JAMES DICARLO: So let me start by first-- I already alluded to this, but let's talk about the problem of vision. This is just one computational challenge that our brains solve, but it's one that many of us are very fascinated by. As you'll hear in the rest of the course, there are other problems that are equally fascinating. But I'm going to talk about problems of vision. I'm going to talk about a specific problem of vision, and that's the problem of object recognition. So I will try to operationalize that for you. And one thing you'll see when I talk is that our field, even though we can be motivated by words like vision and object recognition, we're going to only make progress if we start to operationally define things and then decide in what domain models are going to apply. And I think that's an important lesson that I hope will come across in my talk.

So this is the way computer vision operationally defines part of the problem of object recognition and vision. It's as if you take a scene like this and you want to do things like come up with an answer space that looks like this, where you have noun labels, say a car. And you have what are called bounding boxes around the cars, similarly for people, or buildings, or trees, or whatever nouns that you or DARPA or whoever wants to actually label. Right, so this is just one way of operationalizing vision. But I think it gets at the crux of what we're after, which is, there is what's called latent content in this image that all of us instantly bring to our memories, that we can say, aha, that's a car, that's a building. There are nouns that pop into our heads.

We also know other latent information about these things, like the pose of this car, the position of the car, the size of the car. The key point that I'm going to tell you today about this problem is that that information feels to us that it's obvious, but it's quite latent in the image-- that's implicit in the pixel representation. Those of you who have worked on this problem will understand this and those of you who haven't, I hopefully will give you some flavor for what that problem feels like.

So I want to back up a bit. This is more from a cognitive science perspective, or a human brain perspective, to ask, why would we even bother worrying about this problem of object

recognition? And maybe this is obvious that those of you-- and I don't need to say this, but I like to point out that we think of the representations of the tokens of what's out there in the world as being the substrates of what you might do, what's called higher level cognition, things like memory, value judgments, decisions and actions in the world. Imagine building a robot and having it try to act in the world and it doesn't even really know what's out there. So these are the sort of substrate of these kind of cognitive processes.

Again, from an engineering perspective, these are processes or behaviors. This is just a short list of them that might depend on your good abilities to recognize and discriminate among different objects. I think if you look through this list, you could imagine things that would go terribly wrong if you didn't actually do a good job at identifying what's out there in the world. So that's just to think about, again, as an engineer building a robot. This is a slide I stuck in that I want to connect to this course, the idea that I know many of you are from maybe these backgrounds, or from this background. And when I think about the brain, I have this coin here to say, really these are kind of two sides-- we're studying the same coin from two directions here. And really the question that we have to all be excited about, I hope many of you are excited about it is, how does the brain work? And you could do computer science and not care at all about this question. I think it's a little harder to do these and not care about this question. But it's possible, I guess.

So these are all trying to answer this question. And this is maybe pretty obvious, but when you have biological brains that are performing tasks better than current computer systems, machines that humans have built, then the flow tends to want to go this way. You discover phenomena or constraints over here. These lead to ideas that can be built into computer code that can say, hey, can I build a better machine based on what we discover over here? And many of us who came into the field excited to do this and are still excited of this kind of direction. But an equally important direction is that when you have systems that are matched with our abilities, or that can compute some of the things that we think the brain has to compute, then the flow goes more this way, where there's many possible ways to implement an idea and these become falsifiable. That is, that they can be tested against experimental data to ask which of these many ways of implementing a computation are the ones that are actually occurring in the brain. And that's important if you say you want to build brain-machine interfaces, or fix diseases, or do something that's on the level of interacting with the brain directly.

I hope that you guys keep this picture in mind because I think it's sort of the spirit of the course that both of these directions are important. And it's not as if we work on this for 20 years and then work on this for 20 years. It's really the flow across them that I think is the most exciting to us. So just to connect to that, a little bit of history of where was the field on this problem of visual recognition. I don't know if many of you heard this, but here you are at summer school, so there was a Summer Vision Project-- it was called, at MIT. I used to think this story was apocryphal. In 1966, there was a project that the final goal was object identification, which we'll actually name, Objects by Matching the Vocabulary of Known Objects. So this was essentially a summer project to say, we're going to get a couple undergraduate students together and we're going to build a recognition system in 1966. And this was the excitement of AI, we can build anything that we want. And of course, those of you who know this, this problem turned out to be much, much harder than anticipated.

So sometimes problems that seem easy for us are actually quite difficult. If any of you wants this, I would be happy to share this document with you. It's interesting, the space of objects that they describe things like recognizing-- of course, I would say like coffee cups on your desk. But they also say packs of cigarettes on your desk. So this sort of dates the time of this here. So it's a little bit like Mad Men or something. So now, here we are today. And I guess I just can't help but sort of get excited about, here's this really cool machine that's just amazing that does these computations. The things got-- I can't tell you all this because of the 100 billion computing elements, solves problems not solveable by any previous machine. And the thing, it looks crazy, but it only requires 20 watts of power. Those of you who have seen this slide, I'm not talking about this thing. I'm talking about that thing right there. So this is a scale of what we're after. And we often talk about power, but this is something engineers are especially interested in as they build these systems, is how does our brains solve these problems at such a low wattage, so to speak. This is, again, the spirit of many of the things that I hope that you guys are excited about in the future of this field.

Here's another slide that I pulled out that I often like to show is that, from an engineer's point of view, we often try to say, well, we want to build machines that are as good or better than our brain. So machines today, you guys know this, beat us at many things, straight calculation, they beat us at chess. When I was a grad student, they recently won at Jeopardy. In memory, they've always beaten us. Machines are way better at memory than us in the simple form of memory. Seeing, in pattern matching, go to the grocery store, hey, what's that bar code done? I don't know what that was, but it just scans in and somehow it does pattern matching, right?

So there's forms of vision that machines are way better than us.

But some forms of vision that are more complicated that require generalization, like object recognition, or more broadly, scene understanding, we like to think that we are still the winners at. And even things that we take for granted, like walking, this is quite a challenging problem. So engineers really want to move this over here. So our goal is to discover how the brain solves object recognition. And the reason I put this up is, from an engineering point of view, that just doesn't mean write a bunch of papers in a textbook that says, this part of the brain does it, but actually help to implement a system where this is, at least, matched with us and I assume someday, will be better than us.

And this is also a gateway problem. That is, even if it's just this domain, we think that the systems we're studying might generalize to other, for instance, sensory domains. Gabriel told me you were going to do an auditory, visual comparison session later in the week. That's an engineer's point of view, how do I just build better systems? Let's step back and talk from a scientist's point of view. So this is really now to introduce the talk that I'm going to give you today.

