JOSH TENENBAUM: We're going to-- I'm just going to give a bunch of examples of things that we in our field have done. Most of them are things that I've played some role in. Maybe it was a thesis project of a student. But they're meant to be representative of a broader set of things that many people have been working on developing this toolkit.

And we're going to start from the beginning in a sense-- just some very simple things that we did to try to look at ways in which probabilistic, generative models can inform people's basic cognitive processes. And then build up to more interestingly kinds of symbolically structured models, hierarchical models, and ultimately to these probabilistic programs for common sense. So when I say a lot of people have been doing this, I mean here's just a small number of these people. Every year or two, I try to update this slide.

But it's very much historically dated with people that I knew when I was in grad school basically. There's a lot of really great work by younger people who maybe their names haven't appeared on this slide. So those dot dot dots are extremely serious. And a lot of the best stuff is not included on here.

But in the last couple of decades, across basically all the different areas of cognitive science that cover basically all the different things that cognition does, there's been great progress building serious mathematical-- and what we could call reverse engineering models in the sense that they are quantitative models of human cognition, but they are phrased in the terms of engineering, the same things you would use to build a robot to do these things, at least in principle. And it's been developing this toolkit of probabalistic generative models.

I want to start off by telling you a little bit about some work that I did together with Tom Griffiths. So Tom is now a senior faculty member at Berkeley, one of the leaders in this field-- as well as a leading person in machine learning, actually. One of the great things that he's done is to take inspiration from human learning and develop fundamentally new kinds of probabilistic models, in non-parametric Bayes in particular, inspired by human learning.

But when Tom was a grad student, we worked together. He was my first student. We're almost the same age. So at this point, we're more like senior colleagues than student advisor. But I'll tell you about some work we did back when he was a student, and I was just starting off.

And we were both together trying to tackle this problem and trying to see, OK, what are the prospects for understanding even very basic cognitive intuitions, like senses of similarity or the most basic kinds of causal discovery intuitions like we were talking about before, using some kind of idea of probabilistic inference in a generative model? And at the time-- remember in the introduction I was talking about how there's been this back and forth discourse over the decades of people saying, yeah, rah rah, statistics, and, statistics, those are trivial and uninteresting?

And at the time we started to do this, at least in cognitive psychology, the idea that cognition could be seen as some kind of sophisticated statistical inference was very much not a popular idea. But we thought that it was fundamentally right in some ways. And it was at the time-- again, this was work we were doing in the early 2000s when it was very clear in machine learning and AI already how transformative these ideas were in building intelligent machines or starting to build intelligent machines.

So it seemed clear to us that at least it was a good hypothesis worth exploring and taking much more seriously than psychologists had much before that. That this also could describe basic aspects of human thinking. So I'll give you a couple examples of what we did here.

Here's a simple kind of causal inference from coincidences, much like what you saw going on in the video game. There's no time in this. It's really mostly just space, or maybe a little bit of time. The motivation was not a video game, but imagine-- to put a real world context on it-- what's sometimes called cancer clusters or rare disease clusters. You can read about these often in the newspaper, where somebody has seen some evidence suggestive of some maybe hidden environmental cause-- maybe it's a toxic chemical leak or something-- that seems to be responsible for-- or maybe they don't have a cause.

They just see a suspicious coincidence of some very rare disease, a few cases that seem surprisingly clustered in space and time. So for example, let's say this is one square mile of a city. And each dot represents one case of some very rare disease that occurred in the span of a year.

And you look at this. And you might think that, well, it doesn't look like those dots are completely, uniformly, randomly distributed over there. Maybe there's some weird thing going on in the upper left or northwest corner-- some who knows what-- making people sick. So let me just ask you. On a scale of 0 to 10, where 10 means you're sure there's some kind of thing going on and some special cause in some part of this map.

And 0 means no, you're quite sure there's nothing going on. It's just random. What do you say? To what extent does this give evidence for some hidden cause? So give me a number between 0 and 10.

AUDIENCE: 5.

JOSH TENENBAUM: OK, great. 5, 2, 7. I heard a few examples of each of those. Perfect. That's exactly what people do. You could do the same thing on Mechanical Turk, and get 10 times as much data, and pay a lot more. It would be the same. I'll show you the data in a second.

But here's the model that we built. So again, this model is a very simple kind of generative model of a hidden cause that various people in statistics have worked with for a while. We're basically modeling a hidden cause as a mixture. Or I mean it's a generative model, so we have to model the whole data.

When we say there's a hidden cause, we don't necessarily mean that everything is caused by this. It's just that the data we see in this picture is a mixture of whatever the normal random thing going on is plus possibly some spatially localized cause that has some unknown position, unknown extent. Maybe it's a very big region. And some unknown intensity-- maybe it's causing a lot of cases or not that many.

The hypothesis space maybe is best visualized like this. Each of these squares is a different hypothesis of a mixture density or a mixture model-- which is a mixture of just whatever the normal uniform process is that causes a disease unrelated to space and then some kind of just Gaussian bump, which can vary in location, size, and intensity, that is the possible hidden cause of some of these cases.

And what the model that we propose says is that your sense of this spatial coincidence-- like when you look at a pattern of dots really and you see, oh, that looks like there's a hidden cluster there somewhere. It's basically you're trying to see whether something like one of those things on the right is going on as opposed to the null hypothesis of just pure randomness. So we take this log likelihood ratio, or log probability, where we're comparing the probability of the data under the hypothesis that there's some interesting hidden cause, one of the things on the right, versus the alternative hypothesis that it's just random, which is just the simple, completely uniform density.

And what makes this a little bit interesting computationally is that there's an infinite number of these possibilities on the right. There's an infinite number of different locations and sizes and intensities of the Gaussian. And you have to integrate over all of them. So again, there's not going to be a whole lot of mathematical details here. But you can read about this stuff if you want to read these papers that we had here.

But for those of you who are familiar with this, with working with latent variable models, effectively what you're doing is just integrating either analytically or in a simulation over all the possible models and sort of trying to compute on average how much does the evidence support something like what you see on the right, one of those cluster possibilities, versus just uniform density. And now what I'm showing you is that model compared to people's judgments on an experiment. So in this experiment, we showed people patterns like the one you just saw.

The one you saw is this one here. But in the different stimuli, we varied parameters that we thought would be relevant. So we varied how many points were there total, how strong the cluster was in various ways, whether it was very tightly clustered or very big, the relative number of points in the cluster versus not. So what you can see here, for example, is it's a very similar kind of geometry, except here this is a sort of biggish cluster. And then we're making basically there's four points that look clustered and two that aren't. And in these cases, we just make the four points more tightly clustered.

Here, what we're doing is we're going from having no points that look clustered to having almost all of the points looking clustered and just varying the ratio of clustered points to non-clustered points. Here, we're just changing the overall number. So notice that this one is basically the same as this one.

So again, at both of these, we've got four clustered points and two seemingly non-clustered ones. And here we just scale up in set-- or scale up from four to two, to eight and four. And here we scale it down to two and one, and various other manipulations.

And what you can see is that they have various systematic effects on people's judgments. So what I'm calling the data there is the average of about 150 people who did the same judgment you did-- 0 to 10. What you can see is the one I gave you was this one here. And the average judgment was almost exactly five.

