GR, but wanted to expand! Since the introduction of the cosmological constant did not help either in stabilizing a static universe, or in giving flesh and blood to Mach's principle, he later abandoned it (indeed, he is reputed to have called it "the worst mistake of my scientific life").

It seemed at this point as though one had a large number of rather confusing solutions, none of which was apparently satisfactory. However within the next few years, a very striking result emerged, this time on the observational front, which changed the nature of the discussion completely.

Beginning in 1917, with the first observations made by the new 100-inch Hooker telescope on the Mount Wilson observatory above Los Angeles, and using newly-developed methods of astronomical photography, astronomers suddenly found themselves with a new tool able to probe regions of space well beyond what had previously been possible. Success was not too long in coming. In 1924 Hubble announced that he had definite proof that the M31 nebula (the 'Andromeda nebula') lay far outside the Milky Way, and that it should be seen as an external galaxy. That it should have taken so long was remarkable - already 120 years earlier William Herschel had argued for the existence of external galaxies (which he called 'island universes), on the basis of observations he had made with considerably smaller telescopes, using only his eyes; and many astronomers by 1924 accepted as a matter of course that the 'spiral nebulae' were external galaxies. But Hubble was able to take long-exposure photos, and thereby resolve individual bright supergiant stars in M31 (which we now know to be at a distance of 2.4 million light years from the earth). Amongst these stars were 'Cepheid variables', which oscillate in brightness in a very well-understood way, allowing us to fix their distance by measuring their apparent brightness.

Having now established a way of measuring the distances of the closest galaxies, Hubble then went on to look at very large numbers of them, establishing a classification for them on the basis of their shapes, brightnesses, and sizes. Finally, in 1928-29, he and Humason made the crucial step of looking out to the furthest reaches of the universe then visible, and taking the spectra of these galaxies. Now, as we have seen already in the discussion of EM waves, the spectrum of a body emitting light will be shifted in frequency if it is moving with respect to us - towards the blue if it is moving towards us, and into the red if it is moving away. This is the famous 'Doppler effect', and V. Slipher

had already noticed back in 1912 that almost all galaxies seemed to be receding from the earth. What Hubble and Humason found was that not only were the galaxies were moving away from us, but that the further away they were, the faster they were receding. In fact, it seemed as though the velocity of recession was proportional to the distance, a very striking result.

