ALEX KELL: So, let's talk about the historical arcs of vision and auditory sciences. In the mid 20th century, auditory psychophysics was, like, a pretty robust and diverse field. But currently there are very few auditory faculty in psychology departments, whereas, like, vision faculty are the cornerstone of basically every psychology department in the country.

In a similar vein, automated speech recognition is this big fruitful field, but there's not much kind of broader automated sound recognition. In contrast, computer vision is this huge field. So, I'm just kind of wondering historically and sociologically, how did we get here?

JOSH TENENBAUM: You probably talk about this more.

JOSH MCDERMOTT: Sure, yeah. I'm happy to tell you my take on this. Yes, so it's-- I think there are really interesting case studies in the history of science. If you go back to the '50s, psychoacoustics was sort of a centerpiece of psychology.

And if you looked around at, like, you know, the top universities, they all had really good people studying hearing. And even some of the names that you know, like Green and Swets, you know, the guys that invented signal detection theory, pretty much. They were psychoacousticians. People often forget that. But they literally wrote the book on signal detection theory.

JOSH TENENBAUM: When people talked about separating signal from the noise, they meant actual noise.

JOSH MCDERMOTT: Yeah, and there was this famous psychoacoustics lab at Harvard that everybody kind of passed through. So back then, what constituted psychology was something pretty different. And it was really kind of closely related to things like signal detection theory. And it was pretty low level by the standards of today.

And what happened over time is that hearing kind of gradually drifted out of psychology

departments and vision became more and more prominent. I think the reason for this is that there's one really important-- there are several forces here, but one really important factor is that hearing impairment is something that really involves abnormal functioning at the level of the cochlea.

So the actual signal processing that's being done at the cochlea really changes when people start to lose their hearing. And so there's always been pretty strong impetus coming, in part, from the NIH to try to understand hearing impairment and to know how to treat that. And knowing what happens at the front end of the auditory system really has been critical to making that work.

In contrast, most vision impairments are optical in nature, and you fix them with glasses. Right? So it's not like studying vision is really going to help you understand visual impairment. And so there was never really that same thing. And so, when psychology sort of gradually got more and more cognitive, vision science went along with it. That really didn't happen with hearing.

And I think part of that was the clinical impetus to try to continue to understand the periphery. The auditory periphery was also just harder to work out, because it's a mechanical device, the cochlea. And so you can't just stick an electrode into it and characterize it. It's actually really technically challenging to work out what's happening. And so that just kept people busy for a very long time.

But as psychology was sort of advancing, what people in hearing science were studying kind of ceased to really be what psychologists found interesting. And so the field kind of dropped out of psychology departments and moved into speech and hearing science departments, which were typically at bigger state schools. And they never really got on the cognitive bandwagon in the same way that everybody in vision did.

And then what ends up happening in science is there's this interesting phenomenon where people get trained in fields where there are already lots of scientists. So if you're a grad student, you need an advisor, and so you often end up working on something that your advisor does. And so if there's some field that is under-represented, it typically gets more under-represented as time goes on. And so that's sort of been the case.

You know, if you want to study, you know, olfaction, it's a great idea, right? But how are you

going to do it? You've got to find somebody to work with. There's not many places you can go to get trained. And the same has been true for hearing for a long time. So that's my take on that part of it.

Hynek, do you have anything to say on the computational?

HYNEK HERMANSKY: No, I don't know. I'm thinking, if it is something that also evolved with tools available. Because in the old days it was easier to generate the sounds than to generate images. On the computer now, it's much easier, right? So vision and visual research became sexier.

So I teach auditory perception to engineers. Also I teach a little bit of visual perception. And I notice that they are much more interested in visual perception, especially the various effects. And, you know, because somehow you can see it better.

And the other thing is, of course, funding. I mean, you know, the hearing research's main applications are, as you said, hearing prostheses. These people don't have much money, right? The speech recognition, we didn't get much of the benefits yet from hearing research, unfortunately. So I wonder if this is not also a little bit--

JOSH TENENBAUM: Can I add one? So, maybe-- this is sort of things you guys also gesture towards, but I think in both-- to go back towards similarities and not just differences. Maybe that's what will be my theme. In both vision and hearing or audition, there's, I think, a strong bias towards the aspects of the problem that fit with the rest of cognition.

And often that's mediated by language, right? So there's been a lot of interest in vision on object recognition, parts of vision that ultimately lead into something like attaching a word to a part of an image or a scene. And there's a lot of other parts of vision, like certain kinds of scene understanding, that have been way understudied until recently also, right?

And it does seem like the parts of hearing that have been the focus are speech. I mean, there are lots of people in-- it's maybe not as much as vision or object recognition, but certainly there's a lot of mainline cognitive psychology who studied things like categorization of basic speech elements and things that just start to bleed very quickly into psycholinguistics. Right?

Whereas, the parts of hearing-- at least that's been where a lot of the focus. But the parts of hearing that are more about auditory scene analysis in general, like sound textures or physical events or all the richness of the world that we might get through sound, has been super understudied. Right?

But echoing something Josh said also or implicitly is, just because it might be understudied and you need to find an advisor doesn't mean you shouldn't work on it. So if you have any interested in this, and there's even one person, say, who's doing some good work on it, you should work with them. And it's a great opportunity.

JOSH MCDERMOTT: If I could follow up and just say that my sense of this is, if you can bear with it and figure out a way to make it work, in the long run it's actually a great place to be. It's a lot of fun to work in an area that's not crowded.

ALEX KELL: All right. Obviously, a large-- to kind of transition, a large emphasis of the summer school is on potential symbiosis between, like, machine and, like, engineering and science. And so, what have been the most kind of fruitful interactions between machine perception, either vision or hearing, over the years, and are there any kind of lessons that we can learn in general?

HYNEK HERMANSKY: You know, I went a little bit backwards, quite frankly. I'm trained as an engineer, and I was paid always to build better machines. And in the process of building better machines, you discover that, almost unknowingly, we were emulating some properties of hearing.

So then I, of course, started to be interested, and I wanted to get more of it. So that's how I got into that. But I have to admit that in my field, we are a little bit looked down at.

[INTERPOSING VOICES]

HYNEK HERMANSKY: Because mainly not all that much, because engineers are such that [INAUDIBLE] that if something works, they like it and they don't much want to know why, you know? Or at least they don't talk much about it.

So it's interesting when I'm in engineering meetings, they look at me as this strange kid who works also on hearing. And when I'm in an environment like this, people look at me like a speech engineer. But, I mean, I don't even feel either in some ways.

But what I was wondering when Josh was talking about, is there anybody in this world who works on both? Because, you know, that, I think, is much more needed than anything else. I mean, somebody who is interested in both audio and visual processing and is capable of making this--

JOSH So not both science and engineering, but both vision and audition.

TENENBAUM:

HYNEK HERMANSKY: Yes, that's what I mean, vision and audition, with a real goal of trying to understand both and trying to find the similarities. Because, personally, I got some inspiration from visual research, even when I work with audio. But I don't see much of it anymore.