And if you look at the variance, it looks just like what you saw here. Some people say two or three. Some people say seven. I chose one that was right in the middle.

The interesting thing is that, while you maybe felt like you were guessing-- and if you just listened to what everyone else was saying, maybe it sounds like we're just shouting out random numbers-- that's not what you're doing. On that one, it looks like it, because it's right on the threshold. But if you look over all these different patterns, what you see is that sometimes people give much higher numbers than others. Sometimes people give much lower number than others.

And the details, that variation, both within these different manipulations we did and across them, are almost perfectly captured by this very simple probabilistic generative model for a latent cause. So the model here is-- this is the predictions that model I showed you is making, where again, basically, a high bar means there's strong evidence in favor of the hidden latent cause hypothesis. Some, one, or more-- some cluster-- that low bar means strong evidence for the alternative hypothesis.

The scale is a bit arbitrary. And it's a log probability ratio scale. So I'm not going to comment on the scale. But importantly, it's the same scale across all of these. So a big difference is, it's the same big difference in both cases. And I don't think this is fairly good evidence that this model is capturing your sense of spatial coincidence and showing that it's not just random or arbitrary, but it's actually a very rational measure of how much evidence there is in the data for a hidden cause.

Here's the same model now applied to a different data set that we actually collected a few years before, which just varies the same kinds of parameters, but has a lot more points. And the same model works in those cases, too. The differences are a little more subtle with these more points.

So I'll give you one other example of this sort of thing. Like the one I just showed you, we're taking a fairly simple statistical model. This one, as you'll see, isn't even really causal. This one at least, that I showed you, is causal. The advantage of this other one is that it's both a kind of textbook statistics example, it's one where people do something more interesting than what's in the textbook. Although you can extend the textbook analysis to make it look like what people do.

And unlike in this case here, you can actually measure the empirical statistic. You can go out, and instead of just like positing, here's a simple model of what a latent environmental cause would be like, you can actually go and measure all the relevant probability distributions and compare people not just with a notional model, but with what, in some stronger sense, is the rational thing to do, if you were doing some kind of intuitive Bayesian inference.

So these are, again stuff that Tom Griffiths did with me, in an in and then after grad school. We asked people to make the following kind of everyday prediction. So we said, suppose you read about a movie that's made $60 million to date. How much money will it make in total?

Or you see that something's been baking in the oven for 34 minutes. How long until it's ready? You meet someone who's 78 years old. How long will they live? Your friend quotes to you from line 17 of his favorite poem. How long is the poem? Or you meet a US congressman who has served for 11 years. How long will he serve in total?

So in each of these cases, you're encountering some phenomenon or event in the world with some unknown total duration. We'll call that t, total. And all we know is that t, total, is somewhere between zero and infinity.

We might have a prior on it, as you'll see in a second. But we don't know very much about this particular t, total except you get one example, one piece of data, some t, which we'll just assume is just randomly sampled between zero and t, total. So all we know is that whatever these things are, it's something randomly chosen, less than the total extent or duration of these events.

And now we can ask, what can you guess about the total extent or duration from that one observation? Or in mathematical terms, there is some unknown interval from zero up to some maximal value. You can put a prior on what that interval is. And you have to guess the interval from one sampled point sampled randomly within it. It's also very similar-- and another reason we studied this-- to the problem of learning a concept from one example.

When you're learning what horses are from one example, or when you're learning what that piece of rock climbing question is-- what's a cam-- from one example, or what's a tufa from one example. You can think, there's some region in the space of all possible objects or something, or some set out there. And you get one or a few sample points, and you have to figure out the extent of the region. It's basically the same kind of problem, mathematically.

But what's cool about this is we can measure the priors for these different classes of events and compare people with an optimal Bayesian inference. And you see something kind of striking. So here's, on the top-- I'm showing two different kinds of data here. On the top are just empirical statistics of events you can measure in the world; nothing behavioral, nothing about cognition. On the bottom, I'm showing some behavioral data and comparing it with model predictions that are based on the statistics that are measured on top.

So what we have in each column is one of these classes of events, like movie grosses in dollars. You can get this data from iMDB, the Internet Movie Database. You can see that most movies make $100 million or less. There's sort of a power law. But a few movies make hundreds, or even many hundreds, maybe a billion dollars even, these days.

Similarly with poems, they have a power law distribution of length. So most poems are pretty short. They fit on a page or less. But there are some epic poems, or some multi-page-- many, many hundreds of lines. And they fall off with a long tail.

Lifespans, movie runtimes are kind of unimodal, almost Gaussian-- not exactly. Those red curves, histograms' bars, show the empirical statistics that we measured from public data. And the red curves show just the best fit of a simple parametric model, like a Gaussian or a power law distribution that I'm mentioning.

House of representatives-- how long people serve in the House has this kind of gamma, or particular gamma called an Erlang shape with a little bit of an incumbent effect. Cake baking times-- so remember we asked how long is this cake going to bake for. They don't have any simple parametric form when you go in and look at cookbooks. But you see, there's something systematic there. There's a lot of things that are supposed to bake for exactly an hour. There are some which have a smaller, or a shorter, but broad mode. And then there's a few epic 90-minute cakes out there.

So that's all the empirical statistics. Now what you're seeing on the bottom is people's-- well, on the y-axis, the vertical axis, you have the average-- it's a median-- of a bunch of human predictions for the total extent of any one of these things, like your guess of the total length of a poem given that, basically, there is a line 17 in it. And on the x-axis, what you're seeing is that one data point, the one value of t, which is, all you know is that it's somewhere between zero and t, total.

So different groups of subjects were given five different values. So you see five black dots, which correspond to what five different subgroups of subjects said for each of these possible t values. And then the black and red curves are the model fit, which comes from taking a certain kind of Bayesian optimal prediction, where the prior is what's specified on the top-- that's the prior on t, total. The likelihood is a sort of uniform random density.

So it's just saying t is just a uniform random sample from zero up to t, total. You put those together to compute a posterior. And then you-- the particular estimator we're using is what's called the posterior median. So we're looking at the median of the exterior and comparing that with a median of human subjects. And what you can see is that it's almost a perfect fit. And it doesn't really matter whether you take the red curve, which is what comes from approximating the prior with one of these simple parametric models, or the black one, which comes from just taking the empirical histogram. Although, for the cake baking times, you really can only go for the empirical one. Because there is no simple parametric one. That's why you just see a jagged black line there.

But it's interesting that it's almost a perfect fit. There are a couple-- just like somebody asked in Demis's talk-- there's one or two cases we found where this model doesn't work, sometimes dramatically, and sometimes a little bit. And they're all interesting. But I have time to talk about it. That's one of the things I decided to skip. If you'd like to talk about it, I'm happy to do that. But most of the time, in most of the cases we've studied, these are representative.

And I think, again, all of the failure cases are quite interesting ones. That point to, this is one of the many things we need to go beyond. But the interesting thing isn't just that the curves fit the data, but the fact that the actual shape is different in each case. Depending on the prior of this different classes of events, you get a fundamentally different, or qualitatively different, prediction function. Sometimes it's linear. Sometimes it's non-linear. Sometimes it has some weird shape.

