[SQUEAKING]

[RUSTLING]

[CLICKING]

JOSH MCDERMOTT: So I want to start out today by talking about the generic viewpoint assumption, which is relevant to a few things that we've covered so far in the class. And then we're going to get into object recognition.

So the generic viewpoint assumption, it relates to this idea of an accidental viewpoint. So there's oftentimes these viewpoints that will create a special image on the retina. And in particular, they're special in the sense that they're not robust to small changes in viewpoint.

So here's an example. So this is like a scene in the world where you have these four different surfaces that have different orientations and they're at different depths. But if you view it from exactly the right point, it looks like that image. All right? So everything kind of lines up in exactly the right way.

And there's lots of examples of things like that. So here we've got the good old Necker cube viewed from a generic position. And then here's the accidental view, where if you view it such that two of the corners kind of line up, you get an image that looks like this.

So all of these images, they could be a cube. But interpreting c as a cube would require an accidental viewpoint. And we don't see a cube when we look at this. And for a long time, this was proposed to be consistent with this idea that we prefer a generic viewpoint. We assume what's called a generic viewpoint. So here's some other examples. This is a cube viewed from some different perspectives, generic and non-generic.

Here's another example. Here's an image of something that in the world could be this, could be this, could be this. But in these latter two cases, you've got this accidental view that kind of lines up in a special way with the thing in the world.

This is kind of a cool real-world example of this that was made here in this building. I don't know if you recognize the entrance here. Whoa. What's going on there? It's an accidental view. This was made on April Fool's Day about 10 years ago. And lo and behold.

AUDIENCE: Oh, Wow.

JOSH MCDERMOTT: Yeah. So there was some tape that was put down in a very tricky arrangement such that when you view it from a certain perspective, everything lines up. OK, so that's the idea of an accidental viewpoint, right? And it raises the question of what is it about accidental views that causes us not to assume them?

So intuitively, you kind of look at that image, and it seems the inference that that's a cube doesn't seem like the best idea. But it's hard to make precise. And so there was this beautiful paper that came out in the mid-'90s by Bill Freeman.

So Bill Freeman is a well-known computer vision researcher who works in CSAIL. At the time he was at MERL, Mitsubishi Electric Research Labs. That's like right around the corner. And he kind of came up with this really elegant way of thinking about the generic viewpoint assumption that kind of clarifies everything in a really beautiful way.

And it also kind of links up to some of the Bayesian inference principles that we've talked about elsewhere in this class. And so that's one of the reasons why I kind of wanted to go over it. So in all of these examples that we've been talking about so far, we've been talking about the notion of generic and accidental viewpoints.

But the notion that a scene can be generic or accidental extends to other things as well. So this is an example where this notion of generic and accidental extends to shape and illumination.

So remember, shape from shading is one of the ways by which we infer the three-dimensional shape of things in the world. And so the idea is that if surfaces are Lambertian, then the amount of light that is reflected by a surface depends on the angle between the incident illumination and the surface orientation.

So here's an example image. It's got one of these little gradients that sometimes looks like a bump, sometimes looks like a crater. And this is a set of possible scenes that could generate the image.

And so each one of these is a three-dimensional shape indicated by this grid paired with the direction of illumination. So far when we talked about shape from shading, we talked about the idea that the shape and the illumination could combine in different ways. So the problem is ill-posed. You can only do shape from shading if you know the direction of the illumination.

And so in the classic example of craters and bumps, they can be craters with the illumination coming from one direction or bumps with the illumination coming from the other. So that's what we have on the left and the right, number 1 and number 5.

But they're also like all of these kind of funky shapes, which when paired with exactly the right type of illumination, everything kind of lines up. The illumination lines up with the shape variation in exactly the right way to give you that observed image.

So the idea is that this is kind of like an accidental alignment of the illumination with the shape. It's conceptually similar to the idea of an accidental viewpoint, where things line up in exactly the right way to create a special image.

And this turns out to be kind of an analytically tractable version of the problem with which to think about the kind of general issue. And so the essential idea here-- so the question is, it's intuitively obvious that the good or the correct interpretations of that image is this and this.

These seem like they're not very good interpretations. The question is, why is that? And in what sense are they not good? And the essential intuition is that these situations where you get accidents in the scene interpretation are situations where the image that would be rendered by that accidental scene is very unstable with respect to the scene parameters.

So here we have these two possible shapes that, with the right illumination, can produce the observed image, which was something like this. And what is shown here are five different images. So the one in the middle is the one that is the observed image. And the other ones are the images that would be rendered for slightly different directions of illumination.

And so the point is that for the bottom shape, for the bump, the image that would be generated with these slightly different directions of illumination does not change very much. Those all look the same. So they're slightly different, but the difference is pretty small.

Whereas for the accidental scenario-- so given this shape, if you change the illumination just a hair, the image that you would get from that combination of shape and illumination changes a lot, because you really have to get the shape and the illumination to line up exactly right in order to render that image.

So that's the first idea, is that there's something very different about these scene interpretations, which is this one is stable and this one is unstable. All right. Well, so what, you might say.

So let's think about the relevance of this for good old Bayesian theories of perception. OK, so remember how the probability of a scene interpretation given the image can be formalized with Bayes' theorem. So we generally think of the goal of perception as finding the scene that is most likely given some observed stimulus, like an image.

So that's the posterior probability of the scene given the image. We want to maybe maximize that or find a few scenes for which that's high. So here's the posterior. All right, now, this looks a little different than what you have been used to seeing, because now there are two variables here. There's beta and there's x.

So we're now dealing with a situation where the scene is defined by at least two variables. So in this particular setting, one of those variables is the shape of the surface and the other variable is the direction of the illumination.

