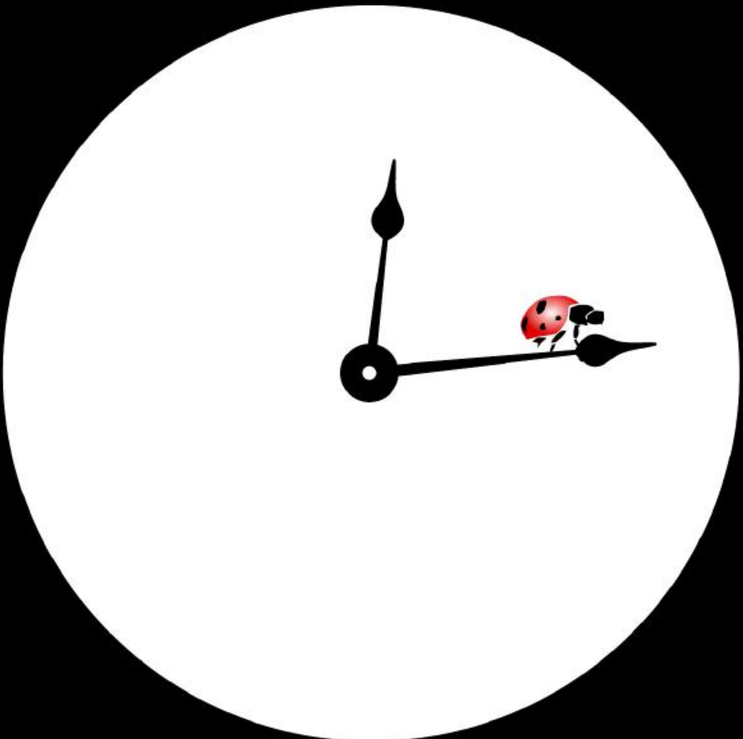
RELATIVITY FOR POETS



CROWELL

Relativity for Poets

Benjamin Crowell

www.lightandmatter.com

Fullerton, California
www.lightandmatter.com

rev. March 4, 2016

Contents

Part I

Space and time

Chapter 1

Galilean relativity

1.1 Galileo versus Aristotle

Once upon a time, there was an ornery man who liked to argue. He was undiplomatic and had a talent for converting allies into enemies. His name was Galileo, and it was his singular misfortune to be correct in most of his opinions. When he was arguing, Galileo had a few annoying habits. One was to answer a perfectly sound theoretical argument with a contradictory experiment or observation. Another was that, since he knew he was right, he freely made up descriptions of experiments that he hadn't actually done.

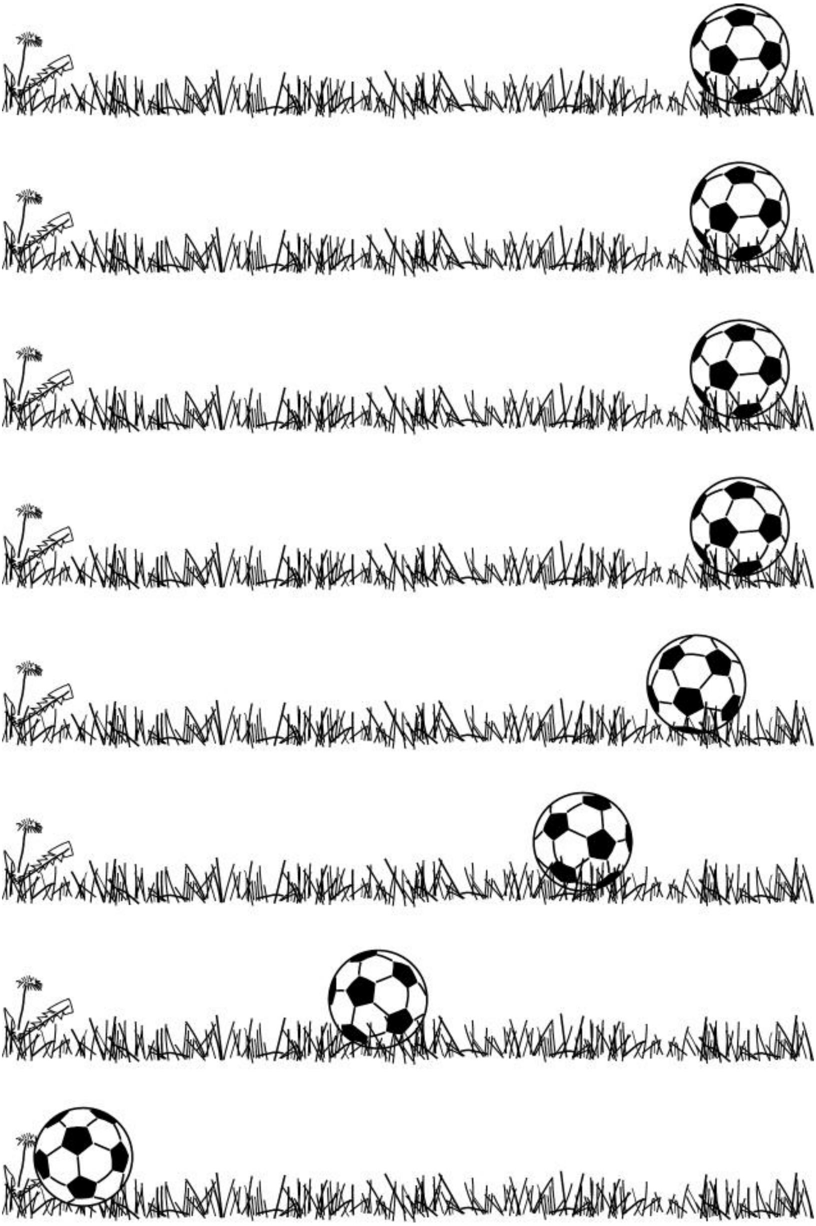
Galileo's true opponent was a dead man, the ancient Greek philosopher Aristotle. Aristotle, probably based on generalizations from everyday experience, had come up with some seemingly common-sense theories about motion.

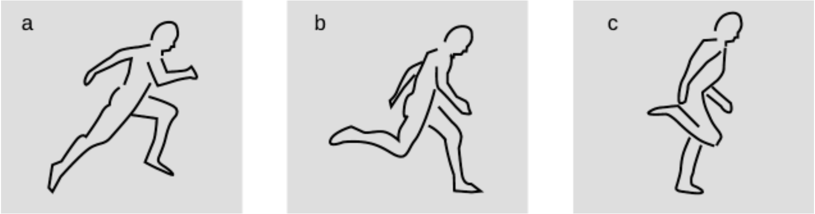
The figure on the facing page shows an example of the kind of observation that might lead you to the same conclusions as Aristotle. It is a series of snapshots in the motion of a rolling ball. Time moves forward as we go up the page. Because the ball slows down and eventually stops, it traces out the shape of a letter “J.” If you get rid of the artistic details, and connect the drawings of the ball with a smooth curve, you get a graph, with position and time on the horizontal and vertical axes. (If we had wanted to, we could have interchanged the axes or reversed either or both of them. Making the time axis vertical, and making the top point toward the future, is a standard convention in relativity, like the conventional orientation of the compass directions on a map.)

Once the ball has stopped completely, the graph becomes a vertical line. Time continues to flow, but the ball’s position is no longer changing. According to Aristotle, this is the natural behavior of any material object. The ball may move because someone kicks it, but once the force stops, the motion goes away. In other words, vertical lines on these graphs are special. They represent a natural, universal state of rest.

Galileo adduced two main arguments against the Aristotelian view. First, he said that there was a type of force, friction, which was the reason that things slowed down. The grass makes a frictional force on the ball. If we let the grass grow too high, this frictional force gets bigger, and the ball decelerates more quickly. If we mow the grass, the opposite happens. Galileo did experiments in which he rolled a smooth brass ball on a smooth ramp, in order to make the frictional force as small as possible. He carried out careful quantitative observations of the balls’ motion, with time measured using a primitive water clock. By taking measurements under a variety of conditions, he showed by extrapolation from his data that if the ramp was perfectly level, and friction completely absent, the ball would roll forever.

Today we have easier ways to convince ourselves of the same conclusion. For example, you’ve seen a puck glide frictionlessly across an air hockey table without slowing down. Sometimes

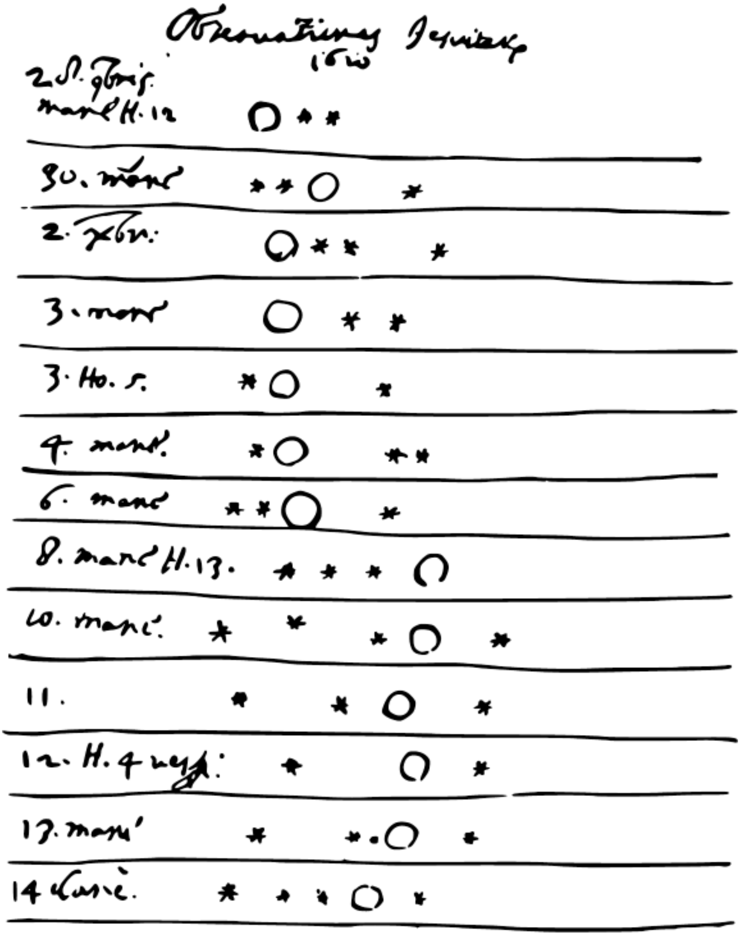




Galileo's analysis may be hard to accept. Running takes us a lot of effort, and it seems as though we need to apply a continuous force to keep going. The figure above shows what's really happening. At b, the runner's heel strikes the ground, and friction slows him down. A forward force in c serves only to recover from the loss in speed.

Galileo's second argument had to do with the fact that we can't judge motion except by comparing with some reference point, which we consider to be stationary. Aristotle used the earth as a reference point. But in 1610, Galileo looked at the planet Jupiter through a telescope that he had invented, and saw something startling. The planet, which looks like a bright and starlike point to the naked eye, appeared as a disk, and accompanying it were four points of light. Observing them from one night to the next, as shown in the notebook page reproduced on the facing page, he saw that they wove their way back and forth, to the east and west of Jupiter's disk, in the plane of the solar system. He inferred that they were moons that were circling Jupiter, and that he was seeing the circles from the side, so that they appeared flat. (The analogy with our diagram of the rolling ball is nearly exact, except that Galileo worked his way down the page from night to night, rather than going up.)

The behavior of these moons is nothing like what Aristotle would have predicted. The earth plays no special role here. The moons circle Jupiter, not the earth. Galileo's observations demoted the earth from its Aristotelian place as the universal reference point for the rest of the cosmos, making it into just one of several planets in the solar system.


Observationes Jesuitarum
1610
25. 9bris. mane H.12 O * *
30. mane * * O *
2. xbris. O * * *
3. mane O * *
3. Ho. s. * O *
4. mane. * O * *
6. mane * * O *
8. mane H.13. * * * O
10. mane. * * * O *
11. * * O *
12. H. 4 vesp: * O *
13. mane' * * * O *
14 Carie. * * * O *

1.2 Frames of reference

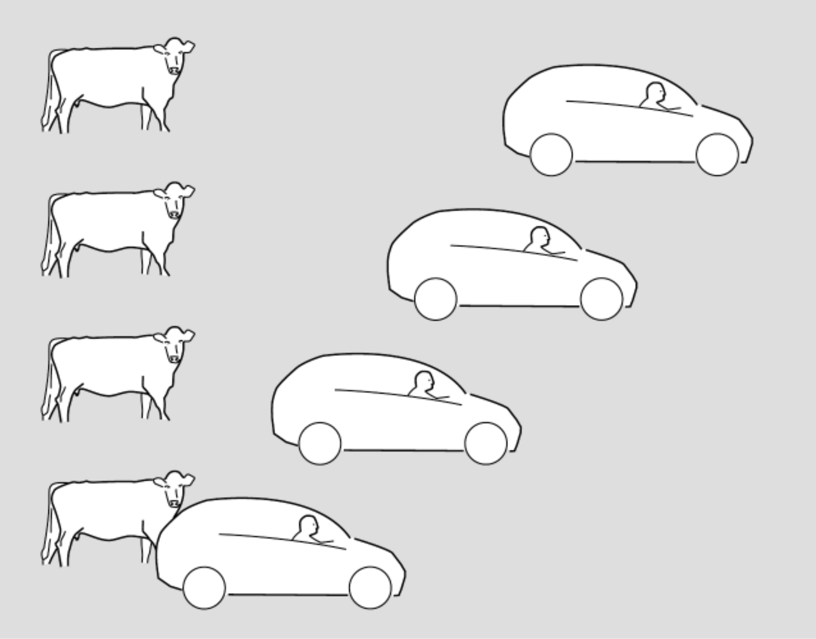
If Aristotle had grown up on Jupiter, maybe he would have considered Jupiter to be the natural reference point that was obviously at rest. In fact there is no one reference point that is always right. Psychologically, we have a strong tendency to think of our immediate surroundings as being at rest. For example, when you fly on an airplane, you tend to forget that the cabin is hurtling through the sky at nearly the speed of sound. You adopt the *frame of reference* of the cabin. If you were to describe the motion of the drink cart progressing down the aisle with your nuts and soda, you might say, "Five minutes ago, it was at row 23, and now it's at row 38." You've implicitly laid out a number line along the length of the plane, with a reference point near the front hatch (where "row zero" would be). The measurements like 23 and 38 are called *coordinates*, and in order to define them you need to pick a frame of reference.

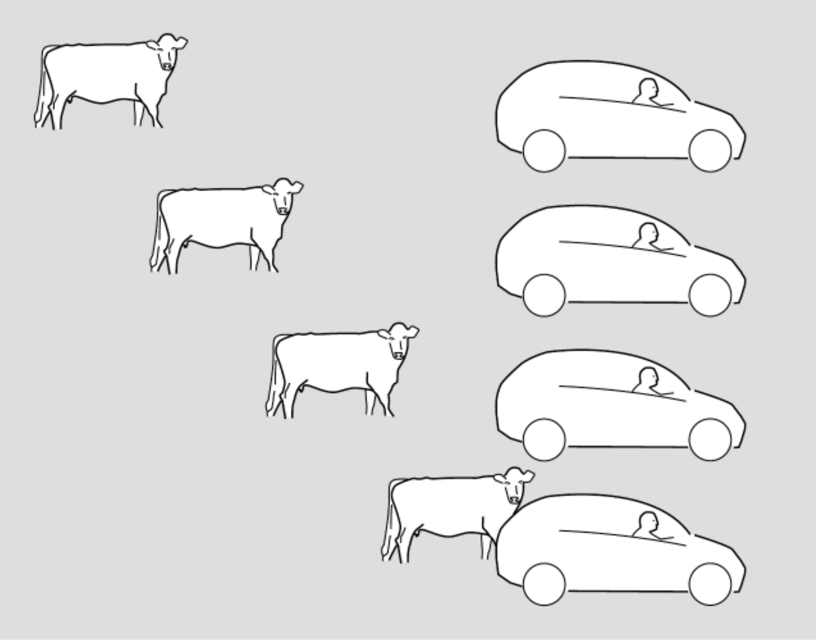
In the top part of the figure on the next page, we adopt the cow's frame of reference. The cow sees itself as being at rest, while the car drives by on the road. The version at the bottom shows the same situation in the frame of the driver, who considers herself to be at rest while the scenery rolls by.

Aristotle believed that there was one special, or *preferred* frame of reference, which was the one attached to the earth. Relativity says there is no such preferred frame.

1.3 Inertial and noninertial frames

That's not to say that all frames of reference are equally good. It's a little like the line from Orwell's satirical novel Animal Farm: "All animals are equal, but some animals are more equal than others." For an example of the distinction between good and bad frames of reference, consider a low-budget movie in which the director wants to depict an earthquake. The easy way to do it is to shake the camera around — but we can tell it's fake.

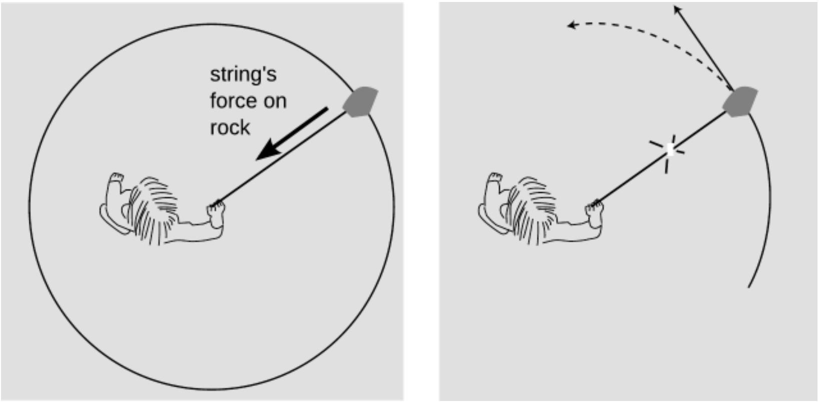




Similarly, when a plane hits bad turbulence, it's possible for passengers to go popping out of their seats and hit their heads on the ceiling. That's why the crew tells you to buckle your seatbelts. Normally we assume that when an object is at rest, it will stay at rest unless a force acts on it. Furthermore, Galileo realized that an object in motion will tend to stay in motion. This tendency to keep on doing the same thing is called *inertia.* One of our cues that the bouncing cabin is not a valid frame of reference is that if we use the cabin as a reference point, objects do not appear to behave inertially. The person who goes flying up and hits the ceiling was not acted on by any force, and yet he changed from being at rest to being in motion. We would interpret this as meaning that the plane's motion was not steady. The person didn't really pop up and hit the ceiling. What really happened, as suggested in the figure, was that the ceiling swerved downward and hit the person on the head. The cabin is a *noninertial frame of reference.* Moving frames are all right, but noninertial ones aren't.

1.4 Circular motion

In the era of Galileo and Newton, there was quite a bit of confusion about whether inertial motion was a broad enough notion to include both motion in a straight line and motion in a circle. Circular motion is not inertial. The figure is an overhead view of a person whirling a rock on the end of a string. When the string breaks, the rock flies off straight. It doesn't keep going in a circle. The inward force from the string is necessary to the circular motion, and when it disappears, we get straight-line motion. Similarly, when we spin a coin on a tabletop, the atoms



in the coin are being held together by attractive electrical forces from the neighboring atoms.

Linear, inertial motion is never detectable without reference to something external. Not so for circular motion. For example, if the room in which you're reading this book started to rotate, you would be able to tell because, for example, you'd feel dizzy. This would be your inner ear acting like a gyroscope. Passenger jets use a type of optical gyroscope in order to sense and maintain their orientation and stay on course. For these reasons, a rotating frame of reference is not considered an inertial frame.

1.5 Addition of velocities

All inertial frames of reference are equally valid. However, observers in different frames of reference can give different answers to questions and get different, correct answers from measurements. For example, suppose that Sparkly Elf Lady is riding her unicorn at 30 kilometers per hour, and she fires an arrow from her bow in the forwar direction. To her, the bow fires the arrow at its normal speed of 20 km/hr. But to an observer watching her go by the arrow is going 50 km/hr. To convert velocities between different frames of reference, we add and subtract.

As a more realistic example, with real-world consequences, Air France flight 447 disappeared without warning on June 1, 2009, over the Atlantic Ocean. All 232 people aboard were killed. Investigators believe the disaster was triggered because the pilots lost the ability to accurately determine their speed relative to the air. This is done using sensors called Pitot tubes, mounted outside the plane on the wing. Automated radio signals showed that these sensors gave conflicting readings before the crash, possibly because they iced up. For fuel efficiency, modern passenger jets fly at a very high altitude, but in the thin air they can only fly within a very narrow range of speeds. If the speed is too low, the plane stalls, and if it's too high, it breaks up. If the pilots can't tell what their airspeed is, they can't keep it in the safe range.

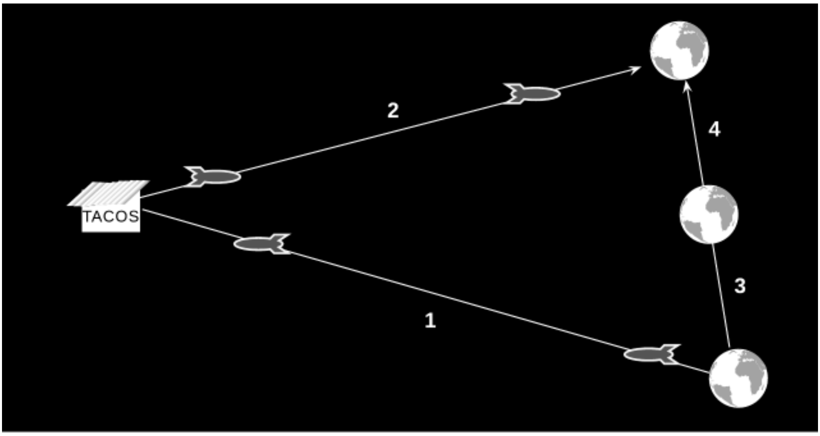
Many people's reaction to this story is to wonder why planes don't just use GPS to measure their speed. The answer has to do with addition of velocities. Passenger jets do have GPS these days, but GPS tells you your speed relative to the ground, not relative to the air. We have

$$\begin{aligned}
&\text{(velocity of the plane relative to the ground)}\\
=&\text{(velocity of the plane relative to the air)}\\
+&\text{(velocity of the air relative to the ground).}
\end{aligned}$$

Given two of these numbers, you could find the third, but GPS only tells you one of these numbers, and the one it tells you is not the one that the pilots on Air France 447 needed to know.

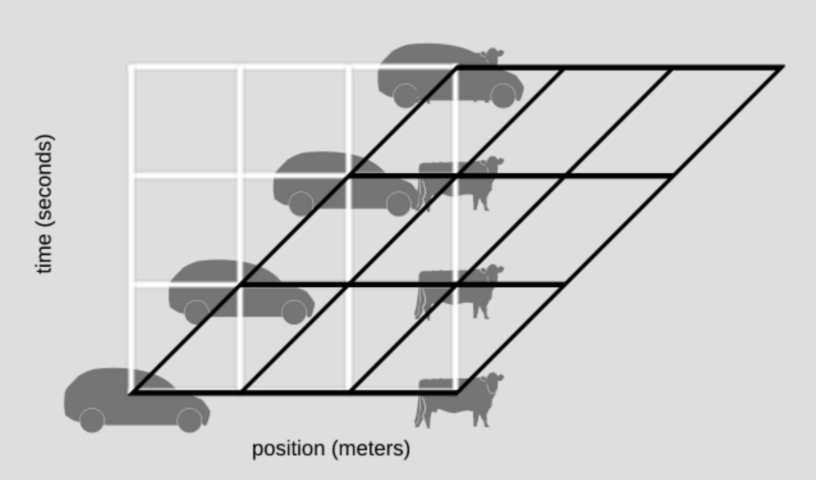
1.6 The Galilean twin paradox

Alice and Betty are identical teenage twins. Betty goes on a space voyage to get a fish taco with cilantro and grated carrots from a taco shack that has the best food in the known universe. Alice stays home. The diagram shows the story using our usual graphical conventions, with time running vertically and space horizontally.



Motion is relative, so it seems that it should be equally valid to consider Betty and the spaceship as having been at rest the whole time, while Alice and the planet earth traveled away from the spaceship along line segment 3 and then returned via 4. But this is not consistent with the experimental results, which show that Betty undergoes a violent deceleration and reacceleration at her turnaround point, while Alice and the other inhabitants of the earth feel no such effect.

The paradox is resolved because Galilean relativity doesn't say that *all* motion is relative. Only inertial motion is relative. When motion is noninertial, we can get detectable physical effects. The earth's motion is inertial, and the spaceship's isn't.



1.7 The Galilean transformation

In section 1.2 on p. 16, we visualized the frames of reference of the cow and the car by making two different diagrams of the motion, one with the cow at rest and one with the car at rest. Rather than making two separate graphs, we can be more economical by drawing a single diagram, but superimposing two different graph-paper grids, as shown above.

The white grid is what the cow would use. The car travels one meter forward for every second of time.

The driver of the car would prefer the black grid. According to this grid, the cow starts at $t = 0$ with a position of 3 meters, and by $t = 3$ seconds, it's gone backward to a position of 0 meters.

What we're doing here is switching back and forth between the coordinates of two different frames of reference. Coordinates are like names that we attach to events. When we switch coordinate systems, reality doesn't change, only the names do. It's just like translating words to a different language. This method of translating between two frames of reference is called the Galilean transformation.

SCIENCE AND SOCIETY

1.8 The Galileo affair

Galileo had a famous showdown with the Church, which ought to teach us something useful about the relationship between science and religion and their roles in society. But having taught about this topic for twenty years, I find that the more I learn about it, the less sure I am of what lessons to draw.

First some background. Galileo's lifetime (1564-1642) coincided exactly with the Counter-Reformation, and during his mature years Europe was consumed by a period of brutal warfare much like what we see today in places like Iraq, with religious and state powers cynically using one another to further their own agendas. In some areas, 75% of the population was killed by the war and its side-effects, including disease and famine. In parallel with the hard work of physical slaughter, there was an intellectual battle going on. New religious orders such as the Theatines were founded, with the mission of defending Italy from the Protestant heresy. In 1559 the Church published an Index of Forbidden Books (abolished in 1966), and in 1588 the Roman Inquisition was established. During the height of the conflict Galileo lived in relative safety near Florence, with a Medici prince as his patron. The security of his position was also seemingly buttressed because his scientific work was approved of by both the Jesuits and his friend Cardinal Barberini, who in 1623 became Pope Urban VIII.

But Galileo was a flamboyant personality and a best-selling author, and he ended up getting in trouble by very publicly advocating the Copernican system of the universe, in which the planets, including the earth, circled the sun. After some initial ignorant fulmination against Galileo from the pulpit, the first real shot across the bow came in the form of an intellectually sophisticated 1616 letter to Galileo from the Theatine priest Ingoli, who had participated in the Accademia dei Lincei, of which Galileo was a leading member. Apologists for the Church love Ingoli's letter, because it zeroes in on two real scientific weaknesses in Galileo's position. This is contrary to the usual picture of the Church persecuting poor Galileo purely because they thought a moving earth contradicted passages from the Bible, e.g., "Tremble before him, all the earth. The world is established and can never be moved," (1 Chronicles 16:30). Although Ingoli does include such theological points, they are downplayed. Ingoli, basing his main arguments on work by the astronomer Tycho Brahe, points out two purely *scientific* problems with Copernicanism:

1. Based on the best available data, the scale of a Copernican universe would have had to be such that the stars were implausibly large (about the size of our entire solar system).

2. If the earth were spinning on its own axis, then there would be detectable effects on the motion of projectiles, but no such effects are observed.

The first argument turns out to have been based on mistaken data, but we can recognize the second as an argument about *relativity.* Let's analyze this with the benefit of modern hindsight, since neither Galileo nor Ingoli had a clear enough concept of inertial motion to be able to attack it definitively.

One's initial reaction might be that motion is relative, so if we see the sun appear to spin around the earth, isn't it just a matter of opinion whether the earth is spinning or the sun revolving around it? From this point of view, arguing about an earth-centered cosmos versus a sun-centered one is like arguing about whether it's really the cow or the car on p. 17 that moves. This was in fact essentially the position taken by the higher Church officials later on: they told Galileo that he could discuss Copernicanism as a mathematical hypothesis, but not as a matter of physical fact.

But this argument is just plain wrong, since inertial motion is not just motion at constant speed, it's motion that is also in a *straight line*. Therefore motion in a circle really *is* more than a matter of opinion, and in fact we *can* detect the effects of the earth's rotation on projectiles. For example, when powerful naval guns fire a shell to the north, in the northern hemisphere, the shell can fly as far as 30 km, and its motion carries it to a higher latitude, at which the earth's rotation has a smaller velocity. (As an extreme case, Santa's candy-cane pole doesn't go anywhere as the earth rotates, it just spins in place.) The projectile retains its original, higher eastward velocity, so to an observer on the earth's surface, it appears to swerve to the east. This is known as the Coriolis effect, and is also the explanation for the rotation of cyclones. So Galileo was right about the earth's spin, but for the wrong reasons. At the time, the state of physics simply wasn't advanced enough to accurately calculate the effect Ingoli described. Such a calculation would have shown that the effect was too small to have been observable with contemporary technology.

The aged Galileo persisted stubbornly in advocating Copernicanism, and in 1633 he was sentenced to house arrest for the rest of his life. Tradition holds that after the sentence was pronounced, he muttered, "And yet, it does move."

So what do we learn from all this? A fairly common interpretation is that the whole conflict was needless. There doesn't have to be any conflict between science and religion. We simply need to keep science, religion, and the state within their separate and proper spheres of authority. Although I've given this interpretation to my students in the past, I can't help feeling now that I was being a Pollyanna. As I write this, the press is reporting on the killing of 12 journalists and 4 Jews by Muslim fundamentalists in Paris, as well as the massacre of 2,000 civilians in northern Nigeria by Boko Haram, an Islamic fundamentalist group whose name means literally "books are a sin." Maybe a more realistic moral to draw from the Galileo story is that religion is an inherently dangerous force in the world, one which tends to wreak havoc in the secular world unless it's backed into a corner with a whip, like a snarling circus lion.

Chapter 2

Einstein's relativity

2.1 Time is relative

In chapter 1 I laid out a beautiful theory of motion, which, by the way, is wrong. To see why it's wrong, we have to think about the assumptions hidden within it about space and time.

Suppose that I drive to Gettysburg, Pennsylvania, and stand in front of the brass plaque that marks the site of the momentous Civil War battle. There I am, at the same place where it all happened. But wait, am I really in the same place? An observer whose frame of reference was fixed to another planet would say that our planet had moved through space since 1863.

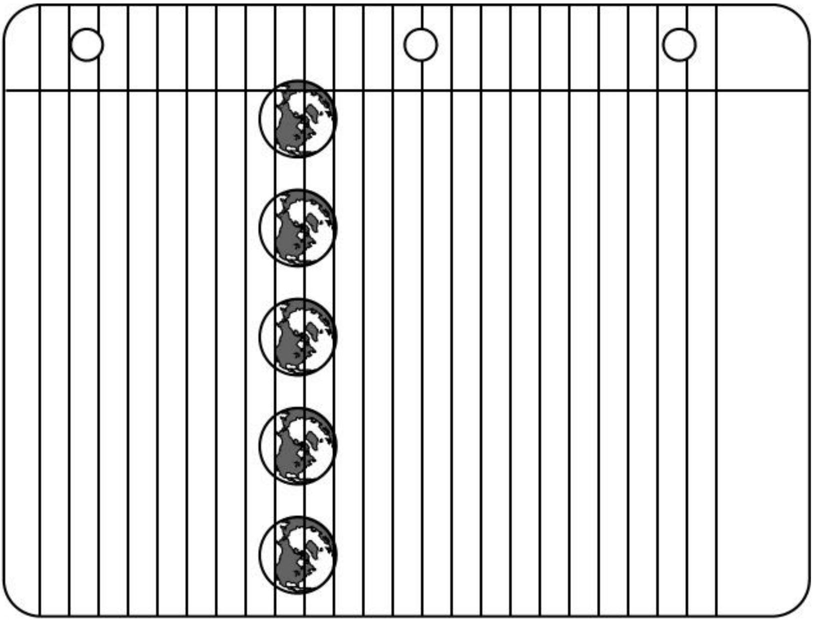
Consider the figure on the facing page. The top diagram shows the inertial motion of the earth as you'd learn to represent it if Aristotle was your schoolteacher. He tells you to use lined notebook paper, because its blue parallel lines represent a state of rest. The series of positions of the earth has to be drawn so that it lines up correctly. The earth is, after all, at rest.

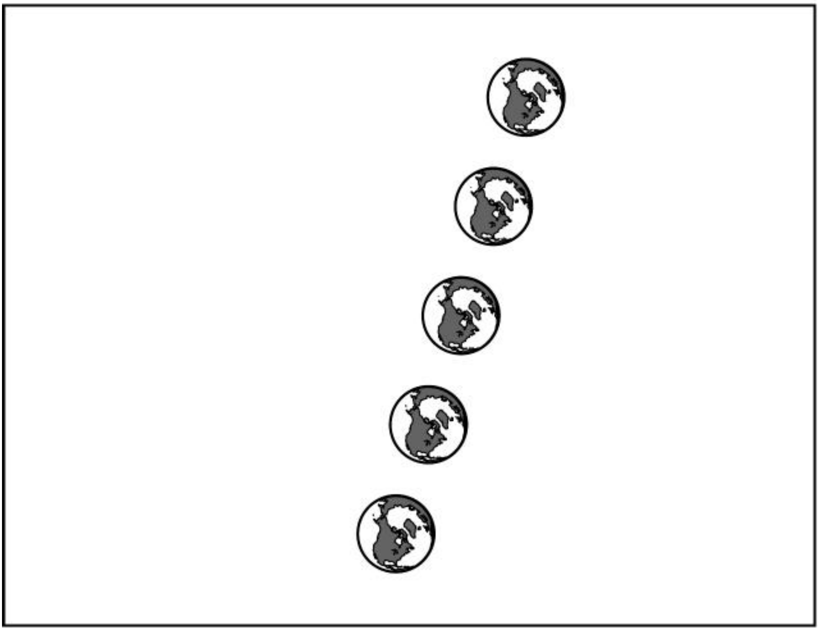
But then your parents pull you out of the strict, structured Aristotelian Academy and enroll you in the hippie-dippy, color-outside-the-lines Galilean School for Groovy Boys and Girls. At the Galilean School, the teachers just hand out blank pieces of white typing paper. Your new teachers tell you it's all right to make your diagram however you like, as long as the earth's positions form a line.

In other words, the things we've been saying about motion can also be taken as claims about the structure of the background, the stage of space and time on which motion is played out. We call this background *spacetime*, and each point on it is an *event*: a certain time and place, such as July 1, 1863 at Gettysburg, Pennsylvania.

Aristotelian spacetime comes equipped with a special structure, represented by the blue lines on the notebook paper. They represent a preferred frame of reference, and they always allow us to decide unambiguously whether or not two events happened at the *same place.* Galileo corrected this view by showing that spacetime didn't actually have any such structure. We can't say in any absolute sense whether two events happen at the same place. It's a matter of opinion, because it depends on the frame of reference we pick.

But *both* Galileo and Aristotle implicitly assumed a further type of structure for spacetime, which we can visualize as a set of *horizontal* lines on the paper. Events lying on the same horizontal line are *simultaneous*, and the assumption has been that simultaneity is not a matter of opinion. It's absolute.





Neither Galileo nor Aristotle realized that this was a nontrivial assumption, but Isaac Newton (1642-1726) did, and he explicitly stated it:

> Absolute, true, and mathematical time, of itself, and from its own nature, flows uniformly without regard to anything external [...]

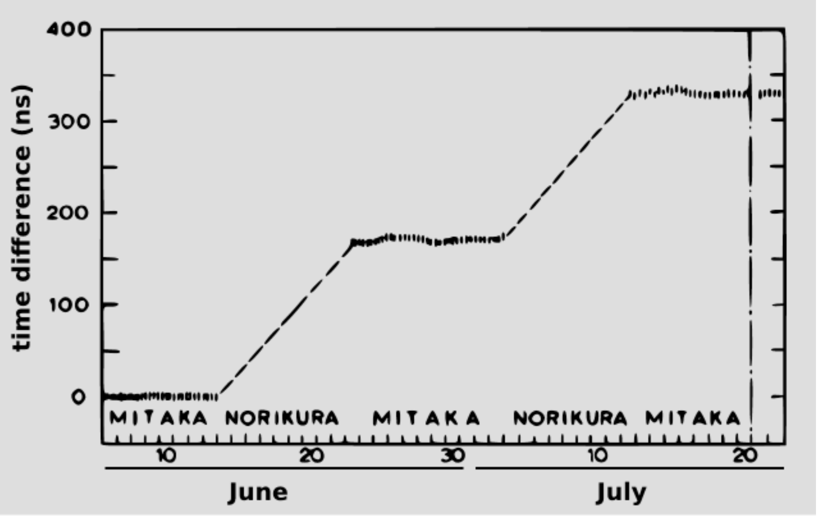
The photo shows a 1971 experiment that proves this hidden assumption to be wrong. There had been plenty of earlier evidence, but this experiment is particularly direct and easy to explain. Physicist J.C. Hafele and astronomer R.E. Keating brought atomic clocks aboard commercial airliners and flew around the world, once from east to west and once from west to east. The clocks took up two seats, and two tickets were bought for them under the name of "Mr. Clock."