And really, quite surprisingly to us, people seem to be sensitive to that. So they seem to predict in ways that are reflective of not only the optimal Bayesian thing to do, but the optimal Bayesian thing to do from the optimal prior, from the correct prior. And I certainly don't want to suggest that people always do this. But it was very interesting to us that for just a bunch of everyday events, and really, the places where this analysis works best are ones, again, where we think people actually might plausibly have good reasons to have the relevant experiences with these everyday events, they seem to be sensitive to both the statistics in the sense of just what's going on in the world and doing the right statistical prediction.

So that's what we did. 10 years ago or so, that was like the state of the art for us. And then we wanted to know, well, OK, can we take these sorts of ideas and scale them up to some actually interesting cognitive problems, like say, for example, learning words for object categories. And we did some of that. I'll show you a little bit of that before showing you what I think was missing there.

I mean, in a lot of ways, this is a harder problem. I mean, it's very similar, as I said. It's basically like, there's just like the problem I just showed you, where there was an unknown total extent or duration, and you got one random sample from it, here there is some un-- imagine the space of all possible objects-- could be a manifold or described by a bunch of knobs. I mean, these are all generated from some computer program. If these were real, biological things, they would be generated from DNA or whatever it is.

But there's some huge, maybe interestingly structured, space of all possible objects. And within that space is some subset, some region or subset, somehow described that is the set of tufas. And somehow you're able to grasp that subset, more or less, if you get its boundaries, to be able to say yes or no as you did at the beginning of the lecture from just, in this case, a few points-- three points-- randomly sampled from somewhere in that region. It would work just as well if I showed you one of them, basically.

So in some sense, it's the same problem. But it's much harder, because here, the space was this one dimensional thing. It was just a number. Whereas here, we don't know what's the dimensionality of the space of objects. We don't know how to describe the regions. Here we knew how to describe the regions. They were just intervals with a lower bound at zero and an upper bound at some unknown thing. And the hypothesis space of possible regions was just all the possible upper bounds of this event duration.

Here we don't know how to describe this space. We don't know how to describe the regions that correspond to object concepts. We don't know how to put a price on those hypotheses. But in some work that we did-- in particular, some work that I did with Fei Xu, who is also a professor at Berkeley. We were colleagues and friends in graduate school. We sort of did what we could at the time.

So we made some guesses about what that hypothesis space-- what that space might be like, what the hypothesis space might be like, how to put some priors, and so, on there. Used exactly the same likelihood, which was just this very simple idea that the observed examples are a uniform random draw from some subset of the world. And you have to figure out what that subset is. And we were able to make some progress.

So what we did was we said, well, like in biology, perhaps-- and if you saw-- how many people saw Surya Ganguli's lecture yesterday morning? Cool. I sort of tailored this for assuming that you probably had seen that. Because there's a lot of similarities, or parallels, which is neat. And it's, again, part of engaging on generative models and neural networks. As you saw him do, you'll get my version of this.

So also, like he mentioned, there are actual processes in the world which generate objects-- something like this. We know about evolution-- produces basically tree-structured groups, which we call species, or genus, or something like that, or just taxa, or something. There's groups of organisms that have a common evolutionary descent. That's the way a biologist might describe it. And we know, these days, a lot about the mechanisms that produce that. Even going back 100 or 200 years, say, to Darwin, we knew something about the mechanisms that produced it, even if we didn't know the genetic details, ideas of something like mutation, variation, natural selection as a kind of mechanistic account, about right up there with Newton and forces.

But anyway, scientists can describe some process that generates trees. And maybe people have some intuition, just like people seem to have some intuitions about these statistics of everyday events, maybe they have some intuitions, somehow, about the causal processes in the world, which give rise to groups and groups and subgroups. And they can use that to set up a hypothesis space.

And the way we went about this is, we have no idea how to describe people's internal mental models of these things, but you can do some simple-- there are simple ways to get this picture by just basically asking people to judge similarity and doing hierarchical clustering. So this is a tree that we built up by just asking people-- getting some subjective similarity metric and then doing hierarchical clustering, which we thought could roughly approximate maybe the internal hierarchy that our mental models impose on this. Were you raising your hand or just-- no. OK. Cool.

We ultimately found this dissatisfying, because we don't really know what the features are. We don't really know if this is the right tree or how people built it up. But it actually worked pretty well, in the sense that we could build up this tree. We could then assume that the hypotheses for concepts just corresponded to branches of the tree. And then you could-- again, to put it just intuitively, the way you do this learning from one or a few examples, let's say that you see those few tufas over there. You're basically asking, which branch of the tree do I think-- those are randomly drawn from some internal branch of the tree, some subtree. Which subtree is it?

And intuitively, if you see those things and you say, well, they are randomly drawn from some branch, maybe it's the one that I've circled. That sounds like a better bet, for example, than this one here, or maybe this one, which would include one of these things, but not the others. So that's probably unlikely. And it's probably a better bet than, say, this branch, or this branch, or these ones, which are logically compatible, but somehow it would have been sort of a suspicious coincidence. If the set of tufas had really been this branch here, or this one here, then it would have been quite a coincidence that the first three examples you saw were all clustered over there in one corner.

And what we showed was that, that kind of model, where that suspicious coincidence came out from the same kinds of things I've just been showing you for the causal clustering example, and for the interval thing, it's the same Bayesian math. But now with this tree-structured hypothesis space, that was actually-- did a pretty good job of capturing people's judgments. We gave people one or a few examples of these concepts that, the examples could be more narrowly or broadly spread, just like you saw in the clustering thing, but just sort of less extensive.

We did this with adults. We did this with kids. And I won't really go into any of the details. but If you're interested, check out these various Xu and Tenenbaum papers. That's the main one there. And you know, the model kind of worked. But ultimately, we found it dissatisfying. Because we couldn't really explain-- we didn't really know what the hypothesis space was. We didn't really know how people were building up this tree.

And so we did a few things. We-- meaning I with some other people-- turned to other problems where we had a better idea, maybe, of the feature space and the hypothesis space, but the same kind of ideas could be explored and developed. And then ultimately-- and I'll show you this maybe before lunch, or maybe after lunch-- we went back and tackled the problem of learning concepts from examples with other cases where we could get a better handle on really knowing what the representations that people were using were, and also where we could compare with machines in much more compelling apples and oranges ways.

In some sense here, there's no machine, as far as I know, that can solve this problem as well as our model. On the other hand, that's, again, it's just very much like the issue that came up when we were talking about-- I guess maybe it was with you, Tyler-- when we were talking about the deep learning-- or with you, Leo-- the deep reinforcement network. A machine that's looking at this just as pixels is missing so much of what we bring to it, which is, we see these things as three-dimensional objects.

And just like the cam in rock climbing, or any of those other examples I gave before, I think that's essential to the abilities that people are doing. The generative model we build, this tree is based not on pixels, or even on ConvNet features, but on a sense of the three-dimensional objects, its parts, and their relations to each other. And so, fundamentally, until we know how to perceive objects better, this is not going to be comparable between humans and machines on equal terms. But I'll show you a little bit later some still pretty quite interesting, but simpler, visual concepts that you can still learn and generalize from one example, but where they are comparable in equal terms.

