TOM MITCHELL: I want to talk about some work that we're doing to try to study language in the brain. Actually, to be honest, this is part of a grander plan. So here is what I'm really doing with my research life. I'm interested in language, and so I'm involved in two different research projects.

One of them is to build a computer to learn to read. And we have a project which we call our Never Ending Language Learner, which is an attempt to build a computer program to learn to read the web. NELL, we call it, has been running nonstop, 24 hours a day since 2010. So it's now five years old. If you have very good eyesight, you can tell that everybody's t-shirt there in the group is wearing a NELL fifth birthday party t-shirt.

But it's an effort to try to understand what it would be like to build a computer program that runs forever and gets better every day. In this case, its job is to learn to read the web. It is getting better. It currently has about 100 million beliefs that it has read from the web. It's learning to infer new beliefs from old beliefs. It's a better reader today than it was last year. It was better last year than it was the year before. It's still not anything like as competent as you and I, but it's one line of research that you can follow if you're interested in understanding language understanding.

The other thread, which is what I'm going to talk about tonight, which is in the bottom half here, is to study how the brain processes language by putting people in brain imaging scanners of different types, and showing them language stimuli, and getting them to read. So I'm going to focus really on the bottom part. But I can't really talk about this honestly unless I fess up to the fact that my goal is for these two projects to collide in a monstrous collision. They haven't yet, although you'll see some signs, I hope, tonight, of some of the cross-fertilisation between the two areas.

When it comes to the brain imaging work, we have a very great team of people. One of them, Nicole Rafidi, is sitting right here. Some of you have already met her this week. And so what I'm going to present is really the group work of quite a few people. And the idea is simple, but here's the brainteaser. Suppose you're interested in how the brain processes language, and

you have access to some scanning machines, then what would you do?

And so we started out by showing people in a scanner stimuli like these. Maybe single words, initially nouns like camera, and drill, and house, and saw. Sometimes pictures, sometimes pictures with words under them. But just showing people stimuli to get them to think about some concept. And then we collect a brain image, like this one, which we collected when a person was looking at this particular stimulus, a bottle.

And this is posterior, this is the back of the head on top. This is the front of the head at the bottom here. And these four slices are four out of about 22 slices of the brain that make up the three dimensional image. And so you can see here what the brain activity looks like-- kind of blotchy-- when one particular person thinks about, bottle.

So you might ask, what does it look like if they think about something else? Well, I can show you what it looks like on the average. If we average over 60 different words, then here's the brain activity. And you can see that it looks a lot like bottle, but maybe there are some differences. And in fact if I subtract out this mean activity from the brain image we get for bottle, then you can see the residue here. There are in fact some differences in the activity we see for bottle compared to the mean activity over many words.

Whether that's signal or noise, I guess you can't tell by looking at this picture. But that's the kind of data that we have if we use fMRI to capture brain activity while people read words. So the first thing you might think of doing if you had this kind of data would be to train a machine learning program to decode from these brain images which word somebody is thinking about it.

And so we, in fact, began that way by training classifiers where we'd give them a brain image. And during training time we would tell them which word that brain image corresponds to. And then after training we could test the classifier to see whether indeed it had learned the right pattern of activity by showing it new brain images and having it tell us, for example, is this person reading the word hammer or bottle.

And, in fact, that works that works quite well. And, in fact, if you try it over several different participants in our study, you can see we get classification accuracies for a Boolean classification problem. Are they reading a tool word like hammer, saw, chisel, or a building were like house, palace, hotel.

Then, depending on the individual person, we can get in the high 90% accuracy or a little worse. In fact, if you ask why it's not the same for all people, it turns out the accuracy that we get correlates very well with measure of head motion in the machine. So a lot of this is noise.

But the bottom line here is good. fMRI actually has enough resolution to resolve the differences in neural activity between, say, thinking about house versus hammer. And machine learning methods can discover those distinctions. So that's a good basis.

And so given that, you can start asking a number of interesting questions. Like we could ask, well, what about you and me? Do we have the same pattern of brain activity to encode hammer, and house, and all the other concepts? Or do each of us do something different?

And we can convert that into a machine learning question, right? We could say, well, what if we train on people on that side of the room. We'll collect their brain data and train our program. Then we'll collect data from these people and try to decode which word they're reading based on the patterns that we learned from those people.

If that works, then that's overwhelming evidence that we have very similar neural encodings of different word meanings. So we tried that and, in fact, it works. In fact, here you see in black, the accuracies, just like on the first slide, of how well we can decode which word a person is reading in black, If we train on data from the same person we're testing on. But in white you see the accuracies we get if we train on no data at all from this person, but instead train on the data from all the other participants.

And you see on average we do about as well with the white bars as we do with the black bars. In fact, in some cases we do better training on other people. That might be, for example, because we get to use more training examples. We get to use all the other participants' data instead of just one participant's data.