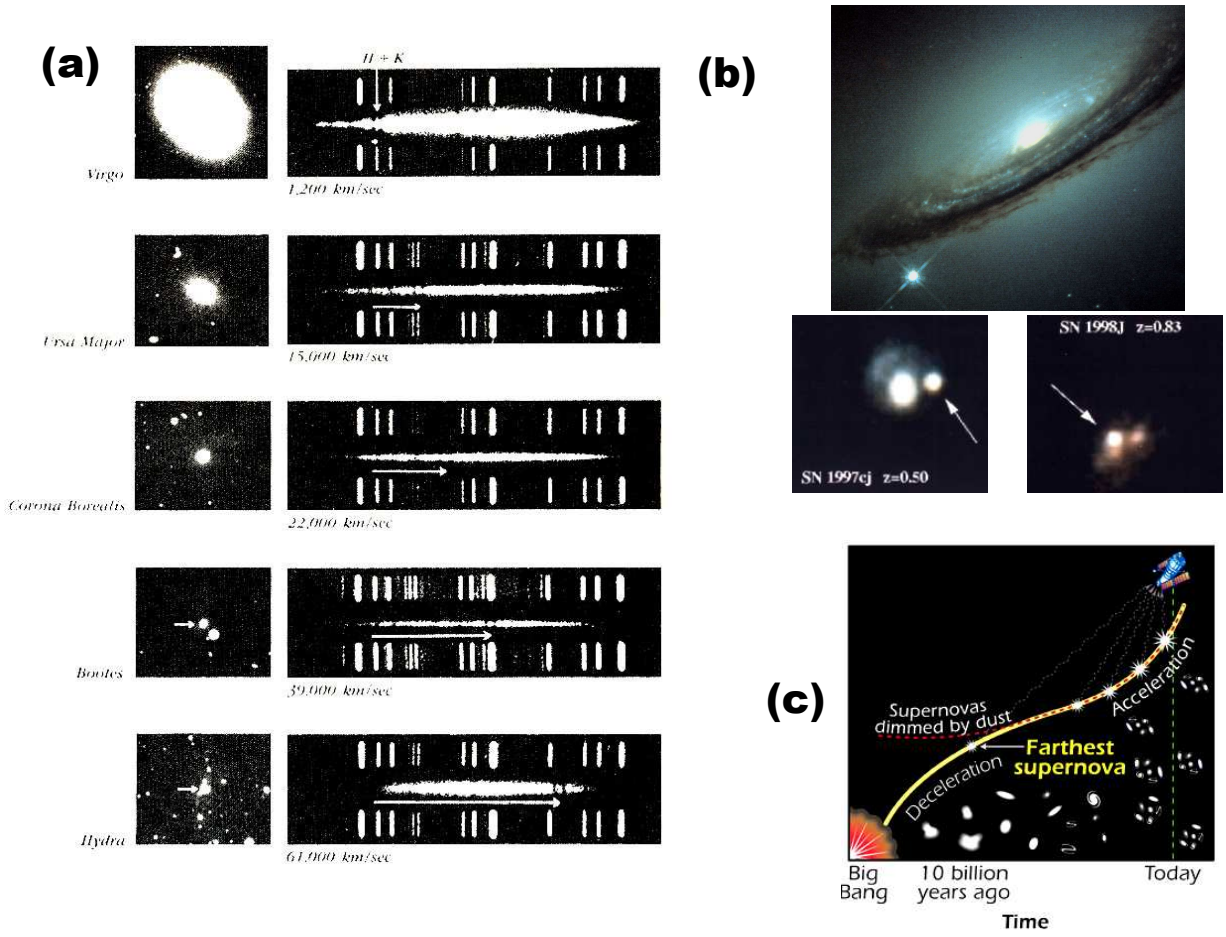

FIG. 9: The expanding universe, as seen from earth. In (a) is shown a selection of photos taken by Hubble and Humason of galaxies in different clusters, along with the optical spectra of these galaxies - the arrow below each one denotes the red shift of a particular double absorption line. In (b) we show at the top a supernova, supernova 1994D (the bright star at lower left) in the outskirts of the galaxy NGC 4526; below this we show 2 supernovae occurring in extremely distant galaxies.

If we look at Fig. 9(a) we can see what Hubble saw in his observations. The photographs show at left, galaxies situated in various clusters at different distances - Hubble could estimate how far away they were simply by using his classification system for galaxies, even though Cepheids were no longer visible at these distances. Then, on the right of Fig. 9(a) we see the spectra of these galaxies, with the velocity of recession shown immediately underneath. You can see that in each spectrum a particular double absorption line is shown by an arrow, shifted ever further to the right (ie., red-shifted) as one goes further away. The furthest of these galaxies, in the Hydra cluster, is receding at 61,000 km/sec from us, ie., at just over 20 % of the velocity of light.

This result had an enormous impact on the astronomy community at that time. First of all, the result was quite incredible - to imagine that everything in the universe was receding from us at these incredible speeds was quite astonishing, and the only possible explanation was that the universe was expanding uniformly. Then, of course, it was quickly realized that this is exactly what GR theory had been saying should happen. Thus the theory of General Relativity had now to be taken very seriously indeed, as a description of the whole universe, which of course now had to be thought of in non-Euclidean terms.