But first I want to tell you a little bit about these-- yet another cognitive judgment, which like the word learning, or concept learning cases, involved generalizing from a few examples. They also involve using prior knowledge. But they're ones where maybe we have some way of capturing people's prior knowledge by using the right combination of statistical inference on some kind of symbolically structured bottle. So you can already see, as-- I mean, just sort of to show the narrative here.

The examples I was giving here, this doesn't require any symbolic structure. All that stuff I was talking at the beginning, about how we have to combine statistical inference, sophisticated statistical inference, with sophisticated symbolic representations, you don't need any of that here. All the representations could just be counting up numbers or using simple probability distributions that statisticians have worked with for over 100 years. Once we start to go here, now we have to define a model with some interesting structure, like a branching tree structure, and so on.

And as you'll see, we can quickly get to lots more interesting causal, compositionally-structured generative models in similar kinds of tasks. And in particular, we were looking for-- for a few years, we were very interested in these property induction tasks. So this was-- it happened to be-- I mean, I think this was a coincidence. Or maybe we were both influenced by Susan Carey, actually.

So the work that Surya talked about, that he was trying to explain as a theoretician-- remember, Surya and Andrew Saxe, they were trying to give the theory of these neural network models that Jay McClelland and Tim Rogers had built in the early 2000s, around the same time we were doing this work. And they were inspired by some of Susan Carey's work on children's intuitive biology, as well as other people out there in cognitive psychology-- for example, Lance Rips, and Smith, Madine. Many, many cognitive psychologists studied things like this-- Dan Osherson.

They often talked about this as a kind of inductive reasoning, or property induction, where the idea was-- so it might look different from the task I've given you before, but actually, it's deeply related. The task was often presented to people like an argument with premises and a conclusion, kind of like a traditional deductive syllogism, like all men are mortal, Socrates is a man, therefore Socrates is mortal. But these are inductive in that there is no-- you can't conclude with deductive certainty the conclusion follows from the premises or is falsified by the premise, but rather you just make a good guess. The statements above the line provide some, more or less, good or bad evidence for the statement below the line being true.

These studies were often done with-- they could be done with just sort of familiar biological properties, like having hairy legs or being bigger than a breadbox. I mean, it's also-- it's very much the same kind of thing that Tom Mitchell was talking about, as you'll start to see. There's another reason why I wanted to cover this. We worked on these things because we wanted to be able to engage with the same kinds of things that people like Jay McClelland and Tom Mitchell were thinking about, coming from different perspectives. Remember, Tom Mitchell showed you his way of classifying brain representations of semantics with matrices of objects and 20-question-like features that included things like is it hairy, or is it alive, or does it eggs, or is it bigger than a car, or bigger than a breadbox, or whatever.

Any one of these things-- basically, we're getting at the same thing. Here there's just what's-- often these experiments with humans were done with so-called blank predicates, something that sounded vaguely biological, but was basically made up, or that most people wouldn't know much about. Does anyone know anything about T9 hormones? I hope so, because I made it up. But some of them were just done with things that were real, but not known to most people.

So if I tell you that gorillas and seals both have T9 hormones, you might think it's sort of, fairly plausible that horses have T9 hormones, maybe more so than if I hadn't told you anything. Maybe you think that argument is more plausible than the one on the right; given that gorillas and seals have T9 or hormones, that anteaters have hormones. So maybe you think horses are somehow more similar to gorillas and seals than anteaters are. I don't know. Maybe. Maybe a little bit.

If I made that bees-- gorillas and seals have T9 on hormones. Does that make you think it's likely that bees have T9 hormones, or pine trees? The farther the conclusion category gets from the premises, the less plausible it seems. Maybe the one on the lower right also seems not very plausible, or not as plausible. Because if I tell you that gorillas have T9 hormones, chimps, monkeys, and baboons all have T9 on hormones, maybe you think that it's only primates or something. So they're not a very-- it's, again, one of these typicality-suspicious coincidence businesses.

So again, you can think of it as-- you can do these experiments in various ways. I won't really go through the details, but it basically involves giving people a bunch of different sets of examples, just like-- I mean, in some sense, the important thing to get is that abstractly it has the same character of all the other tasks you've seen. You're giving people one or a few examples, which we're going to treat as random draws from some concept, or some region in some larger space.

In this case, the examples are the different premise categories, like gorillas and seals are examples of the concept of having T9 hormones. Or gorillas, chimps, monkeys, and baboons are an example of a concept. We're going to put a prior on possible extents of that concept, and then ask what kind of inferences people make from that prior, to figure out what other things are in that concept. So are horses in that same concept? Or are anteaters? Or are horses in it more or less, depending on the examples you give? And what's the nature of that prior?

And what's good about this is that, kind of like the everyday prediction task-- the lines of the poems, or the movie grosses, or the cake baking-- we can actually sort of go out and measure some features that are plausibly relevant, to set up a plausibly relevant prior, unlike the interesting object cases. But like the interesting object cases, there are some interesting hierarchical and other kinds of causal compositional structures that people seem to be using that we can capture in our models.

So here, again, the kinds of experiments-- these features were generated many years ago by Osherson and colleagues. But it's very similar to the 20 questions game that Tom Mitchell used. And I don't remember if Surya talked about where these features came from, that he talked a lot about a matrix of objects and features. I don't know if he talked about where they come from. But actually, psychologists spent a while coming up with ways to get people to just tell you a bunch of features of animals.

This is, again, it's meant to capture the knowledge that maybe a kid would get from maybe plausibly reading books and going to the zoo. We know that elephants are gray. They're hairless. They have tough skin. They're big. They have a bulbous body shape. They have long legs. These are all mostly relative to other animals. They have a tail. They have tusks. They might be smelly, compared to other animals-- smellier than average is sort of what that means. They walk, as opposed to fly. They're slow, as opposed to fast. They're strong, as opposed to weak. It's that kind of business.

So basically what that gives you is this big matrix. Again, the same kind of thing that you saw in Surya's talk, the same kind of thing that Tom Mitchell is using to help classify things, the same kind of thing that basically everybody in machine learning uses-- a matrix of data with objects, maybe as rows, and features, or attributes, as columns. And the problem here is-- the problem of learning is to say-- the problem of learning and generalizing from one example is to take a new property, which is a new concept, which is like a new column here, to get one or a few examples of that concept, which is basically just filling in one or a few entries in that column, and figure out how to fill in the others, to decide, do you or don't you have that property, somehow building knowledge that you can generalize from your prior experience, which could be captured by, say, all the other features that you know about objects.

So that's the way that you might set up this problem, which again, looks like a lot of other problems of, say, semi-supervised learning or sparse matrix completion. It's a problem in which we can, or at least we thought we could, compare humans and many different algorithms, and even theory, like from Surya's talk. And that seemed very appealing to us.

What we thought, though, that people were doing, which is maybe a little different than what-- or somewhat different-- well, quite different than what Jay McClelland thought people were doing-- maybe a little bit more like what Susan Carey or some of the earlier psychologists thought people were doing-- was something like this. That the way we solve this problem, the way we bridged from our prior experience to new things we wanted to learn was not, say, by just computing the second order of statistics and correlations, and compressing that through some bottleneck hidden layer, but by building a more interesting structured probabilistic model that was, in some form, causal-- in some form-- in some form, compositional and hierarchical-- something kind of like this.

