EERO SIMONCELLI: I'm going to talk about a bunch of work that we've been doing over the last-- it's about four years, on trying to understand basically, that terra incognita that Gabrielle just mentioned that lies between V1 and IT. I brought this back with me from the Dolomites, where I was last week with my family. And when you sit and you look at it and that image comes into your eyes and gets processed by your brain, there's a lot of information there. It's a lot of pixels.

And the question that I'm going to start with is, where does it go? You have all this information. It's flooding into your eyes all day every day for your entire lifetime. Obviously, you don't store it all there. Your head doesn't inflate until it gets to the point of explosion.

So where does it go? And as a theorist's diagram of the brain-- a square with rounded corners. In comes the information, and there are really only three options. You either act on the information, do something with it, sensory motor loops, or for complex organisms especially, a fair amount of it you might actually try to remember. You might hold on to, and we heard about that earlier today.

But this really only accounts for, I think, a fairly small portion of what goes on, because a lot of it you throw away. You have to. You really don't have a choice. You have to summarize it, squeeze it down to the relevant bits that you're going to hold on to or act on, and the rest of it you just toss.

So the question is, how can we exploit that basic fact? It's an obvious fact. It has to be true. How do we exploit that to understand something about what the system does and what it doesn't do?

And there's a long history to this, and in fact, since I come from vision and most of my work is centered on vision and auditory system to some extent, the vision scientists were the first to recognize the importance of this. And it really is a foundational chunk of work in the beginning of the field that set in motion a lot of things that we currently know about vision. And so I'm going to just-- for those of you that don't know that story, I'm going to give a very, very brief

reminder of what that is, because I think it's an absolutely fantastic scientific story. And then from there, I'll talk about texture.

So two examples-- I'm going to quickly say something about trichomatic color vision, and then I'm going to talk about texture, and then we'll go into V2 and metamers and other things. So trichromacy-- Newton figured out that light comes in wavelengths. He split light with a prism. There's the picture drawing of him splitting light coming in through a hole in the wall. He split it with a prism into wavelengths, saw a rainbow, did a lot of experiments to recognize that you could take that rainbow and reassemble it into white light, but you couldn't further subdivide it, and basically gave us the foundations for thinking about light and spectral distributions.

In the 1800s, a group of people that were combined physicists, mathematicians, and psychologists all rolled into one-- and there were quite a few of them. Helmholtz was one of them. Grassmann was one of the most important ones. I'll mention him again in a moment-- figured out something peculiar about human vision-- that even though there was this huge array of colors in the wavelengths in the spectrum, that humans actually had these deficits that we were not able to actually sense or discriminate things that it seemed like we should be able to.

And it boiled down in the end, after a lot of study and discussion and theorizing, to this experiment, which is known as a bipartite color matching experiment. So on the left side of this little display, here's a gray annulus. In the middle is a circle.

On the left side of this is light coming from some source. It has some spectral distribution illustrated here. This is all a cartoon, but just to give you the idea of how this works. On the right side are three primary lights.

And the job of the observer in this experiment is to adjust, let's say, sliders or knobs in order to change the intensity of these three lights to make the light on the right side of this split circle look the same as the light on the left side. And it turns out that-- so just to be clear, so these three things have their own spectral distributions. They might look like that, for example.

And when the observer comes up with the knob settings, they're going to produce something that might look like that. This is just a sum of the three copies of these spectra weighted by the knob settings. So this is a linear combination of three spectral distributions. And intentionally, I've drawn this so that they don't look the same, because that's the whole point of the

experiment.

It turns out that humans-- you can do this experiment, and that any human with normal color vision can make these settings so that these two things are absolutely indistinguishable. They look identical, and yet they knew even in the mid-1800s that these two things have very different spectra, and I've drawn it that way intentionally. So the point is that humans are obviously-- even though we can see all the bands of the spectrum, we can see all the colors-- we actually have this deficiency in terms of noticing the difference between these two things. So how can that be?

And I think and hope that most of you know the answer to that question, because you're using devices every day that are exploiting this fact. But the bottom line is that in the 1850s Grassmann laid down a set of rules. Grassmann was a mathematician. He actually developed a large chunk of linear algebra in order to explain and understand and manipulate these ideas.

And he pointed out that-- he actually had a set of laws that he laid out, and I won't drag you through all of that. But in the end, what all of those laws amounted to, taking into account all of the evidence that he had, he laid down these laws. And what it amounted to is that the human being, when setting these knobs, was acting like a linear system. The human was taking an input, which is a wavelength spectrum, and adjusting the knobs. And the settings of the knobs were a linear function of the wavelength spectrum that was coming into the eye.

And it's a remarkable and amazing fact, if you know that the brain is a highly non-linear device, how is it that a human can act like a linear device? And the answer is that basically the human, taking this thing in and making the knob settings, has a front end that's linear and is doing a projection of the wavelength spectrum onto basically three axes. And those three measurements-- that process is linear.

Everything that happens after that, which is complicated and non-linear and involves noise and decisions and all kinds of motor control and everything else-- as long as the information in those original three measurements is not lost, then the human is going to basically act like a linear system, in terms of doing this matching. So Grassmann realized this. The theory that he set out and that others then elaborated on perfectly explained all the data for normal human subjects, that lights that appear identical but had physically distinct wavelength spectra could be created, and they called these metamers-- two things that are physically different but look the same.

