PATRICK WINSTON: You know, it's unfortunate that politics has become so serious.

Back when you were little it was a lot more fun.

You could make fun of politicians.

Here's a politician some of you may recognize.

But it's convenient to be able to vary what this particular politician looks like.

For example, we can go from a cookie baker to radical.

[LAUGHTER] PATRICK WINSTON: We can go from superwoman to bimbo.

[LAUGHTER] PATRICK WINSTON: Socialite-- I put socialite into this.

There she is.

Or we can move the slider over the other way to bag lady.

Alert, asleep, sad, happy.

How does that work?

I don't know.

But I bet by the end of this hour you'll know how that works.

And not only that, you'll understand something about what it takes to recognize faces.

It turns out to some theories of face recognition are based on the same principles that this program is based on.

But you can kind of guess what's happening here.

There are many stored images and when I move those sliders it's interpolating amongst them.

So that's how that works.

But the main subject of today is this matter of recognizing objects.

Faces could be the objects, but they don't have to be.

This could be an object that you might want to recognize.

And I want to talk to you a little bit about the history of this problem and where it stands today.

It's still not solved.

But it's an interesting exercise to see how the attempts at solution have evolved slowly over the past 30 years.

So slowly, in fact, that I think if someone told me how long it would take to get to where we are 30 years ago I think I would have hung myself.

But things do move slowly.

And it's important to see how slowly they move.

Because they will continue to move slowly in the future.

And you have to understand that that's the way things work sometimes.

So to start this all off, we have to go back to the ideas of the legendary David Marr, who dropped dead from leukemia in about 1980.

I say, the gospel according to Marr, because he was such a powerful and central figure that almost anything he said was believed by a large collection of devotees.

But Marr articulated a set of ideas about how computer vision would work that started off by suggesting that with the input from the camera, you look for edges.

And you find edge fragments.

And normally they wouldn't be even as well-drawn as I've done them now.

Or as badly drawn as I've done them now.

But the first step, then, in visual recognition would be to form this edge-based description of what's out there in the world.

And Marr called that the primal sketch.

And from the primal sketch, the next step was to decorate the primal sketch with some vectors, some surface normals, showing where the faces on the object were oriented.

He called that the two and a half D sketch.

Now why is it two and a half D?

Well, it's sort of 2D in the sense that it's still camera-centric in its way of presenting information.

But at same time, it attempts to say something about the three-dimensional arrangement of the faces.

So the speculation was that you couldn't get to where you wanted to go in one step.

So you needed several steps to get from the image to something you could recognize.

And the third step was to convert the two and a half D sketch into generalized cylinders.

And the idea is this.

If you have a regular cylinder, you can think of it as a circular area moving along an axis like so.

So that's the description of a cylinder.

A circular area moving along an axis.

You can get a different kind of cylinder if you go along the same axis but you allow the size of the circle to change as you go.

So for example, if you were to describe a wine bottle.

It would be a function of distance along the axis that would shrink the circle appropriately to match the dimensions of a wine bottle.

A fine burgundy, I perceive.

In any case, this one once converted into a generalized cylinder, when matched against a library of such descriptions, results in recognition.

Great theory, based on the idea that you start off by looking at edges and you end up, in several steps of transformation, producing something that you could look up in a library of descriptions.

Great idea.

Trouble is, no one could make it work.

It was too hard to do this.

It was too hard to do that.

And the generalized cylinders produced, if any, were too coarse.

You couldn't tell the difference between a Ford and a Chevrolet or between a Volkswagen and a Cadillac.

Because they were just too coarse.

So although it was a great idea based on the idea that you have to do recognition in several transformations of representational apparatus, it just didn't work.

So much later, maybe 15 years later or so, we get to the next part of our story.

Which is the alignment theories, most notably the one produced by Shimon Ullman, one of Marr's students.

So the alignment theory of recognition is based on a very strange and exotic idea.

It doesn't seem strange and exotic to mechanical engineers for a while, because they're used to mechanical drawings.

But here's the strange and miraculous idea.

Imagine this object.

You take three pictures of it.

You can reconstruct any view of that object.

Now, I have to be a little bit careful about how I say that.

