JAMES DICARLO: I'm going to shift more towards this decoding space than we talked about, the linkage between neural activity and behavioral report. And I introduced that a bit. You just saw that there's some population powerful activity in IT. And I'm going to expand on that a bit here. But sort of stepping back, when you think about it again, what I call an end to end understanding, going from the image all the way to neural activity to the perceptual report, one of the things we want to do, again, is just define a decoding mechanism that the brain uses to support these perceptual reports. Basically what neural activity are directly responsible for these tasks?

And I'll come back to later this encoding side. It's like, you know, and notice I'm putting these in this order, right? So once you know what the relevant aspects of neural activity are in IT, or wherever you think they are, then that sets a target for what is the image to neural transformation that you're trying to explain? Not predict any neural response, but those particular aspects of the neural response. So that's what I mean by the relevant ventral stream patterns of activity.

So we start here. We work to here, and then we work to here, rather than the other way around. OK, so I'm going to try it again. Keep with the domain I set up. I talked about core recognition. I now need to start to define tasks.

I'm going to talk about specific tasks that are, for now, let's call them basic level nouns. I'm actually going to relax that to subordinate tasks in a minute. But here they are. Car, clock, cat. These are not the actual nouns. I'll show you the ones we use. But just to fix ideas, we're imagining a space of all possible nouns that you might use to describe what you just saw. And I'm going to have a generative image domain. So I now have a space of images here. I'm not just going to draw these off the web. We're going to generate our own image domain that we think engages on the problem, but gives us control of the latent variables.

So I'll show you that now. So the way we're going to do this is by generating one foreground object in each image that we're going to show. And we just did this by taking 3-D models like

these-- this is a model of a car. We can control it's other latent variables beyond its identity. So this is a car. It has a particular car type. So there's a couple of latent variables about identity here that relate to the geometry.

Then there's these position-- other latent variables like position, size, and pose that I mentioned, that are unknowns that make the problem challenging. And we can then just, like, render this thing. And we could place it on any old background we wanted to. And what we did was we tended to place them on uncorrelated naturalistic backgrounds. And that creates these sort of weirdish looking images. Some of them may look sort of natural, hence, this looks pretty unnatural.

But the reason we did this. Why would you do this? So-- so we did this because we could add a generative space. And because it was-- so we know what's going on with the latent variables we care about. And we also, when we built this, it was challenging for computer vision systems to deal with this, even though humans could naturally-- you know, they don't have advantage of any contextual cues here because by construction, these are uncorrelated. We just took natural images and would randomly put objects on them.

But this was enough to fool a lot of the computer vision systems at the time that tended to rely on the contextual cues. Like blue in the background signals or being an airplane, we didn't want those kind of things being done. We wanted the actual extraction of object identity. And again, humans could do it quite well. So that's why we ended up in this sort of maybe this no man's land of image space, which is not very simple, but not ImageNet just pulled off off the web.

And so that's how we got there. And just to give you a sense that this is actually quite doable for humans, I'll show you a few images. I won't even cue you what they are. I'm going to show them for 100 milliseconds. You can kind of shout out what object you see.

AUDIENCE: Car.

AUDIENCE: [INAUDIBLE]

JAMES DICARIO: Right. So see, it's pretty straightforward, right? And those look weird, you can do that quite well. And you know, here's the kind of images that we would generate. This would be-- so when we think of image bags, we think of partitions of image space. This is some images that would correspond to faces. These are all images of faces under some transformations. Again,

different backgrounds. These are not faces. These are other objects again, under transformations.

And we can have as many of these as we want. We call this one-- this distinction, when shown for 100 milliseconds-- is one core recognition test. Discriminate face for not face. Here is a subordinate task. This is beetle from not beetle. This is a particular type of car. You can see it's more challenging. Again, we don't show these images like this. This is just to show you the set. We show them one at a time.

And so let me now go ahead and say, we're going to try to make a predictive model using that kind of image space to see if we can understand what are the relevant aspects of neural activity that can predict human report on an image space? And when I say we, I mean Naiib Maiai and Ha Hong, who are post-doc and graduate student that were in the lab that led this experimental work. And Ethan Soloman and Dan Yamins also contributed to the work.

So what we did was to try to record a bunch of IT activity to measure what's going on in the population as I showed you earlier, but now in this more defined space where we're going to collect a bunch of human behavior to compare possible ways of reading IT with the behavior of the human. This is how we started. We're now doing monkeys-- where we're recording and the monkey's doing a task. But what we did here was we just passively fixating monkeys, compared with behaving humans. And as I showed you earlier, monkeys and humans have very similar patterns of behavior.

So what we record from IT, in this case, we were using array recording electrodes. These are chronically implanted. This shows them here. You implant them during a surgery, as kind of is shown here. Down in the IT cortex. You can get their size here. There are about hundred-- there's actually 96 electrodes on each of them. They typically yield about half of the electrodes having active neurons on them. So you get, you know, on the order of 150 recording sites. And you can lay them out. You can lay-- we would typically lay out three of them across IT and V4 to record a population sample out of IT.

And we would do this across among multiple monkeys. And here's an example of the kind of data we would get. This is 168 IT recording sites. This is similar to what I showed you earlier. This is the mean response in a particular time window out of IT, similar to what I showed you earlier in that study with Gabriel. And what we do here is, I'm just showing you to give you feel. That's one image. Here is eight more-- here's seven more images. And these are just the

population vectors in a graphic form. And but we actually collected nearly 25-- this is 2,560 images. This is sort of the mean response data of this 168 neurons. And now you have this again, this rich population data. And you can ask, what's available in there to support these tasks? And how well does it predict human patterns of performance on those tasks?