At this point one might have imagined that the idea of the Big Bang would have taken hold rather quickly. Indeed, remarkably, in 1927 the Belgian cosmology theorist (and Catholic priest) Georges Lemaitre had already predicted the expansion of the universe and the redshift phenomenon seen by Hubble and Humason, arguing from solutions to the

GR equations found by Friedmann and himself. Then in 1931, Lemaitre went the whole way and described what we now call the "Big Bang scenario", arguing that the universe had begun as a tiny and incredibly dense 'primeval atom'.

However, interestingly, the Big Bang scenario would not take hold for a long time. As mentioned before, few people worked in this field, and although a number of extremely talented theorists worked on the topic of cosmology (Einstein, Tolman, Gamow, Hoyle, Bondi, Lifshitz, Khalatnikov, etc.) in the long period between 1931 and the late 1950's, not much obvious progress was made. The main reason for this, apart from the war, was a lack of new observations, and an apparent lack of relevance of this field to more 'down to earth' questions. Another reason was that neither Einstein nor many of the other theorists involved really took the idea of a 'spacetime singularity' seriously - indeed, Einstein simply rejected this consequence of his equations as unphysical.

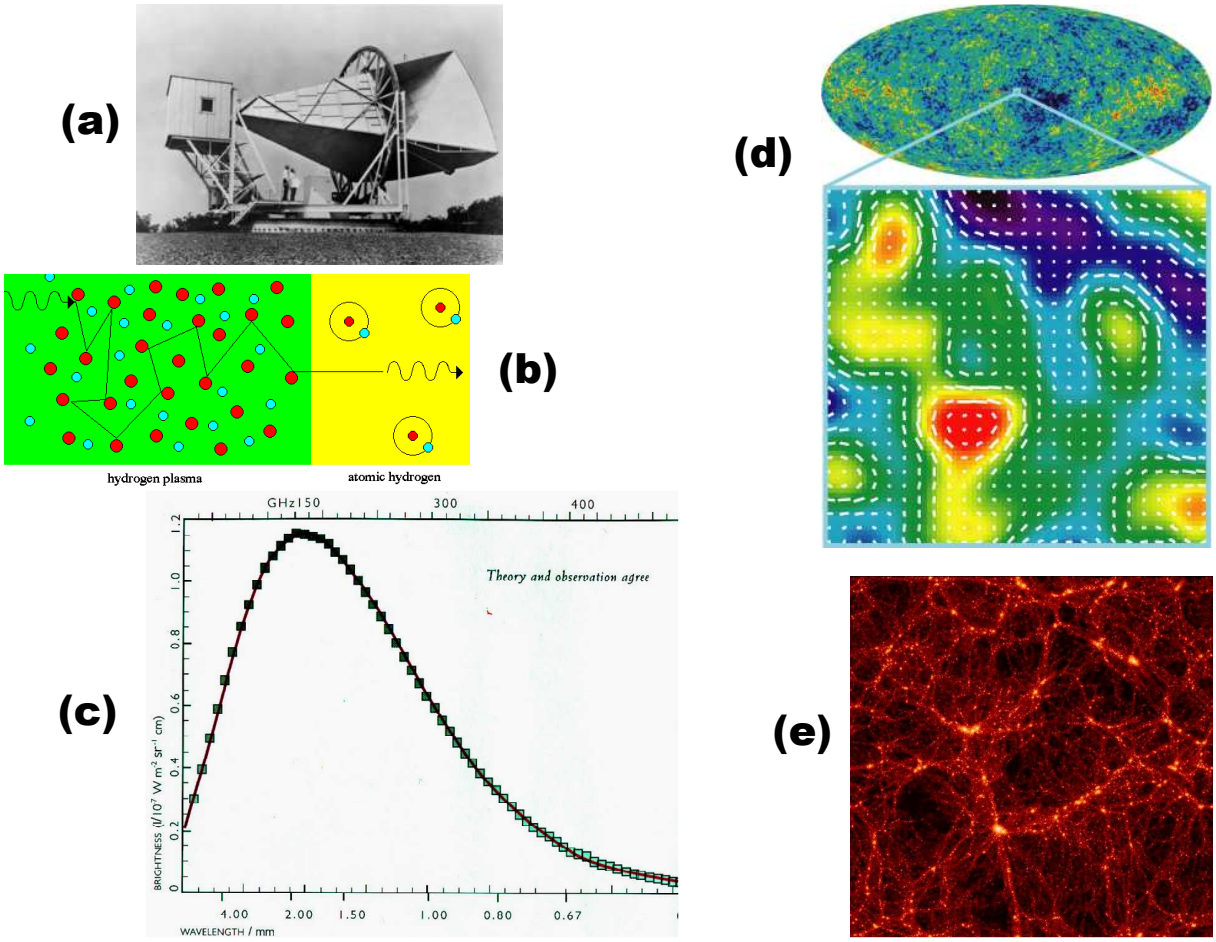

FIG. 10: The microwave background. In (a) we see the riginal microwave detector of Penzias and Wilson. In (b) the scenario for the origin of this radiation is shown - once the Big Bang has cooled to roughly 3000 K, protons and electrons combine to form H atoms, and the radiation in the universe is free to propagate over enormous distances. In (c) the observed intensity spectrum of the microwave background radiation is shown, along with the theoretical prediction; and in (d) we see WMAP measurements of the very small directional non-uniformities in this radiation. In (e) the current distribution of mass-energy in a large section of the universe is shown.

Finally, however, things began to change. One reason for this was that the invention of radio telescopes, and their perfection during the war, gave astronomers a new tool which allowed them to look even further than before (and in 1948 the 200-inch Hale optical telescope on Mount Palomar also came on line). By the late 1950's it was becoming clear that Hubble's law continued to be obeyed to very great distances indeed - one could observe galaxies several billions of years old, receding at large fractions of the velocity of light. The obvious implication was that the expansion had been going on for a very long time, and so a Big Bang scenario now had to be seriously contemplated. A second development was purely theoretical, and originated in highly mathematical studies by R Penrose, beginning int he early 1960's - these eventually resulted in the 'singularity theorems' to be described in section C, which showed that not only could singularities exist in GR, but that they more or less *had to*.