When they got back, Hafele and Keating observed that there was a discrepancy between the times measured by the traveling clocks and the times measured by similar clocks that stayed home at the U.S. Naval Observatory in Washington. The effect was small enough that it has to be expressed in units of nanoseconds (ns), or billionths of a second. The east-going clock lost time, ending up off by -59 ns, while the west-going one gained 273 ns. Although the effects were small, they were statistically significant compared to the clocks' margins of error of about ±10 ns.

Actually it makes sense that the effects were small. Galileo's description of spacetime had already been thoroughly tested by experiments under a wide variety of conditions, so even if it's wrong, and we're going to replace it with some new theory, the new theory must agree with Galileo's to a good approximation, within the Galilean theory's realm of applicability. This requirement of backward-compatibility is known as the *correspondence principle.*

It's also reassuring that the effects on time were small compared to the three-day lengths of the plane trips. There was therefore no opportunity for paradoxical scenarios such as one in which the east-going experimenter arrived back in Washington before he left and then convinced himself not to take the trip. A theory that maintains this kind of orderly relationship between cause and effect is said to satisfy causality.

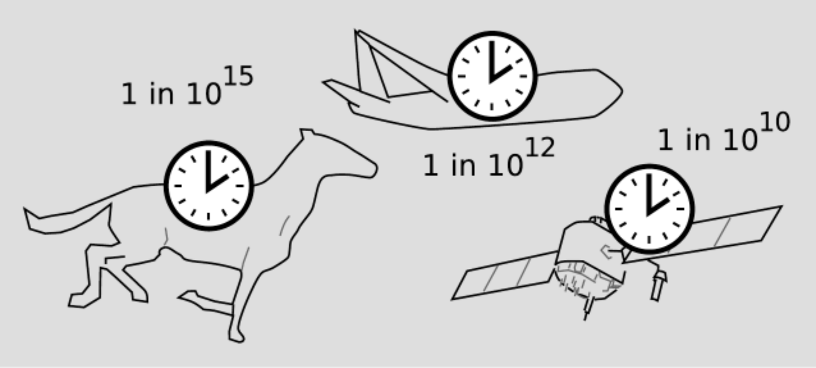
Hafele and Keating were not working in the dark. They already had access to an improved theory of spacetime, which was Albert Einstein's theory of relativity, developed around 1905-1916. They were testing relativity's predictions (and they were not the first to do so). Let's work backward instead, and inspect the empirical results for clues as to how time works. The east-going clock lost time, while the west-going one gained. Since two traveling clocks experienced effects in opposite directions, we can tell that the rate at which time flows depends on the motion of the observer. The east-going clock was moving in the same direction as the earth's rotation, so its velocity relative to the earth's center was greater than that of the clock that remained in Washington, while the west-going clock's velocity correspondingly reduced. The fact that the east-going clock fell behind, and the west-going one got ahead, shows that the effect of motion is to make time go more slowly. This effect of motion on time was predicted by Einstein in his original 1905 paper on relativity, written when he was 26.



If this had been the only effect in the Hafele-Keating experiment, then we would have expected to see effects on the east-going and west-going clocks that were equal in size. In fact, the two effects are unequal. This implies that there is a second effect involved, simply due to the planes' being up in the air. This was verified more directly in a 1978 experiment by Iijima and Fujiwara, in which one clock was kept in Mitaka, a suburb of Tokyo, while an identical clock was driven up nearby Mount Norikura, left motionless there for about a week, and then brought back down and compared with the one at the bottom of the mountain. Their results are shown in the graph. This experiment, unlike the Hafele-Keating one, isolates one effect on time, the gravitational one: time's rate of flow increases with height in a gravitational field. Einstein didn't figure out how to incorporate gravity into relativity until 1915, after much frustration and many false starts. The simpler version of the theory without gravity is known as special relativity, the full version as general relativity. We'll restrict ourselves to special relativity for now, and that means that what we want to focus on right now is the distortion of time due to motion, not gravity.

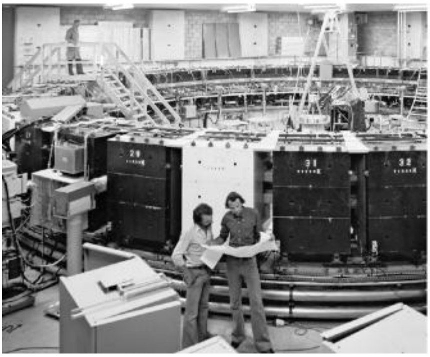
By the way, the effects described in these atomic clock experiments could have seemed obscure to laypeople in the 1970s, but today they are part of everyday life, because the GPS system depends crucially on them. A GPS satellite in orbit experiences effects due to motion and gravity that are both much larger than the corresponding effects in the Hafele-Keating and Iijima experiments. The satellites carry atomic clocks, and beam down time-stamped radio signals to receivers such as the ones used by motorists and hikers. By comparing the time stamps of signals from several different satellites, the receiver can calculate how long the signals took to travel to it at the speed of light, and therefore determine its own position. The GPS network started out as a US military system, and this was why Hafele and Keating were able to get funding from the Navy. There is a legend that the military brass in charge of the program weren't quite sure they believed in all the crazy relativity stuff being spouted by the longhaired physics professors, so they demanded that the satellites have special software switches designed into them so that the relativity corrections could be turned off if necessary. In fact, both special and general relativity are crucially important to GPS, and the system would be completely broken without them.

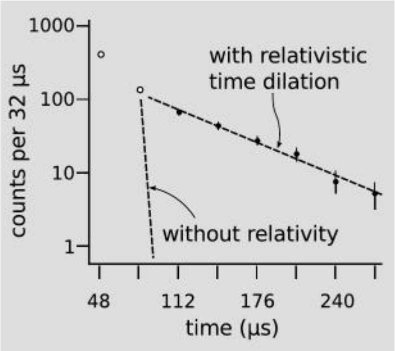
We can now see in more detail how to apply the correspondence principle. The behavior of the clocks in the Hafele-Keating experiment shows that the amount of time distortion increases as the speed of the clock's motion increases. Galileo lived in an era when the fastest mode of transportation was a galloping horse, and the best pendulum clocks would accumulate errors of perhaps a minute over the course of several days. A horse is much slower than a jet plane, so the distortion of time would have had a relative size of only one part in $1,000,000,000,000,000$ (10^{15}) — much smaller than the clocks were capable of detect-



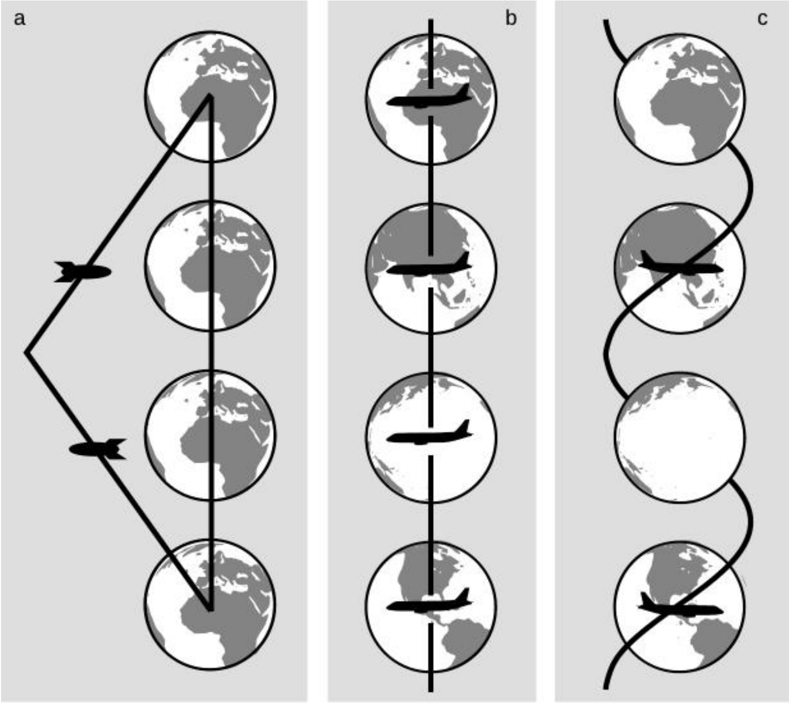
ing. At the speed of a passenger jet, the effect is about one part in 1,000,000,000,000 (10^{12}), and state-of-the-art atomic clocks in 1971 were capable of measuring that. A GPS satellite travels much faster than a jet airplane, and the effect on the satellite turns out to be about one part in 10 billion (10^{10}). The general idea here is that all physical laws are approximations, and approximations aren't simply right or wrong in different situations. Approximations are better or worse in different situations, and the question is whether a particular approximation is good enough in a given situation to serve a particular purpose. The faster the motion, the worse the Galilean approximation of absolute time. Whether the approximation is good enough depends on what you're trying to accomplish. The correspondence principle says that the approximation must have been good enough to explain all the experiments done in the centuries before Einstein came up with relativity.

But I don't want to give the impression that relativistic effects are always small. If we want to see a large time dilation effect, we can't do it with something the size of the atomic clocks Hafele and Keating used; they would become deadly missiles with destructive power greater than the total megatonnage of all the world's nuclear arsenals. We can, however, accelerate subatomic particles to very high speeds. For experimental particle physicists, relativity is something you do all day before heading home and stopping off at the store for milk. An early, low-precision experiment of this kind was performed by Rossi and Hall in 1941, using naturally occurring cosmic rays. The figure shows a 1974 particle-accelerator experiment of a similar type.





Particles called muons (named after Greek μ, "myoo") were produced by an accelerator at CERN, near Geneva. A muon is essentially a heavier version of the electron. Muons undergo radioactive decay, lasting an average of only 2.197 microseconds (millionths of a second, coincidentally also written with the Greek μ). The experiment was actually built in order to measure the magnetic properties of muons, but it produced a high-precision test of time dilation as a byproduct. Muons were injected into the ring shown in the photo, circling around it until they underwent radioactive decay. At the speed at which these muons were traveling, time was slowed down by a factor of 29.33, so on the average they lasted 29.33 times longer than the normal lifetime. In other words, they were like tiny alarm clocks that self-destructed at a randomly selected time. The graph shows the number of radioactive decays counted, plotted versus the time elapsed after a given stream of muons was injected into the storage ring. The two dashed lines show the rates of decay predicted with and without relativity. The relativistic line is the one that agrees with experiment.



2.2 The principle of greatest time

A paradox now arises: how can all of this be reconciled with the fact that motion is relative? If clocks run more slowly because they're in motion, couldn't we determine a preferred rest frame by looking for the frame in which clocks run the fastest?

To clear this up, it will help if we start with a simpler example than the Hafele-Keating experiment, which involved three different sets of clocks, motion in circles, and the motion of the planes being superimposed on the rotation of the earth. Let's consider instead the relativistic version of the Galilean twin paradox from section 1.6, p. 21, reproduced here in panel a of the figure. Recall that in this thought experiment, Alice stays at home on earth while her twin Betty goes on a space voyage and returns. The relativistic version of this story is in fact what people normally have in mind when they refer to "the" twin paradox, and it has an especially strange twist to it. According to relativity, time flows more slowly for Betty, the traveling twin. The faster her motion is, the bigger the effect. If she flies fast enough, we can produce a situation in which Betty arrives home still a teenager, while Alice has aged into an old woman!

But how can this be? In Betty's frame of reference, isn't it the earth that's moving, so that it should be the other way around, with Alice aging less and Betty more?

The resolution of the paradox is the same as in the Galilean case. Alice's motion is inertial, and we can see this because her track across the diagram, called her *world-line* in relativistic parlance,[1] is a straight line. Betty's motion isn't inertial; it consists of two segments at an angle to one another. In fact, there is a nice way of stating the relativistic rule for time dilation in these terms. The *principle of greatest time* states that among all the possible world-lines that a clock can follow from one event to another, the one that gives the greatest elapsed time is the inertial one. The shortened time interval is said to be *dilated.*

Analyzing the Hafele-Keating experiment becomes almost as easy if we simplify by pretending that the west-flying plane was going at exactly the right speed to cancel out the earth's rotation. (In reality this was probably roughly, but not quite, true.) Then this plane actually stood still relative to the center of the earth, whose motion is inertial. Therefore the plane's motion is inertial, as shown in panel b of the figure. The east-going clock flies in a circle, with its airspeed added on to the rotation of the earth to give double the velocity. As shown in panel c, its world-line is not an inertial straight line. Although panels b and c are drawn separately for clarity, the clocks really did start out together at the beginning and have a reunion at the end, as in the principle of greatest time. Therefore the inertial, west-going will record more elapsed time than the east-flying plane.

[1] The odd-sounding term is a translation from German of a phrase intended to mean "line *through* the world," i.e., through spacetime.

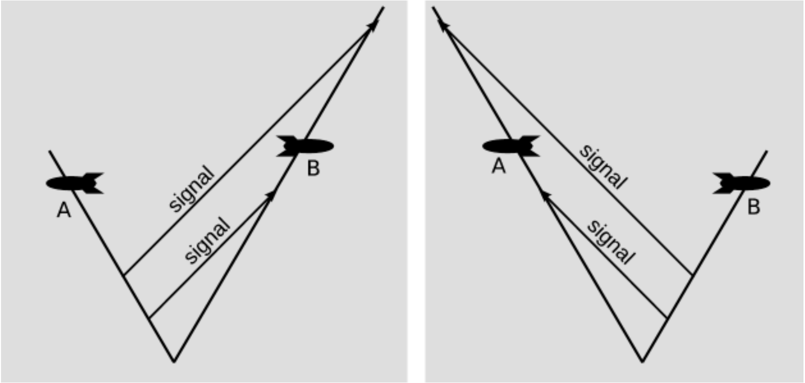
2.3 A universal speed limit

So far we've considered experiments that seem to fall into two different categories. In the atomic clock experiments, the clocks were together to start with, were separated, and were then reunited and compared again. These dovetail nicely with the structure of the principle of greatest time. But no such reunion happens, for example, in the muon experiments. When I first learned relativity, I got terribly confused about the no-reunion experiments. According to Einstein, if observers A and B aren't at rest relative to each other, then A says B's time is slow, but B says A is the slow one. How can this be? If A says B is slow, shouldn't B say A is fast? After all, if I took a pill that sped up my brain, everyone else would seem slow to me, and I would seem fast to them.

Suppose, for example, that Betty is on the outward-bound leg of her interstellar taco run. She puts the ship on cruise control and gets on the phone with Alice. Talking on the phone, can't they now establish who's actually slow and who's fast? Shouldn't the slow one sound like Darth Vader to the fast one, and conversely shouldn't the slow one hear the fast one's voice as that of a chipmunk who's been using helium and crack cocaine? There would then be an objective difference between Alice and Betty's frames — but this wrecks the logic of relativity, which says that all inertial frames are supposed to be equally valid.

The key is that we've been implicitly assuming that the interstellar cell phone is an instantaneous method of communication, so that it establishes the truth of the matter: who is really slow and who is really fast. We are forced to the opposite conclusion, that the phones do *not* send signals that propagate at an infinite speed — in fact, that *no* form of communication, no method of cause and effect, can operate at a distance without a certain time lag. The whole logical structure of relativity falls apart unless we assume a universal speed limit for all motion in the universe. We refer to this speed with the letter c.

The figure shows what actually happens to Alice and Betty

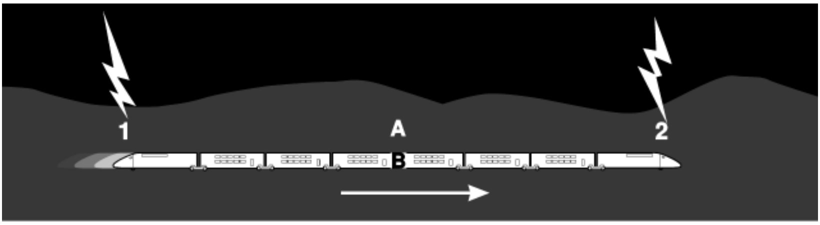


if, for example, one twin makes two hand claps near her phone, separated by a one-second interval. If the two signals traveled at infinite speed, then their world-lines would be horizontal — they would be received at the same time they were sent. But because they actually travel at a finite speed (let's say exactly at c), they are sent at one time and arrive at some later time, and the signals' world-lines have some slope. We conventionally choose the time and distance scales on our diagrams so that this is a 45-degree angle. We can see on the diagram that when Alice sends the two hand claps, Betty receives them at times that are spread to more than one second apart, but exactly the same thing happens when we flip the diagram. The situation is completely symmetric, and each twin perceives the other's transmission as having been slowed down.

The universal speed limit c is also the speed at which light travels. In the metric system, which is designed to handle times and distances on the human scale, c is a huge number, about $300,000,000$ meters per second (3×10^8 m/s in scientific notation). This is an example of the correspondence principle at work. One of the reasons that we don't notice the effects of Einstein's relativity ("relativistic" effects) very often in everyday life is that the speeds at which we walk, drive a car, and throw baseballs are so small compared to c.

2.4 Einstein's train

We've seen that if simultaneity isn't absolute, then there must be a universal speed limit c. The converse is also true: if c is universal, then simultaneity must be relative. The figure shows a famous thought experiment used by Einstein to present this idea.



A train is moving at constant velocity to the right when bolts of lightning strike the ground near its front and back. Alice, standing on the dirt at the midpoint of the flashes, observes that the light from the two flashes arrives simultaneously, so she says the two strikes must have occurred simultaneously.

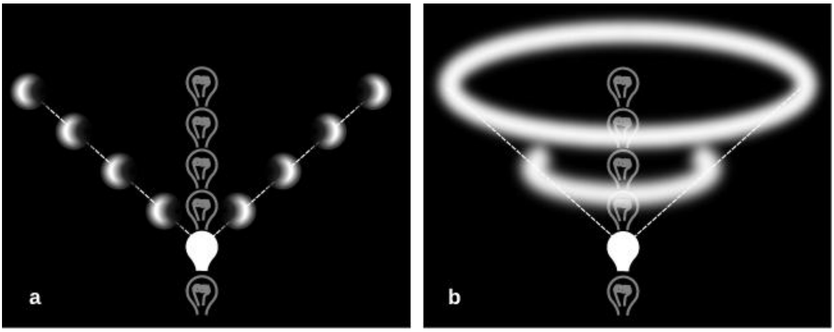
Bob, meanwhile, is sitting aboard the train, at its middle. He passes by Alice at the moment when Alice later figures out that the flashes happened. Later, he receives flash 2, and then flash 1. Since the light from the flashes travels at c, and c is *universal*, it has the same value in Bob's frame of reference as in Alice's. Bob knows that the two flashes traveled at the same speed as each other. Therefore he infers that since both flashes traveled half the length of the train, and flash 2 arrived first, flash 2 must have occurred first. The two events that in Alice's frame are simultaneous are not simultaneous to Bob.

2.5 Velocities don't simply add

It's crucial to the example of the train in section 2.4 that the speed of light is the same for Bob as for Alice. We saw in section 1.5, p. 19, that according to the Galilean description of spacetime, velocities add in relative motion. This is not the case according to Einstein's relativity. If it were, then a velocity c in one frame wouldn't equal c in another frame. It wouldn't be universal. Einstein later recalled that when he was trying to create special relativity, he got stuck on this point for about a year. It stumped him that "the concept of the invariance of the velocity of light ... contradicts the addition rule of velocities," which he had been assuming would still be true. Finally he realized that "Time cannot be absolutely defined, and there is an inseparable relation between time and [...] velocity." After this he finished his seminal paper on special relativity within five weeks.

Addition of velocities had, of course, already been used successfully in countless experiments and practical applications in the course of the several centuries since Galileo. The correspondence principle tells us that simple velocity addition must be a good approximation in the conditions under which it had already been tested. But all of these experiments had been ones involving material objects at speeds low compared to c. We'll see later how to combine velocities correctly, but that method will have to give results nearly the same as straight addition when the velocities are small. For example, in section 1.5 we added velocities of 30 km/hr and 20 km/hr to get 50 km/hr. The correct relativistic result is 49.9999999999999995 km/hr.

That was for velocities small compared to c. If we go to the opposite extreme, combining a velocity that *equals* c with any other velocity should simply give back c, since c is universal. A high-precision test of this prediction was performed in 1964 by Alväger *et al.* In their experiment, subatomic particles called mesons were produced, in motion at 99.98% of c. These particles are radioactive and decay in flight, producing flashes of light. If Galilean addition of velocities were correct, then a flash emitted in the forward direction at 100% of c relative to the decaying particle would be expected to move with a velocity of 199.98% of c, i.e., almost twice the normal speed of light. The experiment actually found them to travel at 100.00% of c, as predicted by relativity, to within the experiment's precision of ±0.01%.



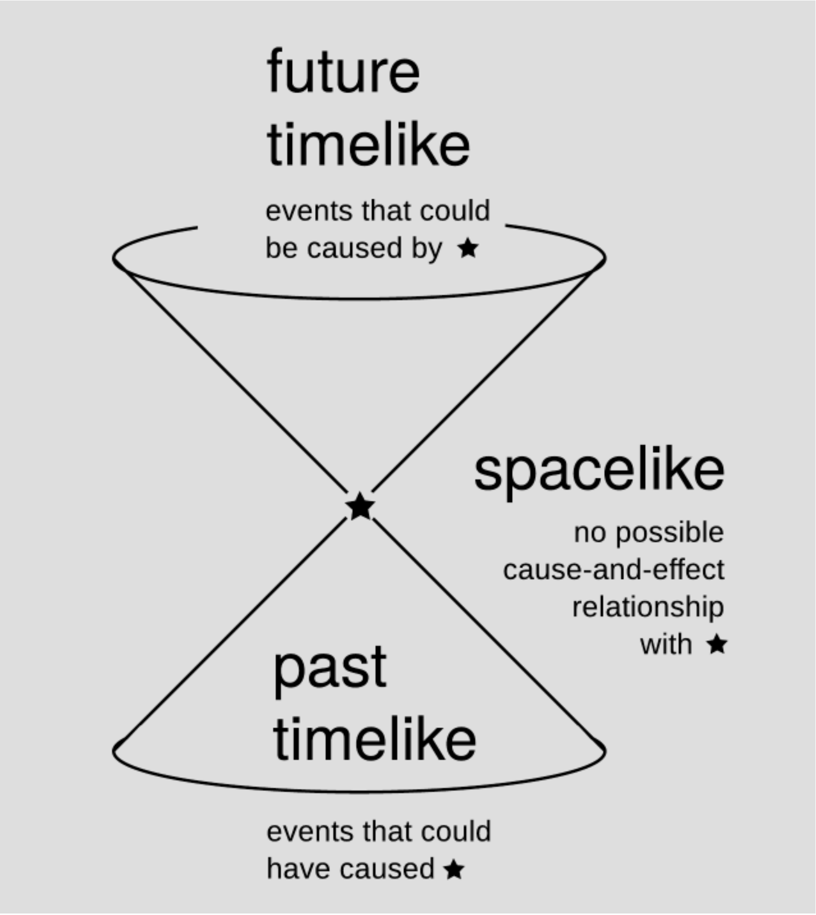
2.6 The light cone

In figure a above, a light bulb blinks on momentarily and then back off. Flashes of light spread out from it at c to the right and left, tracing lines which, because of our choice of units, make 45-degree angles. These lines form a geometrical figure, which is called the *light cone*, for reasons that may be more clear from figure b, which attempts to represent two dimensions of space as well as the time dimension.

You might not guess from this description that the light cone is fundamentally important, or that it doesn't have much to do with light. Remember, c isn't really the speed of light, it's the ultimate speed limit on cause and effect. We can draw a light cone centered on any event, such as the one labeled $\star$ in the figure on the next page. If $\star$ is the Russian Revolution, then events inside $\star$'s future light cone can be described as all the ones that we could reach, starting from the Russian Revolution, by traveling at less than the speed of light. A signal could be sent from $\star$ to any of these events, and $\star$ could be the *cause* of any of these events. In Einstein's theory of space and time, the light cone plays the same fundamental geometrical role as the circle in Euclidean geometry.

Events inside $\star$'s light cone are said to have a *timelike* relation to $\star$, because they are separated from it by more time than space. For example, one of these events could be separated in time from

⋆ by 100 years, but in space by only 50 light-years (the distance light travels in 50 years). Then a spaceship could get there by going at half the speed of light. Timelike relationships can be future-timelike or past-timelike. Events outside ⋆'s light cone are said to be *spacelike* in relation to ⋆, and events on the surface of the light cone as *lightlike*.



SCIENCE AND SOCIETY

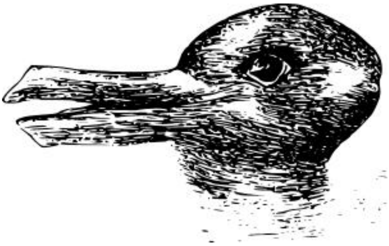
2.7 Scientific revolutions

Special relativity overthrew Galilean relativity, a scientific theory that physicists had wholeheartedly believed in for hundreds of years. Does an example like this mean that physicists don't know what they're talking about?

Some social scientists deny that the scientific method even exists, claiming that science is no more than a social system that determines what ideas to accept based on an in-group's criteria. If so, it would seem difficult to explain its effectiveness in building such useful items as CD players and sewers — accomplishments that seem to have eluded voodoo and astrology. The extreme social-relativist attitude was effectively skewered in a famous hoax carried out in 1996 by New York University physicist Alan Sokal, who wrote a nonsense article titled "Transgressing the Boundaries: Toward a Transformative Hermeneutics of Quantum Gravity," and got it accepted by a cultural studies journal called *Social Text*.

Most physicists claim that their science does tell us objective truths about the world, and they would invoke the correspondence principle (p. 31) in order to explain an example such as the replacement of Galilean relativity with special relativity. From this point of view, science is not merely a matter of fashion. If a theory such as Galilean relativity became accepted in physics, it was because it was supported by a vast number of experiments. It's just that experiments never have perfect accuracy. The old experiments weren't all wrong. They were right, within their limitations. This point of view is associated with the philosopher of science Karl Popper (1902-1994). Trained as a psychologist during the era when Freud was a dominant figure, Popper proposed *falsifiability* as the criterion for distinguishing science from non-science. According to Popper, real science makes itself vulnerable by proposing theories that predict new things, and that could be disproved if the predictions turned out to be false. He saw Freudianism as nonscientific, since it was so elastic that it could be made to retroactively "explain" any facts whatsoever.

It can be difficult to shoehorn real science into Popper's system. Galileo got in trouble for saying that the solar system revolved around the sun rather than the earth, and in order to advocate this, he had to argue that motion was relative, since otherwise we would observe dramatic effects of the earth's motion in our environment at the earth's surface. But in fact the sun-centered Copernican system did not make more accurate predictions of the motion of the planets than the earth-centered one. By Popper's criterion, the whole argument was unscientific.



A controversial alternative to Popper comes from Thomas Kuhn (1922-1996), who argues that science operates in two different modes, a normal one in which knowledge is steadily accumulated, and revolutionary steps, or *paradigm shifts.* According to Kuhn, Galilean relativity and heliocentrism were a paradigm shift that only much later led to a theory that was superior to the old one in terms of falsifiability. Kuhn claims that theories from after a paradigm shift are incommensurable with theories from before it, meaning that they ask different questions and require different world-views of their users. He compared this to an optical illusion that the brain can interpret in two different ways. Kuhn would probably reject my analysis in terms of the correspondence principle as a fairy-tale version of history. His critics see him as attacking scientific realism, the notion of scientific truth, and the authority of science.

Chapter 3

The Lorentz transformation

3.1 Surveying space and time

The human brain is highly specialized for visual processing, and we have the intuitive feeling that we can always look at a scene and get a snapshot of *what is*, a continuously updated flicker of *now* on our retinas. Metaphorically, we even apply this intuition to things that aren't actually visible. "I see what you mean," we say. "Let's look at the situation." "Are we seeing eye to eye?"

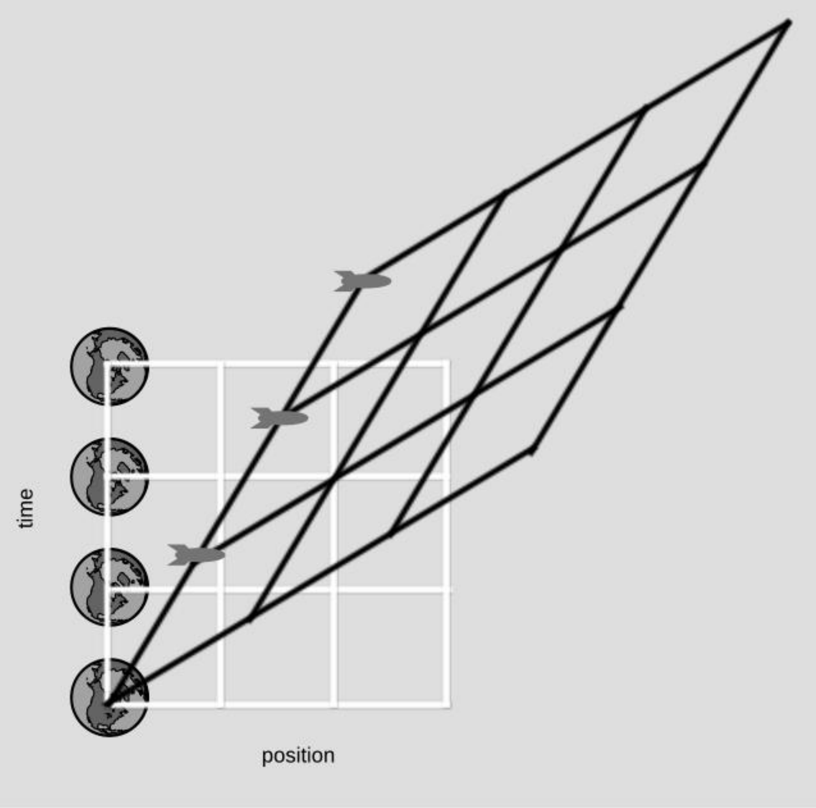
Just as Galileo's contemporaries struggled to overcome their intuitive resistance to the principle of inertia, people today need to wrestle with this powerful intuition, which is wrong and arises from the fact that the light coming to our eyes goes so fast that it seems to arrive instantaneously. When you look at the night sky, you're seeing the stars as they were in the past, when their light started on its journey to you. A star may have already gone through its death throes and exploded, but still appear to us as it was before. Within an atom, the subatomic particles move at a substantial fraction of c, and attract and repel each other through electric and magnetic forces. The force experienced by a particular electron is not a force created by the other particles at their current locations, but at the locations they had before the time lag required for the propagation of the force. In the movie Star Wars, a planet is destroyed by the bad guys, and Obi Wan Kenobi says, “I felt a great disturbance in the Force, as if millions of voices suddenly cried out in terror and were suddenly silenced.” But if “the Force” operated according to the same rules as the known fields of force in physics (electricity, gravity, etc.), there would have been a delay of many years before he could have gotten the signal.

Once we do receive a signal from some distant event, we can always apply a correction to find out when it really happened. For example, Chinese astronomers recorded a supernova in the year 1054, in the constellation Taurus. We now know that the star that exploded was about 6000 light years from earth. (A light year is the distance light travels in one year.) Therefore the explosion must have really happened around the year 5000 B.C. The “now” wasn’t really now, it was a long time ago. This correction technique is called *Einstein synchronization.* But even with Einstein synchronization, the example of the train, section 2.4, p. 40, shows that simultaneity is relative. If observers A and B are not at rest relative to one another, they have different definitions of “now.”

3.2 The Lorentz transformation

Subject to these limitations, the best we can do is to construct the relativistic version of the Galilean transformation (section 1.7, p. 22). The relativistic one, shown in the figure on the next page, is called the Lorentz transformation, after Hendrik Lorentz (1853-1928), who partially anticipated Einstein’s work, without arriving at the correct interpretation. The distortion of the graph-paper grid is a kind of smooshing and stretching.

The most noticeable feature of the Lorentz transformation is that it doesn’t maintain simultaneity. The horizontal lines on the white grid connect points that are simultaneous in the earth’s



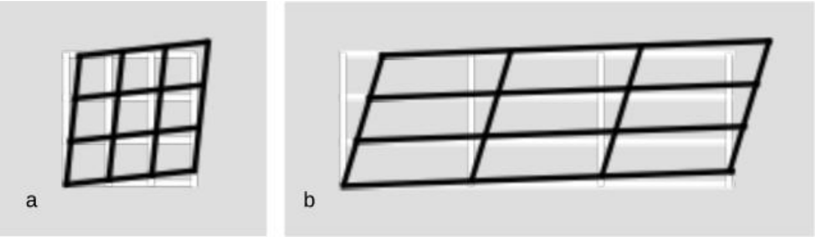
frame of reference, but they are not parallel to the corresponding lines on the black grid, which represents the rocket ship's frame.

One thing that does coincide between the two grids is the 45-degree diagonal, which is a piece of the light cone. This makes sense, because the two observers should agree on c. As usual, the units of the graph are chosen so that c lies at this slope. For example, if the units of time are years, then the distances have to be in light-years.

As in the Galilean case, the slope of the black grid's left edge relates to the velocity of the spaceship relative to the earth. In this example, the spaceship takes 5 time units to travel 3 distance units, so its velocity relative to the earth is 3/5 of c.

3.3 Correspondence principle

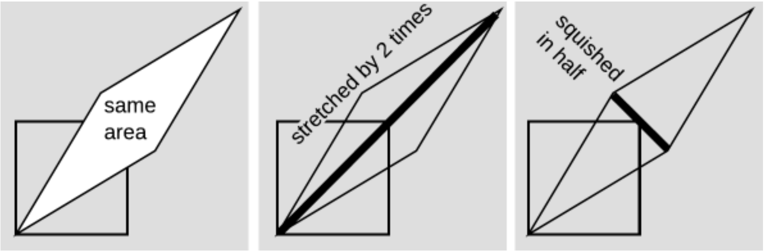
The correspondence principle requires that the Lorentz transformation (p. 49) match up with the Galilean transformation (p. 22) when the velocity of one frame relative to the other is small compared to c. This does work out properly because of the special relativistic units used in our graphical representation of the Lorentz transformation. Let's say that we're using seconds as our unit of time, so that our unit of distance has to be the light-second, the distance traveled by light in one second. A light-second is a huge distance, roughly equal to the distance from the earth to the moon.



The figure above shows an example of what happens in a nonrelativistic experiment. In panel a, we have a Lorentz transformation for a velocity that's small compared to c, in relativistic units. In panel b, we stretch the horizontal scale quite a bit, although nowhere near as much as we would need to do in order to convert from, say, light-seconds to meters. The result is nearly indistinguishable from a Galilean transformation.

3.4 Scaling, and the squish and stretch factors

The physical arguments we've made so far are enough to pin down every feature of the Lorentz transformation except for an over-all scaling factor. Given the white grid, we could seemingly take the black one and blow it up or reduce it like a photograph.

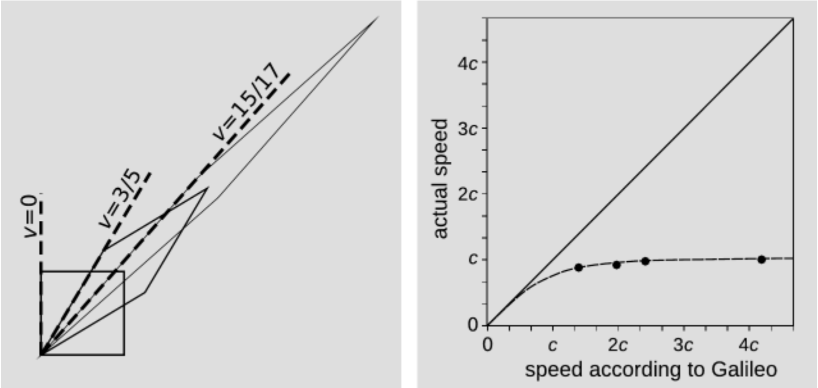


This scale can be fixed by using the fact, proved in section 3.11, p. 61, that *areas* have to stay the same when we carry out a Lorentz transformation.

This makes it possible to give a very simple geometrical description of the Lorentz transformation. It's done by stretching one diagonal by a certain amount and squishing the other diagonal by the inverse of that factor, so that the total area stays the same. The figure shows the case where the velocity is 3/5 of c, which happens to give a stretch by a factor of 2 and a squish by 1/2.

3.5 No acceleration past c

The stretch and squish factors give us an easy way to visualize what happens if we try to keep accelerating an object past the speed of light. Since the numbers are simple for a velocity of 3/5, let's imagine that we take an object and accelerate it to 3/5 of c relative to the earth. The corresponding Lorentz transformation is stretched and squished by a factor of 2. Now let's say that we continue accelerating it until it's moving at an additional 3/5 of c relative to *that* motion. If velocities added the way Galileo imagined, then this would give 6/5 of c, which is greater than c. But what will actually happen here is that we'll have stretched the long diagonal by a factor of 4 and shrunk the short one by the same factor (left panel of the figure below). This is a Lorentz transformation that still corresponds to some velocity less than c. (It turns out to be 15/17 of c.) No matter how many times we

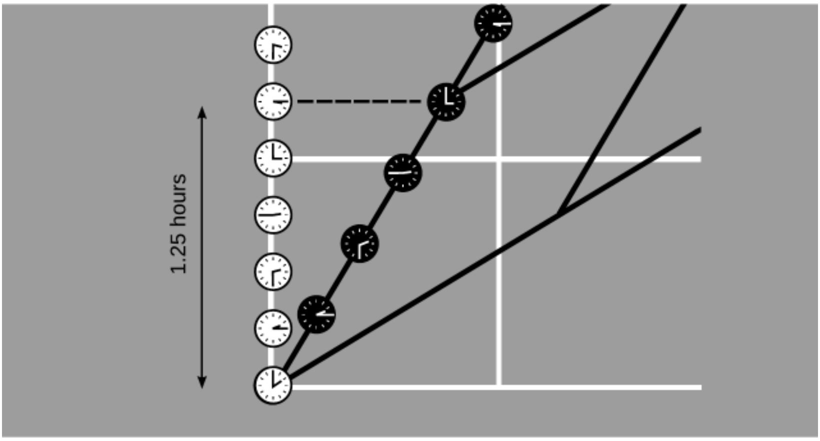


do this, the velocity will still be less than c. Thus no continuous process can accelerate an object past the speed of light, although this argument doesn't rule out a discontinuous process such as Star Trek's transporter.