So in this study, that's all we were asking to do. We're trying to do more and more recently. But let me show you what all we were trying to do is to say, look. One thing we observed, even though you saw that car-- you could do car, you could do faces. It seemed like you were doing 100%. Turns out you're better at some things than others. So discriminate-- this is a deep prime map of humans. So red means good performance. High D prime. You know, a D prime of 3 is something like-- I don't know, psychophysicists in the room may correct me. A D prime of 3 is sort of on the order of 90 some 95% correct, in that range. So these are very high performance levels when you get up to 5. 0 is chance. So 50%-- well this is an eight way task. So one over 8% correct.

So the subjects were doing either eight way basic level tasks, or eight way subordinate cars, or eight way faces. And these are the D prime levels under different amounts of variation of those other latent variables position size and pose. Don't worry about those details. What I want you to see is the color here. So look, it's tables versus-- discriminating tables from all these other objects. You do that at a very high D prime. Discriminating beetles from other cars, you do it at slightly lower D prime.

You can see this, specially at a high variation, you're actually starting to get down to lower performance. And faces-- one face versus another face, you're actually quite poor at that. You're a little bit better than chance. But it's actually quite challenging in 100 milliseconds without hair and glasses to discriminate those 3-D kind of face models. I showed you Sam and Joe earlier as examples. You're actually quite challenging to do that for humans in that domain of faces.

So, what I want to show you here is you have this pattern of behavioral performance. You have all this IT activity. This is humans. This is monkeys. And what we wanted to do is say, look. We can use this pattern. This is very repeatable across humans. Can we use this repeatable behavioral pattern to understand what aspects of this activity could map to that? And again, this pattern is reliable. I just said that. And it's not as if you can predict this pattern by just running classifiers on pixels or V1. In fact, I'll show you that a minute.

But we thought there's some aspects of IT activity that would predict this. And we wanted to try to find those aspects to-- so, again, this was motivated by that study I showed you earlier. So which part of the IT population activity could predict this behavior over all recognition tasks? We're seeking a general decoding model that would work. Here's some specific tasks. But we'd like it to be-- work over any task that we could imagine testing humans within this domain of taking 3D models, putting them under variation. Work over that entire domain. That was what we were hoping to do.

So again, I'll briefly take you through this. Because I already showed you this earlier. Again, we've previously shown that you could kind of take this kind of state space, and say hey, can you separate images of faces from non-faces, using these simple linear classifiers, which are essentially weighted sums on the IT activity? And now we wanted to ask, could this predict human behavioral face performance, and monkey, because again, they're very similar.

And not only would this class of decoding models that was motivated by the earlier work predict this task, but would predict car detection? Would the same model predict car one versus car two? That's a subordinate task. And all such tasks. Again, over the whole domain, can you take a same decoding strategy and take the data and say, I'm going to just learn on a certain number of training examples, build a classifier, and then I'll say that's my model of how the human does every one of these tasks. And if that's true, then it should perfectly predict that pattern or performance that I just showed you earlier.

And so here was again, this was the working hypothesis. Passively evoked spike rates using single fixed time scale that are spatially distributed, because they're sampled over IT, over a single fixed number of non-human-- of non-human primate cortex. So a single number of neurons. And learn from a reasonable number of training examples. So all of that is a decoding class of models that we thought might work. And if this is correct-- this is what I just said-- it should predict the behavioral data that we collect. For example, the D prime data I just showed you. But also more fine grained behavioral data in principle.

So I want to just step back to make it clear that it's not obvious that this should work, right? I mean, it depends-- in the audience, I get people on completely different sides of this, whether this should work or not. So, you know, one thing is, like, well look, it's passively evoked. You heard Gabriel say, well, you didn't like passive tasks. And I agree with that. In the ideal world, the animal will be actively doing the task.

And then you'd say, well I'll measure while the animal's doing the task. That's going to be your best chance of prediction. But we also saw earlier that that passively evoked monkey still-- you know, nobody would argue that a passively evoked retinal data is not going to be somewhat applicable to vision. And you know, the question is, how much of those arousal effects show up in a place like IT cortex, which is typically high? Which is high up in the ventral stream.

So you could argue both sides of this. But it's possible that attentional arousal mechanisms are needed to make this a good predictive linkage between that to sort of activate IT in this sort of crude way, if you like.

Some people have pointed out that you need the trial by trial coordinated spike timing structure to actually make good predictions, that those are critical. Some people have pointed out that you have to kind of assign different parts of IT to particular roles, which is a prior on the decoding space. For instance, that you could believe that biologically, an animal's born. There's some tissue that's going to be dedicated to faces. You have to wire those neurons downstream to that tissue. And that means you're going to restrict the decoding space, rather than just letting them learn from the space of IT as if they collected samples off of all of IT. So I think some people implicitly believe that even if it's not stated quite that way.

IT does not directly underlie recognition. You could imagine that. I mean, it's not for sure known. And some lesions of IT don't produce deficits in recognition. That's a possibility. Maybe you need too many training examples. Monkey neural codes cannot explain human behavior. You know, again, but I already showed you monkeys and humans are very similar.

So these are the reasons that you might say this is negative, and might not work. And probably already have guessed that I'm telling all these negatives because it turns out this simple thing works quite well for the grain of behavior that I've shown you so far. And here's my evidence of that. So this is actual behavioral performance out of humans that I showed you earlier. This is mean D prime. This is the predicted behavior or performance of taking a classifier, reading from that IT population data that I've shown you, which gives a predicted D prime. Here is-- we first chose a decoder. We had to match things like the number of neurons. We had to get it in the ballpark, so-- because again, there's a free variable, as I showed you earlier. There's at least one. But for now, let's think of matching the number of neurons to get you near the diagonal, so that you have sufficient number of neural recordings to say, how well do you do on a face detection task?