First of all, some of the vertexes are not visible in the views that you have.

So, of course, you can't do anything with those.

So let's say that we have a transparent object where you can see all the vertexes.

If you have three pictures of that, you can reconstruct any view of that object.

Now I have to be a little careful about how I say that, because it's not true.

What's true is, you can produce any view of that in orthographic projection.

So if you're close enough to the object that you get perspective, it doesn't work.

But for the most part, you can neglect perspective after you get about two and a half times as far away as the object is big.

And you can presume that you've got orthographic projection.

So that's a strange and exotic idea.

But how can you make a recognition theory out of that?

So let me show you.

Well, here's one drawing of the object, I need two more.

Let's see.

Let's have this one.

And maybe one that's tilted up a little bit.

It's important that these pictures not be just rotations on one axis.

Because they wouldn't form what you might think of as a kind of basis set.

So there are three pictures.

We'll call them a, b, and c.

And then we want a fourth picture.

Which will look like this.

It doesn't have to be too precise.

And we'll call that the unknown.

And what we really want to know is if the unknown is the same object that these three pictures were made from.

So let me begin with an assertion.

I'll need four colors of chalk to make this assertion.

What I want to do is I want to pick a particular place on the object, like this one.

And maybe the same place on this object over here.

Those are corresponding places, right?

So I can now write an equation that the x-coordinate of that unknown object is equal to, oh, I don't know, alpha x sub a plus beta x sub b plus gamma x sub c plus some constant, tau.

Well, of course, that's obviously true.

Because I'm letting you take those alpha, beta, gamma, and tau and make them anything you want.

So although that's conspicuously obviously true, it's not interesting.

So let me take another point.

And of course, I can write the same equation down for this purple point.

And now that I'm on a roll and having a great deal of fun with this, I can take this point and make a blue equation.

And you know I'm destined to do it, so I've got one more color.

I might as well use it.

Let's just make sure I get something that works here.

That's this one, that's this one.

I hope I've got these correspondences right.

STUDENT: [INAUDIBLE].

PATRICK WINSTON: Have I got one off?

STUDENT: [INAUDIBLE].

PATRICK WINSTON: Which color?

STUDENT: Blue.

[INAUDIBLE].

PATRICK WINSTON: OK.

So this one goes with this one, goes with this one.

Is that one wrong?

STUDENTS: Yeah.

PATRICK WINSTON: Oh, oh, oh.

Of course this one, excuse me, goes down here.

Right?

And then this one is off as well.

I wouldn't get a very good recognition scheme if I can't get those correspondences right.

Which is one of the lessons of today.

OK.

Now I've got them right.

And now that equation is correct.

I think I've got this one right already.

So now I can just write that down.

I'm on a roll, I'm just copying this.

So those are a bunch of equations.

And now the astonishing part is that I can choose alpha, beta, gamma, and tau to be all the same.

That is, there's one set of alpha, beta, gamma, and tau that works for everything, for all four points.

So you look at that puzzled.

And that's OK to be puzzled.

Because I certainly haven't proved it.

I'm asserting it.

But right away, there's something interesting about this and that is that the relationship between the points on the unknown object and the points in this stored library of images are related linearly.

That's true because it's orthographic projection.

Linearly related.

So I can generate the points in some fourth object from the points in three sample objects with linear operations.

Christopher?

STUDENT: Is that the x-coordinate of-- PATRICK WINSTON: It's the x-coordinate.

Christopher asked about the x-coordinates.

Each of these x-coordinates are meant to be color coded.

It gets a little complicated with notation and stuff.

So that's the reason I'm color coding the coordinates.

So the orange x sub u is the x-coordinate of that particular point.

STUDENT: In 3D space?

PATRICK WINSTON: No.

Not in 3D space.

In the image.

STUDENT: So it's a 2D projection of it?

PATRICK WINSTON: It's a 2D projection of it, an orthographic projection.

OK?

So we're looking at drawings.

And those coordinates over there are the two-dimensional coordinates in the drawing.

Just as if it were on your retina.

STUDENT: [INAUDIBLE] vertexes on the 3D projection or can curved surfaces also?

PATRICK WINSTON: So he asked about curved surfaces.