And so we're assuming that this is a setting in which one of those variables is something that you particularly care about. So in this particular case, you want to estimate what the shape is. But there's also this generic variable or parameter, the illumination that's going to affect whether or not the shape can explain the image.

So this is just Bayes' theorem. So we've got the posterior here. Anybody remember what is this thing called? y is the image. What's that? What is this? Probability of the image given the scene interpretation, yeah?

AUDIENCE: The likelihood.

JOSH MCDERMOTT: Yep, that's the likelihood. So it just looks a little different because you've got two variables here. And remember, what is this?

AUDIENCE: Prior.

JOSH MCDERMOTT: That's the prior, yeah. So you've got two priors, one over the shape and one over the illumination. So this is just the thing that you're used to seeing. It just looks a little different because there's now two variables.

OK, so here's the catch. So let's just suppose that what we really want to do here is estimate shape. So really what you want to do is maximize this. You want to find the shape that's most probable given the image.

So how do you get that? Well, you take the joint distribution of shape and illumination, and you integrate with respect to the generic variable, the illumination, which is x. So if you've got a joint distribution and you want to get the marginal distribution-- so in this case, you start out with the distribution of beta and x and you want to just get the distribution over beta-- you integrate over the other variable.

So this is just standard probability theory. And so this is just the integral of this thing with the quantities that don't depend on x kind of brought out in front of the integral. So then what's inside the integral?

Well, we've got our likelihood here and then the prior probability of the illumination. But the key thing here that's going to be really important is the likelihood. So what is this integral doing? Well, you can think of it as adding up the likelihood here for all different possible values of the illumination.

All right, so let's review the likelihood. And I'll say there was a question on the last exam asking about the likelihood, and not everybody got it right. Some people kind of got it partially right. So this is a good thing to review.

So the key idea of the likelihood is it measures the extent to which particular scene parameters can account for an observed image y. So typically, likelihood functions will assume a Gaussian probability. And that's what's written here.

So it's our likelihood. And it just says that it equals this thing. But the key thing to note here is that we've got y minus this function f of x and beta. So remember, x and beta are the variables that describe the scene. x is the illumination, beta is the shape, and y is the observed image.

f here is a rendering function, which gives the image created by the generic-- that's the illumination-- and scene parameters-- that's the shape-- x and beta So what is this doing? Well, it's a comparison between the observed image and the image that would be generated by these hypothesized scene parameters.

So that's what the likelihood does. It says, for a given hypothesis about the world, what is the stimulus that that would generate? And how similar is that to the observed stimulus? So this is just one example of a likelihood function.

So the key thing is the difference between the observed image and the rendered image that you would get for particular values of illumination and shape. And the fact that this is inside this normal distribution is because we're assuming Gaussian noise in this case. But the key idea is that it's just taking a difference here.

OK. So we're trying to compute the probability of a given shape given the image. We've got to integrate the likelihood here over illumination. And each value of the likelihood is going to be comparing the observed image to the image that's predicted by particular values of illumination and shape.

So that predicted image is what is shown up here. So these images are the output of a computer graphics program that takes shape and illumination and generates an image. So we're taking an integral, remember, over the illumination direction. And that integral is going to add up the likelihood for each of these images.

The likelihood will be high whenever the rendered image is similar to the observed image. So the key idea here is that in this scenario here for this shape, you're adding up a bunch of likelihoods that are going to be consistently high, because irrespective of the exact direction of the illumination, you get out an image that is pretty close to the observed image.

Whereas for the funny shape, the likelihood is only high for this one special direction of illumination. And for all other slightly different directions of illumination, the image that is rendered is very different from the observed image, and so the likelihood will be low.

And so the consequence is that when you do this integral, for the funny shapes, this sum or integral is going to end up being relatively low. And that's what happens. So here, we end up with the total score. It's like the posterior of all of these different shapes.

And so the crater and the bump come out as pretty probable. And then the funny shapes come out as less probable. What questions do you have about that? Yeah.

AUDIENCE: Like, what are these numbers?

JOSH MCDERMOTT: There's some constant of proportionality that is going to turn them into probabilities. I don't remember. Yeah, it's proportional to the posterior. I don't know what the-- I forget what the units are.

AUDIENCE: [INAUDIBLE] you expect the ones they have to do with 1/2?

JOSH MCDERMOTT: To be?

AUDIENCE: Sort of close to 1/2?

JOSH MCDERMOTT: Ah, if this was actual probability? Yeah, it's not actual probability. It's something that's proportional to that. Yeah. So what matters is the relative height of these things.

AUDIENCE: So it would be linearly proportional to the actual number?

JOSH MCDERMOTT: I think so. It's been a long time since I read the paper in detail, so I don't remember. What really matters here is just the fact that these are a lot higher than that. So I'm just trying to give you the intuitive kind of explanation here. Yeah. Other questions?

So here's another example. So this is an image of a piece of sculpture. And it's been corrupted by noise. And so here we have two different possible shapes that could account for the image.

What you see, obviously, is this. But this kind of crazy-looking shape, if lit from exactly the right direction, would generate something very, very close to that image. But the key concept is that because the illumination has to be exactly right for that shape to generate the image, if you integrate out illumination, this thing ends up being a lot more likely. That's what you see.

All right. So the take-home message here is that we don't actually really need a generic viewpoint assumption. No assumption is really needed. Interpretations under generic views or generic anything are just more likely given standard probabilistic inference, assuming that we integrate over variables that we don't need to estimate precisely.

And so that's actually a slightly non-trivial statement. So all of this is positing that when you are seeing shape, you're integrating over illumination. It's like you don't really care about the illumination. You're not trying to estimate that. You're just trying to estimate shape. And that is maybe an assumption about the content of perception, which is non-trivial. Yeah.