This was codified. It took many, many decades. Things moved slower back then. We don't have these rapid, Google-style overturns of scientific establishment within a year or two.

It took until the 1930s to actually build this into a set of standards that were used in the engineering community to generate and create color film, color devices, eventually color video, color monitors, color projectors, color printers-- everything else that we use. And these specifications were to allow the reproduction of colors so that they looked the way they were supposed to look.

So you record color with a camera. It turns out that your camera is also only recording three color channels, just like your eye, and then you have to be able to re-render that on another device. And these standards specify how to do that.

The surprising thing in the whole story is-- so this is 1850s. Well, we go back to Newton. It was a 1600s. Then in the 1850s, when we're getting this beautiful theory that's very, very precise, this gets built into engineering standards. And it's not until 1987 that it actually gets verified in a mechanistic sense.

And I like to tell this story, because I think it's a reminder that aiming always for the reductionist solution is not necessarily the right thing to do. This is a very beautiful piece of science that was done at Stanford, actually, by Baylor, Nunn, and Schnapf. They took cones from a macaque-- I think originally they worked with turtles, but then macaque, sucked them up into a glass micro-pipette, shined monochromatic lights through them, and measured their absorption properties. And they found these three functions for three different types of cones and verified, basically, that these three absorption spectra perfectly explained the data from the 1800s.

So this is an amazing thing, if you can have a theory and a set of behavioral experiments that make very precise and clear predictions that then get verified and tested in a mechanistic sense more than 100 years later, and they come out basically perfect. So it's an astounding, astounding sequence, in my view.

So what we wanted to do is to set out trying to do the same kind of thing for pattern vision. And we're going to do that by thinking about texture. So what's a texture? A texture is an image that's homogeneous with repeated structures.

So each of these are examples of texture. That's a piece of woven basket. This is tree bark.

That's a herringbone pattern, and these are some sort of nuts or stones. And each of these has the property that there's lots of repeated elements with some variability. Sometimes there's more variability, sometimes there's less variability, but there's usually at least some.

And of course, these things are ubiquitous. When I started working on this, which is about 15 years ago-- maybe a little bit more, 16 years ago-- I started photographing things that I saw as I walked around, and textures are everywhere. Most things are textured. The world is not made up of plain-- of Mondrians. It's not made up of things that are plain, blank colors separated by sharp edges. It's made up of textures, and often the boundaries between things are boundaries between things that are textured objects, like the seats in the auditorium, for example.

So how is it that we can go about thinking about this in terms of metamers and representation in, let's say, the visual system? And the idea really comes from Julesz, who proposed in 1962 a famous theory that he later abandoned. The theory goes like this. First of all, he said the thing that we're going to do to try to describe textures is we're going to use statistics.

And why statistics? Because these are supposed to be variables, so I need some stochasticity. But I also want something that's homogeneous, so I'm going to average or measure things averaged across the entire image. That's the statistical side of it.

And he proposed that, well, if I start by measuring just pixel statistics-- say single pixel statistics, pairwise pixels statistics, maybe triples of pixels, eventually I should reach a point where I've made enough measurements to actually sufficiently constrain the texture such that any two textures that have the same statistics up to that-- whatever that order is, should look the same to a human being. And he didn't talk about this in physiological terms, but I think in the background is the notion that humans are actually measuring those statistics, and if you can get them right-- if you can make two images have the same statistics, and that's the only thing that humans are measuring, then those two images will look the same.

So Julesz goes ahead with this, and eventually constructs by hand, because he did everything with binary patterns constructed by hand-- he constructs these two examples that are identical. He first falsifies the theory at n equals 2, and then he tries to do third-order statistics. And he comes up with these two examples-- counter-examples to the theory.

These are matched in terms of their third-order statistics. It's not easy to see that or realize that, but it's true. If you take triples of pixels, and you take the product of those three, and you

average that over the image, these two things are identical, but they look very different. And if you draw samples of each of these, it's very easy to label them as, let's say, A or B into these two categories. Here's another example that came out a bit later by Jack Yellott. These two things also are matched up to third-order.

So Julesz decides that the theory is a failure, and he abandons it. And he begins a new theory, which is the theory of textons, which is a much less precisely-specified theory that has to do with laying down-- basically, it's a generative model, if you like. Everybody's fond of generative models these days, except for me. And he comes up with a generative model-- ah, and maybe Tommy.

He comes up with the generative model, which is to lay down many copies of a small, repeating unit, which he called the texton. And so he came up with this method of generating texture images, which he went to town on, and he made lots of examples. The problem is that that wasn't a description of how to analyze texture images or how a human would analyze texture images, and so it became very difficult to bridge that gap. And I think, in my view, that the theory really never succeeded, and he should have stuck with the initial theory.

Anyway, but that gave us an opportunity. So we went back many years later-- this is around 1999. I had a fantastic post-doc, Javier Portilla, who came from Spain, and we started thinking about texture and started putting together a model that was Juleszian in spirit, but a little bit different, because we wanted to build in a little bit of what we knew about physiology. Now, Julesz knew about physiology, because Hubel and Wiesel were doing all those experiments in V1 in the late '50S and the early '60s, but he really didn't incorporate that into his thinking.

