The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or to view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

PROFESSOR: OK, in that case, let's start-- I want to begin by giving a quick review of where we were last time. And then we'll pick up from there. Today's lecture, the main subject will be the-- see if this works. The main subject will be the spacetime metric, which is what we'll begin by talking about. And later, I hope we'll be doing the geodesic equation.

Last time, we began by talking about open universes. And we got to open universes by way of closed universes. And we started with this Robertson-Walker form of the closed universe metric, which holds for k greater than 0 describing a closed universe. And then we said if we want to describe an open universe, we can use the same equation, but let k be less than 0.

And this metric has a name, the Robertson-Walker metric, which applies for k being positive, negative, or 0 as a special case. When k is 0, this just becomes the Euclidean metric written in polar coordinates. So it's a flat space for k equals 0.

Then we addressed the question of, why should we believe this as a proper description of an open universe? We know how to write it. And we could make it look a little better perhaps by introducing a kappa, which is minus k. So that kappa could be positive when k is negative.

To answer that question, we had to define what criteria we had in mind for the metric we were looking for. And what we're trying to do is write down a metric that will describe a homogeneous and isotropic universe. Because from the beginning of course, we said that those are the key properties that describe, to a good approximation, the universe that we live in.

So what we want to know is that this metric is homogeneous and isotropic when kappa is positive and k is negative, the new case. For the closed universe case, we already knew those things because in the closed universe case, it's obvious because we know we got it from a sphere. And a sphere is clearly homogeneous and isotropic from the start.

So looking at this metric for kappa positive, we see immediately that it's obviously isotropic, at least about the origin. Because if we just sit at the origin and looked at different angles-- theta and phi-- the theta phi dependence of this metric is simply given by that expression. And this is exactly the metric of the surface of a sphere whose radius happens to be a of t times little r. And we know that that sphere is isotropic.

It doesn't look manifestly isotropic because when you put theta phi coordinates on the surface of a sphere, you choose a north pole and measure everything from that. So your coordinates break the isotropy. But you know perfectly well that the sphere itself is completely isotropic.

So isotropy is settled about the origin. And if we're soon going to prove homogeneity, it's enough to know that it's isotropic about the origin because homogeneity will demonstrate that all points are equivalent. So if it's isotropic about the origin, we will ultimately know that's isotropic about all points.

So homogeneity is the hard thing. How do we convince ourselves that this metric is homogeneous? Now, if we look at it, it doesn't look homogeneous. It certainly looks like the origin is special. That was the case for the closed universe Robertson-Walker metric as well. So it's certainly not decisive.

But it also doesn't prove that it's homogeneous by itself. So we have to figure out how we would prove that it's homogeneous. And here, we only sketched an argument. We didn't really go through it in detail because it gets messy. But I think logic of the argument is clear. And it's, I think, rather persuasive that this argument would work if you wrote out all the equations.

So let me go through the argument again. We start by thinking about how we would demonstrate that the closed universe was homogeneous in a mathematical way using algebra rather than words. And if we wanted to prove that the closed universe was homogeneous using algebra, we would set out with the goal of trying to show that any point-- and we'll let r sub 0, theta sub 0, pi sub 0 denote an arbitrary point.

What we'd like to show is that any point is equivalent to the origin. And by equivalent to the origin, what we really mean is that we could define a new coordinate system where this arbitrary point would become the origin, and the metric would look just like it looked to start with. And then this new arbitrary point will be playing exactly the same role that the origin played in the first place. So we're looking for a transformation which will map r0, theta 0, and phi 0 to the origin while maintaining the form of the metric.

And we really know how to do that, because we know from the beginning that this universe is homogeneous because of the way we constructed it as the coordinization of the surface of a sphere. And the sphere is manifestly homogeneous. You can rotate any point on the sphere into any other point just by performing a rotation, which certainly does not change anything about the metric on the sphere. So we want to basically take advantage of that fact.

And we could imagine-- and we can even carry it out if we have to, but I only want to imagine it. I want to imagine constructing a map from r, theta, and phi to some new r prime, theta prime, phi prime. We want to map the entire space. But we want it to have the property that the special point-- r0, theta 0, and phi 0-- gets mapped to the origin.

So we want to construct this general mapping which has the property that our special point is mapped to the origin. And we can do that in three steps. And they're shown schematically here. We first simply transform from our r, theta, and phi coordinates to the four coordinates x, y, z, w that we, in fact, started with, the four coordinates that describe the euclidean four dimensional space in which this three dimensional sphere is embedded.

Once we have the four dimensional space, we can perform euclidean rotations in that four dimensional space. And we can perform any rotation we want. And we, in principle, know how to write those out in detail.

And we can choose the rotation, which maps r0, theta 0, and phi 0, keeping track of where it went, to the values of x, y. Z, w that will ultimately correspond to the origin of our coordinate system when we get back to r prime, theta prime, and phi prime. So we can arrange for that to happen by choosing the right rotation here.

And once we've rotated, we can then define r prime, theta prime, and phi prime in terms of x prime, y prime, z prime, and w prime in exactly the same way as we did in the first place when we didn't have primes. We just used the same formulas again. And that will ensure that the metric and r prime, theta prime, and phi prime will be just the metric that we've had, because it's determined from the euclidean metric in the four dimensional space in exactly the same way.