And the answer is that you have to find corresponding points on the object.

So if you have a totally curved surface and you can't identify any corresponding points, you lose.

But if you consider our faces, there are some obvious points, even though our face are not by any means flat like these objects.

We have the tip of our nose and the center of our eyeballs and so on.

So if that's true, what does that mean about recovering alpha, beta, gamma, and tau?

Can we find them?

[INAUDIBLE], what do you think?

How do we go about finding them?

You're nodding your head in the right direction.

[LAUGHTER] STUDENT: It's four equations and-- PATRICK WINSTON: Splendid.

It's four equations and four unknowns.

Four linear equations and four unknowns.

So obviously, you can solve for alpha, beta, gamma, and tau if you know that these equations are correct.

So how does that help us with recognition?

It helps us with recognition because we can take another point, let me say this square point here and this corresponding square point here and this corresponding square point here, and what can we do with those three points now?

We've got alpha, beta, gamma, and tau, so we can predict where it's going to be in the fourth image.

So we can predict that that square point is going to be right there.

And if it isn't, we're highly suspicious about whether this object is the kind of object we think it is.

So you look at me in disbelief.

You'd like me to demonstrate this, I imagine.

STUDENT: Yeah.

PATRICK WINSTON: OK.

Let me see if I can demonstrate this.

So I'm going to do this in a slightly simplified version.

I'm only going to allow rotation around the vertical axis.

And just so you know I'm not cheating, there's a little slider here that rotates that third object.

Let's see, why are there just two known objects and one unknown?

Well that's because I've restricted the motion to rotation around the vertical axis and some translation.

So now that I've spun that around a little bit, let me pick some corresponding points.

Oops.

What's happened?

Wow.

Let me run that by again.

OK.

So there's one point I've selected from the model objects.

The corresponding point over here on the unknown is right there.

I'm going to be a little off.

But that's OK.

So let me just pick that one and then that corresponds to this one.

Krishna, would you like to specify a point so people know I'm not cheating.

Pick a point.

Pick a point, Krishna.

STUDENT: Oh, the right?

PATRICK WINSTON: The right?

STUDENT: Yeah.

PATRICK WINSTON: This one?

STUDENT: Yep.

PATRICK WINSTON: Oops.

OK, let's pick it out on the model first.

Now pick it over here.

Boom.

So all the points are where they're supposed to be.

Isn't that cool?

Well, let's suppose that the unknown is something else.

This is a carefully selected object.

Because the points are all the correct positions vertically, but they're not necessarily the correct positions in the other two dimensions.

So let me pick this point, and this point, and this point, and this point.

And Krishna had me pick this point.

So let me pick this point.

So if it thinks that the unknown is one of these obelisk objects, then we would expect to see all of the

corresponding points correctly identified.

But boom.

All the points are off.

So it seems to work in this particular example.

I find the alpha and beta using two images.

And I predict the locations of the other points.

And I determine whether those positions are correct.

And if they are correct, then I have a pretty good idea that I have in fact identified the object on the right as either an obelisk or an organ, depending on which of the model choices and the unknown choices I've selected.

So the only thing I have left to do is to demonstrate that what I said about this is true.

So I'm going to actually demonstrate that what I said about this is true using the configuration in this demonstration.

Because it's much too hard for me to remember matrix transformations for generalized rotation in three dimensions.

So here's how it's going to work.

The z-axis is going up that way.

Or, it's going to be pointing toward you.

And what I'm going to do is I'm going to rotate around this axis.

And what I want to do is I want to find out how the x-coordinate in the image of the points move as I do that rotation.

So here's the x-axis.

This is the coordinate that you can see.

Here is the y-axis.

That's in depth, so you can't tell how far away it is.

And the z-axis-- x, y, z-axis must be pointing out that way toward you.

So now I'm going to consider just a single point on the object and see what happens to it.

So I'm going to say to myself, let's put the object in some kind of standard position.

I don't care what it is.

It can be just random, just spin it around.

Some position, we'll call that the standard position, S. And that means that the x-coordinate of the standard position is x sub s.

And the y-coordinate of the standard position is y sub s.

And now I'm going to rotate the object three times.