But by far the most important development came with the publication in 1965 of experimental results by Penzias

and Wilson, two engineers working for the Bell telephone laboratories in New Jersey. They had been perfecting a new kind of microwave receiver, shown in Fig. 10(a) (at that time the use of microwaves was in its infancy). They had noticed that there was a noise source they simply could not get rid of, even after cleaning the antenna of all possible extraneous sources (including pigeon droppings on the main antenna horn). Finally they realised, from the way in which the noise varied with the rotation of the earth, that the noise was actually coming from the sky, and that its intensity depended on the wavelength of the microwaves in a very characteristic way (see Fig. 10(c)). The intensity pattern was in fact that of a 'black body', ie., a homogenous object in thermal equilibrium at a particular temperature, radiating energy thermally over a range of frequencies/wavelengths.

By great luck the theoretical group of Dicke and Wilkinson, at Princeton university only a few miles from Bell Labs, was in 1964-65 actually working on the idea of observational tests for the Big Bang, and were quickly able to explain to Penzias and Wilson what they had actually seen. Dicke and Wilkinson were actually working on an idea first cooked up by Gamow in 1945, which had been considerably elaborated by Zeldovich and co-workers by the early 1960's; it is illustrated in Fig. 10(b). In the first stages of the Big Bang the universe was so hot that it could only exist in a kind of hot expanding soup of radiation and elementary particles like electrons, protons, etc. However as it expanded it cooled, and eventually the protons and electrons were able to combine to form stable Hydrogen (H) atoms. Now this caused a crucial change - for while EM radiation finds it very difficult to move through a soup of charged particles (as we already saw, EM waves are scattered or absorbed/emitted very easily by charges), it has no trouble moving through a gas of neutral particles. Thus, very suddenly, the radiation found itself free to move over very large distances. We now know that this happened roughly 360,000 years after the Big Bang, when the temperature had dropped to roughly 3,000 K (ie., roughly 2,700 C). In fact, almost all of the radiation released at this time has been propagating freely ever since throughout the universe - so that we in affect have a 'photograph' of this radiation, and what it was emitted from, from a time only 360,000 years after the Big Bang. The radiation has been 'stretched' by the subsequent expansion of the universe, increasing its wavelength and lowering its energy, so now its temperature has fallen to only 2.7 K, ie., only 2.7^o above absolute zero.