AUDIENCE: My question is in regards to the noise. Because we saw that in the other photo, it was noisy. And that's why there could be two interpretations. Is there a certain percentage, certain amount of noise in which those two kind of perceptions would be more equal in our brains? Like, the thing gets noisier and noisier, and eventually we can't really tell the difference?

JOSH MCDERMOTT: Yeah. OK, so first of all, I should say that in this particular case, the noise here was added to actually cause this shape to better account for the image than this one. So given exactly the right illumination, the likelihood is actually higher for this than for this.

And that's what this is supposed to show. So these are the images that would be rendered by these two shapes. And so the point is that this one doesn't actually model the noise, whereas this one does.

So it's just sort of a contrived scenario that's supposed to make the point that even if you have a situation where you could, in principle, better account for the image with this crazy interpretation, once you take into account the fact that you need to integrate over illumination, this is going to end up being the preferred explanation. Yeah.

OK. So key concept here is, again, the idea that we can think about perception as finding the most likely scene interpretation, given an image or sound or whatever kind of sensory input you're dealing with, that we're trying to maximize the posterior, that you can decompose the posterior into the likelihood and the prior.

That the likelihood involves a comparison between the observed stimulus and the stimulus that would be generated by a hypothesis. And then in this particular application of this, there's this additional wrinkle that if you've got multiple variables in your scene interpretation and you're trying to estimate one of them, then you iterate over the other one.

And then that kind of causes this generic viewpoint assumption. And so the other thing that I'll say is that there's lots of different applications of Bayesian approaches to perception. We've talked about some applications where the work is really being done by the prior.

So for instance, when we talked about motion perception, the critical thing to explaining some of these motion illusions is that there's a prior that favors slow speeds. That kind of does a lot of the work for you.

This is a case where actually the prior is not really super relevant. In fact, I think they assumed a uniform prior. You could imagine maybe making this more realistic by actually having a prior that favors illumination from above, which maybe is a little bit more realistic.

But here, really, the work is done by the likelihood and by essentially probability theory, just basically saying that you have to integrate over the other variable that you're not trying to estimate. But it's the same basic kind of mathematical framework and the same kinds of ingredients.

So this same kind of idea comes up a lot in a lot of different places in perception. There's some interesting ones with illusory contours that are kind of fun to look at and think about. So if you look at these two displays, the one on the left, people typically see much stronger illusory contour than the one on the right.

Is that what you all are seeing? Very clear kind of illusory square on the left, not so clear on the right. And you can at least qualitatively think about this in terms of generic and accidental views. So this particular display-- so let me back up.

One way to think about what we see when we look at this is the idea that there are two main possible explanations for the image. One is the kind of thing that most people see here, which is like, well, there's just a bunch of little patterns in the image that are all kind of just lying there.

And the other is what people typically say they see when they look at this, which is that there's a white square which is kind of on top of some other stuff. Now, this one could, in principle, not have the white square and could just be explained by these features in the image. This one could, in principle, also have the white square.

So the idea is that if the explanation of this stimulus is that there was a white square here, you would have to posit an accidental dental viewpoint. And specifically-- so imagine there's a white square here and you change your view.

Then the image is going to change, because the square is now going to overlap with these features differently than it otherwise would. So you need to posit an accidental viewpoint here. Whereas in this case, that's not so true.

Like, from a different viewpoint, well, you still get line endings. They just kind of move around a little bit. But the image doesn't really qualitatively change. So another case where this idea of generic and accidental views seems to have some explanatory power.

And the framework that we just talked about gives us some mathematical basis for thinking about that, which is the idea is that you don't normally care about exactly what the viewpoint is, so you're going to integrate over that in order to estimate the things that you are trying to estimate.

Here's another one. This one is maybe a little bit stronger. Same idea. So pretty strong illusory contours on the left and not so strong on the right. OK. OK. Any questions about that? Yeah.

AUDIENCE: Could you explain how an accidental view is necessary for the stimulus on the right?

JOSH MCDERMOTT: Well, yeah, so the notion is that-- so the question is, all right, why don't you actually infer that there's a square here? And in order to actually evaluate the probability of that interpretation, the proposal is that, well, you have to integrate over different viewpoints.

And the point is that let's imagine that there was a square here. There would be one particular viewpoint where the combination of the square and all of these funny things would generate this image.

And the idea is that in a world where there's a white square kind of hovering out in front of this thing, if the viewpoint changes a little bit, well then the square is going to start to occlude these corners. And the image that you would get from that potential scene would be very different from the one that is actually observed.

So it's exactly analogous to that shape from shading example, where there's this one viewpoint where with the white square, you get exactly the observed image. And then for other viewpoints, you get something that deviates quite a lot from the observed image. Whereas with this example, that's not the case, because you don't have these funny corners here that would be able to get occluded or not. Does that make sense?

All right. Let's talk about object recognition. So this is a simplified view of visual processing. So we spent the first part of the section of the class on vision, sort of talking about what we called early vision.

You can think of that as feature extraction. You got all these filters in your visual cortex that make all these useful measurements of orientation and spatial frequency and binocular disparity and wavelength and stuff like that.

We then talked about all of these processes that we kind of loosely think of as mid-level vision, many of which are different forms of shape estimation or things that might be relevant to recognizing what's there-- color estimation, shape from contour, from shading, from texture.

So these all kind of, from a certain perspective, give you information about three-dimensional structure and other properties of the world. And we often think that all of that stuff is in the service of recognition, telling us what's out there in the world. And that often is thought to be the realm of high-level vision.

And again, these are all very loose terms. And in reality, there's kind of lots of interactions and feedback and things like that. But it's still often kind of a useful division, at least in the kinds of things that people study.

All right, so what happens when we recognize things? So here's an image. You can look at this image and name lots of things that you see, some of the things that are circled here.

