

The following content is provided under a Creative Commons license. Your support will help MIT OpenCourseWare continue to offer high quality educational resources for free. To make a donation or view additional materials from hundreds of MIT courses, visit MIT OpenCourseWare at ocw.mit.edu.

REBECCA SAXE: There's a whole bunch of limitations of Haxby-style correlations-- one of them is that all the tests are binary. The answer you get for anything you test is that there is or is not information about that distinction, so there's no continuous measure here. It's just that two things are more-- they are different from one another or they are not different from one another. And so once people started thinking about this method it became clear that this is actually just a special case of a much more general way of thinking about fMRI data. So this particular method-- using spatial correlations-- is very stable and robust, but it's a special case of a much more general set.

And here's the more general idea. The more general idea is that we can think of the response pattern to a stimulus in a set of voxels, for example-- the voxels in a region-- we can think of that response pattern as a vector in voxel space. So every time you present a stimulus you get the response of all the voxels. Now, instead of thinking of that as a spatial pattern, think of that as a vector in voxel space. Every voxel defines a dimension, and the position in voxel space is how much activity in each of those voxels there was. Can everybody do that mental transformation? This is, like, the key insight that people had about MVPA-- is we had been thinking about everything in space-- in the space of cortex-- but instead of thinking of a spatial pattern on cortex, treat each voxel as a dimension of a very multi-dimensional space.

Now, the response to every stimulus is one point in voxel space. OK? As soon as you think of it that way, then your mental representation of fMRI data looks like that. Right? So your mental representation of fMRI data used to be a BOLD response, and then it was a spatial pattern of a cortex, and now it's a point in voxel space. And if you can follow those three transformations then you realize that a set of points in a multi-dimensional space is the kind of problem that all of machine learning for the last 20 years has been working on. Right? And so everything that has ever happened in machine learning could now be used in fMRI, because-- well, almost-- because machine learning has just absolutely proliferated in both techniques and problems and solutions to those problems for handling data sets where you have no idea where the

dataset came from, but it's now represented as multiple points in a multidimensional space.

And so that's what happened about five years ago-- is that people realized that we could think of fMRI as the response to every stimulus as a point in voxel space. A set of data is a set of points in voxel space. Now, do anything you want with that. And the first most obvious thing to do is to think of this as a classification problem. OK? So we created conditions in our stimuli or dimensions in our stimuli, so now we can ask, can we decode those conditions? Can we find clusters? Can we find dimensions? Right? All the standard things that people have done when you had points in multi-dimensional spaces.

And so, again, the most common thing people now do is now that you think of fMRI data that way, try linear classification of the categories or dimensions that you're interested in, and typically using standard machine learning techniques. So think of training a classifier on some of your data and testing it on independent data and trying to find the right classification techniques that can identify whatever distinction you're interested in in the data set that you built. And so, the way that this one looks is that you take some-- now, voxels are on the y-axis of this heat map, so we have whatever that is 80, 100 voxels in a region-- maybe more, and for every stimulus you have the response in every voxel to that stimulus. Right? So each of those columns now is a representation of where that stimulus landed in voxel space, and you have a whole bunch of instances.

And so now what you're going to do is use the training to learn a potential linear classifier that tells you what was the best way to separate the stimuli that came from one labeled set versus the stimuli that came from some other labeled set. And the test of that is going to be-- take a new stimulus or new response and use the classification you learned to try to decode which stimulus that that came from, and measure your accuracy. And so the new measure of the information in fMRI is going to be classification accuracy. Does that makes sense-- the people with me? OK, because that's where a lot of fMRI is right now-- is now thinking about responses to stimuli as points in voxel space and the problem as one of classification accuracy in independent data.

OK. Here's one experiment that we did where we use classification. So now, another thing just to note is that in this context you're often trying to classify a single trial. Right? So in our case, we're always trying to classify a single trial, so we've gone from partitioning the data to two halves and asking about similarity, to training on some of the data, and now classifying single independent trials. OK. So here's a case where we tried to do that, and it was an extension of

the stuff that I just showed you that you could classify seeing versus hearing, and so we tried to replicate and extend that. So we told people's stories like this-- there's a background-- so Bella's pouring sleeping potion into Ardwin's soup, where her sister, Jen, is waiting. They're holding their breath while he starts to eat. The conclusion of the story is always going to be the same. Bella concludes that the potion worked, and then we tell you based on what evidence she made that conclusion. Another case here is going to be-- Bella stared through the secret peephole and waited. In the bright light she saw his eyes close and his head droop, so that her evidence for the conclusion that the potion has worked.