By about 1930, particle accelerators had progressed to the point at which relativistic effects were routinely taken into account. In 1964, W. Bertozzi at MIT did a special-purpose experiment as an educational demonstration to test relativity using an electron accelerator. The accelerator was powerful enough to have accelerated the electrons to many times the speed of light, had Galileo been right. As shown in the right panel of the figure, the actual speeds measured just got closer and closer to c. The experimental data-points are in good agreement with the dashed line, which shows the predictions of special relativity.

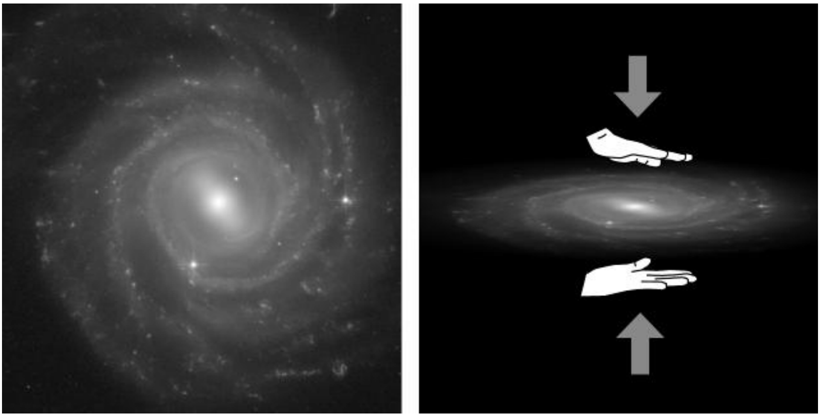
3.6 Time dilation

The figure shows a close-up of a Lorentz-transformation graph, covering just a little more than one unit of time — one hour. The white and black clocks are in motion relative to one another. If the observer in the white frame sends signals back and forth and goes through the process of Einstein synchronization (dashed line), she finds out that at the moment when the black clock



shows one hour of elapsed time, her own clock has measured 1.25 hours. (This is again for our favorite numerical example where the velocity is 3/5 of c.) This ratio of 1.25, or 5/4, is the factor by which each of the two observers says that the other one's time has slowed down. There is a mathematical symbol for this ratio, and a formula that would allow us to plug in the 3/5 and get the 5/4 out, but the focus of this book isn't on that kind of numerical calculation. For our purposes, drawing the graph is just as reasonable a way of working out the result.

In this example, the velocity was moderate compared to c, and the time-dilation factor was moderately big. For velocities small compared to c, the correspondence principle tells us that time-dilation factor must get close to 1, i.e., the effect goes away. As the velocity gets closer and closer to c, the time-dilation factor blows up to infinity. This is the scientific background for science fiction stories such as the Planet of the Apes movies, in which thousands of years pass back home while astronauts aboard a spaceship experience only a few months. This kind of motion very close to the speed of light is referred to as *ultrarelativistic.* In theory, if we had enough energy, we could accelerate ourselves to very close to the speed of light and watch the stars grow cold and the cosmos slip into senescence.



3.7 Length contraction

A science fiction story with better dramatic possibilities might be one in which our protagonists zoom across the galaxy and have various adventures, all within a human lifetime. Our galaxy is about 10,000 light-years in size, so since the ship can't go faster than c, the voyage would take at least 10,000 years according to people back on earth. But due to time dilation, this time interval would seem shorter to the travelers. If they moved at ultrarelativistic speeds, the time dilation could be extreme enough so that they could live long enough to complete the trip within their lifetimes.

But what does this all look like according to the voyagers? To them, their own time seems normal, and it's the people back on earth who are slowed down. So how can they explain the fact that they traveled 10,000 light-years such a short time? The answer is that the distance doesn't seem like 10,000 light-years to them. The Lorentz transformation describes distortions of both time and space. To the astronauts, the galaxy appears foreshortened, as suggested by the figure. This shortening of distances, called length contraction, is by the same ratio as the time dilation factor. Summarizing these results, we have the following:

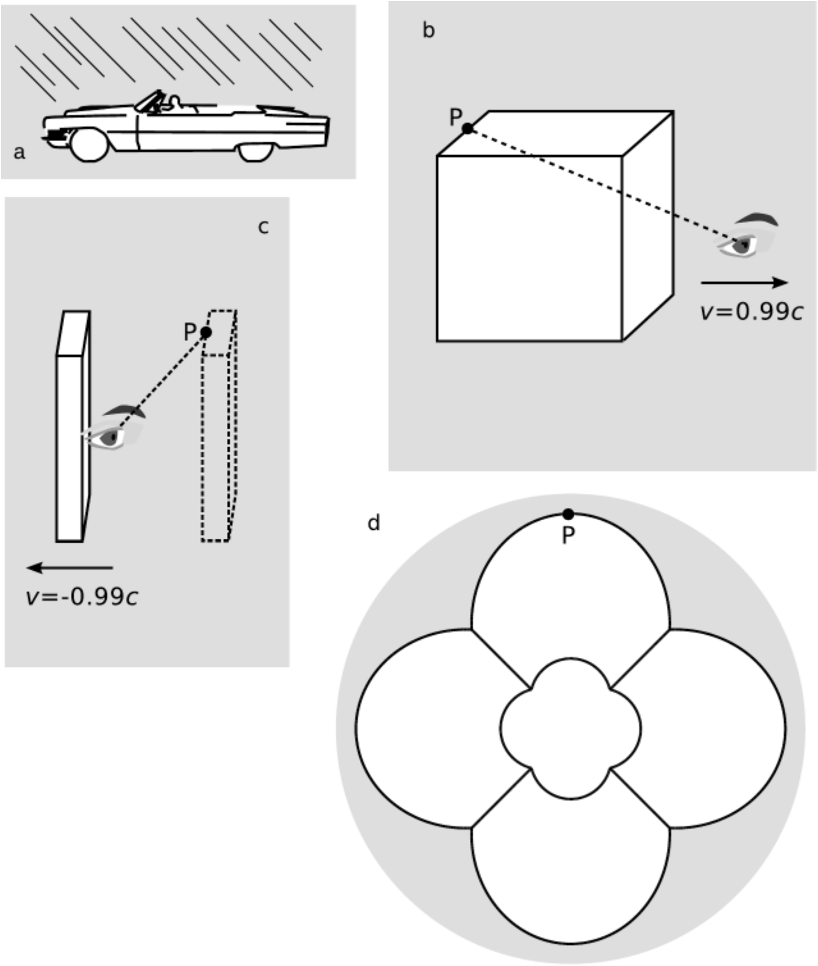
Time dilation
A clock appears to run fastest to an observer at rest relative to the clock.

Length contraction
A meter stick appears longest to an observer who is at rest relative to it.

3.8 Not what you see

By the way, you shouldn't get the impression from words like "appears" that these are the only effects that one actually *sees* with optical observations using the eye or a camera. What we mean here is what an observer finds out from a procedure such as Einstein synchronization, which involves sending signals back and forth, completing a surveying process, and working backward to eliminate the delays due to the time the signals took to propagate. In an actual optical measurement, these delays are present, and they have an effect on what we see.

Furthermore, the rays of light that come to us suffer an effect called relativistic aberration, which makes them seem like they're coming from a different direction than they really are. To understand this effect, imagine, as in panel a of the figure on the following page, that you're riding in a convertible with the top down, at 40 km/hr. Meanwhile, rain is falling straight down at 40 km/hr, as measured in the frame of reference of the sidewalk. But in your frame, the rain appears to come down at a 45-degree angle, not from straight overhead. Relativistic aberration is usually a small effect because we don't move eyes or cameras at velocities that are comparable to the speed of light. However, the effect does exist, and astronomers routinely take it into account when they aim telescopes at high magnification, because the telescope is being carried along by the motion of the earth's surface.


a
b
P
v=0.99c
c
P
v=-0.99c
d
P

Panels b-d of the same figure show a visualization for an observer flying through a cube at 99% of the speed of light. In b, the cube is shown in its own rest frame, and the observer has already passed through. The dashed line is a ray of light that travels from point P to the observer, and in this frame it appears as though the ray, arriving from an angle behind the observer's head, would not make it into her eye. But in the observer's frame, c, the ray is at a forward angle, so it actually does fall within her field of view. The cube is length-contracted by a factor about 7. The ray was emitted earlier, when the cube was out in front of the observer, at the position shown by the dashed outline.

The image seen by the observer is shown in panel d. Note that the relativistic length contraction is not at all what an observer *sees* optically. The optical observation is influenced by length contraction, but also by aberration and by the time it takes for light to propagate to the observer. The time of propagation is different for different parts of the cube, so in the observer's frame, c, rays from different points had to be emitted when the cube was at different points in its motion, if those rays were to reach the eye.

3.9 Causality

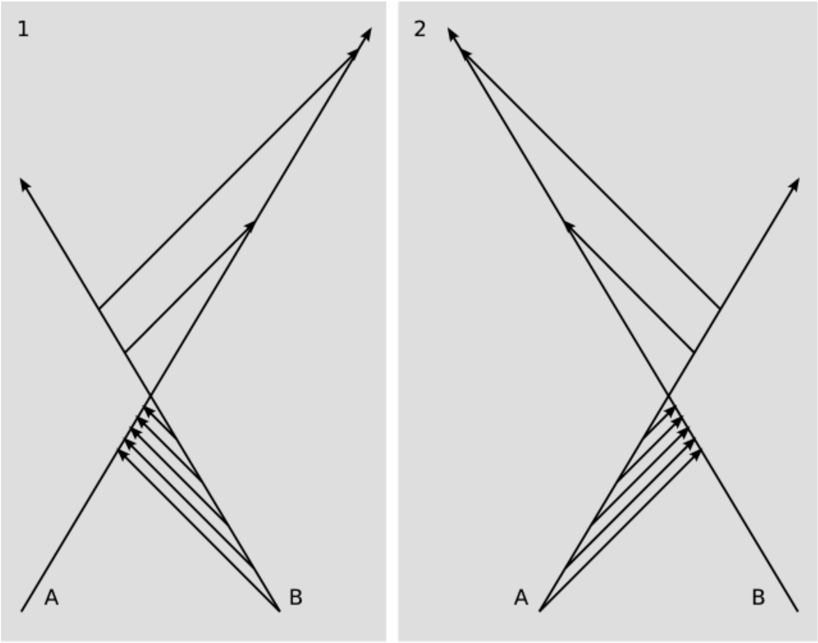
Although time dilation is a kind of time travel, it's always time travel into the future, so special relativity avoids scenarios such as the Robert Heinlein story "— All You Zombies —," in which a woman has a sex-change operation, goes back in time, has sex with a woman who he doesn't realize is his younger self, and becomes the father of a baby who turns out to be himself. If event A causes a later event B in a certain frame of reference, then B must be inside A's future light cone. The Lorentz transformation doesn't change the light cone, so B is guaranteed to be later than A in all other frames of reference as well.

3.10 The stretch factor is the Doppler shift

If you hear little kids playing with toy race cars, they'll often imitate the sound they've heard on TV as the cars go by: the buzz of the engine starts out a little higher in pitch as it approaches, then shifts a little lower as it goes away. This is called a Doppler shift. The size of the effect depends on how fast the car is moving compared to the speed of sound. We don't notice Doppler shifts as much for ordinary cars, because their speeds are much smaller compared to the speed of sound. The revolutions of the spinning motor are like the ticks of a clock. The rate at which these sound impulses arrive at the ear of a stationary observer is distorted because between the emission of one impulse and the next, the distance to the ear has changed.

We can also observe Doppler shifts with light rather than sound. The concept turns out to be very closely related to the situation illustrated in the figure on p. 59. This was a spacetime diagram for a situation in which Alice and Betty, flying away from each other in spaceships, try and fail to establish whose time is really slow and whose fast. In that experiment, Alice clapped her hands twice, one second apart, next to her cell phone, and Betty heard the hand claps more than one second apart. Although Betty might conclude from this that Alice's time is slow, the situation is symmetric, and Alice gets the same results if Betty sends hand-claps to her.

The figure on the facing page shows what happens if we *repeat* the hand claps over and over. The claps begin while Alice and Betty are approaching one another, and continue after they do their flyby and recede. While they're approaching, the hand claps are received closer together in time, and this can be interpreted as a combination of two effects: their earlier and earlier arrival as the distance shrinks, and a time dilation effect, which only partially counteracts the distance effect. As they recede, the distance effect causes the time intervals to be wider, and the time dilation effect cooperates with the diatance effect to make the intervals even longer.



But now suppose that the signal is not a series of claps but simply a radio wave. We previously thought of the diagonal world-lines as the hand-clap signals traveling through space. Now we can think of them as the peaks of the wave traveling through space. During the receding part of the motion, the peaks are farther apart in time when received, so the receiver perceives the wave as having slower vibrations (a lower frequency). This is similar to the Doppler shift of the race car's sound, but what matters now is how fast the ships are traveling compared to the speed of *light*.

The mathematical description of this type of Doppler shift is very simple. The motion of A relative to B is at some speed, and at this speed the Lorentz transformation has a certain stretch and squish factor (section 3.4, p. 50). Suppose that the stretch and squish factors are 2 and 1/2. Then the Doppler shift changes the rate at which the signals are received by a factor of 2 if A and

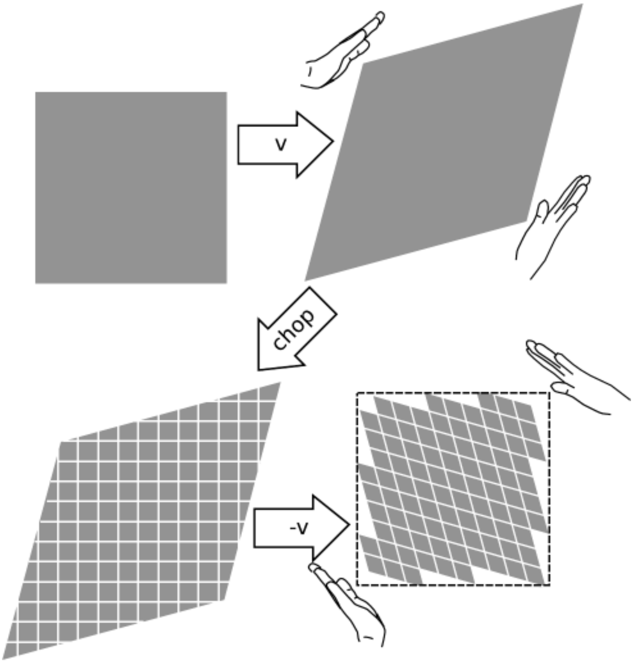
B are approaching one another, or by 1/2 if they're receding.

Because the velocities in everyday life are so small compared to the speed of light, the Doppler shifts are very small. For example, if a jet plane is flying directly toward us at 300 meters per second, the light that comes from it to our eyes is shifted up in frequency by a factor of 1.000001, which is only a ten thousandth of a percent, so although the light is shifted a little bit toward the blue end of the spectrum, the change isn't noticeable. As it recedes from us, the frequencies get multiplied by 0.999999, which is a drop of a ten thousandth of a percent, and we have a tiny shift toward the red.

But there is at least one example in everyday life where Doppler shifts matter. When you use GPS, your receiver is getting signals from satellites that are moving about ten times faster than a jet plane. Although this is still very small compared to the speed of light, GPS depends on extremely accurate time measurements, so the Doppler shifts are big enough so that if we didn't take them into account, the system wouldn't work at all.

Because the Lorentz transformation preserves the area of a square as it distorts it into a parallelogram, the stretch and shift factors are always inverses of one another: 2 and 1/2, 7 and 1/7, and so on. This exact relationship would not be true without relativity, so testing it provides a test of relativity. Physicists using specialized equipment can measure Doppler shifts extremely accurately, so these tests are among the strictest to which anyone has ever subjected the theory. The first such test was done by Ives and Stilwell in 1938, and the results were as predicted by relativity. A modern, more precise version by Saathoff *et al.* in 2003 used lithium ions moving at 6.4% of c in a particle accelerator. The forward and backward Doppler shifts were measured to be 1.06592034 and 0.93815641, respectively, and if you multiply these two numbers together, you'll see that the product equals 1 to nine digits of precision, which is impressive by anyone's standards. The spectacular agreement with theory has made this experiment a lightning rod for anti-relativity kooks.

As we'll see in chapter 13, p. 165, Doppler shifts are the



method we use to measure the motion of distant galaxies toward or away from us, and this was the technique by which we discovered that the universe was expanding.

OPTIONAL

3.11 Proof that area stays the same in a Lorentz transformation

This optional section gives a proof that Lorentz transformations don't change area in the $t - x$ plane

We first subject the square in the figure to a transformation with velocity v, and this increases its area by a factor F, which we want to prove equals 1. We chop the resulting parallelogram up into little squares and finally apply a $-v$ transformation, i.e., a Lorentz transformation with the same speed but in the opposite direction. This changes each little square's area by a factor G, so that the whole figure's area is also scaled by G. The final result is to restore the square to its original shape and area, so $FG = 1$. But we must have $F = G$, because space is symmetrical, and we can't treat one direction differently than theother. Since F and G are both positive, they must both equal 1, which was what we wanted to prove.

Chapter 4

The measure of all things

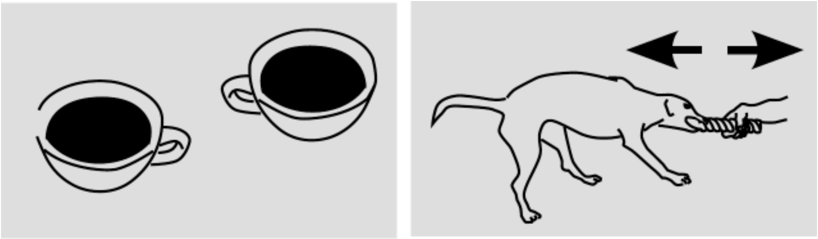
4.1 Operationalism

In college, I was a geeky teenager who was high on math and physics. My father and stepmother lived close enough that I could stop by for dinner on the weekend, and often when I showed up my stepmother would sit me down in their tiny kitchen, pour glasses of white wine for me and herself, and engage me in conversation. Having majored in English, with a lifelong interest in ballet, she must have decided to try to at least partially ungeekify me. On one of these occasions, I announced to her, trying to be interesting and provocative, that nothing really existed unless you could measure it. “What about love?” she asked. I can’t remember whether I replied that love could be measured or that love didn’t exist — probably the latter, since it would have sounded more provocative.

Despite my immature understanding of it, *operationalism*, developed by Bridgman around 1927, is a legitimate approach to understanding what we mean by the words we use to describe the world around us. The operationalist approach works like this. What is time? Time is what a clock measures. That is, measuring time requires carrying out certain measurement operations, and, according to Bridgman, “we mean by any concept nothing more than a set of operations; the concept is synonymous with the corresponding set of operations.” It may not work for love, but it does work pretty well for time.

4.2 Galilean scalars and vectors

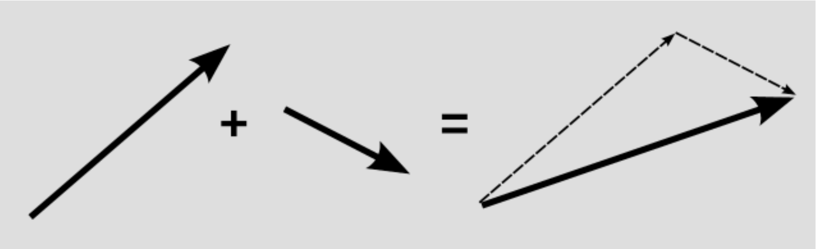
Most of the things we want to measure in physics fall into two categories, called vectors and scalars. In Galilean relativity, a scalar is something that doesn’t change when you turn it around, while a vector does change when you rotate it, and the way in which it changes is the same as the way in which a pointer such as a pencil or an arrow would change. For example, temperature is a scalar, because a hot cup of coffee doesn’t change its temperature when we turn it around. Force is a vector. When I play tug-of-war with my dog, her force and mine are the same in strength, but they’re in opposite directions. If we swap positions, our forces reverse their directions, just as an arrow would.



The simplest example of a Galilean vector is a motion from one place to another, called a displacement vector.

I chose the coffee and tug-of-war examples because they were intuitively appealing, but we don’t have to depend on intuition to check that they’re right. For example, temperature can be defined operationally as what a thermometer measures, so one can actually put a thermometer in a cup of coffee, rotate the cup, and check that the thermometer reading stays the same.

Scalars are just numbers, and we do arithmetic on them in the usual way. With vectors, the rules are a little different. In particular, there is a special method for adding vectors, as shown in the figure below. We represent them with two arrows, slide the arrows so that the tip of one is at the tail of the other, and then draw the result as the arrow that goes from the tail of the first vector to the tip of the second one. If we hike from A to B, and then from B to C, vector addition shows us how we could have achieved the same result by hiking in a straight line from A to C, without the detour to B. When we add two force vectors, we get a single force vector that would give the same result, if it acted on a certain object, as the two individual vectors would have given if they had both acted on that object.



A vector has a magnitude, which means its size, length, or amount. Rotating a vector can change the vector, but it will never change its magnitude. In fancy language, we say that the magnitude of a vector is *invariant* under rotations. A meter-stick is still a meter-stick, no matter how we turn it around.

4.3 Relativistic vectors and scalars

But in relativity, a meter-stick might not be a meter long. Its length could be Lorentz contracted. This is a sign that we need to rethink the idea of vectors and scalars a little in order to make them useful in relativity. In relativity we think of space and time as a single, unified stage on which events play out. The basic example of a vector is now a displacement in spacetime. Such a displacement could take me from here to Buenos Aires, or it could take me from now to next week. Imagine that you have a magic doorway that can take you to any time or place. You have to twiddle some knobs to tell it where and when. The settings on the knobs specify a relativistic displacement vector.

Such a vector could change not just because an observer viewed it from a different direction, but also because the observer was in a different state of motion. The direction change is a rotation. The change in motion is referred to as a "boost." A relativistic scalar is something that doesn't change under rotations or boosts. A good example of a relativistic scalar is electric charge, the property that makes your socks cling together when they come out of the dryer. Although time is a Galilean scalar, it's not a relativistic scalar, because when we do a boost, a time interval can suffer time dilation. A time interval by itself is neither a relativistic vector nor a relativistic scalar; it could be *part* of the description of a spacetime displacement.

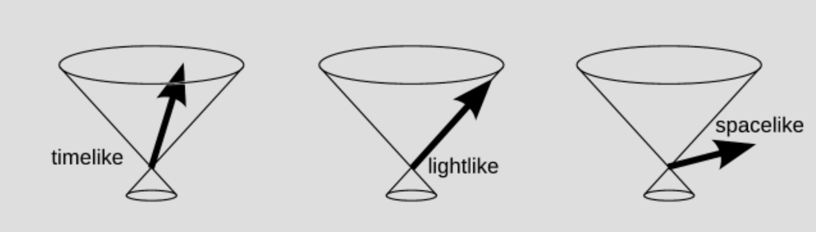
4.4 Measuring relativistic vectors

We would like to be able to measure relativistic vectors and describe their magnitudes in the same way we did with Galilean vectors. We can do that, because Einstein's relativity, unlike Galileo's, gives us a way of expressing times and distances in the same system of units.

An example is the relativistic displacement vector from your birth to your death. As you live your life, your body traces out a world-line through spacetime. You can't go faster than c, so for example the displacement from your birth to your death is guaranteed to have a time component that is bigger than its spatial part. If you live to be 80 years old, your place of death is guaranteed to be less than 80 light-years from your place of birth. (In the earth's frame of reference, it's much, much less.) In the terminology of section 2.6, p. 42, these events are therefore timelike in relation to each other. The vector connecting them is likewise called a timelike vector.

The magnitude of a timelike vector is defined as the time elapsed on a clock that moves inertially along that vector. This clock-time is called *proper time.* The word "proper" is used here in the somewhat archaic sense of "own" or "self," as in "The Vatican does not lie within Italy proper." Suppose that an alien living in a distant galaxy is observing you through a powerful telescope. Because the universe is expanding, she sees our galaxy as moving away from her at a significant fraction of the speed of light. Your hypothetical lifetime of 80 years is dilated as seen in her frame of reference, and may be 200 years to her. But if she knows your motion, she can correct for that, and infer that a clock moving along with you would only have registered an elapsed time of 80 years. You and the alien agree on the magnitude of your birth-to-death vector: it's 80 years.

If we imagine the alien as moving relative to us (and us relative to her) at a speed very, very close to c, the correction factor can get very large. It might take a trillion years of telescope observing for the alien to see you finish brushing your teeth this morning. If some thing, say a ray of light, moves at *exactly* c, then its proper time becomes zero for any observer. In this sense, a ray of light experiences no time at all. This type of displacement is lightlike, and its magnitude is defined to be zero, even if the vector itself isn't zero.



Spacelike vectors are of less interest because they can't trace out the world-line of any thing. Although it is possible to define a magnitude for a spacelike vector, we won't go into that now because it leads to complications involving plus and minus signs.

Part II

Matter and $E = mc^2$

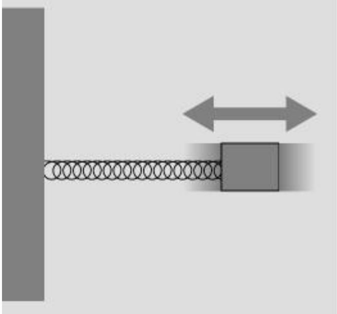
Chapter 5

Light and matter before Einstein

5.1 How much does matter really matter?

Everyone knows $E = mc^2$, even if they don't know what the letters stand for. You already know about c, so let's move on to m. It stands for mass, and most dictionaries define it as something like the quantity of matter. This definition falls into the category referred to by the physicist Wolfgang Pauli as "not even wrong," meaning that it doesn't even mean enough for us to decide whether it's right. The basic problem is that it's not an operational definition. As an operational definition, pick up the nearest handy heavy object, such as your backpack full of textbooks. Hold it firmly in both hands, and shake it back and forth. What you're feeling is a sensation of its mass. If you were in outer space, where there is no gravity, you'd get exactly the same results, because mass doesn't depend on gravity. This is in fact a technique sometimes used by astronauts to check on their body mass while they're in orbit. They sit in a chair attached to a spring, and measure how fast the vibrations are.

As an *interpretation* (not a definition), we can say that mass is a measure of how much inertia an object has: how hard it is



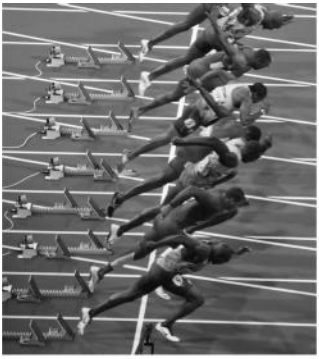
to change its state of motion.

To define a numerical scale of mass, we can just start with a standard object, and do this kind of vibration measurement on the standard object. (The metric kilogram unit, for example, is defined by a standard platinum-iridum cylinder housed in a shrine in Paris.) If the results are the same for the standard object as for our object of unknown mass, then the unknown has a mass of one unit. If three copies of the standard object bundled together give the same results as our unknown, then its mass is three units.

5.2 How hard is the hit, how loud is the bang?

Mass measures how hard it would be to change an object's motion, but we would also like to have some way to measure how much motion it actually *has*. We could measure its velocity, but that would give the same measure of "motion" to world-champion sprinter Usain Bolt as to my skinny little spaniel-greyhound mutt, who can run at about the same pace. Bolt is a big, burly guy, and if I had to choose whether to get hit by him or my dog, running at full speed, I'd choose the dog. What we're groping for here is something called *momentum*. As an operational definition, momentum can be defined by taking the moving object (say my dog), colliding it with a target (me), and seeing how fast the target recoils. If stopping Usain Bolt makes me recoil ten times faster than stopping my little dog, then Bolt has ten times more momentum. This definition is exactly the one used to measure momentum in particle-accelerator experiments. At nonrelativistic speeds (i.e., speeds much lower than c), experiments show that momentum is proportional to the object's mass multiplied by its velocity — but this is *not* the definition of momentum, as some people would have you think, and in fact is only an approximation for nonrelativistic speeds. In nonrelativistic physics, momentum is a vector.

One of the most fundamental known laws of physics is that momentum is *conserved*, meaning that the total amount of it always stays the same. For example, if my dog runs into me and I stop her motion, her momentum vanishes, but I recoil with the same amount of momentum, so that the same amount still exists. In the photo, the sprinters initially have zero momentum. Once they kick off from the starting blocks, they have momentum, but the planet earth recoils at a very small speed in the opposite direction, so that the total is still zero.



Energy and momentum.

Energy, but zero total momentum.

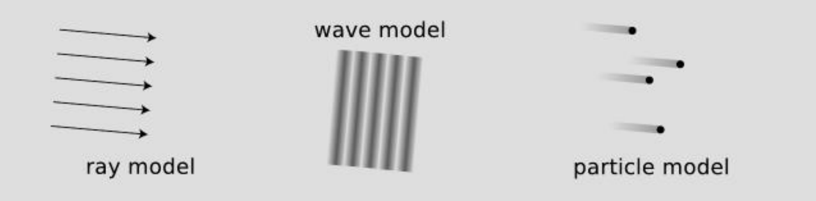
As a measure of "oomph," however, momentum can't be the whole story. The firework in the photo has pieces that move in all directions symmetrically. Since momentum is a vector, the momenta of the pieces cancel out. We want some other number that measures how much of a bang was stored up in the gunpowder and released as heat, light, sound, and motion when the gunpowder exploded. This is called *energy*, and it's the E in $E = mc^2$. In nonrelativistic physics, energy is conserved, and its conservation is essentially the only way of defining it operationally. For example, if we define a unit of energy as a certain amount of heat, then the energy content of a battery can be determined by using it to heat something. Whatever amount of energy was gained as heat, that same amount must have been lost by the battery.

In nonrelativistic physics, energy is a scalar. For example, if we rotate a car, there is no change in the amount of energy stored in its gas tank.

Kinetic energy is the term for the energy that a material object has because of its motion. Although I intend this to be a book with almost no equations, it will be helpful to introduce an equation for kinetic energy. Since conservation laws are additive, kinetic energy must be proportional to the mass of an object. For example, if two cars are moving at the same speed, but one has twice the mass, it must have twice the energy. Experiments at nonrelativistic speeds show that kinetic energy is also approximately proportional to the square of the velocity, i.e., the velocity multiplied by itself. So for example if we drive a car twice as fast, it has *four* times the kinetic energy. (This makes driving at high speeds more dangerous than people intuitively expect.) Finally, if we measure mass, velocity, and energy in their metric units of kilograms, meters per second, and joules, then there is also a conventional proportionality factor of 1/2. Putting all the factors together, we have the formula $K = (1/2)mv^2$. The presence of the square is also connected to the fact that energy is a scalar. If we pick a certain direction and call that positive, then velocities in the opposite direction are considered negative. But driving a car in the opposite direction doesn't negate its energy — if it did, you could fill your gas tank back up by driving that way! Mathematically, when we multiply a negative number by itself, we get a positive result, so the kinetic energy comes out the same.

5.3 Light

As a physics student, Einstein was taught that the universe consisted of two ingredients: matter and light. Because we can't touch and feel light, and can't manipulate it with our hands, its properties can be hard to pin down. But it often suffices to work with some *model* of light. The ray model is the simplest. Rays of light were probably one of the things Euclid had in mind when he imagined perfectly straight lines. The wave model, discussed in section 12.3, p. 158, has the advantage of providing a straightforward explanation of color: the different rainbow colors are different wavelengths of light. The particle model seems to contradict the wave model, but in fact can be reconciled with it; today light is believed to have both wave and particle properties.



In the ray model of light, the only way your eyes can see anything is if a ray goes into your eye. When you look directly at a luminous object such as a candle flame, rays travel straight from the flame to your eye. When you look at a nonluminous object such as your hand, rays from a source of light such as the sun hit your hand, are reflected, and then enter your eye.

Although physicists had good descriptions of both light and matter before Einstein, there were subtle contradictions between them, which bothered him as a young student. In the following chapter this will lead us directly to the famous $E = mc^2$.

Chapter 6

$E = mc^2$

We now trace essentially the same chain of reasoning that Einstein used in arriving at his most famous equation.

6.1 Energy shift of a beam of light

A ray of light always moves at c, and this leads to some big differences in behavior between the energy of matter and the energy of light. If a boxer throws a harder punch, his fist has more energy because of its greater speed. But we can't make a flashlight brighter by making its beam come out faster.

This leads to the following relativistic question, which turned out to be crucial to Einstein. Suppose we move toward an oncoming beam of light at a significant fraction of c, as in a head-on collision between two cars. What happens to the beam's energy when we switch to this frame of reference?

Since its speed is the same in the new frame, we might imagine that its energy was unchanged as well. After all, if you don't change the speed of a baseball or a bullet, you don't change its energy. But this can't be right, for the following reason. The law of conservation of energy is delicately constructed so that if energy is conserved in one frame of reference, it's conserved in other frames as well. In a typical energy-transformation process, such as the emission of a beam of light by a flashlight, we have a bunch of different forms of energy before the transformation, and a similar list afterwards. The sum of all the initial forms of energy has to equal the sum of all the final forms. If we switch to a different frame of reference, the numerical values of many of these numbers change. (For example, if the flashlight was at rest in the original frame, with zero kinetic energy, then in the other frame, where it's moving at a substantial fraction of c, it has more energy than a nuclear warhead.) If energy is to be conserved in the new frame as well, then all of the various forms of energy must be different in the new frame, and the differences must be such that the books still balance. There is no way to make this work if we insist that a beam of light keep the *same* energy in all frames.

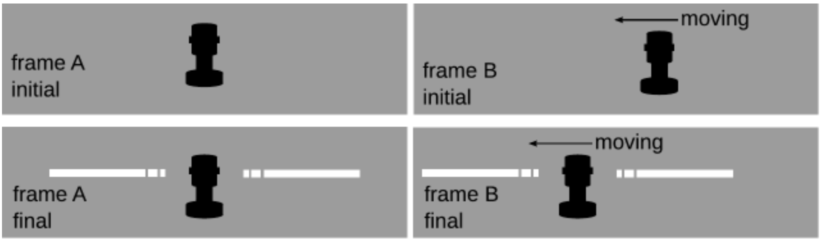
So a beam of light must be brighter to an observer rushing into the beam. How much brighter? The effect turns out to be that the energy gets multiplied by the stretch factor introduced in section 3.4, p. 50. For example, if the observer moves at 3/5 of c, then the stretch factor equals 2, and the beam has twice as much energy. This use of the stretch factor is pretty much the only possibility that makes sense, because we could have a chain of observers, with B in motion relative to A, and C in motion relative to B. When we combine velocities like this, the stretch factors multiply, but so do the intensity factors for a beam of light. For example, if B says the beam is twice as bright as A, and C says it's twice as bright as B, then C says it's four times as bright as A.

By the way, we do see this effect in real life. Because of the expansion of the universe, distant galaxies are moving away from us at speeds comparable to c. This makes their light dimmer and harder to detect through a telescope. (There is also a color shift toward the red end of the spectrum, which in the wave model of light is described as a lengthening of the wavelength. This wavelength effect is called the Doppler shift.)

6.2 Einstein's argument

We now have all the necessary ingredients to go through Einstein's argument leading to $E = mc^2$, and to interpret what the equation means. In Einstein's 1905 paper on this topic, he gave an argument using algebra, but I'll present it using a numerical example instead.

Suppose that in Alice's frame of reference A, we have a battery-powered lantern that is initially at rest. The lantern is set up so that it shoots one beam of light to the right, and simultaneously another beam of equal intensity to the left. We turn the lantern on and then back off, and each of the two escaping beams of light carries away 0.5 units of energy. Although the beams of light do have momentum, momentum is a vector, so the momenta in opposite directions cancel out, and the lantern doesn't recoil. Since the lantern is still at rest after the emission of the beams, its kinetic energy both before and after emission is zero. Therefore by conservation of energy, the $0.5 + 0.5 = 1$ unit of energy in the light must have come from the dissipation of 1 unit of energy that had been stored in the battery.



Now suppose that this same process is observed by Betty, who is moving to the right at 3/5 of c relative to Alice and the lantern. In Betty's frame B, the leftward beam is increased in energy by a factor of 2, while the rightward beam has its energy cut in half. Their energies in frame B are 0.25 and 1.0 units. But now we have a problem making the books balance on conservation of energy. The lantern has some kinetic energy, but we don't expect it to change its mass or velocity when the light is emitted, so we would think the kinetic energy would remain unchanged between the initial and final state, and we would be able to ignore it for purposes of conservation of energy. Initially, we have:

1 unit of energy stored in the battery

After emission, we have:

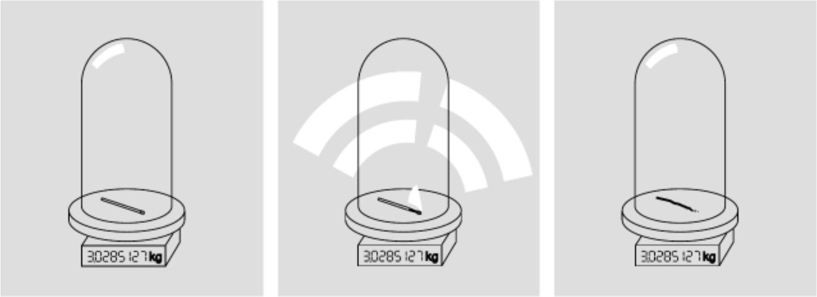
0 energy stored in the battery
+0.25 units of energy in the right-going beam
+1.00 units of energy in the left-going beam

We have a discrepancy of 0.25 units. If we don't patch things up somehow, conservation of energy will be violated. That is of course logically possible, and there were in fact a couple of occasions in the early twentieth century in which physicists did seriously consider abandoning energy conservation as a fundamental law of physics. Einstein decided to do something trickier. He decided to *redefine* energy. He proposed the following:

> **Equivalence of mass and energy**
> Mass and energy are equivalent, and the conversion between mass and energy units is given by the equation $E = mc^2$.