This process is usually pretty effortless for humans. And you really just look around the image and you can see all this stuff. So the fact that it's pretty effortless for us for a long time kind of belied the challenge of the problem.

So this is a really famous memo that was put out here at MIT in 1966. How many of you have seen this before? OK, a couple people. Yeah. So this often gets trotted out as an example of how our perceptual competencies are deceptively-- they can seem a lot simpler than they actually are.

So this is a memo that was issued by the Artificial Intelligence Group here at MIT a long time ago. And the proposal was to solve the problem of vision one summer in 1966. So it's called the Summer Vision Project. This was assigned to a student.

So it says the final goal is object identification, which will actually name objects by matching them with a vocabulary of known objects. All right. And so, of course, this was a long time ago. And it's only been very, very recently that we've had machine vision systems that can do anything close to this.

So it's very easy for humans. And people didn't initially understand the complexity of the problem. So humans and other primates, when we solve object recognition, we're mostly using the central 10 degrees of vision, because we just kind look around at stuff that we want to recognize.

So we make eye movements around scenes. Can make three to four saccades a second. So you can think of object recognition as consisting of these brief snapshots, maybe 200 milliseconds. You look from one thing to the next. And you're mostly able to recognize what's there from these brief snapshots.

All right. So why is the problem hard? So one reason why it's a difficult problem is that there are lots of different kinds of objects that you can recognize. So this is just a small subset of possible things, different types of animals, different types of objects. There's lots and lots of things that you can ascribe names to.

Another reason why it's hard is that the same physical source in the world-- so the same kind of thing-- can produce many different types of images. So you can vary the position, the size, the pose, the illumination that's directed towards an object, like the background that it's on.

And the ability to tolerate this kind of variation is normally referred to as invariance. And we talked about this at the very start of the class. One of the things that's hard about perception is that you have to be able to recognize things invariant to all these sources of variation.

So the images that are produced by an object are highly variable. This can be due to occlusions, illumination changes, viewpoint differences, non-rigid deformations. So here you're able to tell that one of these faces on the right is the same as the one on the left, and that one is different.

So as a consequence of this, the same object can produce wildly different images depending on the viewing conditions. So these are all the same pair of shoes. But the pixel array that you get in each of these cases is very different. And also different exemplars of a given object category can also generate very different images. But you can tell that all of these things are shoes.

So what parts of the brain are involved in visual recognition? Everybody remembers how we think of the visual system very coarsely as being organized into dorsal and ventral pathways. And we think of the ventral pathway as mediating our recognition abilities.

And there's now lots and lots of evidence in support of this. So the initial clues to this came from lesion studies. So these are individuals who have brain damage. So if you have damage to the temporal lobe, either in humans or monkeys, that leads to something called visual agnosia, which is the inability to recognize objects.

And in monkeys in particular, there's some pretty powerful dissociations between the consequences of lesions to the temporal lobe and lesions to the parietal lobe. So these are two example kinds of tasks that were used in the service of this.

So temporal lobe lesions leads to deficits on shape discrimination tasks. So if you show a monkey a particular shape before the trial and then ask it to pick which one is similar to it with temporal lobe lesions, that task gets really difficult. By comparison, if the task here is to choose the object that's close to the landmark-- in this case, the cylinder-- parietal lesions cause this task to become difficult.

So this is an example of the responses of a patient with temporal lobe damage, studied by Martha Farah, a neuropsychologist. So the task here is to look at the object on the left and then choose the object in the row that's to the right of it that has the same shape. And you can see that the person is not able to choose the correct shape. Yeah.

AUDIENCE: Does the person ever adapt or is it they permanently can never tell an object?

JOSH MCDERMOTT: Yeah, I mean, there's usually some degree of recovery that I think can vary in the extent. So the brain is remarkably plastic, right? And so typically, most of the cases where a human has brain damage of this sort are the result of stroke.

There's some kinds of other things that can cause it, but it's mostly stroke. So the blood flow to a particular part of the brain is cut off. It's deprived of oxygen. Neurons die if they're deprived of oxygen. And so there will be some part of the brain that can be damaged.

Can be pretty small, can be pretty big. It varies. And so typically, my understanding is that you can have very profound deficits, like in the immediate aftermath of that. And then things often get better. So there's some degree of partial recovery.

So in some cases, we think that's because other parts of the brain can end up mediating some of the functions that the damaged part would carry out. It's often not complete. I think it depends a little bit on which area is damaged and how big it is and things like that. So I think this patient actually had had the brain damage for quite a while, and so I think these deficits are fairly permanent. But I don't remember, to be sure.

So this is a really interesting case of a patient who is unable to identify line drawings of common objects and also couldn't copy them, but could draw from memory. So this is a situation where they were shown these two line drawings, asked to name them, and they couldn't, asked to copy them, and basically couldn't.

Although it's kind of interesting that you can see what looks like maybe they see some of the texture, like on the page, and are trying to replicate that. But when told or when asked to please draw an apple, like from memory, they're able to do something reasonable, or please draw a book, they're able to do something reasonable.

So the inference from this is that what is damaged here is the recognition machinery. So they can't look at the image and generate a representation of shape from the visual input. But it appears that they have some form of intact memory for the shape and are able to draw from it. So pretty interesting.

So we think, as a consequence of the evidence that I just showed you and lots of other kind of similar pieces of evidence, that our recognition abilities are mediated by the visual ventral stream shown here. Yeah.

AUDIENCE: On the previous slide, if you give them color and ask them to copy a filled-in color drawing of an apple, how do they perform if it's [INAUDIBLE]?