So what we did is build a very simple model. It's just dumb, stupid, simple, in which we took this description of V1 neurons. So these are oriented receptive fields. The idea is that this is a description of a neuron that takes a weighted sum of the pixels with positive and negative lobes. And it has a preferred orientation, because the positive and negative lobes have a particular oriented structure. And then it takes the output of that weighted sum and runs it through some rectifying, nonlinear function.

And here's another, and this is a classic thing that Hubel and Wiesel described for a simple cell. And here's another one, which is a complex cell. And this one basically does two of these and combines them. I'm trying to avoid the details here, because they're not critical for understanding what going to show you.

So then we took those things and we said, well, what if we measure joint statistics of those things over the image? So we're going to take not just these filters, but of course, we're going to do a convolution. That is, we're going to compute the response of this weighted sum at different positions throughout the image. We're going to rectify all of them.

Now we're going to take joint statistics. What do I mean by that? Just correlations, basically-- second-order statistics of the simple cells, of the complex cells, of the cross statistics between them. And these statistics are between different orientations and different positions and also different sizes.

And given that large set of numbers-- and typically for the images that we worked with back then, these were typically on the order of 700 numbers. So we have an image over here, which is say, tens of thousands or hundreds of thousands of pixels, being transformed through this box into a set of, let's say, 700 numbers. So 700 summary statistics to describe this pattern.

And then the question is, how do we test the model? And for testing the model-- most people, when they test models like this, they do classification. This should sound very familiar these days, with the deep network world.

They take a model, and then they run it on lots of examples. And they ask, well, do the examples that are supposed to be the same kind of thing, like the same tree bark-- do they come out with statistics that are similar or almost the same as each other? And can I classify or group them or cluster them and get the right answer when trying to identify the different examples?

We decided that that was a very-- at least at the time, a very weak test of this model, because this is a high-dimensional space, and we had only, let's say, on the order of hundreds of example textures. And hundreds of-- that sounds like a lot of textures-- a couple hundred textures, but if the outputs live in a 700 dimensional space, then it's basically nothing. We're not filling that space. And for those of you that are statistically-oriented, you know that there's this thing called the curse of dimensionality. The number of data samples that you need to fill up a space goes up exponentially with the number of dimensions.

So this was really bad news, and we decided that it was going to be a disaster to just do classification-- that pretty much any set of measurements would work for classification. So we

were looking for a more demanding test of the model. And for that, we turned to synthesis.

So the idea is like this. So you take this image. You run it through the model. You get your responses.

Now we're going to take a patch of white noise. We're going to run it through the same model, and then we're going to lean on the noise, push on it-- push on all the pixels in that noise image until we get the same outputs. So this is sometimes called synthesis by analysis.

This is not a generative model, but we're using it like a generative model. We're going to draw samples of images that have the same statistics by starting with white noise and just pounding on it until it looks right. And pounding on it means, for those of you that want to know, measuring the gradients of the deviation away from the desired output and just moving in the direction of the gradient. I'm giving you the quick version of this.

A little bit more abstractly, we can think of it this way. There's a space of all possible images. Here's the original image. It's a point in this space.

We compute the responses of the model, which is a lower dimensional space-- a smaller space. That's this. Because this is a many to one mapping and it's continuous, there's actually a manifold-- a continuous a collection of images over here, all of which have the same exact model responses. And what we're trying to do is grab one of these.

We want to draw a sample from that manifold. If the theory is right-- if this model is a good representation of what humans see and capture when they look at textures, then all of these things should look the same. That's the hypothesis.

And the way we do it, again, is to start with a noise seed-- just an image filled with noise. We project it onto the manifold. We push it onto this point. We can test that, because we can, of course, measure the same things on this image and make sure that they're the same as this image, and that's our synthesized image. So that's a abstract picture of what I told you on the previous slide.

And then finally, the scientific or experimental logic is to test this by showing it to a human observer. So we have the original image, and then we compute the model responses. We generate a new image, and we ask the human, do these look the same?

And if the model captures the same properties as the visual system, then two images with

identical model responses should appear identical to a human. So that's the logic. And any strong failure of this indicates that the model is insufficient to capture what is important about these images.

So it works, or I wouldn't be telling you about it. Here is just a few examples. There are hundreds more on the web page that describes this work.

On the top are original photographs-- lizard skin, plaster of some sort, beans. On the bottom are synthesized versions of these. The lizard skin works really well. The plaster works quite well. The beans a little less so.

And it depends-- whether it works well or not depends on the viewing condition. So if you flash these up quickly, people might be convinced that they all look really great. If you allow them to inspect them carefully, they can start to see deviations or funny little artifacts. So it's a partial success.

And I should point out that it also provides a pretty convincing success on Julesz' counter-examples. So these are examples. This is synthesized from that, and this is synthesized from that, and they're easily classifiable.

And there's fun things you can do with this. You can fill in regions around images. So if you take this little chunk of text here and you measure the statistics, and you say, fill in the stuff around it with something with has the same statistics, but try to do a careful job of matching up at the boundaries, you can create things like this. So you can read the words in the center, but the outside looks like gibberish.

Each one of these was created in the same way. So the center of each of these is the original image, and what's around it is synthesized. So it works reasonably well.