And then, here's all the other tasks. This is those 64 points that I showed you earlier. Here's some examples like fruit versus other things, car versus other things. And you should see that all these points kind of line up along this diagonal, which says, wow, this is actually quite predictive, that I can take this simple thing and predict all the stuff that we've collected so far.

And so let me now kind of be more concrete about what is the inferred neural mechanism that we're testing here? Well, I'll show you in a minute. This is, for each new object, we think what happens is some downstream observer, a downstream neuron, randomly samples roughly 50,000 single neurons, spatially distributed over all of IT, not biased to any compartments. Listens to each IT sites.

When I say listen in this case, we think could average over 100 milliseconds. We're not sure about this. This is just the version that's shown here. Learn an appropriate weighted sum of those IT spiking. And then listen at 10%. That's basically, once you learn, there's a heavily weighted about 10% of the IT neurons are heavily weighted for each of the tasks. That's just an observation that we have in our data. But this is trying to map it to neuroscientist language from these decoder versions out of IT.

So what that is a model that says, learn weighted sums of 50,000 random average 100 milliseconds single unit responses distributed over all IT. So a bunch of stuff in here is what your model is sort of encapsulating. That's still too long. So I made a little acronym out of that. And that caught Laws of RAD IT decoding mechanism. So this is just to say there's a hypothesis of how everything might work, but now can be make predictions for other objects and could potentially be falsified.

So, so far, this model works quite well over these tasks. And in fact, the correlation is 0.92. You might look at this and say, oh, it's not perfect. But it turns out that that's about the level that which humans differ from each other. So it's passing a Turing test, that this mechanism read off of the monkey IT hides in the distribution of the human population that we're asking to also perform these same tasks. So it can't be distinguished from being a human in these tasks. You guys, watch "X Machina?" Wasn't that a movie I saw? Doesn't pass that test. Passes just a simple core recognition test. But so that was a Turing test of this.

So OK, so, this is here that I quantified. So this is human to human consistency. That's the range I just mentioned that, you've got to get into here to pass our Turing test on this. And that's a decoding mechanism I just showed you. There's other ways of reading out of IT that

don't pass. There's ways of reading out of V4, which you recorded from-- none of them we've tried are able to get you to this here.

That doesn't mean V4 isn't involved. V4 is the feeder to IT. It just means you can't take simple decodes off of V4 and naturally produces this pattern. And that's similar for like, pixels or V1 representations. So lower level representations don't naturally predict this pattern of behavior. And even some computer vision codes that we tested at the time, as you can see, if those of you know these older computer vision models didn't do this. But more recent computer vision models actually do. And I'll show you that at the end. OK.

So, this is a little bit for the aficionados to tell you how we got there as we increase the number of units in IT, that drives performance up. So as you read more and more units out of IT, you get better and better performance. That's also true out of V4. But I'm trying to show you this here, is it's like, not the absolute performance that is the good thing to compare a model with actual behavioral data. It's the pattern of performance, which we call the consistency with the humans. That's that correlation along that diagonal that I showed you earlier, that tasks that are hard for the models are also hard for the humans. Tasks that are easy for humans are also easy for the models. And you could imagine doing that, not just at the task level, but at the image level as well.

And anyway, that's what's quantified here. And you see that when you get up to around you know, about 100-- I showed you 168 recordings out of IT. This point right there is about 500 IT features. And taking you through some things that maybe I won't have time for, that's actually how we approximate that 50,000 single IT neuron number. That's an inference from our data based on if we didn't actually record 50,000 single neurons.

But from these kind of plots, we're able to make a pretty good guess that this kind of model right here would produce-- would land right there. To be consistent with humans, and would get the absolute level of performance which humans matched. And you know, the models we tried out of V4, this is one example of them. They can get performance. But they can never-- they don't match this pattern of performance naturally. They over perform on some tasks, and under-perform on others. They sort of reveal themselves as not being human like by being too good at some things, right? So that's a way to fail the Turing test. OK.

Maybe I'll skip through this, it's sort of the same thing. This is about training examples. If those of you guys care about this, I could kind of take you through how we-- there's actually a family

of solutions in there. And I'm just telling you about one of them for simplicity.

So, let me then just take it down to another grain. So that was the pattern of performance, it's actually naturally predicted by this first decoding mechanism that we tried. But what about the confusion pattern? So not just the absolute D primes for each of these tasks, but there's finer grained data, like how often an animal is confused with a fruit, or an animal's confused with a face. These are the confusion pattern data here. I'm sorry I don't have the color bars up. All I'm going to need you to do is say, well these are the confusion patterns that we predicted.

And this is what is the predicted confusion pattern, if I gave the machine, the IT, these ground truth labels. And it predicts this. This is what actually happened in human data. And what I want to sort of look at this and this, and say, there actually look quite similar. Their noise corrected correlation is 0.91. So they were still quite good at predicting confusion patterns. Although this did not hold up fully. We're only at 0.68. I say only. Some people would say this is success. We're only at 0.68 on high variation. So there's a failure here of the model. That should be at 1, because it's noise corrected. So there's something about this that's not quite right at predicting the confusion patterns of humans at high variation images. And that to us, that's an opening to push forward, right?