Let's see what happens when we apply this idea to the example of the lantern. Initially, the battery contains 1 unit of energy. This is equivalent to an amount of mass $m = E/c^2$. In metric units, where c is a huge number, this comes out tiny. For example, 1 joule of energy would be equivalent to about 0.00000000000000001 kilograms (or 10^{-17} kg in scientific notation). But when we switch to Betty's frame, the lantern is moving at a very high speed, so this miniscule amount of mass could end up contributing quite a bit of kinetic energy. How much extra kinetic energy do we get? We can't immediately answer this question because although we have a formula for kinetic energy (p. 74), that formula is only a nonrelativistic approximation. The speed of 3/5 of c that we're dealing with in our numerical example is pretty high, so we shouldn't expect the nonrelativistic formula to work exactly. If we nevertheless go ahead, take this mass and the velocity of $(3/5)c$, and plug them into the formula, we get an energy of 0.18 units, which is close to, but not exactly equal to, the 0.25 units that we need in order to make conservation of energy work. An advantage of Einstein's use of algebra over our numerical approach is that his method makes it easier to demonstrate that at low velocities, where the nonrelativistic formula for kinetic energy is known to be a good approximation, the discrepancy in conservation of energy is eliminated. It is eliminated at higher velocities as well, but only if we use the correct relativistic equation for kinetic energy.

6.3 Correspondence principle



The correspondence principle plays a crucial role here. We get to rewrite the rules, redefining energy, but our new rules have to be approximately consistent with the old ones in the situations where the old rules had already been checked by experiments. The figure shows an example. The match is lit inside the bell jar. It burns, and energy escapes from the jar in the form of light. After it stops burning, all the same atoms are still in the jar: none have entered or escaped. The burnt match is lighter, but only because some of its atoms are now in the form of gases such as carbon dioxide. The gases were trapped by the bell jar, so they still contribute to the total weight. The figure shows the outcome expected before relativity, which was that the mass measured on the balance would remain exactly the same. Mass is conserved.

Now Einstein claims that the energy of the light escaping from the jar is equivalent to some mass. How can this be? Experiments like this had been done by physicists as far back in history as Lavoisier, at the time of the French Revolution. But recall our numerical estimate above. A joule is actually a pretty reasonable estimate of the energy released by a burning match, and we saw that a joule is only equivalent to an extremely small amount of mass. Lavoisier was right, to within the precision of the balances he had available. He simply didn't have a scale that measured weights to seventeen decimal places.

6.4 Mass to energy and energy to mass

In Star Trek's transporter technology, people are converted to beams of energy and then reconstituted in another place. Before Einstein, mass and energy had been thought to be separately conserved, so this would be impossible. According to relativity, they are equivalent, and we only have conservation of the *total* amount of mass plus energy (with the two quantities converted into compatible units before addition). The sum is called mass-energy. It's as though we believed that Euros and Swiss Francs were completely different things, until one day we found out that there was a money changing shop that would let us trade in one for the other.

Although nobody has yet succeeded in teleporting an object, we do observe processes in nature that convert mass to energy and energy to mass. The most common examples involve the forms of radioactivity that are present naturally in our environment.

Some naturally radioactive substances in the earth produce positrons, which are like electrons but have the opposite electric charge. A form of antimatter, positrons annihilate with electrons to produce gamma rays, a form of short-wavelength light.

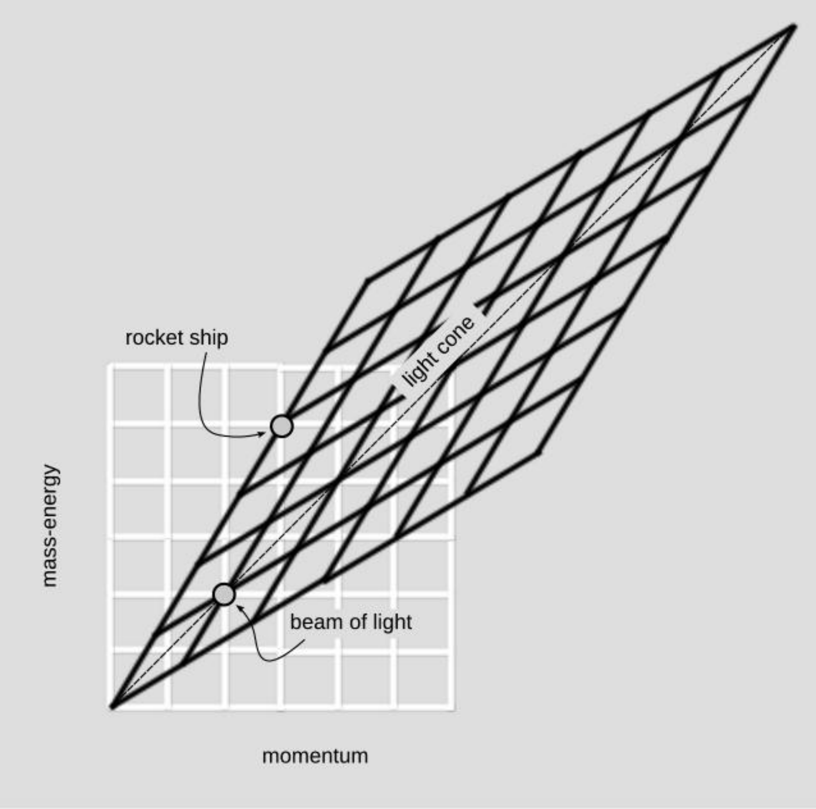
Positron annihilation forms the basis for the medical imag-

ing technique called a PET (positron emission tomography) scan, in which a positron-emitting chemical is injected into the patient and mapped by the emission of gamma rays from the parts of the body where it accumulates. Because the gamma rays fly off back to back, the detector can reconstruct the line along which they flew out and therefore the direction to the source of the radioactivity. In a typical application of PET, a patient with cancer is given a form of radioactive glucose, which emits positrons. The glucose tends to become concentrated in the cancerous tissue, allowing an image of the tumor to be reconstructed.

Conversion of energy to matter also occurs naturally. For example, we are naturally surrounded not just by radiation from minerals in the earth but also by radiation from outer space, called cosmic rays. Some of these cosmic rays are gamma rays, which can interact with matter by converting part of their energy into the creation of an electron and a positron.

6.5 The energy-momentum vector

We saw earlier that spatial displacement wasn't a valid relativistic vector, and to make a valid one, we had to combine it with a time interval. A similar approach works for mass-energy and momentum. To get a valid relativistic vector, we have to combine



mass-energy with momentum.

The figure shows how this plays out in a couple of examples involving different frames of reference. A point on the graph represents a thing with a certain energy and a certain momentum.

The lower dot represents a beam of light, and logically enough, it lies on the light cone. That is, its energy and momentum are numerically equal, if we choose units in which c equals 1. In terms of ordinary metric units, we have to include c as a conversion factor, and the momentum of a beam of light equals its energy divided by c. Because c is a big number in metric units, the result is a small momentum, and this is why a flashlight doesn't kick like a gun.

In the frame of reference represented by the white grid, the beam of light has an energy of 2 units and a momentum of 2 units. The black grid is the frame of reference of an observer chasing the beam at 3/5 of c. To this observer, the beam has energy and momentum both equal to 1.

The upper dot represents a rocket ship. The black grid represents the frame of reference of an observer aboard the ship. To that observer, the ship is at rest. By symmetry, its momentum can point neither to the right (positive) nor to the left (negative), so it must be zero, and zero is the value we read off of the graph. In the ship's own frame, it isn't moving, so it has zero kinetic energy. But its mass-energy isn't zero, it's 4 units. The ship's mass is 4 units, and in this frame that's the only kind of mass-energy it has.

In the white frame, the ship has 3 units of momentum and 5 units of energy. This energy is due to both the ship's 4 units of mass and an additional 1 unit of kinetic energy.

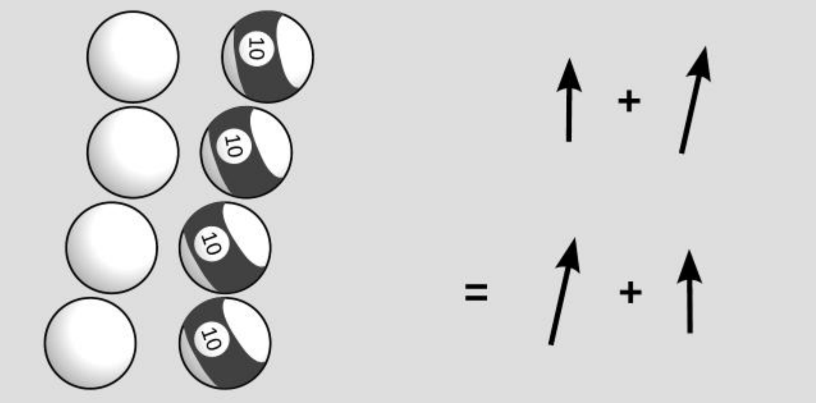
Since the ship's energy-momentum vector is timelike, by definition its magnitude equals its size as measured in the frame where it's at rest. Its magnitude is 4 units, which is its mass. The magnitude is the same in all frames of reference, so an observer in the white frame of reference agrees that its mass is 4. This is how we *define* mass in relativity: as the magnitude of the energy-momentum vector.

The beam of light has an energy-momentum vector that is lightlike, so its magnitude is zero. This is why we say that a beam of light has zero mass. Again, all observers will agree on this fact.

In general, matter is made out of particles whose energy-momentum vectors are timelike, because they have nonzero mass. Light, as well as certain more exotic creatures such as gluons and gravitational waves, has zero mass and therefore a timelike energy-momentum vector. A hypothetical third category of particles, called tachyons, would have a spacelike energy-momentum vector. Tachyons probably don't exist, which is a shame because the universe would be a wilder place if we had them. In section

6.6, p. 88, I'll say a little about searches for tachyons and what tachyons might be like.

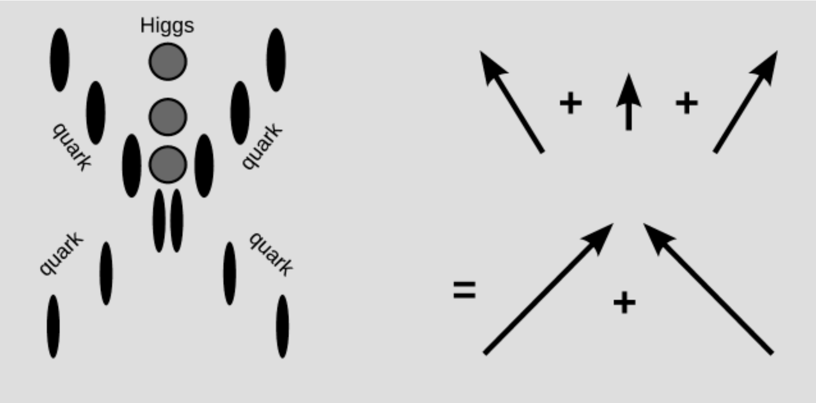
Money isn't much fun if you can't spend it, and energy-momentum isn't much fun if you don't ever change it from one form to another. The first figure below shows a collision between two pool balls, with a diagram showing how their energy-momentum vectors add up. As actually happens in this kind of head-on shot (unless you use a lot of spin on the cube-ball), the cue ball stops dead, transferring all of its motion to the ball it hits. (The diagram is actually a little unrealistic, because I've drawn the balls moving at about 20% of c, which would not only destroy them on impact but also kill everyone within a considerable radius.)



One way to check that energy-momentum is conserved here is to set up each pair of vectors tip-to-tail as in the method of vector addition shown in section 4.2, p. 64. An easier way to verify it is that the two sides of the equation are the same except that the order of the things being added up has been reversed; vector addition doesn't care about the order, just as in ordinary addition, where $1 + 3$ is the same as $3 + 1$.

Most of what we see in this diagram is simply the mass of the balls. The moving ball does however have a little bit of energy (extra height) and some momentum (leaning to the right, so that its spacelike part isn't zero).

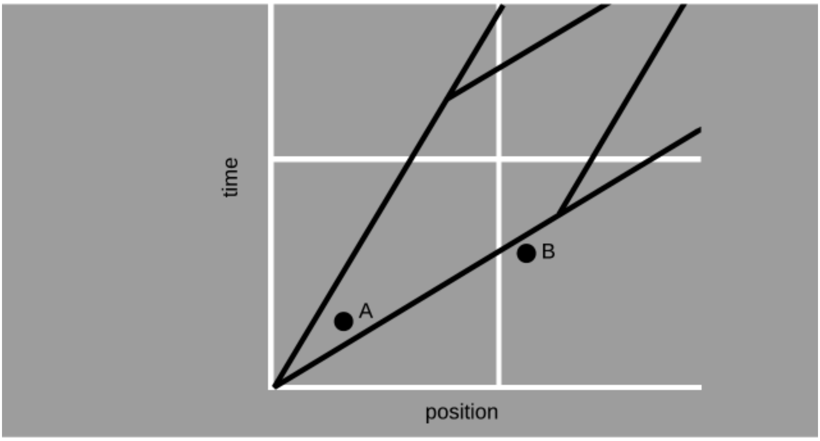
Now here's a diagram of a somewhat more realistic example, one that really requires relativity. This is a head-on collision of two subatomic particles called quarks, such as occurs at the Large Hadron Collider near Geneva, Switzerland. The quarks are actually parts of larger particles — protons — but on this scale we only see one quark from one proton colliding with one quark from another proton. We don't actually know if quarks are pointlike or have a finite size, but I've drawn them here as having some bulk to them, and as ovals rather than circles to represent the effect of Lorentz contraction. In this example, some of the energy of the incoming quarks happens to produce something called a Higgs particle. That is, energy has been converted into matter via $E = mc^2$. The Higgs particle was predicted to exist in the 1960s, and the LHC finally confirmed its detection in 2012. When the LHC was being designed, it was expected to be almost certain to find the Higgs. In fact, the fear was that the Higgs was the only thing it would find at all, and that appears to be what has happened. If so, then many particle physicists will be very depressed, but no doubt they will use it as a justification for another project, even larger and more expensive than the LHC.



6.6 Why can't I have tachyons?

The energy-momentum vector is timelike for matter and lightlike for massless phenomena such as light. A hypothetical third category of particles, called tachyons ("TACK-y-ons"), would have a spacelike energy-momentum vector. They would always go *faster* than c, and their mass would be an imaginary number, like the square root of -1.

In section 3.5, p. 51, we saw both theoretical and experimental evidence that no continuous process of acceleration can boost a material object past c. That didn't, however, address the question of whether one could surpass c through some discontinuous process, such as the "jump to hyperspace" in Star Wars. This loophole now appears to be closed off. We observe that mass is a permanent, fixed property of a material object, and it therefore seems that a material object like the Millenium Falcon could not go faster than c, or else it would have to be transformed into a cloud of tachyons.



But that doesn't mean that tachyons don't exist. If they did, it would be cool. Consider the figure above. In the frame of reference represented by the white grid, someone uses a burst of tachyons to send a signal from event A to event B. According to this observer, B is in A's future, but so far away that there would have been no way for anything to get from A to B except by going faster than c.

But according to the observer represented by the black grid, B is in A's past, since B happens at a time less than zero (i.e., before the time arbitrarily chosen as a reference), while A happens at a time greater than zero.

This means that if tachyons exist, we can automatically use them to send signals back in time. (Personally, I'd tell myself to sell all the stocks in my retirement account in 2007, then buy in 2009.)

But do they exist? This is a question that can only be answered by searching experimentally. The most obvious experimental signature of tachyons would be motion at speeds greater than c. Negative results were reported by Murthy and later in 1988 by Clay, who studied showers of particles created in the earth's atmosphere by cosmic rays, looking for precursor particles that arrived before the first gamma rays.

One could also look for particles with spacelike energy-momentum vectors. Alvager and Erman, in a 1965 experiment, studied the radioactive decay of thulium-170, and found that no such particles were emitted at the level of 1 per 10,000 decays.

Some known subatomic particles, such as neutrinos, don't interact strongly with matter, and are therefore difficult to detect directly. It's possible that tachyons exist but don't interact strongly with matter. If so, then it might be possible to infer their existence indirectly through missing energy-momentum in nuclear reactions. This is how the neutrino was first discovered. An accelerator experiment by Baltay in 1970 searched for reactions in which the missing energy-momentum was spacelike, and found no such events. They put an upper limit of 1 in 1,000 on the probability of such reactions under their experimental conditions.

For a long time after the discovery of the neutrino, very little was known about its mass, so it was consistent with the experimental evidence to imagine that one or more species of neutrinos were tachyons, and Chodos *et al.* made such speculations in 1985. A brief flurry of reawakened interest in tachyons was occasioned by a 2011 debacle in which the particle-physics experiment OPERA mistakenly reported faster-than-light propagation of neutrinos; the anomaly was later found to be the result of a loose connection on a fiber-optic cable plus a miscalibrated oscillator. An experiment called KATRIN, currently nearing the start of operation at Karlsruhe, will provide the first direct measurement of the mass of the neutrino, by measuring very precisely the missing energy-momentum in the decay of hydrogen-3.

Part III

Gravity

Chapter 7

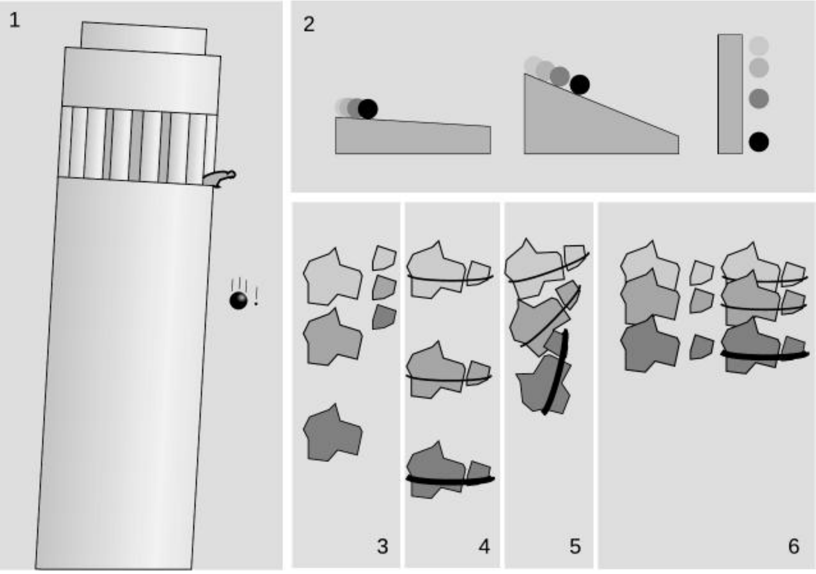
Newtonian gravity

7.1 Galileo and free fall

Galileo had an astounding knack for zeroing in on scientific issues that we still consider crucial four centuries later. Special relativity can be traced directly back to his insistence on the principle of inertia, as opposed to the Aristotelian idea that motion naturally stops if no force is applied. A similar Galileo-Aristotle controversy holds within it the seed of general relativity.

Aristotle thought that an object falling in air was analogous to one sinking in water, and claimed that objects fell at a certain speed, which would be proportional to the object's weight and would also depend on both the object and the medium through which it was falling. In some ways this is a reasonable description, but it's very complicated and in some ways inaccurate. It's both simpler and more fundamentally important to consider the case in which the frictional force from the surrounding medium is negligible, which is what happens when the density of the medium is low, and the falling object is relatively dense and heavy. You might want to give this a try yourself right now. You probably have some handy objects of contrasting weight, such as a coin and a shoe. Stand up and release them side by side.

My own shoe is about 50 times heavier than the nickel I had handy, but it looks to me like they hit the ground at exactly the same moment. So much for Aristotle! Galileo, who had a flair for the theatrical, did the experiment by dropping a bullet and a heavy cannonball from a tall tower[1] (panel 1 of the figure).



Nor does a given object fall with a single definite speed. Starting from rest, it speeds up and speeds up. The ingenious Galileo, who had only primitive clocks, convinced himself of this by doing experiments with balls rolling down inclined planes. By extrapolation, we expect the same to hold when the plane is tipped all the way up to the vertical, so that the object is falling freely (panel 2). In modern units, it turns out that objects on earth gain very nearly 10 meters per second (m/s) of speed in every second, so that if we release a rock at rest (0 m/s), then after one second of falling its speed is 10 m/s, after two seconds 20 m/s, and so on. This rate of speeding up is called the rock's *acceleration*, and it is 10 meters per second every second, i.e., 10

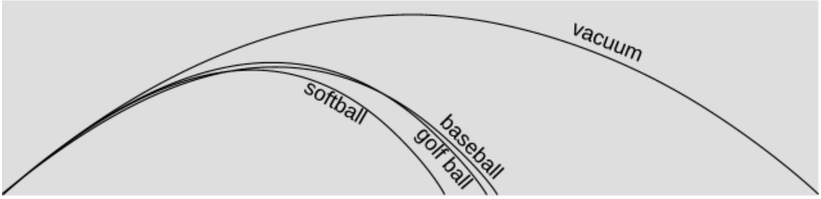
[1] His contemporary and first biographer Viviani says that it was the Leaning Tower of Pisa, although some historians are skeptical about this.

meters per second per second, also written as 10 m/s^2.

Not only did Galileo disprove Aristotle empirically, but he also pointed out a logical contradiction in Aristotle's own reasoning. Aristotle said that heavier objects fell faster than lighter ones (panel 3). If two rocks are tied together, that makes an extra-heavy rock, which should fall faster, 4. But Aristotle's theory would also predict that the light rock would hold back the heavy rock, resulting in a slower fall, 5. In reality, we find that all of these objects fall with the *same* acceleration, 6.

This universality of free fall turns out to have surprisingly far-reaching consequences. Eventually, starting from little more than this fact, we will see that gravity can be described as not being a force at all, but rather a warping or *curvature* of spacetime.

Free fall includes more possibilities than falling straight down. If we watch Galileo's experiment in a frame of reference moving vertically or horizontally, we may see the balls as going up rather than down, or flying along curved arcs. Therefore projectiles must also have universal motion, in the absence of friction. The figure below shows computer simulations of the trajectories of a golf ball, baseball, and softball, all shot at the same initial speed and angle. (The baseball hit in this simulation goes a distance of 200 meters, roughly a home-run distance.) Although the trajectories of the three types of balls are not the same in air, their trajectories would be exactly the same (and considerably longer) in vacuum.



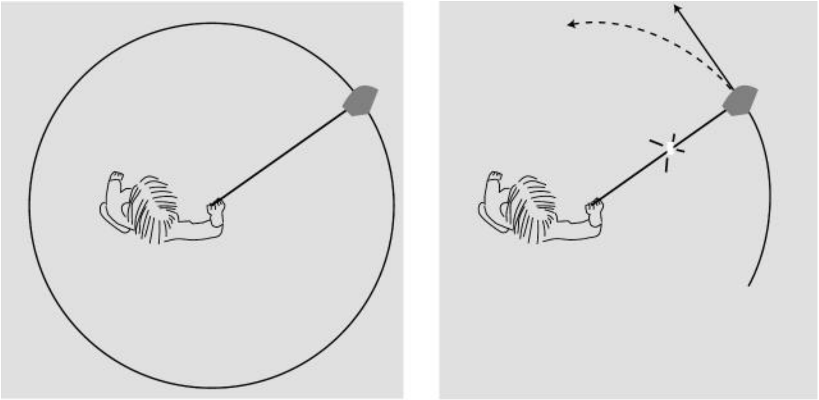
The universality of free fall has been tested in many different experiments over the years. In a silly educational demonstration, Apollo 15 astronaut David Scott dropped a feather and a geologist's rock hammer side by side. Online, you can watch video footage in which the two objects fall to the lunar surface side by side. One can vary not just the weight but also the composition of the falling body, verifying, for example, that a brass ball and a lead ball fall with the same motion. It was realized early on — by Galileo himself, in fact — that the same fundamental issue could be probed by experiments that appeared superficially very different and that allowed higher precision. For example, if gravity produces different accelerations in free-falling balls made of brass and lead, then pendulums with bobs made of brass and lead should also oscillate at different rates. Experiments of this type were done by Eötvos around 1900, and later refined by Braginskii, Dicke, and their co-workers to give a precision of about one part in 100 billion. Attempts are even under way to test whether the universality extends to antimatter. It's conceivable that antimatter is actually *repelled* by the earth, so that it would fall *up*, like the imaginary "upsydaisium" from the Rocky and Bullwinkle cartoon series.

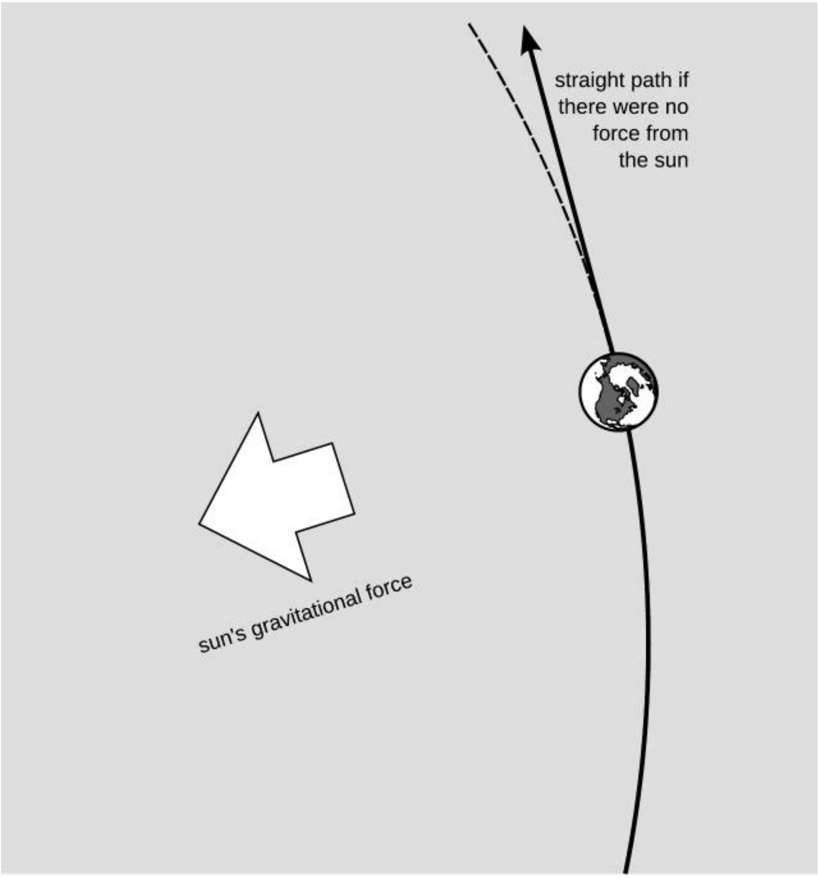
7.2 Orbits

You can count yourself lucky to live in a universe where free fall seems to be universal. At this moment, both you and the planet earth are in a state of free fall, being pulled in by the sun's gravitational attraction. As a result, Aristotle would have expected the earth to accelerate toward the sun much more violently than your body, with the result that the moment you stepped outside, the world would fall out from under you and you would go flying off into outer space. Ouch!

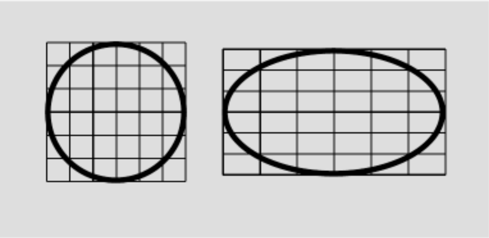
The earth revolves around the sun in an approximately circular orbit. On p. 19 we saw that circular motion requires a force; in the figure on the facing page, if the string breaks the rock will fly off in a straight line. Similarly, the sun's gravitational force is what keeps the earth from zooming away.

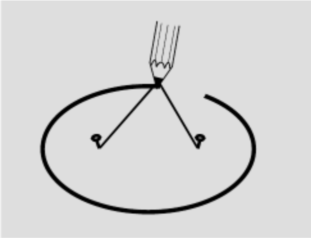
Understanding the motion of the planets was originally considered important partly because in ancient Europe, one of the




straight path if there were no force from the sun
sun's gravitational force

best ways to make a living doing astronomy was to cast horoscopes — at the time, there was no clear distinction between the science of astronomy and the pseudoscience of astrology. After the Copernican-Galilean revolution (see section 2.7, p. 44), the planets were described as revolving around the sun, and the observational astronomer Johannes Kepler (1571-1630) arrived at a complete and fairly accurate mathematical characterization of their motion in terms of three laws. Two of these laws describe the speeds of the planets' motion, while the remaining one states that each planet's orbit is an ellipse, with the sun lying at a certain well-defined off-center point called a focus. One way of defining an ellipse is that it's a circle that has been distorted by shrinking and stretching along perpendicular axes. An ellipse can also be constructed by tying a string to two pins and drawing a curve as shown in the figure, with the pencil stretching the string taut. Each pin constitutes one focus of the ellipse. Just as a square is a special type of rectangle, a circle is a special case of the ellipse. In a circle, the two foci coincide.

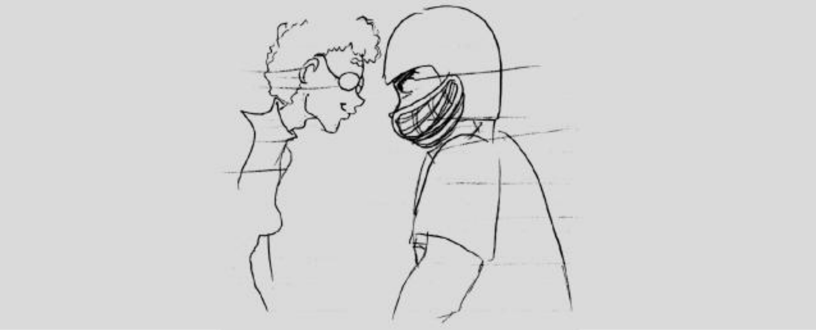




7.3 Newton's laws

Newton firmed up Galileo's work on motion by defining a set of natural laws that applied both to objects on earth and to the heavens — a radical idea at the time. We'll soon see that these laws give a simple and natural explanation for the universality of free fall. Newton was also able to use his laws to prove all three of Kepler's laws for the motion of the planets, although this proof is a longer and more mathematical, and we won't go into it here.

In Newton's description, objects push and pull on each other through attractive and repulsive forces. Gravity is a specific type of force, a universal attraction between any two objects that is proportional to their masses. When objects interact through a force, each one's momentum is changed — in fact, we can *define* force as the rate of change of momentum. The changes in momentum are like transferring money from one bank account to another; the total amount of momentum is conserved.



When an object has more inertia — more mass — it takes less motion to express the same amount of momentum. For example, if I throw a golf ball at an empty can, the can recoils violently, while the same hit on an aircraft carrier produces an imperceptibly small recoil. The momentum absorbed by the can and the aircraft carrier are the same. In the figure, a professional football player collides head-on with a little old lady. Momentum is conserved, so whatever momentum she gains is just enough to balance out what he loses. Therefore their forces on each other are equal. But the effects on their motion are not equal. Her mass is much less, so the change in her motion is much more violent.

In summary, Newton's description of motion consists of the following three laws:

1. If no force acts on an object, its momentum doesn't change.
2. If forces act on an object, its momentum changes at a rate given by the total force.
3. Forces always occur in pairs. If object A exerts a force on object B, then B is also exerting a force on A. These forces are in opposite directions and have equal magnitudes.

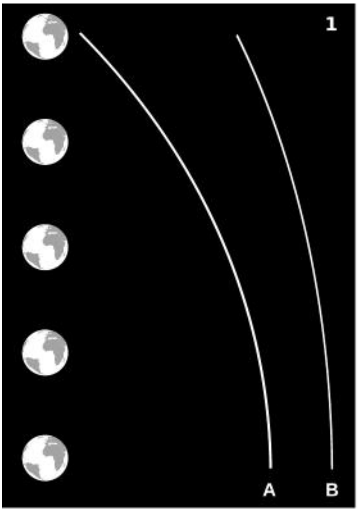
Newton's first law is a special case of the second law. The third law is essentially a statement of conservation of momentum; the rate at which A loses momentum to B equals the rate at which B gains it from A.

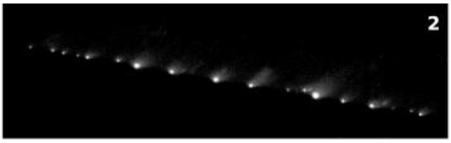
Newton's laws give an explanation for the universality of free fall. If my shoe is 50 times more massive than a nickel, then the earth's gravitational force on the shoe is 50 times greater. But the shoe's inertia is also 50 times greater, so that when I drop shoe and coin together, their responses to the force are exactly the same. But Newton's explanation is a sort of non-explanation explanation. It depends on the fact that an object's inertia is proportional to its gravitational mass. That is, we have two separate concepts of mass, and Newton can't tell us why they're equal. He just observes that they are. When I shake a lead brick back and forth, I sense its inertia, or inertial mass. When I put it on a scale, I measure its gravitational mass. There is no known reason why these two things *have* to be related; it just seems that they are. Besides gravity, we have other fundamental forces, such as electricity. These forces have no such property.

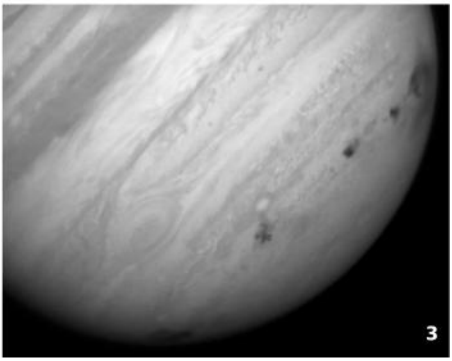
7.4 Tidal forces

If the shoe and the coin start out right next to each other, they stay right next to each other as they fall. A different situation is shown in figure 1. In this spacetime diagram, objects A and

B start out at unequal distances from the earth. Because A is closer, it feels a stronger gravitational force, so it accelerates more rapidly. The result is that A and B become more and more separated from each other.







Figures 2 and 3 show a spectacular example of this effect. In 1994, a comet named Shoemaker-Levy 9 approached the planet Jupiter on a collision course. A comet is a loosely packed dirty snowball, so there is not much holding it together. As this comet approached Jupiter, the small difference in Jupiter's gravity between the front and back of the comet was enough to pull it apart, and the fragments subsequently separated from each other even more, forming the string of pearls shown in figure 2. The fragments hit Jupiter in succession, and because Jupiter spins on its own axis, the impacts occurred at different places in the jovian cloud-tops. Figure 3 shows the resulting string of bruises, each about as big as the planet earth.

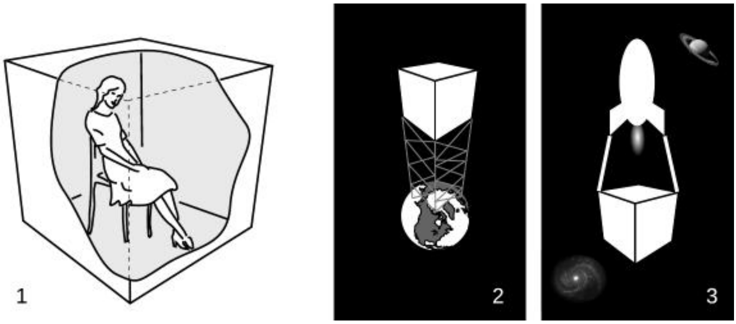
Chapter 8

The equivalence principle

8.1 The equivalence principle

The universality of free fall has the following very subversive consequences. Consider Alice, the girl in figure 1. She is shown in a cutaway view, sealed inside a well-lit and air-conditioned box, which happens to be her favorite place to go and think about physics. (It's small, but it's private and it's all hers.) She feels pressure from the seat of the chair, and this would normally suggest that she was experiencing a gravitational force, perhaps from a nearby planet (figure 2). But Alice can't infer this without peeking outside at her surroundings. It's equally possible that the box is in deep space (figure 3), with no gravity whatsoever, but is being towed by a rocket ship, so that it accelerates constantly. No possible experiment done inside the box can tell her which of these is the case.

For example, suppose that she drops a shoe and a coin side by side. If she's really experiencing gravity, then due to the universality of free fall, they will hit the floor at the same time. But this is exactly the same observation she would make in the rocket-tow scenario: the shoe and coin would continue moving

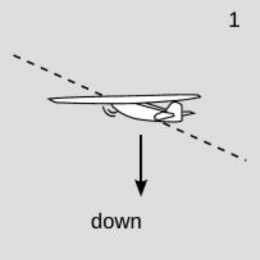


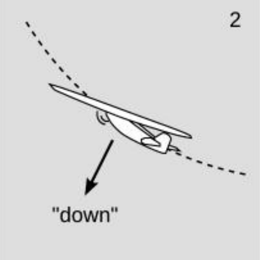
forward at constant speed, but the accelerating floor would catch up with them and hit them both at the same time.

From the perspective of Newton and Galileo, this is deeply disturbing. According to one scenario, the box is an inertial frame of reference. Based on the other, it's not.