So when you're a scientist, what's our job? We say we want to understand. We all write that, understand. What does that mean? Well, what it really means if you boil it down, and I would love to discuss this if you like, is that you have some measurements in some domain. So you can think of this as a state space here. This is like the position of the planets today. And this is like the position of the planets tomorrow. Or you could say, this is the DNA sequence inside a cell. And this is some protein that's going to get made. So you're searching for mappings that are predictive from one domain to another. And we can give lots of examples of what we call successful science, where that's true. This is the core of science is to predict, given some measurements or observations, what's going to happen either in the future or some other set of measurements. So predictive power is the core of all science and the core of understanding. And I think it would be fun if you want to debate that, that you think there's another way. But this is what I come to in thinking about this problem.

And the reason I'm bringing this up is because the accuracy of this predictive mapping is a measure of the strength of any scientific field. And some fields are further along than others. And I would say ours is still not very far along. Our job is to bring it from a nonpredictive state to a very predictive state. And so that means building models that can be falsified and that can predict things. And you'll hear that through my talk. As Gabriel mentioned, what we try to do is

build models that can predict either behavior or neural activity. And that's what we think is what progress looks like.

So now let's translate this to the problem I gave you, which is the problem of vision or more generally object recognition. You could imagine, there's a domain of images. So just to slow down here, just so everybody's on the same page, each dot here might be all the pixels in this image. In this dot, all the pixels in this image. So there's a set of possible pixel-images that you could see. And we imagine that they give rise to, in the brain, some state space. Think of this as the whole brain for now, to just fix ideas, that you could imagine that this image, one you're looking at, it gives rise to some pattern of activity across your whole brain. And this image gives rise to a different pattern of activity across your whole brain. And loosely, we call this the neural representation of this thing.

But then what we do is somehow when we ask you for behavior reports, there's a mapping between that neural state space and what we measure as the output. Whether you say it or write it, you might say, that's a face, these are both faces, if I asked you for nouns among them. OK, so this is another domain of measurement. So now you can see I'm setting up the notion of predictivity. And what we want to do is, we have this complex thing over here of images that somehow map internally into neural activity and then somehow map to the thing we call perceptual reports. And notice I've already put things that we call nouns that we usually associate with objects, cars, face, dogs, cats, clocks, and so forth. OK, so understanding this mapping in a predictive sense is really a summary of what our part of the field is about. And again, accurate predictivity is the core product of the science that underlies our ability to build a system like this-- many of you are interested, to fix a system like this, or to perhaps even augment our own systems. If we want to inject signals here and have them give rise to percepts, we have to know how this works.

A big part of the field of vision is spent-- a lot of the last three decades, working on the mapping between images and neural activity. That's usually called encoding, predictive encoding mechanisms. And it's driven by Hubel and Wiesel's work. The people saw this as a great way forward. It's like, let's go study the neurons and try to understand what in the image is driving them. That is, what's an image computable model in the world that would go from images to neural responses? The other part is that there's some linkage, we think, between the neural activity and these reports. And notice, this is actually why most of us get into neuroscience because you notice this arrow is two-way. This is actually quite deep here. From

an engineer's point of view, you go, well, there's got to be some mapping between the neural activity and the button presses on my fingers or my saying the word noun. There's some causal linkage between this and the things that we observe objectively in a subject. But this is where philosophers debate about like, well, you know in some sense these are sort of two sides of the same coin. We say our own perception, there's some aspects of the internal activity that are the thing that we call awareness or perception. Now I'm not going to get into all that, but I just want to point out that if you're just building models, you can't approach that. It's this sort of strange thing between neurons and these reported states that many of us are fascinated by. So this is called predictive decoding mechanisms. For me, it's all going to be operationalized in terms of reports from humans or animals. And I'll not do that philosophical part, but I thought I'd mention that for those you like to think about those things.

So for visual object perception, I want to point out that, again, the history of the field has been mostly here. This link has been neglected or dominated by weakly predictive word models. That doesn't mean they're not useful starting points, but they're weakly predictive. And so a weakly predictive word model would be-- and for temporal cortex, a part of the brain I'm going to tell you about today, does object recognition. That model has been around for a long time. It is somewhat predictive because it says, you take that out and all object recognition will get destroyed, would be a prediction. Turns out that doesn't actually happen. We can discuss that. But it doesn't tell you how it does it, how to inject signals, which tasks are more or less affected, so that's what I mean by weakly predictive. It's a word model. Face neurons do face task, that's probably true to some extent. But again, it doesn't tell us-- it's more tight. It sort of says, oh, I'll take out these smaller regions and there'll be some set of tasks that involve faces. I don't know, I won't say anything about other tasks. So that's a somewhat more strongly predictive model, but still pretty weakly predictive. And my personal favorite that comes in from reviewers a lot is, attention solves that. So this is just a statement that-- just to be on the lookout for word models that don't actually have content in terms of prediction. I don't know what that means. I read this as, hand of God reaches in and solves the problem. So there's got to be an actual predictive model that can be falsified.

OK, so I don't mean to doubt the importance of these. Before people start giving me a hard time, there are attentional phenomena, there are face neurons, there is an IT, that's what we study. I'm just trying to emphasize for you that we need to go beyond word models into actual testable models that make predictions, that would stand even if the person claiming those models is no longer around, it would make a prediction.

Let me try to define a domain. I said we're going to try to define stuff. It's hard to define stuff. It's big, vision, it's a big area. Object recognition, I sort of said it vaguely. And when I say this, I include faces as an object, a socially important when. You'll hear this from Winrich I think. But I want to say, to try to limit it even further, that's still a big domain. And so we tried early on to reduce the problem even further to something that is more, again, naturalistic, that we think can give us more traction, this predictive sense. So we started by saying, when you take a scene like this and you analyze it, you may not notice it but your ventral stream, really your retina has high acuity in say the central 10 degrees. There's anatomy that I'll show you later that the ventral stream is especially interested in processing the central 10 degrees of information. So that's about two hands at arm's length, for those you see in the room. So you may have the sense that you know what's out there, but you don't really. You kind of stitch that together. And lots of people have shown this, the way you stitch this together is making rapid eye movements around, called saccades, followed by fixations, which are 200 to 500 milliseconds in duration. You don't really see during this time here. It's not as if your brain shuts down, it's just that the movement is too fast for your retina to really keep up with this. So you make these rapid eye movements, you fixate, fixate, fixate. And what you do is, that brings this sort of sampled scene to the central 10 degrees that might look something like this.