Once to form the a picture, once to form the b picture, and once to form the c picture.

And you can make those choices.

Those can be anything, right?

So let's say that the a picture is out here.

So that's the a picture.

The B picture is out here.

And the unknown is up that way.

And so what I want to know depends on these vectors.

We'll call that theta sub a, and this is theta sub b.

And this one is theta sub u.

So I would like to know how x sub a depends on x sub s and y sub s.

And I can never remember how to do that, because I can never remember the transformation equations for rotation.

So I have to figure it out every time.

And this is no exception.

So what I'm going to say is that this vector that goes out to S consists of two pieces.

There's the x part and the y part.

And I know that I can rotate this vector by alpha sub a by rotating this vector and rotating that vector and adding up the results.

So if I rotate this vector by alpha sub a, then the contribution of that to the x-coordinate of a is going to be given by the cosine of theta sub a multiplied by x sub s.

So you can just exaggerate that motion, say, well if I pitch it up that way then the projection down on the x-axis is going to be this length of the vector times the cosine of the angle.

Now there's also going to be a dependence on y sub s.

Let's figure out what that's going to be.

I've got this vector here.

And I'm going to rotate it by theta sub a as well.

If I rotate that by theta sub a and see what the projection is on the x-axis, that's going to be given by the sine of the angle.

But it's going the wrong way, so I have to subtract it off.

So that's how I don't have to remember what the signs are on these equations.

Well, that was good.

And now that I'm off and running I can do what I did before.

It makes it easy to give the lecture.

Because this is going to be x sub b is equal to x sub s times the cosine of theta sub b minus y sub s times the cosine of theta-- oh, you're letting me make mistakes.

Shame.

I can generally tell by all the troubled looks.

But there should be some shouting as well.

That's the sine and that's the sine.

And one more time.

x sub u is equal to x sub s times the cosine of theta sub u minus y sub s times the sine of theta sub u.

And I forgot the b up there.

So there are some equations.

And we don't know what we're doing.

We're just going to stare at them awhile and see if they sing us a song.

So let's see if they sing us a song.

What about x sub a and x sub b?

These are things that we see in the image.

These are things that we can measure.

What about all those cosines and sines of theta a's and theta b's.

Well, we have no idea what they are.

But one thing is clear.

They're true for all of the points on the object.

Because when we rotate the object around by angle theta, we're rotating all of the points through the same angle, right?

So with respect to any particular view of the object-- here we are in the standard position.

Here we are in position a.

The vectors to all of the points on the object are rotated by the same angle when we go from the standard position to the a position.

So that means that for all of the images in this particular rendering, with a particular rotation by theta a, theta b, and theta u, those are constants.

Now remember this is for a particular theta a, a particular theta be, and a particular theta u.

As long as we're talking about all of the points for each of those rotations, those angles and cosines are going to be the same for all possible points on the object.

OK.

So now we go back to our high school algebra experts and we say, look at these first two equations, We've got two equations and what we can now construe to be two unknowns.

What are the unknowns that are left?

We can measure a and b.

Whatever the cosines are, they're the same for all the pictures.

So if we treat those as constants, then we can solve for x sub s and y sub s.

Right?

We can solve for x sub s and y sub s in terms of x sub a and x sub b and a whole bunch of constants.

But, I don't know, a whole bunch of constants, let's see.

We can gather up all of those cosines and ratios of sines and cosines and all that stuff and put them all together.

Because they're all constants.

And then we can do this.

We can say x sub u is equal to-- well, it's going to depend on x sub a and x sub b.

And by the time we wash or manipulate or screw around with all those cosines, we can say that the multiplier for x sub a is some constant alpha and the multiplier for x sub b is some constant beta.

So that's not a slight of hand.

That's just linear manipulation of those equations.

And that's what we wanted to show, that for orthographic projection, which this is-- there is no perspective involved here, we're just taking the projection along the x-axis-- we can demonstrate for this simplified situation that that equation must hold.

Now I want to give you a few puzzles.

Because this stuff is so simple.

Suppose I allow translation as well as rotation.

What's going to happen?

STUDENT: You just get the tau.

Basically, you get a constant.

