JOSH TENENBAUM: I'm going to be talking about computational cognitive science. In the brains, minds, and machines landscape, this is connecting the minds and the machines part. And I really want to try to emphasize both some conceptual themes and some technical themes that are complimentary to a lot of what you've seen for the first week or so of the class.

That's going to include ideas of generative models and ideas of probabilistic programs, which we'll see a little bit here and a lot more in the tutorial in the afternoon. And on the cognitive side, maybe we could sum it up by calling it common sense. Since this is meant to be a broad introduction-- and I'm going to try to cover from some very basic fundamental things that people in this field were doing maybe 10 or 15 years ago up until the state of the art current research-- I want to try to give that whole broad sweep.

And I also want to try to give a bit of a sort of philosophical introduction at the beginning to set this in context with the other things you're seeing in the summer school. I think it's fair to say that there are two different notions of intelligence that are both important and are both interesting to members of this center in the summer school. The two different notions are what I think you could call classifying, recognizing patterns in data, and what you could call explaining, understanding, modeling the world.

So, again, there's the notion of classification, pattern recognition, finding patterns in data and maybe patterns that connect data to some task you're trying to solve. And then there's this idea of intelligence as explaining, understanding, building a model of the world that you can use to play on and solve problems with. I'm going to emphasize here notions of explanation, because I think they are absolutely central to intelligence, certainly in any sense that we mean when we talk about humans. And because they get kind of underemphasized in a lot of recent work in machine learning, AI, neural networks, and so on.

Like, most of the techniques that you've seen so far in other parts of the class and will continue to see, I think it's fair to say they sort of fall under the broad idea of trying to classify

and recognize patterns in data. And there's good reasons why there's been a lot of attention on these recently, particularly coming from the more brain side. Because it's much easier when you go and look in the brain to understand how neural circuits do things like classifying recognized patterns.

And it's also, I think with at least certain kinds of current technology, much easier to get machines to do this, right? All the excitement in deep neural networks is all about this, right? But what I want to try to convince you here and illustrate a lot of different kinds of examples is how both of these kinds of approaches are probably necessary, essential to understanding the mind.

I won't really bother to try to convince you that the pattern recognition approach is essential, because I take that for granted. But both are essential and, also, that they essentially need each other. I'll try to illustrate a couple of ways in which they really each solve the problems that the other one needs-- so ways in which ideas like deep neural networks for doing really fast pattern recognition can help to make the sort of explaining understanding view of intelligence much quicker and maybe much lower energy, but also ways in which the sort of explaining, understanding view of intelligence can make the pattern recognition view much richer, much more flexible.

What do you really mean? What's the difference between classification and explanation? Or what makes a good explanation? So we're talking about intelligence as trying to explain your experience in the world, basically, to build a model that is in some sense a kind of actionable causal model.

And there's a bunch of virtues here, these bullet points under explanation. There's a bunch of things we could say about what makes a good explanation of the world or a good model. And I won't say too much abstractly. I'll mostly try to illustrate this over the morning.

But like any kind of model, whether it's sort of more pattern recognition classification style or these more explanatory type models, ideas of compactness, unification, are important, right? You want to explain a lot with a little. OK? There's a term if anybody has read David Deutsch's book *The Beginning Of Infinity.* He talks about this view in a certain form of good explanations as being hard to vary, non-arbitrary. OK.

That's sort of in common with any way of describing or explaining the world. But some key features of the models we're going to talk about-- one is that they're generative. So what we

mean by generative is that they generate the world, right?

In some sense, their output is the world, your experience. They're trying to explain the stuff you observe by positing some hidden, unobservable, but really important, causal actionable deep stuff. They don't model a task.

That's really important. Because, like, if you're used to something like, you know, end to end training of a deep neural network for classification where there's an objective function and the task and the task is to map from things you experience and observe in the world to how you should behave, that's sort of the opposite view, right? These are things whose output is not behavior on a task, but whose output is the world you see.

Because what they're trying to do is produce or generate explanations. And that means they have to come into contact. They have to basically explain stuff you see.

OK. Now, these models are not just generative in this sense, but they're causal. And, again, I'm using these terms intuitively. I'll get more precise later on. But what I mean by that is the hidden or latent variables that generate the stuff you observe are, in some form, trying to get at the actual causal mechanisms in the world-- the things that, if you were then to go act on the world, you could intervene on and move around and succeed in changing the world the way you want. Because that's the point of having one of these rich models is so that you can use it to act intelligently, right?

And, again, this is a contrast with a approach that's trying to find and classify patterns that are useful for performing some particular task to detect oh, when I see this, I should do this. When I see this, I should do that, right? That's good for one task.

But these are meant to be good for an endless array of tasks. Not any task, but, in some important sense, a kind of unbounded set of tasks where given a goal which is different from your model of the world-- you have your goal. You have your model of the world. And then you use that model to plan some sequence of actions to achieve your goal.

And you change the goal, you get a different plan. But the model is the invariant, right? And it's invariant, because it captures what's really going on causally.

And then maybe the most important, but hardest to really get a handle on, theme-- although, again, we'll try to do this by the end of today-- is that they're compositional in some way. They

consist of parts which have independent meaning or which have some notion of meaning, and then ways of hooking those together to form larger wholes. And that gives a kind of flexibility or extensibility that is fundamental, important to intelligence-- the ability to not just, say, learn from little data, but to be able to take what you've learned in some tasks and use it instantly, immediately, on tasks you've never had any training for. It's, I think, really only with this kind of model building view of intelligence that you can do that.

I'll give one other motivating example-- just because it will appear in different forms throughout the talk-- just of the difference between classification and explanation as ways of thinking about the world with thinking about, in particular, planets and just the orbits of objects in the solar system. That could include objects, basically, on any one planet, like ours. But think about the problem of describing the motions of the planets around the sun.

Well, there's some phenomena. You can make observations. You could observe them in various ways. Go back to the early stages of modern science when the data by which the phenomena were represented-- you know, things like just measurements of those light spots in the sky, over nights, over years.

So here are two ways to capture the regularities in the data. You could think about Kepler's laws or Newton's laws. So just to remind you, these are Kepler's laws. And these are Newton's laws.