That's OK evidence, and we can vary that in a bunch of ways. So one is we can change the modality for evidence. Instead of seeing something, she can hear something. So for example, she pressed her ear against the door and waited. In the quiet she heard the spoon drop and a soft snore. So that's similar content of information, but arrived at through a different modality. Or we can change how good her evidence is, and so in this case we did it by saying, she tried to peer through a crack in the door. In the dim light she squinted to see his eyes closed. OK, so that's less strong perceptual evidence for the conclusion that the potion has worked. OK.

And so now what we're going to ask is-- if we train on one set of stories, on the pattern of activity in a brain region for stories that vary on either of these dimensions, one at a time either vary on modality or vary on quality-- in a new test set, can we decode that dimension? Yeah, and the first answer is we can-- both of them.

One thing about this is that this measure isn't binary anymore. So since we're doing for every stimulus we're asking whether we can classify that stimulus or not-- we can get for every subject, so we can get for each item the probability-- for each subject for each item we get a measure of whether it was classified correctly or not, so across objects, across items-- we know for every item the probability of it being correctly classified or not. And then we can ask, is that related to other continuous features of that item? So in this case what we can say, for example, is the quality dimension-- how good your evidence is for the belief that you conclude-- that's a continuous metric-- though it's a continuous feature. It can be judged continuously by human observers, so for each item we can ask, how good is the evidence for the conclusion for this specific story?

That judgment by human observers of how good the evidence is, continuously predicts the probability of that item being classified as being good evidence or bad evidence-- even over above the label that we gave it. So if you regress out the labels there's a continuous predictor.

So something, like, imagine a neural population that responds-- a sub population that responds more the better the evidence, continuously, so that classification gets better as you get further out on that dimension. It's also not redundant across brain regions, so there's different information in different brain regions. And this is just to show you that in two other brain regions-- so in the right STS we can decode quality, but not modality, and in the left TPJ we can decode modality, but not quality. And the left TPJ we've replicated a bunch of times. In the DMPFC we can't decode modality or quality, but we can decode valence, which is the thing I told you the right TPJ doesn't decode. And then if we go back and look at valence in this dataset-- we can only decode valence in the DMPFC.

So this is, to me, this is starting to get cool, right? Three features of other people's mental states represented differentially in different brain regions. This distinction between the more epistemic stuff-- like modality and quality which is represented in the TPJ, and valence which is represented in the DMPFC-- I think is real and deep and hints at one of the really most important distinctions within our theory of mind that I mentioned at the very beginning-- between epistemic states and affective or motivational states.

So what's cool about classification analyzes? They have all the same properties as the Haxby-style analyses in principle, because they're actually just a generalization of the Haxby analyses, except that they're a lot less robust, because what you're trying to classify as single trials are single items. And so noisy data collapses faster in these classification strategies than in Haxby-style analyses where you're averaging. But otherwise, those are the same two techniques. What's nice about the classification analyses is you can get item specific outcomes, right? So you can say, for a specific item, how likely it is to be classified as one thing or another? And this is where I started the talk before, which is that in both of these cases we think of a hypothesis and test it sequentially. And so the representational similarity matrix tests whole hypothesis spaces instead of single features.

Classification and Haxby-style stuff are ways to think of a future or dimension that might be represented in a brain region you care about, and test whether or not it's represented. So they're a way of thinking of a hypothesis and testing it, and thinking of hypothesis and testing it-- and that's what I mean by sequentially. So you can think of, does the right TPJ represent the difference, for example, between Grace poisoning the person knowingly and poisoning the person unknowingly? The answer to that is yes, it does, but that's one hypothesis. And then we can come up with another hypothesis, and then another hypothesis. And what's interesting

about representational dissimilarity matrices-- one of the versions of MVPA people use these days-- is that it takes a different approach. So instead of trying to think of one hypothesis and test it, it proposes a hypothesis space and tests the space as a whole, and that gives you both different sensitivity and strengths and different weaknesses.