JOSH MCDERMOTT: So I'm pretty sure that somebody with agnosia would have pretty normal color vision, but I don't know for a fact. Yeah. You certainly see the opposite dissociation. Like, we saw examples of people with achromatopsia, where you can suffer from brain damage that will impair color vision, but leave object recognition and most other aspects of vision pretty normal. So I'm pretty sure the opposite can be true. Yeah.

AUDIENCE: Does it impair color vision in the eyes or does it impair your understanding of color?

JOSH MCDERMOTT: Oh, achromatopsia that we discussed results from cortical damage. So again, somebody has a stroke, so the eyes are fine. And the cones are fine and normal. It's just that the cortical machinery that is mediating the conscious perception of color and presumably color constancy and things like that, that's been damaged. Yeah. Yeah.

AUDIENCE: I know not everyone is an artist, but the drawings from memory aren't exactly the best. Is that related to the [INAUDIBLE]?

JOSH MCDERMOTT: I don't know. Yeah, I'm not sure. So Martha Farah has a nice book. I think it might be called*Visual Object Recognition.* It's this short little book that you can find in the library, if you want to read more about these patients. Yeah.

AUDIENCE: [INAUDIBLE] copies being that they cannot recognize what the model is supposed to be. Is that the inference from this experiment?

JOSH MCDERMOTT: I think the inference is that they're unable to take visual input and generate a representation of object shape. And so that renders them unable to then name the object, because the naming is dependent on having a representation of object shape.

And it also means that they have a hard time copying the object, presumably because they can't relate what they are drawing to what's on the page. They're not able to extract a shape representation from that.

AUDIENCE: If they were shown the model and they also told that this is an apple, for example, would they draw something closer to a copy?

JOSH MCDERMOTT: Yeah, I don't know. I mean, they can draw from memory, right? So presumably if you told them to draw an apple, they would do something like that. Yeah. All right. So we got the ventral stream.

So we think of the ventral stream as culminating in inferotemporal cortex or IT. And then there's a few other stages. So we've got a pixel image as input. There's Retinal Ganglion Cells. That's RGS. The LGN, that's the visual part of the thalamus. V1, primary visual cortex. V2, V4, and IT, kind of major components of the ventral stream.

And one way to think about this is that each of these visual stages is performing a transformation on the input. So remember, we previously introduced the idea that you could use convolution as a way to think about what a particular type of receptive field would do if it was applied, say, at every location in an image.

That's what a big set of retinal ganglion cells or a big set of simple cells would do. So you can kind of think of that transformation as kind of like a convolution operation with a particular set of filters. So each of these regions is doing something to the representation.

And then there's also potentially feedback and recurrent connections. And these numbers here are the response latencies. So it's about 60 milliseconds from when the image hits the retina to when you get the initial response of V1. And about 100 milliseconds to when you get a response in IT.

So one kind of early important result on human object recognition indicated that recognition is pretty fast. So this was a study by Simon Thorpe and colleagues. And the experiment involved looking at a very rapidly-presented sequence of images.

So each one of these would are presented one at a time, boom, boom, boom, boom, boom. They'd be up for 20 milliseconds. And the task is just to press a key if you see an animal. So you would press the key there, because that's a fish. Press this key there. I'm not sure what that is, but maybe you would press it there too. All right. So you press keys for animals.

So this is a histogram of the reaction time. So it's like how long it takes people to press the button from the time in which they see the image. So the average reaction time is something like 400 milliseconds.

And there's a speed-accuracy trade off. So if the reaction time is shorter, people tend to do a little bit less accurate. If they take a little bit longer to make the decision, they're more accurate.

All right. But you can't really draw much of an inference from that 400 millisecond number about the time it takes the visual system to recognize the object, because that reaction time involves a lot of things other than recognition. In particular, you have to press the button.

So once your visual system generates some recognition, then somehow it's got to communicate to motor cortex that you need to press the button. Motor cortex has got to send the motor command, travels down your nerves, and eventually the muscles contract and you press the button. OK?

So there's a lot of that that's going into this 400 milliseconds. And so what they did in this study was instead to just let the brain reveal when there was information about recognition made available.

So this is a brain response called an ERP or an Event-Related Potential. And the way that you make this measurement is you put a bunch of electrodes on the scalp. So you're recording scalp potentials.

And this is showing-- so there's a lot of electrodes. And this is one way to depict the voltage that's being measured at different points on the skull. You can see that it evolves over time in some particular way. This is not what's so important.

So here, this is a particular part of that brain response. And what they've done here-- so event-related potential, so what that means is that you're recording these continuous signals. And then you average them in some particular way, typically with respect to a stimulus.

So what they're doing in this case is they're taking all of the times when an animal was presented on the screen and then averaging the brain response that occurred after that. So it's kind of like a spike-triggered average, except it's a stimulus-triggered average, where the stimulus-- you're averaging it separately, depending on whether it's an animal or whether it's a non-animal.

So you get one average brain response for all the animal images and one average brain response for all the non-animal images. And so this is what you get. So this is time. This is the response. Whether it's positive or negative doesn't really mean anything because you just have these electrodes and you're measuring differences in voltage.

And what you really care about is at what point in time there will be some difference in the brain depending on whether the image is an animal or a non-animal. And what does this show? Well, you get this response here.

And they're very, very similar up until looks like about 150 milliseconds. At that point, the animal and the non-animal curves diverge. And so this curve up above them is the difference between the two. So you can see that that is close to zero until about 150 milliseconds. And then you see a difference between the responses to animals and non-animals.

So at 150 milliseconds, the brain is able to discriminate whether you're looking at an animal or a non-animal. And the idea is that because this is being averaged across a gazillion different images, that probably is indicative of some fairly abstract property of the image.

All right. So what's significant about that number of 150 milliseconds? Well, the latency of the V1 response is about 70 milliseconds in humans. And so this result implies that there's pretty fast recognition of objects, potentially mostly feedforward.