So those are 200 millisecond snapshots across that scan path. And I'll play it for you one more time. Now, you should notice that there's one or more objects in each and every image that you probably said, oh, there's a sign. There's a person. There's a car. You might have gotten two out of each one. But you were sort of extracting, at least intuitively to me, at least one or more foreground or central objects when I show you those images. And that ability to do what I just showed you there, we think is the core of how you analyze or build up a scene like this, at least how the ventral stream contributes. And therefore, we call that core recognition, which I defined as a central 10 degrees of visual field, 100 to 200 millisecond viewing duration. And again, it's not all of object recognition, but we think it's a good starting point.

And a way that we probably got into this is because of a rapid serial visual presentation movies from the 70's. Molly Potter showed this really nicely. This is a movie that I've been showing for 15 years now. Notice that this is just a sequence of images where there is typically one or more foreground objects. And you should be quickly mapping those to memory, even though I'm not telling you what to expect. Like Leaning Tower of Pisa, right, I'm not going to tell you that you're going to see Star Wars characters-- well, I just did. But you quickly are able to map those things to some noun or even a more precise subordinate noun. I know this is Yoda. So

our ability to do that, we're very, very good at that. Notice you didn't need a lot of pre-cueing, yet you're still able to do that. And that is really what fascinates us about vision and object recognition in particular. Even without featural attention or pre-cueing, you're able to do a remarkable amount of processing. And I think that's a great demonstration of that.

And just to quantify this for you, because sometimes people say, well you're showing it too short. Your vision system doesn't do much. Here's an eight-way categorization task I'll show you later under range of transformation. These are just the example images of eight different categories of objects. It doesn't really matter what I much do here, you get a very similar curve. And that is, you get most of the performance gain in about the first 100 milliseconds. This is accuracy, you're about 85% correct. This is a challenging task, as I'll show you earlier. It looks easy here, but it's quite challenging. 85% correct, if I let you look at the image longer, up to two seconds, you can bump up to around 90's. So there is some gain with longer viewing duration, but you get-- chance is 50, so you get this huge ability. And we're not the first to show this. This is just to show you in our own kind of task that the data I'm going to tell you about, where we show the image for 100 or 200 milliseconds, this is the typical primate viewing duration that I pin this on. We use this for reasons of efficiency. But you see, the performance is similar across that time. You get a lot done. Your visual system does a lot of work in that first glimpse. And that's core recognition that we are trying to study here. And I know it's not all of object recognition or all of vision, but it's now, we think, a much more defined domain that we can make progress on. And that's what we've been working on. And that's essentially what I'm going to talk about today.

So think of vision, object recognition, within that core recognition. This is David Marr. David and Tommy Poggio, I studied with a long time. And Tommy wrote the introduction to David's-- if you guys haven't read this book, *Vision*-- has anybody, guys know this book? It's really a classic book in our field. It's the first couple chapters that are the part you should really read. That's the best part of the book. And one of the things that you take from this book, that I think David and Tommy helped to lay out a long time ago, is that there is this challenge of level. I think one of the things I take from this is, they would try to define three clean levels. It turns out not to be this clean in practice. But there's one level called computational theory, what's the goal, what's appropriate, what's the logic, and by what strategy can it be carried out. There's another level which is, OK, now once you decide that, how should you represent the data? How can you implement an algorithm to do it? And then there's this actually, how do you run it, how do you build it in hardware?

And neuroscientists often come in, they're like, I'm going to study neurons and it's sort of like jumping into your iPhone and saying, I'm going to study transistors. They often tend to start at the hardware level. And I think that's the biggest lesson you take from this like, oh wait, there's something going on here, these transistors are flying. And you make some story about it if you were recording from the brain or measuring transistors in my iPhone. But I think the important point to take from this is it helps to start thinking about what's the point of the system. What might it be doing? How might you solve that problem? And that leads you then to algorithm. And then you think about representations. So it's sort of a top down approach, rather than just digging into the brain and hoping that the answers will emerge. So I'm going to try to give you that top down approach in this problem that I'm talking about. I've already given you a bit of it by introducing you to the problem. I'll say a little bit more about that and step down a little bit this way. And so this kind of thinking, I think, is important to making progress in how the brain computes things.

So here's a related slide that I made a long time ago that, again, I pulled out for you guys, that I think helps bridge between what I just said about the Marr levels of analysis and whether you're a neuroscientist or cognitive scientist, and are a computer vision or machine learning person. So the first is, what is the problem we're trying to solve? So that's Marr computational level one. So computational vision-- now operationally, you'll hear folks in machine learning, they might say, well, there's some benchmarks, that's good. There's a ImageNet Challenge or whatever challenge they want to solve. Sometimes they'll say, well the brain solves it. That's not good because they didn't really define the problem. Neuroscientists will say, well, it's something like perception or behavior or there's some sort of behavior that they imagined, although characterizing that behavior is not usually their primary goal. But I think there is at least some progress in that regard. Now what does a solution look like? This is really just to talk about language.

So useful image representations for machine learning, like what we might call features-- but neuroscientists will talk about explicit neuronal spiking populations. You heard this in Haim's talk. He was using these words interchangeably. Again, this may be obvious to you guys, but I thought it's worth going through. So this is like Marr level two, representation. How do we instantiate these solutions? So this is still level two algorithms, or mechanisms that actually build useful feature representations. Neuroscientists will think about neuronal wiring and weighting patterns that are actually executing those algorithms. This is what we think is a

bridging language there. And then there's this deeper level that came up in the questions, which is, how would you construct it from the beginning? Learning rules, initial conditions, training images, are words that are used here. There is a learning machine. Here, neuroscientists talk about plasticity, architecture, and experience. But again, those are similar questions just with different language. And I'm doing this because I think the spirit of this course is to try to build these links at all these different levels here.

OK, so hopefully that kind of helps orient you to how we think about it. Let me just go and say, I want to talk about number one. What is a problem we're trying to solve and why is it hard? I said, object recognition is hard and I showed you that MIT Challenge and it was difficult. Maybe it's hard because there's lots of objects. Who thinks that's why it's hard? Who thinks that's not why it's hard? You think computers can list a bunch of objects? It's easy, right? This is what I said about memory, it's a big long list of stuff. Computers are good at that. There's going to be thousands of objects. A list of objects is not a hard thing for a machine to do. What's hard is that each object can produce an essentially infinite number of images. And so you somehow have to be able to take some samples of certain views or poses of an object, this is a car under different poses, and be able to generalize or to predict what the car might look like in another view.