So I'll work through an example in which we did this. I told you that I would come back to thinking about other people's feelings, and in this experiment we took different kinds of things that people could feel as one subspace of theory of mind. So our stimuli, in this case, are 200 stories about people having an emotional experience. And we're going to look at-- what can we understand about how your brain represents those different-- your knowledge that lets you sort out people's experiences in those cases. OK, so it's hard in the abstract, let's do it in the concrete. So in the behavioral version of this test I give you a list of 20 different emotions-- jealous, disappointed, devastated, embarrassed, disgusted, guilty, impressed, proud, excited, hopeful, joyful, et cetera-- so you have 20 different choices. And I'm going to tell you a single story about a character you don't know, and something they experienced-- very briefly-- and what I want you to think to yourself is, which emotion did they experience in that case?

OK? So here's one. After an 18 hour flight, Alice arrived at her vacation destination to learn that her baggage, including camping gear for her trip, hadn't made the flight. After waiting at the airport for two nights, she was informed that the airline had lost her luggage and wouldn't provide any compensation. How many people think she felt joyful? How many people think that she felt annoyed? How about furious? OK, so furious is the modal answer and annoyed is the most likely second choice answer to that case.

Here's a different one. Sarah swore to her roommate that she would keep her new diet. Later, she was in the kitchen, she took a bite of a cake she had bought for the dinner party. When her roommates arrived home to find that she'd eaten half the cake and broken her diet. How many people think that she would feel disgusted? Terrified? Embarrassed? OK. And just to give you a sense of how fine grained your knowledge is in this case, think about this difference. In this case she swore she would keep her diet and then broke it, right? What about the difference between-- she first ate a cake and then swore she would keep her diet. Right? That's a totally different texture to the story. OK. So we have incredibly fine grained knowledge of how a description of a situation predicts an overall emotion.

You can see that in a behavioral experiment, so what I'm showing you here is-- on the y-axis the emotion that we intended when we wrote the story-- so ten stories for each category for

200 stories-- on the x-axis is the percent of participants picking that label. And so the first thing is that 65% of the time, people pick the label we attend. If instead, you ask, take half the subjects to determine a modal answer and the other half of the subjects as the test set, you get the same answer. There's about 65% agreement on the right, single label out of 20. That's, of course, way above chance, which is 5%, so people are quite good at this. And the off-diagonal is also meaningful, so that also contains information-- the second best answer, right? So annoyed as opposed to furious, for example.

OK, so that's a huge amount of rich knowledge about other people's experiences from these very brief descriptions of events to a very fine grained classification of which emotion they're experiencing. And one way to look at these data is to ask, OK, well, that's knowledge that we have-- where is that knowledge in the brain? That's sort of a first question you could ask, and you could ask it by just saying, if we try to use-- so in this case, we're going to do train and test. So we train a classifier on a patch of cortex, based on five examples from each condition, and then we test on the remaining half of the data. And we just ask, based on the pattern of activity in a patch can you get above chance classification in the independent data? For every patch where that's true we put a, sort of, bright mark, and then ask, where in the brain is the relevant decoding that would let you be above chance on this distinction? The answer is in exactly the same brain regions that I have been talking about and showed you before, that is where there's above chance classification, and then overlaid on the standard belief versus photo task in green.

So within the brain regions involved in theory of mind or social cognition are the brain regions that can above chance classify in this 20 way distinction. And then this is just looking inside each one of those. Inside each of the regions-- that's four of the regions in that group that I showed you before-- you can do above-- with using just the pattern of activity in that brain region you can do above chance classification on this 20 way distinction. And there's a hint that that information is somewhat non-redundant, because if you combine information across all of them you do slightly better than if you use any one of them alone.

OK, so now the question is, how can we study what knowledge is represented in each of these brain regions? Right? So we know that there's some information about that 20 way classification, but can we learn anything about the representation of emotions in those brain regions using fMRI? And that's where the representation dissimilarity matrices come in as a strategy. OK, so the question is, how might you represent the knowledge that you have of

what Alice is experiencing, for example, in this story? What's a possible hypothesis? And the way representational dissimilarity matrices work as a strategy for fMRI analyses is that what you should do is think of multiple different hypotheses about how that knowledge could be represented.