I won't really go through the details. Probably, all of you know these or have some familiarity. The key thing is that Kepler's laws are laws about patterns of motion and space and time. They specify the shape of the orbits, the shape of the path that the planets trace out in the solar system. Not in the sky, but in the actual 3D world-- the idea that the orbits, the planets, are an ellipse with the sun at one focus.

And then they give some other mathematical regularities that describe, in a sense, how fast the planets go around the sun as a function of the size of the orbit and the fact that they kind of go faster at some places and slower at other places in the orbit, right? OK. But in a very important sense, they don't explain why they do these things, right? These are patterns which, if I were to give you a set of data, a path, and I said, is this a possible planet or not-- maybe there's a undiscovered planet. And this is possibly that, or maybe this is some other thing like a comet.

And you could use this to classify and say, yeah, that's a planet, not a comet, right? And, you

know, you could use them to predict, right? If you've observed a planet over some periods of time in the sky, then you could use Kepler's laws to basically fit an ellipse and figure out where it's going to be later on. That's great.

But they don't explain. In contrast, Newton's laws work like this, right? Again, there's several different kinds of laws. There's, classically, Newton's laws of motion.

These ideas about inertia and F equals MA and every action produces an equal and opposite reaction, again, don't say anything about planets. But they really say everything about force. They talk about how forces work and how forces interact and combine and compose-- compositional-- to produce motion or, in particular, to produce the change of motion. That's acceleration or the second derivative of position.

And then there's this other law, the law of gravitational force, so the universal gravitation, which specifies in particular how you get one particular force. That's the name of the force we call gravity as a function of the mass of the two bodies and the square distance between them and some unknown constant, right? And the idea is you put these things together and you get Kepler's law.

You can derive the fact that the planets have to go that way from the combination of these laws of motion and the law of gravitational force. So there's a sense in which the explanation is deeper and that you can derive the patterns from the explanation. But it's a lot more than that.

Because these laws don't just explain the motions of the planets around the sun, but a huge number of other things. Like, for example, they don't just explain the orbits of the planets, but also other things in the solar system. Like, you can use them to describe comets.

You can use them to describe the moon going around the planets. And you can use them to explain why the moon goes around the Earth and not around the sun in that sense, right? You can use them to explain not just the motions of the really big things in the solar system, but the really little things like, you know, this, and to explain why when I drop this or when Newton famously did or didn't drop an apple or had an apple drop on its head, right?

That, superficially, seems to be a very different pattern, right? It's something going down in your current frame of reference. But the very same laws describe exactly that and explain why the moon goes around the Earth, but the bottle or the apple goes down in my current experience of the world.

In terms of things like causal and actionable ideas, they explain how you could get a man to the moon and back again or how you could build a rocket to escape the gravitational field to not only get off the ground the way we're all on the ground, but to get off or out of orbiting around and get to orbiting some other thing, right? And it's all about compositionality as well as causality. In order to escape the Earth's gravitational field or get to the moon and back again, there's a lot of things you have to do.

But one of the key things you have to do is generate some significant force to oppose, be stronger, than gravity. And, you know, Newton really didn't know how to do that. But some years later, people figured out, you know, by chemistry and other things-- explosions, rockets-- how to do some other kind of physics which could generate a force that was powerful enough for an object the size of rocket to go against gravity and to get to where you need to be and then to get back.

So the idea of a causal model, which in this case is the one based on forces, and compositionality-- the ability to take the general laws of forces, laws about one particular kind of force that's generated by this mysterious thing called mass, some other kinds of forces generated by exploding chemicals-- put those all together is hugely powerful. And, of course, this as an expression of human intelligence-- you know, the moon shot is a classic metaphor. Demis used it in his talk.

And I think if we really want to understand the way intelligence works in the human mind and brain that could lead to this, you have to go back to the roots of intelligence. You've heard me say this before. And I'm going to do this more by today.

We want to go back to the roots of intelligence in even very young children where you already see all of this happening, right? OK. So that's the big picture. I'll just point you. If you want to learn more about the history of this idea, a really nice thing to read is this book by Kenneth Craik.

He was an English scientist sort of a contemporary of Turing, also died tragically early, although from different tragic causes. He was, you know, one of the first people to start thinking about this topic of brains, minds, and machines, cybernetics type ideas, using math to describe how the brain works, how the mind might work in a brain. As you see when you read this quote, he didn't even really know what a computer was. Because it was pre-Turing, right?

But he wrote this wonderful book, very short book. And I'll just quote here from one of the chapters. The book was called *The Nature Of Explanation.* And it was sort of both a philosophical study of that and how explanation works in science, like some of the ideas I was just going through, but also really arguing in very common sense and compelling ways why this is a key idea for understanding how the mind and the brain works.

And he wasn't just talking about humans. Well, you know, these ideas have their greatest expression in some form, their most powerful expression, in the human mind. They're also important ones for understanding other intelligent brains.

So he says here, "one of the most fundamental properties of thought is its power of predicting events. It enables us, for instance, to design bridges with a sufficient factor of safety instead of building them haphazard and waiting to see whether they collapse. If the organism carries a small scale model of external reality into its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilize the knowledge of past events in dealing with the present and future and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it."

So he's just really summing up this is what intelligence is about-- building a model of the world that you can manipulate and plan on and improve, think about, reason about, all that. And then he makes this very nice analogy, a kind of cognitive technology analogy. "Most of the greatest advances of modern technology have been instruments which extended the scope of our sense organs, our brains, or our limbs-- such, our telescopes and microscopes, wireless calculating machines, typewriters, motor cars, ships, and airplanes."

Right? He's writing in 1943, or that's when the book was published, writing a little before that. Right? He didn't even have the word computer. Or back then, computer meant something different-- people who did calculations, basically.

But same idea-- that's what he's talking about. He's talking about a computer, though he doesn't yet have the language quite to describe it. "Is it not possible, therefore, that our brains themselves utilize comparable mechanisms to achieve the same ends and that these mechanisms can parallel phenomena in the external world as a calculating machine can parallel with development of strains in a bridge?"