So the idea is that you're able to recognize objects, at least at the level of animals versus non-animals, with one pass through this processing pipeline. And so these numbers here are for the macaque. They would be longer for the humans, because humans have a bigger brain. So 150 milliseconds is probably pretty close to the time at which things get to IT in the first pass.

All right. And so this suggests that at least there's some component of recognition that we can think of as a pretty feedforward process. It's not to say that all of recognition is always like that, but there's some component of it that is potentially mostly feedforward and pretty fast. Any questions about that?

OK, so here's the ventral pathway plotted in a slightly different way. So we've got these different areas. And these are the latencies in the macaque. We've already talked quite a lot about this stage here, V1. So remember, in V1, we've got simple cells and complex cells.

So simple cells are orientation selective. Complex cells are orientation selective, but they're more tolerant to position. So they have some simple form of invariance. We've talked a little bit about V2. So one of the things that you see in V2 is some more sophisticated forms of selectivity for edges, like that you get responses to illusory contours.

There's some sensitivity to figure ground relations, border ownership cells you find there. We talked a little bit about V4. So we talked about how lesions of V4 tend to often produce deficits in color perception. But V4 also kind of exhibits neurons that have some interesting selectivity for shape.

So here's one fairly well-known study that presented neurons in V4 with these kind of funny-looking gratings, just different kinds of complicated patterns. And essentially, the finding is that a lot of the neurons are tuned to specific types of these patterns.

So this one really liked these few patterns the most. This one really likes spiral shapes. So just kind of you get slightly more complicated shape selectivity. So if you did this in V1, you wouldn't see anything like this, because the responses there would just be determined by the approximate dot product between one of those simple cell receptive fields and these patterns.

Here's another study from Ed Connor which created a basis for shapes. Essentially the shapes can be curved and then pointy in these different directions. So you get this big variation of these simple shapes. You can measure the response of V4 neurons and find that individual neurons can be-- the responses can be explained fairly well in terms of properties of these particular shapes. So this particular one is it really likes pointy things that are pointing up, looks like.

So you move up the ventral stream, you see selectivity for more complicated kinds of things culminating in inferotemporal cortex. So a few things to know about inferotemporal cortex. One is that it's dominated by the central visual field.

So these are the results of a tracer study from Leslie Ungerleider's lab, where they injected tracers in either the foveal or the peripheral part of V4, and then looked at where they landed. And so what you can see-- so this is different parts of inferotemporal cortex here.

And the main take-home message is that there's a lot of red. So the injections to the foveal part of V4 end up really labeling inferotemporal cortex pretty thoroughly. And there's not as much green. So that's consistent with this idea that object recognition is really predominantly a function of central vision, because you just look at the stuff that you want to recognize. And so IT is really dominated by inputs from the central part of vision.

And so there are lots of these famous reports of pretty complicated forms of stimulus selectivity in this part of the brain. So Charlie Gross was a professor at Princeton who pioneered the study of inferotemporal cortex. And a couple of our faculty here in this department worked in his lab, Bob Desimone and Earl Miller.

And these are different stimuli that were presented when they were recording from neurons in the IT. And it looks as though the neuron is really sensitive to the shape of a hand. And so this is an excerpt from one of their papers.

They say the use of these stimuli was begun one day when, having failed to drive a unit with any light stimulus, we waved a hand at the stimulus screen and elicited a very vigorous response from the previously unresponsive neuron. We then spent the next 12 hours testing various paper cutouts in an attempt to find the trigger feature for this unit.

When the entire set of stimuli used were ranked according to the strength of the response that they produced, we could not find a simple physical dimension that correlated with this rank order. However, the rank order of adequate stimuli did correlate with similarity for us to the shadow of a monkey hand. So selective for some complicated kind of meaningful structure.

Here are some other examples. So this is another early study that kind of found another example of a neuron that's tuned to the shape of a hand. Here's an example where there's some neurons that seem to be selective for faces. So the idea is that they're tuned to these very specific combinations of features. They're highly selective.

Now, one of the other things that we've talked a lot about in the context of the visual cortex is the idea that the cortex is organized into maps. So one of the main principles we talked about is retinotopy, where you have maps of the visual field kind of laid out in the visual cortex.

And so you don't see very clear retinotopy in inferotemporal cortex, in part because the receptive fields are very large. But there's nonetheless some functional organization. So this is the result of a study that was done here in this department where they presented a very large set of images.

So this is the object ID number. There's like 77 images of just kind of complicated shapes-- monkey faces, cars, elephants, just like a big heterogeneous set of object shapes. And there was an electrode that's kind of inserted into the brain. And they measure the responses to each one of these shapes at different points along that red line.

And so each one of these rows is the response that's measured at a different point kind of in the brain. And presumably, these are multi-unit recordings. So they reflect the responses of a bunch of neurons, but in some spatial neighborhood.

And so you just get some big vector for every one of these recording sites, a 77 dimensional vector. And then what is plotted on this graph here is the similarity of those vectors as a function of the separation between the recording sites.

And the point is that this kind of drops. So when the spatial separation is small that you have fairly similar responses, even though they're complicated and hard to describe in words or in terms of physical stimulus dimensions, and then as you get to further points in the brain, the responses kind of become more distinct. So there's some degree of spatial organization.

All right. So another general principle of organization in the ventral pathway is that receptive field sizes increase as you move up the pathway. So in particular by the time you get to anterior inferotemporal cortex, you have these very large receptive fields, even larger than what you get kind of a little bit earlier in inferotemporal cortex. So big receptive fields.

Individual IT neurons typically show some degree of tolerance to variation in position and size. So this is an early study by Nikos Logothetis. This graph is plotting the extent of response to a preferred stimulus at a whole bunch of different sizes.