So this does it. And we could, in principle, do it all in detail. And we would get a concrete expression for r prime, theta prime, and phi prime in terms of r, theta, and phi that would have the property that we want of mapping the arbitrary point that we chose to become the origin of the new system.

So the point is that once you've written that all those equations, you know that they work for k positive. But in the end, you'd just have a set of equations that define r prime, theta prime, and phi prime in terms of r, theta, and phi. And those equations are just as valid for k negative as they are for k positive.

And the fact that the metric will be unchanged is also just as valid for k positive and k negative, because the metric is really just determined by derivatives of the new coordinates with respect to the old coordinates. And those are all just algebraic expressions. And if an algebraic expression is an equality for one sine of k, it will be an equality for the other sine of k.

So I think we have good faith-- although it'd be more convincing perhaps to actually write out all equations, but I think we have good faith that the same map will work for k less than 0. And by "work," it means that it will show that any point could be mapped to the origin by a metric-preserving transformation, which is the key thing, which is what we need to show that the space is actually homogeneous even though when we write in these coordinates, it doesn't look homogeneous.

So does that make sense to everybody? Are there any questions at this point?

The next thing we did-- ah, I'm sorry. I guess we-- point out that we're not going to actually show this explicitly because the algebra involved in the steps does get very complicated. I wanted to mention a few other facts about this Robertson-Walker metric. One important fact, which we will not show-- to show it would take approximately another lecture. It's not an incredibly deep mathematical fact, but it requires establishing some formalism to handle descriptions of curved spaces without yet knowing what the metric is going to be.

But in any case, it can be shown-- and we're going to store this in back of our heads-- that any three dimensional, homogeneous isotropic space can be described by this Robertson-Walker metric. Now, it's important to realize that that does not mean that the Roberson-Walker metric is the only way to write down a metric for a homogeneous and isotropic space. You could choose different coordinate systems that would make things look different.

But the point is that for any homogeneous and isotropic three dimensional space, it is always possible to assign coordinates who make the metric the Roberson-Walker metric, which means that if you understand the Robertson-Walker metric, you need to understand anything else. Any homogeneous and isotropic space can be described this way.

Next, we pointed out last time-- and did a short calculation to demonstrate for ourselves-- that for k greater than 0, the universe is finite. And that's clear from the beginning, because it was described as a surface of a sphere and the surface of a sphere is finite.

But for k less than or equal to 0, the variable little r of the Robertson-Walker metric can become arbitrarily large. That by itself does not imply that the space is necessarily infinite. But you could also calculate the distance from the origin as a function of little r. And that, you can show, becomes arbitrarily large. And that does mean that the space is literally infinite in size. So for the flat case or the open case, the Robertson-Walker metric describes an infinite universe.

And next, so we mentioned-- and this is a homework problem, or an optional homework problem on the current set-- that the Gauss-Bolyai-Lobachevsky geometry is actually simply the open universe in the two dimensional case rather than three dimensional case. So in the language of the Robertson-Walker metric, I think it's much easier to describe than in the coordinate system that Felix Klein invented. But it's the same space. And on the homework set, you can work out the mapping between the Klein coordinates and the Robertson-Walker coordinates to see that they're the same space.

Any questions? OK. Next, we changed subjects and started talking about the topic that we'll be continuing now because we did not finish this discussion-- the discussion of how to go from this space metric that we now understand to the spacetime metric, which is the fundamental quantity of general relativity and which we'll be using to describe our model universes.

Time ends up not playing an important role in what we'll be talking about. But nonetheless, it is an important part of the basic formalism of general relativity. And for some questions, it's crucial how time enters the metric. So we will discuss how time enters the metric.

So we want to generalize the metric from a spatial metric to a spacetime metric. And the first thing that means is that we want to understand the relativistically invariant interval between two points in spacetime. A point in spacetime is also called an event.

And we begin with special relativity because that's what all this is a generalization of. In special relativity, one can define the Lorentz invariant distance between two events. Here, the events are A and B. And the coordinates of those events are xA, yA, zA, and tA for event A. And as you would obviously guess, xB, yB, zB, and tB for event B. And the Lorentz invariant interval between those events is s squared sub AB.

And the first and most important thing about this interval is that it is Lorentz invariant. That is, you can compute it in any inertial frame, in any Lorentz frame, and it will have the same value even though the different pieces of it will have different values. The differences will cancel out.

And in the end, when you calculate the sum of these four terms to make s squared sub ab, you will find they'll have the same value in any Lorentz frame. And we're not going to show that fact. But we're going to use that fact, a very well known fact to anybody who studied special relativity in a reasonably way.

It's important to think a little bit about what the meaning of this peculiar quantity is that mixes space and time. And I think the easiest way to think about the meaning of it is to think about special cases. And it's a real number which could be positive, negative, or 0. And those are the special cases I want to think about-- positive, negative, or 0.

So if s squared is positive, it means that the separation between the two events is what's called spacelike. It's dominated by the spatial term. And if that's the case, it's always possible to find a frame where the two events are simultaneous. And in that frame, s squared is just the square of the distance between the two events, so has a clear interpretation. It's just the distance in the Lorentz frame in which they occur simultaneously.

Similarly, if s squared is negative, it means it's dominated by the negative time term. And in that case, the separation is called timelike, as you'd guess. And it has the property that it's always possible to find a Lorentz frame in which the two events occur at exactly the same point in space. And in that frame, s squared is equal, up to a factor of minus c squared, to the time separation between the two events. So s squared measures the time separation between the two events in the frame at which they happen in the same place. And s squared is actually minus c squared times the time separation squared.