And there are some basic questions which I would like to ask-- and maybe I should have sent it in-- which is like, I don't even know which most are the similar and different in audio and vision. Should I look at time? Should I look at modulations? Should I look at the frequencies? Should I look at the spatial resolution? And so on and so on.

So I don't know if somebody can help me with this. I would be very happy to go home knowing that. I mean, sometimes I suspect that spatial resolution and frequency resolution in hearing are similar. I'm thinking about modulations in speech. Josh talked a lot about it.

But there must be a modulation in vision also. But, of course, we never studied much of the vision of the moving pictures. A lot of vision research was fixed. Basically, the images, right? It was a little bit like, in speech, we used to study vowels. We don't do it anymore, because the area of speech is something very, very different.

The same thing is in image processing. I think that now it's getting more and more, because, again, I mean I'm going back to availability of the machines. You can do some work on moving images and on video and so on and so on. But I don't know how much of that is happening.

And so my question is really, are there any similarities, and on which level they are? [INAUDIBLE] There is a time, there is a spatial frequency, there is a carrier frequency, there are modulations in speech. I don't know. I mean, I would like to know this. I mean, that's something somebody can help me.

DAN YAMINS: Oh, I was-- I had-- I think those are interesting questions, but I actually-- I'm sure I don't have the answers at this point. But I was going to say something a little-- you know, going back to the original general question, which is, again, you know, this sort of thinking about it from a historical point of view I think is helpful.

In the long run, I think what's happened over the past 50 years is that biological inspiration has been very helpful at injecting ideas into the engineering realm that end up being very powerful. Right? I mean, I think we're seeing kind of the arc of that right now in a very strong way.

I mean, you know, in terms of vision and audition, the sort of algorithms that are most dominant are ones that were strongly biologically inspired. And there had been a historical arc over, I think, a period of decades, where, first the algorithms were sort of biologically inspired back in the '50s and '60s, and then they were not biologically inspired for a while. Like, the biology stuff didn't seem to be panning out.

That was sort of the dark ages of, kind of, neural networks. And then more recently that has begun to change. And, again, biologically inspired ideas seem to be very powerful, for creating, you know, algorithmic approaches. But the arc is a very long one, right?

And so it's not like, you know, you discover something in the lab and then the next day you would go implement it in your algorithm and suddenly you get 20% improvement on some task. All right? That's not realistic. OK? But if you're willing to have the patience to wait for a while, and sort of see ideas sort of slowly percol up, I think they can be very powerful.

Now, the other direction is really also very interesting, like using the algorithms to understand neuroscience, right? That's one where you can get a lot of bang for your buck quickly, right? And it's sort of like-- it's like that's a short-term high, right?

Because what happens is that you take this machine that you didn't understand and you apply it to this problem that you were worried about and that is sort of scientifically interesting, and suddenly you get 20% improvement overnight. Right?

That is feasible in that direction, e.g., taking advances from the computational side or on the algorithmic side and applying them to understanding data in neuroscience. That does seem to have been borne out, but only much more recently. So, there wasn't much of that at all for many decades. But more recently that has begun to happen.

And so I think that there is a really interesting open question right now as to which thing, which direction, is more live. Which one is leading at this point, I think, is a really interesting question. Maybe neither of them is leading. But I think that certainly on the everyday, on-the-ground experience, as somebody who is trying to do some of both, it feels like the algorithms are leading the biology. OK?

Are leading the neuroscience, to me, in the sense that I feel like, in the short run at least, things are going to come out of the community of people doing algorithms development that are going to help understand neuroscience data before specific things are going to come out

of the neuroscience community that are going to help make better algorithms, like in the short run. OK?

Again, I think the long run can be different. And I think that's a really deep open research program question, is which tasks are the ones that you should choose such that learning about them from a neuroscience and psychology point of view will help, in the five- to 10-year run, make better algorithms. And if you can choose those correctly, I think you really have done something valuable. But I think that's really hard.

ALEX KELL: Yeah. I want to push back. I like how you said how the engineering is really helping the science right now, but the science-- the seed that science planted in the engineering. Like, what you're talking about is just CNNs, basically.

DAN YAMINS: Well, I think not entirely, because I think that there are some ideas in recurrent neural networks.

ALEX KELL: OK, sure. Neural networks generally. But the point is the ideas that were kind of inspired from that have been around for decades and decades and decades. Are there other kinds of key examples besides, like, the operations that you throw into a CNN? The idea of convolution and the idea of layered computation-- these are obviously very, very important ideas, but, like, what are kind of other contributions that science has given engineering besides--?

DAN YAMINS: Well, Green and Swets. I mean, the thing that he mentioned earlier about Green and Swets is another great example.

ALEX KELL: Yeah.

DAN YAMINS: Right.? Psychophysics helped understand signal detection theory. But that's much older, but that's a very clear example.

HYNEK HERMANSKY: Signal detection theory didn't come from Green and Swets. It came from Second World War.

DAN YAMINS: I was just thinking of all the work--

HYNEK HERMANSKY: They did very good work obviously, and they indeed were auditory people.

DAN YAMINS: And they were actually-- they were doing a lot of work during-- government--

JOSH MCDERMOTT: They formalized a lot of stuff.

DAN YAMINS: Yeah, and they did a lot--

HYNEK HERMANSKY: Yeah, I don't want to take anything away from them.

DAN YAMINS: But, you know, it's interesting. There's this great paper that Green and Swets have where they talked about their--

HYNEK HERMANSKY: --was engineering.

DAN YAMINS: They talked about their military work, right? And they did-- they actually worked for the military, just, like, determining which type of plane was the one that they were going to be facing. And so, yeah, I agree that came out of that.

HYNEK HERMANSKY: If I want to still-- if I can still spend a little bit of time on engineering versus science, we also are missing one big thing, which is, like, Bell Labs. Bell Labs was the organization which paid people for doing-- having fun. Doing really good research.

There was no question that, at the time, Bell Labs were about speech and about audio. So there was-- a lot of things were justified. And even, like, Bela Julesz and these people-- they pretended they are working on perception because the company wanted to make more money on the telephone calls. This has gone. Right?

Both. Speech is gone. Bell Labs is gone. And maybe image is high-- image processing is in, because the government is interested in finding various things from the images and so on and so on. So, a lot of that is funding.

Since you mentioned neural networks, it never stops amazing me that people would call artificial neural networks anything similar to biology. I mean, the only thing which I see similar there maybe are now these layered networks and that sort of things.

ALEX KELL: I think a lot of the concepts were inspired by that. I don't think it was, like, directly-- like, I don't think anyone takes it as a super serious--

HYNEK HERMANSKY: But there I still have maybe one point to make. Most of the ideas which are now being explored in neural networks are also very old. The only thing is that we didn't have the hardware. We didn't have the means, basically, of doing so.

So technology really supported this, and suddenly, yeah, it's working. But to some people it's not even surprising that it's working. They say, of course. They say, we couldn't do it.