These ideas lead relativists to the following point of view. Gravity is not really a force. Any observer who thinks she's experiencing a gravitational field can equally well interpret her observations by switching to a different frame of reference, a *free-falling* frame of reference, in which there is no gravity. These free-falling frames are the ones that deserve to be called inertial. (By this definition, the room in which you're reading this book is not an inertial frame.) Summarizing these observations, we say that the existence of a gravitational field is equivalent to choosing an accelerated frame of reference. This is called the *equivalence principle*, and it is the single most important feature of Einstein's theory of general relativity.

The equivalence principle has an important application to flying an airplane. The pilot of a plane cannot always easily tell which way is up. The horizon may not be level simply because the ground has an actual slope, and in any case the horizon may not be visible if the weather is foggy. One might imagine that the problem could be solved simply by hanging a pendulum and observing which way it pointed, but by the equivalence principle the pendulum cannot tell the difference between a gravitational field and an acceleration of the aircraft relative to the ground — nor can any other accelerometer, such as the pilot's inner ear. For example, when the plane is turning to the right, accelerometers will be tricked into believing that "down" is down and to the left. To get around this problem, airplanes use a device called an artificial horizon, which is essentially a gyroscope. The gyroscope has to be initialized when the plane is known to be oriented in a horizontal plane.[1]





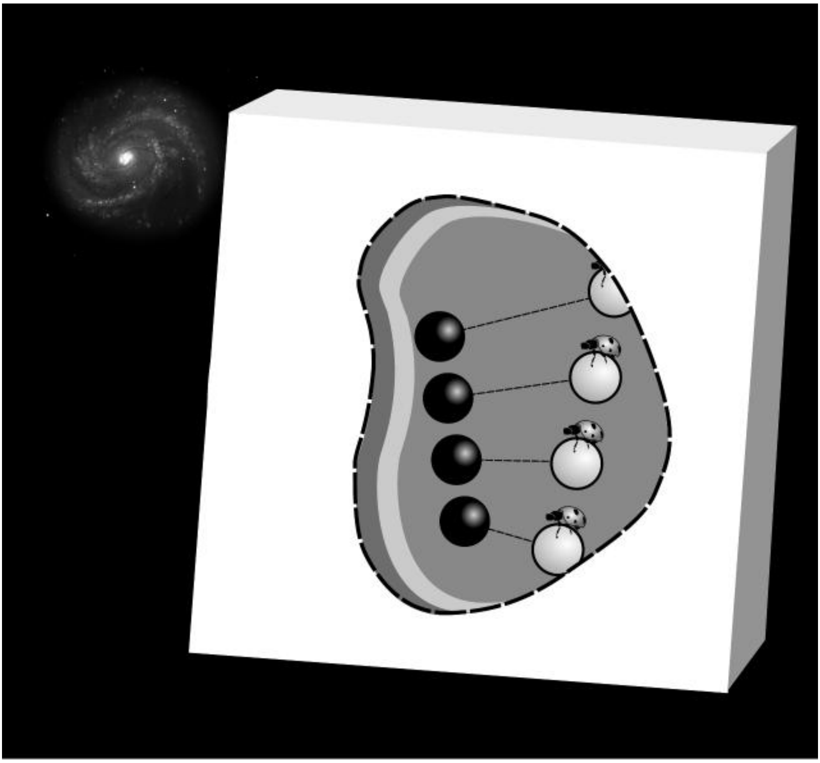
8.2 Solving the chicken-and-egg problem of inertial frames

An advantage of equivalence principle is that it allows us to say for sure whether a certain frame of reference is inertial. In the Newton-Galileo theory, there was a crippling chicken-and-egg problem with this. For example, on p. 17 we visualized the frames of reference of a cow and a passing car. Suppose we're already convinced that the cow's frame is inertial. The cow sees the car moving along in a straight line, at a speed that appears to be constant in the cow's frame. Therefore the cow says that the car's motion is inertial — the car isn't going over bumps,

[1] No gyroscope is perfect, so over time it will drift. For this reason the instrument also contains an accelerometer, and the gyroscope is always forced into agreement with the accelerometer's average output over the preceding several minutes. If the plane is flown in circles for several minutes, the artificial horizon will be fooled into indicating that the wrong direction is vertical.

running into a tree, or anything like that. Similarly, if the driver is already convinced that the car's motion is inertial, then she can tell that the cow, who to her is moving backward at constant speed, also has a good inertial frame.

But now perhaps you've noticed the difficulty. Given one inertial frame of reference, we can tell which other frames are also inertial. But how do we get started? Similarly, imagine that you were in Tanzania, and needed to make sure that the money you were handling was valid currency. If you had a 1000 shilling note from a reputable source, you could compare other bills with it and make sure they looked the same. But without a known-good example to start with, you'd have no way of telling whether shopkeepers were all handing you Monopoly money as change.



Let's say the girl Alice in the box on p. 104 wants to know whether she's in an inertial frame. To do this, she needs to know whether forces are causing her to accelerate. She knows about the equivalence principle, so she doesn't consider gravity to be a force. Therefore the question of whether gravity is acting on her doesn't even arise. She does notice, however, that the seat of the chair is touching her and making a force on her. She knows that this force is accelerating her, so she knows her frame is noninertial.

For an example of a frame that we know *is* inertial, imagine building a box with layers of shielding, as suggested in the figure on the preceding page, where the box is free-falling in outer space. One layer of shielding is aluminum foil to keep out any external electric forces. Another layer provides shielding against magnetism. (There are special products sold for this purpose, with trade names such as mu-metal.) We shield against all the forces except for the "force" of gravity, we don't consider to be a force. Now inside the box we introduce a black ball and a white ball. The roles played by the two balls are completely symmetrical, but if we like, we can think of one ball as playing the role of an observer. For this reason I've drawn a ladybug clinging to the white ball. The bug can, if she likes, define herself to be at rest, so that in her frame of reference only the black ball is moving. She knows she's shielded from all forces, so she knows her frame of reference is inertial. The chicken-and-egg problem is solved. Once she knows that her frame, tied to the white ball, is inertial, she can determine that the black ball's motion is inertial as well.

8.3 Apparent weightlessness

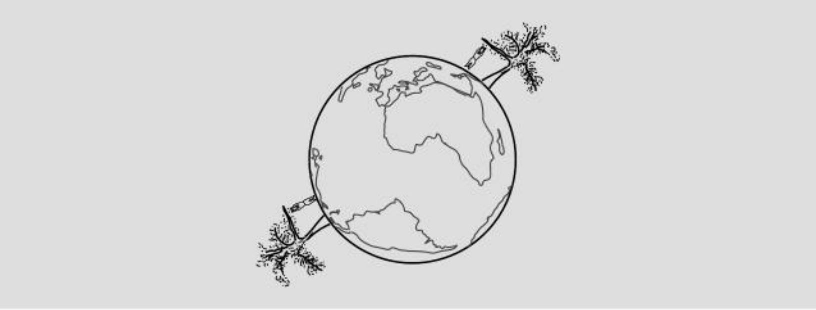
This may all seem pretty bizarre, so let's connect it to the real world. At an amusement park near where I live in Southern California, there is a ride called the Supreme Scream. Riders are raised in chairs up to the top of a tower, then dropped toward the ground in free fall. When I tried this ride, I got a queasy feeling as though my guts were floating rather than staying tied down — probably the same sensation that the Apollo astronauts got when they experienced weightlessness by traveling very far from the earth. A popular practice among people who get on this ride is to take a coin and release it in the air in front of them after the free fall has begun. An observer on the ground describes the coin as falling along with the rider, but to the rider, the coin appears to float weightlessly. The rider is doing a local experiment in an inertial (free-falling) frame of reference. In this frame, gravity doesn't exist.

The Supreme Scream is 95 meters tall, which limits the period of free fall to about four seconds. If we want to experience apparent weightlessness for longer we have to fall from higher. For example, NASA maintains a plane, affectionately known as the Vomit Comet, for training purposes. The plane climbs steeply and then free-falls along an trajectory that in the earth's frame is an arc of a parabola. The rise and fall together last about 25 seconds. The photo shows physicist Stephen Hawking on a ride in the Vomit Comet.

Exactly the same thing is going on for astronauts aboard the International Space Station. As in the case of the Vomit Comet, their free fall carries them not straight down but along an arc (according to an observer in the earth's frame). Because they're moving at such a high speed (about eight kilometers per second), the arc is a very broad one — broad enough to carry them around the world in a circular orbit, without ever hitting the ground.

8.4 Locality

The equivalence principle guarantees that we can always define an inertial frame of reference, by choosing a free-falling object as a point of reference. It would be convenient if we could define a single inertial frame big enough to cover the entire universe, but we can't. In the figure, two girls simultaneously drop down from tree branches — one in Los Angeles and one in Mumbai. The girl in LA defines an inertial frame, but if she tries to expand her frame of reference to cover her counterpart in Mumbai, the Mumbai girl does not appear to have inertial motion: to the LA girl she looks like she is falling *up* (with twice the normal acceleration of gravity as viewed in an earth-fixed frame).



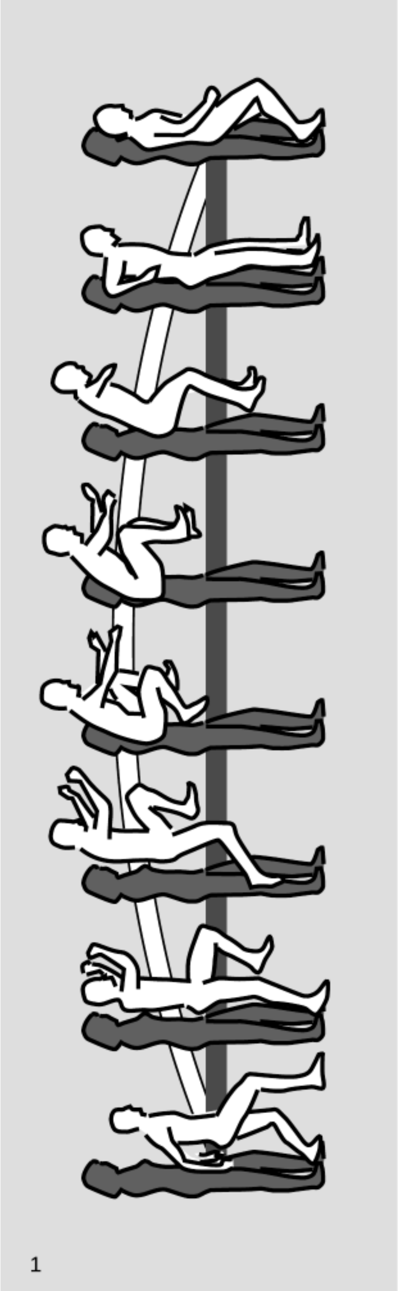
This example shows that the equivalence principle is a purely local thing. *Locally*, we can always adopt a free-falling frame of reference in which gravity doesn't exist, so that in this sense gravity is only an illusion. But we can't do the same thing *globally*, i.e., across all of time and space. Gravity is after all inescapably real, if we look at large scales. Similarly, tidal forces cannot be dismissed as illusions; for example, they destroyed of the comet in figure 2 on p. 101.

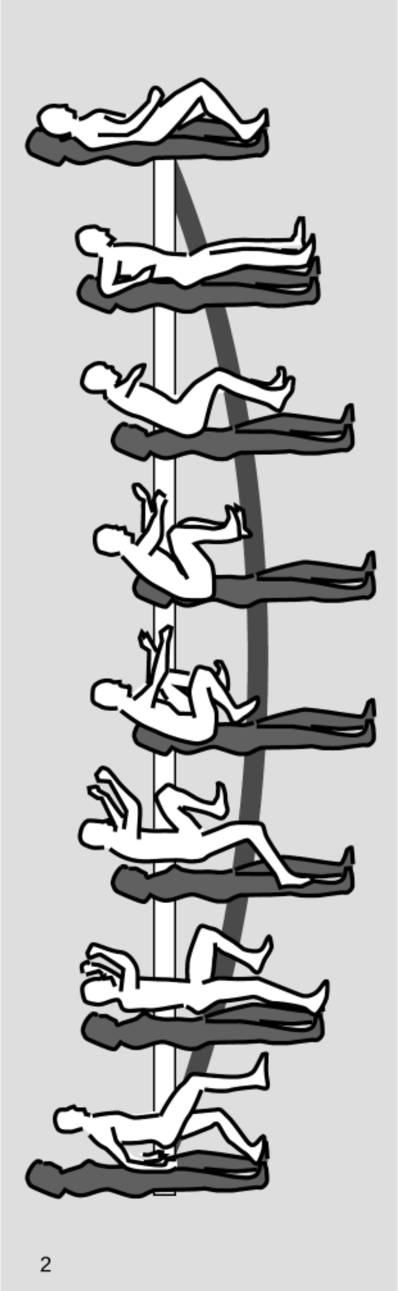
8.5 Gravitational time dilation

The equivalence principle makes powerful and surprising predictions. In figure 1, Amy (gray) stands on the ground while her twin brother Bob (white) jumps straight up in the air next to her. Because we're using our standard graphical conventions for spacetime diagrams, it looks like Bob is doing a long jump, but remember, this is just a straight up-and-down jump — the vertical axis here is time. If we didn't know about the equivalence principle, we'd expect that slightly less time would pass for Bob than for Amy, since he's moving and she's not; we'd think there would be time dilation.

But the equivalence principle makes exactly the opposite prediction. Based on the equivalence principle, we've been led to define *free-falling* frames as inertial. Therefore the more natural depiction of the situation is the one in figure 2, where we adopt Bob's frame of reference. Bob is not experiencing any forces (remember, gravity isn't a force), so his frame is inertial. According to Bob, Amy's frame is noninertial, because she does experience a force from the floor on her feet. That force is accelerating her. Her world-line is the one that is curved, not his. Based on the principle of greatest time, he will experience more time than she does. Less time will pass for her, not him.

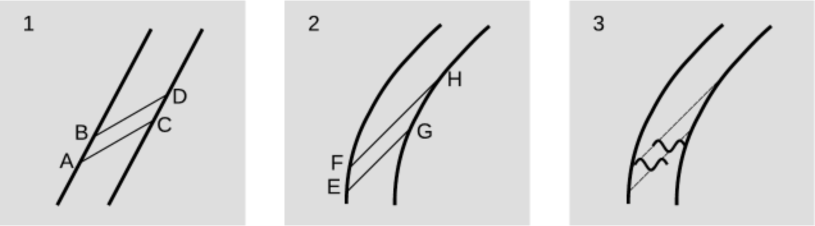
A more practical version of this experiment was carried out in 1976 by the Gravity Probe A experiment, in which an atomic clock was launched straight up (relative to the ground) aboard a space probe, to an altitude of about 10,000 km. As predicted by the equivalence principle, the clock aboard the free-falling rocket experienced a little more time (about half a part per billion) than similar clocks on the ground.


1


2

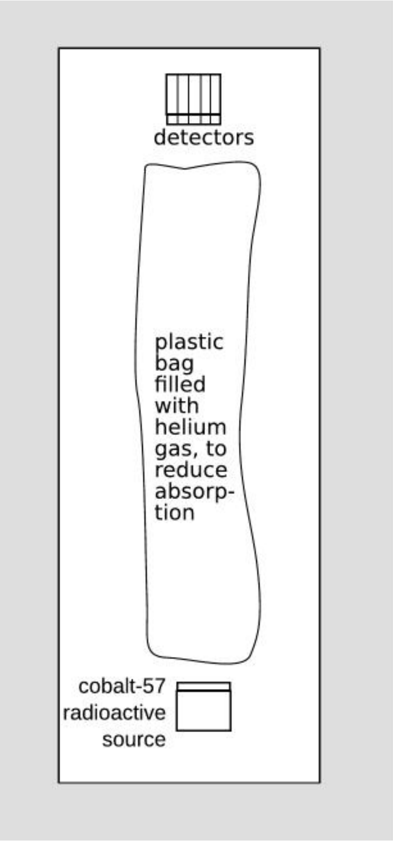
8.6 Time dilation with signals

Gravity Probe A crashed into the Atlantic Ocean at the end of its flight and was destroyed, so it couldn't be reunited with the ground-based clocks and compared with them again side by side. Instead, radio signals were sent back and forth between the probe and the ground station. To understand how this works, let's simplify the situation and go back to the scenario of Alice in her noninertial box (p. 103).



Suppose that a light source on the floor of Alice's box emits two upward flashes of light, separated by a small interval of time, so that one chases the other as they rise toward the ceiling, where they can be detected. In an inertial frame, we say that the box is accelerating upward. Therefore after the first flash arrives at the detector, the continued acceleration of the box causes the second flash to take a little extra time to arrive. It matters that the box is accelerating, not just moving. Figure 1 shows the case where the box's motion is inertial rather than accelerated. Flashes are emitted from the floor at A and B, and detected at the ceiling at C and D. (You have to visualize the box tipped over sideways.) The time intervals AB and CD are the same. But the accelerated case, shown in figure 2, is different: GH is longer than EF. Because the speed of light is big, this effect would be small, but the time interval between the arrival of the flashes would be slightly lengthened compared to the time between their emission.

If we're willing to take the equivalence principle seriously, then we should expect exactly the same effect in a laboratory that sits in a gravitational field. That is what was observed in a famous experiment by Pound and Rebka at Harvard in 1959. Figure 3 above shows the basic idea. The experiment was done in a 22 meter tower in one of the university's lab buildings. Rather than sending two discrete flashes of light up from the floor to the ceiling, they used the fact that light is a wave, so that two successive wave crests play the role of the two flashes. The light waves used were gamma rays emitted in the decay of a radioactive source. The figure below shows a simplified diagram of the setup, and photos of the two experimenters at the top and bottom of the tower.


detectors
plastic bag filled with helium gas, to reduce absorption
cobalt-57 radioactive source

The effect Pound and Rebka were trying to measure was extremely small. The time of flight for a gamma ray, moving at the speed of light, to travel 22 meters is only about 70 nanoseconds (billionths of a second). If we describe the tower in a free-falling, inertial frame of reference, then its upward acceleration at 10 meters per second squared makes it move upward by only about three millionths of a billionth of a meter (3×10^{-14} m) during this time of flight. The lengthening of the time interval between the arrivals of the two flashes is thus extremely small, and detecting it required several tricky techniques. The result was exactly as predicted by the equivalence principle.

An observer in a free-falling, inertial frame of reference explains the results of the experiment as in figures 2 and 3 on p. 112: we have two successive signals, and they travel at the same speed c, but they have to travel unequal distances so the reception of the second one is delayed.

But an observer in the noninertial frame of the earth's surface does not have this explanation available. In this frame, the tower was not moving, so the signals traveled equal distances. This observer is forced to conclude instead that time itself runs at different rates at the top and bottom of the tower. This is similar to the gravitational time dilation effect discussed in section 8.5. The cobalt-57 nucleus in the Pound-Rebka experiment is acting like a tiny clock. When it emits a gamma ray, the successive wave crests are like the clock making two ticks. An observer at the top of the tower, receiving the ticks, observes them to be abnormally far apart. This is exactly what we would expect if the clock was running slow because of a slower flow of time at the bottom of the tower.

This may seem like a rash generalization, and indeed there were probably quite a few physicists back in the 1950s who would have expressed some doubt that a result like Pound and Rebka's would also be observed with "real" clocks that count and measure time units, as opposed to microscopic "clocks" that are merely subatomic particles. But remember that such an effect was observed, for example, in 1978 by Iijima (p. 32) in an experiment using one atomic clock at the base of a mountain and another clock at the top.

Summarizing all of these examples, we conclude that time runs more slowly when you're low down in a gravitational field, faster if you're higher up. This summary also works as an alternative explanation for the thought experiment in which Bob jumps while his twin sister Amy stands on the ground. Analyzing figure 2 on p. 111, we previously said:

> Bob's motion is inertial while Amy's is not. Therefore by the principle of greatest time, Amy experiences less time than Bob.

An equally valid explanation, leading to the same numerical predictions, is:

> Amy experiences less time for a combination of two reasons. (1) As seen in Bob's inertial frame, she is the one who is moving. Motion causes time dilation. (2) Amy is also lower in a gravitational field than Bob is. For this reason as well, Amy experiences less time. These two effects reinforce each other, and their sum is the observed amount by which time was slowed.

By the way, there seems to be a very common misconception about gravitational time dilation, which is that it involves time slowing down where gravity is stronger. Not true! It's *height* that matters. For example, consider the scenario in figure 3 on p. 104, where Alice's box is being towed by the spaceship. If she measures the fake "gravity" near the floor of the box and near the ceiling, she will get exactly equal results. If she duplicates the Pound-Rebka experiment inside her box, she will see an effect, because of the difference in *height*. Or imagine sending a clock to the center of the earth. At that location, the planet's gravity pulls equally in all directions, so by symmetry there is zero gravitational force. We would nevertheless find that this clock was *slow* compared to one on the earth's surface.

Chapter 9

Curvature of spacetime

This chapter is called “curvature of spacetime,” which sounds pretty scary, but if you’re willing to wholeheartedly accept the equivalence principle, it isn’t really that bad.

9.1 What is straight?

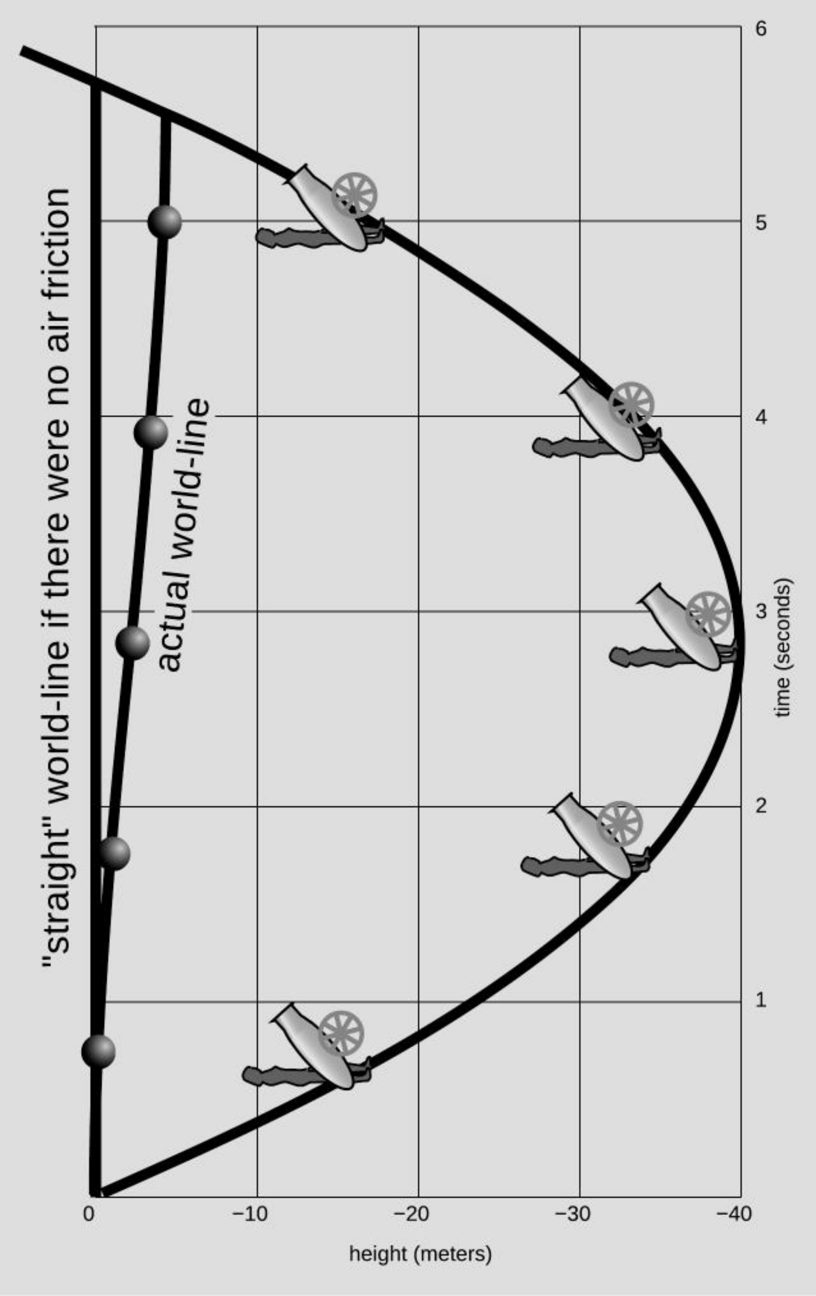
What do we mean when we say something is straight rather than curved? Say Maria wants to check that a line on a piece of paper is straight. She’ll probably use a ruler. But how does she know the ruler is straight? To check that the ruler was straight, she might hold it up to her eye and sight along it. That amounts to checking its straightness against the straightness of a ray of light. Now how does she know that a ray of light is straight?

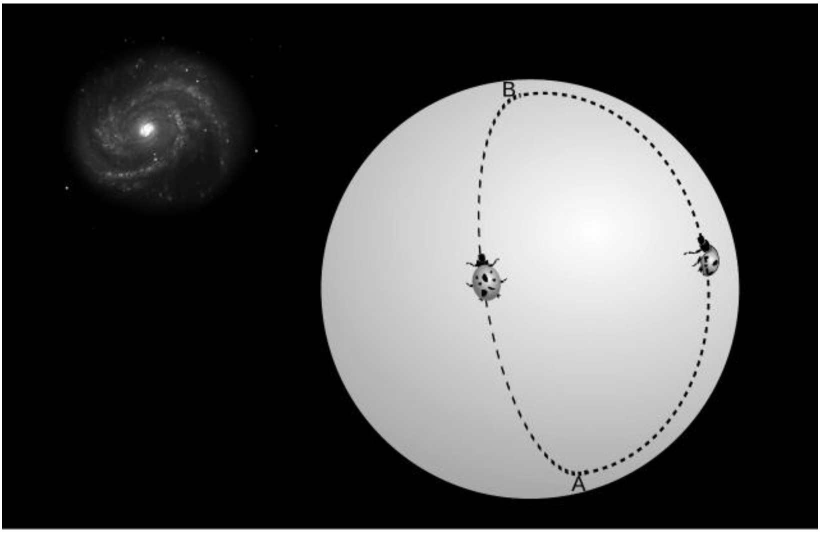
We’ll never get anywhere unless we go ahead and *define* something as straight to start with. Maria ended up resorting to a ray of light. A ray of light is a *thing* that, in this example, we assume was not being acted on by any force. In other words, the best we can probably do is to define a straight line as the path of an object that is moving inertially. That object could be a flash of light from a laser, or it could be a baseball. And since we’re talking about relativity, we want to define straightness in the context of spacetime, i.e., we want to take this as a definition of what it means for an object's world-line to be straight.

So "straight" just means "inertial." The figure on the facing page shows a realistic example. The woman fires her toy cannon, which shoots a fist-sized steel ball bearing upward at about the speed and angle of a home-run baseball.[1] The spacetime diagram shows one dimension of time and one dimension of space (the vertical one); the horizontal dimension has been suppressed, which is why the ball looks like it's going to hit her on the head at the end. The diagram is constructed in an inertial frame of reference that coincides with the initial rest frame of the ball. In other words, the point of reference is a hypothetical projectile shot just like the ball bearing, but for which air friction is negligible, so that no forces act on it. (Remember, we don't count gravity as a force.) In this frame, the woman, the cannon, the air, and the dirt all zoom downward, turn around, and then come back up; this is sort of like the view you'd see of the ground if you were a bug riding on the perfectly inertial ball. Because of air resistance, the ball's trajectory is slightly curved; it is swept downward by the air rushing down past it, and then decelerated again as the air comes back up.

The principle of greatest time (section 2.2, p. 36) says that of all the possible world-lines that a clock can follow from one event to another, the one that gives the greatest elapsed time is the inertial one. This leads to an analogy between straightness in space and straightness in spacetime. In space, a straight line gives the smallest distance between two points. In spacetime, a straight world-line gives the largest proper time between two events. The only change is that *smallest* switches to *largest*. The reason for the switch is simple. Of the three world-lines in the figure, the cannon's takes the *most* ink to draw, and this means that its world-line is the one that least resembles an inertial one. A clock moving along with the cannon would therefore suffer the most time dilation, and show the *least* elapsed time.

[1] This was simulated for a steel ball of 4 cm diameter, shot at an angle of 40 degrees above horizontal and an initial speed of 45 meters per second.


"straight" world-line if there were no air friction
actual world-line
time (seconds)
1
2
3
4
5
6
height (meters)
0
−10
−20
−30
−40



9.2 The parable of the bugs

The close analogy between spatial straightness and straightness in spacetime leads to the following parable, illustrated in the figure. Two blind bugs get in an argument.

"A Flock of Seagulls was the greatest band ever."

"Oh, they were totally derivative. I'd say Sonic Youth —"

The argument escalates. Words are said that can't be taken back. Embittered, the bugs set off from A in different directions to find new urban spaces with better brewpubs and more access to sustainably grown kale. Unexpectedly, they run into each other again at a drum circle in a public park in B.

"What are you doing here?"

"Me? I went straight and never looked back."

"That's not possible. I went straight too, and we left A traveling in different directions."

Because I've drawn the figure in a certain way, the explanation of the mystery seems obvious to us. But the bugs are blind and cling to the surface of their world, so their mental picture of their universe is two-dimensional. They can nevertheless work out what has happened. Their two straight paths have formed a kind of shape that you never learned about in kindergarten. The shape with four sides is a square, three sides make a triangle, and two is —

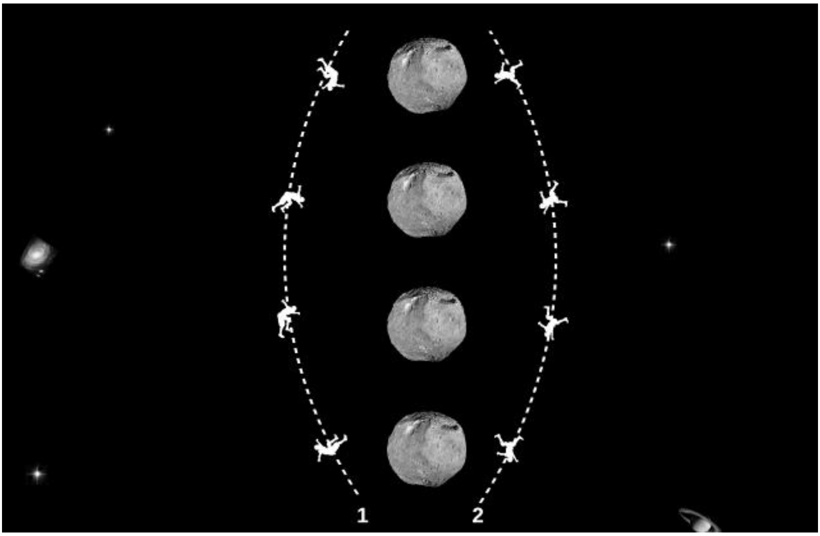
"Oh yeah," says one bug knowingly, "it's a lune. Really obscure shape, you probably never heard of it."

You can't draw a lune on a flat surface using straight lines. If a lune exists, it means that the surface is curved. What's crucial is that the bugs can detect, understand, and measure this curvature without knowing that their two-dimensional world is embedded in a three-dimensional space. I could rewrite the story so that the third dimension *didn't* actually exist in the bugs' universe, and all the events would be the same.

The bugs can even define numerically how curved their world is. One good measure of curvature would be to take the angle at the top or bottom of a lune, and divide by its area. This is essentially[2] the same as a measure of curvature called the Gaussian curvature. We can imagine making the bugs' world less and less curved by making the sphere bigger and bigger. As it gets very big, its surface becomes very flat, the distance between A and B gets very big, the area of the lune also gets very big, and the Gaussian curvature gets very small (because dividing by a big number produces a small result).

Although I presented this using a fairy-tale scenario, we can encounter all of these concepts on the earth's surface. Because the earth is big, its curvature is small, and we don't notice in everyday life that it's not flat. A motorboat cruising without turning, or an airplane flying straight and level, will trace a path on the earth's curved surface that is in some sense straight — it's as straight as any path could be. Such a path is called a *geodesic*, and the same term as also been coopted by relativists for use in spacetime, where it describes the world-line of an object moving inertially.

[2] "Essentially" means we don't worry about constant factors such as the one that arises from choosing the units of measurement for the angle.



9.3 Gravity as curvature

The figure on this page shows the parable of the bugs translated from space into spacetime. Spacesuited astronauts 1 and 2 are initially flying off in opposite directions. Each astronaut can verify that she is not being acted on by any forces — gravity isn't a force — and therefore knows that her world-line is "straight," or, in fancy terminology, a geodesic. Another way of saying this is that each astronaut is continuously experiencing apparent weightlessness. This is the equivalence principle at work.

But after some time has passed, they notice that they are no longer receding from one another.

"Ah, Two, this is One, over."

"Roger, One. Hey, have you done some kind of course correction? You're not moving away from me anymore."

"No, Two. I'm going straight. But you must have fired your jets, because you've stopped moving away from me."

The situation is similar to the one with the girls in Mumbai and Los Angeles on p. 109. The equivalence principle is local, so locally, the effects of gravity can be made to vanish by adopting an inertial (free-falling) frame of reference. But globally, that doesn't work. There is a planet in between the astronauts, and even though its surrounding gravitational field can be considered to be an illusion *locally*, it is inescapably real *globally*. It will have real, observable effects, because each astronaut ends up falling back toward the planet and crashing into the surface.

This is what we mean by curvature of spacetime. We can have two straight lines in spacetime that are initially diverging, but that later reconverge. Just as the bugs are forced to conclude that their world is curved, the astronauts will attribute their unfortunate fates to the fact that their spacetime is curved. The curvature is being caused by the gravity of the planet. In relativity, this is how we describe gravity: gravity *is* curvature of spacetime.

9.4 Deflection of starlight

Since the mass of a ray of light is zero, Newton would have expected it to be immune to gravity. But in general relativity, several different arguments lead us to expect that a ray of light can be deflected by a gravitational field:

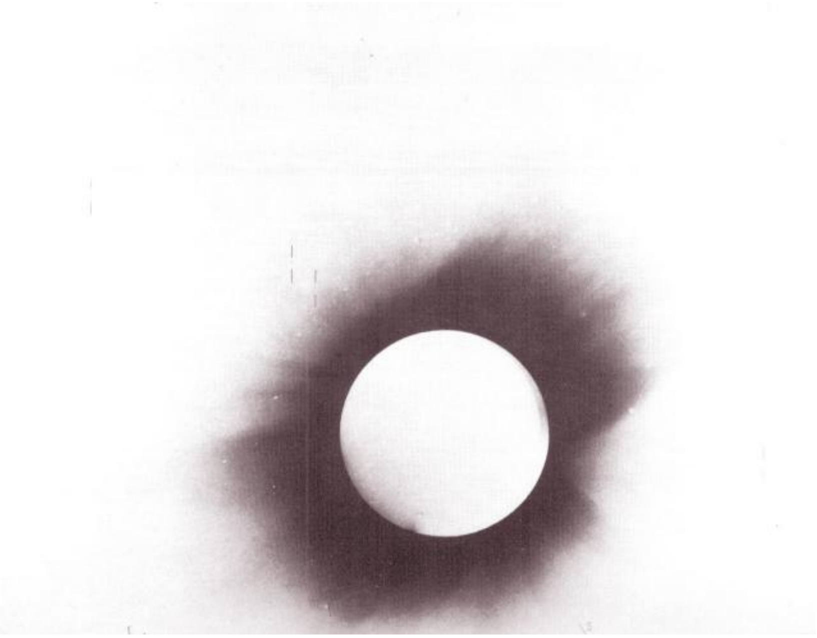
1. A ray of light does have energy, and energy is equivalent to mass.

2. Suppose, for example, that we send a ray of light in the horizontal direction here on earth. In a free-falling frame, gravity doesn't exist, and by symmetry there is no reason for the ray to bend in any direction. But if the ray is straight in a free-falling frame, then it is curved in a frame fixed to the earth's surface.

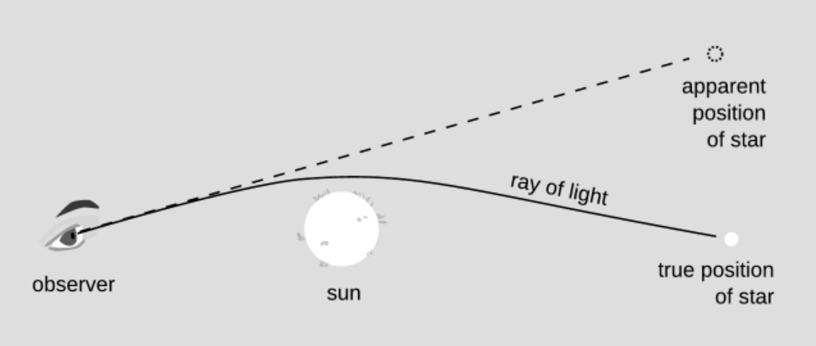
3. Consider the following analogy with special relativity. Suppose that time dilation affected some clocks but not others, and length contraction affected some rulers but not others. Then we would not be able to interpret the Lorentz

> transformation as describing a distortion of spacetime itself. Similarly, suppose that gravity affected the motion of some objects (rocks) but not others (flashlight beams). We would then have good reason to abandon general relativity, which describes gravity as a curvature of spacetime itself.

The first important experimental confirmation of relativity came in 1919 when stars next to the sun during a solar eclipse were observed to have shifted a little from their ordinary position, as shown in the photograph below. (If there was no eclipse, the glare of the sun would prevent the stars from being observed.) Starlight had been deflected by the sun's gravity. The picture is a photographic negative, so the circle that appears bright is actually the dark face of the moon, and the dark area is really the bright corona of the sun. The stars, marked by lines above and below them, appeared at positions slightly different than their normal ones.