You can also do fun things like this. So these are examples where-- I told you we started from white noise, and then pushed it onto the manifold, but we can actually start from any image. So if we start from these images-- these are three of my collaborators-- two of my students and my collaborator Tony Movshon.

If we start with those as starting point images, and we use these textures for each of them, we arrive at these images, where you can still see some of the global structure of the face. Because the model is a homogeneous model, it doesn't impose anything on global structure. And so if you seed it with something that has particular global structure or arrangement, it will

inherit some of that. It'll hold onto it.

Anyway, this is just for fun. Let's get back to science. So now, here's an example of Richard Feynman. This is Richard Feynman after he's gone through the blender.

You can see pieces of skin-like things and folds and flaps, but it's all disorganized. Again it's a homogeneous model. It doesn't know anything about the global organization of this photograph.

But what we want to know is-- so do we have a model that's just a model for the perception of homogeneous textures, or can we actually push it a little bit and make it, first of all, a little more physiological, and second of all, maybe a little bit more relevant for everyday vision? For me, standing here and looking at this scene, how do I go about describing something like this that's going on when I'm looking at a normal scene? So let's go through thinking about how to do this.

So I'm going to jump right to this diagram of the brain again. So V1 is in the back of the brain. The information that comes into your eyes goes through the retina, the LGN, back to V1. And then it splits into these two branches, the dorsal and the ventral stream. The ventral stream is usually associated with spatial form and recognition and memory.

So I'm going to think about the ventral stream, and we're going to try to understand what this model might have to say about processing in the ventral stream. I'm going to rely on just a few simple assumptions. First, that each of these areas has neurons, and that they respond to small contents or regions of the visual input. They're known as receptive fields. Most of you know that.

In each visual area, I'm going to assume that those receptive fields are covering, blanketing the entire visual field. So there's no dead spots. There's no spots that are left out. Everything is covered nicely.

And in fact, we know that this is true, for example, starting in the retina. So this is a cartoon diagram to illustrate the inhomogeneity that's found in the retina. So the receptive field sizes in the retina grow with eccentricity. And it turns out that that starts in the retina, but that's true, actually, all the way through the visual system and throughout the ventral stream, in particular.

And this diagram is showing these little circles are about 10 times the size of your midget

ganglion cell receptive fields in your retina. So you looking-- if you fixate right here in the center of this, these things are about 10 times the size of your receptive fields. And that's been long thought to be the primary driver of your limits on acuity, in terms of peripheral vision.

So in particular, if you take this eye chart, which is modified by-- this was done by Richard Anstis back in the '70s, and you lay it out in this fashion, these things are about 10 times the threshold for visibility and recognition of these letters. And so you can say that the stroke widths of the letters are about matched to the size of these ganglion cells, and it works, at least qualitatively-- that things are scaling in the right way, in terms of acuity, and in terms of the size that the letters need to be for you to recognize them.

And you can make pictures like this. This is after Bill Geisler, who showed that if you foveate-- if you fixate here, in fact, you can't see the details of the stuff that's far from your fixation point, and if you blur it, people don't notice. You can actually add high frequency noise to it, alternatively, and people won't notice that either. Because those receptive fields are getting larger and larger, and you're basically blurring out the information that would allow you to distinguish, let's say, these two things. When you look right at it you can see it, but if you keep your eye fixated here, you won't notice it.

So let's work off of those ideas-- the idea of these receptive fields that are getting larger with eccentricity, that are covering the entire visual field. And let's notice the following-- so this is data taken-- physiological data from several papers that were assembled by Jeremy Freeman, who was a grad student in my lab. And here you can see the center of the receptor fields versus the size of the receptive fields.

And you can see that in the retina-- I already showed you on the previous slide that it grows with eccentricity, but it's actually very slow compared to what happens in the cortex. V1, the receptor fields grow at a pretty good clip. V2, they grow about twice as fast as that, and V4 twice as fast again.

Another way of saying this-- at any given receptive field location relative to the fovea-- let's say 15 degrees, the receptive fields in V1 are of a given size. It's on the order of 0.2 to 0.25 times. The diameter is 0.2 to 0.25 times the eccentricity. The receptive fields in V2 are twice that size, so about 0.45 times the eccentricity, and the receptive fields in V4 are twice that again.

In cartoon form, it looks something like this. So here's V1. Lots of cells and small-ish receptive fields growing with eccentricity. Here's V2. They're bigger. They grow faster.

Here's V4. And by the time you get to IT-- Jim DiCarlo was here a bunch of days ago, and he probably told you this-- almost every IT cell includes the fovea as part of its receptive field. They're very large, and they often cover half the visual field.

So now we have to figure out what to put inside of these little circles in order to make a model, and I'm going to basically combine-- smash together the texture model that I told you about, which was a global homogeneous model, with this receptive field model. I'm going to basically stick a little texture model in each of these little circles. That's the concept.

So how do we do that? Well, we're going to go back to Hubel and Wiesel. Hubel and Wiesel were the ones that said you make V1 receptive field simple cells out of LGN cells by just taking a bunch of LGN cells that line up. Here they are-- center surround receptive fields from the LGN, which are coming off of the center surround architecture of the retina. You line them up, you add them together, and that gives you an oriented receptive field, like the ones that I showed you earlier.