DAN YAMINS: I think what was surprising to them was that it didn't work for so long and that people were very disappointed and upset about that. And then, you know-- but I agree that basically there's all these, like-- all the ideas are these 40-year-old or 50-year-old ideas that people had thought of, typically many of them coming out of the psychology and neuroscience community a long time ago but just couldn't do anything about it. And so that takes a long time to bear fruit, it feels like.

GABRIEL KREIMAN: So I have more questions rather than answers, but to try to get back to a question about vision and hearing and how we can synergistically interact between the two. First, I wanted to lay out a couple of biases and almost religious beliefs I have on the notion that cortex is cortex, meaning that there's a six-layer structure, that there are some patterns of connectivity that have been described both in the vision-- visual cortex as well as auditory cortex.

They are remarkably similar, and we have to work with the same type of hardware in both cases. The type of plasticity learning rules that have been described are very similar in both cases. So there's a huge amount of similarity to the biological level. We use a lot of the same vocabulary in terms of describing problems about invariants and so on.

And yet, at the same time, I wanted to raise a few questions, particularly to demonstrate my ignorance in terms of the auditory world and get answers from these two experts here. I cannot help but feel that maybe there's a nasty possibility that there are differences between the two, in particular the role of timing has been somewhat under-explored in the visual domain.

We have done some work on this, some other people have. But it seems that timing plays a much more fundamental role in the auditory domain. Perhaps the most extreme example is sound localization, where we need to take into account micro-second differences in the arrival of signals within the two ears.

I don't know anything even close to that in the visual domain. So that's one example where I

think we have to say that there is a fundamental difference.

Now thinking more about sort of the recognition questions that many of us are interested in, I think, again, timing seems to play a fundamental role in the auditory domain. But I would love to hear from these two experts here.

I easily come up with questions about what is an object in the auditory domain that's sort of defined in a somewhat heuristic way in the visual domain? But we all sort of agree on what objects are. And I don't know what the equivalent is in the auditory domain.

And how much attention should we pay to the fact that the temporal evolution of signals is a fundamental aspect in the auditory world, which we don't really-- by and large, we don't really think about too much in the visual domain. With that said, I do hope that at the end we will find similar fundamental principles and algorithms, because, as I said, cortex is cortex.

JOSH MCDERMOTT: I can speak to some of those issues a little bit. Look, I think it can be-- I mean, it's an interesting and fun and, I think, often useful exercise to try to map, kind of, concepts from one modality onto another.

But, again, at the end of the day, the purpose of perception is just to figure out what's out there in the world, and the information that you get from different modalities, I think in some cases it just tells you about different kinds of things. So sound is usually created when something happens, right? It's not quite-- it's not quite the same thing as there being an object there off of which light reflects.

I mean, sometimes there's, in some sense which there's an object there. Like a person, right? Persons producing sound. But, oftentimes, the sound is produced by an interaction between a couple of different things. So, really, the question is sort of what happened, as much as what's there.

And so, you could probably try to find things that are analogous to objects, but it's-- in my mind it may just not be exactly the right question to be asking about the sound.

JOSH TENENBAUM: Can I just comment on that one? Yeah, I mean, I think, again, this is a place where Gabriel and I have somewhat different biases, although, again, it's all open. But an object to me is not a visual thing or an auditory thing. An object is a physical thing, right?

So those of you who saw Liz Spelke's lectures on this, this is very inspiring to me that from

very early on infants have a concept of an object, which is basically a thing in the world that can move on its own or be moved. And the same principles apply in vision but also haptics.

And, you know, it's true that the main way we perceive objects is not through audition, but we can certainly perceive things about objects from sound and often just, echoing what Josh said, it's the events or the interactions between objects that make sounds. They make the-- physically cause sounds. And so it's often what we're learning from sound is--

GABRIEL KREIMAN: But maybe if I could ask-- I don't disagree with your definition of objects, a la Spelke and so on. But I guess in the auditory domain, if I think about speech, you know, are we talking about phonemes? Are we talking about words? I mean, if we talk about Lady Gaga or Vivaldi, are we talking about a whole piece of music, a measure, a tone, a frequency? These are things that--

JOSH TENENBAUM: So, structure, sort of structure more generally.

GABRIEL KREIMAN: What's the unit of computation that we should think about algorithmically? In the same way that Dan and us and many others think about algorithms that will eventually have labels and objects, for example. I mean, what are those fundamental units? And maybe the answer is all of them, but--

JOSH TENENBAUM: Well, speech is really interesting, because from one point of view, you could think of it as, like, what-- it's basically, like-- it's an artifact, right? Speech is a thing that's created through biological and cultural evolution, manipulating a system to kind of create these artificial event categories, which we can call phonemes and words and sentences and so on.

And, you know, surely there was audition before there was speech, right? So, it seems like it's building on a system that's going to detect a more basic notion of events, physical interactions, or things like babbling brooks or fires or breezes. And then animal communication. And it hacks that, basically, both on the production side and the perception side.

So it's very interesting to ask what's the structure? What's the right way to describe the structure in speech? It probably seems most analogous to something like gesture, you know? That's a way to hack the visual system to create these events visually. Salient changes in motion, whether for just non-verbal communication or something in sign language. It's super interesting, right?

But it's, again-- I wouldn't say-- the analog-- speech isn't a set of objects. It's a set of

structured events, which have been created to be perceivable by a system which was evolutionarily much more ancient one, perceiving object interactions and events.

JOSH MCDERMOTT: But I also think it's a case that-- yeah, there's a lot of focus on objects and vision, but it's certainly the case that vision is richer than just being about objects, right? I mean, you have-- there-- right? I mean, there's-- I think in some sense, the fact that you are posing the question, it's a reflection of where a lot of work has been concentrated on.

But, yeah, there's obviously-- you know, you have scenes, there's stuff, not just things, right? And the same is true in audition. And the difference is just that there isn't really as much of a focus on, like, things, only because those are not--

GABRIEL KREIMAN: Here's the fundamental question I'm trying to raise, as well as the question about timing. In the visual domain, let's get away from objects and think about action recognition, for example. And that's one domain where you would think that, well, you have to start thinking about time. It's actually extremely challenging to come up with good stimuli that you cannot recognize-- where you cannot infer actions from single frames. And I would argue, but--

JOSH TENENBAUM: Let's talk about events.

GABRIEL KREIMAN: But let me say one more thing. But please correct me if I'm wrong. I would argue that in the auditory domain, it's the opposite. It's very hard to come up with things that you can recognize from a single incident.

JOSH MCDERMOTT: Sure.

GABRIEL KREIMAN: You need time. Time is inherent to the basic definition of everything. In the visual domain-- again, we've thought about time and what happens if you present parts of objects asynchronously, for example. And you can disrupt object recognition or action recognition in that way. But it's sort of-- again, you can do a lot without time or without thinking too seriously about time. Then maybe, I don't know-- time is probably not one of your main preoccupations, I suspect, in the visual domain.

HYNEK HERMANSKY: I'm not sure, because one of the big things which always strikes me in vision is the saccade and the fact that we are moving eyes, and the fact that it's possible even to lose the vision,

basically, if you really fix the things on the retina, and so on and so on. So vision probably figured out different ways of introducing time into perception, basically moving eyes and maybe in sounds. Indeed, it's happening more, like, already out there.