PATRICK WINSTON: Yeah, you add a constant, tau.

But what do we need to do in order to solve it?

STUDENT: Subtract them [INAUDIBLE].

You subtract two equations and then [INAUDIBLE].

PATRICK WINSTON: Let's see, now we've got three unknowns, right?

I don't know tau.

I don't know x sub s.

And I don't know y sub s.

So I need another equation.

Where do I get the other equation.

STUDENT: [INAUDIBLE].

PATRICK WINSTON: From another picture.

That's why up there I needed four points.

That covers a situation where I've got three degrees of rotation and translation.

Here I got by with just two pictures in this illustration.

That one involved a tau translational element, so I needed three pictures.

And this one's got full rotation, so I needed four.

So great idea, works fine.

The trouble is it doesn't work so fine on natural objects.

It works fine on things that are manufactured because they all have identical dimensions.

So if I made a million of these in a factory, I'd have no trouble recognizing them.

Because all I'd have to do is take three pictures, record the coordinates of some of the points, and I'd be done.

The trouble is the natural world isn't like this.

And you aren't like this either.

I don't know, if I'm trying to recognize faces, it's not that easy to do all this.

First of all, it's a little difficult to find the exact point, the exactly corresponding points.

I made a mistake in doing it myself.

And if the computer made a mistake it would certainly make an error.

Because it would be using non-corresponding points to make the prediction.

So it would be way off.

But this is still in the tradition of working from local features in the objects toward recognition.

So having looked at that theory, we also find it a little wanting.

It works great it some circumstances, doesn't seem to solve the whole recognition problem.

Years pass.

Shimon Ullman comes up with another theory that's not so much based on edge fragments or the location of particular features but rather on correlation.

Taking a picture of, say, Krishna's face, taking a picture of the whole class, and then using that as a kind of correlation mask, running it all over the picture of the class, seeing where it maximizes out.

Now that's vague.

I'll explain when I'm talking about [INAUDIBLE] correlation in a minute.

But it's basically saying, if I have a picture of Krishna, where do I find him?

I'll find him in one place.

But you know what?

Krishna doesn't look like anybody else.

So I might not find any other faces.

And if my objective is to find all the faces, then maybe that idea won't work either.

Or, to take another example, here's a dollar bill.

We haven't had raises in quite well, so this is my last one.

It's got a picture of George Washington on it.

And I can look all over the class.

And if I use this is as a face detector, I'd be sorely disappointed.

Because I wouldn't find any faces.

Because thank God, nobody looks exactly like George Washington.

So the correlation wouldn't work very well.

So that idea's a loser.

But wait a minute.

I don't have to look for the whole face.

I could just look for eyes.

And then I could look for noses and maybe mouths.

And maybe I could have a library of 10 different eyes and 10 different noses and 10 different mouths.

Would that idea work?

Probably not so well.

The trouble with that one is, I'd find eyeballs in every doorknob.

There's just not enough stuff there to give me a reliable correlation.

So let's make this a little more concrete by drawing some pictures.

Halloween is approaching.

So here's a face.

All right?

Here's another face.

So those might be faces in my pre-recorded library of pumpkin faces.

Now along comes this face.

What's going to happen?

Well, I don't know.

Let's draw yet another face.

I don't know, that could be a pretty weird pumpkin face, I suppose.

But I mean it to be something that doesn't look very much like a face.

So if I'm doing a complete correlation with either of these faces in my library, neither one will match this one very well.

If I'm looking for fine features like eyes, then I've got these eyes everywhere.

So it doesn't help very much.

So you can see where I'm going.

And you can reinvent Ullman's great idea.

What is it?

We don't look for big features, like whole faces.

We don't look for small features, like individual eyes.

We look for intermediate features, like two eyes and a nose, or a mouth and a nose.

So when we do that, then we can say, now, here are two eyes and a nose.

Well, that's found in this one.

And what about the combination of that nose and that mouth?

Oh, that's over here.

But neither of those features can be found in the fourth image.

So that's the Goldilocks principle.

When you're doing this sort of thing, you want things that are not too small and not too big.

I've got the Rumpelstiltskin principle up there, too, by the way.