This is what's called the invariance problem. and it's due to the fact that, again, there's identity preserving image variation. This is why the bar code reader in your supermarket works fine, because the code is always laid out very simply. But when you have to be able to generalize across a bunch of conditions, potentially things like background clutter, even more severely occlusion, things you heard from Gabriel, or you may even want to generalize across the class of cars where the cars have slightly different geometry but they're still cars, these kind of generalizations are what make the problem hard. So I'm lumping them all together in what we call the invariance problem.

Many of you in the room know this is the hard problem. And I think that hopefully it fixes ideas of, that's what you should think about. It's not the number of objects, but it's the fact that it has to deal with that invariance problem. Haim was talking about manifolds, and this is my version of that. So this is to introduce you to the problem of, why that invariance problem-- what it looks like or feels like. I'm not going to give you math on how to solve it. It's just a geometric feel for the problem.

So if you imagine you're a camera-- or your retina, which is capturing an image of an object,

let's call this a person, I think I called him Joe. So when you see this image of Joe, and this is the retina, so now this is a state space of what's going on in your retina. So it's a million retinal ganglion cells. Think of them as being an analog value out of each, so this is a million dimensional state space. So when you see this image of Joe, he activates every retinal ganglion cell, some a lot, some a little, but he's some point of that million dimensional space. OK, everybody with me? If everybody's heard all this before and wants me to go on, everybody wave your hand and I'll move on.

AUDIENCE: No, it's good.

JAMES DICARLO: Keep going, OK. So the basic idea is that if Joe undergoes a transformation, like a change in pose, what that does is, it's only a 1 degree of freedom I'm turning under the hood one of those latent variables. If I had a graphics engine, I'm changing the pose of latent variables. It's only one knob that I'm turning, so to speak. And that means there's one line through here as Joe projects across these different images here. And I'm ignoring noise and things. This is just the deterministic mapping onto the retinal ganglion cells. So Joe goes--

[MOVING NOISE]

--and he goes over here. And if I turn the other knob, he goes over here. And so I could imagine, if I turned those two knobs of two axis opposed always possible and plotted this in the million dimensional state space, there'd be this curved up sheet of points, which you could think of Joe's identity manifold over those two degrees of view change. It's only two dimensions, it's hard to start showing more than this. But it's this curved up sheet of points. Everybody with me so far? You don't actually get to see all those. You could imagine a machine actually running them all, but you don't really get to see them. You've got to get samples of them. But there's some underlying manifold structure here.

Now, what's interesting and what's important to point out is that this thing, even though I've drawn it and it's a little curve, but it's highly complicated in this native pixel space. It's all curved up and bending all over the place. And the reason that matters, and this is what Haim introduced you to, is that if you want to be able to separate Joe from another object, say not Joe, another person say, then you need a representation.

I showed you retinal ganglion cells. This is another imaginary state space where you can take simple tools to extract the information. And the simple tools that we like to use are linear classifiers. But you can use other simple tools. Haim used the exact same description to you

guys in his talk, that you have some linear decoder on the state space that can say, oh, they can separate cleanly Joe from not Joe. So these manifolds are nicely separated by a separating hyperplane. That's what these tools tend to do is they like to cut planes. This is one thing they like to do, or they want to find locations or regions, like compact regions in this space, depending on what kind of tool you use. But you don't want the tool having to do all kinds of complicated tracing through this space. That's basically the original problem itself. So what you need is, you have a simple tool box, which we think of as downstream neurons. So a linear classifier, as an approximation, it's like a dot product. It's a weighted sum, which is what we think, neuroscientists, of downstream neurons doing. So it's a weighted sum. And if we want an explicit representation in some neural state space, then we need to be able to take weighted sums of some population representation to be able to separate Joe from not Joe, and Sam from Jill, and everything from everything else that we want to separate.

If we had such a space of neural population, we'd call that a good set of features or an explicit representation of object shape. And for any aficionados here, it's not just cleanly linear separation, it's actually being able to find this with a low number of training examples. So that turns out to be important. But it helps to fix ideas to think about linear separation, ideally with a low number of training examples. So that's a good representation. And notice, I'm starting to mix up terms here. I am assuming, when I talk about shape, that that will map cleanly to identity, or what you might call broadly, category. That's another topic I won't talk about, if you just think about the shape of Joe, or separating one geometry from another.

Now, here's a simulation that my first graduate student, Dave Cox, who's now at Harvard, did. This is a number of years old. This takes these two face objects, render them under changes, and view. And then he actually simulated the manifolds in a 14,000 dimensional space. And then he wanted to visualize it. And because we wanted to try to make the point that these manifolds of these two objects are highly curved and highly tangled, this is a three dimensional view. Remember, it's sitting on a 14,000 dimensional simulation space. You can't view that space. This is a three dimensional view of it. And the point is that it's like two sheets of paper being all crumpled up together and they're not fused. They look fused here because it's in three dimensions. But they're not actually fused. But they're complicated, you can't easily find a separating hyperplane to separate these two objects. We call these tangled object manifolds. And really, they're tangled due to image variation. Remember, if I didn't change those knobs of view or position or scale, there would just be two points in the space and it would be easy. That's the easy problem of listing objects. But if they have to undergo all this

transformation, they become these complicated structures that need to be untangled from each other.

So the problem that's being solved is, you have this retina sampling data, like a camera on the front end, where things look complicated with respect to the latent variables, in this case shape or identity, Sam or Joe. And that they somehow are transformed, as Haim mentioned, they're transformed by some non-linear transformation, some other neural population state space, shown here, where the things look more like this. The latent variable structure is more explicit, that you can easily take things like separating hyperplanes to identify things like shape, which again, roughly corresponds to identity or other latent parameters, like position and scale. You maybe haven't thrown away all these other latent parameters. And if I have time, I'll say something about that so you don't just get identity. But if you can untangle this, you would have a very nice representation with regard to those originally latent parameters. That's the dream of what you'd like to do. It's like reverse graphics, if you will.

So this is what we call an untangled explicit object information. And we think it lives somewhere in the brain, at least to some degree. And I'll show you the evidence for that later on. So what you have then is you have a poor encoding basis, the pixel space. And somewhere in the brain is a powerful encoding basis, a good set of features. And as Haim mentioned, as I already said, this must be a non-linear transformation because the linear transformations are just rotations of that original space. So now let's go down to-- actually this would be Marr level three. Let's go to instantiation. Let's get into the hardware here. We're supposed to be talking about brains. So I'm going to give you a tour of the ventral stream.