This observation was crucial - it was the confirmation of yet another key theoretical prediction, and in fact the microwave background has been giving us important insights ever since. In Fig. 10(d) we see results from the WMAP space probe, which show the non-uniformities in the microwave background around the sky (the whole sky being shown in the top part of this figure, and a blow-up in the lower figure, colour-coded for intensity, and also showing the polarization of the radiation). The non-uniformity is greatly exaggerated - it amounts to no more than about a fraction 10^{-5} (ie., $1/100,000$, or a thousand of a percent). However it is crucial, because it these tiny inhomogeneities that then grew, under their own self-attraction, and over the subsequent few hundred million years after the microwave radiation was released, into giant concentrations of mass-energy. These huge masses continued to grow, attracting and gradually 'eating each other', and eventually became the first galaxies. The formation and subsequent evolution of these initial galaxies is a matter for current research - as they continued to absorb each other in colossal collisions, enormous shock waves formed which then collapsed into the first stars - less than 500 million years after the Big Bang, the galaxies lit up with billions of stars. The early universe was a violent place - most of the first galaxies (and there would have been trillions of them) were destined to be swallowed by others, in a deadly game of collisions in which the larger galaxies triumphed, and continued to grow. The final result is shown in Fig. 10(e), where we see that our universe is at present a very inhomogeneous place - mass-energy is organized in giant streamers and filaments, joining together vast 'superclusters' of galaxies, with huge voids in between. Our own 'Local Group' of galaxies, dominated by the Milky Way and by M31, but containing another 30-40 galaxies (most of which are 'dwarf galaxies), is but a minor part of the 'Virgo supercluster, whose centre is 50 million light years away from us. Some of the galaxies in the central region of the Virgo cluster are really vast - the single monster galaxy M87 has a mass of some 2.7 trillion suns, and is thus some 6-7 times as massive as the entire Local Group.

For a long time it was thought that the expansion of the universe had been proceeding more or less uniformly with time since the Big Bang itself, so that for some reason everything seemed to be very finely balanced between an 'open universe', in which the expansion continued forever, and a closed universe, in which it would eventually recollapse. For many years, attempts had continued to pin down which of these 2 alternatives was the correct one (or whether, for some bizarre and unknown region, the balance was exact). From these measurements an age for the universe has gradually emerged - we now believe that the Big Bang took place some 13.8 billion years ago.

However, in the last 15 years a completely new twist has emerged. This is the story of what some have called 'quintessence' (this name is derived from the Roman term - *quinta essentia* - for Aristotle's 'fifth element' or 'cosmic aether'). In their search for a way of measuring the distances of the most remote galaxies we can see (either directly or by gravitational lensing), astronomers hit on the idea of using supernovae in these galaxies as 'standard candles' (ie., standards of brightness). It so happens that there is one kind of supernova, called a 'Type Ia supernova', for which one expects the same total luminosity for all members of this type. Supernovae are extremely rare, but a large galaxy like our own will have one roughly once every 20 years. They are so incredibly luminous (shining with a brilliance up to 15 billion suns) that they can be seen as far away as an entire galaxy. Thus we can use them to measure the

distance of any galaxy in which we see one exploding - see Fig. 9(e). The top photo in Fig. 9(e) shows a supernova in NGC 4526. This is a galaxy near the central regions of the Virgo supercluster, and is 55 million light years away - it can be seen in a medium-sized telescope. Notice the brilliance of the supernova - its luminosity is perhaps 10% of that of the central region (the hub) of the entire galaxy. Then, the 2 lower photos in Fig. 9(e) show 2 galaxies whose distances are some 5 billion and 8 billion light years respectively, at the moment when a supernova in each galaxy is at peak luminosity.