But, you know, I had one joint project actually where we tried to work on audio-visual recognition. And it was the project about recognizing unexpected things. And that was a big pain initially, because, of course, vision people thinking one way or auditory people thinking another way. But eventually we ended up with the time and with the surprises and with the unexpected and with the priors.

And there's been a lot of similarities between audio and visual world, you know? So that's why I was maybe saying in the beginning, people should be more encouraged-- now I'm looking at the students-- to look at both. I mean, don't just say I'm a visual person and I just want to know a little bit about speech or something.

No. I mean, these things are very interesting. And, of course I mean in auditory world, there are problems that are very similar to visual problems. And in the visual world there are very similar problems to auditory work. You just take a speech and take a writing, right? And be it handwriting or being even printed things. I mean, these things communicate messages, communicate information, in a very similar way.

So, I would just say I got a little bit excited because I finished my coffee. But I would just say, let's look for the similarities rather than differences, and let's be very serious about it. Like, sort of say, oh, finally I found something. Like, for instance, I give you one little example. We had a big problem with a perceptual constancy when you get linear distortions in the signal.

And I just accidentally read some paper by David Marr at the time, and I didn't understand it. I have to say I actually missed [INAUDIBLE] a little bit. But still, it was a great inspiration. And I came up with an algorithm which ended up to be a very good one. Well, at the time. I mean, I was being beaten many times. But, you know, let's just look for the similarities. That's what I'm somehow, maybe arguing.

And that was also my quest-- like, I don't even know what is similar and different in auditory and visual signals. So, find-- certainly maybe-- on a certain level, it must be the same, right? The cortex is very, very similar. So I believe that, indeed, at the end we are getting information into our brain, which is being used for figuring out what's happening in the world. And there are these big differences at the beginning. I mean, the senses are so different.

JOSH TENENBAUM: Could I nominate one sort of thing that could be very interesting to study that's very basic in both vision and audition, of where there are some analogies? Which is certain kinds of basic events that involve physical interaction between objects. Like, I'll try to make one right here. Right? OK. So there was a visual event, and it has a low level signal-- a motion signal. There was some motion over here.

Then there was some other motion that, in a sense, was caused. There was some sound that went with it. There was the sound of the thing sliding on the table and then the sound of the collision. We have all sorts of other things like that, Right?

Like, I can drop this object here, and it makes a certain sound. And so there's very salient, low levelly detectable, both auditory and visual signals that have a common cause in the world. One thing hitting another. It's also the kind of thing which-- I don't know if Liz mentioned this in her lecture. I mentioned this a little bit.

Even very young infants, even two-month-olds, understand something about this contact causality, that one object can cause another object to move. It's the sort of thing that Shimon has shown-- Shimon Ullman has shown. You can, in a very basic way, use this to pick out primitive agents, like hands as movers.

So this is one basic kind of event that has interesting parallels between vision and audition, because there's a basic thing happening in the world, an exertion of force between one moving object when it comes into contact with another thing.

And it creates some simultaneously detectable events with analogous kinds of structure. I think a very basic question is, you know, if we were to look at the cortical representation of a visual collision event and the auditory side of that, you know? How do those work together? What are similarities or differences in the representation and computation of those kind of very basic events?

HYNEK HERMANSKY: If I still may, obvious thing to use is use vision to transcribe the human communication by speech. If somebody wants a lot of money from Amazon or Microsoft or Google or government, you know, work on that. Because there is a clear visual channel, which is being used very heavily, you know?

Not only that. I move the hands and that sort of thing. If somebody can help there, I mean, that

would be great. And it's actually a relatively very straightforward problem. I'm not saying simple. But it's well defined. Because there is a message, which is being conveyed in a communication by speech. And it's being used. I mean, lips are definitely moving, unless you are working with a machine. And hands are moving unless you are a really calm person, which none of us is. And so this is one--

JOSH TENENBAUM: Just basic speech communication.

HYNEK HERMANSKY: Basic speech communication, as Martin [INAUDIBLE] is saying. That would be great, really.

JOSH MCDERMOTT: I mean, it's also worth saying, I think, you know, most of perception is multimodal, right? And you can certainly come up with these cases where you rely on sound and have basically no information from vision and vice versa, right? But most of the time, you get both and you don't even really think about the fact that you have two modalities. You're just, you know-- you want to know what to grab or whether to run or whether it's safe to cross the street and, you know--

HYNEK HERMANSKY: Of course, the thing is that you can switch off one modality without much damage. That's OK, because in most of the perception this is always the case. You don't need all the channels of communication. You only need some. But if you want to have a perfect communication, then you would like to use it. But I absolutely agree that the world is audiovisual.

JOSH TENENBAUM: This is a comment I was going to add to our discussion list, which I shared with Alex but maybe not the rest, is I think it's a really interesting question, what can be understood about the similarities and differences in each of these perceptual modalities by studying multimodal perception?

And to put out a kind of a bold hypothesis, I think that, for reasons that you guys were just saying, because natural perception is inherently multimodal. And it's not just these ones. It also involves touch and so on. I think that's going to impose strong constraints on the representations and computations in both how vision and audition work.

The fact that they have to be able to interface with a common system, what, you know, I would think of as a kind of physical object events system. But, however you want to describe it, the fact of multimodal perception's pervasiveness, the fact that you can switch on or off sense modalities and still do something, but that you can really just so fluently, naturally bring them

together into a shared understanding of the world, that's something we can't ignore, I would say.

GABRIEL KREIMAN: Why are people so sure that in everyday life, most things are multimodal? I'm not really sure how to quantify that. But is there any quantification of this?

JOSH MCDERMOTT: No, I don't know of a quantification. All I mean is that, most of the time, I mean, you're listening and you're looking and you're doing everything you can to figure out what happened, right? I mean, it's like, you know, you want to know if there's traffic coming, right? I mean, there's noise that the cars make. You also look, you know? You do both of those. And you probably don't even really think about which of them you're doing.

GABRIEL KREIMAN: No. I'm not talking about the most of the time part. Yes, that's a very good example of multimodal experience. I can cite lots of other examples where I'm running and listening to music and they're completely decoupled. Or I'm working on my computer.

JOSH TENENBAUM: You don't listen to music when you're driving, right?

GABRIEL KREIMAN: I do, but--

GABRIEL KREIMAN: No, but no. But, I mean, not in the way that, like-- sure, you listen to music, obviously. You listen to music when we're driving, but we try-- it's sort of important that it doesn't drown out all other sounds.

GABRIEL KREIMAN: I'm just wondering to what extent this--

JOSH TENENBAUM: Ok, fine.

ALEX KELL: And how much of that is, like, kind of the particular, like, the modern-- like, in the contemporary world you can actually decorrelate these things in a way that in the natural world you can't. Like, if you are a monkey, these things would probably be a lot more correlated than you are as a human in the 21st century. Like, there would be a [INAUDIBLE] physical world causing the input to both your modalities in a way that you can break now, right? Like, I don't know. That feels--

GABRIEL KREIMAN: You may be right. I haven't really thought deeply about this.