So this is a strategy going forward as we have an initial guess of how you read out of IT. It looks pretty good for first grain test. But now we can turn the crank harder. We need more neural data. We need more psychophysics, finer grained measurements to sort of distinguish among, not just say IT's better than V4 or those other representations. But what exactly about the IT representation? Is it 100 milliseconds? What time scale? Maybe those synchronous codes do matter. Some of those things that I put on there earlier might start to matter when we push the code-- push this even further.

So what I take home here is that you do quite well with this first order rate code reads out of IT. But now there's an opportunity to try to dig in and say, well at what point do they break down? And what kind of decoding models are you going to replace them with? And that's what we're trying to do.

I've told you that IT does good at identity. But remember I said earlier on, remember I showed you those manifolds, and said there's other latent variables like position and scale. And I said those don't get thrown away. They just get unwrapped, right? Remember that manifold picture I showed earlier? And so one of the things we've been doing recently is asking, because we

built these images, we know these other latent variables, like position and pose-- that was one of the advantages of building the images this way. And we've been asking how well IT encodes those other latent variables about the pose of the object, the position of the object.

And to make-- let me just skip through. To make a long story short, IT actually encodes-- not only has information about these kind of variables, which is really not surprising, because others have shown that there's information about those kind of things before. But that's sort of what's on here. Everything what I'm showing here, here's IT V4 simulated V1 in pixels. And always, everything goes up along the ventral stream for the other variables, which may be non-intuitive to some of you. I mean, because position is supposed to be V1. But position of an object in a complex background is better at IT. That's one example.

But all these latent variables go up along the ventral stream in terms of their ease of decoding. But what I'm most excited about is that if you do this comparison with humans again, you actually get this sort of, again, pretty decent, not quite as tight correlation, between the human-- actual measured behavioral performance on making estimates of those other latent variables, and the predicted behavioral performance out of IT. And again, much better correlations. It's not perfect. So again, there's some gap here, some failure of understanding. But much better than if you read out of V4, V1 or pixel.

So this says that the representation again isn't just an identity thing. It seems like this could be representational underlie some of these other judgments, at least at the central 10 degrees for sort of foreground objects as we've been measuring here. That's the-- don't worry about the details on here-- that's the upshot of what I'm trying to say with this slide. But I just wanted to put that out there so you didn't forget that you haven't thrown away all this other interesting stuff about what's out there in the scene.

OK. Let me kind of-- I've sort of alluded to this a bit. I want to come back to kind of now, this is like Marr level 3 stuff, right? So you have this idea of what you're trying to solve. You have a decode-- you have an algorithm that's a decoder on a basis, that's trying-- that looks like it predicts pretty well. It's not perfect. There's work to be done there. But it actually does quite well. Now what does that mean on the physical hardware level? So that's Marr level 3.

So you think-- here's how I visualize it. You have IT cortex, which I mean AIT and CIT. So it's about 150 square millimeters in a monkey. And remember I told you there was about 1 millimeter scale of organization? I showed you that earlier. And others have shown-- I showed

this earlier, too-- that there's sort of face regions. So I've drawn them just for sort of for scale here, just a schematic. That they're slightly bigger organizations, they're 2 to 5 millimeter.

So I think of IT as being this sort of like 100 to 200 little-- similar to Tanaka. This is not a new conceptual idea. But there's sort of just the simple version would be each millimeter does exactly the same thing, is a feature. And if you sample off of that, you take 5,000 neurons, but they're really sampling from only about 150 IT features at 1 millimeter scale. Remember, I don't know if you caught that. But I showed 150-- 101-- 150. I showed you 168 IT neurons predicted the pattern of human performance. I showed that a few slides ago. But I told you the real number of neurons is probably 50,000. Most of those are redundant copies of that 168 dimensional feature set. That's how we think about it.

So you could imagine, it's just a redundant set of about-- I like to think of about 100 features in IT which are sampled maybe randomly downstream neurons that are then learned. So when you learn faces versus other things, hey, there's lots of good information about faces versus other things. And these face patches, that's how they're defined. But those neurons are going to lean heavily-- this downstream neuron is going to lean heavily on those neurons. And then these-- so that would make these regions causally involved.

So that doesn't mean you had to pre-build in anything here. You just learn this at a downstream version. And you would get something that looks like it would explain our data. So we like that, because it captures that case. But it also captures the more general case. If you learn cars, you're going to sample from a different subset of neurons. But you're following the same learning rule. That's what I said earlier on.

So you end up-- we think this is the initial state. This is when you learn objects. And so what we think is a post learning, what you have is again, about 100 to 150 IT sub regions, each at 1 millimeter scale, that are supporting a number of noun tasks read off this common basis here. That's the model that we like, given the kind of data that I've been showing you. The post learning model, as we call it.

So the reason I'm bringing this up is probably for the neuroscientists to fix ideas about how we think about IT as a basis set. And this is-- I think Haim sort of set this up nicely, he sort of implied similar things. That somebody downstream reads from it. OK. But now, we have a more-- you know, we're starting to have a more concrete model, that we now, I'm trying to start to be physical about it, about the size of these regions connecting to earlier data, how

many there are. So we're gaining inference on that from these different experiments.

And now, if you believe this, it starts to make a prediction of what's-- now we can do causality, right? Somebody mentioned that earlier. And so, one of the things we've been doing recently is if we can start to silence-- look, the way I've drawn this, this bit of tissue for-- this is just schematic-- is somehow involved in this task and that task. Face task and car task. But this bit of tissue, only face task. And that bit of tissue, only car task. And this bit of tissue, neither. So if you believe that, you had the tools, you should be able to go in and start to silence little bits of IT. And you should get predictable patterns out of the behavioral deficits of the animal when you make those manipulations, right? Everybody follow that? Right? OK.