So we would love to know how this brain solves it. This is the human brain. This is a non-human primate. This is not shown to scale. This is blown up to show you it's a similar structure, temporal lobe, frontal lobes, occipital lobe. There is a non-human primate. We like this model for a number of reasons. One reason that we like it is that they are very visual creatures, their acuity is very well matched to ours. In fact, even their object recognition abilities are actually quite similar to our own. This may be surprising to you, but let me just show you some data for that. This is actually data from Rishi Rajalingham, in my lab. It says, impressed, but this just came out. This is the confusion matrix patterns of humans trying to discriminate different objects under those transformations that I showed you earlier, where they're not just seeing images, but they have to deal with these invariances. And this is rhesus monkey data trying to do the same thing. And the task goes, I'll give you a test image and then

you get choice images. Was it a car or a dog? I'll show you an image, what choice was it, a dog or a tree? And you're trying to entertain many objects all at once, and you get an image under some unpredictable view and unpredictable background, and then you have to make a choice.

So this is the confusion difficulty. And when you look at this, it's intuitive that these are sort of geometry similar. Camel is confused with dog, and tank is confused with truck, and that's true of both monkeys and humans. And to some level, this shouldn't be surprising to you. The same tasks that are difficult for humans are difficult for monkeys because probably they share very similar processing structures. They don't have to bring in a bunch of knowledge about tanks are driven by people or that, they just have to say, was there a tank or a truck. And under those conditions, they make very similar patterns of confusion. And these patterns are very different from those that you get when you run classifiers on pixels or low level visual simulations. But they're very similar to each other, in fact, are statistically indistinguishable, monkeys and humans, on these kind of patterns of confusion.

OK, so that's one reason we like this subject, the monkey model, is that the behavior is very well matched to the humans. The other reason is that we know from a lot of previous work that I alluded to, that some studies have shown that lesions in these parts of the brain can lead to deficits in recognition task. So again, we think the ventral stream solves recognition. So we know a weak word model of where to look, we just don't know exactly what's going on there. Just to orient you, these ventral areas, V1, V2, V4, and infer temporal cortex, or IT cortex-- IT projects anatomically to the frontal lobe to regions involved in decision and action, and around the bend to the medial temporal lobe to regions involved in formation of long-term memory. Because these are monkeys and not humans, and Gabriel mentioned this in his talk, we can go in and we can record from their brains, and we can perturb neural activity in their brains directly. And we can do that in a systematic way. This is the advantage of an animal model as opposed to a human model.

OK, as neuroscientists now, we've taken a problem, translated it to behavior, taken that behavior into a species we can study, we know roughly where to look, and now we want to try to understand what's going on. So as engineers, we take these curled up sheets of cortex and think of them as I've already been showing you, as populations of neurons. So there's millions of neurons on each of these sheets. I'll give you numbers on a slide coming up. There's some sort of processing that may be common here, I put these T's in, there might be some common

cortical algorithm processing forward this way. There's also inter-cortical processing. And there's also some feedback processing going on in here. So all that's schematically illustrated in this slide that I'll keep bringing up here when we talk about these different levels of the ventral stream. Now I'm most going to be talking about IT cortex here at the end.

Why do we call these different areas? One reason is that there's a complete retina topic map, a map of the whole visual space in each of these different levels. In retina, there's one. In LGN-- in the thalamus, there's another. In V1, there's another map. In V2, there's another map. In V4, there's another map. In IT, it's less clear that it's retinotopic, we're not even sure that IT is one area. Maybe we'll have time, I'll say more about that detail. So it's not that retinotopic in IT, except the most posterior parts of IT. But that's why neuroscientists divide these into different areas.

So a key concept, though, for you computationally is, think of each of these as a population representation that's retransforming the data from that complicated space to some nicer space. And it's doing this probably in a stepwise, gradual manner. So IT is believed to be that powerful encoding basis that I alluded to earlier, where you have these nice flattened object manifolds. And I'll show you the evidence for that.

This is recently from a review I did that gives more numbers on these things. And I've sized the areas according to their relative cortical area in the monkey. Here's V1, V2, V4, IT. IT is a complex of areas. And I'm showing you these latencies. These are the average latencies in these different visual areas. You can see, it's about 50 milliseconds from when an image hits the retina until you get activity in V1. 60 in V2, 70-- there's about a 10 millisecond step across these different areas. So it's about 100 millisecond lag between an image it's here, and you start to see changes in activity at this level up here that I'm referring to. When I say IT, I'm referring to AIT and CIT together. That's my usage of the word IT for the aficionados in the room. And that's about 10 million output neurons in IT just to fix numbers. In V1 here, you have like 37 million output neurons. There's about 200 million neurons in V1, similar in V2. And many of you probably heard about other parts of the visual system. Here's MT, many of you probably heard about MT. So you can see it's tiny compared to some of these areas that I'm talking about here.

I'm going to show you some neural dam-- I'm just going to give you a brief tour of these different areas, so brief, it's almost cartoonish. But at least those of you who haven't seen this should at least be exposed. So in the retina-- you guys know in the retina there's a bunch of

cell layers in the retina. The retina is a complicated device. I think of it as a beautiful camera. So you're down in the retina. To me, the key thing in the retina is in the end you've got some cells that are going to project back along the optic nerve. So these are the retinal ganglion cells, they actually live on the surface. The light comes through, photo receptors are here, there is processing in these intermediate layers, and then there's a bunch of retinal ganglion cell types. There's thought to be about 20 types or so.

The original physiology, there are two functional central types where they have on center or off center. Let's take an on center cell, you shine light in the middle of a spot-- now this is a tiny little spot on the retina, the size depends on where you are in the visual field. But you shine a little bit of light in the center, the response goes up. See the spike rate going up here. Put light in the surround, the response rate goes down. So it has an on center, off surround profile. And then there's a flip type here. So that's the basic functional type. When you think about the retina, it is tiled with all of these point detectors that have some nice center surround effects. There's some nice gain control for overall illumination conditions. But my toy model of the retina, it's basically a really nice pixel map coming back down the optic track to the LGN.

OK, I'm going to skip the LGN and go straight to V1. People have known for a long time, functionally V1 cells they have sensitivity to especially edges. They have what's called orientation selectivity. Hopefully this isn't new to you guys. Here's a simple cell in V1. If you shine a bar of a light on it inside it's receptive field-- does everyone know what a receptive field is? I don't want to go-- OK. It's OK if you ask, because I want to make sure you guys are OK.

So the receptive field, you shine a bar light in it, turn it on in the right orientation, gives good response out of the cell. Move it off this position, now not much response, there's a little bit of an off response here. Change the orientation, nothing happens. Full field illumination, nothing happens. OK, so this is called selectivity. That is, there's some portion of the image space that it cares about. It doesn't just respond to any light at that spot like the pixel wise, retinal ganglion cell would.