Because I meant to mention that Marr was a genius at naming things.

And even though many of his theories have faded, he's still known for these names like primal sketch and two and a half D sketch because he was such an artist at naming the concepts.

He even got credit for a lot of stuff that he didn't do.

Not because he was deliberately trying to get it inappropriately, but just because he was so good at naming stuff.

So we had the Rumpelstiltskin principle back then.

And now we have the Goldilocks principle.

Not too big, not too small.

But that leaves us with the final question, which is, so if what we want to do is look for intermediate-size features, how do we actually find them in a sea of faces out there?

See, I might have a library, I might take 10 of you and record your eyes.

Take another ten, record your mouths.

And they may be put together in a unique way for each of you out there.

But it's likely that I'll fin Lana's eyes somewhere else in a crowd.

And Nicola's mouth somewhere else in a crowd.

So how do we in fact go about finding them?

And I mentioned the term correlation a couple of times now.

Let me make that concrete.

So let's consider a one-dimensional face that looks like this.

Which is a signal.

And I'm going to consider a one-dimensional image.

And in that one-dimensional image I've got a facsimile of the face.

And the question is, what kind of algorithm could I use to determine the offset in the image where the face occurs?

So you can see that one possibility is you just do an integral of the signal in the face and the signal out here over the extent of the face and see how it multiplies out.

Or, to make it less lawyerly and more MITish, let's say that what we're going to do is we're going to maximize over some parameter x the integral over x of some face, which is a function of x and the image g, which is a function of x minus that offset.

So when the offset, t, is equal to this offset, then we're essentially multiplying the thing by itself and integrating over the extent of the face.

And that gives you a very big number if they're lined up and a very small number if they're not.

And it's even true if we add a whole lot of noise to the images.

But these are images.

They're not one dimensional.

But that's OK.

It's easy enough to make a modification here.

We're going to maximize over translation parameters x and y.

And these are no longer functions of just x, they're also functions of y.

Like so.

So that's basically how it works.

We won't go into details about normalization and all that sort of thing because that's the stuff of which other courses remain the custodians.

So would you like to see a demonstration?

OK.

All right.

So without realizing it, Nicola and Erica have loaned us their pictures.

And they are embedded in that big field of noise.

And it's pretty easy to pick out Erica and Nicola, right?

Because we are actually pretty good at picking faces out of these images.

So let's add some noise.

It's a little harder now.

What I'm going to is I'm going to run this correlation program over this whole image using Nicola's face as a mask

and seeing where the correlation peaks up, in spite of all the noise that's in there.

Boom, there he is.

I don't know, maybe we can find Erica too.

I forgot where she was.

I can't find her.

There she is.

Unfortunately the parameters aren't very good here.

Do you see that?

Let me get another version of this.

I'll just do some real-time programming.

I've been trying to reset the parameters so that the images in the demonstration comes out clearly up there.

Let's see if this works a little better.

OK, so let's add some noise.

And let's find Erica.

There she is.

There are some other things that look a little bit like Erica.

But nothing looks quite exactly like Erica.

So let's try Nicola's eyes.

So they stand out pretty clearly against the background.

Let's see if we can find Erica's eyes.

So they stand out pretty clearly against the background.

Notice that it also gets Nicola's eyes.

So two eyes is an intermediate-size constraint.

It's loose enough that it will match more than one person.

But it's not so loose that it's as bad as looking for one eye.

See, they're all over the place.

So two eyes and a nose, a mouth and a nose, that would be even better as an intermediate feature.

But it doesn't matter what the best ones are, because you can work that out experimentally.

So that's how correlation works.

And it's just amazing how much noise you can add and it'll still pick out the right stuff.

There's Nicola.

Boom.

Very clear.

Want to add some more noise?

I don't know, I can see it, but that's because I'm a pretty good correlator, too.

Boom.

I don't know, let's add some more noise.

It's just hard to get rid of it.

It's just amazing how well it picks it out.

That's good.

That's cool.

Now, but the reason that this is 30 years and we're still not done is there are still some questions.

This is recognizing stuff straight on.

How is it I can recognize you in the hall from the side?

Nobody knows.

One possibility is that you have an ability to make those transformations.