So a first hypothesis, which is deep in the literature on emotions, is that we represent other people's emotional experience in terms of two fundamental dimensions of emotional experience-- valence and arousal. Have you guys heard of valence and arousal as the two fundamental-- OK. So this hypothesis says, when we think about emotions-- our own or other people-- we put emotions in a two dimensional space, which is, how good or bad did it make you feel, and how intense was it? OK. So terrified is negative and very intense. Lonely is negative, but not that intense. Right? That's the idea. Happy is positive and somewhat intense. Thrilled is happy and more intense. So that idea is that there's these two basic dimensions of emotional experience, and so one thing we can do is have each of our stories, like this one-- we can have people tell us in that story was she feeling positive or negative? How positive or negative, and how intensely? And so, for each individual story we can have a representation of it as a point in that space. And if you use just that, you can classify our 200 stories reasonably well-- not as well as people can, but still reasonably well. OK, so the 200 stories do clump into lumps in that two dimensional space.

But another idea is that valence and arousal seem not to capture the full texture of the 20 categories that we originally have. It's not that we can't embed 20 categories in two dimensions-- you obviously can have 20 clusters in a two dimensional space. But we had the intuition that it's not a two dimensional space-- that those two dimensions don't capture all the features that people have and know about when they use the stimuli. And so, based on another literature called appraisal theory-- what we tried to do is capture some of the abstract knowledge that people have about these situations that lets them identify which emotion it is. And we did that by having them rate each of these stories on a bunch of abstract event features. So those event features are things like-- was this situation caused by a person or some other external force. So I hope you guys have the sense that if your luggage gets lost on your way to the trip-- it's different if that was airline incompetence versus a tornado, right? Does everybody have that intuition? The emotion is different. OK.

So that's an important abstract feature of our knowledge of other people. Was it caused by you yourself? If you left your luggage at home, that's different from if airline incompetence

caused you not to have your luggage. And does it refer to something in her past? Is she interacting with other people? That makes a really big difference, for example, in pride and embarrassment-- whether other people are around. How will it affect her future relationships? So things that potentially cause harm to future relationships feel very different from things that are just annoying right now but will end. So these are abstract features and they encapsulate things we know about emotion relevant features of the situations people find themselves in. So we came up with 42 of these and we had every story rated on all of those dimensions. And of course, we can, again, classify the stories as 20 clusters in a 42 dimensional space, right? Again, of course we can. But the question is [INAUDIBLE] this is those data. This is just every set of 10 stories and their average rating on our-- oh, 38-- on our 38 appraisal features, so that creates a 38 dimensional space.

Here the idea is-- for each category-- like, for all the stories about being jealous-- you can get-- for, let's say, for the two dimensions of valence and arousal-- the average value of valence and the average value of arousal, right? So that's a point in a two dimensional space-- the stories about being jealous. OK. Then you take the stories about being terrified. What's their valence and arousal? So that's another point in a two dimensional space. And then you take the distance between them, and that number goes in a representational dissimilarity matrix. So the further away you are in a two dimensional space, the more dissimilar. And you could do the same thing in a 42 dimensional space, a 38 dimensional space, any dimensional space you want-- what you need to know is just how far away you are. And so what a representational dissimilarity matrix has in it is for every pair. So the jealous stories versus the grateful stories-- the number in that cell is the distance from the mean position in your space of all the jealous stories to the mean position in your space of all the grateful stories. Does that make sense? And that could be true of any dimensionality.

When you know these 38 features-- so this is behavioral data-- when you know the 38 features of these emotions, the green bar is how well you can classify new items, just behaviorally. So if I give you a new item and all I tell you is its value in these 38 dimensions, how well can you tell me back which emotion category it comes from? The best you could possibly do is 65%, because that's what human observers do in all of our-- so the reality is the human observers-- the features come from human observers, so our ceiling's going to be 65%, and the answer is about 55%. OK. And you can take that in two different ways. One tendency is to say, wow, we know a lot of the key features that go into emotion attribution. I think, Amy, who I did this work with, had a tendency to feel that way. And I think, wow, we

thought of 38 things and we still didn't think of all the important things. Like, what are those other things that we didn't think of that explain the rest of the variation? So you could feel either way about this, but in any case, once you know the position of one of these stories in the 38 dimensional space of these features, you know a lot about which emotion category it came from.