So now there's this complex cell that's also in V1, which maintains this orientation selectivity across a change in position, as shown here, also across some changes in scale. So it maintains it, meaning that you have this tolerance-- so that's called position tolerance, for position. You can move the bar around it, still likes that oriented bar. But you change its angle and it goes down, so it still maintains the same selectivity here but it has some tolerance. So you get this build up of some orientation sensitivity followed by some tolerance.

And there are models from Hubel and Wiesel that they thought that you could build this first and then you build these out of these, that's the simple version. And here they are. These are the Hubel and Wiesel models, how you build these and like operators to build selectivity from pixel-wise cells with an and like operator lining these up correctly. You can imagine oriented tuned cells built this way. There's evidence for this in physiology that this is how these are constructed. The tolerance of these complex cells is thought to build by a combination of simple cells. And there's some evidence for this. And this is again, all the way from Hubel and Wiesel, who won a Nobel Prize for this and related work in the 1960s. And then there were a bunch of computational models that are really inspired by this and I think are still the core models of how the system works. And some of the original ones that were written down are Fukushima in the '80s, and then Tommy Poggio and others built what's called an HMAX Model, you guys have probably heard about, that's off of these similar ideas, much more refined and much more matched to the neural data. But I'm just try to point out that these kind of physiological observations are what inspired this class of largely feedforward models that you heard about much today.

So that's a brief tour of V1. Now, what's going on in V2? For a long time, people thought it was hard to tell the difference from V1 and V2. And I just thought I'd show you guys, this is a slide I stuck in, this is from Eero Simoncelli and Tony Movshon. And I think you guys have Eero teaching in the course a bit later, so he may say some of this. But V2 cells have some sensitivity to natural image statistics that V1 cells don't. And maybe I'll see if I can take you through this. So the way that they did this is you can simulate-- so this is all driven off of work that Eero and Tony have done-- especially Eero has done on texture synthesis. So you have these original images, and if you run them through a bunch of V1-like filter banks, and then you take a new image, a random seed, which is like white noise, and you try to make sure that it would activate populations of V1 cells in a similar way, there's a large set of images that would do that because you're just doing summary statistics, but these are some examples of them. For this image, this is one that one might look like. So you can see, to you, it doesn't look the same as this. But to V1, these are metamers, they're very similar in the summary statistics in V1. And then you start taking cross products of these V1 summary statistics and then you try to match those. And what's interesting is you start to get something that looks, texture wise, much more like this original image. And this is a big part of what Eero and others did in that work. And the reason I'm showing you this is that Tony's lab has gone and recorded in V1 and V2 with these kinds of stimuli, and the main observation they have is that V1 doesn't

care whether you show it this or this.

To V1, these are both the same, which says we have the summary statistics for V1 right in terms of the average V1 response. That's all I'm showing you here. The paper, if you want it, is much more detailed. But you go to V2 and there's a big difference between this, which V2 cells respond to more, and this, which they respond to less. And really one inference you can take from this is that V2 neurons apply a repeated-- another and like operator on V1. That's a simple inference that these kinds of data seem to support . And they also tell you that these and-like operators, these conjunctions of V1 statistics tend to be in the direction of the statistics of the natural world, that's naturalistic statistics. Now lots of controls haven't been done here to narrow in exactly what kinds ands, but that's the spirit of where the field is in trying to understand V2. Everybody thinks it has something to do with corners or a more complicated structure. But this is a way that current in the field to try to move these image computing models forward in V1 and V2. And Tony likes to point out that this is one of the strongest differences that you see between V1 and V2, other than the receptive field sizes. So I think that's quite some exciting work if you don't know about it on V2.

OK, then you get up into V4 and things get much murkier. So what's going on in V4? Well, let me just briefly say that one of my post-docs-- this is more recent work just because it builds on that earlier work. This is Nicole Rust, when she was a post-doc in the lab, compared V4. She actually compared it to IT. I'll skip that. But she was using these Simoncelli scrambled images. These are actually the texture images from-- these are the original images and these are the texture versions. So this should look like a textured version of that. You can see that these algorithms don't actually capture the object content of these images. And what Nicole actually showed is that similar to what you just saw there, in the earlier work like V1, V4 doesn't care about the differences between these. It responds similarly, as a population, to this and this, and this and this, and this and this. But IT cares a lot about this versus this. So this is just repeating the same theme, the general idea that you have and -like operators that we think are aligned along the ventral stream that are tuned to the kind of statistics that you tend to encounter in the world. And this is some of the evidence for it in V2, and then later in V4, and IT, and Nicole's work, if you piece that all together.

When you go to a place like V4, remember V4 is now like three levels up. And what does V4 do? Look, this is Jack's work in 1996. This is from Jack Gallant when he was working with David Van Essen. And people had some ideas that maybe there are these certain functions

that V4 neurons like, and they would show these-- the same thing people have done in V2, they would show a bunch of images like this and figure out, well, does it like these Cartesian gratings or these curved ones. And you know what, you get out of this is, you could tell some story about it, but you get a bunch of responses out of it. The color indicates the response. And you kind of look at it, and people would tell some stories, but it really was just kind of like tea leaves. Here's a bunch of data, we don't really know what these V4 neurons were doing. This was a science paper, so you could go back and read it.

And then Ed Connor and Anitha Pasupathy worked together a few years after that to try to figure out more about what V4 neurons do. And they did things like take images like this, which were isolated, and try to cut them into parts, like curved parts, pointy parts, curved, concave, convex. And this was motivated off of some psychology literature. And they would define these based on the center of the object. So this wasn't an image computable model, it was just a basis set that they built around these silhouette objects. And so they made this basis set about any kind of silhouetted object they like here. They hypothesized that they could fit the responses of V4 neurons in this basis set. And this was their attempt to do it. They could actually fit quite well. And that's kind of what's being shown here.

Here's the response of a V4 neuron. The color indicates the depth of the response. You can see, this is sort of like that previous slide, you're looking at tea leaves. It looks complicated, but under this model they were able to, in the shape space, explain about half of the response variants of V4 neurons. The upshot is, that V4 curve is about some combination of curves. And then later, Scott Brincat, with Ed, went on into posterior IT and showed that maybe some combinations of these V4 cells could fit posterior IT responses quite well. So if you read the literature in V4 and IT, you'll come across these studies. And they are important ones to look at. Unfortunately, they don't give you an image computable model of what these neurons are doing. But it's some of the work that you should know about if you want to look in V4 or early IT, so I'm telling it to you.