And now the models give you a framework to build those predictions and to also estimate the magnitude of those effects that you should see. And so that's what we've been doing more recently. And I'll just give you a taste of this, because this is really ongoing. But I think it connects to what Gabriel said earlier about now there are these tools available to do that.

Oh, I put that in from an earlier talk where-- I think Google has a thing called Inception. And I don't know-- was it Google? Or somebody has it-- you can't do Inception unless you're actually in a brain. So are you going to try to insert-- the reason we do this is my student that is working on it really wants to inject signals in the brain. There's a dream about VMI, right? Could you kind of inject a percept? And to do that, you're going to need to do experiments like this. And you understand this hardware to interact with it. It's something we talked about earlier.

So actually-- and Tonegawa's lab has some cool Inception stuff on memory. But this is like inserting an object/person. So to do that, this has been a dream for many of us for a long time. Can we reliably disrupt performance by suppressing 1 millimeter bits of IT? So to do that, what we're doing is testing a large battery of tasks and a battery of suppression patterns. So not just sort of saying, can we affect face tasks or one task? But let's imagine we test a battery of tasks. And then, we-- and the idea where we'd have a whole bunch of tasks and we'd do every bit of IT one by one, and then in combination, and we'd sort of get all that data and figure out what's going on, right? That's sort of the dream, right? So we're trying to build towards that dream. Do you guys get it? Right. I mean, I don't know. And then we're motivated by this kind of idea here.

So to build-- so we started-- I'm just going to give you a quick tour of we have tools to start to

do this. You know, this is our recording, we can localize what we're recording two very fine grain using x-rays. So we know exactly where we're recording the IT to like about 300 micron resolution. So that's why I'm putting this slide up. And what we're interested in is going, if I silence this bit of IT, or that bit of IT, or that bit of IT, so actually do this experiment, what happens behaviorally?

And Arash Afraz is a post-doc in the lab, started these actual experiments. And one of the things Arash did was to first say, let's see if we can get this silencing of optogenetics tool to work in our hands. And the reason we were so excited about that is because we think lesions, if we can make temporary brief silencing, that that will give it much more reliable disruption of behavior that then, if we started to try to inject signals, which would be our dream, but that seems too risky to us. We just want to say, what is a temporary lesion of each bit of IT do?

And optogenetics is cool, because there's no other technique that can briefly silence-- temporarily silence activity. You can do pharmacological manipulations, but those last for hours. So this could briefly silence bits of IT. And that's why we were excited about it. We also did pharmacological manipulation as a reference to get started. But what we're doing is trying to silence 1 millimeter regions of IT using light delivered through optical fibers as the recording electrode. And to silence bits of neurons here.

And so what Arash did was first show that you can actually silence neurons in this way. So if you guys haven't seen optogenetics plots, this is data from our lab. What's quite cool about this, again, is you have the same images are being presented. So this green line should be up here.

But Arash turns a laser on right here, shines light on there. And there's some opsins expressed in the neurons in that local area. And you can see it just sort of shuts the thing down, and it sort of deletes or blocks this. You have the same input coming in. But you can sort of delete it here. And this is another example. These are some pretty strong examples. It's not always this strong. But this is, again, you can see we can return back to normal right away, right? So this is a 200 millisecond silencing. You could go even narrower than that.

But so this is what we had done so far. And again, what we did was say, look. This is a risky tool. This is it not going to work at all. So Arash just wanted to test something that was likely to work. And so we picked a face task because there was a lot of evidence of spatial clustering of faces that you'll hear from Winrich and you also known in the literature.

So what Arash did was to say, we picked a task of discriminating males from females. We put in our notion of invariance. It's not just do this image access. But you have to do it across a bunch of transformations. In this case, its identity as a transformation. So you're saying, all of these are supposed to be called male, and all these are called female. And he wanted you to distinguish this from this. That's what he trained a monkey to do.

And just to give you the upshot, is that, we do all this work, we silence the bits of cortex. And here's the big take home. You get a 2% deficit of single one millimeter silencing of bits of IT cortex. Parts of IT cortex, not all of IT cortex, produce a 2% deficit. Here's the animal running at 80%, 6% correct. These are interleaved trials where we silence some local bit of IT. You get a 2% deficit. That's true only in the contralateral field, not that ipsilateral field, for the aficionados.

You might look at this 2% and go, well, that's tiny. But we looked at it, this is exactly what's predicted by the models that we were talking about. It's right in the range of what should happen. And so this, to us, is really quite cool. This is highly significant. And now we sort of are in position to start to say, OK these tools work. They do what they're supposed to. And now we can start to expand that task space. So this result has been published recently, if you're interested in this.

And here is one of the ways we're going forward is that Rish Rajaingham, the one doing those tasks in the monkey I showed you earlier. Silencing different parts of IT. This is now with muscimol, different bits of IT-- these are different tasks, lead to different patterns. That's what these dots are here-- different patterns of deficits. And if you go back to the same location, you get the same pattern of deficits. So this is only 10 tasks. But I think it hopefully gives you the spirit of what we're trying to do. And again, this is only muscimol, which doesn't have all the advantages of optogenetics. But this is what we're were building towards here. So I'm just giving you the sort of state of the art.

So our aim is to measure the specific pattern of behavioral change induced by the suppression of each IT sub region, ideally testing many of them, and then compare with the model predictions.