And what he's saying is that the brain is this amazing kind of calculating machine that, in some

form, can parallel the development of forces in all sorts of different systems in the world and not only forces. And, again, he doesn't have the vocabulary in English or the math really to describe it formally. That's, you know, why this is such an exciting time to be doing all these things we're doing is because now we're really starting to have the vocabulary and the technology to make good on this idea.

OK. So that's it for the big picture philosophical introduction. Now, I'll try to get more concrete with the questions that have motivated not only me, but many cognitive scientists. Like, why are we thinking about these issues of explanation? And what are our concrete handles?

Like, let's give a couple of examples of ways we can study intelligence in this form. And I like to say that the big question of our field-- it's big enough that it can fold in most, if not all, of our big questions-- is this one. How does the mind get so much out of so little, right?

So across cognition wherever you look, our minds are building these rich models of the world that go way beyond the data of our senses. That's this extension of our sense organs that Craik was talking about there, right? From data that is altogether way too sparse, noisy, ambiguous in all sorts of ways, we build models that allow us to go beyond our experience, to plan effectively.

How do we do it? And you could add-- and I do want to go in this direction. Because it is part of how when we relate the mind to the brain or these more explanatory models to more sort of pattern classification models, we also have to ask not only how do you get such a rich model of the world from so little data, but how do you do it so quickly?

How do you do it so flexibly? How do you do it with such little energy, right? Metabolic energy is an incredible constraint on computation in the mind and brain.

So just to give some examples-- again, these are ones that will keep coming up here. They've come up in our work. But they're key ones that allow us to take the perspective that you're seeing today and bring it into contact with the other perspectives you're seeing in the summer school.

So let's look at visual scene perception. This is just a snapshot of images I got searching on Google Images for, I think, object detection, right? And we've seen a lot of examples of these kinds of things.

You can go to the iCub and see its trainable object detectors. We'll see more of this when Amnon, the Mobileye guy, comes and tells us about really cool things they've done to do object detection for self-driving cars. You saw a lot of this kind of thing in robotics before.

OK. So what's the basic sort of idea, the state of the art, in a lot of higher level computer vision? It's getting a system that learns to put boxes around regions of an image that contains some object of interest that you can label with the word, like person or pedestrian or car or horse, or various parts of things. Like, you might not just put a box around the bicycle, but you might put a box around the wheel, handlebar, seat, and so on. OK.

And in some sense, you know, this is starting to get at some aspect of computer vision, right? Several people have quoted from David Marr who said, you know, vision is figuring out what is where from images, right? But Marr meant something that goes way beyond this, way beyond putting boxes in images with single-word labels.

And I think you just have to, you know, look around you to see that your brain's ability to reconstruct the world, the whole three-dimensional world with all the objects and surfaces in it, goes so far beyond putting a few boxes around some parts of the image, right? Even put aside the fact that when you actually do this in real time on a real system, you know, the mistakes and the gaps are just glaring, right? But even if you could do this, even if you could put a box around all the things that we could easily label, you look around the world.

You see so many objects and surfaces out there, all actionable. This is what this is when I talk about causality, right? Think about, you know, if somebody told me that there was some treasure hidden behind the chair that has Timothy Goldsmith's name on it, I know I could go around looking for the chair. I think I saw it over there, right?

And I know exactly what I'd have to do. I'd have to go there, lift up the thing, right? That's just one of the many plans I could make given what I see in this world.

If I didn't know that that was Timothy Goldsmith's chair, somewhere over there there's the Lily chair, right? OK. So I know that there's chairs here. There's little name tags on them.

I could go around, make my way through looking at that tags, and find the one that says Lily and then, again, know what I have to do to go look for the treasure buried under it, right? That's just one of, really, this endless number of tasks that you can do with the model of the world around you that you've built from visual perception. And we don't need to get into a

debate of, you know-- here, we can do this in a few minutes if you want-- about the difference between, like say for example, what Jim DiCarlo might call core object recognition or the kind of stuff that Winrich is studying where, you know, you show a monkey just a single object against maybe a cluttered background or a single face for 100 or 200 milliseconds, and you ask a very important question.

What can you get in 100 milliseconds in that kind of limited scene? That's a very important question. But the convergence of visual neuroscience on that problem has enabled us to really understand a lot about the circuits that drive the first initial paths of some aspects of high-level vision, right?

But that is really only getting out the classification or pattern detection part of the problem. And the other part of the problem, figuring out the stuff in the world that causes what you see, that is really the actionable part of things to guide your actions the world. We really are still quite far from understanding that at least with those kinds of methods.

Just to give a few examples-- some of my favorite kind of hard object detection examples, but ones that show that your brain is really doing this kind of thing even from a single image. You know, it doesn't just require a lot of extensive exploration. So let's do some person detecting problems here. So here's a few images.

And let's just start with the one in the upper left. You tell me. Here, I'll point with this, so you can see it on the screen. How many people are in this upper left image? Just tell me.

AUDIENCE: Three.

AUDIENCE: About 18.

JOSH TENENBAUM: About 18? OK. OK. Yeah. That's a good answer, yeah. There are somewhere between 20 or 30 or something. Yeah. That was even more precise than I was expecting.

OK. Now, I don't know. This would be a good project if somebody is still looking for a project. If you take the best person detector that you can find out there or that you can build from however much training data you find labeled on the web, how many of those people is it going to detect?

You know, my guess is, at best, it's going to detect just five or six-- just the bicyclists in the front row. Does that seem fair to say? Even that will be a challenge, right?

Whereas, not only do you have no trouble detecting the bicyclists in the front row, but all the other ones back there, too, even though for many of them all you can see is like a little bit of their face or neck or sometimes even just that funny helmet that bicyclists wear. But your ability to make sense of that depends on understanding a lot of causal stuff in the world-- the three-dimensional structure of the world, the three-dimensional structure of bodies in the world, some of the behaviors that bicyclists tend to engage in, and so on. Or take the scene in the upper right there, how many people are in that scene?

AUDIENCE: 350.

JOSH TENENBAUM: 350. Maybe a couple of hundred or something. Yeah, I guess. Were you counting all this time? No. That was a good estimate

AUDIENCE: No.