Using measurements of supernovae in some 40 different galaxies at large distances, 2 groups came in 1998 to a startling conclusion - the expansion of the universe has not only not been uniform since the Big Bang, it has actually gone through 2 periods - one of deceleration in the first 5-7 billion years, followed since then by an *acceleration* in the expansion. Thus, not only does it look as though the universe may be open, but it is getting 'more open' as time goes on. The question is of course - what is *causing* this acceleration? At the moment we simply do not know - the behaviour looks roughly what one would expect if there was a non-zero cosmological constant, but the reason for this is a mystery.

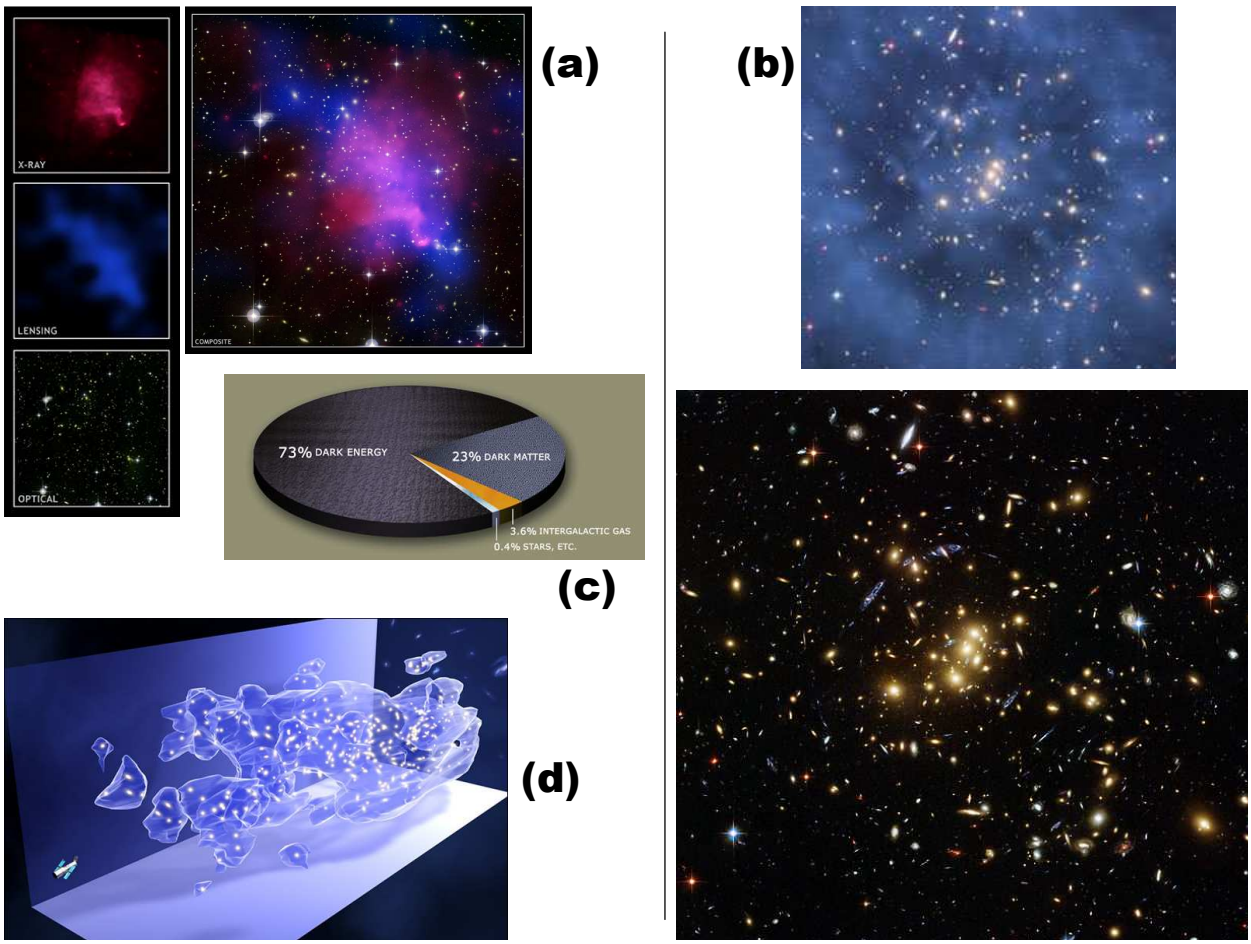

FIG. 11: Dark Matter in the Universe. In (a) we see images of the Abell 520 cluster; at left these are analyzed into X-ray, gravitational lensing, and optical components (see left), with a composite of all three shown on the right. In (b) we see images of the CL0024+17 cluster, with the gravitational lensing contribution at the top, and the optical component below. In (c) we show the contributions of the various different components to the total mass-energy of the universe; and in (d) we see a density map of the dark component to the mass of the universe out to a distance of some 7 billion light years from the earth, in one direction.

B.2(b) Dark Matter and Dark Energy

Mysteries often come in multitudes - and in modern cosmology, this certainly seems to be the case. The story of dark matter begins with measurements of the orbital velocity of stars around their galactic centres. By looking at how the orbital velocity of a star varies with the distance from the centre, astronomers can deduce how much mass in the galaxy is contained in the region closer to the centre than that star (this is basically just Kepler's 3rd law). Slowly

a curious result began to emerge - there was far more mass in the galaxy than was visible, and its density seemed to fall off rather slowly as one moved away from the centre (as compared to the gas, dust and stars in the galaxy, which are concentrated in a thin disc, with a massive bulge near the centre). Whatever this extra 'stuff' was, it clearly had mass-energy (since it created a gravitational field) but did not appear to interact with either matter or radiation in any other way (so that it was invisible).