And finally, if s squared is 0, that's called a lightlike separation. And it means that the two events are separated by just the right distance so that if a light pulse left one, it would just arrive at the other location at exactly the time of that event. So the two events could be joined by a light pulse that travels at the speed of light. And that's the significance of s squared being 0.

OK, any questions? OK-- actually, I don't want to that slide. We'll get back to that.

OK, that finishes the review of last lecture. Now, what we want to do is continue talking about spacetime intervals and how we fit that into the metric, which will ultimately describe distances in both space and time. So I'm going to work on the blackboard for awhile now.

So the formula that we're starting with-- and we'll put dot dot dot here and minus c squared tA minus tB squared. The dot dot dot means the y term and the z term which you can probably imagine are there without my writing them. From this expression, knowing that what we want to do is to take advantage of these geometrical ideas introduced by people like Gauss which described distances in terms of infinitesimal distances between infinitesimally close points, we can write the analogous equation for an infinitesimal distance.

And that becomes ds squared is equal to dx squared plus dy squared plus dz squared minus c squared dt squared. So this would be the Lorentz invariant separation between infinitesimally separated events. And this is what we're going to try to generalize to our curved space situation. So this would be the metric for special relativity. It's called the Minkowsky metric.

OK, now I want to move into the general relativity generalization of this idea. And general relativity makes use of the idea that Gauss originally suggested that distances should always be quadratic functions of the coordinate differentials. So we're going to keep that. Einstein kept that.

Now, in talking about coordinate differentials, we should emphasize here that in general relativity, unlike special relativity, coordinates are just arbitrary labels for points in spacetime. In special relativity, coordinates actually measure distances and times directly, which is why the metric is so simple. You don't really need the metric in special relativity. The coordinates themselves will tell you the distances and the times.

But in general relativity, that will not be the case. There's no way to do that for a curved space or a curved spacetime. So in general relativity, the coordinates are just arbitrary labels of points in spacetime. And to know anything about actual distances, you have to look at the metric. The coordinates themselves don't tell you the actual distances.

This immediately implies something about the kinds of coordinate transformations that you might want to think about. In special relativity, we have a privileged set of coordinates, namely the coordinates of Lorentz frames, of inertial frames. And the physics is simple when described in terms of those coordinates. In principle, you can use any coordinates you want, even in special relativity. But you never do, because the physics is so much simpler in the inertial coordinates that there's never any motivation for using any other coordinate systems.

But in general relativity, there is no privileged coordinate system. And it's very common to make all kinds of transformations of coordinates in the context of general relativity. And the formalism is set up so you could make any coordinate transformation you want, and it's just thought of as a relabeling of the points in space and time. And the formalism of general relativity works for an arbitrary labeling of points in spacetime. So in general relativity, any coordinate transformation is allowed.

But there's an important feature of these coordinate transformations-- is that when we make a coordinate transformation, we're always going to readjust our metric so that ds squared between any two nearby spacetime points has the same value in the new coordinate system that it had in the old. We will always change our metric to reflect our changes of coordinates. So ds squared must have the same value in any coordinate system. So the statement is that ds squared is coordinate-invariant.

OK, to define what we mean by ds squared, which if you notice, I haven't quite done yet. It's going to be, of course, the analog of what we've been talking about in special relativity. In special relativity, we did have this special class of observers, inertial observers, observers whose measurements of length and time really corresponded to the inertial frames and whose observations are related to each other by Lorentz transformations.

It's important to start out by asking is there any class of observers in general relativity which might play the same roles-- the observers that sort of define the measurements that you want to talk about.

And it's clearly a little bit more complicated in general relativity. The inertial observers of special relativity are characterized by the statement that there are no forces acting on them. So they just travel at a constant velocity. And you can always go to a frame where that velocity is 0 and you an talk about the rest frame of any inertial observer.

In general relativity, we need to distinguish to some extent between non-gravitational forces and gravitational forces. Non-gravitational forces, like say, electrical forces, are treated in general relativity in a way that's fundamentally similar to the way that such forces are treated in special relativity. But gravity is treated totally differently. Gravity is really just going to be described by the metric of spacetime-- by the distortion of spacetime.

And we already know, by way of simple examples, that if general relativity actually works to describe the universe that we've been talking-- which it'd better or we'd be in trouble-- we have a system where if we just look at the co-moving observers, each co-moving observer has no non-gravitational forces acting on him. He's just sitting still as far as he's concerned. But nonetheless, these co-moving observers are accelerating relative to each other as the universe expands and as that expansion changes its expansion rate, which we've already calculated.

So if there's going to be any observers that are going to play the role of inertial observers, it's presumably going to be a class that includes these co-moving observers. And the question of whether or not there are gravitational forces acting on the co-moving observers ends up depending on your point of view.

Each co-moving observer would think that there's no gravitational forces acting on him. He would just be standing still. But he would see all these other co-moving observers accelerating relative to him. So he would say that there are gravitational forces acting on these other observers.

So gravitational forces in general activity becomes coordinate-dependent ideas. And the Hubble expansion is one example of that where every co-moving observer would consider himself to be unaccelerating but would see all the other co-moving observers accelerating.