So this is degrees of visual angle from small to big. And the response is pretty consistent. This is the stimulus presented at different locations on the screen. Maybe there's some spatial selectivity, but it's quite broad. OK. So these are all various kinds of clues about what is presumptively the locus of visual recognition.

All right. What we're going to talk about now is how we can think about what these tuning properties kind of might be doing in a more rigorous sense. So I want to pose to you the question of what is the computational function of the ventral pathway, or really of any system that is attempting to do recognition?

And in order to pose this question, we're going to think about the representation of images. And this generalized could be any kind of stimulus, but we're going to be talking about it here in vision. We're going to think about the representation of images in a population of neurons.

So the idea is that you look at an image, and then some stage of your visual system generates a representation of that image via the responses of its neurons. So we can think of the representation of one image as a point in a very high-dimensional space where the axes of that space are the responses of different neurons.

There might be thousands of these neurons, so it could be a thousand-dimensional space. So you see an image and there's a response generated in your brain. Now in actuality, it's more complicated than that, because maybe the response evolves over time. But let's for the moment just assume that you get one number, which is maybe the average response.

All right. And the other thing that we're going to keep in mind here is that we're talking about the problem of object recognition. And one of the challenges of object recognition is the need to be invariant to all of the different factors that can cause the image that's produced by an object to change.

So in this particular case, it's the viewpoint. So there's this person, evidently named Joe, who's being seen from different views. And those different views create different images. So each one of those images will generate a representation, which again, in this framework, is a point in this high-dimensional space of neural responses.

And the collection of all of those points, we're going to call that a manifold that corresponds to Joe. And so the proposal here is that what we would like to achieve with our brain, in particular with our visual system, is a good set of visual features that has this property that we're going to call an explicit representation of object shape.

And so this property is going to be that when you look at all of the images corresponding to one particular type of objects-- in this case, Joe-- they're going to reside in one part of this high-dimensional space. Whereas the images of some other kind of object, Sam, will be in another part of the space.

And we will evaluate whether the manifold for Joe and the manifold for Sam are really separate by testing whether we can separate them with a plane, or a hyperplane in this case, because it's a high-dimensional space.

And you can evaluate that with a linear classifier. So a linear classifier by definition is based on a hyperplane. And it has this hyperplane. And when points are on one side of the hyperplane, it says it's class 1. And when they're on the other side of the hyperplane, it says it's class 2.

OK, so that's the proposal. This hypothesis is that the visual system should be performing transformations that produce representations like this, so that the images corresponding to one kind of thing end up in one part of this representation space, and the images corresponding to other parts of things lie in another part of that space, such that you can separate them with a plane and thus classify things with a linear classifier.

OK, so what is a linear classifier? So a linear classifier uses projection onto a vector. So the classifier is defined by a direction, in this case b, and a threshold. So you take some observed response to an image. Could be a. You project it onto the classification vector.

And that projection, depending on which side of the classification plane it is, would determine whether you're looking at class 1 or class 2. All right? So it just involves dot products and thresholds.

And it's natural to think that you could potentially instantiate this with a neuron. So we often think of what a single neuron might be doing as computing a dot product. So think of it like synaptic weights and taking an input and multiplying it by all the weights, and then adding those up.

And that gives you the membrane potential. And then there's a threshold inside the soma, which determines whether you respond or not. Again, that's a super oversimplified description of what a neuron actually does. But again, very loosely could be mapped on to this kind of math. So we often think that neurons could potentially be instantiating linear classifiers.

All right. So here's the problem of vision. So we talked about how recognition is hard because there's all these ways in which images of the same kind of thing can vary. And another way of stating that is that the manifolds that correspond to different types of objects, when viewed in the pixel space, in the input space, are very tangled

So in the pixel space, what makes recognition hard is that all of the images of one person are not cleanly separated from all the images of another person. So those manifolds are all tangled up. This is why you need a visual system.

So the hypothesis is that what the visual system is doing is performing transformations on the representations of the images to end up with something where the representations of all the images of one kind of thing are in one part of the space, and all the others are in the other part of space.

And we call that explicit. So the notion of making something explicit is that you make it easy to read off from the neural population or some other kind of representation. And easy to read off is typically instantiated as being able to use a very simple classifier-- and a linear classifier is the simplest type of classifier-- and being able to extract that information.

So that's the notion of something being explicit. So the idea is that the identity of the individual in the pixel input is implicit. So it's there because the visual system just gets that image as input and is able to recognize something. But it's not explicit.

It's not easy to read off. But then you perform this series of transformations and you generate these representations where by hypothesis, things are explicit. So that's the idea. Now, we haven't said what these transformations are.

And in fact, it's kind of non-obvious how to write down transformations that would do this. But we can ask the question of whether or not this actually happens in the human or the monkey visual system.

OK, so what I'm going to tell you about here is a test of this idea in inferotemporal cortex. So this is an experiment where monkeys were presented with a big set of 78 objects from eight categories.

So the objects could vary in their scale and position and also in what kind of thing they were. So here's a monkey. Here's a tank. Here's some kind of toy. OK, so lots of different objects. They're presented very briefly. The monkeys are fixating. And you're recording from a whole bunch of neurons in the inferotemporal cortex while the monkeys are looking at these objects.

So this is the outcome of the neurophysiological experiment. So we've got 78 images here. So the columns represent images and the rows represent recording sites. So you kind of loosely think of those as neurons. In actuality, each one of these is more than one neuron.

But whether or not it's one neuron or several neurons is not really critical for our purposes. It's a different little bit of the brain. And so you can see that particular recording sites respond more to particular images than others. Just this big complicated set of responses.

So now what we're going to try to ask here is whether or not this neural representation makes information about objects explicit. So remember, the idea is that we're going to think of a particular image now as being represented in a point in-- well, here we're seeing 63 recording sites.