With the advent of gravitational lensing, it became possible to map dark matter in a much more sophisticated way, on scales of galaxy clusters and even much larger. The results are spectacular and of enormous importance.

To see what is involved, let's start by looking at Fig. 11(a). This shows a cluster of galaxies, the cluster Abell 520 (sometimes known as the "Train wreck cluster"), which is situated at a distance of 2.65 billion light years from us; the galaxies are tiny smudges on the main composite picture (not to be confused with the foreground stars, which are in our own galaxy). The 3 inset photos show (a) an X-ray image in red of the cluster - this picks out those parts of the cluster volume containing very high-temperature gas, which emits copious high-energy radiation; then (b) in blue, a mass/energy density image, produced by looking at the lensing images of much more distant galaxies; and (c) an optical image, showing visible light (and thus much of the ordinary matter). We see immediately that most of the mass/energy in this cluster is neither in the same location as the ordinary matter, nor as the high-temperature gas. Now the high-temperature gas we see is actually a relic of a collision between 2 clusters that took place some time ago - the modern 'train wreck cluster' is what is left behind now, in the form of an amalgamated cluster. This is the reason for the high-temperature gas - all galaxies and clusters of galaxies contain very large amounts of gas and dust which has not condensed into stars, gathered into massive clouds concentrated in the central regions of the cluster, as well as in each galaxy. The collision between the 2 clusters would have brought the 2 clouds into collision at high velocity, liberating a great deal of heat and high-energy radiation - the gas is still very hot and still radiating X-rays. But we see that most of the mass of the cluster is not on the region of the gas - and in fact most of it is not radiating at all, at any wavelength. This is the dark component.