The diagram below shows an explanation of the effect. A ray of light from the star arrives from a different direction than the one from which it would have come if the sun's gravity had not interfered. This leads to a change in the star's apparent position in the sky.

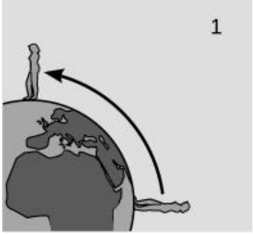


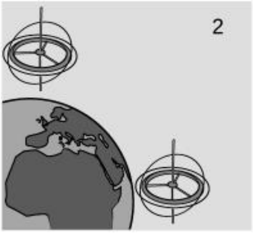
9.5 No extra dimensions

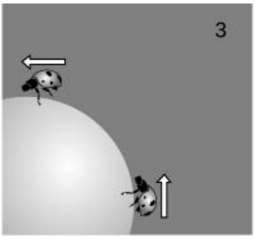
In the parable of the bugs (section 9.1, p. 118), we visualized their universe as a two-dimensional surface embedded in a three-dimensional space. The third dimension just happened to be invisible to the bugs. In science, we usually try not to invoke the existence of things that can't be empirically observed, such as invisible pink unicorns (cf. section 4.1, p. 63). To the bugs, *there are no extra dimensions.* They exist in two dimensions of space and one dimension of time, abbreviated as $2 + 1$. Similarly, you and I live in $3 + 1$ dimensions. Although we may find it useful, when we want to conceptualize curvature, to draw pictures like the pictures of the bugs on a sphere, that doesn't mean that our universe actually has extra dimensions that are invisible to us. Nobody has ever succeeded in observing a fifth dimension, although attempts have been made using particle accelerators (Franceschini *et al.*, `arxiv.org/abs/1101.4919`).

9.6 Parallel transport

This unobservability of extra dimensions has some interesting consequences. In figure 1, a woman walks from the earth's equator to the north pole, and as she does so, her body rotates by 90 degrees relative to the stars. If she was living in the bugs' universe, the stars wouldn't exist. But she isn't, they do, and she can detect the rotation, e.g., when she was at the equator, the pole star was on the horizon, but once she got to the north pole, it was directly overhead.







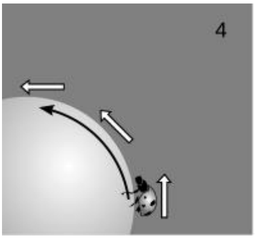
Even if the weather was always cloudy, so that she could never observe any external reference points, it would still be possible for her to verify, using only local measurements, that she had rotated. For example, as in figure 2, she could carry a gyroscope with her.[3] When she starts out at the equator, the gyroscope's axle is perpendicular to her body. When she gets to the north pole, it's parallel to her spine. Since the gyroscope is supposed to maintain its orientation, she can infer that it was her body that rotated.

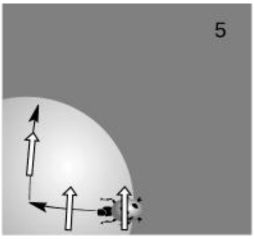
But consider how this would work out for the bug in figure 3. We can imagine that the bug owns a smartphone with a built-in gyroscope. But because there is no third spatial dimension in the bug's reality, the smartphone's screen always lies in the plane of the surface, and if it displays a fixed direction as an arrow on the phone's screen, that arrow always lies in that plane as well. If

[3] The figure shows a toy mechanical gyroscope, which wouldn't actually be practical for this purpose, but many smart phones contain a tiny gyroscope fabricated as part of a silicon chip.

the bug travels 90 degrees around the sphere, it's not possible to obtain the same result as in figure 2, because that would require the arrow to point in a direction that doesn't exist. The bug simply sees the arrow point steadily in the direction that the bug is going. We say that the arrow has been *parallel-transported* along the surface of the sphere. Parallel transport means taking a vector and moving it to a new location along a certain path, without allowing any external influence to twist the vector (or change its magnitude).

Similarly, suppose that the bug had a frictionless hockey puck. He could push the puck forward, and it would glide with a certain momentum vector. No external influence is deflecting the puck, so its momentum vector is parallel-transported.





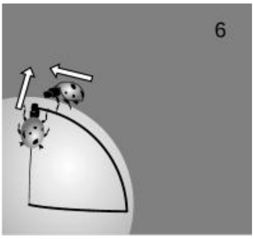
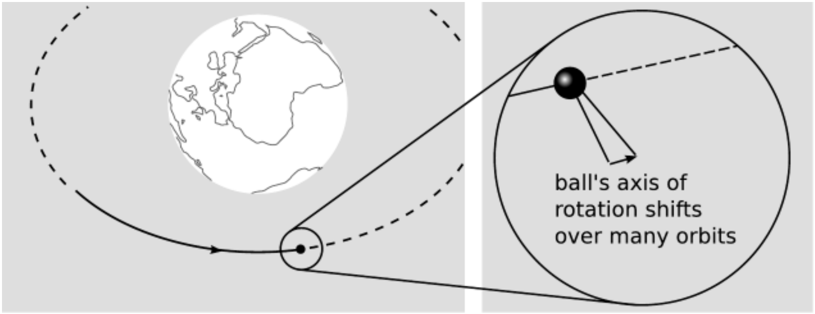


Figure 4 duplicates figure 3: our bug parallel-transports his cell-phone gyroscope from what we would describe as his world's "equator" up to its "north pole." Meanwhile, in figure 5, his friend starts from the same point and ends up at the same destination, but via a different path: he first travels west 90 degrees along the equator, then goes north to the pole. When the bugs start out, they make sure that their gyroscopes agree. When they are reunited, they find that the gyroscopes are out of whack. Each bug could claim that the other's gyro is broken, but in reality what has happened is an effect of the curvature of their world. In fact, this is a good general definition of what we mean by curvature: when curvature is present, parallel-transport from point A to point B can give a result that depends on the path chosen.

This path-dependence of parallel transport was verified directly in 2005 in a satellite-based experiment called Gravity Probe B. On board the satellite was a gyroscope in the form of a silicon ball spinning frictionlessly in a vacuum inside a box. The satellite was orbiting the earth, so on a spacetime diagram we would visualize its world-line as a helix. Parallel-transporting the gyroscope along this helical world-line gives a different result than if its path through spacetime had not been orbiting. If the gyroscope had then been brought down from orbit and compared with one left on the surface of the earth, they would have disagreed by a small amount, due to the earth's gravity: the earth's gravitational field is a curvature of the surrounding space. In reality, the gyroscope was too delicate to be handled in this way without disrupting its motion, so the effect was observed as a change in the orientation of its axis relative to the stars. Over the one-year duration of the mission, the axis twisted around by an angle of about 0.002 degrees, in good agreement with the predictions of general relativity.



Chapter 10

Matter

In Newton's nonrelativistic description (ch. 7, p. 93), gravity was a universal attraction between masses and other masses. It was a force, and it acted instantaneously from a distance. The relativistic description is completely different. Matter, such as the planet in the figure on p. 122, causes curvature in the spacetime around it. This curvature then influences the motion of matter. The relativistic picture is called Einstein's theory of general relativity. The word "general" is used because in special relativity, the curvature of spacetime is assumed to be zero; general relativity generalizes the theory to allow the curvature to be nonzero. Relativist John Wheeler summarizes this in a slogan:

> "Spacetime tells matter how to move; matter tells spacetime how to curve."

Section 13.1 further fleshes this out.

Wheeler's description requires us to define what we mean by "matter." Newton would have said that mass was the correct way to measure the quantity of matter. But for the reasons discussed in section 9.4, "matter" in general relativity should be taken to mean mass-energy, not just mass.

One of the most surprising predictions of general relativity is the existence of black holes, which are formed when matter is sufficiently compressed by gravity. To understand how this can happen, we need to understand how matter should be expected to behave under extreme conditions. Luckily this turns out to be unexpectedly easy. Understanding the *detailed* properties of matter is the highly technical province of specialists such as metallurgists, physical chemists, and condensed-matter physicists. But, surprisingly, relativity itself makes some more *general* predictions about matter that are independent of its detailed structure.

10.1 No nongravitating matter

One such prediction is that there is no such thing as matter that is immune to gravity; anything that has mass-energy *is* affected by gravity. We see objects in everyday life that superficially seem to have such an immunity. Birds can fly, but that's not because they ignore gravity. Birds exploit forces from the air acting on their wings. The same kind of explanation holds for a speck of dust floating in the air, the microscopic drops of water that form clouds, helium balloons, or air molecules themselves. In particular, if we use a vacuum pump to suck the air out of a flask, the flask becomes lighter, because the air that was removed had weight. If air didn't have weight, the earth's gravity wouldn't keep the planet's thin film of atmosphere from escaping, and our world would be an airless body like the moon. This universal susceptibility to gravity even extends to things that nobody but a relativist would call matter, such as the rays of sunlight in the 1919 test of general relativity described on p. 124.

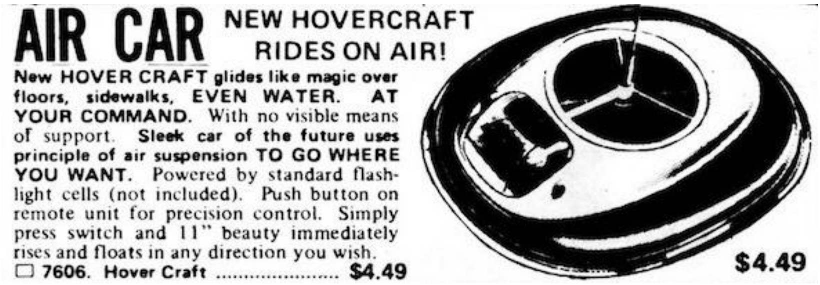
Suppose, on the contrary, that we *could* find some form of exotic matter — call it FloatyStuff$^{\text{TM}}$ — that had the ordinary amount of inertia, but was completely unaffected by gravity. If FloatyStuff existed, it would cause us to doubt the equivalence principle. The basis for the equivalence principle is the universality of free fall, which is violated by FloatyStuff.

To dramatize this a little more, say that alien gangsters land in a flying saucer, kidnap you out of your back yard, konk you on the head, and take you away. When you regain consciousness, you're locked up in a sealed cabin in their spaceship. You pull your keychain out of your pocket and release it, and you observe that it accelerates toward the floor with an acceleration that seems quite a bit slower than what you're used to on earth, perhaps a third of a gee. There are two possible explanations for this. One is that the aliens have taken you to some other planet, maybe Mars, where the strength of gravity is a third of what we have on earth. The other is that your keychain didn't really accelerate at all: you're still inside the flying saucer, which is in interplanetary space and accelerating at a third of a gee, so that it was really the deck that accelerated up and hit the keys.

There is absolutely no way to tell which of these two scenarios is actually the case — unless you happen to have a chunk of FloatyStuff in your other pocket. If you release the FloatyStuff and it hovers above the deck, then you're on another planet and experiencing genuine gravity; your keychain responded to the gravity, but the FloatyStuff didn't. But if you release the FloatyStuff and see it hit the deck, then the flying saucer is accelerating through outer space. In this situation, you are able to tell the difference between a gravitational field and an accelerated frame of reference, which according to the equivalence principle are supposed to be equivalent things.

Thus the existence of matter that is immune to gravity is forbidden by the equivalence principle. It is an ironclad prediction of general relativity that no such form of matter exists. If any such material were ever discovered, general relativity would be disproved.

For similar reasons, we can never shield against gravitational fields in the same way that we can with electrical, magnetic, or other forces. If we could, then we could take any ordinary piece of matter, such as a brick, and turn it into a gravity-immune object by wrapping it in gravitational shielding. This is why, in our operational definition of an inertial frame of reference (section 8.2, p. 105), we shielded against all forces *except* for gravity, which would have been impossible.



AIR CAR NEW HOVERCRAFT RIDES ON AIR!

New HOVER CRAFT glides like magic over floors, sidewalks, EVEN WATER. AT YOUR COMMAND. With no visible means of support. Sleek car of the future uses principle of air suspension TO GO WHERE YOU WANT. Powered by standard flashlight cells (not included). Push button on remote unit for precision control. Simply press switch and 11" beauty immediately rises and floats in any direction you wish.

□ 7606. Hover Craft $4.49

$4.49

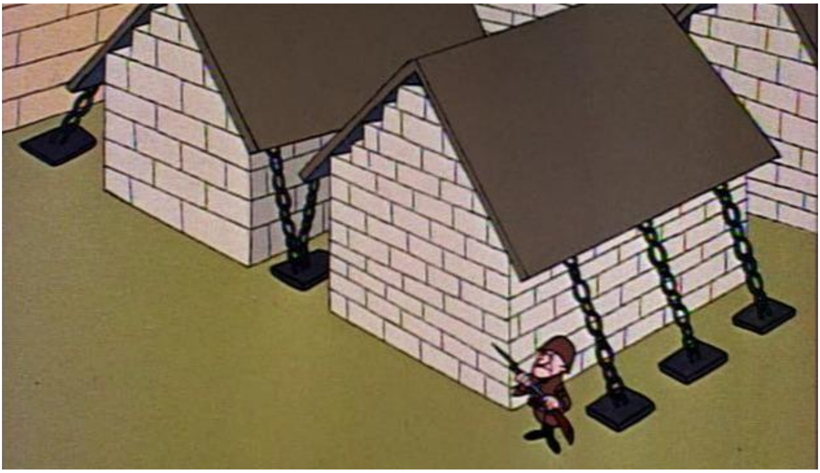
As an example, the figure shows an old ad that was frequently found in comic books. If the ad claimed that the car worked by antigravity, then we could be pretty sure that it was a scam, since it would violate general relativity. But a careful reading of the text shows that the car is supposed to work by floating on a cushion of air, which is more scientifically plausible, and in fact, people who bought the car way back when have posted descriptions on the internet stating that the car did work. (But note that it's only 11 inches long, so you can't ride in it.)

I've structured this book so that the next few chapters revolve around black holes. As we'll see in more detail later, there are some pop-culture ideas about black holes that turn out to be right, and others that are wrong. One of the correct ones is that once you've fallen into a black hole, you can never escape. The nonexistence of gravitationless matter and gravitational shielding help to explain why this should be so. If you had a flying car that was made out of gravitationless matter, it certainly does seem like it would be able to escape a black hole's gravity.

10.2 No repulsive gravity

The black holes that we observe are believed to have formed by the runaway gravitational collapse of stars or clouds of gas and dust. This would be much like the formation of objects such as the earth and the sun, except that the collapse would have progressed much further, to extremely high densities. When the idea of black holes was first proposed (long before they had actually been observed), various objections were raised to their formation, one of them being that we don't necessarily know the properties of matter under such extreme conditions. By analogy, suppose that we place two hydrogen atoms at some small distance from each other, say a distance equal to a few of their own diameters. They will experience an attractive electrical force and will "fall" toward each other. But once they get close enough, so that their electron clouds start to overlap significantly, the attraction turns into a repulsion. The collapse stops, and we end up with a stable hydrogen molecule, the form of the gas that is actually encountered normally.

Is it possible, then, that under certain conditions gravity might also behave repulsively, possibly preventing the formation of black holes? A silly pop-culture version of this, from the Rocky and Bullwinkle cartoon series, is the fictional substance upsidaisium, which falls up rather than down. The figure shows a soldier guarding a military stockpile of upsidaisium bricks, which have to be kept staked down so they don't fly up into the sky.



If upsidaisium were discovered, it would violate the universality of free fall, and therefore invalidate the equivalence principle and disprove general relativity.

But there is a way to get repulsive gravity without throwing general relativity down the toilet. Recall that mass is defined not in terms of weight but as a measure of inertia (section 5.1, p. 71). Suppose that we have a brick whose *mass* is *minus* one kilogram. This is admittedly very strange. For example, if I hit such a brick with a baseball bat toward the east, it would accelerate to the west. If I whacked a car with it, the brick's momentum would be in the opposite of the expected direction, so the sheet metal would pucker outward rather than forming an inward dent. But if this brick also had negative weight (like upsidaisium), it would fall down, just like normal matter: since it responds to forces in the opposite of the usual way, the upward force of gravity would accelerate it downward. Thus we could have repulsive gravity while preserving the equivalence principle.

The nonexistence of repulsive gravity, or negative mass-energy, is therefore not completely ruled out by general relativity, and we have to posit it as an extra assumption. Giving an exact mathematical statement of this assumption is complicated and beyond the scope of this book, but the foregoing discussion should make it clear what the general idea is. It turns out that there is more than one way of giving an exact formulation, and the different ways are not equivalent to one another; some are stricter requirements and some are weaker. These are referred to as *energy conditions.* If we look at the ordinary forms of matter that surround us in our solar system, or that we have been able to test in the laboratory, all of them obey all of the energy conditions. A notable exception is *dark energy*, a recently discovered, mysterious phenomenon that seems to pervade the universe, but that has a negligible effect except at cosmic scales.

10.3 Finite strength of materials

A final possible objection to the possibility of runaway gravitational collapse is that with many materials, the more we compress them, the more they resist compression. For example, pumping up the tire of a bicycle is relatively easy at first, but as the pressure inside the tire gets greater and greater, it gets harder and harder to pump. Now suppose we had a powerful mechanical pump. It could compress air to much greater pressures than we could achieve using a hand pump, but at a certain point, something new would happen: under sufficient pressure, air turns from a gas into a liquid. (A similar example is the carbon dioxide canisters hidden behind the soda dispensers at restaurants. The CO_2 inside has been pressurized so much that it is a liquid.) A liquid has qualitatively different properties than a gas, and in particular it strongly resists further compression. This is why, when you squeeze a water balloon, it just bulges outward in the gaps between your fingers.

Under higher and higher pressures, matter is believed to pass through a series of abrupt changes similar to the gas-liquid transition — more about this in section 11.1, p. 137. With each such transition, it crunches down into a denser and yet more exotic state. Our scientific knowledge is always limited, and these are conditions that we can't replicate in the laboratory. Is it possible that beyond the states of matter currently hypothesized there is some other, as yet unsuspected, form of matter that is even more dense, and that represents the ultimate possible state of compression of matter?

Surprisingly, relativity allows us to answer this question, and the answer is no. It tells us that there must be a limit to the ability of matter to resist compression.

The key here is to consider the speed at which vibrations travel in a material. When you pop a champagne bottle, the disturbance in the air spreads outward as an audible sound. The speed of sound in air is fast — several hundred meters per second. In a more incompressible substance like water, the speed of sound is about five times greater. This is because stronger forces now resist compression, and these stronger forces cause the water to accelerate more violently. If a substance had infinite resistance to compression, then it would be perfectly rigid, and the speed of sound in that substance would be infinite. But relativity tells us that this is impossible. No signal or physical effect can propagate at a speed greater than *c*.

We therefore conclude that there is some limit to the ability of matter to withstand compression. Every substance, like the crushed cars in the photo, will compress if squished hard enough. Thus it seems that if a sufficiently massive body collapses gravitationally, the result must be a runaway collapse.

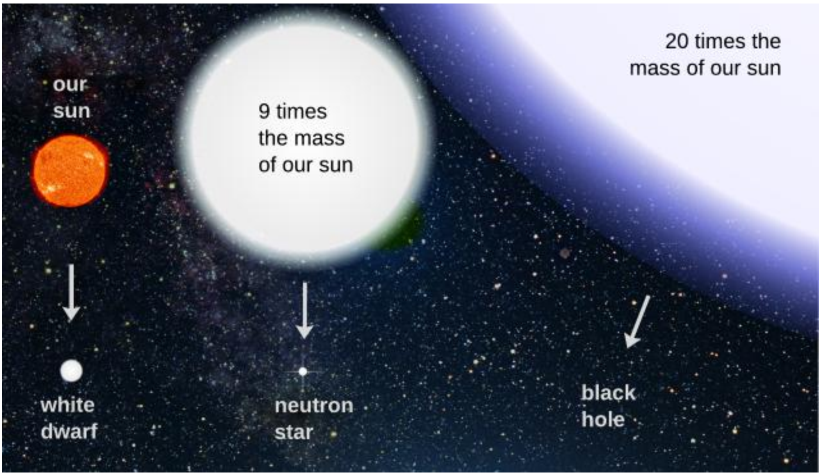
Chapter 11

Black holes

11.1 Collapse to a singularity

Let's talk about death. Our sun is very close to the end of the time during which it will be the right brightness to support life on our planet. As it nears total depletion of the hydrogen fuel that it burns, it will pass through a series of complicated changes, at first subtle and gradual but then more rapid and violent. Astronomers know about these changes because other stars, further down the road, provide snapshots of the process. Near the beginning of the sun's process of dying, our little planet will be completely sterilized, and later it may be destroyed outright.

Once the fuel gives out, our sun loses its source of heat. Just as the heat inside a hot air balloon keeps it inflated, the heat energy produced in our sun is currently what prevents it from crumpling. With the heat turned off, there will be nothing to oppose gravity, and the sun will begin to collapse gravitationally.



But this will not, according to our current understanding, be a runaway collapse. Our sun is after all a pretty small and insignificant star, and its gravity is not exceptionally strong. In the end, we think matter's resistance to compression will be strong enough to slow the collapse to a stop, leaving our dead star in a very dense form called a white dwarf. A new and final equilibrium will have been achieved. White dwarf matter has the fantastic density of about a ton per cubic centimeter, far beyond anything we can achieve in the laboratory.[1] But this is not pure theoretical extrapolation. Our galaxy is littered with white dwarfs, which are the most common type of stellar corpse.

In 1930, at the age of 19, the astrophysicist Subrahmanyan Chandrasekhar (1910-1995) demonstrated in a theoretical calculation that white dwarf matter could not stand up to more than

[1] The white dwarf in the figure is not drawn to scale; it would actually be about the size of the earth, which would be too small to see on the page.

a certain amount of pressure. Inside a big enough dying star, the atoms themselves begin to collapse under the pressure of gravity, with the electrons and protons fusing to form neutrons. The result is a ball made mostly of pure neutrons, called a neutron star. Observational evidence for the existence of neutron stars came in 1967 with the detection by Bell and Hewish at Cambridge of a mysterious radio signal, repeating at an interval of exactly 1.3373011 seconds. The signal's observability was synchronized with the rotation of the earth relative to the stars, rather than with legal clock time or the earth's rotation relative to the sun. This led to the conclusion that its origin was in space rather than on earth, and Bell and Hewish originally dubbed it LGM-1 for "little green men." The discovery of a second signal, from a different direction in the sky, convinced them that it was not actually an artificial signal being generated by aliens. Bell published the observation as an appendix to her PhD thesis, and it was soon interpreted as a signal from a neutron star. Neutron stars can be highly magnetized, and because of this magnetization they may emit a directional beam of electromagnetic radiation that sweeps across the sky once per rotation — the "lighthouse effect." If the earth lies in the plane of the beam, a repeating signal can be detected, and the star is referred to as a pulsar.

As we saw in section 10.3, p. 134, relativity predicts that we can't go on discovering more and more compressed forms of matter. (It's a bit like the situation of Columbus and Magellan: they may have been surprised to come across two or three new continents, but they knew that the world was round, so at some point there had to be an end to what there was to discover.) For a sufficiently massive star, we expect that gravity will win and matter will lose. There will be a runaway collapse. The end result is expected to be a black hole: a star with all its mass squeezed down to a space smaller than an atom of ordinary matter. This tiny object, which is a mathematical point according to general relativity, is called a *singularity.* It represents a point where either the laws of physics break down completely, or general relativity stops making sense and should not be trusted.

11.2 The event horizon

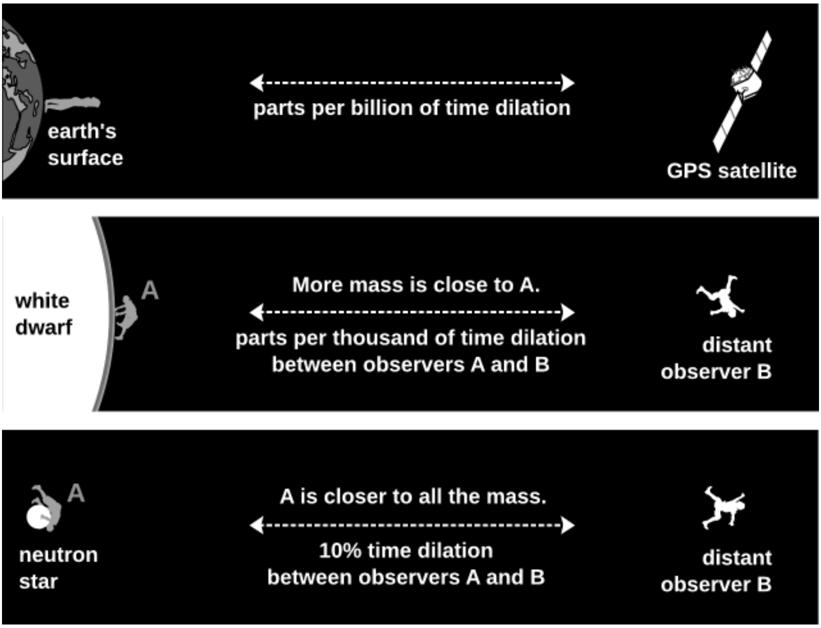
The most distinctive property of black holes is that they are expected to be black. To see why, consider what we know about gravitational time dilation (section 8.5, p. 110).

If we compare the flow of time at point A on the earth's surface with its rate of flow at some distant point B such as a GPS satellite, we get a difference, but it's very small, only a part per billion perhaps. We can make the effect a little bigger by choosing B to be farther away, but we rapidly reach a point of diminishing returns, because the earth's gravity dies out rapidly with distance. We could also try making point A underground, but again this doesn't produce any vastly increased effect. The reason is the same: the earth's gravity gets weaker as you go deep inside it. If this seems surprising, consider what you would feel if you could go to the point exactly at the center of the earth: the attraction of the earth's mass would pull on you equally in all directions, so there would be no gravity at all. So we can maximize the gravitational time dilation of A relative to B by putting A at the earth's center and B infinitely far away, but the result will still be in the parts-per-billion range. The ultimate reason for this is simply that the earth has a certain size, and we can't get any closer to all of its mass than that.

With a white dwarf we can do quite a bit better. The mass is a million times greater, so we can get a time dilation effect almost in the parts-per-thousand range. The detection of this effect was one of the early experimental tests of relativity.

What about a neutron star? Whereas a white dwarf is about the size of the earth, a neutron star is the size of a city. The mass isn't much more, but we can get *closer* to all of it. The result is time dilation of something like ten percent.

Now you can see where all of this is leading. Because a black hole is essentially a pointlike object, there is no limit on how close you can get to all of its mass — if you dare. Although the math is too complex to go into here, it turns out that if you get within about 30 kilometers of a ten-solar-mass black hole, the


earth's surface
parts per billion of time dilation
GPS satellite
white dwarf
A
More mass is close to A.
parts per thousand of time dilation between observers A and B
distant observer B
A
neutron star
A is closer to all the mass.
10% time dilation between observers A and B
distant observer B

time dilation becomes *infinite*.

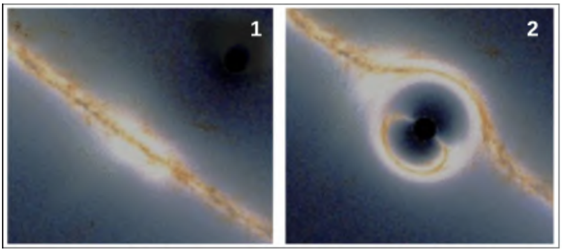
Now let's think about what that means. We have an imaginary sphere centered on our black hole's singularity, with a radius of 30 km. The surface of this sphere is called the *event horizon*. Suppose that you fly down to the event horizon in your rocket ship and then try to transmit a radio message back home. You speak the first word of your message, and then the second word. To you, the time between these words is a fraction of a second. But to a distant observer receiving this message, the time interval between the words equals eternity. In other words, it is not possible to transmit information to a distant point from the event horizon or anywhere inside it.

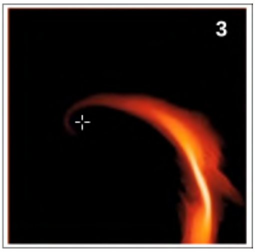
Similarly, suppose that what is emitted from the event horizon isn't an information-bearing signal but simply a light wave. At emission, the time between one vibration of the wave and the next is very short, but to a distant observer it's an infinite time. Clearly the light wave simply isn't escaping.

11.3 Detection

Because light can't escape from inside the event horizon, we can never hope to observe black holes by seeing them shine, as normal stars do. There are nevertheless several methods that have either been used already or may be used in the future for detecting black holes.

One method, which is so far purely hypothetical, is gravitational lensing, as shown in figures 1 and 2. Figure 1 shows a telescope's magnified view of a distant galaxy. Now suppose that, by pure luck, a black hole happens to lie very nearly along our line of sight to this galaxy. As in the 1919 test of general relativity described on p. 124, any rays of light that happen to pass very close to the black hole (but without entering its event horizon) will be bent by the black hole's gravity. The result would be a type of distortion of the image, as in figure 2. Although gravitational lensing has been detected many times, lensing by a black hole has not yet been observed.





A far easier method is to exploit the fact that, as in the simulated view in figure 3, black holes are often actively feeding on nearby matter. In this simulation, an entire star has wandered too close to the black hole, which is at the position marked with crosshairs. The star has been broken up by tidal forces (section 7.4, p. 100) and is now falling toward the black hole as a cloud of gas. Internal friction causes the gas to become so hot that it glows in the x-ray part of the spectrum. These x-rays are a highly distinctive sign of a black hole. The object known as Cygnus X-1 is the best-studied example. This X-ray-emitting object was discovered by a rocket-based experiment in 1964.

Around the turn of the 21st century, new evidence was found for the prevalence of supermassive black holes near the centers of nearly all galaxies, including our own. Near our galaxy's center is an object called Sagittarius A*, detected because nearby stars orbit around it. The orbital data show that Sagittarius A* has a mass of about four million solar masses, confined within a sphere with a radius less than 22 million kilometers. There is no known astrophysical model that could prevent the collapse of such a compact object into a black hole, nor is there any plausible model that would allow this much mass to exist in equilibrium in such a small space, without emitting enough light to be observable.

The existence of supermassive black holes is surprising. Gas clouds with masses greater than about 100 solar masses cannot form stable stars, so supermassive black holes cannot be the endpoint of the evolution of heavy stars. Mergers of multiple stars to form more massive objects are generally statistically unlikely, since a star is such a small target in relation to the distance between the stars. Once astronomers were confronted with the empirical fact of these supermassive black holes' existence, a variety of mechanisms was proposed for their formation. Little is known about which of these mechanisms is correct.

11.4 Light cones

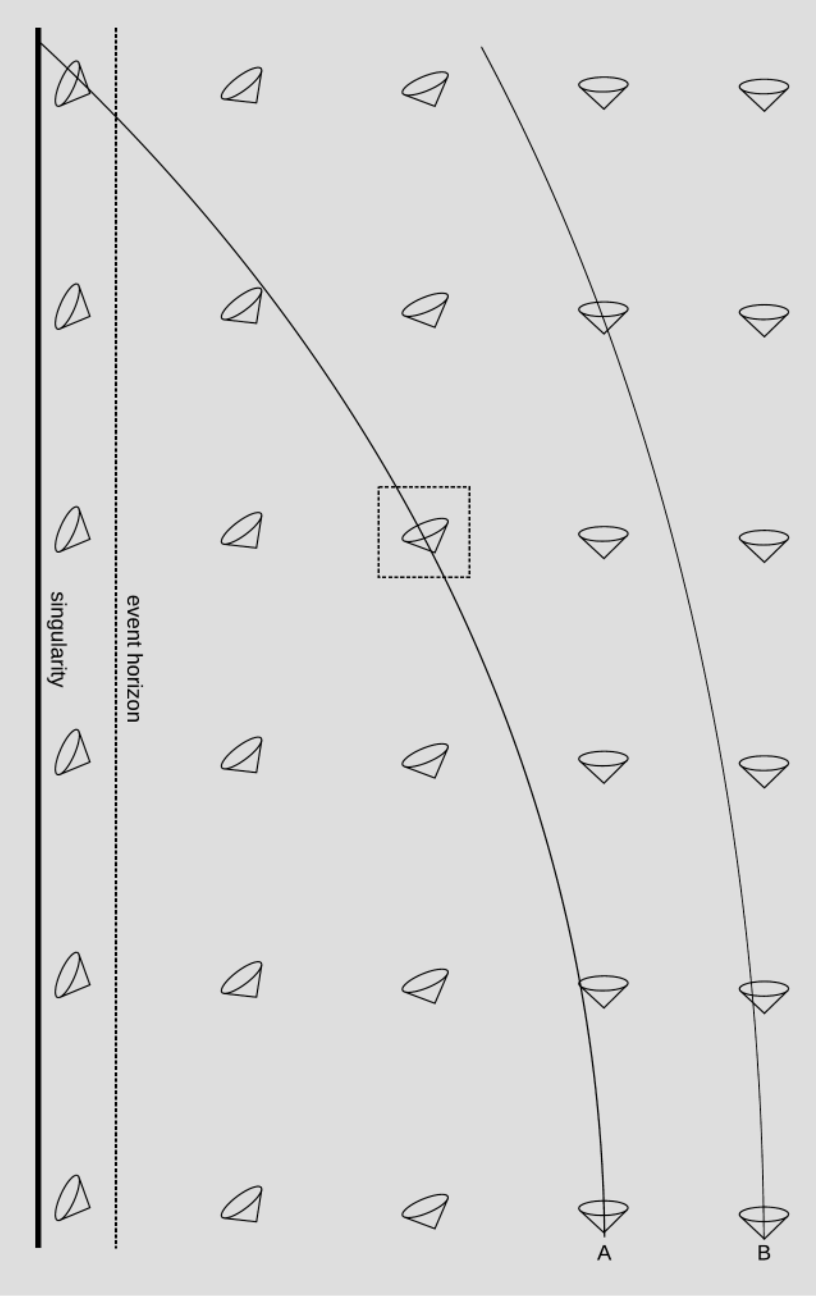
A typical scheme for evading the one-way nature of the event horizon would be to put a space probe on the end of a rope, lower the rope down through the horizon, and then pull it back out. The plan won't work, because tidal forces will break the rope. Why can't we just use a stronger rope? The answer is that, as we've seen in section 10.3, p. 134, relativity imposes ultimate limits on the strengths of materials. This example demonstrates a more general fact, which is that the event horizon is not just a barrier that is difficult to escape through. The reason it's called the *event* horizon is that it marks a fundamental disconnection of cause and effect. Events inside the horizon can never cause events outside.

The figure on the facing page shows a good way to visualize this. It's a standard spacetime diagram, with the time axis running vertical and space horizontal. The black hole is stationary in this frame of reference, so the singularity's world-line is a vertical line, and so is the event horizon's. Astronauts Alice and Bob start at rest, at unequal distances from the black hole, and under the influence of gravity they begin to accelerate toward it in free fall.

Since they're so close to a black hole, you might imagine that they would feel very heavy, but they're free-falling, so they both experience apparent weightlessness (section 8.3, p. 107). This is the equivalence principle at work. Locally, gravity is an illusion that we only experience if we adopt a frame of reference that's not inertial (not free-falling).

Suppose Alice carries out physics experiments in the small region of spacetime indicated by the dashed square. She will find absolutely nothing unusual, and everything will be explainable in terms of *special* relativity, because the curvature of spacetime is not detectable on such a small scale. (Similarly, we don't worry about the curvature of the earth when driving around town.) In particular, she will be able to define a future light cone for herself, as special relativity says she can always do.

But now consider the situation shown near the top of the diagram, when Alice has crossed the event horizon. We know that if she emits a ray of light from this position, aiming it outward, it will not escape the black hole. But to Alice, in her frame of reference, special relativity is still locally valid, and the ray of light travels, as light always does, at c. How can this be, if the light isn't going to escape? The resolution of the paradox is that if we draw her light-cones on this type of diagram, they're tipped. In her local frame of reference, light moves at the expected speed in both the inward and outward directions, and if she were drawing the light cone, she'd draw it upright. But to us, in our global view, it's tipped.


singularity
event horizon
A
B

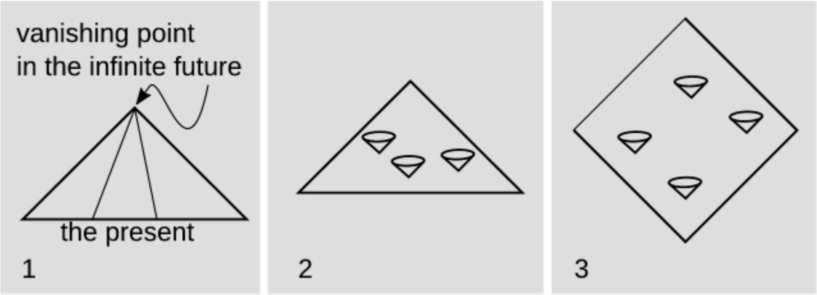
The light cone, as we've already seen, is not fundamentally about light. It's about mapping out the possible cause-and-effect relationships between events. As Alice crosses the event horizon, her entire future light cone tips over so far that it all points inward, not outward. Alice can no longer physically escape, nor can she send a signal back to Bob. For her, there is no future other than the region inside the black hole.

11.5 Penrose diagrams

The figure with the tipping light cones on p. 145 is one way of visualizing the curved spacetime around a black hole. In this section I introduce a different method called a Penrose diagram, named after mathematician and relativist Roger Penrose, which is in some ways nicer. It starts from the simple idea of a vanishing point, demonstrated in the image below of the view through a car's windshield.