I'm saying there's this domain, and I want to sort of sample the whole domain. So far, I've given you only just samples of tasks in the domain. But we're really trying to define the domain. And I'm just-- I'm going to skip through this just to give you the punchline, is that we

do a whole bunch of behavioral measurements. We presented this work before. It's like, this is now up to three million Mechanical Turk trials. It seems to us that we can embed all objects, even subordinate objects, of the type of task that I've been telling you, in roughly, in essentially a 20 dimensional space. So there's 20 dimensions. We think we infer that humans are projecting to about 20 dimensions to do these kind of, the tasks that we've shown here. Which is sort of smaller, but eerily close to that in the order of magnitude to that 100 or so features that I've been talking about.

So that's where-- regardless of whether-- these are some of the dimensions and how we're projecting them. Again, I won't take you through this, because I think we've already used up enough time and I want to get on to this part. But we're trying to define a domain of all tasks where we can sort of predict what would happen across anything within that domain. And that raises questions of the dimensionality of that domain. And there were behavioral methods to do that. And we've been doing some work on that. So I'll just leave it at that. And if you guys have questions, we can talk about that some more.

I want to sort of in the time I really have left is to talk about the encoding side of things, because I promised you guys I would get to this. Unless people have any more burning questions on this decoding side. So far I've been talking about the link between IT and perception. Now I'm going to switch gears and talk about this other side. Which is, so I talked about this. And that tells us that the mean rates in IT are something that seem to be highly predictive. I showed you at least one model that has the laws of RAD IT model. But now, it's like now, we can turn to the encoding side and say, we need to predict the mean rates of IT. And that should be our goal if we want to explain images to IT activity.

So, these would be called predictive encoding mechanisms. So, now you guys have heard about deep convolutional networks. If not, you've heard about them already, you'll probably hear about them some more. So we started messing around in 2008. This is a model inspired-- I mentioned this family of models before. Hubel-Wiesel, Fukushima, and there's a whole HMAX family of models, that really was the inspiration of this larger-- this large family of models, that have this repeating structure that are now really the sort of modern day deep convolution networks really grew out of all of this earlier work.

And so we started exploring the family in 2008. And just, this is a slide that you've already sort of seen a version of this from Gabriel where you know, for when you take an image, you pass it through a set of operators. So you have filters. So these are dot products over some

restricted spatial restricted region, like receptive fields. You have a non linear area, like a threshold and a saturation. You have pooling operation. Then you have a normalization.

So you have all these operations happen here. And that produces a stack. So think of like, if there are four filters here, like four orientations, you get four images, you have one image in, you have four images out. But if you had 10 of these, you'd get 10 of these out. Then you repeat this here, right? And so as you keep adding more filters, this stack just keeps getting bigger and bigger. And it keeps, because you're spatially pooling, it keeps getting narrower and narrower, right? So you go from this image to this sort of deep stack of features that has less retinatopy. It still has a little bit of retinotopy.

And that, you can see, has been exactly a very good model why people liked it of how people think about the ventral stream. So these models typically have thousands of feat-- visual neurons or features at the top level. Just to give you a sense of scale of how they're run. And just to take you through, you know, I guess maybe you'll hear about this, if you haven't already.

Each element has like, a filter, has a large fan in. Like these are like neuroscience related things. They have non-linearities, like thresholds of neurons. Each layer is convolutional, which means you apply the same filters across visual space. Which is like retinotopy, that is a view on cell that is oriented here. There'll be another view on cell that's in another spatial position, same orientation, different spatial position. That's what the convolutional models are just an implementation of that idea of copying the same filter type across the retina. And there's a deep stack of layers. These are all things that I think are commensurate with the ventral stream anatomy and physiology.

So, but one of the key things that those who work with these models know is that, they have lots of unknown parameters that are not determined from the neurobiology. Even though the family of models is well described, what are the exact filter weights? What are the threshold parameters? How exactly do you pool? How do you normalize? There's lots of parameters when you build these things, essentially thousands of parameters, most of them hidden in the weight structure here. Which, if you think about, the first layer, that would be like, should I choose Gabor filters? Or should I do some other-- you know Haim was talking about random weights, right? So there's choices there. There are lots of parameters.

So the upshot is, there's a big-- that's why I call it a family of models. And how do you choose

which one is the right one, so to speak? Or is there a right one? Or maybe the whole family is wrong, right? These are the interesting discussions. So, what I like about it is, at least when you set it, it's a model. It makes predictions. And then you can test it. So it's at least a model. And it predicts the entire-- you know, if you start to map these, you say this is V1, this is V2, this is V4. It predicts the full neural population response to any image across these areas. So it's a strongly predictive model once built.

So that's nice. But now you have to determine how am I going to build it? How do I set the parameters? So how do we do that? Well, there's lots of ways you could do it. And I'll tell you the way we chose to do it. Which was to just not use any neural data. It was just to use optimization methods to find specific models to set the parameters inside this model class. And we chose an optimization target. This is a little bit, again, inspired from a top down view of what the system's doing. What are the visual tasks that we suppose the ventral stream was supposed to solve? Which I already told you, we think it's invariant object recognition. That's what makes the problem hard. So we tried to optimize models to solve that.

And essentially when we're doing that, we're kind of doing the same thing that computer vision is trying to do, except we're doing it in our own domain of images and tasks that we set up. But we essentially, there's a meeting between computer vision and what we were trying to do here. And when I say we, this is work by Dan Yamins, a post-doc in the lab, and Ha Hong, a graduate student.

And what we did was to just try to simulate again, as I did earlier. We took these simple 3-D objects. We could render them, just as before, place them on naturalistic background. And then we just built models that would try to discriminate bodies from buildings from flowers from guns. So they would have good feature sets that would discriminate between these things.