And this is a good example of a hierarchical generative model. There's three layers of structure here. The bottom layer is the observable layer. So the arrows in these generative models point down, often, usually, where the thing on the bottom is the thing you observe, the data of your experience. And then the stuff above it are various levels of structure that your mind is positing to explain it.

So here we have two levels of structure. The level above this is sort of this tree in your head. The idea-- it's like a certain kind of graph structure, where the objects, or the species, are the leaf nodes. And there's some internal nodes corresponding, maybe to higher level taxa, or groups, or something. You might have words for these, too, like mammal, or primate, or animal.

And the idea is that there's some kind of probabilistic model that you can describe, maybe even a causal one on top of that symbolic structure, that tree, that produces the data that's more directly observable, the observable features, including the things you've only sparsely observed and want to fill in. And then you might also have higher levels of structure. Like if you want to explain, how did you learn that tree in the first place, maybe it's because you have some kind of generative model for that generative model.

So here I'm just using words to describe it, but I'll show you some other stuff in a-- or I'll show you something more formal a little bit later. But you could say, well, maybe the way I figure out that there's a tree structure is by having a hypothesis-- the way I figure out that there's that particular tree-structured graphical model of this domain is by having the more general hypothesis that there is some latent hierarchy of species. And I just have to figure out which one it is.

So you could formulate this as a hierarchical inference by saying that what we're calling the form, the form of the model, it's like a hypothesis space of models, which are themselves hypothesis spaces of possible observed patterns of feature correlation. And that, that higher level knowledge, puts some kind of a generative model on these graph structures, where each graph structure then puts a generative model on the data you can observe. And then you could have even higher levels of this sort of thing. And then learning could go on at any or all levels of this hierarchy, higher than the level of experience.

So just to show you a little bit about how this kind of thing works, what we're calling the probability of the data given the structure is actually exactly the same, really, as the model that Surya and Andrew Saxe used. The difference is that we were suggesting-- may be right, may be wrong-- that something like this generative model was actually in your head. Surya presented a very simple abstraction of evolutionary branching process, a kind of diffusion over the tree, where properties could turn on or off. And we built basically that same kind of model. And we said, maybe you have something in your head as a model of, again, the distribution of properties, or features, or attributes over the leaf nodes of the tree.

So if you have this kind of statistical model. If you think that there's something like a tree structure, and properties are produced over the leaf nodes by some kind of switching, on-and-off, mutation-like process, then you can do something like in this picture here. You can take an observe a set of features in that matrix and learn the best tree. You can figure out that thing I'm showing on the top, that structure, which is, in some sense, the best guess of a tree structure-- a latent tree structure-- which if you then define some kind of diffusion mutation process over that tree, would produce with high probability distributions of features like those shown there.

If I gave you a very different tree it would produce other patterns of correlation. And it's just like Surya said, it can be all captured by the second order statistics of feature correlations. The nice thing about this is that now this also gives a distribution on new properties. So if I observe-- because each column is conditionally independent given that model. Each column is an independent sample from that generative model. And the idea is if I observe a new property, and I want to say, well, which other things have this, well, I can make a guess on using that probabilistic model. I can say, all right, given that I know the value of this function over the tree, this stochastic process, at some points, what do I think the most likely values are at other points?

And basically, what you get is, again, like in the diffusion process, a kind of similarity-based generalization with a tree-structured metric, that nearby points in the tree are likely to have the same value. So in particular, things that are near to, say, species one and nine are probably going to have the same property, and others maybe less so. And you build that model. And it's really quite striking how much it matches people's intuitions.

So now you're seeing the kinds of plots I was showing you before, where-- all my data plots look like this. Whenever I'm showing the scatterplot, by default, the y-axis is the average of a bunch of people's judgments, and the x-axis is the model predictions on the same units or scale. And each of these scatterplots is from a different experiment-- not done by us, done by other people, like Osherson and Smith from a couple of decades ago.

But they all sort of have the same kind of form, where each dot is a different set of examples, or a different argument. And what typically varied within an experiment-- you vary the examples. And you fix constant the conclusion category. And you see, basically, how much evidential support to different sets of two or three examples gives to a certain conclusion. And it's really, again, quite striking that-- sometimes in a more categorical way, sometimes in a more graded way-- but basically, people's average judgments here just line up quite well with the sort of Bayesian inference on this tree-structured generative model.

These are just examples of the kinds of stimuli here. Now, we can compare. One of the reasons why we were interested in this was to compare, again, many different approaches. So here I'm going to show you a comparison with just a variant of our approach. It's the same kind of hierarchical Bayesian model, but now the structure isn't a tree, it's a low-dimensional Euclidean space. You can define the same kinds of proximity smoothness thing.

I mean, again, it's more a standard in machine learning. It's related to Gaussian processes. It's much more like neural networks. You could think of this as kind of like a Bayesian version of a bottleneck hidden layer with two dimensions, or a small number of dimensions. The pictures that Surya showed you were all higher dimensions than two dimensions in the latent space, or the hidden variable space, of the neural network, the hidden layer space. But when he compress it down to two dimensions, it looks pretty good.

So it's the same kind of idea. Now what you're saying is you're going to find, not the best tree that explains all these features, but the best two-dimensional space. Maybe it looks like this. Where, again, the probabilistic model says that things which are relatively-- things that are closer in this two-dimensional space are more likely to have the same feature value. So you're basically explaining all the pairwise feature correlations by distance in this space.

It's similar. Importantly it's not as causal and compositional. The tree models something about, possibly, the causal processes of how organisms come to be. If I told you that, oh, there's this-

- that I told you about a subspecies, like whatever-- what's a good example-- different breeds of dogs. Or I told you that, oh, well, there's not just wolves, but there's the gray-tailed wolf and the red-tailed wolf. Red-tailed wolf? I don't know.

Again, they're probably similar, but they might-- one red-tailed wolf, whatever that is, more similar to another red-tailed wolf, probably has more features in common than with a gray-tailed wolf, and probably more to the gray-tailed wolf than to a dog. The nice thing about a tree is I can tell you these things, and you can, in your mind-- maybe you'll never forget that there's a red-tailed wolf. There isn't. I just made it up. But if you ever find yourself thinking about red-tailed wolves and whether their properties are more or less similar to each other than to gray-tailed wolves, or less so to dogs, or so on, it's because I just said some things, and you grew out your tree in your mind. That's a lot harder to do in a low-dimensional space.

And it turns out that, that model also fits this data less well. So here I'm just showing two of those experiments. Some of them are well fit by that model, but others are less well fit. Now, that's not to say that they wouldn't be good for other things. So we also did some experiments. This was experiments that we did. Oh, actually, I forgot to say, really importantly, this was all worked done by Charles Kemp, who's now a professor at CMU. And it was part of the stuff that he did in his PhD thesis.

So we were interested in this as a way, not to study trees, but to study a range of different kinds of structures. And it is true, going back, I guess, to the question you asked, this is what I was referring to about low-dimensional manifolds. There are some kinds of knowledge representations we have which might have a low-dimensional spatial structure, in particular, like mental maps of the world. So our intuitive models of the Earth's surface, and things which might be distributed over the Earth's surface spatially, a two-dimensional map is probably a good one for that.