So let me go on to IT, which is what I want to talk about for the rest of today. Again, I'm talking about AIT and CIT. And I'll just quickly say that the anatomy, again, suggests that the IT is the central 10 degrees. And even though V1, V2, and V4 cover the whole visual field, if you make injections in V4, that's shown here, where you make injections in the more peripheral parts of the V4 representation, which is up here, that you don't get much projection into IT, which is here. You don't see much green color, whereas, you make projections in the center part of V4,

these red sites here, you see much more coverage into IT, which is shown here.

So when I say 10 degrees, that's rough. Everything in biology is messy. But this is some of the evidence, beyond recordings, there's anatomical evidence that as you go down into IT, you are more and more focused on the central 10 degrees. OK, let me talk about a little bit of the history of IT recordings. This is when people got excited about IT, in the 70s. This is work by Charlie Gross, who's one of the first people to record an IT cortex. And I'll show you what they did here. This was in an era where, remember, Hubel and Wiesel had just done their work in the '60s. And they recorded from the cat visual cortex. And they had found these edge cells, and they ended up winning the Nobel Prize for that. So it was the heyday of like, let's record and figure out what makes cells go. So they were brave enough to put an electrode down an IT cortex in 1970 and said, what makes this neuron go. Remember, that's an encoding question, what's the image content that will drive this neuron. And it's fun to just look back on this and what they were doing. So they didn't have computer monitors. They were actually waving around stimuli in front of the animals. This is an anesthetized animal on a table. This is a monkey. Actually, they started with a cat and then they later went to monkey. The use of these stimuli was begun one day when, having failed to drive a unit with any light stimulus-- that probably means spots of light, edges things that Hubel and Wiesel had been using. We waved a hand at the stimulus screen, they waved in front of the monkey, and elicited a very vigorous response from the previously unresponsive neuron.

And then we spent the next 12 hours-- so the animal's anesthetized on the table, their recording from this neuron. It's 12 hours because nothing's moving, so you can record for a long period of time. So singular neuron, they're recording, listening to the spikes. We spent the next 12 hours testing various paper cut outs in attempt to find the trigger feature. You can see, that's a Hubel and Wiesel idea, what makes this neuron go. What's the best thing, that's become a lot of what the field spent time doing. Trigger feature for this unit, when the entire stimulus set were 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, that means psychophysical judged, to the shadow of a monkey hand. So these are their rank order of the stimuli. And they say look, it looks like it's some sort of hand neuron. That's all I know how to describe it. I can't find some simple thing on here.

So this kind of study then launched a whole domain where people started to go in to record

these neurons and they found interesting different types. Bob Desimone, who worked with Charlie Gross, later showed much more nicely under more controlled conditions, yes, there are indeed neurons that respond. You can see more to these hand-- this is the post stimulus time histogram, lots of spikes, lots of spikes, lots of spikes-- respond more to these hands than to these other kind of stimuli here. So you could say, these neurons have tuned to specific combinations of high selectivity.

You'll hear from Winrich that others had shown that you could record some of the neurons are really like faces that you could find, and not so much hands. So you could find neurons that seem to have some interesting selectivity in IT cortex. And then others later went on to show in a number of studies-- this is from Nico Logothetis' work of a number of years later. It's just one example that this selectivity had some tolerance to, say, the position of the stimulus, that's what's shown here. The fact that these bars are high just means that it tolerates movement in where the-- sorry, this is size, degrees of visual angle. This is position, moving the stimulus around. So this was known for a number of years that there's some tolerance to position and size changes at least.

OK, so I'm putting these up and you say, there's some selectivity and there's some tolerance. And that should remind you of what we already said in V1, there's some selectivity, simple cells. There's some tolerance, complex cells. So you have the same themes here, just different kinds of types of stimuli being used. Then people really went on, in the 80s especially, and said, let's go after this trigger feature. And Tanaka's group really went after this really hard. Tanaka's group would find the best stimulus they would find, dangle a bunch of objects in front of a recorded neuron, find the best out of a whole set of objects, and then they try to do a reduction. They'd try to figure out, how can I reduce this. This is their attempt to reduce the stimulus to its features without lowering the neural response. So high response, high response, high response, high response, high response, suddenly I do this, the response drops. I do this, the response drops. And they have lots of examples of this. And they want you to try to get to the simplest thing that could capture the response.

And when they did this, they would take stimuli like this, and end up with stimuli that looked like that. Now, many of you should probably start to wonder here, there's lots of paths for stimulus space. It's not clear that these are elemental in any way. There's lots of ways that you can show with modeling that you can get easily lost in this space of navigating around here. This is just, again, a history of the work. This is the kind of things that people were doing. And then

from that, they presented what we think of as the ice cube model of IT, that I think is actually still a very reasonable approximation. They not only showed that neurons tended to like certain relatively reduced stimulus features, not full objects, but that they are gathered together. So these are millimeter scale regions of IT that nearby neurons, within a millimeter or so, have similar preferences. They're not just scattered willy-nilly throughout the tissue. When you go record nearby neurons, they're similar. So there's some mapping within IT cortex. This is schematic here. This is optical imaging data of IT cortex also from Tanaka's group that show you that these different blobs of tissue get activated by different images shown here. And I'm just showing you the scale of this, it's around a little less than a millimeter.

And our lab has evidence of this too. So there's some sort of spatial organization in IT, but we really don't really yet understand the features, these elemental features yet, or at least, not at this time. Then later, there's lots of beautiful work in IT. Again, I'm probably not telling you all of it. Some of the most exciting work recently-- and you'll hear about this from Winrich, that people started to use fMRIs. So Doris Tsao and Winrich Freiwald and Marge Livingstone all together started to use fMRI data to compare faces versus objects. This was motivated from human work, by work like Nancy Kanwisher lab and others. What they found was that in monkeys, you could find different parts that would show up, what are called face patches, where you have a relative preference for faces over objects. Again, I don't want to take all of Winrich's talk here, but you have these different patches here. And then what's really cool is, you go in and record from these patches and then you find a very enriched locations for face neurons. And these enriched locations were known from a number of other studies. But this is a nice correlation between functional imaging and this enrichment of these face cells. And that's what's shown here, that these neurons respond mostly to faces and not so much other objects. Although, you see they still sort of respond to these. So this kind of says fMRI and physiology are telling you similar things. It also tells you there's some spatial clumping, at least for face-like objects, at a scale of a few millimeters or so, the size of these patches. OK, so that's larger scale organization.

This is data from our own lab that shows the same thing. Maybe I'll just skip through this in the interest of time-- that we can map and record the neurons very precisely, map them spatially and compare that with fMRI. So this is just a larger field of view maps of the same idea. So what we have then, just to wrap up this whirlwind tour of the ventral stream, is that we had some untangled explicit information. And what I want to try to convince you of now, is that-- I've told you about the ventral stream, but I'm going to try to tell you that, in IT cortex, this is a

powerful representation for encoding object information. And then we'll take a break because we've already probably been going a while. Yeah, about 10 more minutes and then we'll take a break. So what I've told you is, I've led you up the ventral stream, I've given you a bit of the history, so now let's talk about IT more precisely.