The other famous example, which is part of the original motivation of general relativity, is the famous Einstein elevator, which is also discussed in Ryden's textbook. If we have an elevator box, we could imagine letting the elevator fall. There was a rope there, but somebody cut it. And the elevator's now falling.

And we have a person in it. And the person in my-- the version of the story that I have in the lecture notes, the person's holding a bag of groceries. And if the elevator is falling freely and we ignore any air resistance or any other kind of friction so the elevator's falling at exactly the freefall rate, inside the elevator, everything will be falling with the same acceleration.

The person could lift his feet up off the floor, and he would just hover there. He would feel no gravity pushing him towards the floor. And similarly, he could let go of the bag of groceries, and they will just appear to float in front of him as long as he's undergoing this freefall. So the effects of gravity have been completely removed.

On the other hand, from outside the elevator, if we use the frame of reference of the Earth, we said that there very definitely is a force of gravity acting here. It's just acting the same on all the objects. And this gets elevated into the equivalence principle of general relativity.

And maybe before I annunciate that, I should consider the other case here. This is one example of how things work. A similar situation can involve the same elevator, but this time, let's have it just be sitting on the floor of the building that it's located in. In that case, the person inside would feel himself pushed against the floor by gravity. If he was holding his bag of groceries, he would notice he has to apply force to the bag of groceries to stop the groceries from falling to the floor.

He would say that he's being acted on by the force of gravity. That would be the natural description.

But we can consider an analogous case where we have the same elevator in empty space with a rocket ship up here that I didn't allow myself room to draw tied by cables to the elevator. And if the rocket ship accelerated with acceleration little g, the person inside the elevator would feel himself pressed against the floor in exactly the same way as you would here.

So again, we have a situation where there's gravity in one case and no gravity in the other case, but no difference in what the person inside would feel. And that is what becomes this principle of equivalence, which says that the physics of the accelerating frame of the elevator in that analogy-- so this is accelerating frame but with no gravity-- is equivalent to feeling the gravitational field of the Earth. In short, if you were living inside the elevator, you cannot tell which of those two pictures describe the world that you're actually part of.

And this is a very deep principle. It has very strong implications. It really does mean that everything you'd ever want to know about how gravity affects physical systems can be described by understanding how accelerations affect the physical systems. So it reduces the questions of what gravity does to just understanding what happens when you're in an accelerating coordinate system.

OK, this also opens the door for the question I began with-- is there a special class of observers here? And we can identify a special class of observers. But the special class is not observers which have no forces acting on them, which is what we would have said in special relativity. But rather, the special observers are the observers who have no non-gravitational forces acting on them, like the co-moving observers in our model of the universe.

But gravitational forces, you could never say if they're there or not because they're always there in some frames and not there in other frames. So we have no control or no way of making any frame-invariant statements about the force of gravity.

So what we'll be interested in as our primary observers and which we're going to use to define things in this class of observers with no non-gravitational forces-- and those will be called free-falling observers. And they're called free-falling because you have no way of knowing whether they're just observers for which there is no gravity, which would be an example of a free-falling observer by our definition.

But the situation is indistinguishable from this one, where free-falling has its obvious meaning-- that the guy there is falling relative to the Earth's frame. But he's freely falling. And therefore, he does not feel, relative to his environment, any forces whatever.

Now, I should emphasize that this equivalence principle holds only in small regions. In principle, it only holds in infinitesimal regions because there are, in gravitational systems, what we call tidal effects, where a tidal effect simply means that the gravitational field is never completely uniform. And if the gravitational field is not uniform, you do not completely cancel it by going into the accelerating frame of the elevator.

But in any infinitesimal region, you can always cancel the effects of gravity by going into a properly accelerating frame. And that's what the equivalence principle says.

OK, are there any questions about that? OK, this being said-- it was a long prelude-- we can now define what ds squared is supposed to represent. And the answer is simply that ds squared has the same meaning as in special relativity except that inertial observers are replaced by free-falling observers.

OK, so let's review what exactly that means. It means that if ds squared is positive, it means that there will always be a class of free-falling observers for whom those two events will occur at the same time. And ds squared will be the distance between those two events as measured by those inertial observers-- bah, I said "inertial"-- free-falling observers.

And similarly, if ds squared is negative, it means there will be a class of free-falling observers for whom those two events will occur at the same location. And ds squared will measure, up to a factor of minus c squared, the time separation squared between those two events. And it will again be the case that if ds squared is 0, it will mean that the two events are separated by just the right distance so that a light pulse can travel from one to the other.

OK, any questions about that? It's was kind of a long-winded discussion. But I think it does pay to actually understand what df ds means rather than just to write down a formula for it and say that's like special relativity.

OK, having said all this, our next goal is to figure out how time enters the Robertson-Walker metric to give us a spacetime metric instead of just a spatial metric, which we already have written down. And I'm going to write down the answer and then describe why that has to be the right answer. I think it's the easiest way to handle it here.

So the right answer is that when we incorporate time and think of this as a metric for spacetime, ds squared is going to be minus c squared dt squared plus a squared of t times dr squared over 1 minus k r squared-- right now, it's just the same spatial part that we had before-- plus r squared d theta squared plus sine squared theta d phi squared. End parentheses. End curly brackets. So all I've done is I've added a minus c squared dt squared term to the metric.

Now, why is this the right metric? I'm going to first consider two special cases, which will verify some of the terms there. And then I want to also discuss why there aren't any other terms besides the ones that we know have to be there.