JOSH TENENBAUM: Yeah, OK. The scene in the lower left, how many people are there?

AUDIENCE: 100?

JOSH TENENBAUM: 100-something, yeah. The scene in the lower right?

AUDIENCE: Zero.

JOSH TENENBAUM: Zero. Was anybody tempted to say two? Were you tempted to say two as a joke or seriously? Both are valid responses.

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: Yeah. OK. So, again, how do we solve all those problems, including knowing that one in the bottom-- maybe it takes a second or so-- but knowing that, you know, there's actually zero there. You know, it's the hats, the graduation hats, that are the cues to people in the other scenes. But here, again, because we know something about physics and the fact that people need to breathe-- or just tend to not bury themselves all the way up to the tippy top of their head, unless it's like some kind of Samuel Beckett play or something, *Graduation Endgame*-- then, you know, there's almost certainly nobody in that scene. OK.

Now, all of those problems, again, are really way beyond what current computer vision can do and really wants to do. But I mean, I think, you know, the aspect of scene understanding that really taps into this notion of intelligence, of explaining modeling the causal structure of the world, should be able to do all that. Because we can, right?

But here's a problem which is one that motivates us on the vision side that's somewhere in between these sort of ridiculously hard by current standards problems and one that, you know, people can do now. This is a kind of problem that I've been trying to put out there for computer vision community to think about it in a serious way. Because it's a big challenge, but it's not ridiculously hard.

OK. So here, this is a scene of airplane full of computer vision researchers, in fact, going to last year's CVPR conference. And, again, how many people are in the scene?

AUDIENCE: 20?

JOSH TENENBAUM: 20,50? Yeah, something like that. Again, you know, more than 10, less than 500, right? You could count. Well, you can count, actually. Let's try that.

So, you know, just do this mentally along with me. Just touch, in your mind, all the people. You know, 1, 2, 3, 4-- well, it's too hard to do it with the mouse. Da, da, da, da, da-- you know, at some point it gets a little bit hard to see exactly how many people are standing in the back by the restroom. OK.

But it's amazing how much you can, with just the slightest little bit of effort, pick out all the people even though most of them are barely visible. And it's not only that. It's not just that you can pick them out. While you only see a very small part of their bodies, you know where all the rest of their body is to some degree of being able to predict an act if you needed to, right?

So to sort of probe this, here's a kind of little experiment we can do. So let's take this guy here. See, you've just got his head.

And though you see his head, think about where the rest of his body is. And in particular, think about where his right hand is in the scene. You can't see his right hand. But in some sense, you know where it is.

I'll move the cursor. And you just hum when I get to where you think his right hand is if you could see, like if everything was transparent.

AUDIENCE: Yeah.

AUDIENCE: Yeah.

JOSH TENENBAUM: OK. Somewhere around there. All right, how about let's take this guy. You can see his scalp only and maybe a bit of his shoulder. Think about his left big toe. OK? Think about that. And just hum when I get to where his left big toe is.

AUDIENCE: Yeah.

AUDIENCE: Yeah.

JOSH TENENBAUM: Somewhere, yeah. All right, so you can see we did an instant experiment. You don't even need Mechanical Turk. It's like recording from neurons, only you're each being a neuron. And you're humming instead of spiking.

But it's amazing how much you can learn about your brain just by doing things like that. You've got a whole probability distribution right there, right? And that's a meaningful distribution.

You weren't just hallucinating, right? You were using a model, a causal model, of how bodies work and how other three-dimensional structures work to solve that problem. OK. This isn't just about bodies, right?

Our ability to detect objects, like to detect all the books on my bookshelf there-- again, most of which are barely visible, just a few pixels, a small part of each book, or the glasses in this tabletop scene there, right? I don't really know any other way you can do this. Like, any standard machine learning-based book detector is not going to detect most of those books. Any standard glass detector is not going to detect most of those glasses. And yet you can do it.

And I don't think there's any alternative to saying that in some sense, as we'll talk more about it in a little bit, you're kind of inverting the graphics process, In computer science now, we call it graphics. We maybe used to call it optics. But the way light bounces off the surfaces of objects in the world and comes into your eye, that's a causal process that your visual system is in some way able to invert, to model and go from the observable to the unobservable stuff, just like Newton was doing with astronomical data.

OK. Enough on vision for now, sort of. Let's go from actually just perceiving this stuff out there

in the world to forming concepts and generalizing. So a problem that I've studied a lot, that a lot of us have studied a lot in this field, is the problem of learning concepts and, in particular, one very particular kind of concept, which is object kinds like categories of objects, things we could label with a word.

It's one of the very most obvious forms of interesting learning that you see in young children, part of learning language. But it's not just about language. And the striking thing when you look at, say, a child learning words-- just in particular let's say, words that label kinds of objects, like chair or horse or bottle, ball-- is how little data of a certain labels or how little task relevant data is required. A lot of other data is probably used in some way, right?

And, again, this is a theme you've heard from a number of the other speakers. But just to give you some of my favorite examples of how we can learn object concepts from just one or a few examples, well, here's an example from some experimental stimuli we use where we just made up a whole little world of objects. And in this world, I can teach you a new name, let's say tufa, and give you a few examples.

And, again, you can now go through. We can try this as a little experiment here and just say, you know, yes or no. For each of these objects, is it a tufa? So how about this, yes or no?

AUDIENCE: Yes.

JOSH TENENBAUM: Here?

AUDIENCE: No.

JOSH TENENBAUM: Here?

AUDIENCE: No.

JOSH TENENBAUM: Here?

AUDIENCE: No.

JOSH TENENBAUM: Here?

AUDIENCE: No.

JOSH TENENBAUM: Here?

AUDIENCE: Yes.

JOSH TENENBAUM: Here?

AUDIENCE: No.

JOSH TENENBAUM: Here?

AUDIENCE: Yes. No. No. No. No. Yes.

JOSH TENENBAUM: Yeah. OK. So first of all, how long did it take you for each one? I mean, it basically didn't take you any longer than it takes in one of Winrich's experiments to get the spike seeing the face. So you learned this concept, and now you can just use it right away.