And in more of a computational diagram, you might draw it like this. So here's an array of LGN inputs coming in. We're going to take a weighted sum of those. Black is negative. White is positive. So we add up these three guys, we subtract the two guys on either side, and then we run that through a rectifying nonlinearity that's a simple cell.

Hubel and Wiesel also pointed out that you could maybe create-- or suggested that you create complex cells by combining simple cells. This is the diagram from their paper in 1962. And so we can diagram that like this.

Here's basically three of these simple cells. They're displaced in position, but they have the same orientation. We halfway rectify all of them, add them together, and that gives us a complex cell.

So it's interesting to note that the hook here is going to be that this is an average of these. An average is a statistic. It's a local average. So we're going to compute local averages, and we're going to call those statistics-- i.e. statistics, as in used in the texture model.

So let's do that. So here's the V2 receptive field. Open that up. Inside of that is a bunch of V1 cells, here all shown at the same orientation. In reality, they would be all different orientations and different sizes. And now we're going to compute those joint statistics, just like I did in the

texture model, and that's going to give us our responses.

We're going to have to do that for each one of these receptive fields. So there's a lot of these. It's not 700 numbers anymore. It's reduced, because it's-- so there's details here. It's reduced, but there's a lot of these, so it's quite a lot of parameters.

And these local correlations that I told you we were going to compute here can be re-expressed, actually, in a form that looks just like the simple and complex cell calculations that I showed you for V1. So in fact, if you take these V1 cells, and you take weighted sums of these guys, and you half-wave rectify them and add them, you get something that's essentially equivalent to the texture model that I told you about. So that's pretty cool, because it means that the calculations that are taking us from the LGN input to V1 outputs have a form, a structure which is then repeated when we get to V2.

We do the same kind of calculations-- linear filters, rectification, pooling or averaging. And so that, of course, has become ubiquitous with the advent of all the deep network stuff. But the idea here is that we can actually do this kind of canonical computation again and again and again and produce something that replicates the loss of information and the extraction of features or parameters that the human visual system is performing.

So this canonical idea, I think, is important, and it's something that we've been thinking about for a long time-- linear filtering that determines pattern selectivity, some sort of rectifying non-linearity, some sort of pooling. And we usually also include some sort of local gain control, which seems to be ubiquitous throughout the visual system and the auditory system in every stage, and noise, as well. And we're currently, in my lab, working on lots of models that are trying to incorporate all of these things in stacked networks-- small numbers of layers, not deep-- shallow, shallow networks for us-- in order to try to understand their implications for perception and physiology.

This was just a description of a single stage, and then, of course, you have to stack them. And there are many people that have talked about that idea. This is a figure from Tommy's paper with Christof, I think-- 1999. And Fukushima had proposed a basic architecture like this earlier. And so I think this has now become-- you barely even need to say it, because of the deep network literature.

So how do we do this? Same thing I told you before. Take an image, plop down all these V2 receptive fields. By the way, I should have said this at the outset-- this is drawn as a cartoon.

The actual receptive fields that we use are smooth and overlapping, so that there are no holes. And in fact, the details of that are that since we're computing averages, you can think of this as a low pass filter, and we try to at least approximately obey the Nyquist theorem, so that there's no aliasing-- that is, there's no evidence of the sampling lattice, for those of you that are thinking down those lines. If you were not thinking down those lines, I'll just say the simple thing, which is that they're not little disks that are non-overlapping, because then we would be screwing everything up in between them. They're smooth and overlapping so that we cover the whole image, and all the pixels in the image are going to be affected by this process.

So we make all those measurements. It's a very large set of measurements. And now we start with white noise, and we push the button. And again, push simultaneously on the gradients from all those little regions until we achieve something that matches all the measurements in all of those receptive fields.

The measurements in the receptive fields are averaged over different regions. So the ones that are in the far periphery are averaged over large regions, and so those averages are throwing away a lot more information. The ones that are averaged near the fovea are throwing away a small amount of information. When you get close enough to the fovea, they're throwing away nothing. So the original image is preserved in the center, and then it gets more and more distorted as you go away from the fovea.

So the question is, does that work for a human? Is it metameric? The display here is not very good, but I'll try to give you a demonstration of it to convince you that it does work. You have to keep your eyes planted here, and I'm going to flip back and forth between this original picture, which was taken in Washington Square Park, near the department. And I'm going to flip between this and a synthesized version. You have to keep your eyes here, at least for a bunch of flips. Hello.

Here we go. Keep your eyes fixated. Those two images should look the same. It's going back and forth, A, B, A, B, and they should look the same. I think for most of you, and for most of these viewing distances, it should work.

And now if you look over here, you'll see that they actually are not the same. That's about the size of a V2 receptive field, and it is the same two images. I'm not cheating here, in case anybody's worried. I'm just flipping back and forth between the same two images.

And you can see that the original image has a couple of faces in that circle, but the synthesized one, they're all distorted, the same way Feynman was when I showed you his photograph. But again, the point here is that these two are not metamers when you look right at this peripheral region, but when you keep your eyes fixated here, they're pretty hard to distinguish. This is right at about the threshold for the subjects that we ran in this experiment, so it should be basically imperceptible to you.