Turning now to Fig. 11(b), we find a rather different picture here. This cluster, CL0024+17, is located some 5 billion light years from earth - as we see from the lower image, gravitational lensing images of more distant galaxies are quite obvious, and a very detailed map of the mass/energy density can be produced (top image in blue). Here we see something remarkable - much of this mass energy density is in a ring-like structure, away from the centre of the cluster, with some central concentration as well. Theorists have found this image very puzzling - the best current ideas suggest that the system is actually 2 or more clusters in the process of colliding right now, and/or that CL0024+17 is actually a cluster located at the junction of 3 dense 'filaments' of dark matter/energy.

Pictures like this can be repeated - we now have detailed maps of a few clusters, and much lower resolution maps of large volumes of the universe, extending out to large distances from us (see Fig. 11(d). The net result of these surveys can be seen both in Fig. 11(c), which shows a 'pie chart' representation of the various contributions to the mass-energy of the universe, and Fig. 10(e), which shows the inferred density of mass/energy in a large section of the universe, according to one particular model. The striking conclusion is that very little of the universe is in the form of ordinary matter or radiation - in fact, only 4% of its, and of this, only 0.4% is in the form of stars (the rest being gas and dust). The rest is either in a very high energy 'dark' component (usually called 'dark energy'), or in a cooler dark component called dark matter. This dark component is not homogeneously distributed - indeed, it exist in filaments and lumps, with very large voids in between. The same of course is true of the galaxies - and this is for good reason, because the clusters of galaxies are gravitationally quite tightly tied to the dark component, which exercises by far the strongest gravitational field.

Let us now stand back and take a look at what we have here. The discovery of dark mass/energy has actually cleared up some important questions. For example, it was always a puzzle that galaxies and stars existed at all. I mentioned above that they condensed out of the initially very small inhomogeneities that existed at the moment the microwave background radiation was released, 360,000 yrs after the Big Bang. But in fact it had long been a mystery how this could have happened - there was simply not a high enough density of matter to cause such a condensation. However the dark component solves this mystery - theory shows that if we increase the density by a factor of 25, then the dark component will father quickly enough into lumps, dragging the ordinary matter with it. Notice that the dark mass/energy and ordinary matter/radiation are not necessarily always in synch with each other - as we saw in the pictures of Abell 520 and CL0024+17, matter and radiation can cut adrift from the dark component, at least for a while. But it is now clear that the evolution of the universe has been largely controlled by the dark component, simply because it so totally dominates the mass/energy of the universe. Notice also that it provides one possible explanation for the acceleration of the expansion of the universe mentioned earlier, in the form of a contribution to the cosmological constant.

However, dark matter/energy poses a very big question for physics - for we simply do not know what it is! Thus, 96% of the mass/energy of the universe is in a form which has essentially no contact with the matter/radiation component of the universe (ie., that component which we are a part of), except through its gravitational effects. Even these are hard to see because dark matter/energy is so diffuse - we have yet to find it in clumps or singularities in

the same way that we see matter (in the form of stars, black holes, etc.). At the present time it is not clear if any theoretical model we have is capable of describing this new form of mass/energy.

Let us now summarize. In the last 50 years, General Relativity has come of age. It began as an intellectual triumph of extraordinary elegance and beauty, which seemed however to most physicists to be of little practical interest. It has, since then, survived every experimental test, and indeed these tests have developed into useful observational tools for understanding the universe. The ideas in GR, initially seen as radically new and strange, have now become a central pillar of physics, standing alongside quantum mechanics. This is not just the case in astrophysics. In fact, many of the ideas embodied in GR were crucial in the developments of our modern understanding of sub-atomic particles - although we still have yet to find a way of unifying GR with quantum mechanics. Nevertheless, only a century after the discovery of General Relativity, it has opened a door for us to the entire universe.

So far in this document I have only discussed some of the 'traditional' tests of GR, and some of its less spectacular consequences. However the real test of General Relativity, and its finest triumph (along with the prediction of the Big Bang) has been in the study of 'spacetime singularities', otherwise known as 'Black Holes'. In the next document, we discuss this topic, and how it has revolutionized astrophysics in the last 40 years.