It's far less than a second of actual visual processing. And there was a little bit of a latency. This one's a little more uncertain here, right? And you saw that in that it took you maybe almost twice as long to make that decision. OK.

That's the kind of thing we'd like to be able to explain. And that means how can you get a whole concept? It's a whole new kind of thing.

You don't really know much about it. Maybe you know it's some kind of weird plant on this weird thing. But you've got a whole new concept and a whole entry into a whole, probably, system of concepts. Again, several notions of being quick-- sample complexity, as we say, just one or a few examples, but also the speed-- the speed in which you formed that concept and the speed in which you're able to deploy it in now recognizing and detecting things.

Just to give one other real world example, so it's not just we make things up-- but, for example, here's an object. Just how many know what this thing is? Raise your hand if you do. How many people don't know what this thing is? OK. Good.

So this is a piece of rock climbing equipment. It's called a cam. I won't tell you anything more

than that. Well, maybe I'll tell you one thing, because it's kind of useful.

Well, I mean, you may or may not even need to-- yeah. This strap here is not technically part of the piece of equipment. But it doesn't really matter. OK.

So anyway, I've given you one example of this new kind of thing for most of you. And now, you can look at a complex scene like this climber's equipment rack. And tell me, are there any cams in this scene?

AUDIENCE: Yes.

JOSH TENENBAUM: Where are they?

AUDIENCE: On top.

JOSH TENENBAUM: Yeah. The top. Like here?

AUDIENCE: No. Next to there.

JOSH TENENBAUM: Here. Yeah. Right, exactly. How about this scene, any?

AUDIENCE: No.

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: There's none of that-- well, there's a couple. Anyone see the ones over up in the upper right up here?

AUDIENCE: Yeah.

JOSH TENENBAUM: Yeah. They're hard to see. They're really dark and shaded, right? But when I draw your attention to it, and then you're like, oh yeah. I see that, right?

So part of why I give these examples is they show how the several examples I've been giving, like the object concept learning thing, interacts with the vision, right? I think your ability to solve tasks like this rests on your ability to form this abstract concept of this physical object. And notice all these ones, they're different colors.

The physical details of the objects are different. It's only a category of object that's preserved. But your ability to recognize these things in the real world depends on, also, the ability to recognize them in very different viewpoints under very different lighting conditions.

And if we want to explain how you can do this-- again, to go back to composability and compositionality-- we need to understand how you can put together the kind of causal model of how scenes are formed. That vision is inverting-- this inverse graphics thing-- with the causal model of something about how objects concepts work and compose them together to be able to learn a new concept of an object that you can also recognize new instances of the kind of thing in new viewpoints and under different lighting conditions than the really wonderfully perfect example I gave you here with a nice lighting and nice viewpoint. We can push this to quite extremes. Like, in that scene in the upper right, do you see any cams there?

AUDIENCE: Yeah.

JOSH TENENBAUM: Yeah. How many are there?

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: Quite a lot, yeah, and, like, all occluded and cluttered. Yeah. Amazing that you can do this. And as we'll see in a little bit, what we do with our object concepts-- and these are other ways to show this notion of a generative model-- we don't just classify things. But we can use them for all sorts of other tasks, right?

We can use them to generate or imagine new instances. We can parse an object out into parts. This is another novel, but real object-- the Segway personal thing.

Which, again, probably all of you know this, right? How many people have seen those Segways before, right? OK. But you all probably remember the first time you saw one on the street.

And whoa, that's really cool. What's that new thing? And then somebody tells you, and now you know, right?

But it's partly related to your ability to parse out the parts. If somebody says, oh, my Segway has a flat tire, you kind of know what that means and what you could do, at least in principle, to

fix it, right? You can take different kinds of things in some category like vehicles and imagine ways of combining the parts to make yet other new either real or fanciful vehicles, like that C to the lower right there. These are all things you do from very little data from these object concepts.

Moving on and then both back to some examples you saw Tomer and I talk about on the first day in our brief introduction and what we'll get to more by the end of today, examples like these. So Tomer already showed you the scene of the red and the blue ball chasing each other around. I won't rehearse that example. I'll show you another scene that is more famous.

OK. Well, so for the people who haven't seen it, you can never watch it too many times. Again, like that one, it's just some shapes moving around. It was done in the 1940s, that golden age for cognitive science as well as many other things.

And much lower technology of animation, it's like stop-action animation on a table top. But just like the scene on the left which is done with computer animation, just from the motion of a few shapes in this two-dimensional world, you get so much. First of all, you get physics.

Let's watch it again. It looks like there's a collision. It's just objects, shapes moving. But it looks like one thing is banging into another. And it looks like they're characters, right?

It looks like the big one is kind of bullying the other one. It's sort of backed him up against the wall scaring them off, right? Does you guys see that?

The other one was hiding. Now, this one goes in to go after him. It starts to get a little scary, right? Cue the scary music if it was a silent movie. Doo, doo, doo, doo, doo, OK.

You can watch the end of it on YouTube if you want. It's quite famous. So I won't show you the end of it. But in case you're getting nervous, don't worry. It ends happily, at least for two of the three characters.

From some combination of all your experiences in your life and whatever evolution genetics gave you before you came out into the world, you've built up a model that allows you to understand this. And then it's a separate, but very interesting, question and harder one. How do you get to that point, right?

The question of the development of the kind of commonsense knowledge that allows you to parse out just the motion into both forces, you know, one thing hitting another thing, and then

the whole mental state structure and the sort of social who's good and who's bad on there-- I mean, because, again, most people when they see this and think about a little bit see some of the characters as good and others as bad. How that knowledge develops is extremely interesting. We're going to see a lot more of the more experiments, how we study this kind of thing in young children, next week.

And we'll talk more about the learning next week. We'll see how much of that I get to. What I want to talk about here is sort of general issues of how the knowledge works, how you deploy it, how you make the inferences with the knowledge, and a little bit about learning. Maybe we'll see if we have time for that at the end.

But they'll be more of that next week. I think it's important to understand what the models are, these generative models that you're building of the world, before you actually study learning. I think there's a danger if you study learning. Without having the right target of learning, you might be-- to take a classic analogy-- trying to get to the moon by climbing trees.