That was a demo, just to convince you that it seems to work. We did an experiment, because we wanted to do more than just show that it sort of works. We wanted to figure out whether we could actually tie it to the physiology in a more direct way, so what we did is we generated stimuli where we used different receptive field size scaling.

So this is a plot. Along this axis is going to be-- just to get you situated, along this axis is going to be models that are used to generate stimuli with different receptive field size scaling. That's the ratio of diameter to eccentricity-- diameter of the receptive field to the eccentricity distance from the fovea. And along here is going to be the percent correct that a human is able to correctly identify-- the way we did this, it's called an ABX experiment. So we show one image, then we show another image, then we show a third image. And we say, which image does the third one look like?

So we're going to plot percent correct here. And if we use a model with very small receptive fields, then we get syntheses that look like this. This one has very little distortion. There's a little bit of distortion around near the edges, but it's pretty close to the original.

If we use really big receptor fields, then we get a lot of distortion. Things really start to fall apart. And somewhere in between-- so far to the right on this plot, we expect people to be at 100% noticing the distortions, and far to the left on this plot, we expect them to be at chance. We expect them to not be able to tell the difference.

And that's exactly what happens. This is an average over four observers. And you can see that the performance, the percent correct starts at around 50%, and then climbs up and asymptotes.

So what's more, we can now do something-- this is a little bit complicated to get your head around. We're using this model to generate the stimuli, and this is the model parameter plotted along this axis. Now we're going to use the model again, but now we're going to use the model as a model for the observer.

So there's two models here. One is generating the stimuli. The other one, we're going to try to fit-- we're going to ask, if we used a second copy of the model to actually look at these images and tell the difference between them, what would its receptive fields have to be in order to match the human data?

And I'm not going to drag you through the details, but the basic idea is that allows us to produce a prediction-- this black line-- for how this model would behave if it were acting as an observer. And by adjusting the parameter of the observer model, we can estimate the size of the human receptive fields. So the end result of all of this is we're going to fit a curve to the data, and it's going to give us an estimate of the size of the receptive fields that the human is using to do this task.

And that is right here. In fact, it's right at the place where the curve hits the 50% line. That's the point where the human can't tell the difference anymore, and that's the point where we think an observer would be-- where the receptive fields of the stimulus would be the same size as the receptive fields of the observer. So that's what we're looking for.

And when we do that for our four observers, they come out very consistent. So here's a plot of the estimated receptive field sizes of these observers. All four of them-- 1, 2, 3, 4, and the average over the four.

And nicely enough-- remember, I told you that we know something about the receptive field sizes in-- these are macaque monkey. And if we plot those on the same plot, these color bands are the size of the receptive fields in a macaque, now combined over this large set of data from a whole bunch of different papers. Jeremy went through incredibly painstaking work to try to put these all into the same coordinate system and unify the data sets.

And so the height of each of these bars tells you-- they're error bars on how much variability there is, where we think the estimates are. And you can see that the answers for the humans are coming right down on top of V2. So we really do think that the information that is being lost in these stimuli is being lost in V2, and it seems to match the receptive field sizes at least of macaque monkey.

We were worried that this might depend a lot on the details of the experiment. So for example, we thought, well, what if we give people a little more information? For example, what if we let them look at the stimulus longer?

So the original experiment was pretty brief-- 200 milliseconds. What if we give them 400 milliseconds? And so up here are plots for the same four subjects. The original task is in the dark gray, and you can see the curves for each of the subjects.

When we give them more time, what you notice is that, in general, they do better. So generally, the light gray curves-- 1, 2, 3-- are above the dark gray curves. They get higher percent correct.

But the important thing is that each of these curves dives down and hits the 50% point at the same place. In other words, what we interpret this to mean is that the estimate of the receptive field sizes is an architectural constraint, and we can estimate the same architectural constraint under both of these conditions, even though performance is noticeably different, at least for these three subjects.

This one, it's really quite a big, big improvement. This subject is doing much, much better on the task when we give them more time. And yet, this estimate of receptive field sizes is pretty stable, so we thought this was a pretty important control.

And down below is another control. That was a bottom-up control. This is a top-down control. People have talked about attention being very important in peripheral tasks, so we now gave the subjects an attentional cue-- a little arrow at the center of the display that pointed toward the region of the periphery where the distortion was largest in a mean-squared error sense.

So we measure little chunks of the peripheral image and look for the place where there's the biggest difference, and we tell them to pay attention to that part of the stimulus. They're not allowed to move their eyes. We have an eye tracker on them the whole time, so they're not allowed to look at it. But we're telling them, try to pay attention to what's, let's say, in the upper left.

And again, the result is quite similar. Their performance improves noticeably, at least for these three subjects. This one, again, is the most dramatic performance improvement. Nobody gets worse. This subject basically stayed about the same.

But again, the estimates of receptive field size are quite stable. So our interpretation is attention is boosting the signal, if there is a signal, that allows them to do the task. But if they're at chance and there's no signal, attention does nothing, which is why that when you get to 50%, all these points coalesce. All the curves are hitting 50% at the same place.

One last control-- we wanted to convince ourselves that really it was V2, and it wasn't just luck that we happened to get that receptive field size that matched the macaque data. So we did a control experiment where we tried to get the same result for V1. So this time, we just measure local oriented receptive fields like Hubel and Wiesel described, and we average them as in a complex cell over different sized regions.