[INTERPOSING VOICES]

GABRIEL KREIMAN: I'm not [INAUDIBLE]

JOSH MCDERMOTT: It would be interesting to compute some statistics of this.

GABRIEL KREIMAN: I'm not disputing the usefulness of multimodal perception. I think it's fantastic. I'm just wondering. I think vision can do very well without the other auditory world. And vice versa.

DAN YAMINS: We could just close our eyes right now, all of us, and we'd have a fine panel for a while.

JOSH TENENBAUM: But many of the social dynamics would be invisible, literally.

JOSH MCDERMOTT: No, I think you'd probably get a lot of reciprocity. It's an open question.

JOSH TENENBAUM: You'd get some, but, like, there's a difference. Have you ever listened to a radio talk show? Sometimes these days the shows are broadcast on TV and also-- and it's, like, when you watch you're like, oh my-- like, you have a totally different view of what's going on. Or, like, if you're there in the studio.

I mean, I totally agree that these are all open questions, and it would be nice to actually quantify, for example, what to me is this often subjective experience. Like, sometimes if the sound is, you know-- I don't know. You turn off the sound on something where you're used to having the sound, it changes your experience, right?

Or you turn on the sound in a way that you had previously watched something, right? Like, you could do experiments where you show people a movie without the sound and then you turn on the sound. You know, in some ways transform what they see and in some ways not. So, maybe the right thing to say is more data is needed.

DAN YAMINS: But don't you guys think, though, that, like, even independent of multimodal, there's still

actually a lot of even more basic questions to be asked about similarity and differences? Like, I mean, just from a very-- from my point of view since that's the only one I'm usually able to take, like, you took a bunch of convolutional neural networks and you train some of them on vision tasks and some of them on audition tasks, right?

And you figured out which architectures are good for audition tasks and which are good for vision tasks. See if the architectures are the same, and if indeed the architectures are fairly similar, then, like, looking at the differences between the features at different levels. I mean, I know that that's a very narrow way to interpret the question, but it's one. And there's probably a lot that can be--

JOSH TENENBAUM: You guys have been doing that. What have you learned from doing that?

ALEX KELL: We haven't done it that exhaustively.

DAN YAMINS: We haven't done it that exhaustively. But suffice it to say that the hints are, I think, very interesting. Like, you begin to see places where there are clear similarities and clear differences and asking, like, where did the divergence occur? Are there any underlying principles about what layers or what levels in the model those divergences start to occur? Can you see similarities at all layers or do you start to see sort of a kind of a clear branching point? Right?

Moreover, like what about lower layers, right? I mean, you start to actually see differences in sort of frequency content in auditory data and differences between that and visual data that seem to emerge very naturally from the underlying similarities. You know, underlying differences between the statistics. But still, downstream from there, there are some deep similarities about extraction of objects of some kind or other. You know, auditory objects, potentially.

And so I think that's a very narrow way of posing the question. And I don't say that everybody should pose it that way by any means. But I just think that before we get to multimodal interaction, which is interesting, I think there's just this huge space of clear, very concrete ways to ask the question of similarities and differences that are-- like, almost no matter what you'll find, you'll find something interesting.

JOSH You're saying if we enlarge the discussion from just talking about vision audition to other parts

TENENBAUM: of cognition, then we'll see more of the similarities between these sense modalities, because they will be the differences that stand out in relief with respect to the rest of cognition.

Yeah, I mean, I think that's a valuable thing to do, and it connects to what these guys were saying, which is that there's a sense in which this-- you know, something like these deep convolutional architecture seem like really good ways to do pattern recognition, right?

This is what I would see as the common theme between where a lot of the successes happened in vision and in audition. And I don't think-- and, again, everybody here has heard me say this a bunch of times-- I think that pattern recognition does not exhaust, by any means, intelligence or even perception. Like, I think even within vision and audition, there's a lot we do that goes beyond, at least on the surface, you know, pattern recognition and classification.

It's something more like building a generative model. Maybe this is a good time to-- that's another theme you wanted to bring in. But, you know, something about building a rich model of the world and its physical interactions.

And, to me, you know, and, again, something Dan and I have talked a lot about it and I think it's-- you know, you've heard some of this from me, and Dan has got some really awesome work in a similar vein of trying to understand how, basically, deep pattern recognizers-- to me, that's another way we could call deep convolutional pattern-- or just deep invariant pattern recognizers, where the invariance is over space or time windows or whatever it is that deep convolutional-- you know, these are obviously important tools.

They obviously have some connection to not just the six layer cortex architecture but these multiple-- you know, the things that goes on in, like, the ventral stream, for example. I don't know, the auditory system as well. But it's going on from one cortical area to a next. A hierarchy of processing. That seems to be a way that cortex has been arranged in these two sense modalities in particular to do a really powerful kind of pattern recognition.

And then I think there's the question of, OK, how does pattern recognition fit together with model building? And, you know, I think in other areas of cognition you see a similar kind of interchange, right? It might be-- like, this has come up a little bit in action planning-- like, model-based planning versus more model-free reinforcement learning. And those are, again, a place where there might be two different systems that might interact in some kind of way.

I think pattern recognition also is useful all over-- you know, where cognition starts to become

different from perception, for example. There's so many ways, but things like when you have a goal and you're trying to solve a problem, do something. Pattern recognition is often useful in guiding problem solving, right? But it's not the same as a plan, right?

So, I don't know if this is starting to answer your question, but I think this idea of intelligence more generally as something like-- I mean, the way Laura put it for learning, the same idea, she put it as, like, goal directed or goal constrained-- how did she put it?-- problem solving or something like that, right? That's a good way to-- if you need one general purpose definition of cognition, that's a good way to put it.

And then, on the other hand, there's pattern recognition. And so you could ask, well, how does pattern recognition more generally work and what have we learned about how it works in the cortex or computationally from studying the commonalities between these two sense modalities? And then how does pattern recognition play into a larger system that is basically trying to have goals, build models of the world, use those goals to guide its action plans on those models?

ALEX KELL: On the public of convolutional neural networks and deep learning, like, they are reaching, like, kind of impressive successes and they might eliminate some similarities and differences between the modalities. But, in both cases, the learning algorithm is extremely non-biological.

And I was wondering if any of you guys-- like, infants don't need millions of examples of label data to learn what words are. So I was wondering if you guys have any kind of thoughts on how to make that algorithm more biologically possible?

DAN YAMINS: I would go to what Josh said earlier, which is you look at those real physically embodied environment. You look for those low level cues that can be used to, like, be a proxy for the higher level information, right? And then what you really want is--

ALEX KELL: Can you be a little more specific? What do you mean?

DAN YAMINS: Well, do you want to be--

JOSH TENENBAUM: I mean, some people have heard this from Tommy and others here about, like, sort of kinds of natural supervision, right? I mean, several people have talked about this, right? Is that what you're getting at? The idea that, often, just tracking things as they move in the world gives you a lot of extra effectively labeled data. You're getting lots of different views of this microphone now, or whatever, for walking around the stage or all of our faces as we're rotating.