They had about 350 in total. So it's a point in a 350-dimensional space. Different images are different points. And the question is whether we're dealing with an implicit representation-- so this would be an example where the green points correspond to one object and the red points correspond to another object.

And so here they're a little bit intermingled. You can't like lay down a plane and classify these things linearly. Here, you can. These are linearly separable. So there's a plane that puts the green dots on one side and the red dots on the other.

So we call this an explicit representation and this an implicit representation. And of course, it could be much worse than this. The red and the green could be completely intermingled in principle.

So how do we test this? Well, you take the neural responses and you fit a classifier for a particular type of object. And I'm not going to tell you exactly how you fit that. But there's established ways to do that.

And then you can test it on some separate images and see whether it classifies things correctly. So you're trying to find the best hyperplane that's going to divide the red and the green dots for some subset of the images. And then you're going to see whether it gets them correct on some held-out images.

And this is the results of that. So the y-axis here shows classification performance. And the x-axis is the number of sites that the classifier can operate on. So you can ask, how does the ability to classify things vary with the number of sites? And it gets better.

And the point is that-- and there's two curves here, because there's two kinds of classification that can be done. There's categorization, which I think is like whether you have the general class of object correct, so like if it's a monkey versus something else.

And then identification, which is whether you have the specific exemplar within that class correct. And so the identification task is harder, which is why the blue curve is below the yellow curve. But the point is that in both cases, the classification performance from the brain responses just using a linear classifier increases with the number of recording sites.

So this implies that information about objects is made explicit in inferotemporal cortex. And so critically, you can't do this, or at least not as well, in earlier visual areas. And so this is one graph that kind of shows this.

So in this analysis, it's the same kind of experiment that I just showed you from a different paper. So you measure responses to a very large number of objects and a very large number of neurons. But here this was done in inferotemporal cortex as well as V4.

So that's the green curve. And so in addition-- so they measured two things here. One is performance, so how well objectively you can classify the object images. But the other is consistency with human judgments.

So they showed these images to humans and had them classify the images. And humans are good at this task, but they're not perfect. And so you can measure the consistency of these brain classifiers with human judgments.

So on the x-axis is classification performance and on the y-axis is consistency. And so the blue curve here shows the results of this analysis for recordings from inferotemporal cortex. The different dots are different numbers of recording sites.

And so as the number of recording sites increases, the classification performance increases and the consistency with human judgments increases. So the idea is that with a modest number of sites, like a couple hundred sites, you can classify things about as well as people can and with pretty good consistency with human judgments.

By comparison, if you do the exact same thing in an earlier stage of the ventral stream, V4, even with-- I guess the most they analyzed here is 128, and the classification performance is above chance, but not very good. And things are not very consistent with human judgments. And the trajectory of this curve suggests that if you continued to analyze more neurons, it's not going to get dramatically better.

So this again suggests that infratemporal cortex contains information about object identity, but in particular its making object identity explicit in the sense that different images that are of the same type are in a particular part of the representational space, and linearly separated from other kinds of images.

And that enables recognition because with a linear classifier, as you could potentially implement with a downstream neuron, you can tell which kind of object you're dealing with just from the neural responses. So the conclusion is that IT neurons mediate object recognition by making object identity explicit, which again means easily read out, which again means that you can linearly classify it.

So the key concepts here are this idea that sensory systems consist of this cascade of transforms that change the representation, and that those transformations may have either evolved or been learned to make information that matters to us, for instance, about objects in the case of visual recognition explicit, such that things of the same type of object category are kind of in the same part of the representational space.

And so we're thinking of the representation here as this high-dimensional space where the different axes of the responses of different neurons. Questions about that? Yeah.

AUDIENCE: So is learning to tell apart similar things just developing the procedures in order to make them explicit?

JOSH MCDERMOTT: Yeah, so that is a great question. It's like, well, what happens if, say, we invent a new type of object that nobody's ever seen before? And you get a lot of experience with that object such that you're really good at recognizing it. Is the representation changing to make that a lot more explicit?

So that's a very plausible hypothesis. To my knowledge, we do not have very good evidence for that at the moment. So I think a lot of this work that I talked to you about was conducted in the lab of Jim DiCarlo here in the building. I had dinner with him the other night, and we were actually talking about exactly that issue.

And I think his stance on this is that there are not very many clear effects of learning. But I think it's an open issue. So people have tried to look at it a little bit, but hasn't really definitively been looked at. But it's a very plausible hypothesis. Yeah.

Yeah. And so one kind of pretty interesting question that I think is still kind of open is the extent to which-- like, these responses are baked into the visual system. Does evolution give us this set of featural transforms and we're mostly born with it? Is it mostly learned after development?

It's very hard to look at, just because it's kind of hard to do these experiments in very young organisms. So that's kind of an open question. But there's a lot of interest in that right now. Different people disagree about the extent to which these things are innate versus learned.

OK. So when we come back, we'll talk about another wrinkle in all of this, which is pieces of evidence for specialization of the recognition machinery in the brain. In particular, evidence that neurons that are responsive to certain classes of objects seem to be segregated in the brain.

So a lot of the evidence for this in humans comes from our own Nancy Kanwisher via fMRI, who has accumulated a lot of evidence for these different pieces of the brain that are functionally specialized for fairly specific things. And maybe the best example is faces.

So this is a picture that is showing a few well-known brain areas-- the fusiform face area that's very responsive to faces, the parahippocampal place area that's very responsive to pictures of places, the extrastriate body area that responds to bodies, and then MT, which is responsive to motion.

So these are examples of functional specialization. And so that's another kind of interesting property of our visual recognition mechanisms. And so we'll talk a little bit about that when we come back next time.