So now this is work from my own lab. You go in and record IT. You go record extracellularly. You travel down into IT cortex, which is down here. And you record from this. And similar to what you saw, another version of what you saw from Charlie Gross or Bob Desimone, you show a bunch of images. And they could be arbitrary images. You take an IT recording site, and see these little dots, those are action potential spikes out of a particular IT site. And these are repeatable. You have some Poisson variability here. But you see that there's more spikes here, there's little more here, less here, less there. These images are all randomly interleaved when you collect the data, as I'll show you in a minute. And you go to different sites and it likes different images. So there is certainly some image selectivity. This should not be surprising because I already showed you this from previous work. This is just data from our own lab.

You can also see now that you are looking closely at the time lag, remember, I said around 100 milliseconds stimulus on. Stimulus off, the stimulus is actually off before the spikes actually start to occur out here in IT because, again, there's a long time lag, 100 milliseconds. OK, so that's what the neural responses look like. I don't know if you guys can hear this, maybe I should have hooked up audio. Maybe you might be able to hear-- this is actually a recording that Chou Hung did when he collected his data in my lab for the early studies we did in the lab. I don't know if you guys can hear.

[STATIC]

[BEEP]

[BEEP]

[BEEP]

Those high beeps are the animal getting reward for fixating on that dot. You're not even going to be able to parse that. I mean, you hear the spikes clicking by, those--

[STATIC]

Those are action potentials. And I don't expect you to look at anything like, oh, it's a face

neuron, or whatever. I just want you to get a feel for how those data were originally collected. This is a pretty grainy video. But you get the idea. You collect data like that. And again, you can find selectivity in those population patterns, as I just showed you. But then, Gabriel and Tommy and I, so the three of us, I think all in this room, way back when in 2005 said, well look, the population of IT might have good, useful information for solving this difficult object manifold tangling problem. It might be a good explicit representation. So we did a, what I call, early test of this idea. We took this simple image set from eight different categories that we had chosen. And there's good stories of why we chose those objects, if you like to hear them. But let me just say, simple objects, we moved them across position and scale, and we collected the responses of IT of a bunch of sites to changes to all these different visual images.

And we showed them as I just showed you. We just showed them for 100 milliseconds. This is this core recognition regime, were just showing them for 100 milliseconds. And then we show another one, and they're just randomly interleaved. And from this, what you do is you could get a population set of data where we recorded 350 IT sites. Here's a sample of 63 sites. This is 78 images, the mean neural response here is the mean response to an image. This is 78 of the images we showed. There's nothing for you to read into here to say, other than, you have this rich population data. And now our question is, well, what lives in this population data that we've collected. Is it explicit with regard to categories? So we come back to what I showed you earlier about those tangled manifolds and said, we need simple decoding tools. Can a simple decoding tool look at that population and tell me what's out there? And again, we were using linear classifiers at the time, because we took that, as you heard from Haim as our operational definition of what a simple tool is. And if it could decode information about the object identity, then we'd say, well, that means, by that operational definition, this is explicit, available, accessible information, or just generally good.

So if you imagine that the activity-- this is schematic. Each dot, this is neuron one, neuron two, and you could have a bunch of IT neurons. But if you can separate any object from all the other object, these points represent the population response to each image of an object. Remember, there's many images of each object. But if you could linearly separate that, that would mean it was explicit. And if you had a hard time separating it, this would be implicit. These are like tangled object manifold. This is Inaccessible, or bad, information. So we just-- we, and when I mean we, I mean Chou Hung, who led the study. Gabriel, Tommy, and I did this. We took the response of an image, like this one. It produced a population vector. Again, we recorded a bunch of neurons. We recorded them sequentially and then pieced together

this population vector. So these are the spikes simulated off a population of IT. We could do various things. In fact, I think Gabriel did everything possible, as I remember at the time. And one of the things we did was just count spikes. One of the simple things, that turns out to work quite well, is count the spikes over 100 milliseconds. So this neuron counts spikes. That gives you a number, one number here, count spikes get one number. So you have n neurons, you get n numbers. So it's a point in a n dimensional state space where n is the number of neurons. And then we had already pre-divided the images into different categories, as shown here.

These are the categories. And again, we just asked how well you could do faces versus non-faces, toys versus non-toys, so on and so forth. These are old slides. But you get the idea, is that basically, you don't need that many sites to already get to very high levels of performance on both categorization and identification. The interesting thing about this was that you could solve simple forms of this invariance problem in this representation quite easily. That if you just trained on the central objects, the center and size, the simple three degree size center position, and test it on the same thing, just held out repeats of this data, you did quite well. That's a baseline. But what's interesting is you test at different position and scale. And then you also do almost nearly as well. So you naturally generalize to these other conditions by training on these simple conditions.

So this is evidence that the population is a good basis set for solving these kind of problems. A few number of training examples on this population then generalizes, well, across conditions makes the problem hard. So again, we published that a long time ago. This was an early step to say, look, the phenomenology looks right for the story that I've been telling you so far. You can't do this easily in earlier visual areas like V1, or simulated V1 or V4. And we later show that a number of ways.

This is consistent with work I was showing you with Logothetis position tolerance, size tolerance, the selectivity. It's really just an explicit test of the idea population encoding. So the take home here is that there's this explicit object representation in IT. I didn't prove to you that this is the link, this predictive model to decoding yet. We're going to talk about that next. But this was some of the important population phenomenology that we did. What I try to tell you today-- hopefully I've introduced you to the problem of visual object recognition and the way we restricted it to core object recognition. We talked a lot about predictive models as being the goal, although I haven't presented much to you yet. Hopefully, that's the second part of the

talk. I've given you a tour of the ventral stream. But it was a poor tour. I'm sure everybody i work with would say that you've neglected all this work because there's no way I can do that all in even a whole week. I just tried to hit some of the highlights for you.

And I told you that the IT population seems to have solved a key problem, this sort of invariance problem that I set up. And one way to step back and say, over the last 40 years or so, from those early studies of Charlie Gross or even Hubel and Wiesel, we, the field of ventral stream physiology, we've largely described important phenomenology. Even that last study is population phenomenology. And so now we need these more advanced models. So the next phase of the field is developing and testing these predictive models that I've motivated at the beginning, but I haven't given you much of yet. So this was hopefully a bit of history and set context to where we are.