And we generate stimuli that are just matched for the average responses of the V1 cells. We don't do all the statistics on top of that that represents the V2 calculation. We're just doing average V1 responses.

When we do that-- we generate the stimuli, we do the same experiment, we get a very different result in light gray here. So you can see that these curves are always higher than the other ones, but they also hit the axis at a much, much smaller value, usually by about a factor of two, which is just right, given what I told you before about receptive field sizes.

So if we go back and we combine all the data on one plot-- down here are the V1 controls. They're about the right size for V1. And up here is the original experiment and the two controls that I told you about-- the extended presentation and the directed attention, and those are all pretty much lying in the range of V2.

We think this has a pretty strong implication for reading speed. When you read, your eyes hop across the page. You do not scan continuously. You hop. And when you hop, here's an example of the kind of hops you do when you're reading. There's an eye position, and the typical hop distance would be about that-- from here to there.

This is the same piece of text. We've synthesized it as a metamer using this model, just to illustrate the idea that the chunk of stuff that you can read around that fixation point, it's about right. It matches what you would expect for the kind of hopping that you could do.

Your reading speed is limited by the distance of those hops, and the distance of those hops is limited by this loss of information. So you can't read anything beyond maybe this I and this N. And in order to read it, you hop your eyes over here, and now you get most of this word. You can make out the rest of an "involuntarily."

So there's an interesting implication here, which is that you can potentially increase reading speed by using this model to optimize the presentation of text. And now that we can do these

things electronically, you can imagine all kinds of devices where the word spacing and the line spacing and the letter sizes and everything else could change with time and position on the display. So you don't have to just put things out as static arrays of characters. You could now imagine jumping things around and rescaling things. You could imagine designing new fonts that caused less distortion or loss of information, et cetera.

So this is just going back to the trichromacy story that I told you. I told you that once they figured out the theory, and they had all the psychophysics down, the next thing that happened is all that engineering. They came up with engineering standards, and they used it to design devices and specify protocols for transmitting images, for communicating them, for rendering them. I think that this has that kind of potential. And this theory is too crude right now, but if you had a really solid theory for what information survived in the periphery, you can really start to push hard on designing devices and designing specifications for devices for improved whatever.

Sometimes you want to improve things. Sometimes you want to make things harder to see, like in this example. So you want to build camouflage. You go in, you take a bunch of photographs of the environment, and then you say, let's design a camouflage that best hides itself when it's not seen directly within this environment. So you could use these kinds of loss of information to exploit things or to aid things in terms of human perception.

So let me say just a few things about V2, and then maybe I should stop. So this work that Jeremy and I did in building this model for metamers, which is a global version of the texture model that operates in local regions, led us to start asking questions about what we could learn by actually measuring cells in V2. And we joined forces with Tony Movshon, who is the chair of my department and a longtime collaborator and friend. And we started a series of experiments to try to explore presentations of texture to V2 neurons to try to understand what we could learn about the actual representations of V2. And these are all done in macaque monkey.

And I should also mention that V2 is-- it's been studied for a long time. Hubel and Wiesel wrote a very important paper about V2 in 1965, which was quite beautiful, documenting the properties that they could find. But the thing that's interesting about this is that V1 didn't really crack until Hubel and Wiesel figured out what the magic ingredient was. And the magic ingredient was orientation.

Before Hubel and Wiesel, people have been poking at primary visual cortex, showing little spots of light and little annuli-- all the things that worked really well in the retina and the LGN, and they were not getting very interesting results. They were saying, well, the receptive fields are bigger and there are hot spots, positive and negative regions, but the cells are not responding that well.

And when Hubel and Wiesel figured out that orientation was the magic ingredient-- and the apocryphal story is that they did that late at night, and they figured it out when they were putting a slide into the projector, and they had forgotten to cover the cat's eyes. And they put the slide into the projector, and the line at the edge of the slide went past on the screen--

TOMMY: it was broken.

EERO SIMONCELLI: It was broken. Ah, I always thought it was the edge of the slide. I've fibbed, and Tommy has corrected me that it was something broken in the slide.

But in any case, the point is that a boundary went by, and they heard-- so they played the spikes through a loudspeaker. This is what most physiologists did in those days, and even still a lot do. Certainly, in Tony's lab you can always walk in there and hear the spikes coming over the loudspeaker.

Anyway, they heard this huge barrage of spikes, more than they had ever heard from any cell that they had recorded from, and that was the beginning of a whole sequence of just fabulous work. And using that tool-- very simple and very obvious in retrospect, but absolutely critical for the progress. The stimuli matter is the point, and making the jump to the right stimuli changes everything.

So V2 for the last 40 years has been sitting in this difficult state where people keep throwing stimuli at it. They try angles. They try curves. They try swirly things. They try corners. They try contours of various kinds, illusory contours.

And throughout all of this, the end story is V2 cells have bigger receptive fields, many of them respond to orientation, some of them respond to particular combinations of orientation, but it's usually a small subset, and the responses are weak. And that's really what the literature has looked like for 40 years. So what we were after is, can we drive these cells convincingly and in a way that we can document is significantly different than what we see in V1? That was the goal-- find a way to drive most of the cells and to drive them differently than what one would

expect in V1.