So here we considered a similar kind of concept learning from a few examples task, where we said-- but now we put it like this. We said, suppose that a certain kind of Native American artifact has been found in sites near city x. How likely is it also to be found in sites near city y? Or we could say sites near city x and y, how about city z.

And we told people that different Native American tribes maybe had-- some lived in a very small area, some lived in a very big area. Some lived in one place, some another place. Some lived here, and then moved there. We just told people very vague things that taps into people's probably badly remembered, and very distorted, versions of American history that would basically suggests that there should be some kind of similar kind of spatial diffusion process, but now in your 2D mental map of cities.

So again, there's no claim that there's any reality to this, or fine-grained reality. But we thought it would sort of roughly correspond to people's internal causal generative models of archeology. Again, I think it says something about the way human intelligence works that none of us are archaeologists, probably, but we still have these ideas.

And it turned out that, here, a spatially structured model actually works a lot better. Again, it shouldn't be surprising. It's just showing that actually, the way-- the judgments people make when they're making inferences from a few examples, just like you saw with the predicting the everyday events, but now in the much more interestingly structured domain, is sensitive to the different kinds of environmental statistics.

There it was different power laws versus Gaussian's of cake bake-- or of movie grosses versus lifetimes or something. Here it's other stuff. It's more interestingly structured kinds of knowledge. But you see the same kind of picture. And we thought that was interesting, and again, suggests some of the ways that we are starting to put these tools together, putting together probabilistic generative models with some kind of interestingly structured knowledge.

Now, again, as you saw from Surya, and as Jay McClellan and Tim Rogers worked on, you can try to capture a lot of this stuff with neural networks. The neat thing about the neural networks that these guys have worked on is that exactly the same neural network can capture this kind of thing, and it can capture this kind of thing. So you can train the very same hidden multilayer neural network with one matrix of object and features. And the very same neural network can predict the tree-structured patterns for animals and their properties, as well as the spatially-structured patterns for Native American artifacts and their cities.

The catch is that it doesn't do either of them that well. It doesn't do as well as the tree-structured models do for peop-- when I say either, it doesn't do that well, I mean, in capturing people's judgments. It doesn't do as well as the best tree-structured models do for people's concepts of animals and their properties. And it doesn't do as well as the best spacial structures. But again, it's in the same spirit as the DeepMind networks for playing lots of Atari games. The idea there is to have the same network solve all these different tasks.

And in some sense, I think that's a good idea. I just think that the architecture should have a more flexible structure. So we would also say, in some sense, the same architecture is solving all these different tasks. It's just that this is one setting of it. And this is another setting of it.

And where they differ is in the kind of structure that-- well, they differ in the fact that they explicitly represent structure in the world. And they explicitly represent different kinds of structure. And they explicitly represent that different kinds of structure are appropriate to different kinds of domains in the world and our intuitions about the causal processes that are at work producing the data. And I think that, again, that's sort of the difference between the pattern classification and the understanding or explaining view of intelligence.

The explanations, of course, go a lot beyond different ways that similarity can be structured. So one of the kind of nice things-- oh, and I guess another-- two other points beyond that. One is that to get the neural networks to do that, you have to train them with a lot of data. Remember, Surya, as Tommy pushed him on in that talk, Surya was very concerned with modeling the dynamics of learning in the sense of the optimization time course, how the weights change over time. But he was usually looking at infinite data. So he was assuming that you had, effectively, an infinite number of columns of any of these matrices. So you could perfectly compute the statistics.

And another important thing about the difference being the neural network models and the ones I was showing you is that, suppose you want to train the model, not on an infinite matrix, but on a small finite one, and maybe one with missing data. It's a lot harder to get the-- the neural network will do a much poorer job capturing the structure than these more structured models. And again, in a way that's familiar with-- have you guys talked about bias-variance dilemma?

So it's that same kind of idea that you probably heard about from Lorenzo. Was it Lorenzo or one of the machin learni-- OK. So it's that same kind of idea, but now applying in this interesting case of structured estimation of generative models for the world, that if you have relatively little data, and sparse data, then having a more structured inductive bi-- having the inductive bias that comes from a more structured representation is going to be much more valuable when you have sparse and noisy data.

The key-- and again, this is something that Charles and I were really interested in-- is we wanted to-- like the DeepMind people, like the connectionists, we wanted to build general purpose semantic cognition, wanted to build general purpose learning and reasoning systems.

And we wanted to somehow figure out how you could have the best of both worlds, how you could have a system that relatively quickly could come to get the right kind of strong constraint-inductive bias in some domain, and a different one for a different domain, yet could learn in a flexible way to capture the different structure in different domains. More on that in a little bit.

But the other thing I wanted to talk about here is just ways in which our mental models, our causal and compositional ones, go beyond just similarity. I guess, since time is short-- well, I was planning to go through this relatively quickly. But anyway, mostly I'll just gesture towards this. And if you're interested, you could read the papers that Charles has, or his thesis. But here, there's a long history of asking people to make these kind of judgments, in which the basis for the judgment isn't something like similarity, but some other kind of causal reasoning.

So for example, consider these things here. Poodles can bite through wire, therefore German shepherds can bite through wire. Is that a strong argument or weak? Compare that with, dobermans can bite through wire, therefore German shepherds can bite through wire. So how many people think that the top argument is a stronger one? How many people think the bottom line is a stronger one?

So that's typical. About twice as many people prefer the top one. Because intuitively-- do I have a little thing that will appear? Intuitively, anyone want to explain why you thought so?

AUDIENCE: Poodles are really small.

JOSH TENENBAUM: Poodles are small or weak. Yes. And German shepherds are big and strong. And what about dobermans?

AUDIENCE: They're just as big as German shepherds.

JOSH TENENBAUM: Yeah. That's right. So they're more similar to German shepherds, because they're both big and strong. But notice that something very different is going on here. It's not about similarity. It's sort of anti-similarity. But it's not just anti-similarity. Suppose I said, German shepherds can bite through wire, therefore poodles can bite through wire. Is that a good argument?

AUDIENCE: No. It's an argument against.

JOSH TENENBAUM: No. It's sort of a terrible argument, right? So there's some kind of asymmetric dimensional reasoning going on. Or similarly, if I said, which of these seems better intuitively; Salmon carry some bacteria, therefore grizzly bears are likely to carry it, versus grizzly bears carry this, therefore salmon are likely to carry it. How many people say salmon, therefore grizzly bears? How many people say grizzly bears, therefore salmon?

How do you know? Those who-- yeah, you're right. I mean, you're right in that's what people say. I don't know if it's right. Again, I made it up. But why did you say that, those of you who said salmon?

AUDIENCE: Bears eat salmon.

JOSH TENENBAUM: Bears eat salmon. Yeah. So assuming that's true, so we're told or see on TV, then yeah. So anyway, these are these different kinds of things that are going on.

And to cut to the chase, what we showed is that you could capture these different patterns of reasoning with, again, the same kind of thing, but different. It's also a hierarchical generative model. It also has, the key level of the hierarchy is some kind of directed graphical structure that generates distribution on observable properties. But it's a fundamentally different kind of structure. It's not just a tree or a space. It might be a different kind of graph and a different kind of process.