The parallel lines painted on the asphalt appear to converge toward a vanishing point in the distance. This is simply because distant objects look smaller, which is a type of distortion, but it's a distortion that we're accustomed to and don't normally think about. In fact, it's a useful distortion. The ancient Egyptians didn't understand how to handle perspective in their art, so they couldn't render a landscape the way it's shown in this figure, simulating the eye's view. Without the useful distortion of perspective, we wouldn't be able to simultaneously perceive both the distant mountains and a nearby feature like a bush. If the bush were rendered on the same scale as the mountain, we would either have to make the bush microscopic, and too small to see, or we'd have to make the mountains so big that they wouldn't fit in the picture. By the way, the road also has a vanishing point behind your head, which is shown in the rear-view mirror. You don't normally see this one, but if you had eyes in back as well as in front, like some jumping spiders, it would look the same to you as the one in front.

In a Penrose diagram, shown in figure 1 below, we have the same idea, but this is now a spacetime diagram, so the vanishing point at the top doesn't represent a point infinitely far in space but one infinitely far in the future. As in perspective art, we distort scales in order to fit everything in. Even though the vanishing point is infinitely far in the future, we draw it just a few centimeters away from the bottom of the diagram. (The distance scale also gets more and more distorted as we move out toward the corners on the sides. These corners represent infinitely distant space at the present time.) The two thinner lines in the diagram are analogous to the painted white lines in the photo of the road. They are parallel, i.e., they always maintain the same distance away from one another in space.

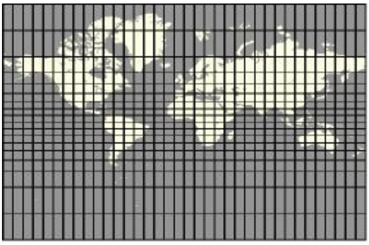


Keep in mind that all we've drawn so far is a Penrose diagram for flat spacetime, i.e., special relativity, with no gravity. There are no black holes lurking in our landscape. All of the distortions are distortions in the way we've *draw* the diagram, not real curvature of spacetime.

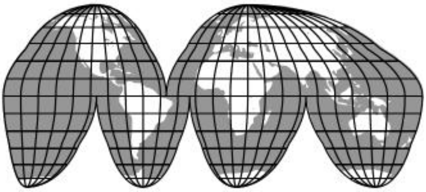
Figure 2 shows some light cones superimposed on the Penrose diagram. Despite all the distortions, the light cones don't tip. This is the big advantage of a Penrose diagram. Light cones spell out the logic of possible cause and effect relationships, so they're very important. When we refrain from tipping them, we make it easier to look at the diagram and reason about the light cone.

Figures 1 and 2 started from the present moment in time and went forward into the future. But we can also represent the past on the same Penrose diagram, as in figure 3, making the result into a diamond. This is like adding in the rear-view mirror in the original photo of the landscape.

So far this has all been in flat spacetime. Before we go into curved spacetime, let's warm up by thinking about ordinary curved surfaces, like the surface of the earth. Because the world is round, we can never reproduce it perfectly on a flat map. To flatten its curved surface, we must distort, fold, or cut it. The figure below shows two of the many possible ways of accomplishing this.



Grid is square. Sizes are distorted.

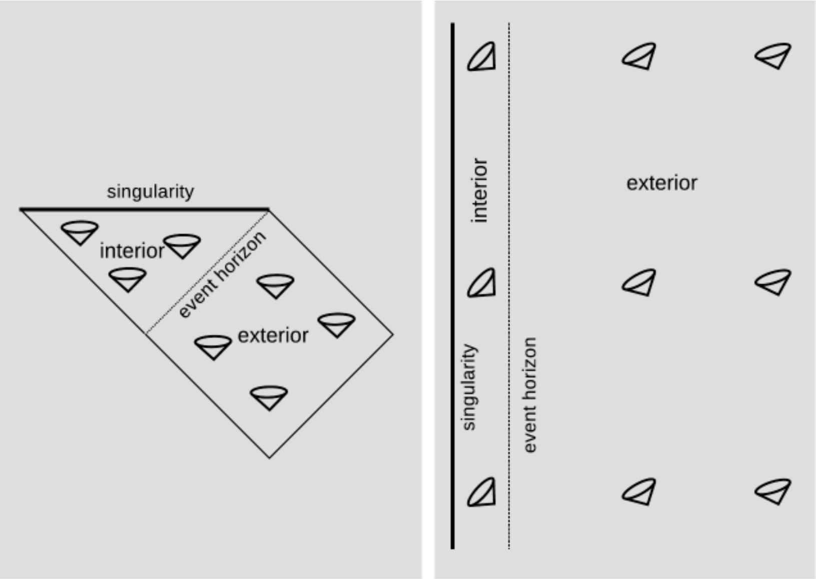


Grid is not square. Sizes are less distorted.

In the version on the left, the latitude-longitude grid is rendered such that where the lines intersect, they always form right angles. We pay for this feature with an extreme distortion of lengths and areas. For example, Greenland looks almost as big as Africa, whereas in reality Africa is much larger.

The version on the right has its own pros and cons. Lengths and areas are much more accurately represented, but the latitude-longitude boxes have funny nonrectangular shapes.

Something exactly analogous happens when we represent curved spacetime on a flat diagram. Two options are shown in the diagram below. The one on the right is just like the one we drew in section 11.4, with the light cones tipping over drunkenly. This tipping is analogous to the non-perpendicularity of the lines of latitude and longitude in the right-hand map of the earth. As in the earth map, we have the advantage that the sizes of things make sense: the interior region of the black hole is of finite size, whereas the exterior region is much bigger — limited in size only by the edge of the page.



The figure on the left is a Penrose diagram for a black hole. Here the light cones are all lined up in an orderly way, which is analogous to the orderly rectangular structure of the grid lines in the left-hand map of the world. This is an overwhelming advantage, since the light cone dictates all of the possible cause-

and-effect relationships. For this reason, professional relativists use Penrose diagrams much more frequently than the ones with tipping light cones. The disadvantage of the Penrose diagram is that sizes are not represented accurately. For example, the triangle representing the interior of the black hole looks half as big as the diamond showing the whole outside universe.

11.6 Tests of general relativity in weaker fields

Figure 1 on the facing page contrasts the predictions of Newton's laws and general relativity for an object whose orbit brings it inside the event horizon of a black hole. They are qualitatively different. According to Newton, the event horizon doesn't even exist, so nothing special happens. The object's orbit is a Keplerian ellipse. But according to relativity, the object can never come back out once it has passed the event horizon. It impacts the singularity at the center.

Relativity also predicts smaller deviations from Newtonian behavior under less extreme conditions. Figure 2 shows that for a body in an elliptical orbit that only comes moderately close to the black hole, the main relativistic effect is a gradual twisting of the ellipse's orientation. This is called precession. By the way, these diagrams were traced from a cool browser-based simulation, which you can play with at `m4r35n357.github.io/orbits/`.

If we are to observe such an effect in reality, we want to look for the strongest possible gravitational field. This makes it logical to examine the orbit of the planet Mercury. Mercury's orbit is significantly elliptical, and just as importantly it comes closer to the sun than any other planet, therefore experiencing stronger relativistic effects.

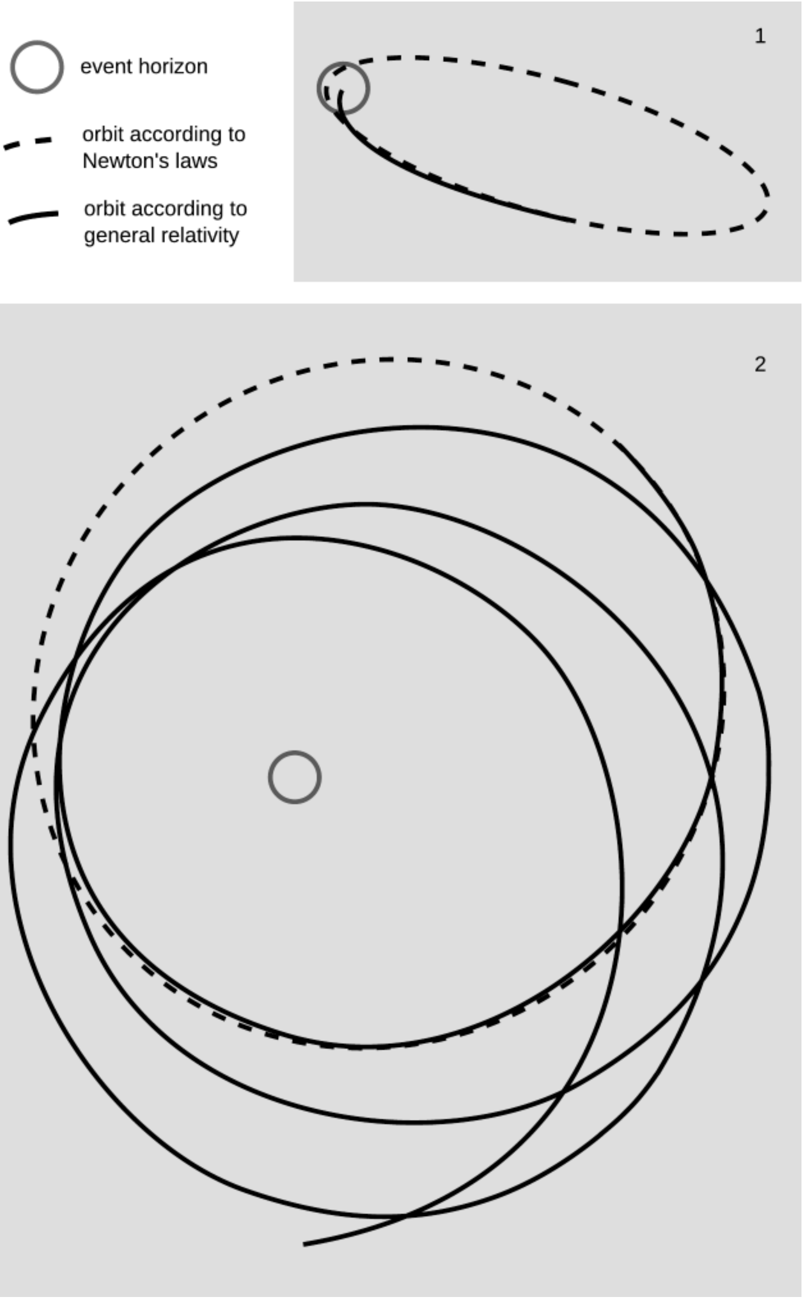
Before Einstein published the general theory of relativity, it had already been observed that Mercury's orbit precessed very slightly, by 1.6 degrees per century. Nearly all of this effect had been explained by Newtonian effects such as the gravitational influence of Jupiter and the slightly nonspherical shape of the sun. But after all of these had been accounted for, there remained an unexplained 0.012 degrees per century. One of the first things Einstein did after finding the field equation of general relativity was to calculate this anomalous precession. It came out right, and he later recalled, “I was beside myself for several days with joyous excitement.”

11.7 Speed of light

The equivalence principle tells us that locally, gravity is an illusion. To any local, free-falling observer, the laws of physics and the properties of spacetime seem absolutely normal. The laws of physics tell us that the speed of light is a universal constant, and an observer sees nothing abnormal in local experiments that measure the speed of light.

But the equivalence principle is only a local thing, so globally things look different. Recall that on a spacetime diagram, the speed of an object is measured by how much its world-line tilts away from vertical, and the speed of light is indicated by a 45-degree tilt. In the diagram in section 11.4, the fact that the light cones are tipped make it look as though light isn’t moving at the speed of light! In fact you are free to tilt your copy of the book to any angle you like, and by doing this you can view any particular light cone the way an observer in that local area would prefer to draw it — upright and law-abiding.

The basic problem here is that when we try to visualize everything-that-is all-at-once, we run up against the hard realities of relativity. An observer O has no technique of measurement that can tell her unambiguously the state of motion of some distant thing T. Such a measurement technique would require sending signals from T to O, and these signals could be distorted by effects such as Doppler shifts. Furthermore, we would have to ask “motion relative to what?” The local points of reference near T are different than the ones near O. A good example of such


event horizon
orbit according to Newton's laws
orbit according to general relativity
1
2

an ambiguity is the ambiguity of parallel transport (section 9.6, p. 126). In any local region of spacetime, we can define velocity vectors for things like basketballs and rays of light. But if we want to ask how a distant observer would perceive such a velocity vector, we have no unambiguous way to answer the question. This would require parallel-transporting the velocity vector over to where the observer is sitting, but parallel-transport is ambiguous and path-dependent.

Chapter 12

Waves

12.1 No action at a distance

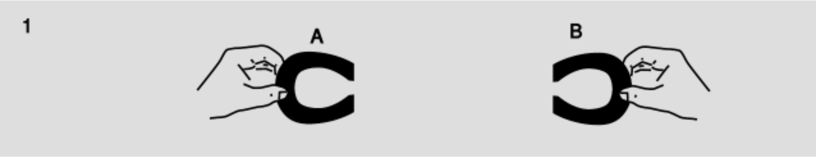
The Newtonian picture of the universe has particles interacting with each other by exerting forces from a distance, and these forces are imagined to occur without any time delay. But as we've seen, relativity forbids this. The universal speed c is the maximum speed at which cause and effect can propagate. This blanket prohibition applies to any force at all. In Star Wars, Obi Wan Kenobi senses the destruction of the planet Alderaan and says, "I felt a great disturbance in the Force, as if millions of voices suddenly cried out in terror, and were suddenly silenced." No matter what kind of force "the Force" is, this is impossible. The planet is light-years away, so any information about its destruction can only arrive after a time delay of years.

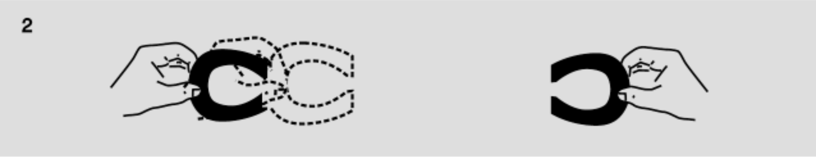
12.2 Fields

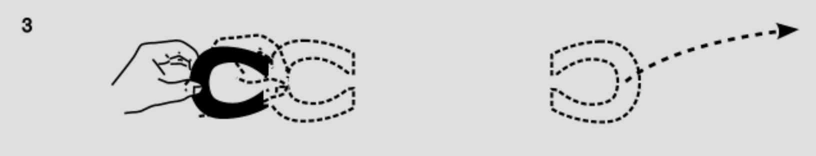
Since forces can't be transmitted instantaneously, it becomes natural to imagine force-effects spreading outward from their source like ripples on a pond, and we then have no choice but to impute some physical reality to these ripples. We call them fields, and they have their own independent existence. Gravity is transmitted through the gravitational field. Besides gravity, there are other fundamental fields of force such as electricity and magnetism. Ripples of the electricity and magnetism are light waves.

Even empty space, then, is not perfectly featureless. It has measurable properties. This concept made a deep impression on Einstein as a child. He recalled that as a five-year-old, the gift of a magnetic compass convinced him that there was "something behind things, something deeply hidden."

The smoking-gun argument for this strange notion of traveling force ripples comes from the fact that they carry energy. In figure 1, Alice and Betty hold magnets A and B at some distance from one another. If Alice chooses to move her magnet father from Betty's, as in figure 2, Alice will have to expend some energy to fight against the magnetic force, burning off some of the calories from that chocolate cheesecake she had at lunch.







But now suppose, as shown in figure 3, that Betty decides to play a trick on Alice by tossing magnet B far away just as Alice is getting ready to move magnet A. We have already established that Alice can't feel magnet B's motion instantaneously, so the magnetic forces must actually be propagated by a magnetic *field.*

Of course this experiment is utterly impractical, but suppose for the sake of argument that the time it takes the change in the magnetic field to propagate across the diagram is long enough so that Alice can complete her motion before she feels the effect of B's disappearance. She is still getting stale information about B's position. As she moves A to the left, she feels resistance, because the field in her region of space is still the field caused by B in its *old* position. She has burned some chocolate cheesecake calories, and it appears that conservation of energy has been violated, because these calories can't be properly accounted for by any interaction with B, which is long gone.

If we hope to preserve the law of conservation of energy, then the only possible conclusion is that the magnetic field itself carries away the cheesecake energy. In fact, this example represents an impractical method of transmitting radio waves. Alice burns calories by moving magnet A, and that energy goes into the radio waves. Even if B had never existed, the radio waves would still have carried energy, and Alice would still have had to do work in order to create them.

Although I referred to magnets in this argument, I never appealed to any properties of the magnetic force, so it follows that force-ripples in *any* field carry energy. For example, suppose that super-powerful aliens, angered by the advent of disco music, come to our solar system on a mission to cleanse the universe of our aesthetic contamination. They apply a force to our sun, causing it to go flying out of the solar system at a gazillion miles an hour.

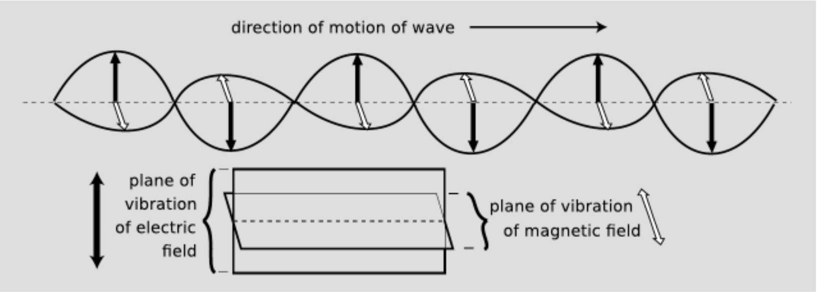
Newton expects the gravitational force of the sun on the earth to *immediately* start dropping off, and and since sunlight takes eight minutes to get from the sun to the earth, the change in gravitational force would be the first way in which earthlings learn the bad news — the sun would not visibly start receding until a little later.

Newton is wrong. Both effects would travel at c, and they would be detected simultaneously on earth.

12.3 Electromagnetic waves

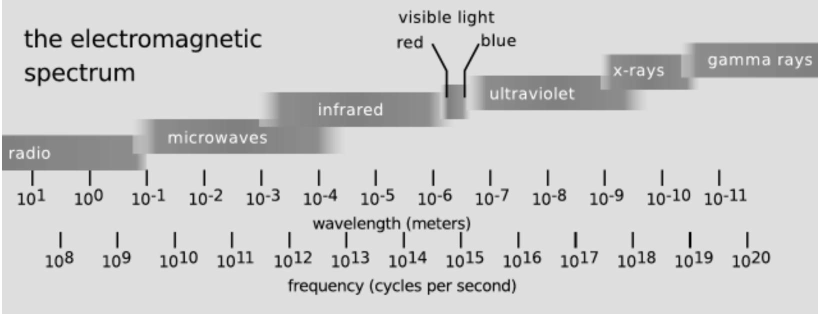
“And God said, Let there be light: and there was light.” Light is pretty important to us, as implied by this passage’s prominent placement in the creation myth. Light is a vibration of the electric and magnetic fields, and to make sense out of that statement, you need to have at least some idea of what electric and magnetic fields are. The operationalist philosophy (section 4.1, p. 63) is that this means knowing how to measure them. A magnetic compass measures the direction of the magnetic field. A crude method of measuring an electric field is to rub a balloon against your hair in order to put some electric charge on it, and then release the balloon; any ambient electric field will cause the balloon to accelerate in a certain direction, which is the direction of the field.

Although physicists didn’t realize it until late in the game, it turns out that electric and magnetic fields are very closely related. If an observer in frame of reference A detects only an electric field, an observer in frame B, moving relative to A, sees a mixture of magnetic and electric fields. Similarly, a field that is purely magnetic in one frame is a mixture in another frame. For related reasons, a changing electric field creates — or “induces” — a magnetic field, and vice versa. This is the principle behind devices such as generators and transformers.


direction of motion of wave
plane of vibration of electric field
plane of vibration of magnetic field

Vibrations in these fields will be emitted by any electric charge that shakes back and forth. For example, in a candle flame, the high temperature indicates rapid vibration of the atoms and charged subatomic particles in the flames. When you transmit radio waves from the radio antenna hidden inside your cell phone, electrons are wiggling back and forth between the ends of a hidden antenna.

Because of the close relationship between the two fields, any vibration of the electric field produces a vibration of the magnetic field, and vice versa. Therefore there are no purely electric or purely magnetic waves. As shown in the diagram, the waves are structured so that the fields vibrate in directions that are perpendicular both to each other and to the direction in which the wave is moving. Unlike sound waves and water waves, electromagnetic waves are not vibrations of any physical medium.

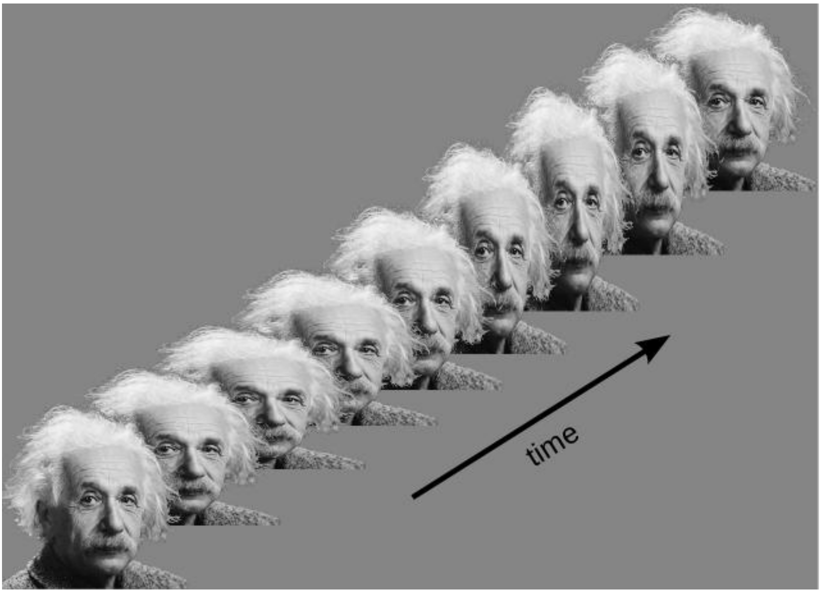


A wave has a wavelength, which is the distance from one crest to the next, and a frequency, which is the number of wave crests that pass over a certain position in one unit of time. A higher frequency means a shorter wavelength. Visible light is an electromagnetic wave that has a frequency within a certain range, with the blue and red ends of the rainbow differing by about a factor of two. Visible light is only one small segment of the vast electromagnetic spectrum, which includes such superficially disparate phenomena as radio waves and x-rays.

Electromagnetic waves carry energy, as you can tell when you use a microwave oven to heat your food.

12.4 Gravitational waves

The equivalence principle tells us that the gravitational field can, at least locally, be considered to be an illusion: for a free-falling observer, there is no field. This suggests the possibility that waves in the gravitational field might also be illusions. In fact, Einstein tried to publish a paper in 1936 making exactly this claim. The paper was wrong, and was rejected based on the report of the Physical Review's anonymous referee. (Einstein held a grudge, and never submitted another paper to that journal.) In nontechnical language, the important point is that although *locally* we can always consider a gravitational field to be fictitious, that doesn't hold on larger scales. On larger scales, we can have tidal forces (section 7.4, p. 100), which are objectively real; for example, they can cause water to rise higher on a beach.



The figure shows what would happen if a strong enough gravitational wave passed through Albert Einstein's head. The effect is to alternately squish and stretch him along two axes perpendicular to the direction in which the wave is moving. Because Einstein's head isn't perfectly flexible, it resists these distortions, and the result is frictional heating; just like electromagnetic waves, gravitational waves transport energy. A more realistic way of detecting a gravitational wave is to replace Einstein's head with a kilometer-long vacuum tube, and monitor any stretching or squishing of the tube using a laser; there is a ground-based project called LIGO already in operation to try to do this, and a European Space Agency project to launch a space-based system called LISA.

There's good news and bad news about detecting gravitational waves.

First the good news. In order to produce strong gravitational waves, you need large masses (such as stars) moving very violently (preferably at speeds close to the speed of light). This means that if astronomers can detect gravitational waves coming from outer space, they can learn about rare awesomely apocalyptic processes, such as the collision of two black holes. Also, gravitational waves do not interact very strongly with matter, so the universe is essentially transparent to them. Our view of the universe through the gravitational-wave spectrum would never be blocked by clouds of gas or dust, as is often the case with electromagnetic waves. In particular, the very early universe (less than about 100,000 years after the big bang) was full of hot, ionized gas, which was completely opaque to light, so gravitational waves are probably the only way we could ever hope to "see" the early universe directly. Direct detection of any gravitational waves would also be a spectacular confirmation of one of general relativity's most important predictions.

The bad news is the same as the good news. Processes violent enough to produce strong gravitational waves are uncommon. If such events happen, they happen rarely and typically at vast distances from us. The weak interaction of gravitational waves with matter also means that detectors made out of matter are not very sensitive to them. Therefore it's very difficult to get such a detector to work well enough to give a definite blip. That's why it took until 2016, a hundred years after Einstein predicted gravitational waves, for their first direct detection, by the LIGO collaboration. The detector was actually two detectors, one in Louisiana and one in Washington state. Two were needed because each of the detectors by itself is essentially the most sensitive vibration detector ever built, and would be overwhelmed by the ordinary background of subsonic rattles and buzzes caused by sources like wind, animals, and passing trucks. With two detectors, physicists could look for a vibration that occurred simultaneously in both, thousands of kilometers apart. The event that caused the 2016 signal is believed to have been the collision of two black holes, about a billion light-years away. For the brief time leading up to the collision, the inward-spiraling black holes radiated energy in the form of gravitational waves at a rate that is believed to have been greater than the power emitted as visible light by all the stars in the visible universe.

Part IV

Cosmology

Chapter 13

For dust thou art

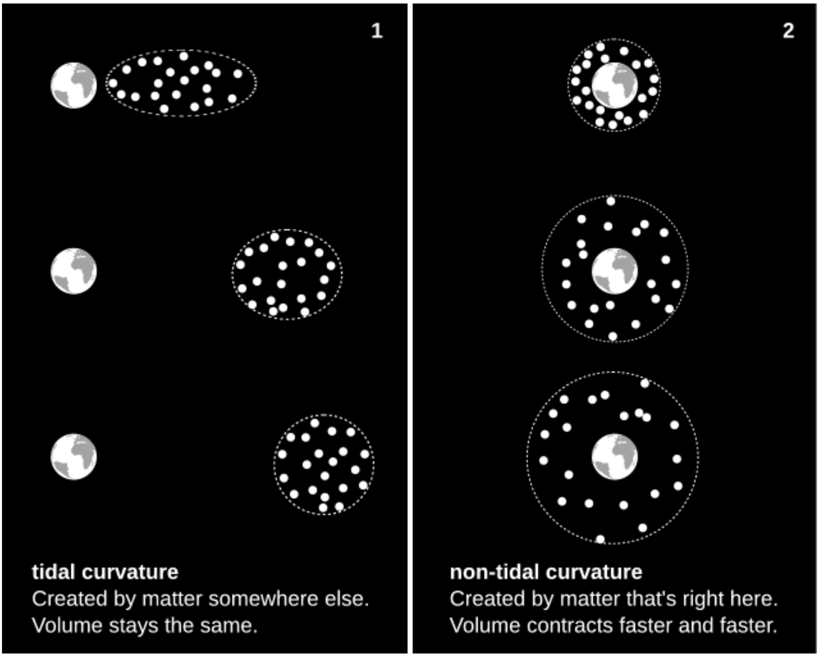
> In the sweat of thy face shalt thou eat bread, till thou
> return unto the ground; for out of it wast thou taken:
> for dust thou art, and unto dust shalt thou return.

Replace "dust" with "atoms," and this verse from Genesis tells a physics story in which matter is conserved, and atoms are recycled as they pass through various forms. "Dust" also has a technical meaning in relativity: a swarm of material objects, all moving relative to each other at speeds that are small compared to c. A flock of galaxies, for example, would be "dust" to a cosmologist, with the galaxies as the dust particles. In this chapter we look at cosmological models in which the matter is all in the form of dust, which until about 1998 was thought to be a a realistic assumption.

13.1 The Einstein field equation

In John Wheeler's formulation (ch. 10, p. 129), "spacetime tells matter how to move, and matter tells spacetime how to curve." So far, we've applied these ideas to black holes, but to apply them to cosmology we need to flesh out Wheeler's description a little more. Whereas a black hole is 100% empty space, the cosmos is more of a uniform soup. Whether we look at a black hole from a safe distance or are brave enough to dive in past the event horizon, the effects we see are all due to the matter contained in the singularity, which is at a distance from us. But when we study the cosmos, we can't go outside and observe it from far away — as far as we know, there is no outside. We're inside, like a fly in the soup.

This distinction between matter-here and matter-over-there is directly coded into the Einstein field equation, which I'll describe here in words and pictures rather than mathematical symbols.


1
2
tidal curvature
Created by matter somewhere else.
Volume stays the same.
non-tidal curvature
Created by matter that's right here.
Volume contracts faster and faster.

In figure 1, someone releases a bag of marbles near the earth, in a spherical cloud that is initially at rest and at some distance from our planet. As in the real-world example of Comet Shoemaker-Levy (section 7.4, p. 100), the cloud stretches out due to tidal stresses; the marbles that start out closer to the earth experience stronger gravity and accelerate more quickly than the ones farther away. But this is not the only effect on the shape of the cloud. The whole cloud is angling in along a cone whose tip is at the center of the earth, and therefore the cloud simultaneously gets *narrower* in the transverse direction. The result turns out to be that although the cloud is distorted into an ellipse, its volume stays exactly the same.

The cloud of marbles in figure 2 has been released so that it encompasses the earth. This is gravity due to a mass that is present, not absent. The cloud contracts in all three dimensions. Its volume gets smaller, and its rate of contraction gets faster as time goes on, because the marbles are *accelerating*, not just moving at constant speed.

Although these pictures are presented in terms of a cloud of physical particles, their only role was to free-fall along geodesic paths and let us visualize that set of geodesics. These world-lines are initially parallel, because the marbles start at rest, but as time goes on they start to converge. When lines start out parallel but later start to converge or diverge, we're dealing with curvature (recall section 9.3, 122). What figures 1 and 2 really show us, then, is a distinction between two types of *curvature*: tidal curvature caused by matter-over-there as distinguished from non-tidal curvature caused by matter-right-here.

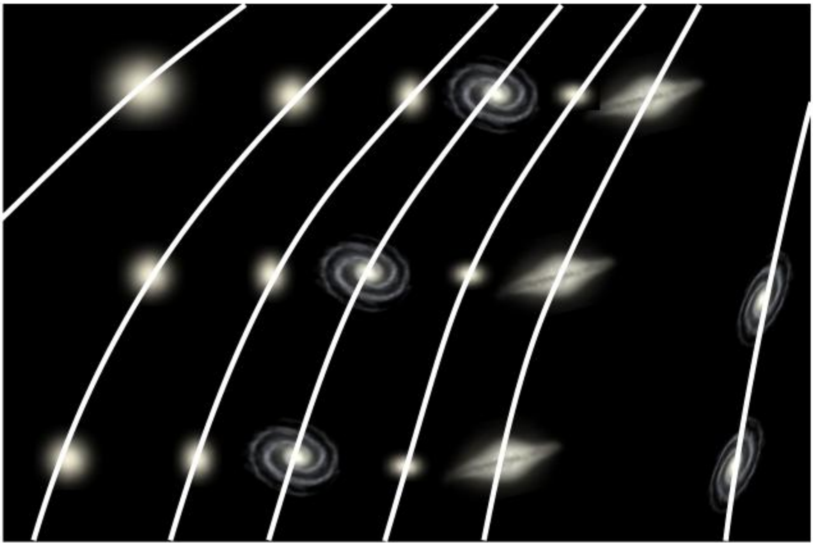
The main idea of the Einstein field equation, translated into words, is that it relates the non-tidal curvature to the matter-right-here. Ordinary matter with positive mass-energy creates non-tidal curvature that is positive, meaning that a cloud tends to pull together. If it's initially at rest, it contracts faster and faster. If it's initially expanding, its expansion slows down; it may or may not reach a point where it stops and recontracts.

If there is no matter-right-here (other than that of the marbles themselves, which we assume to be negligible), we have only tidal curvature. A common reason for tidal curvature is that there is some matter-over-there, as in figure 1. This is not the only possibility, however. For example, gravitational waves are also a type of tidal curvature; perhaps you noticed the similarity between figure 1 and the illustration on p. 160 of gravitational waves stretching and squishing Einstein's head.

How do we measure "matter" for the purposes of the field equation? Roughly speaking, we should use mass-energy, but in fact the *pressure* of the matter also comes into play in a mathematical way that is too complex to describe here. It turns out, however, that this pressure effect can be safely ignored if the relative speeds of the matter particles are small compared to c, which is part of what we mean by the definition of dust.

13.2 Conservation

Many of the most fundamental laws of physics we have are conservation laws, which say that stuff can't just appear or disappear. For example, let's say in a cosmological context that "stuff" means galaxies. Of course there is no specific law of conservation of galaxies, but counting galaxies is easy visually, and if an entire galaxy did spontaneously wink out of existence (scary!), it would certainly violate conservation laws such as conservation of energy-momentum.



The figure shows a simple way of visualizing this on a spacetime diagram. Galaxies are entering and exiting the stage, but that doesn't mean they're being spontaneously created or destroyed. In fact we can verify the conservation of galaxies by counting how many world-lines enter the rectangular figure and how many exit. Seven come in (six through the bottom and one through the left side) and seven go out (six out the top and one out the right side). Because 7 = 7, we have conservation of galaxies.

This counting technique is in fact a mathematically precise and complete way of defining some of the fundamental conservation laws of physics, such as conservation of electrical charge. When you use a laptop computer, you aren't using up the charge of the battery, although people may say it that way in popular speech. The same electrons stay in the laptop computer the whole time; they only get shuffled around, not created or used up. If we imagine hypothetically the physically impossible situation in which one of the electrons gets used up, it would be like a situation in which one of the world-lines simply terminated in the middle of the picture. That world-line would have entered the picture but never exited, and the mismatch in the two counts is how we would detect the violation (and presumably get the Nobel prize for it).

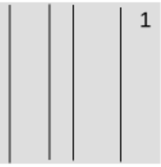
I don't know of a nonmathematical argument to prove it, but conservation of energy-momentum is directly implied by the Einstein field equation; without it, the equation breaks down and becomes mathematical nonsense. This is a firm prediction of general relativity, so if any observation ever does disprove conservation of energy-momentum, even by some tiny amount, then general relativity will have been disproved as well.

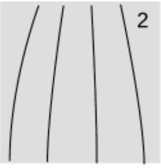
Conservation of energy-momentum is similar to the other examples above, but with the difference that energy-momentum is a vector, not a scalar (section 4.3, p. 65, and section 5.2, p. 72). This leads to an important loophole in general relativity's conservation of energy-momentum. If we draw our rectangular diagram too big, then we have a problem verifying that the energy-momentum vectors coming into the box add up to the same amount as what emerges. This is because if we want to compare two vectors, we have to do it using parallel transport (9.6, p. 126). But when spacetime is curved, parallel transport is ambiguous. Therefore the conservation law only holds if we draw a box small enough so that the curvature of spacetime doesn't matter. In other words, conservation of energy-momentum in general relativity is only a *local* conservation law.

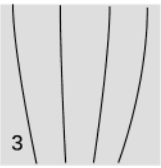
13.3 Change or changelessness?

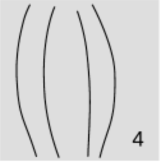
Now we have two tools: the Einstein field equation and conservation of energy-momentum. When Einstein invented these in 1915, scientists had a preexisting preference for uniformitarianism, which was the idea that the universe had always existed, would always exist, and would, on the average, always look the same. Uniformitarianism was a reaction against catastrophism, a natural history dominated by dramatic events such as massive volcanic eruptions. The debate between the two schools of thought had revolved around scientific questions such as whether California's spectacular Yosemite Valley had been formed by a sudden subsidence or slow glacial erosion. The subtext was a suspicion that catastrophism was an attempt to shoehorn Biblical events such as Noah's flood into the geological record.

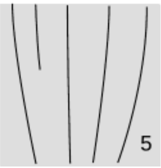
The Einstein field equation doesn't allow for a static cosmos with dust as its contents. Suppose that we had such a universe, and consider a small region within it, containing a sampling of galaxies (dust particles). The definition of dust tells us that the galaxies are moving at nonrelativistic speeds relative to their neighbors, so we can adopt a frame of reference tied to one of the galaxies, and in this frame the speeds of the others will also be approximately zero, as in spacetime diagram 1 in the figure. We've picked out a set of particles, initially at rest, very much like one of the clouds of marbles in the figures on p. 166. But unlike the marbles, our galaxies have a significant amount of mass of their own. That's matter-right-here, so the cloud will start to shrink faster and faster, as shown in figure 2. But this is a contradiction to our assumption that the universe was unchanging. What general relativity *does* allow is a universe that is contracting, expanding (figure 3), or expanding and then recontracting (figure 4).











Einstein realized this early on, but the idea of a contracting or expanding universe had never been suggested on any theoretical or empirical grounds, so he rejected it and focused his ingenuity on fixing what he saw as a problem with his theory. It was not until the 1930s that evidence began to appear that the universe actually was expanding (section 3.10, p. 58). The theory was worked out in full by the Russian physicist Alexander Friedmann, and the Belgian priest Georges Lemaître proposed that if one extrapolated backward, there must have been a time when all of the matter in the universe had emerged from a “cosmic egg” — what we now call the big bang.