And then this is the correlation to the neural RDM data that I showed you. And so, again, what I showed you is, so observer's knowledge-- that's everything that we know that lets us classify a story. Valence and arousal is the yellow bar-- that's just these two features of the story, and they're both less good than this intermediate thing, which is the 38 dimensional space. And one question is, like, do I really think it's 38 dimensions? No, definitely not. That was just the set of all the things that we could think of. How many dimensions is it, really? Again, I don't know, really, but I can tell you that the best ten dimensions capture most of the information from the 38 dimensions. So what we've discovered so far is ten really important dimensions of your knowledge of emotion. I don't, again, think that means that our knowledge of emotion is ten dimensional. Lots of this is limited by the set of stimuli that we chose, the resolution of the data that we have, and so forth and so on. But in these data you need something on the order of ten dimensions to get close to human performance or close to the genuinely differential signal in the neural data.

If you take one thing away from this talk about the methods used in representational dissimilarity matrix-- really only one thing. Here's the one thing I want you to know-- the dimensionality of the theory that generated your representational dissimilarity matrix does nothing for you in the fit to your data. Nothing at all. It's a parameter-free fit. OK? So anybody to whom those words mean anything, this will be important, so I want you to actually know this. Representational dissimilarity matrices provide a parameter-free fit to the data, and therefore, the dimensionality of the theory that generated the representational dissimilarity matrix has nothing to do with the fit of the data.

You can probably notice I should have ordered this better. Valance has two dimensions, the observers has a lot of dimensions-- I don't know how many, but a lot more than 38. We know that because 38 doesn't explain all their data. So as you go up in-- and in principle, having more dimensions doesn't help in the set. You can correctly see that they overfit rather than fitting the data, and here's why. Because the way you build a representational dissimilarity matrix is, out of no matter how many dimensions you have in your data set for every pair of

stimuli, you take one number, and then a representational dissimilarity matrix encodes the relationships among those numbers. OK? So jealousy is more similar to irritation than it is to pride. By how much? OK? And those relative differences is all you have. You have nothing else, and so there's no parameters. Right? You have the same amount of information in a representational dissimilarity matrix that you generated from a one dimensional theory, a two dimensional theory, a 38 dimension theory, and an infinite dimensional theory. The size of the theory doesn't make any difference, because what you get in the end is exactly the same thing-- the relative distance between every two points in the set.

There's a few things to say about-- so one thing to say about the representational dissimilarity analysis that I just showed you is that it tells you that the 38 dimensional theory is better than the valence theory. Like, the event feature theory is better than the valence theory, but it doesn't tell you why. Right? It doesn't tell you whether any specific one of those features is capturing variance in any specific one of those regions. It tells you that that whole set was better than this other whole set, and maybe this is where you're getting at. It's much less good for trying post-hoc things for saying, but why? Which aspect of that theory was better than the valence and arousal? It gives you an all things considered answer, not a dimension specific answer. That's one thing that is a limit in the way you should use representational dissimilarity analyses.

There's two key problems that I think bear reflecting on about MVPA, and one of them is a catastrophe and the other one is an incredibly deep puzzle. And I think I should just say them right away before you get too excited, because all of this stuff was really exciting and now I'm going to tell you a catastrophe and a puzzle. Here's the catastrophe. The catastrophe is that you can't make anything of null results. OK, now, here's why. Because when I say that you can decode something from an MVPA analysis, what I mean is that at the scale of voxels, there's some signal in terms of which voxels relatively higher or relatively lower in response to the stimuli. Right? So in voxel space or in spatial space, whichever one of those you find helpful-- if that at the level of voxels we could cluster these stimuli. And what that says is that they are something like distinct populations in this region, responding across that feature dimension, and they're spatially segregated enough that we could pick up on them with fMRI.

But who cares if they're spatially segregated enough that we could pick up with them with fMRI, right? fMRI is the scale of a millimeter. And there could be many, many, many things that are represented by populations of neurons within a region that are not spatially organized

at the scale of a millimeter. Not only could there be-- there absolutely, definitely are. There's a whole bunch of things that we already know are really important properties of neural representations of things we care about, and we know that their spatial scale is not high enough that they can be picked up on with fMRI.