So to be a little bit more technical, the things I showed you with the tree and the low-dimensional space, they had a different geometry to the graph, but the same stochastic process operating over it. It was, in both cases, basically a diffusion process. Whereas to get the kinds of reasoning that you saw here, you need a different kind of graph. In one case it's like a chain to capture a dimension of strength or size, say. In the other case, it's some kind of food web thing. It's not a tree. It's that kind of directed network.

But you also need a different process. So the ways-- the kind of probability model to find that out is different. And it's easy to see on the-- for example-- on the reasoning with these threshold things, like the strength properties, if you compare a 1D chain with just symmetric diffusion, you get a much worse fit people's judgments than if you'd used what we called this drift threshold thing, which is basically a way of saying, OK, I don't know. There's some mapping from strength to being able to bite through wire. I don't know exactly what it is. But the higher up you go on one, it's probably more likely that you can bite-- that you can do the other.

So that provides a wonderful model of people's judgments on these kind of tasks. But that sort of diffusion process, like if it was like mutation in biology, then that would provide a very bad model. That's the second row here. Similarly, this sort of directed kind of noisy transmission process on a food web does a great way of modeling people's judgments about diseases, but not a very good way of modeling people's judgments about these biological properties. But the tree models you saw before that do a great job of modeling people's judgments about the properties of animals, they do a lousy job of modeling these disease judgments.

So we have this picture emerging that, at the time, was very satisfying to us. That, hey, we can take this domain of, say, animals and their properties, or the various things we can reason about, and there's a lot of different ways we can reason about just this one domain. And by building these structured probabilistic models with different kinds of graphs structures that capture different kinds of causal processes, we could really describe a lot of different kinds of reasoning. And we saw this as part of a theme that a lot of other people were working on.

So this is-- I mentioned this before, but now I'm just sort of throwing it all out there. A lot of people at the time-- again, this is maybe somewhere between 5 to 10 years ago-- more like six or seven years ago-- we're extremely interested in this general view of common sense reasoning and semantic cognition by basically taking big matrices and boiling them down to some kind of graph structure. In some form, that's what Tom Mitchell was doing, not just in the talk you saw, but remember, he said there's this other stuff he does-- this thing called NELL, the Never Ending Language Learner. I'm showing a little glimpse of that up there from a *New York Times* piece on it in the upper right.

In some ways, in a sort of at least more implicit way, it's what the neural networks that Jay McClelland, Tim Rogers, Surya were talking about do. And we thought-- you know, we had good reason to think that our approach was more like what people were doing than some of these others. But I then came to see-- and this was around the time when CBMM was actually getting started-- that none of these were going to work. Like the whole thing was just not going to work. Liz was one of the main people who convinced me of this. But you could just read the *New York Times* article on Tom Mitchell's piece, and you can see what's missing.

So there's Tom, remember. This was an article from 2010. Just to set the chronology right, that was right around-- a little bit after Charles had finished all that nice work I showed you, which again, I still think is valuable. I think it is capturing something about what's going on. It was very appealing to people, like at Google, because these knowledge graphs are very much like the way, around the same time, Google was starting to try to put more semantics into web search-- again, connected to the work that Tom was doing.

And there was this nice article in the *New York Times* talking about how they built their system by reading the web. But the best part of it was describing one of the mistakes their system made. So let me just show this to you. About knowledge that's obvious to a person, but not to a computer-- again, it's Tom Mitchell himself describing this. And the challenge of, that's where NELL has to be headed, is how to make the things that are obvious to people obvious to computers.

He gives this example of a bug that happened in NELL's early life. The research team noticed that-- oh, let's skip down there. So, a particular example-- when Dr. Mitchell scanned the baked goods category recently, he noticed a clear pattern. NELL was at first quite accurate, easily identifying all kinds of pies, breads, cakes, and cookies as baked goods. But things went awry after NELL's noun phrase classifier decided internet cookies was a baked good.

NELL had read the sentence "I deleted my internet cookies." And again, think of that as, it's kind of like a simple proposition. It's like, OK. But the way it parses that is cookies are things that can be deleted, the same way you can say horses have T9 hormones. It's basically just a matrix. And the concept is internet cookies. And then there's the property of can be deleted, or something like that. And it knows something about natural language processing. So it can see-- and it's trying to be intelligent. Oh, internet cookies. Well, maybe like chocolate chip cookies and oatmeal raisin cookies, those were a kind of cookies. Basically, that's what it did. Or no, actually did the opposite. [LAUGHS]

It said-- when it read "I deleted my files," it decided files was probably a baked good, too. Well, first it decided internet cookies was a baked good, like those other cookies. And then it decided that files were a baked goods. And it started this whole avalanche of mistakes, Dr. Mitchell said. He corrected the internet cookies error and restarted NELL's bakery education. [LAUGHS] I mean, like, OK. Now rerun without that problem.

So the point, the lesson Tom draws from this, and that the article talks about, is, oh, well, we still need some assistance. We have to go back and, by hand, set these things. But the key thing is that, really-- I think the message this is telling us is no human child would ever make this mistake. Human children learn in this way. They don't need this kind of assistance. It's true that, as Tom says, you and I don't learn in isolation either. So, all of the things we've been talking about, about learning from prior knowledge and so on, are true.

But there's a basic kind of common sense thing that this is missing, which is that at the time a child is learning anything about-- by the time a child is learning anything about computers, and files, and so on, they understand well before that, like back in early infancy, from say, work that Liz has done, and many others, that cookies, in the sense of baked goods, are a physical object, a kind of food, a thing you eat. Files, email-- not a physical object. And there's all sorts of interesting stuff to understand about how kids learn that a book can be both a no-- a novel is both a story and it's also a physical object, and so a lot of that stuff.

But there's a basic common sense understanding of the world as consisting of physical objects, and for example, agents and their goals. You heard a little bit about this from us, from me and Tomer, on the first day. And that's where I want to turn to next. And this is just one of many examples that we realized, as cool as this system is, as great as all this stuff is, just trying to approach semantic knowledge and common sense reasoning as some kind of big matrix completion without a much more fundamental grasp of the ways in which the world is real to a human mind, well before they're learning anything about language or any of this higher level stuff, it was just not going to work, in the same way that I think if you want to build a system that learns to play a video game, even remotely like the way a human does, there's a lot of more basic stuff you have to build on. And it's the same basic stuff, I would argue.

A cool thing about Atari video games is that, even though they were very low resolution, very low-bit color displays, with very big pixels, what makes your ability to learn that game work is the same kind of thing that makes the ability, even as a young child, to not make this mistake. And it's the kind of thing that Liz and people in her field of developmental psychology-- in particular, infant research-- have been studying really excitingly for a couple of decades. That, I think, is as transformative for the topic of intelligence in brains, minds, and machines as anything.

So that's what motivated the work we've been doing in the last few years and the main work we're trying to do in the center. And it also goes hand-in-hand with the ways in which we've realized that we have to take what we've learned how to do with building problematic models over interesting symbolically-structured representations and so on, but move way beyond what you could call-- I mean, we need better, even more interesting, symbolic representations. In particular, we need to move beyond graphs and stochastic processes defined over graphs to programs. So that's where the probabilistic programs come back into the mix.