How about this? Just to give one example that is familiar, because we saw this wonderful talk by Demis-- and I think many people had seen the DeepMind work. And I hope everybody here saw Demis' talk.

This is just a couple of slides from their *Nature* paper, where, again, they had this deep Q-network, which is I think a great example of trying to see how far you can go with this pattern recognition idea, right? In a sense, what this network does, if you remember, is it has a bunch of sort of convolutional layers and of fully connected layers. But it's mapping.

It's learning a feedforward mapping from images to joystick action. So it's a perfect example of trying to solve interesting problems of intelligence. I think that the problems of video gaming AI are really cool ones.

With this pattern classification, they're basically trying to find patterns of pixels in Atari video games that are diagnostic of whether you should move your joystick this way or that way or press the button this way or that way, right? And they showed that that can give very competitive performance with humans when you give it enough training data and with clever training algorithms, right? But I think there's also an important sense in which what this is doing is quite different from what humans are doing when they're learning to play one of these games.

And, you know, Demis, I think is quite aware of this. He made some of these points in his talk and, informally, afterwards, right? There's all sorts of things that a person brings to the problem of learning an Atari video game, just like your question of what do you bring to learning this.

But I think from a cognitive point of view, the real problem of intelligence is to understand how learning works with the knowledge that you have and how you actually build up that knowledge. I think that at least the current DeepMind system, the one that was published a few months ago, is not really getting that question. It's trying to see how much you can do without really a causal model of the world.

But as I think Demis showed in his talk, that's a direction, among many others, that I think they realized they need to go in. A nice way to illustrate this is just to look at one particular video game. This is a game called *Frostbite.*

It's one of the ones down here on this chart, which the DeepMind system did particularly poorly on in terms of getting only about 6% performance relative to humans. But I think it's interesting and informative. And it really gets to the heart of all of the things we're talking about here.

To contrast how the mind system as well as other attempts to do sort of powerful scalable deep reinforcement learning, I'll show you another more recent result from a different group in a second. Contrast how those systems learn to play this video game with how a human child might learn to play a game, like that kid over there who's watching his older brother play a game, right? So the DeepMind system, you know, gets about 1,000 hours of game play experience, right?

And then it chops that up in various interesting ways with the replay that Demis talked about, right? But when we talk about getting so much from so little, the basic data is about 1,000 hours of experience. But I would venture that a kid learns a lot more from a lot less, right? The way a kid actually learns to play a video game is not by trial and error for 1,000 hours, right?

I mean, it might be a little bit of trial and error themselves. But, often, it might be just watching someone else play and say, wow, that's awesome. I'd like to do that. Can I play? My turn. My turn-- and wrestling for the joystick and then seeing what you can do.

And it only takes a minute, really, to figure out if this game is fun, interesting, if it's something

you want to do, and to sort of get the basic hang of things, at least of what you should try to do. That's not to say to be able to do it. So I mean, unless you saw me give a talk, has anybody played this game before?

OK. So perfect example-- let's watch a minute of this game and see if you can figure out what's going on. Think about how you learn to play this game, right? Imagine you're watching somebody else play. This is a video of not the DeepMind system, but of an expert human game player, a really good human playing this, like that kid's older brother.

[VIDEO PLAYBACK]

[END PLAYBACK]

OK. Maybe you've got the idea. So, again, only people who haven't seen before, so how does this game work? So probably everybody noticed, and it's maybe so obvious you didn't even mention it, but every time he hits a platform, there's a beep, right? And the platform turns blue. Did everybody notice that? Right.

So it only takes like one or two of those, maybe even just one. Like, beep, beep, woop, woop, and you get that right away. That's an important causal thing.

And it just happened that this guy is so good, and he starts right away. So he goes, ba, ba ba, ba, ba, and he's doing it about once a second. And so there's an illusory correlation.

And the same part of your brain that figures out the actually important and true causal thing going on, the first thing I mentioned, figures out this other thing, which is just a slight illusion. But if you started playing it yourself, you would quickly notice that that wasn't true, right? Because you'd start off there.

Maybe you would have thought of that for a minute. But then you'd start off playing. And very quickly, you'd see you're sitting there trying to decide what to do.

Because you're not as expert as this person. And the temperature's going down anyway. So, again, you would figure that out very quickly. What else is going on in this game?

AUDIENCE: He has to build an igloo.

JOSH He has to build an igloo, yeah. How does he build an igloo?

TENENBAUM:

AUDIENCE: Just by [INAUDIBLE].

JOSH TENENBAUM: Right. Every time he hits one of those platforms, a brick comes into play. And then what, when you say he has to build an igloo?

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: Yeah. And then what happens?

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: What, sir?

AUDIENCE: [INAUDIBLE]

JOSH TENENBAUM: Right. He goes in. The level ends, he gets some score for. What about these things? What are these, those little dust on the screen?

AUDIENCE: Avoid them.

JOSH TENENBAUM: Avoid them. Yeah. How do you know?

AUDIENCE: He doesn't actually [INAUDIBLE].

AUDIENCE: We haven't seen an example.

JOSH TENENBAUM: Yeah. Well, an example of what? We don't know what's going to happen if he hits one.

AUDIENCE: We assume [INAUDIBLE].

JOSH TENENBAUM: But somehow, we assume-- well, it's just an assumption. I think we very reasonably infer that there's something bad will happen if he hits them. Now, do you remember of some of the other objects that we saw on the second screen? There were these fish, yeah. What happens if he hits those?

AUDIENCE: He gets more points

JOSH TENENBAUM: He gets points, yeah. And he went out of his way to actually get them. OK. So you basically figured it out, right? It only took you really literally just a minute of watching this game to figure out a lot.

Now, if you actually went to go and play it after a minute of experience, you wouldn't be that good, right? It turns out that it's hard to coordinate all these moves. But you would be kind of excited and frustrated, which is the experience of a good video game, right? Anybody remember the *Flappy Bird* phenomenon?

AUDIENCE: Yeah

JOSH TENENBAUM: Right. This was this, like, sensation, this game that was like the stupidest game. I mean, it seemed like it should be trivial, and yet it was really hard. But, again, you just watch it for a second, you know exactly what you're supposed to do. You think you can do it, but it's just hard to get the rhythms down for most people.