So, when you pointed to the biological implausibility of the standard way of training deep networks, I think a lot of people are realizing-- and this was the main idea behind Tommy's conversion to now be a strong prophet for people learning instead of being a critic, right?-- was that, the issue of needing lots of labeled training, that's not the biggest issue. There's other issues, like backpropagation as a mechanism of actually propagating error gradients all the way down to a deep network. I think that troubles more people.

DAN YAMINS: I have quite the opposite view on that.

JOSH TENENBAUM: OK.

DAN YAMINS: Yes. I agree that it's true that the specific biological plausibility of a specific deep learning, like backpropagation algorithm is probably suspect. But I suspect that by the same token, there are somewhat inexact versions that are biologically plausible or more plausible anyway that could work pretty well.

I think that's less like-- let me put it this way. I think that's a flashy question. I think if you actually end up solving that both from an algorithm point of view and maybe, more importantly, seeing how that's implemented in a kind of real neural circumstance, you'll win the Nobel Prize. But, I mean, I think that-- I feel like that's something that will happen, right?

I think that there is a bigger question out there, which is, you know-- I do think that from an algorithmic point of view which things that people don't yet know how to do, how to replace, like, millions of heavily semantic training examples with those other things, right? Like, the things that you just mentioned a moment ago, like, the extra data.

Like, it hasn't actually been demonstrated how to really do that. And I feel like the details of getting that right will tell us a lot about the signals that babies and others are paying attention to in a way that's really conceptually very interesting and, I think, not so obvious at this point how that's-- I think it'll happen too, but it will be conceptually interesting when it does in a way that I think that--

JOSH TENENBAUM: Both are pretty interesting.

DAN YAMINS: Yeah.

JOSH TENENBAUM: Some people are more worried about one or the other. But, yeah.

DAN YAMINS: Exactly. And, personally, I would say that, from an algorithm point of view, I'm more interested in that second one, because I think that will be a place where the biology will help us teach how to do better algorithms.

JOSH TENENBAUM: Learning about the biological mechanism backpropagation seems less likely to transform our algorithms. Although, again, if you ask Andrew Sax-- he's one person. He's been here.

He's think-- that's the question he most wants to solve, and he's a very smart person. And I think he has some thoughts on that. But I-- my sympathies are also with you there. I think there are other things besides those that are both biologically and cognitively implausible that need work, too, so those-- but those are two of the main ones that--

HYNEK HERMANSKY: I think you are touching something very interesting, and one of the major problems with machine learning in general, as I see it, which is like use of transcribed or untranscribed data. And I think that this is one direction, which actually, specifically in the speech, it's a big, big, big problem because, of course, data is expensive, so-- unless you are Google. But even there, you will want to have it transcribe data. You want to know what is inside, and clearly, this is not what--

JOSH TENENBAUM: You guys have this thing in speech. I think this is I'd like to talk more about this because you guys have this thing, particularly at Hopkins, in speech that you call zero resource speech recognition.

HYNEK HERMANSKY: Right, that's--

JOSH TENENBAUM: And I think this is a version of this idea, but it's one of the places where studying not just neuroscience. But cognition and what young children do, the ability to get so much from so little is a place where we really have a lot to learn on the engineering side from the science.

HYNEK HERMANSKY: Yes, I mean, [INAUDIBLE] that I could speak about it a little bit more in depth. But definitely, this is the direction I'm thinking about, which is like what do you do if you don't know what is in the signal, but you know there is a structure, and you know that there is information you need. And you have to start from the very scratch, figure out where information is, how is it coded,

and then use it in the machine. And I think it's a general problem in the same thing as in region.

JOSH TENENBAUM: Maybe-- you're asking several different questions. I mean, I don't know if-- have people in the summer school talked about these instabilities It's an interesting question. People are very much divided on what they say. And I do think that generative models are going to come out differently there.

But, again, I don't want to say generative models are better than discriminatively trained pattern recognizers. I think, particularly for perception and a lot of other areas, what we need to understand is how to combine the best of both. So in an audience where people are just neural networks, rah, rah, rah, rah, rah, and that's all there is, then I'm going to be arguing for the other side. But that's not that I think they are better. I think they have complementary strengths and weaknesses.

This might be one. I think pretty much any pattern classifier, whether it's a neural network or something else, will probably be susceptible to these kind of pathologies where you can basically hack up a stimulus that's arbitrarily different from an actual member of the class that gets classified. Basically, if you're trying to put a separating surface between two or n finite classes-- I was trying to see how to formulate this mathematically. I think you can basically show that it's not specific to neural networks. It should be true for any kind of discriminatively trained pattern classifier.

I think generative models have other sorts of illusions and pathologies. But they're definitely going to be-- my sense is they're going to be some of the ways that any pattern classifier is susceptible, that generative models won't be susceptible to. And there will be others that they will be susceptible to. But it's sort of an orthogonal issue.

But I think the illusions that generative models are susceptible to, are generally going to have, like, interesting, rational interpretations. It's going to tell you something. They're less likely to be susceptible to just completely bizarre pathologies that we look at and are like, I don't understand why it's seeing that. On the other hand, they're going to have other things that will frustrate us. And my inference algorithm is stuck. I don't understand why my Markov chain isn't converging. If that's the only way you're going to do inference, you'll be very frustrated by the dynamics of inference. And that's where, hopefully, some kind of pattern recognition system will come to the rescue.

And if we just look at anecdotal experience, both in speech and vision, there are certain kinds of cases where, like, you know, in a passing rock, you suddenly see a face, or a noise. Or in a tree people will see Jesus's face on arbitrary parts of the visual world. And also sometimes in sound. You hear something. So this idea of seeing signal in noise, we know humans do that.

But, for example, there are ways to get deep confidence in vision to see-- you can start off with an arbitrary texture. Have you seen this stuff? And massage it to look like-- like you can start off with a texture of green bars and make it look like a dog to the network, and it doesn't look like a dog to us. And we're never going to see a dog in a periodic pattern of green and polka dotted bars.

JOSH MCDERMOTT: But the reason you can do that is because perfect access to the network. Right? And if you had perfect access to visual stimuli [INAUDIBLE].

JOSH TENENBAUM: Sure. Sure. But I'm just saying-- these don't-- well, I don't think so.

DAN YAMINS: Of course there's going to be visual illusions in every case. The question is whether or not they're going to make sense to humans--

JOSH TENENBAUM: Right. And I think some of them--

ALEX KELL: --as a test of whether or not that model--

JOSH TENENBAUM: If you learned something from that.

DAN YAMINS: --is a real model.

JOSH TENENBAUM: Just to be clear, the ones that the convnets are susceptible to that say a generative model or a human isn't-- they're not signs that they're fundamentally broken. Rather they're signs of any discriminatively trained pattern recognizer. I would predict, whether it's good or bad, it's signs of the limitations of pattern recognition.