So again, you already saw this. And I'm trying to close the loop back to what we're doing in

CBMM. I've given you about 10 to 15 years of background in our field of how we got to this, why we think this is interesting and important, and why we think we need to-- why we've developed a certain toolkit of ideas, and why we think we needed to keep extending it. And I think, as you saw before, and as you'll see, this also, in some ways-- I think we're getting more and more to the interesting part of common sense.

But in another way, we're getting back to the problems I started off with and what a lot of other people at this summer school have an interest in, which is things like much more basic aspects of visual perception. I think the heart of real intelligence and common sense reasoning that we're talking about is directly connected to vision and other sense modalities, and how we get around in the world and plan our actions, and the very basic kinds of goal social understandings that you saw in those little videos of the red and blue ball, or that you see if you're trying to do action recognition and action understanding.

So in some sense, it's gotten more cognitive. But it's also, by getting to the root of our common sense knowledge, it makes better contact with vision, with neuroscience research. And so I think it's a super exciting development in what we're doing for the larger Brains, Minds, and Machines agenda. So again, now we're saying, OK, let's try to understand the way in which-- even these kids playing with blocks, the world is real to them. It's not just a big matrix of data. That is a thing in their hands. And they have an understanding of what a thing is before they start compiling lists of properties.

And they're playing with somebody else. That hand is attached to a person, who has goals. It's not just a big matrix of rows and columns. It's an agent with goals, and even a mind. And they understand those things before they start to learn a lot of other things, like words for objects, and advanced game-playing behavior, and so on.

And when we want to talk about learning, we still are interested in one-shot learning, or very rapid learning from a few examples. And we're still interested in how prior knowledge guides that, and how that knowledge can be built. But we want to do it in this context. We want to study in the context of, say, how you learn how magnets work, or how you learn how a touchscreen device works-- really interesting kinds of grounded physical causes.

So this is what we have, or what I've come to call the common sense core. Liz, are you going to talk about core knowledge at all? so there's a phrase that Liz likes to use called core knowledge. And this is definitely meant to evoke that. And it's inspired by it. I guess I changed it a little bit, because I wanted it to mean something a little bit different. And I think, again, to anticipate a little bit, the main difference is-- I don't know. What's the main difference?

The main difference is that, in the same way that lots of people look at me and say, oh, he's the Bayesian guy, lots of people look at Liz and say, oh, she's the nativist gal or something. And it's true that, compared to a lot of other people, I tend to be more interested, and have done more work prominently associated with Bayesian inference. But by no means do I think that's the whole story. And part of what I tried to show you, and will keep showing you, is ways in which that's only really the beginning of the story.

And Liz is prominently associated, and you'll see some of this, with really fascinating discoveries that key high level concepts, key kinds of real knowledge, are present, in some sense, as early as you can look, and in some form, I think, very plausibly, have to be due to some kind of innately unfolding genetic program that builds a mind the same way it builds a brain. But just as we'll hear from her, that's, in some ways, only the beginning, or only one part of a much richer, more interesting story that she's been developing.

But for that, among other reasons, I'm calling it a little different. And I'm trying to emphasize the connection to what people in AI call common sense reasoning. Because I really do think this is the heart of common sense. It's this intuitive physics and intuitive psychology. So again, you saw us already give an intro to this. Maybe what I'll just do is show you a little bit more of the-- well, are you going to talk about the stuff at all?

LIZ SPELKE: I guess. Yeah.

JOSH TENENBAUM: Well, OK. So this is work-- some of this is based on Liz's work. Some of this is based on Renée Baillargeon, a close colleague of hers, and many other people out there. And I wasn't really going to go into the details. And maybe, Liz, we can decide whether you want to do this or not. But what they've shown is that, even prior to the time when kids are learning words for objects, all of this stuff with infants, two months, four months, eight months-- at this age, kids have, at best, some vague statistical associations of words to kinds of objects. But they already have a great deal of much more abstract understanding of physical objects.

So I won't-- maybe I should not go into the details of it. But you saw it in that nice video of the baby playing with the cups. And there's really interesting, sort of rough, developmental timelines. One of the things we're trying to figure out in CBMM is to actually get much, much clearer picture on this. But at least if you look across a bunch of different studies, sometimes by one lab, sometimes up by multiple labs, you see ways in which, say, going from two months to five months, or five months to 12 months, kids seem to-- their intuitive physics of objects is getting a little bit more sophisticated.

So for example, they tend to understand-- in some form, they understand a little bit of how collisions conserve momentum, a little bit, by five months or six months-- according to one of Baillargeon's studies-- in the sense that if they see a ball roll down a ramp and hit another one, and the second one goes a certain distance, if a bigger object comes, they're not too surprised if this one goes farther. But if a little object hits it, then they are surprised. So they expect a bigger object to be able to move it more than a little object.

But a two-month-old doesn't understand that. Although a two-month-old does understand-- this is, again, from Liz's work-- that if an object is colluded by a screen, it hasn't disappeared, and that if an object is rolling towards a wall, and that wall looks solid, that the object can't go through it, and that if it somehow-- when the screen is removed, as you see on the upper left-- appears on the other side of the screen, that's very surprising to them. I think-- I'm sure what Liz will talk about, among other things, are the methods they use, the looking time methods to reveal this.

And I think there's really-- this is one of the two main insights that I, and I think our whole field, needs to learn from developmental psychology, is how much of a basic understanding of physics like this is present very early. And it doesn't matter whether it's-- in some sense, it doesn't matter for the points I want to make here, how much or in what way this is innate, or how the genetics and the experience interact. I mean, that does matter.

And that, that's something we want to understand, and we are hoping to try to understand in the hopefully not-too-distant future. But for the purpose of understanding what is the heart of common sense, how are we going to build these causal, compositional, generative models to really get at intelligence, the main thing is that it should be about this kind of stuff. That's the main focus.

And then the other big insight from developmental psychology, which has to do with how we build this stuff, is this idea sometimes called the child as scientist. The basic idea is that, just as this early commonsense knowledge is something like a scientific theory, something like a good scientific theory, the way Newton's laws are a better scientific theory than Kepler's laws because of how they capture the causal structure of the world in a compositional way. That's another way to sum up what I'm trying to say about children's early knowledge.

But also, the way children build their knowledge is something like the way scientists build their knowledge, which is, well, they do experiments, of course. We normally call that play. That's one of Laura's Schulz's big ideas. But it's not just about the experiments. I mean, Newton didn't really do any experiments. He just thought.

And that's another thing you'll hear from Laura, and also from Tomer, is that a lot of children's learning looks less like, say, stochastic gradient descent, and more like scratching your head and trying to make sense of, well, that's really funny. Why does this happen here? Why does that happen over there? Or how can I explain what seemed to be diverse patterns of phenomena with some common underlying principles, and making analogies between things, and then trying out, oh, well, if that's right, then it would make this prediction.

And the kid doesn't have to be conscious of that the way scientists maybe are. That process of coming up with theories and considering variations, trying them out, seeing what kinds of new experiences you can create for yourself-- call them an experiment, or call them just a game, or playing with a toy, but that dynamic is the real heart of how children learn and build the knowledge from the early stages to what we come to have as adults. Those two insights of what we start with and how we grow, I think, are hugely powerful and hugely important for anything we want to do in capturing-- making machines that learn like humans, or making computational models that really get at the heart of how we come to be smart.