So two cases that I'll tell you about because you should care about them-- one is face responses in the middle temporal region that Doris and Winrich study for face representations in monkeys. It's one of the middle ones. In that one there's face features that can tell you how far apart the pupils are, how high the eyebrows are-- did Winrich show you this amazing data? Totally, amazing, beautiful, feature space of face identity representation? One of the most strikingly beautiful things I've ever seen. And he already knows-- he and Doris already know-- that there's no spatial relationship at all between the property that one neuron signals and its distance from other neurons that signal other properties. There's no spatial organization at all. So if you know that right here is a neuron that responds to eye width, you know nothing more about the neuron next to its preferred property than a neuron a centimeter away. There's no spatial structure to which feature I give a neuron response to, which means that you absolutely could not and cannot pick up on that in fMRI, which Doris has shown. This feature structure information cannot be picked up on with fMRI, even though it is there and really important.

Another example is valence and coding in the amygdala. The amygdala contains some neurons that respond to positively valenced events and other neurons that respond to negatively valenced events, and they are as spatially interleaved as physically possible-- that's what Kay Tye's data shows. You couldn't get them more spatially interleaved than they are. They are as close together as the size of the neurons allow. So you absolutely will never be able to decode with fMRI in those populations-- the amygdala-- that there are different populations for positive and negatively valenced events, but there are. OK. So that means that when you see something in fMRI it's probably there, but when you don't see it in fMRI you don't know that it's not there. And the reason why that's a total catastrophe is if it means that when I tell you that a region codes A and not B-- I don't know that it doesn't code B. And when I tell you the region that this thing is coded in region A and it's not coded in region B-- I don't know that it's not coded in region B.

So I can never show you a double dissociation. I can never show you a single dissociation. I can never show you a dissociation at all. All I can say for sure is that the spatial scale of the information is different between one region and another, or between one piece of information

and another, and we have no reason to believe that that matters at all. Right? Really important things are encoded at very fine spatial scales. And so any time I tell you-- which I told you a bunch of times because I think it's really cool-- that there's a difference in what feature is encoded where, you have no reason to believe me. And that's the catastrophe. It's a total catastrophe. If you can't make distinctions, you can't make any conclusions at all.

I'll just briefly say the other thing that's a problem with this, which is that this idea of similarity space-- the idea that you should think of a concept, like jealous, as a point in a multidimensional space, and what it means to think of somebody as jealous is to think of them as a certain distance from irritated and angry and proud and impressed-- that idea has been thoroughly undermined in psychology and psychophysics and computational cognition. It's really a bad theory of concepts. It can't do any of the work that concepts are supposed to do. One of the most important things they can't do is compositionality. It can't explain the way concepts compose, which is absolutely critical to the way that we think and even more critical to the way that we think about other people's minds, because every thought you have about somebody else's mental state is a composition of an agent, a mental state, and a content. And so, this whole way of thinking about concepts as points in multi-dimensional spaces works, but shouldn't work. And that's the other problem with this whole endeavor.

OK. There's a bunch of things that we're doing with this that I will just briefly mention in case people want to think about it or know about it. The two things I'm really excited about-- one is adding temporal information, so looking at the change in information in brain regions over time, and how they influence one another. And that's my post-doc Stefano Anzellotti's project. And another thing that I'm excited about is that-- to the degree that you take these positive claims as something interesting, which I actually still do in spite of all my end of the world talk-- one thing that I think is really neat is the idea of increasingly differentiable representational spaces. So two sets of stimuli that produce clusters that are not separable-- for example, in voxel or neural space-- and making them increasingly distinct.

So Jim DiCarlo calls this unfolding a manifold. Right? That idea, which is Jim DiCarlo's model of the successive processing in stages from V1 to V2 to V4 to IT-- I think that's a really cool model of conceptual development. That what you might have is originally neural responses that can't separate stimuli along some interesting dimension-- that unfold that representational space to make them more dissimilar as you get that concept more-- or that dimension or feature of the stimuli more distinctively represented. And so we've tried a first version of this

with justification-- so kids between age seven and 12 get better and better at distinguishing people's beliefs that have good and bad evidence. And we've shown that that's correlated with a neural signature in the right TPJ getting more and more distinct over that same time and those same kids. And so I think thinking of representational dissimilarity as a model of conceptual change, while certainly wrong, is probably really powerful, and I'm very excited about it.

And the last thing I will do is thank the people who did the work, especially everybody in my lab, and two PhD students-- Jorie Koster-Hale and Amy Skerry and you guys. Thank you.