DAN YAMINS: Or signs of the limitation of the type of tasks that are being used of which the recognition is being done. If you replaced something like categorization with something like the ability to predict geometric and physical interactions x period of time in the future, maybe you'd end up

with quite different illusions. Right? There's something very brittle about categorization that could lead to, sort of null space as being very broad.

JOSH TENENBAUM: Exactly. That's what I mean. By pattern recognition, I mean pattern classification in particular. Not prediction but classification.

DAN YAMINS: Right. But I don't think it's yet known whether the existence of these, sort of fooling images or this kind of weird allusions, e.g. the models are bad or do not pick out the correct resolutions. I don't know whether-- people are not totally sure whether that's like, the networks need to have feedback, and that will be what you really need to solve it. It's at that broad level of mechanism.

Or is it like, the task is wrong? So it's sort of a little bit less bad. Or maybe like Josh said, it's like the easiest thing would be, well, actually if you just did this with the neural system, you'd find exactly the same thing. But we don't have access to it, so we're not finding it. Right? And so I think it's not totally clear where it is yet. Right? It's a great question, but I feel like the answers are murky right now.

ALEX KELL: Yeah. OK. On the topic of feedback, I wanted to kind of move over-- and Gabriel talked about feedback during his talk. And there's really heavy kind of feedback in both of these modalities, where, like, in hearing as Josh talked about, it goes all the way back to, it can alter the mechanics of the cochlea, of the basilar membrane. That's pretty shocking. That's pretty interesting.

So what is the role-- Gabriel talked about a couple of specific examples where feedback would actually be useful. Can you say something more broadly about, in general when is feedback useful across the two modalities? Do we think they are kind of specific instances-- can we talk about specific instances in each?

GABRIEL KREIMAN: Throughout the visual system there is feedback essentially all over except for the retina. Throughout the auditory cortex, again this feedback and recurrent connections all over.

We've been interested in a couple of specific apps in situations where feedback may be playing a role. This includes visual search. This includes pattern completion, feature-based attention. I believe, and hopefully Josh will expand on this that these are problems that at least at a very superficial level also exist in the auditory domain, and where it's tempting to think that feedback will also play a role.

More generally, you can mathematically demonstrate that any network with feedback can be transformed into a feed-forward network just by decomposing time into more layers. So I think ultimately, feedback in the cortex may have a lot to do with, how many layers can you actually fit into a system the size of the head that it has to go through-- interesting places at some point. And there are sort of physical limitations to that more than fundamental computational ones.

At the heart of this is the question of, how many recurrent computations do you need? How much feedback you actually need, how many recurrent loops you need. If that involves only two or three loops, I think it's easy to convert that into a feed forward network that will do the same job. If that involves hundreds of iterations and loops, it's harder to think about a biological system that will accomplish that.

But at least at the very superficial level, I would imagine that--

JOSH TENENBAUM: Can I ask a very focused version of the same question, or try to, which is, what is the computational role of feedback in vision and audition? Like when we talk about feedback, maybe we mean something like top-down connections in the brain or something like recurrent processing. Just from a computational point of view of the problem we're trying to solve, what do we think its roles are in each of those?

GABRIEL KREIMAN: So more and more, I think that's the wrong kind of question to ask. If I ask you--

JOSH TENENBAUM: Why?

GABRIEL KREIMAN: What's the role of feed-forward connections?

JOSH TENENBAUM: Pattern recognition.

GABRIEL KREIMAN: There is no role of feed-forward connections.

JOSH TENENBAUM: No. On the contrary. Tommy has a theory of it. You know, you have another theory. Something like very quickly trying to find invariant features-- trying to very quickly find invariant

features of certain classes of patterns. That's a hypothesis. It's pretty well supported.

GABRIEL KREIMAN: There's a lot of things that happen with feed-forward.

HYNEK HERMANSKY: If you want a feed so that you can make things better somehow, I mean, you need a measure of goodness first. I mean, otherwise I mean, how to build-- I agree with you that you can make a very deep structure which will function as a feedback thing.

But always what worries me the most, and in general a number of cognitive problems, is, how do I provide my machine with some mechanism which tells the machine that output is good or not? If the output is bad, if my image is making no sense, there is no dog but it's a kind of weird mix of green things and it's telling me it's a dog, I need feedback.

That's the point I need the feedback, I believe. And I need to fix things. Josh is talking about tuning the cochlea. Yeah, of course that means that sharpening the tuning is possible. But in communication, I mean, if things are noisy I go ahead and close the door. There's the feedback to me. But I know, as a human being, I know information is not getting through, and I do something about it. And this is what we--

JOSH TENENBAUM: That's great You just gave two good examples of, I think, just to generalize those, right, or just to say them in more general terms. One role of feedback that people have hypothesized is like in the context of something like analysis by synthesis. If you have a high level model of what's going on, and you want to see, does that really make sense? Does that really explain my low level data? Let me try it out and see that.

Another is, basically saying, the role of the feed-forward connections is a kind of invariant pattern recognizer, and tuning those, tuning the filters, tuning the patterns to context or in particular in a contextual way to make the features more diagnostic in this particular context. Those are two ideas. And they are probably others.

HYNEK HERMANSKY: Yeah, you gave the wonderful example with analysis by synthesis. But even there, we need an error measure. And I'm not saying that least mean squared error between what I generate and what I see is the right one.

JOSH TENENBAUM: So you think feedback if it helps tune the error measure.

HYNEK HERMANSKY: Well, no. It's a chicken and egg problem. I think I need the error measure first, before I even start using the feedback. Because, you know, feedback, we can talk about it. But if you want to implement it, you have to figure out, what is the error measure or what is the criteria? And that I recognize my output is bad or not good enough. Obviously it will a little bit bad. Right? I am not good enough and I have to do something about it. Once I know it, I know what to do, I think. Well maybe not me, but my students, whatever.

This is one of the big problems in which we are working on actually-- figuring out, how can I tell that output from my neural net or something is good or bad? And so far I don't have any answer.

JOSH TENENBAUM: But I think it's neat that those two or three different kinds of things-- they are totally parallel in vision and audition. They're useful for both, engineering-wise, and people have long proposed them both in psychology and neuroscience.

GABRIEL KREIMAN: Generally I think there's a lot to be done in this area, I think. And the notion of adding-- I mean, a lot of what's been happening in commercial networks is sort of tweaks and hacks here and there. If there are fundamental principles that come from studying recurrent connections and feedback from the auditory domain, from the visual domain, those are the sort of things that, as Dan was saying, could potentially sort of lead to major jumps in performance or in our conceptual understanding of what these networks are doing. I think it's a very rich area of exploration, both in vision and the auditory world.

ALEX KELL: And wanted to ask Josh one more thing-- the common constraints thing. You were kind of saying like, it seems like there would be common constraints and there are probably consequences of those. What are some kind of specific consequences that you think would come out? Like, how can we think about this? To the extent that there is kind of a shared system, what does it mean?