So first, let's just consider the case-- case one will just be dt equals 0. If there's no time separation between the two events, then we're only interested in spatial separations. And we've already talked about how to describe spatial separations in a way which makes the description homogeneous and isotropic. And we said that this is the most general way of describing spatial separations that are in a space which is homogeneous and isotropic.

So from what we said previously, this has to be the answer. That's how we describe homogeneous isotropic spaces. So when dt vanishes, it just reduces to the case we've already discussed. Simple enough.

Case two, which involves a little bit of new thinking-- suppose dr equals d theta equals d phi equals 0 so that only time changes. OK, this describes the situation about our co-moving observers. They're sitting at fixed spatial coordinates and evolving in time. And we've already said that the thing that we call cosmic time is simply time as measured by the wrist watches of our co-moving observers.

So t, if you want t to be cosmic time, which we do-- we're trying to describe a metric for our spacetime as we've already described it. We're now just trying to write a metric for it. So t should be cosmic time. So t should be the time as measured on the wrist watches of the co-moving observers.

And that's exactly what this metric says. It says that if ds squared defines the measurements of our free-falling observers, which is what we said is the definition of ds squared, that it is just equal to minus c squared times the change in the coordinate time. And coordinate time means cosmic time because that's the coordinate system we're using. So putting in the minus c squared dt squared term is the only way that it can be so that the wrist watches of our co-moving, free-falling observers measure the same thing that the coordinate measures, which is what we defined cosmic time to be in the first place. So I think that justifies this term.

And notice, if you had any coefficient here other than c squared, there'd be a multiplicity offset between what the wrist watches of your observers are measuring and what cosmic time is ticking off. And we're not allowing that, because we define cosmic time to be the time measured by the wrist watches.

OK, so that takes care of these two cases. And I think it implies that these terms have to be here in exactly the form that we've written. And we could stop now and pretend that we've solved the whole problem. But I always like to be what I consider to be thorough. So I like to sort of imagine the questions that could pop up if people were inquisitive.

So you might imagine that you have some difficult roommate who says, why can't I put some other term, and what else could there be? The only thing that we've left out here are terms that involve products of dt with either dr, d theta, or d phi. So what about terms like the product of dr times dt or d theta times dt or d phi times dt? Question mark.

So one possible answer is I looked in a book and it wasn't there. But that's not the best of all possible answers. It's good to understand why things are in books and why other things or not.

So you might want to construct an argument of why these terms have to be absent. And the reason why those terms have to be absent is because if they were there, they would violate isotropy. Roughly speaking, the notion is that if you have a dt times some d spatial coordinate, that singles out a certain direction in spatial coordinates space because dr is not the same as minus dr. Dr points in a certain direction.

To be more explicit about that, in the notes, I discuss a thought experiment which basically gives a concrete realization of the asymmetry that I just discussed. So to see how those terms explicitly violate isotropy, we can imagine a thought experiment where we start by thinking about some particular point in space. And we'll give it coordinates r, theta, and phi. And I'll assume r is non-zero.

We could then imagine that two people sitting at this point-- and in the Lewis Carroll spirit, I call them Tweedledee and Tweedledum-- can decide to do an experiment by first synchronizing their clocks. And they might as well synchronize t cosmic time, let's say. And then, one of them can go off in the direction of positive r, and the other can go off in the direction of negative r at the same coordinate velocity, which

I'll call v.

And by coordinate velocity, I mean dr dt, because that's the simplest thing to talk about here. It many not be the same as the physical velocity, but we don't care. It'll be the same for both of them.

And the experiment will be that they will each travel until there's some-- until cosmic time-- they're passing a lot of cosmic time clocks as they travel. And they agree to travel until cosmic time ticks off until some chosen final time. And when they each finish the experiment by noticing that the cosmic time clocks now read t sub f, they will look at their own watches and see how much time elapsed. So they're basically measuring time dilation-- how do their wrist watch times, when they move, differ from cosmic time.

And the point is that if we have a dr dt term in the metric, these people get different values because the time that they measure will be what they call ds squared, up to a factor of minus c squared. And that will include this dt dr term. And dr for one of them will be the coordinate velocity they chose times d cosmic time, the amount of cosmic time interval they travel for. They've agreed on that before they take off. And for the other, dr will be negative vc times dt, so cosmic.

So this term will give a different contribution to the ds squared that each of these two entities, Tweedledee and Tweedledum, will measure. And therefore, they'll be measuring different ds squareds, and that means they'll be measuring different things on their wrist watches. And that means they have an asymmetry in directions. By going one direction or the other, they could determine whether their time dilation will be increased or decreased.

And that, if our universe is isotopic, should not be possible. And therefore, if we want to write a metric which describes an isotropic universe, we have to omit the dr dt term. And a completely identical argument implies that we have to also omit d theta dt and d phi dt.

So isotropy implies dt dr term is not allowed. Because otherwise, we would have a

Tweedledee Tweedledum time dilation asymmetry, which we're not allowed to have an isotropic universe. OK, any questions about that?

OK, if you have no questions about that, now we're ready to go onto our next topic. Now that we've described the metric of our universe, there it is-- the full Robertson-Walker spacetime metric for a homogeneous and isotropic universe.

The next thing I'd like to about is how do we calculate motion in a metric in the context of general relativity. And our treatment here will be completely general. We'll learn how to calculate motion in an arbitrary metric. And we'll in fact use the Schwarzschild metric, which describes spherically symmetric objects like stars or even black holes as an example. But our real purpose is to understand things like motion in the universe.