And these were essentially trained by various forms of supervision. Now there's lots of ways you can train these models. I could tell you about how we did it and how others have done it. I think those details are beyond what I want to talk about today. But just, it's a supervised class that's probably not learned in the same way that the brain has learned. Most people don't think so. But the interesting thing is the end state of these models might look very much like the current adult state of the brain. And that's what I want to try to tell you next.

So first, let me show you that when we built these models, this was in 2012. We had a particular optimization approach that we called HMO that was trying to solve these kind of

problems that I showed you earlier on these kind of images. And I showed you IT was pretty good with humans. I showed you its performance was almost up to humans, even with just 168 samples.

And when we first built a model here, we were able to do much better than some of our previous models that-- on these same kind of tasks. So I told you we constructed, because we knew it made these things-- we made these models not do so well. So we built these high invariance tasks to push these models down. And then we had space to build a model that we could do better on. And we called it HMO 1.0. And then we started to say, now we have this model that has been optimized for performance. Let's see how well it does on comparing with neurons. Let's see if its internals look like the neural data.

So here's the model we built, HMO 1.0. It's a deep convolutional network. It has two different levels. It had four levels. It had a bunch of parameters that we set by optimization, that I'm just telling you kind of what we optimized. I didn't tell you-- I'm not telling you any of the parameters. And now, we come back to say, well look. We can show the same images to the model that we showed to the neurons. And then we can compare how well these populations look like that population, or this population looks like that.

And so what we did was, we asked how well can layer four predict IT first? That was the first thing we wanted to do, take the top layer of this model, the last layer before the linear readout of this model. And to do that, you might sort of say, well, wait a minute. The model doesn't have mappings. It has sort of neurons simulated here, neuron 12 or something. And there's some neuron we recorded. But there's no linkage between that neuron and that neuron, right? You have to make that map.

So what we do is we take each IT neuron and treat this as sort of a generative space. You can generate as many simulated IT neurons as you want. You would just ask, let's take this neuron, take some of its data, and try to build a linear regression to this neuron. Treat this as a basis to explain that neuron. And then test the predictive power on the held out IT data. And that's what I'm writing here. That's cross-validation linear regression.

So I'm going to show you predictions on held out data where some of the data were used to make the mapping. And there's lots of ways we chose-- we could make the mapping. And we did essentially all of them. And I could talk about that if you want. But that's this central idea. Take some of your data, say, is this in the linear space spanned by this basis set? So I can I fit

that well with this linear basis here? As a linear map from this basis?

And here's what we actually-- here's what it looks like. Here's the IT neural response of one simulated-- one actual IT neuron in black. This is not time. These are images. I think there's like 1,600 images here. So each black going up and down, you can barely see, is the response, the mean response, to different images. And you see we grouped them by categories, just so, just to help you kind of understand the data. Otherwise, it'd just be a big mess. Because IT neurons do-- you can kind of see they have a bit of category selectivity.

And again, this was known. This neuron seems to like chair images, but not all chair images. It sometimes likes boats and some planes a little bit. And the red line is the prediction of the model, once fit to part of the-- to this neuron. This is the prediction on the held out data for the neuron. You can see the R squared is 0.48. So half the explainable response variance is explained by this model. And again, these are predictions. The images were never seen-- the objects even were never seen by this model before it makes these predictions here.

So this is just saying that the IT neurons live in this space. It's actually quite well captured by the top level, in this case, of this first HMO model we built. I'll show you some other models in a minute.

Here's another neuron that you might call a face neuron because it tends to like faces over other categories. So it might-- it would pass the test of the operational definition of a face neuron. This model, this neuron was well predicted, again, by both its preferred and non-preferred face images by this HMO model. Again, a slightly-- an R squared near 0.5.

Here's a neuron that you would look at the category structure. And you don't even-- you can't really see the categories here. They're still here. But you don't see these sort of blocks. You just see there's sort of some images it likes and some it doesn't. It's hard to even know what's driving this neuron. But it's actually quite well predicted, I think. You don't have the R squared. But it's similar. It's about half the explainable variance. Just another example.

And here is a sort of summary here. If you take-- this is a distribution of the explainable variance for the top level of the model fitting about, I think this is 168 IT sites. Some sites are fit really well, near 100%. Some are fit not as well. The average is about 50%, which is shown here. So this is the median of that distribution here. So the summary take home is about 50% of singularly response variance predicted. And this is a big improvement over previous models I'll show you in a minute.

The other levels of the model don't predict nearly well. So the first level doesn't predict well. Second level better, third level better, the fourth level the best. If you take other models-- these are some of the models I showed you earlier-- they don't fit nearly as well. Here's their distributions and here's their average, their median explained variance. And just to fix-- to just fix ideas, you might think, well look, we built a model that's a good categorizer. So of course it fits IT neurons well. Because IT neurons are categorizers.

Well, here's a model that actually has explicit knowledge of the category. It's not an image computable model, and it's not an easy one. But it's just given that sort of an oracle that's given the category, and how well it explains IT. And you can see, it explains IT much worse than the actual model. So this implies a model is limited by the real-- the architecture puts constraints on the model and how it adds variance that the sustained IT neurons are categories does not easily capture.

So that kind of-- that sort of inspired us to say, OK. What about if we go down and say not just IT, but let's go to V4. Because we had a bunch of V4 data. And so we play the same game in V4. Let's take level three and see if we can predict V4. And here's the IT data I just showed you a minute ago. And here's the V4 data. So the V4 neurons are highly predicted in the middle layer. Layer three is the best predictor of V4. The top layer is actually not so predictive, less predictive of V4 neurons than the middle layers. And the first layer is not so well predictive. And again, the other models are actually, now you can see they're getting on relatively better. You can think of them as sort of lower level models. And they're getting better, which is what you'd expect.