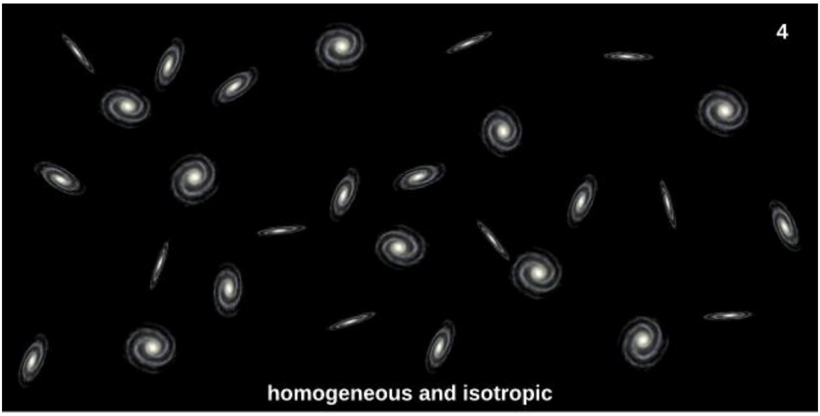
Einsten hated it, and told Lemaître, “Your calculations are correct, but your physics is abominable.” In 1948, the British astronomers Bondi, Gold, and Hoyle came up with an alternative, shown in figure 5, called the steady-state model. According to this idea, the universe expands, and everything gets farther apart, but at the same time atoms are spontaneously springing into existence, so that on the average, the density of matter stays the same. Thus although the universe isn’t static (unmoving), it looks the same at all eras of history. One problem with the theory was that the proposed creation of matter is exactly the kind of nonconservation of energy-momentum that is forbidden by general relativity. Although Hoyle went on trying to get around this and other difficulties with the theory, it was eventually conclusively disproved (see section 13.6, p. 176).

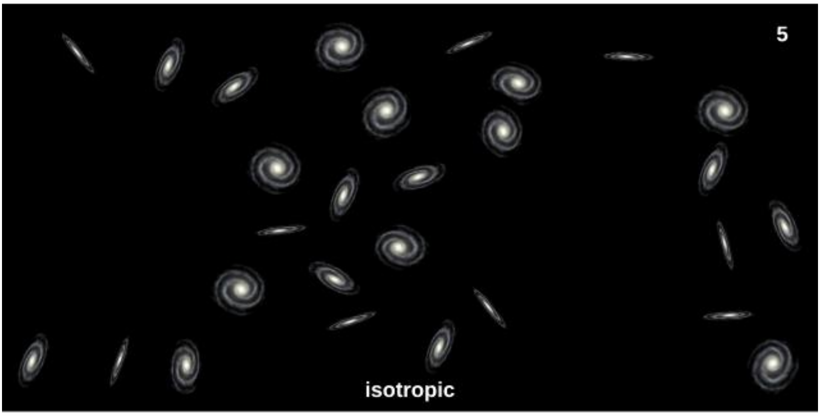
1. Vesto Slipher (1875-1969) used Doppler shifts to find the motion of other galaxies.
2. Henrietta Swan Leavitt (1868-1921) found methods for determining how far away other galaxies were.
3. Edwin Hubble (1889-1953) demonstrated that the universe was expanding.

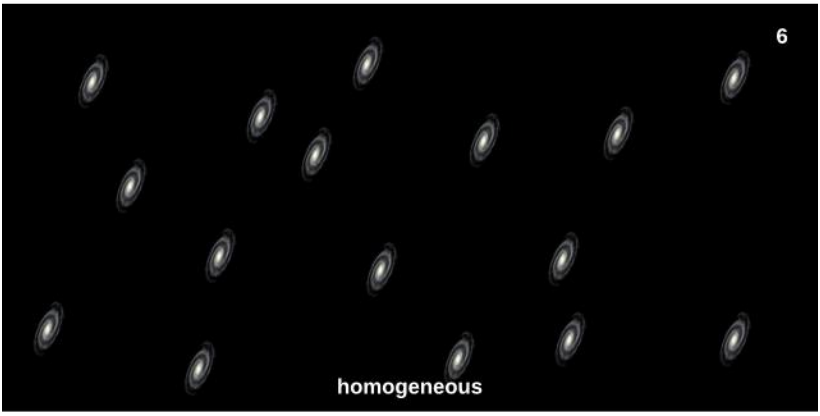
13.4 Homogeneity and isotropy

The night sky looks like a bowl: the eye has no way of judging depth. But once astronomers such as Henrietta Swan Leavitt had developed the appropriate techniques, it became possible to map out the structure of the universe on large scales — the scales on which we could consider galaxies as the "dust" grains in our models. As surveys of the sky became more sensitive, these maps were able to cover larger and larger distance scales, like a cosmic zoom outward. Structure is evident on these maps, including black empty spaces like the holes in Swiss cheese and stringy structures like spiderwebs. These voids and filaments are believed to have been formed by gravitational collapse that occurred during and despite cosmological expansion. But on the very largest scales, beyond a few hundred million light years, we no longer see any structure. The map looks like an undifferentiated, random scattering of galaxies.

It is therefore natural to simplify our cosmological models by making the approximation that the universe is *homogeneous* on the average, i.e., no region looks different from any other region. Furthermore, it does not appear that the universe has any special axis or preferred direction, such as an axis of rotation or a special direction with which spiral galaxies like to align themselves. This additional symmetry is called *isotropy.* The figures on the facing page show visual examples that demonstrate these two symmetries.


4
homogeneous and isotropic


5
isotropic


6
homogeneous

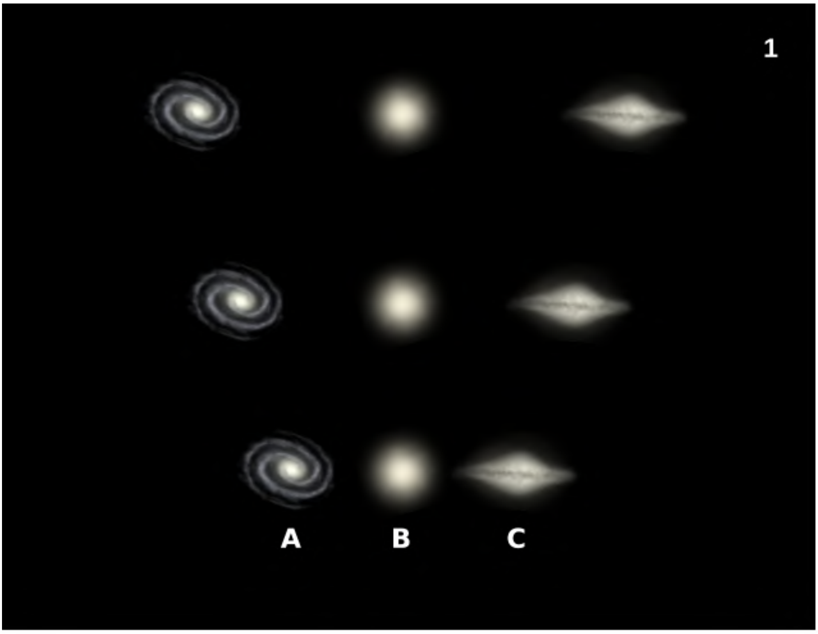
13.5 Hubble's law

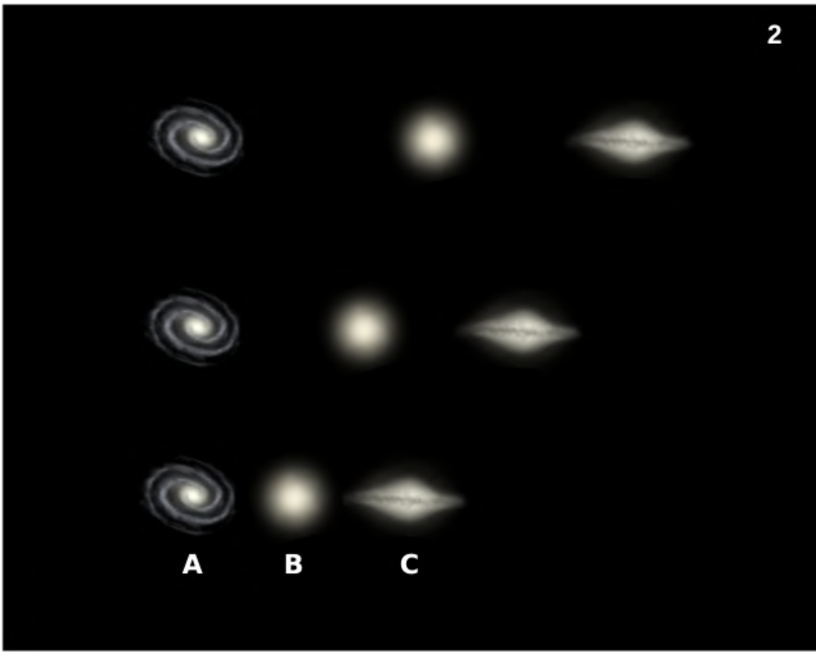
The spacetime diagram in figure 1 on the facing page shows an expanding universe, simplified by focusing on three galaxies that happen to start out equally spaced along a line. This is the frame of an observer in the middle galaxy, B. Galaxies A and C each start out one unit of distance away from B, and over the course of the time shown, they end up two units away.

Why do distances AB and BC both have to grow by the same amount over the same time interval? They don't, and in reality they probably wouldn't, if we were actually talking about three nearby galaxies. But consider this instead as a cartoon representing the structure of the universe on the largest scales. Then by homogeneity, both of these distances have to expand at the same rate.

Figure 2 shows the same situation in a frame where galaxy A is at rest. Initially, observer Alice in galaxy A sees B as being one unit away, and C as lying at a distance of two units. These distances of 1 and 2 double to 2 and 4. Alice says that B is moving away from her at some speed, and C is fleeing at double that speed. (Of course Alice won't live long enough to watch the process play out. She will actually measure these speeds using Doppler shifts — the technique pioneered by Slipher.) The velocity of recession v is proportional to the distance d.

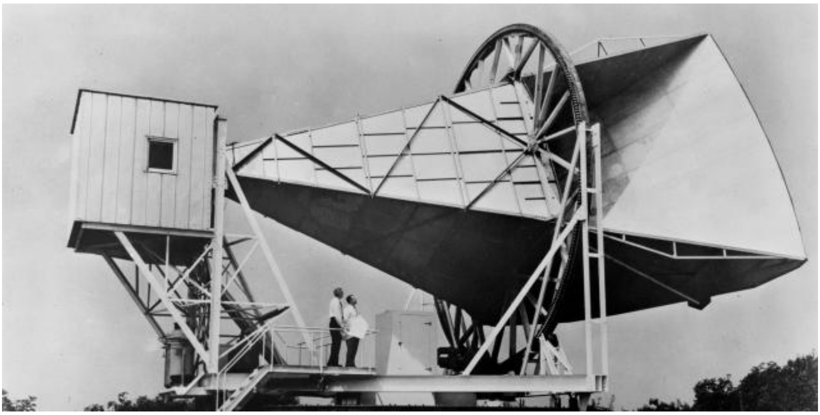
This proportionality is known as Hubble's law, after Edwin Hubble, who worked at the Mount Wilson observatory near Los Angeles. The constant of proportionality is called the Hubble constant, H, so that we have $v = Hd$. Suppose for the moment that, as implied by the diagrams, the motion of each galaxy is inertial; gravity is negligible. Then by using the grade-school equation (speed) = (distance)/(time), we find that if we extrapolate the expansion backward in time, there was a time equal to $1/H$ in the past when all three galaxies were in the same place. That is, the Hubble constant can be interpreted as the inverse of the age of the universe, provided that the expansion has been at a constant rate.


1
A
B
C


2
A
B
C

13.6 The cosmic microwave background

Despite scientists' initial reluctance to accept the big bang model, many independent threads of evidence point to it. For example, our universe has quite a bit of hydrogen-2, also known as deuterium. This is an isotope of hydrogen in which the nucleus has one proton and one neutron. The present-day universe has physical processes in the cores of stars that steadily destroy deuterium, but no processes that produce it. If the universe had been infinitely old and always in its present state, as in the steady-state model, all of the deuterium should have long since been eliminated.



The steady-state model finally went to its grave in 1964 when Arno Penzias and Robert Wilson at Bell Labs in New Jersey found an annoying source of background radiation in the signals picked up by the microwave antenna, shown in the photo, which they had hoped to use for satellite communications and radio astronomy. They tried to fix their apparently broken instrument by rousting out the pigeons that had been nesting in it, and cleaning out the droppings. The signal was still there. Was it human-generated, like radio waves from power lines or cars' spark plugs? But that type of noise would typically change in intensity between day and night. Not only did the signal fail to vary with a

24-hour periodicity, but it also didn't show the telltale variation over 23.93 hours (a sidereal day, the time it takes the earth to spin on its axis) that is characteristic of emissions from a source at one location in the sky.

Eventually they became convinced that these radio waves were coming from outside our galaxy, and uniformly from all directions in the sky. Communication with the physics community at nearby Princeton led them to realize that they had detected the afterglow of the big bang, Doppler shifted from the visible spectrum down into radio. The big bang model accounts not only for the existence of the radiation but also for its detailed characteristics, such as the variation of the intensity with wavelength. The steady-state model is unable to account for these facts. The radiation is referred to as the cosmic microwave background, or CMB.

13.7 A hot big bang

These days the study of the CMB is an entire scientific field of its own, as technical and sophisticated as the study of wine, ballet, plumbing, or rock climbing. Without delving into the details too deeply, we can pick out some straightforward ideas. It's a glow, and hot things glow, so we suspect that the universe used to be hot. The temperature scale we use for this is the Kelvin scale, whose zero is *absolute* zero, the temperature at which all random motion ceases completely. A Kelvin degree (K) is the same size as a Celsius degree, so the Kelvin and Celsius scales are the same except for an additive constant. Everyday temperatures that we experience on our planet are in the ballpark of 300 K.

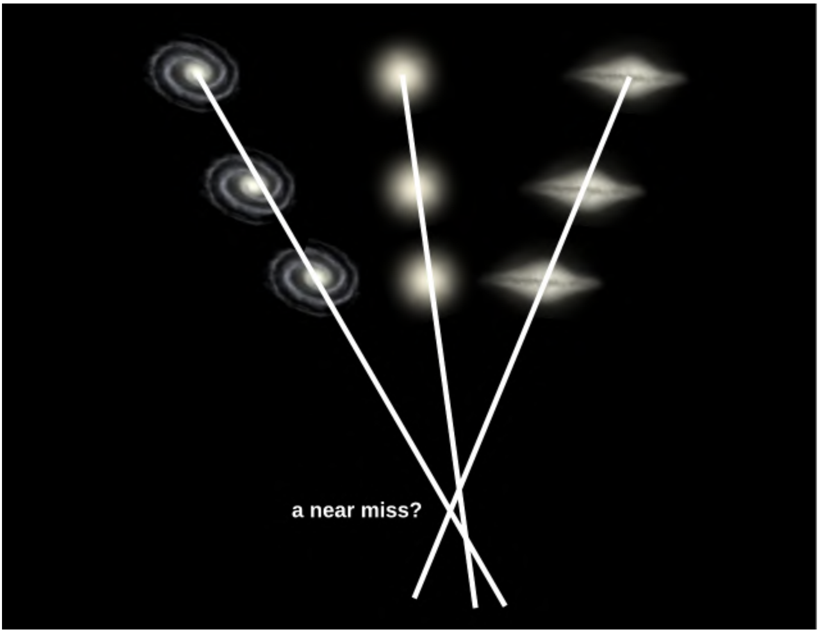
The mixture of wavelengths emitted by a glowing object depends on its temperature. Right now, your body is emitting infrared radiation, which means light with a wavelength a little too long to be visible without infrared goggles. A hot poker glows red. If you heat it more and more, the mixture of wavelengths it emits shifts toward lower and lower wavelengths. Presently, the CMB has a mixture of wavelengths that is typical of a temperature of only 2.7 K, but these wavelengths are not the ones originally emitted. General relativity predicts that as the universe expands, and galaxies get farther apart, the space between them also expands. Because of this expansion, when a light wave travels across the universe for billions of years, its wavelength stretches.

Since most of the normal matter in the universe is hydrogen, we expect that this light was originally emitted by hot hydrogen gas at a temperature of about 3000 K, which is the highest at which hydrogen is transparent; at higher temperatures, the electrons are stripped off of the atoms, and the gas becomes opaque. So the CMB is a relic of an era in which the universe had a temperature of about 3000 K. When originally emitted, the CMB has a yellowish-orange color. The drop of a factor of about 1000 in temperature between then and now indicates that the wavelengths have expanded by 1000 times, and therefore that cosmological expansion since that era amounts to a factor of 1000.

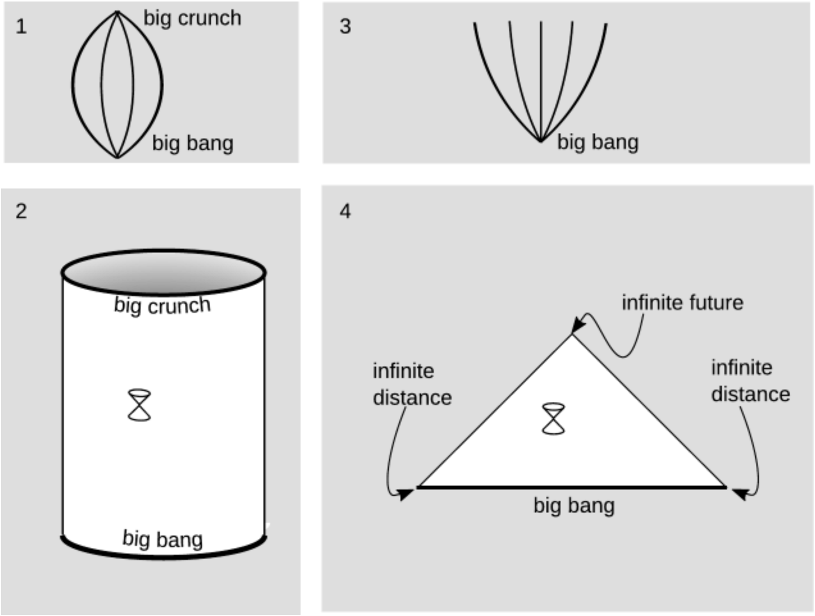
We are led to cosmological models in which the big bang was *hot*. Presumably it was even hotter at times before the CMB was emitted. Support for this extrapolation comes from the mixture of chemical elements and isotopes of those elements that we observe in the present universe. We noted on p. 176 that deuterium (hydrogen-2) is currently being depleted by nuclear reactions in stars. That means that the universe's present supply of deuterium must have originated from nuclear fusion reactions in the first few minutes after the big bang, when the temperature was millions of degrees kelvin.

13.8 A singularity at the big bang

It seems as though the universe keeps getting hotter and denser as we extrapolate back in time closer and closer to the big bang. When Friedman first constructed isotropic and homogeneous cosmological models using the Einstein field equations, he always got results in which the temperature and density blew up to infinity at the big bang. This would indicate a *singularity*, similar to, but different from, a black hole singularity. Cosmologists were at first skeptical as to whether these singularities were real. The assumption of homogeneity was after all only an approximation. They expected something more like the situation shown in the figure, where instead of a big bang we have a big near miss. This expectation was natural based on their knowledge of the universe, because in general astronomical objects are tiny compared to the distances between them, so they make small targets, and collisions are extremely rare.



But in the 1970s, relativist Stephen Hawking (b. 1942) proved the Hawking singularity theorem, which tells us that there was a big bang singularity provided that three assumptions are satisfied: (1) general relativity and the Einstein field equation are valid, (2) matter behaves in a reasonable way by obeying certain energy conditions (p. 134), and (3) present-day conditions in the universe fall a within certain range of parameters. (A similar theorem was proved by Penrose for black hole singularities.) Condition 3 has been verified definitively through astronomical observations. Condition 2 is believed to have been valid in the early universe. Condition 1 is expected to hold up to certain known, very high densities and temperatures, at which quantum mechanics is expected to invalidate general relativity.



Like a black hole singularity, the big bang singularity is spacelike. A black hole singularity becomes an inevitable part of the future light cone of an observer who passes inside the event horizon, but other observers never come in contact with it. The big bang singularity, on the other hand, lies in the past light cone of every observer.

The figures show two possible scenarios according to Friedmann. In figure 1, we see the world-lines of several objects in a universe that has enough mass in it so that gravitational attraction gradually slows down the expansion of the universe and brings it to a halt. After that, the universe recontracts and ends up going out in style, in a fiery implosion called, naturally enough, a big crunch. Figure 2 shows the same cosmology as a Penrose diagram. Because Penrose diagrams don't try to show times and distances to scale, the expansion and recontraction are not shown. What the diagram does show accurately is that in this type of universe, space is finite, and it wraps around on itself.

In figures 3 and 4 we see a possibility in which the universe does not contain enough mass to cause it to recontract. Its expansion is analogous to the motion of a projectile that has been shot up from the earth's surface at greater than escape velocity, so it will never fall back down. This universe is infinite both in space and in time.

The Penrose diagram in figure 4 looks exactly like the one in figure 1 on 147, for the future half of the flat spacetime of special relativity. The past half is missing, because there was no time before the big bang. These two spacetimes are not actually the same — one is curved and the other is flat. But Penrose diagrams aren't designed to show every aspect of curvature. The two spacetimes *do* have the same structure in terms of light cones, and cause and effect.

Figures 2 and 4 accurately show that the big bang was not an explosion that happened at one point in space. The big bang happened everywhere at once. In the case where the universe is infinite, figure 4, it has always been infinite and always will be.

The triangular shape of figure 4 makes it look as though everything in the universe is getting closer together over time, which would be a contraction rather than an expansion. But remember, Penrose diagrams aren't supposed to show distances on a consistent scale. Similarly, the parallel painted lines on the highway in the photo on p. 146 *look* like they're getting closer together, but actually they're not; even if the people who painted the lines had decided to make them get gradually farther and farther apart, they would still appear to the eye to converge at a vanishing point.

It may seem strange to imagine an infinite universe that is expanding. If infinity is already bigger than anything, how can it get bigger than that? This is similar to a famous paradox proposed by the mathematician Hilbert. Suppose we have a hotel with infinitely many rooms, and each of them is occupied by one of the infinitely many dentists attending a convention. While ToothCon is still going on, a soybean farmers' convention begins, and there are also infinitely many farmers attending. If the hotel is the only one in town, and it's already full of dentists, where can the farmers stay? We can accomodate them in the same hotel. All we have to do is move the dentist from room 1 into room 2, and similarly from room 2 to 4, 3 to 6, and so on. Now only the even-numbered rooms are occupied, and the farmers can check in to the odd ones. Similarly, we can have a universe with an infinite supply of cubic centimeters of space, but still have it make sense to talk about doubling its volume over a certain period of time.

Regardless of whether the universe is finite or infinite, the part of it that we can presently *observe* is finite, because it's confined to our past light cone. This is called the *observable universe.* The edge of the observable universe is at a distance from us such that information traveling at the speed of light would have just arrived here today if it had been emitted immediately after the big bang.

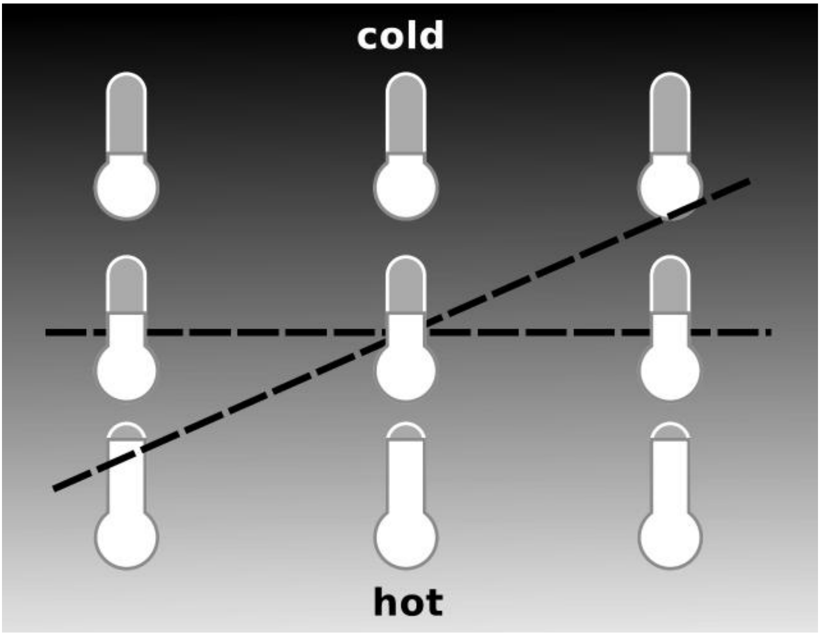
A common solecism is to fail to distinguish between the universe and the observable universe, and to use the same word "universe" for both. For example, there are about 100 billion galaxies in the observable universe, but people will sometimes say that there are this many galaxies "in the universe." This is incorrect. The universe may be infinite, and in that case the number of galaxies in it is presumably infinite as well.

13.9 A cosmic calendar

I've been blithely referring to ideas like the age of the universe and the time at which the CMB was emitted. When we use words like "age" and "time," we need to be careful, because relativity tells us that time is not absolute. For example, the age of the universe is currently estimated at 14 billion years, but whose clock would we use to measure that time? It makes a difference what state of motion the clock is in.

But because the universe has been evolving over time, we have a natural way of getting around this ambiguity. For example, the average temperature of the universe (meaning the average temperature of all the matter and light in it) has gone down over time. In the figure on p. 184, we have a spacetime diagram that shows this cooling process. All three thermometers are at rest relative to the average motion of the nearby galaxies. By homogeneity, all three thermometers measure the same process of cooling, which happens at the same rate according to a clock attached to the thermometer. The result of all this is that we get a preferred notion of simultaneity, indicated by the horizontal dashed line in the figure. We are free to pick some other notion of simultaneity, as suggested by the slanted line, but this is less natural and convenient.

This whole process is similar to how we can estimate the time of year by looking out the window. If there's snow on the ground in my front yard, it's probably December, not July.


cold
hot

Chapter 14

Dark matter and dark energy

Some say the world will end in fire,
Some say in ice.
Robert Frost

How will the world end? Nuclear war? Environmental ruin? Zombie apocalypse? Surprisingly, science has an easier time answering this question for the universe as a whole than for our own planet, which we can so easily touch, see, and feel. How, then, will the *universe* end its days?

14.1 A difficult census

If gravity didn't exist, then the answer would clearly be neither fire nor ice but dilution. Everything is moving farther apart, and as time goes on it gets more spread out. The lumps of matter are separated by more and more empty space. If the expansion continues indefinitely, and at a constant rate, then the universe will eventually get to be like the thinnest soup imaginable.

Gravity does exist. Suppose, as was believed for most of the 20th century, that the universe's matter consists mostly of ordinary atoms (not, say, exotic subatomic particles), which move nonrelativistic speeds relative to other nearby matter. It then qualifies as the “dust” of ch. 13. The Einstein field equation tells us that the expansion is not at a constant rate. Since dust satisfies certain energy conditions (p. 134), the result is as we would expect according to Newton’s description of gravity as an attractive force. The expansion decelerates. We then have the two possible scenarios described in section 13.8, p. 178. Either the universe recollapses into a big crunch, or it goes on expanding forever. To figure out which of these would happen, astronomers devoted a great deal of effort in the 20th century to trying to measure the average density of the matter in the cosmos. Essentially they were trying to take a census of the universe’s contents.

14.2 Dark matter

In our own solar system, the sun accounts for 99.8% of the mass. (The earth’s contribution is only a few ten thousandths of a percent.) Stars are bright objects, so we’d imagine it would be easy to peer through a telescope and map out almost all of the universe’s mass.

If only it were so simple. As discovered by Vera Rubin in the 1970s, stars orbit within their galaxies at high speeds, and if the only mass present was that of the other stars, there would not be enough gravitational attraction to keep them in orbit. They would be ejected, like the sparks flying off of the grinding wheel in the photo. Similar observations apply at the scale of clusters of galaxies. There must be a large amount of nonluminous matter to hold these structures together.

Now the weirdness starts. We have multiple independent lines of evidence telling us that this mysterious matter can’t be made out of atoms:

1. If the universe contained this much matter in the form of atoms, then nuclear fusion reactions in the first few minutes after the big bang would have proceeded more efficiently,

and the universe would contain a higher proportion than is actually observed of elements such as helium.

2. The CMB is very nearly uniform, but does contain tiny irregularities, which are clues as to what happened in the early universe. The existence of non-atomic matter is required in order to make models that reproduce these fluctuations.

3. In merging galaxy clusters, it has been observed that the gravitational forces don't point in the direction they should if they were being produced by the clouds of gas made of normal, atomic matter.

Because we don't know what this exotic matter is, we just call it "dark matter." Our best guess is that it consists of some kind of subatomic particle, similar to the neutrino (which we met on p. 89), that doesn't interact strongly with ordinary matter. (We know that it can't actually be neutrinos, though, because we can detect neutrinos, and there aren't enough of them flying around.)

When the Large Hadron Collider came online in 2010, it was hoped that it would create a slew of new particles, some of which might be the same as the ones in dark matter. That didn't happen, so we remain ignorant. A number of experiments, buried in deep mineshafts to shield against background noise, are under way to detect dark matter directly. As of 2013, the most sensitive experiment has given null results.[1]

14.3 Dark energy

As if this wasn't weird enough, astronomers determined in 1998 that the acceleration of the universe was *accelerating* rather than decelerating. The original evidence came from the Doppler shifts of supernovae (exploding stars) in distant galaxies. Because the light from such galaxies has taken billions of years to reach us, it gives us a window into the past, and allows us to determine how fast the universe *used* to be expanding. Unexpectedly, the data showed that the universe used to be expanding more slowly, and has now sped up.

The statistical quality of the supernova data was not very solid, and many people, including me, were skeptical. Extraordinary claims require extraordinary evidence. But since then, the acceleration has been confirmed by two other independent methods (analysis of CMB fluctuations, and a method called baryon acoustic oscillations).

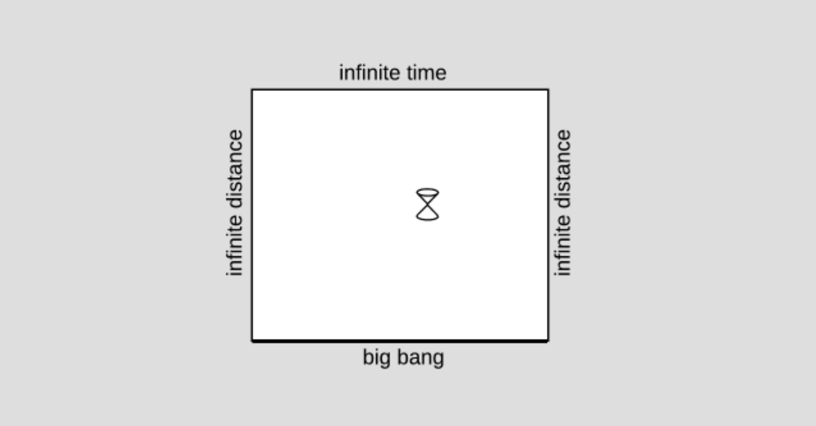
The Einstein field equation tells us that this cannot happen as a result of the gravitational fields of ordinary matter or dark matter. These forms of matter obey the energy conditions — roughly speaking, their mass is positive, rather than negative — and therefore they create a gravitational *attraction* that would *decelerate* the expansion of the universe. The acceleration must be caused by something else, which we refer to as *dark energy.* According to current estimates, the universe is about 68% dark energy, 27% dark matter, and 5% matter made of atoms.

Unlike dark matter and ordinary matter, dark energy is believed to be nearly uniformly distributed across the universe; if it were concentrated in galaxies, then it would cause those galaxies

[1] arxiv.org/abs/1310.8214

to fly apart. In our current models, dark energy is described as a kind of energy that is automatically built in to the structure of empty space. As the universe expands, more space is produced, and this leads to the production of more dark energy, which fuels further expansion. (Surprisingly, the careful mathematical analysis of this process shows that it does not violate the local conservation of energy-momentum described in section 13.2.) In the early universe, the gravitational fields were dominated first by light and later by matter, but we have now entered the universe's final era, in which the dominant effect is that of dark energy.

The figure below shows the Penrose diagram for a universe that, like ours, ends up dominated by dark energy.



In closing, it's worth remarking on the fact that in relativity and cosmology, as is often the case in the sciences, we learn definite answers to questions that most people would imagine science could not answer, while conversely we know very little about things that laypeople would think were elementary and well established. We know all about the detailed history of the universe and its ultimate fate, questions that would traditionally have been relegated to myths or religion. At the same time, we turn out to know nothing at all about what 95% of the universe is made of.

Photo credits

Cover *Clock face:* Wikimedia commons user Lexlexlex, retouched by user Any23cu. **Cover** *Ladybug:* Redrawn from a photo by Wikimedia Commons user Gilles San Martin, CC-BY-SA. **14** *Runner:* Redrawn from photos by Muybridge, 1887. **17** *Car:* Redrawn from a photo by Wikimedia Commons user Auregann, CC-BY-SA. **17** *Cow:* Redrawn from a photo by Wikimedia Commons user Cgoodwin, CC-BY-SA. **18** *Plane:* Redrawn from a photo by Wikimedia Commons user Andy Mitchell, CC-BY-SA. **22** *Cow and car:* As for the figure on p. 17. **35** *Muon storage ring at CERN:* (c) 1974 by CERN; used here under the U.S. fair use doctrine. **36** *Plane:* Redrawn from a photo by Wikimedia Commons user Andy Mitchell, CC-BY-SA. **36** *Rotating earth:* NASA image, public domain; animated gif by Wikimedia Commons user Marvel, CC-BY-SA. **45** *Duck-rabbit illusion:* J. Jastrow, Popular Science Monthly, 1899. **54** *Galaxy:* NASA/STScI/ESA, public domain. **56** *Cadillac:* Redrawn from a photo by That Hartford Guy, CC-BY-SA. **72** *Astronaut:* NASA. **73** *Fireworks:* Kristina Servant, CC-BY. **73** *Sprinters:* Darren Wilkinson, CC-BY-SA. **83** *PET scan:* NIH, public domain. **99** *Football player and old lady:* Hazel Abaya. **101** *Impact sites on Jupiter:* NASA, public domain. **101** *Comet breaking up:* NASA/ESA, public domain. **104** *M100:* European Southern Observatory, CC-BY-SA. **104** *Saturn:* NASA, public domain. **104** *Girl:* Georges Hébert, L'éducation Physique féminine, 1921, public domain. **105** *Airplane:* From a photo by Major James Skitt Matthews (Canada), 1930, public domain. **105** *Artificial horizon:* NASA, public domain. **106** *M100:* European Southern Observatory, CC-BY-SA. **106** *Ladybug:* Redrawn from a photo by Wikimedia Commons user Gilles San Martin, CC-BY-SA. **108** *Hawking aboard Vomit Comet:* Public domain product of NASA. **111** *Standing woman:* Redrawn from work by William Crochot, CC-BY-SA. **111** *Jumping man:* Redrawn by the author from photos by Eadweard Muybridge,ca. 1880. **113** *Pound and Rebka photo:* Harvard University. I presume this photo to be in the public domain, since it is unlikely to have had its copyright renewed. **119** *Standing woman:* Redrawn from work by William Crochot, CC-BY-SA. **120** *Ladybugs:* Redrawn from a photo by Wikimedia Commons user Gilles San Martin, CC-BY-SA. **120** *M100:* European Southern Observatory, CC-BY-SA. **122** *Galaxy and star:* Hubble Space Telescope. Hubble material is copyright-free and may be freely used as in the public domain without fee, on the condition that NASA and ESA is credited as the source of the material. The material was created for NASA by STScI under Contract NAS5-26555 and for ESA by the Hubble European Space Agency Information Centre. **122** *Saturn:* Public domain, NASA. **122** *Vesta (made spherical):* NASA, public domain. **122** *Human figures:* Line art by B. Crowell, CC-BY-SA licensed. Based on a photo by Richard Peter/Deutsche Fotothek, CC-BY-SA licensed.. **124** *Eclipse:* 1919, public domain. **132** *Hovercraft:* Copyright status unknown; used here under the U.S. fair use doctrine. **133** *Upsidaisium:* Copyright Jay Ward Productions; used here under the U.S. fair use doctrine. **136** *Crushed cars:* Wikimedia Commons user IFCAR, public domain. **138** *Sun:* NASA/SDO (AIA), public domain. **138** *Star field:* G. Bacon (NASA/ESA/STScI), public domain. **141** *Satellite:* From a public-domain artist's conception of a GPS satellite, product of NASA. **141** *Woman on earth:* Redrawn from work by William Crochot, CC-BY-SA. **141** *Astronauts:* Line art by B. Crowell, CC-BY-SA licensed. Based on a photo by Richard Peter/Deutsche Fotothek, CC-BY-SA licensed. **142** *Accretion:* S. Gezari, J. Guillochon, STScI/NASA, public domain. **142** *Lensing:* Wikimedia Commons user Alain r, CC-BY-SA. **146** *Rear-view mirror, main landscape:* Bradley Gordon, CC-BY. **146** *Rear-view mirror, landscape in mirror:* Benjamin D. Esham, CC-BY-SA. **148** *Landforms for Goode homolosine:* Wikimedia Commons user Strebe, CC-BY-SA. **148** *Mercator projection:* Wikimedia Commons user Eric Gaba, CC-BY-SA. **148** *Grid for Goode homolosine:* Wikimedia Commons user Strebe, CC-BY-SA. **160** *Einstein:* Oren Jack Turner, 1947, public domain in the U.S. **168** *Galaxies:* Ville Koistinen, CC-BY-SA. **172** *Hubble:* 1920, public domain. **172** *Leavitt:* public domain. **172** *Slipher:* 1905, public domain. **173** *Galaxy:* Ville Koistinen, CC-BY-SA. **175** *Galaxies:* Ville Koistinen, CC-BY-SA. **176** *Penzias-Wilson antenna:* NASA, public domain.

Index