And I guess at this point, I am going to-- load up screen. OK, good.

OK, I'm going to just use the equations from the lecture notes. This particular calculation involves a lot of long equations, so I think doing it on the blackboard would probably be a bit too tedious. So I'm instead going to just lift the equations and talk about them directly from the lecture notes themselves.

So what we're interested in is thinking about a geodesic in some arbitrary metric. And we're going to start with the simplest possible example of the two dimensional spatial metric of the same kind of spaces that Gauss and Bolyai and Lobachevsky were talking about using this notation of differential geometry of thinking of a metric in terms of coordinates, which in this case, we'll initially call x and y. It will generalize perfectly straightforwardly to spacetimes because all the ideas are the same. But it's easiest to start out by thinking about you're simply talking about measuring distances in a two dimensional space.

A geodesic is defined as a line between two points in space which has the property that the length of that line is stationary with respect to any variations. Stationary means the first derivative vanishes. Now, in this two dimensional space example, our stationary lines will also always be minima. That is, you can minimize the distance between two points by finding the shortest possible line.

There is no longest possible line. And there aren't any saddle points either, I don't think. So I think in this case the minima-- the stationary points will always be minimum I believe.

But in general, when we have spacetime metrics, especially when things are not even positive definite, these geodesics, the stationary lines, can be either maxima or minima or saddle points. So you should imagine that all those possibilities are there. However, the equations we'll derive will really just be the equations that say that the first order difference vanishes.

If you vary the path a little bit to first order, the length does not change. And that will be true for maxima, minima, or saddle points. We won't have to care in deriving the equations. So we start by imagining a metric like that.

And the first thing we want to do is just adopt a better notation for the metric. And there are two improvements. The first is to number the coordinates instead of thinking of them as different letters. So instead of talking about x and y, we're going to talk about x super one and x super 2.

Now, these 1's and 2's have the danger of possibly being confused with a power. We probably never write x to the first power, but you might write x super 2 and think of if as x squared. Many times we, of course, do that. So one always has to hope that the context will make it clear what that index refers to. But here, these upper index objects-- those superscripts are just indices. They're not powers.

You might wonder why we tolerate such a crazy notation when we could have written them as subscripts. And then there would not be this confusion. But the answer is that in general relativity, one does make use of both subscripts and superscripts in a slightly different way. And to some extent, you'll see that in what we'll be doing.

So it's useful in general relativity to have two kinds of scripts. And the only places that seem to exist are up and down. So they're superscripts and subscripts. And one simply hopes that there's no confusion with powers.

OK, so step one is number your indices and to number your coordinates. And then instead of writing that the sum of three terms that we had-- and of course, it gets to be much more. If you have four coordinates, it would be 16 terms or 10 terms, depending on how you collected them. But instead of writing that mess, you can write it using the summation notation. Sum from i equals 1 to 2, sum j equals 1 to 2 of g sub ij of the x-coordinates times dx i dx j.

And when you sum over i and j, you're summing over 1 and 2, which means you're summing over x and y. And the sum includes the x, x term, which is now called the 1, 1 term. And the y, y term which is now called the 2, 2 term, and the x, y term, which is now called the 1, 2 or the 2, 1 term. And those are identical to each other. And they just get added.

So that shortens the notation considerably. But then there's one further simplification that was actually introduced by Einstein himself. And it's always called the Einstein summation convention. Notice that in this equation, the letter i appears twice as an index-- as an upper index there and as a lower index on the metric, g sub ij. And the Einstein convention is that whenever you have a repeated index where one is upper and one is lower, you automatically sum over them without writing the summation sign. The summation sign is implied.

So then this equation get simplified to that equation, which is the form that we'll actually be using. And that's as about as simple as it gets.

OK, so far, that's just notation. OK, now what we want to do is to talk about a path between two points. And we want to discuss how we're going to describe the path and how we're going to derive the equations that will tell us that this path has the minimum possible length, which is what we're trying to do. We're trying to find the equations that tell us when a path has an extreme value of the length.

So to describe the path itself, we're going to imagine parameterizing it, which means we're going to think of a function xi of lambda, where xi-- remember, xi means 1 and

2. It means specifying one both x and y as a function of lambda. So this is really two functions of lambda. And this function of lambda is going to show us how the point that traces out this path varies from point A to point B. And we'll allow it to vary as a function of this parameter that we've introduced, lambda. And we'll adopt the convention that lambda is 0 at one end and lambda sub f at the other end.

So xi of 0 will be required to be the coordinates of point A If we want to start at some specified point A, we want to end up at some specified point B. So we'll insist that the coordinates evaluated at lambda sub f are xi sub B. And as long as xi of lambda is a continuous function, which we will also insist on, then xi of lambda will describe a path from A to B, which is what we're trying to do.

OK, to apply the metric and write down an expression for the length of this path, that's what we want to do next. And then we want to figure how to extremize that expression for the length. First step, we need to get an expression for the length. So the metric is written in terms of infinitesimal separations. So we want to imagine dividing this path up into little segments, each corresponding to some d lambda. Each little segment goes from some lambda to some lambda plus d lambda, where d lambda is infinitesimal.