And certainly, this game is a little bit hard to time the rhythms. But what you do when you play this game is you get, from one minute, you build that whole model of the world, the causal relations, the goals, the subgoals. And you can formulate clearly what are the right kinds of plans.

But to actually implement them in real time, but without getting killed is a little bit harder. And you could say that, you know, when the child is learning to walk there's a similar kind of thing going on, except usually without the danger of getting killed, just danger falling over a little bit. OK. Contrast that learning dynamics-- which, again, I'm just describing anecdotally.

One of the things we'd like to do actually as one of our center activities and it's a possible project for students, either in our center or some of you guys if you're interested-- it's a big possible project-- is to actually measure this, like actually study what do people learn from just a minute or two or very, very quick learning experience with these kinds of games, whether they're adults like us who've played other games or even young children who've never played a video game before. But I think what we will find is the kind of learning dynamic that I'm describing. It will be tricky to measure it. But I'm sure we can.

And it'll be very different from the kind of learning dynamics that you get from these deep reinforcement networks. Here, this is an example of their learning curves which comes not

from the DeepMind paper, but from some slightly more recent work from Pieter Abbeel's group which basically builds on the same architecture, but shows how to improve the exploration part of it in order to improve dramatically on some games, including this *Frostbite* game. So this is the learning curve for this game you just saw.

The black dashed line is the DeepMind system from the *Nature* paper. And they will tell you that their current system is much better. So I don't know how much better. But, anyway, just to be fair, right?

And, again, I'm essentially criticizing these approaches saying, from a human point of view, they're very different from humans. That's not to take away from the really impressive engineering in AI, machine learning accomplishments that these systems are doing. I think they are really interesting.

They're really valuable. They have scientific value as well as engineering value. I just want to draw the contrast between what they're doing and some other really important scientific and engineering questions that are the ones that we're trying to talk about here.

So the DeepMind system is the black dashed line. And then the red and blue curves are two different versions of the system from Pieter Abbeel's group, which is basically the same architecture, but it just explores a little bit better. And you can see that the x-axis is the amount of experience. It's in training epochs.

But I think, if I understand correctly, it's roughly proportional to like hours of gameplay experience. So 100 is like 100 hours. At the end, the DeepQ network in the *Nature* paper trained up for 1,000.

And you're showing there the asymptote. That's the horizontal dashed line. And then this line here is what it does after about 100 iterations.

And you can see it's basically asymptoted in that after 10 times as much, there's a time lapse here, right? 10 times as much, it gets up to about there. OK. And impressively, Abbeel's group system does much better.

After only 100 hours, it's already twice as good as that system. But, again, contrast this with humans, both what a human would do and also where the human knowledge is, right? I mean, the human game player that you saw in here, by the time it's finished the first screen, is

already like up here, so after about a minute of play.

Now, again, you wouldn't be able to be that good after a minute. But essentially, the difference between these systems is that the DeepQ network never gets past the first screen even with 1,000 hours. And this other one gets past the first screen in 100 hours, kind of gets to about the second screen. It's sort of midway through the second screen.

In this domain, it's really interesting to think about not what happens scientifically. It's really interesting to think about not what happens when you had 1,000 hours of experience with no prior knowledge, because humans just don't do that on this or really any other task that we can study experimentally. But you can study what humans do in the first minute, which is just this blip like right here.

I think if we could get the right learning curve, you know, what you'd see is that humans are going like this. And they may asymptote well before any of these systems do. But the interesting human learning part is what's going on in the first minute, more or less or the first hour, with all of the knowledge that you bring to this task as well as how did you build up all that knowledge.

So you want to talk about learning to learn and multiple task learning, so that's all there, too. I'm just saying in this one game that's what you can study I think, or that's where the heart of the matter is of human intelligence in this setting. And I think we should study that.

So, you know, what I've been trying to do here for the last hour is motivate the kinds of things we should study if we want to understand the aspect of intelligence that we could call explaining, understanding, the heart of building causal models of the world. We can do it. But we have to do it a little bit differently.

In a flash, that's the first problem, I started with. How do we learn a generalizable concept from just one example? How can we discover causal relations from just a single observed event, like that, you know, jumping on the block and the beep and so on, which sometimes can go wrong like any other perceptual process?

You can have illusions. You can see an accident that isn't quite right. And then you move your head, and you see something different. Or you go into the game, and you realize that it's not just touching blocks that makes the temperature go down, but it's just time.

How do we see forces, physics, and see inside of other minds even if they're just a few shapes

moving around in two dimensions? How do we learn to play games and act in a whole new world in just under a minute, right? And then there's all the problems of language, which I'm not going to go into, like understanding what we're saying and what you're reading here-- also, versions of these problems.

And our goal in our field is to understand this in engineering terms, to have a computational framework that explains how this is even possible and, in particular, then how people do it. OK. Now, you know, in some sense cognitive scientists and researchers, we're not the first people to work on this. Philosophers have talked about this kind of thing for thousands of years in the Western tradition.

It's a version of the problem of induction, the problem of how do you know the sun is going to rise tomorrow or just generalizing from experience. And for as long as people have studied this problem, the answer has always been clear in some form that, again, it has to be about the knowledge that you bring to the situation that gives you the constraints that allows you to fill in from this very sparse data. But, again, if you're dissatisfied with that is the answer, of course, you should be.

That's not really the answer. That just raises the real problems, right? And these are the problems that I want to try to address in the more substantive part of the morning, which is these questions here.

So how do you actually use knowledge to guide learning from sparse data? What form does it take? How can we describe the knowledge? And how can we explain how it's learned?

How is that knowledge itself constructed from other kinds of experiences you have combined with whatever, you know, your genes have set up for you? And I'm going to be talking about this approach. And you know, again, really think of this as the introduction to the whole day. Because you're going to see a couple of hours from me and then also from Tomer more hands on in the afternoon.

This is our approach. You can give it different kinds of names. I guess I called it generative models, because that's what Tommy likes to call it in CBMM. And that's fine.