As a starting point, we succeeded with textures. So basically, we took a bunch of textures. Here are some example textures drawn from the model.

Down below are spectrally-matched equivalents. So these things have the same power spectra, the same amount of energy and different orientation and frequency bands, but they lack all the higher-order statistics that are coming in this texture model that give you nice, clean edges and contours and object-y things, or lumps of objects.

And sure enough-- so here's some example cells. Here's three V1 cells. Here's three V2 cells. And in each of these plots, there's two curves. These are shown over time.

The stimulus is presented here for 100 milliseconds. You see a little bump in the response. And there's a light curve and a dark curve. The light curve is the response to the spectrally-matched noise, and the dark curve is the response to the texture, with the higher-order statistics.

V1 doesn't seem to care is the short answer here, and V2 cares quite significantly. So when you put those higher-order statistics in, almost all V2 cells respond significantly more, and you can see that in these three examples. These are not unusual. That's what most of the cells look like.

So here's a plot, just showing you 63% of the V2 neurons significantly and positively modulated. And by the way, this is averaged over all the textures that we showed them. And if you pick any individual cell, there's usually a couple of textures that drive it really well, and then a bunch of textures that drive it less well. So this effect could be made stronger if you chose only the textures that drove the cell well.

And up here is V1, where you can see that very few of them are modulated by the existence of these higher-order statistics. Oh, here it is across texture category. So now on the horizontal axis is the texture category-- 15 different textures, and you can see, again, that V1 is pretty much very close to the same responses-- dark and light, again, for the spectrally-matched and the higher-order.

And for these three V1 cells, they're basically the same responses for each of these pairs. And for the V2 cells, there are always at least some textures where there's an extreme difference. So this is a really good example. There's a huge difference in response here for these two

textures, but for actually many of the other textures, there's not much of a difference. So sort of a success.

And the last thing I was going to tell you about is that we think-- so this is really fitting, given what Jim DiCarlo told you about, or what I assume he told you about-- this idea of tolerance or invariance versus selectivity. We wanted to know, how can we take what we know about these V2 cells and pull it back into the perceptual domain? How can we ask, what is it that you could do with a population of V2 cells that you couldn't do with a population of V1 cells?

And the thought was if the V2 cells are responding to these texture statistics, then if I made a whole bunch of samples of the same texture, the V2 cells should be really good at identifying which texture that is-- which family it came from. And the V1 cells will be all confused by the fact that those samples each have different details that are shifting around. So the V1 cells will respond to those details, and they'll give a huge variety of responses invariant to re-sampling from that family, and the V2 cells will be more invariant or more tolerant to re-sampling from that family. That was the concept.

And that turns out to be the case, so let me show you the evidence. So here's four different textures, four different-- what we call different families. Here's images of three different examples drawn from each.

So these are just three samples drawn, starting with different white noise seeds. And you can see that they're actually physically different images, but they look the same. Three again. Three again.

And so we got 100 cells from V1 and about 100 cells from V2. The stimuli are presented for 100 milliseconds. We do 20 repetitions each. We need a lot of data.

And what's shown here are just these this 4 by 3 array, but we actually had 15 different families and 15 examples of each. 20 repetitions of each of those. 225 stimuli times 20 repetitions. That's the experiment.

So what we wanted to know is, does the hypothesis hold? And so here's an example. These are responses laid out for these 12 stimuli.

And what you can see is that this is a V1 neuron-- a typical V1 neuron. You can see that the neuron actually responds with a fair amount of variety in these columns. That is, for different

exemplars from the same family, there's some variety. High response here, medium response here, very low response here. And this is for these three images, which to us look basically the same.

So this cell would not be very good at separating out or recognizing or helping in the process of recognizing which kind of texture you were looking at, because it's flopping all over the place when we draw different samples. That, as compared to having to V2 cell-- this is a typical V2 cell, which you can see is much more stable across these columns. This is roughly the same response here, roughly the same here, little bit of variety in this one, roughly the same in this one.

And sure enough, if you actually go and plot this, V2 has much higher variance across families. That's vertically. These are the V1 cells. These are the V2 cells.

And this is the variance across families. This is the variance across exemplars. V2 has higher variance typically across families, and V1 has higher variance across exemplars.

And now if you take the populations of equal size-- 100 of each, and you ask well, how good would I be at taking that population and identifying which family, which kind of texture I'm looking at? And we do this with cross-validation and everything. I can give you the details later, if you want to know.

We find V2 is always better than V1 in doing this task. So we can do a better job in performing this task-- identifying which of these families a given example was drawn from if we look at V2 than if we look at V1. And if we flip that around and we try to do exemplar identification, with 15 different examples of a given family-- if we say, which one was it? It turns out that V1 is better than V2 for that. So we think of this as evidence that V2 has some invariance across these samples, whereas V1 is much more specialized for the particular samples.

This work started with this fantastic post-doc that I had mentioned earlier, Javier Portilla. Jeremy Freeman came into my lab, and we just jumped all over this in making the metamers. Josh McDermott is on here because I usually also play the auditory examples and walk through a little bit of that work, but I'm going to leave that for him. And Corey Ziemba, who's a student who's in the lab right now and is doing a lot of the physiology and did a lot of the physiology that I showed you in Tony's lab. And we were funded by HHMI and also the NIH. So thanks.