And the change in the coordinates over that interval are then just the derivative of xi with respect to lambda-- remember, we have this function, xi of lambda-- times d lambda. This will give us the differential coordinates between any two neighboring points along the line.

Then ds squared is defined in terms of d xi. And we just plug this formula into the expression for ds squared in terms of the infinitesimal separations. So we have the metric. And then where we had previously just d xi, now we have d xi d lambda times d lambda. And similarly, where we previously had just d xj, now we have the derivative of xj with respect to lambda, again, times d lambda. So we have two powers of d lambda appearing in this expression.

ds itself will be the square root of ds squared. In this case, we are talking about positive, definite distances. So we can take the square root. So we put a square root sign over it. And now, we have only one power of d lambda. And this describes the length of the segment that goes from lambda to lambda plus d lambda, the length as defined by the metric.

The full length of the line is obtained just by integrating that from 0 to the final value of lambda. So equation 540 here is what we're looking for-- the expression for the length of the line in terms of the parametrization that we've chosen. OK, any questions about that formula?

OK, next step-- and here's where things get kind of complicated with the algebra, although I think the ideas are still pretty simple. The next step is to figure out how we determine when that expression is at its minimum value. How do we determine when the path has the right properties that we found the minimum length?

So to do that, we want to imagine varying the path. We want to consider comparing the length of the path that we're thinking about to the length of an arbitrary nearby path. And to do that, we can introduce a little bit of extra notation here. Here's the point xA Here's the point xB. x of lambda is the path that we're thinking about. And we're asking the question, is this the path of minimum possible length?

And to do that, we're going to compare it with an arbitrary nearby path. So the arbitrary nearby path is what's called x tilde in this diagram. It starts at the same point xA and ends at the same point xB. But along the way, it deviates by an infinitesimal amount from the original path. And we're going to parametrize that by equation 541a here. The tilde path will be equal to the original path plus a parameter that I'm going to introduce called alpha times a function wi of lambda, where wi of lambda is really an arbitrary function.

And I've introduced this extra parameter alpha just so I could say in a simple way what it means for these paths to be infinitesimally close, which just means that if alpha has an infinitesimal value, the two paths are infinitesimally close. And the function wi, we'll think of being a perfectly finite function with values like 2 and 5, not values that are infinitesimally small.

OK, with this parametrization of our paths, we want to impose one important criteria, which is that the two paths are supposed to start and end at the same points, A and B. And that means that this w super i that describes the derivation between two paths have an advantage at those two end points. Or else, the paths aren't going from the same starting point to the same ending point. And, certainly if you move the endpoints, you can always find a shorter path. There's no geodesic if you allow yourself to move the endpoints.

So we insist therefore that wi of 0, which is wi at the first endpoint, is 0. And similarly, wi of lambda sub f at the other endpoint is also 0. That's what we insist on.

OK, now having set up this formalism, we can now write down a very simple equation that says this path is an extremum. The path is an extremum if ds d alpha is equal to 0 for all choices of wi of lambda. OK, if it's an extremum, it means any small variation, any small variations are proportional to alpha. Any small variation produces 0 derivative. So ds d alpha should equal 0. And that should be the case for any possible deviation if we really have found the minimum possible length. OK? OK with everybody?

OK, now, it's mostly just a lot of gore to get the answer. The key step will be a crucial integration by parts that you'll see in a minute. But let's just go through the algebra together. I'm going to define an auxiliary quantity a of lambda alpha, which is just the metric times the derivatives of the functions. The path length of the deviated path-- these are tilde functions here. So a is the integrand for the length of the perturbed path, the tilde path. So s of x tilde is just the integral of the square root of a, d lambda.

OK, now we need some pieces to carry out our derivative. So I've introduced a few auxiliary calculations here that we can then put into the big calculation. We're going to need the derivative of the metric with respect to alpha. Now, the metric does not depend directly on alpha.

But the metric does depend on x tilde. It's evaluated at the point x tilde for any given value of lambda. And x tilde depends on alpha, because remember, x tilde was equal to the original path plus alpha times this wi, the derivation.

So it's a chain rule problem to figure out what the derivative of gij is with respect to alpha. So it's the derivative of gij with respect to xk times the derivative of xk tilde with respect to alpha-- just straightforward chain rule. And the derivative of x tilde with respect to alpha is just this function, w super i, or in this case, w super k. That's what defines the deviations. So this is our result, then, for the derivative of gij with respect to alpha.

Then, we apply that to differentiating s itself, finding all the alpha's inside that square root. And scroll up a little bit so we can see the definition of a. a consists of gij times the dx's themselves. And the dx's themselves depend on the alphas. So we're going to get terms coming from differentiating those with respect to alpha. And we get a term coming from differentiating the gij with respect to alpha.

So the whole quantity in the integrand here has a square root operating on it. So the derivative of the square root of a quantity is 1 over the square root of the same quantity times the derivative of the quantity, just differentiating the 1/2 power of a. So that gives us a 1/2 and 1 over the square root of a. And then inside here, we have the derivative of a with respect to alpha.

And one of those terms, we've already calculated. It's this multiplied by dx i d lambda dx j d lambda, which come along for the ride. And they lose their tilde because we're trying to calculate the derivative at alpha equals 0. So once we differentiate one factor with respect to alpha, we evaluate the other factors at alpha equals 0. So that's what we've done. We've evaluated the other factors at alpha equals 0.