Like any other approach, you know, there's no one word that captures what it's about. But these are the key ideas that we're going to be talking about. We're going to talk a lot about generative models in a probabilistic sense.

So what it means to have a generative model is to be able to describe the joint distribution in some form on your observable data with some kind of latent variables, right? And then you can do probabilistic inference or Bayesian inference, which means conditioning on some of the outputs of that generative model and making inferences about the latent structure, the hidden variables, as well as the other things.

But crucially, there's lots of problematic models, but these ones have very particular kinds of structures, right? So the probabilities are not just defined in statisticians terms. But they're defined on some kind of interestingly structured representation that can actually capture the causal and compositional things we're talking about, that can capture the causal structure of the world in a composable way that can support the kind of flexibility of learning and planning that we're talking about.

So a key part of how you do this sort of work is to understand how to build probabilistic models and do inference over various kinds of richly structured symbolic representations. And this is the sort of thing which is a fairly new technical advance, right? If you look in the history of AI as well as in cognitive science, there's been a lot of back and forth between people emphasizing these two big ideas, the ideas of statistics and symbols if you like, right?

And there's a long history of people sort of saying one of these is going to explain everything and the other one is not going explain very much or isn't even real, right? For example, some of the debates between Chomsky in language in cognitive science and the people who came before him and the people who came after him had this character, right? Or some of the debates in AI in the first wave of neural networks, people like Minsky, for example, and spend some of the neural network people like Jay McClelland initially-- I mean, I'm mixing up chronology there. I'm sorry.

But you know, you see this every time whether it's in the '60s or the '80s or now. You know, there's a discourse in our field, which is a really interesting one. I think, ultimately, we have to go beyond it. And what's so exciting is that we are being starting to go beyond it.

But there's been this discourse of people really saying, you know, the heart of human intelligence is some kind of rich symbolic structures. Oh, and there's some other people who said something about statistics. But that's like trivial or uninteresting or never going to anything.

And then some other people often responding to those first people-- it's very much of a back and forth debate. It gets very acrimonious and emotional saying, you know, no, those symbols are magical, mysterious things, completely ridiculous, totally useless, never worked. It's really all about statistics. And somehow something kind of maybe like symbols will emerge from those.

And I think we as a field are learning that neither of those extreme views is going to get us anywhere really quite honestly and that we have to understand-- among other things. It's not the only thing we have to understand. But a big thing we have to understand and are starting to understand is how to do probabilistic inference over richly structured symbolic objects.

And that means both using interesting symbolic structures to define the priors for probabilistic inference, but also-- and this moves more into the third topic-- being able to think about learning interesting symbolic representations as a kind of probabilistic inference. And to do that, we need to combine statistics and symbols with some kind of notion of what's sometimes called hierarchical probabilistic models. Or it's a certain kind of recursive generative model where you don't just have a generative model that has some latent variables which then generate your observable experience, but where you have hierarchies of these things-- so generative models for generative models or priors on priors.

If you've heard of hierarchical Bayes or hierarchical models and statistics, it's a version of the idea. But it's sort of a more general version of that idea where the hypothesis space and priors for Bayesian inference that, you know, you see in the simplest version of Bayes' rule, are not considered to be just some fixed thing that you write down and wire up and that's it.

But rather, they themselves could be generated by some higher level or more abstract probabilistic model, a hypothesis space of hypothesis spaces, or priors on priors, or a generative model for generative models. And, again, there's a long history of that idea. So, for example, some really interesting early work on grammar induction in the 1960s introduced something called grammar grammar, where it used the grammar, a formal grammar, to give a hypothesis base for grammars of languages, right?

But, again, what we're understanding how to do is to combine this notion of a kind of recursive abstraction with statistics and symbols. And you put all those things together, and you get a really powerful tool kit for thinking about intelligence.

There's one other version of this big picture which you'll hear about both in the morning and in

the afternoon, which is this idea of probabilistic programs. So when I would give a kind of tutorial introduction about five years ago-- oops, sorry-- I would say this. But one of the really exciting recent developments in the last few years is in a sense a kind of unified language that puts all these things together.

So we can have a lot fewer words on the slide and just say, oh, it's all a big probabilistic program. I mean, that's way simplifying and leaving out a lot of important stuff. But the language of probabilistic programs that you're going to see in little bits in my talks and much more in the tutorial later on is part of why it's a powerful language, or really the main reason.

It's that it just gives a unifying language and set of tools for all of these things, including probabilistic models defined over all sorts of interesting symbolic structures. In fact any computable model, any probabilistic model defined on any representation that's computable can be expressed as a probabilistic program. It's where Turing universal computation meets probability.

And everything about hierarchical models, generative models for generative models, or priors on priors, hypothesis space by hypothesis space, can be very naturally expressed in terms of probabilistic programs, where basically you have programs that generate other programs. So if your model is a program and it's a probabilistic generative model-- so it's a probabilistic program-- and you want to put down a generative model for generative models that can make learning into inference recursively up in higher levels of abstraction, you just add a little bit more to the probabilistic program. And so it's a very both beautiful, but also extremely useful model building tool kit.

Now, there's a few other ideas that go along with these things which I won't talk about. The content of what I'm going to try to do for the rest of the morning and what you'll see for the afternoon is just to give you various examples and ways to do things with the ideas on these slides. Now, there's some other stuff which we won't say that much about.

Although, I think Tomer, who just walked in-- hey-- you will talk a little about MCMC, right? And we'll say a little bit about item four, because it goes back to these questions I started off with also that are very pressing. And they're really interesting ones for where neural networks meet up with generative models.

You know, just how can we do inference and learning so fast and not just from few examples-- that's what this stuff is about-- but just very quickly in terms of time? So we will say a little bit

about that. But all of these, every item, component of this approach, is a whole research area in and of itself.

There are people who spend their entire career these days focusing on how to make four work and other people who focus on how to use these kind of rich probabilistic models to guide planning and decision making, or how to relate them to the brain. Any one of these you could spend more than a career on. But what's exciting to us is that with a bunch of smart people working on these and kind of developing common languages to link up these questions, I think we really are poised to make progress in my lifetime and even more in yours.