But interestingly, this is really exciting to us. Because look, this model was not optimized to fit any neural data other than that last mapping step. All it is is a bio inspired algorithm class, which is the neuroscience sort of view of the feed-forward class of the field. And tasks that we and others hypothesize are important, that the ventral stream might be optimized to solve, and an actual optimization procedure that we applied. And that leads to neural like encoding functions at the top and in the middle layer. So you don't-- so this sort of leads to funny things like saying, what does V4 do? The answer here would be, well, it's an intermediate layer in a network built to optimize these things. That's the way to describe what V4 does, according to this kind of modeling approach.

Now I want to point out, this is only half of the explainable variance. So it's far from perfect.

There's room to improve here. But it's really dramatic how much improvement we got out of these kind of models. And so if you take this sort of-- well, I'll skip this. If you take this back to you know, big picture, what did we do here? What we're doing is we have performance of a model on high end variance recognition tasks. We're saying, this is what we've been trying to optimize.

And what we noticed is that if you plot-- these dots are samples out of that model family. These black dots are other models I showed you. So they're control models that were in the field at the time. And this is the ability of the top-- the model-- the top level of any of the models to predict IT responses. So, you know, how good they are predicting-- this is sort of the median variance explained of single IT responses.

And you see there's a correlation here. If you're better at this, you're better at predicting that. And all we did was optimize this way, which we think of as like, evolution or development. So we're not fitting neural data. We're just optimizing for task performance. And that led in 2012 to a model that I just showed you, explained about half of the IT response variance. OK, so it's like, well, this looks like it's continuing up this way.

OK so if you believe that story, then, that says, if we can optimize further on these kind of tasks, maybe we can explain more variance. And it turned out, we didn't actually need to do that, because again, I said, computer vision was already working on this. And they got a lot more resources. They're already doing it. They're already better than us on this. So here's our HMO model.

This is now Charles Cadieu, a post-doc in the lab. These were models that came out at the time. This is Krizhevski et al. supervision. It's ICLR 2013. They were better than the model that we had built. You know, we were in this restricted image domain, you know, there's lots of reasons why we could say they're better. Regardless, they were better at our own tasks than the models that we had built, right? So they were already ahead of us on the task that we had designed.

And so they were up here, and then they were up here. And so, if you follow that prediction, that means these models might be better predictors of our neural data, right? These guys don't have our neural data. All they're doing is building models to optimize performance on tasks. And but we could take their features from the neural data, play the same game. And we actually explained our response-- data better than our model explained our own data. So this

is a nice statement that is not even in our own lab. Just a continued optimization for those kinds of tasks leads to features that are good predictors of the IT responses. And that's what's shown here. So I think that's what I just said there.

So, Charles took this further and analyzed this in more detail. This is a summary of what I presented in the second half now, showing that IT firing rates are feature based, learned object judgments naturally predict human monkey performance. This is why the laws of RAD IT. I picked a particular model, which is 100 millisecond read on this time window, 50,000 neurons. 100 training examples. That's one particular choice of a decode model, that's just a-- is a current set of decode model that fits a lot of our data, but not all of our data. And we also want to get finer grain data.

The inference is, this might be the specific neural code and decoding mechanism that the brain uses to support these tasks. That's what we'd like to think. But now, we're trying to do systematic causal tests. And we talked a lot about trying to silence bits of IT as one example of that. And the tools are still not where we'd like them to be. But you see we're making progress there.

So the second was I showed the optimization of deep CNN models for invariant object recognition tasks led to dramatic improvements in our ability to predict IT and V4 responses. I showed you our model HMO. But then the convolutional neural networks in the field have already surpassed our predictive ability on our own data. And so the inference is that these encoding mechanisms in these models might be similar to those that work in the ventral stream.

And now, you know, there's a whole sort of area where you can start to think about doing physiology on the models, so to speak. And that problem's almost as hard as doing physiology except on the animal, except that you can gain a lot more data. And so, and this is allowing the field to design experiments to explore what remains, what's unique and powerful about primate object perception. So within core object recognition or perhaps having to extend out of that, I think is now what people are trying to do.

So big picture in terms of us for the future, I've talked about this law's of RAD IT. Can we perturb here and get effects here that are predictable? Can we predict for each image, coding model, and for the optical manipulations? We talked about that. Dynamics and feedback are something that we're interested in. But I haven't talked much at all about. I think that's a good

point, a discussion topic. I can tell you how we're thinking about it. We have some efforts in that regard.

I talked on the encoding side about these kind of deep convolutional networks that map from images. But the dash lines mean they're only 50% predicted. Both of these cases, they're not perfect, right? So there's work to be done there. And one of the really exciting things is here is how these models learn. This supervised way of learning these models is almost surely not what's going on in the brain. So finding more-- less supervised, biologically motivated learning of these models is a good-- is the next step, I think, for much of the field. But what's nice is to have an end state that is much better than any previous end state we'd had before. So that sets a target of what success might look like.

And you know, maybe we can think about expanding beyond core recognition. We can talk in the question period about that. When is the right time to kind of keep working within the domain of core recognition that is set up, versus expanding beyond that? Because there's lots of aspects of object recognition that I didn't touch on here. And that comes up in the questions. I think, there's lots of work to be done within the domain, but there's also interesting directions that extend outside of that domain.