And then, when we differentiate dx i tilde with respect to d lambda, we just get DWI d wi with respect to lambda. And dx j d lambda comes along, now evaluated at alpha equals 0. And similarly, the second term is where we differentiated the second factor here with respect to alpha. And we differentiate with respect to alpha, we bring down the w. So this becomes dx i d lambda dx j d lambda. So this is the expression.

And now we want to simplify it a little bit and figure out how to write down an equation which tells us when it's actually 0. So first, we want to simplify it a little bit. And I guess I want to-- do they fit? Almost. What I want to argue is that these last two terms are really equal to each other up to just rearranging the indices.

Remember, i and j are just being summed over. So we could have called them any letter we wanted, and they would still just be summed 1 and 2. And in particular, we can interchange what we call i with what we call j. And then, these two terms would become identical. And we're allowed to do that because these are just what are called dummy indices. They're just names of indices that are being summed over. And you get the same sum no matter what you call the index you're summing over.

So those terms can be combined, giving us just 2 times either one of those two terms we can keep. And now, we only two terms in our expression, which is not bad. But things are still a little complicated. And what makes them complicated at this point, which is what we have to get rid of, is the fact that w occurs as a multiplicity factor in the first term. But w is differentiated with respect to lambda in the second term. And when it's written that way, there's no direct way you could see what properties w has to have or the other terms have to have so that the expression vanishes.

But the crucial trick for handling that particular issue-- and it's the only real issue in this problem. The rest is just straightforward or sometimes tedious manipulations. The key step is to integrate by parts to turn the derivative of w expression into an expression that is just multiplicative in w. And the miraculous thing is that for the situation we have, there are no boundary terms that arise from that integration by parts.

So here's integration by parts spelled out in gory detail. We're going to use the famous formula that says that the integral of udv is equal to minus the interval at vdu plus the product U times V evaluated at the two endpoints and subtracted. So this is just the standard formula that defines integration by parts. The U is the term that starts out not having derivatives and later acquires derivatives. So that's the 1 over root a gij dxj d lambda of this-- this is the quantity we're trying to calculate.

And the du will just be-- I'm sorry-- dv will just be the factor which is a differential in the original expression, d wi d lambda times d lambda we're going to let be equal to dv. So this u and this dv give us the original integral. This integral is the same as the integral we're trying to evaluate.

Now when we integrate by parts, we apply a derivative to U to write down dU, which means it's just the derivative with respect to lambda of the quantity in brackets times d lambda. That's the du. And dv is just easily integrable. D is equal to wi.

Now, the important thing is to look at the boundary term. Because if the boundary term went on 0, we might not have accomplished anything. But the boundary term is 0 because the boundary term is the product of U times V, and V is wi. And wi vanishes at both boundaries that remember, was just the condition that the path goes between A and B. When you vary the path, you don't vary the points A and B. You only vary the path in between.

So wi vanishes at the endpoint. So that means that v vanishes at the endpoints. So that means if the product of U times V vanishes at the end points. And that means our boundary term, our service term, does not contribute.

So we turned the original integral into another integral where now wi appears as the multiplicity factor and it's no longer differentiated. Lots of other things get differentiated in the process. But wi gets to sit by itself. And that now makes it easy to combine these two terms and see under what circumstances the sum vanishes.

So the integral, after we make this integration by parts on one of the two terms, becomes this expression where now, the w's are always multiplicative. And by rearranging the names of these dummy indices-- as we initially have it, w has a superscript k in the first term and the subscript i and the second term. But one could rearrange these dummy indices-- we could name them anything you want-- so that in both cases, w has the same index. And then you can factor it out. And then you get this marvelous equation, which is now very close to being an equation that we're prepared to deal with.

The question we want to address now is under what circumstances does this vanish for every wi? Now, if we only know that it vanished for some particular wi, then we would not be able to say very much. Because it's very easy for an integral to be nonzero all over the place, literally everywhere except maybe at some isolated points and still integrate to 0. It could be positive in some places, negative in other places and zero only at crossing points and still integrate to 0.

But if this is going to vanish for any wi, then the claim is that the quantity in curly brackets has to vanish identically. And I think the best way to prove that is to say that if the quantity in brackets did not vanish identically, then you could let wi be equal to the quantity in brackets. Remember, if this is non-zero for any wi, we have a contradiction because we're going to require this to vanish for any wi.

So if the quantity in brackets were nonzero, we could let wi be the same quantity. And then we would just have the integral of a perfect square. And then clearly, the integral would not vanish. So that shows that if the quantity in brackets does not vanish, the integral does not vanish, at least for some wi. And that means that if the integral's going to vanish for every wi, which is what we're trying to impose, the quantity in curly brackets has to vanish identically. And that's our conclusion.

So that implies-- and I guess here is where we're going to stop. But we get to the famous boxed equation. And this really is-- we'll simplify it a little bit afterwards next time. But this really is the result. The geodesic equation is that equation, which is just the equation that the quantity that we had in curly brackets vanishes. So if the path that we've chosen has the property that these derivatives are equal to each other-- and notice it depends on the metric and it depends on the path. Because you have dx d lambda appearing everywhere. And x of lambda is the path.

But if this equation holds, then that path is a stationary point. It's an if and only if statement as long as paths are continuous. That is the geodesic equation. It tells us whether or not our path is the minimum. And next time, we will simplify it a bit. And we'll look at examples and understand how the formula works. So I'll see you folks again next Tuesday.