If so, then that alignment theory has a role to play.

Another theory is that, well, after I've seen you once I can watch you turn your head and keep recording what you look like at all possible angles.

That would work.

The trouble is, is there enough stuff in there?

Maybe.

We don't know.

Now what would it take to break this mechanism?

Well, I don't know.

Let's just see if we can break the mechanism.

Let's see if you can recognize some well-known faces.

Who's that?

Quick.

STUDENT: Obama.

PATRICK WINSTON: Oh, that's too easy.

We'll see if we can make some harder ones.

Yeah, that's Obama.

Who's this?

STUDENT: Bush.

PATRICK WINSTON: Oh boy.

You're really good at this.

That's Bush.

How about this guy?

STUDENT: Kerry.

PATRICK WINSTON: OK.

Now I've got it.

Some people are starting to turn their heads.

And that's not fair.

[LAUGHTER] PATRICK WINSTON: That's not fair.

Because you see what's happened is that if this kind of pumpkin in theory is correct, then when you turn the face upside down you lose the correlation of those features that have vertical components.

So if you have two eyes and a nose, they won't match two eyes and a nose when they're turned upside down.

Well, let's see.

We'll try some more.

Who's that?

STUDENT: Gorbachev.

PATRICK WINSTON: Gorbachev.

Who said that?

Leonid, where are you?

This is Gorbachev, right?

You can recognize him because of the little birthmark on the top of his head.

One more.

Who's-- oh, that's easy.

Who is it?

That's Clinton.

How about this one?

Do you see how insulting it is to be at MIT?

That's me.

[LAUGHTER] PATRICK WINSTON: And you didn't even know.

Oh, god.

So this might be evidence for the correlation theory.

But of course, turning the face upside down would make it very difficult to do alignment, too.

So it would break out alignment theory, as well.

Let me get that after class, Was there a mistake, or?

STUDENT: No, no.

I was just curious [INAUDIBLE] stretching would break the correlation.

PATRICK WINSTON: If what would break the structure?

What?

Stretching?

STUDENT: [INAUDIBLE].

PATRICK WINSTON: Elliot asked if stretching would break the correlation.

And the answer is, I think, stretching in the vertical dimension is worse than stretching in the horizontal dimension.

Because you get a certain amount of stretching in the horizontal dimension when you just turn your head.

By the way, since our faces are basically mounted on a cylinder, this kind of transformation might actually work.

That's a sidebar to the answer to your question, Elliot.

But now you say, well, OK, so this is not completely solved.

You can work this out.

But if you really want to work something out, let me tell you what the current questions are in computer vision.

People have worked for an awful long time on this recognition stuff and, to my mind, have neglected the more serious questions.

It's more serious questions are, how do you visually determine what's happening?

If you could write a program that would reliably determine when these verbs are happening in your field of view, I will sign your Ph.D. thesis tomorrow.

There are 48 of them there.

And that is today's challenge.

But since we're short on time, I want to skip over that and perform an experiment on you.

I want you to tell me what I'm doing.

STUDENT: [INAUDIBLE].

PATRICK WINSTON: So the best single-word answer is?

[INAUDIBLE]?

STUDENT: Drinking.

PATRICK WINSTON: OK, this is not a trick question.

OK, the best single-word answer.

Christopher, what do you think?

STUDENT: Toasting.

PATRICK WINSTON: Christopher.

Well, you.

You.

STUDENT: Toasting.

PATRICK WINSTON: What?

Toasting.

OK.

Not a trick question.

What's happening here?

Best single-word answer?

STUDENT: Drinking.

PATRICK WINSTON: Is drinking.

Which pair look more alike?

[LAUGHTER] PATRICK WINSTON: So that cat is drinking and nobody has any trouble recognizing that.

And I believe it's because you're telling a story.

So our power of storytelling even reaches down into our visual apparatus.

So the story here is that some animal has evidently had an urge to find something to drink and water is passing through that animal's mouth.

That's the drinking story.

So even though they look enormously different visually, the stuff at the bottom of our vision system provides enough evidence for our story apparatus so that we can give the left one and the right one different labels and recognize the cat is drinking.

And that's the end of the story.