JOSH TENENBAUM: Well-- so again, I can only, as several of the other speakers said, only give my very, very personal, subjectively biased view. But I think the brain has a way to think about the physical world to represent, to perceive the physical world. And it's really like a physics engine in your head. And that the different-- it's like analysis by synthesis that's a familiar idea probably best developed classically in speech, but in a way that's almost independent of sense modality. I think we have different sensory modalities, and they're all just different projections of an underlying physical representation of the world.

I think that, whether it's understanding simple kinds of events as I was trying to illustrate here, or many other kinds of things, I think, basically, at some one, one of the key outputs of all of these different sensory processing pipelines has to be a shared system that represents the world in three dimensions with physical objects that have physical reality-- properties like mass or surface properties that produce friction when it comes to motion-- roughness. Right.

There's some way to think about the forces, whether the force that one object exerts on another, or the force that an object presents in resisting, when I reach for it, either the rigidity that resists my grasp and that [INAUDIBLE] with it. Or the weight of the object that that requires me to exert some weight to do that.

So I think there has to be a shared representation that bridges perception to action. And that it's a physical representation, and that it has to bridge-- it's the same representations that's going to bridge the different sense modalities.

DAN YAMINS: Yeah, but to make [INAUDIBLE].

JOSH TENENBAUM: More what?

DAN YAMINS: More brain meat-oriented, I think that there's a version of that that could be a constraint that is so strong that you have to have a special brain area that's used as the clearinghouse for doing that common representation.

JOSH TENENBAUM: I'm certainly not saying that.

DAN YAMINS: OK. Right. No, I didn't think you were. But that would be a concrete result. Another concrete result is like effectively that the individual modality structures are constrained in such a way that they have an API that has access to information from the other modality, so that message passing is efficient, among other things. Right? And I think that they can talk to each other.

JOSH TENENBAUM: I think it's often not direct. I think some of it might be, but a lot of it is going to be-- it's not like vision talking to audition, but each of them talking to physics.

DAN YAMINS: Right. Right.

JOSH It's very hard for a lot of--

TENENBAUM:

DAN YAMINS: Right. And that's a third--

JOSH TENENBAUM: Mid-level vision and mid-level audition are hard to talk to each other.

DAN YAMINS: Right. And a third possibility is is that it's actually not really so much of-- it that there's no particular brain area, and they're not exactly talking to each other API directly. It's just that there is a common constraint in the world that forces them to have similar structure or sort of aligned structure for representing the things that are caused by the same underlying phenomenon. And that's like the weakest of the types of constraints that you might have. Right? The first one is very strong.

JOSH TENENBAUM: But it's not it's not vague or content-less. So, Nancy Kanwisher and Jason Fisher have a particular hypothesis. They've been doing some preliminary studies on the kind of intuitive physics engine in the brain. And they could point to a network of brain areas, some premotor, some parietal. You know, it's possible. Who knows. It's very early days. But this might be a candidate way into a view of brain systems that might be this physics engine.

DAN YAMINS: Right. But you are--

JOSH TENENBAUM: Also ventral stream area.

DAN YAMINS: [INAUDIBLE] be something like, if you optimized network's set of parameters to do the joint physics prediction interaction task, you'll get a different result than if you sort of just did each modality separately. And that would be a better-- that new different thing would be a better match to the actual neural response patterns in interesting--

JOSH TENENBAUM: Yeah. I think would make really cool thing to explore.

DAN YAMINS: And that's, I think the concrete way with cache out. And that certainly seems possible.

JOSH TENENBAUM: And it might be, you know, a lot of things which are traditionally called association cortex, right. This is an old idea and I'm not enough of a neuroscientist to know, but a cartoon history is, there are lots of parts of the brain that nobody could figure what they were doing because they

didn't respond in an obvious selective way to one particular sense modality. It wasn't exactly obvious what the deficit was.

And so they can be called the association cortex. That connects to the study of cross-modal association and association to semantics, and this idea of association. It's this thing you say when you don't really know what's going on. But it's quite possible that big chunks of the brain we're calling association cortex are actually doing something like this. They're this convergence zone for a shared physical representation across different perceptional modalities and bridging to action plan.

And a big open challenge is that we can have what feels to us like a deep debate that we can think of as like the central problem in neuroscience of, how do we combine whatever you want to call it-- I don't know, generative and discriminative or model-based analysis by synthesis and pattern recognition synthesis. But actually there's a lot of parts of the brain that have to do with reward and goals.

And again, I thought Laura's talk was a really good illustration of this, understanding perception is representing what's out there in the world, but that's clearly got to be influenced by your goal. And what is certainly a big problem is the relation between those. It says something that most of us who are studying perception don't think we need to worry about that. But I think we should, particularly those of us, again, echoing what Laura said-- if we're studying learning, we definitely, I think, need to think about that more than we do.

HYNEK HERMANSKY: It may not be exactly related to what you are asking, but I don't know. I believe that we are carrying the model of the world in our brain, and we are constantly evaluating the fit of what we expect to what we are seeing. And as long as we are seeing or hearing what we expect, we don't work very hard, basically, because the model is there. You know what I'm going to say. I'm not saying anything special. I look reasonable and so on and so on.

And when the model of the world is for some reason violated, that may be one way how to induce the feedback, because then suddenly I know I should do something about my perception. Or I may just give up and say-- or I become very, very interested. But I think this is a model of the world priors which we all carry with us. It's extremely important and it helps us to move through the world. That's my feeling.

In speech we have a somehow interesting situation that we can actually predict-- say we are interested in estimating probabilities of the speech sounds. But we can also predict them from

the language model. Our language model is learned typically very differently from a lot of texts, and it's a lot of things. And so we had quite a bit of success in trying to determine if the world recognizes if it's working well or not, by comparing what they recognize and expect and what it sees. And as long as these things go together well, it's fine. If there is a problem between these two, we have to start working.

JOSH TENENBAUM: I think, you know, physics is a source of beautiful math and ideas. I think it's an interesting thing to think about, maybe some tuning of some low level mechanisms and in both sensory modalities might be well thought of that way. Right. But I think it is dangerous to apply too much of the physicist's approach to the system as a whole.

This idea that we're going to explain deep stuff about how the brain works as some kind of emergent phenomenon, something that just happened to work that way because of physics. This is an engineered system. Right. Evolution engineered brains. It's a very complicated-- I mean, we have had a version of this discussion before. But I think it's something that a lot of us here are committed to. Maybe not all of us. But the way I see it is, this is a reverse engineering science. And the brain isn't an accident. It didn't just happen. There were lots of forces over many different timescales acting to shape it to have the function that it does.

So if it is the case that there are some basic mechanisms, say maybe at the synaptic level that could be described that way, it's not an accident. They were [INAUDIBLE]. I would call that, you know, biology using the physics to solve a problem. And again there's a long history of connecting free energy type approaches to various elegant statistical inference frameworks.

And it could be very sensible to say, yes, at some levels you could describe that low level sensory adaptation as doing that kind of just physical resonance process. Or nonequilibrium stat mech could describe that. But the reason why nature has basically put that physics interface in there is because it's actually a way to solve a certain kind of adaptive statistical inference problem.

ALEX KELL: All right. Cool. Let's thank our panel.

[APPLAUSE]