But again, the important thing here is, this is very strong evidence that, even though we're all very different people, we have remarkably similar neural encodings when we think about common nouns. Which is something that really, say in the year 2000, I don't think anybody understood.

So I want to kind of wrap up this idea. So I want to go through basically four ideas in this talk. Idea number one is, gee, we could train classifiers to try to decode from the neural activity which word a person is reading. And if we do that, then we can actually ask some interesting

scientific questions, like are the patterns similar across our brains? Does it depend whether it's a picture or a word?

And, in fact, we can think of this technique of training and classifier as-- the way I think of it is it's a way of building a virtual sensor of information content in the neural signal. So I think that fMRI was truly a revolution in the study of the brain, because for the first time we could look inside and see the activity.

But I think these classifiers give us a different thing. Now we can look inside and see not just the neural activity, but the information encoded in that neural activity. And so it's a different kind of sensor. And you can design your own and train it, and then use it to study information represented in the neural signal in the brain. So it kind of opens up a very large set of methods, and techniques, and experiments that we can now run with brain imaging. Where instead of looking just at the activity, we now can look at the information content.

OK, so that's idea number one. We were quite pleased with ourselves and we are doing this work. But in the back of back of our mind was kind of a gnawing question of, well, this is good, now maybe we've trained on a couple of hundred words, so we have a couple hundred different neural patterns of activity. We have kind of a list of the neural codes for a couple of hundred words, but that's not really a theory of neural encodings of meaning. It's a list. What would it mean to have a theory?

Well, scientific theories are logical systems that can make predictions. And if they're interesting theories, they make experimentally testable predictions. So in our case, it would be nice, if we want to study representations of meaning, to have a theory where we could input an arbitrary noun and get it to predict for us what would be the neural representation for that non. At least that would be better than a list. That would be a generative theory or model.

And so we're interested in this. And we worked on this for a while and came up with-- our first version of this looked like this. It's a computational model that was trained. And once it's trained, it would make a prediction for any input word, like telephone, in two steps.

Step one, if you gave it a word like telephone, for example. Step one, it would look up the word telephone in a trillion words of text collected from the web and represent that word by a set of statistics about how telephone is used. In our case, statistics about which verbs co-occurred with that noun. And then in the second step, it would use that vector which approximates the meaning of the input noun as the basis for predicting in each of 20,000 locations in the brain,

how much activity will there be there. So let me push on that a little bit.

So I say in step one, we look up for a word like celery which verbs that occur with. Well, here are the statistics that we get. This is normalized to be a vector of length 1. But you can see for celery the most common verb is eat. And taste is second most common. But celery doesn't occur very often with ride.

On the other hand, airplane occurs a lot with ride, and not very much with manipulate and rub. So these are the verb statistics extracted from the web for two typical nouns. And step one of the model was just to collect statistics for whatever now we give it to make the prediction.

Step two is then to predict at each location in the brain what the neural activity will be there, the fMRI activity, as a function of those statistics we just collected. So for the word celery, now we know it occurs 0.84 with eat and 0.35 with the verb taste. We're now going to make a prediction of this voxel. In particular, the prediction that voxel v is the sum, over those 25 verbs that we're using, of how frequently verb i occurs with the input noun, celery in this case, times some coefficient that we have to learn from training. And this coefficient tells us how voxel v is influenced by co-occurring with verb i. And we have 25 verbs, 20,000 voxels, so we have 500,000 of these coefficients to learn.

We learn them by taking nouns, collecting the brain-- the same data we use to train those classifiers. So we have a collection of nouns and the corresponding brain images. For each of those nouns we can look up the verbs statistics. And then we can train on that data to estimate all these half million coefficients.

When you put those coefficients together, say, for eat, this is actually a plot of the coefficient values. Here's one of those coefficients for the verb eat in a particular voxel right there. So you can think of the coefficients associated with each verb as forming a kind of activity map for that verb. And a weighted linear sum of those verb-associated activity maps gives us a prediction for celery.

You could ask, how well do these predictions work? One way I could answer that is to show you here, when we trained on 58 other nouns, not including celery, not including airplane. And then we had the system predict these novel, to it, words. Celery, it predicted this image. Airplane, it predicted this image. Unbeknownst to it, here are the actual observed images for celery and airplane.

So you can see it correctly predicts some of this structure-- this is, by the way, fusiform gyrus-- but not all the structure. So it captures some of what's going on. I can, in a more quantitative way, tell you how well it's working by-- we can test the program this way. We can say, here are two words you have not seen. Here are two images you have not seen. One of them is celery, one is airplane. You, the program, tell me which.

If it was just working at chance, it would get an accuracy of 50%. If you just guess randomly, you'll get half of those right by chance. In its case, averaged over nine different subjects in the experiment, we get 79% accuracy. So what does this mean? What this means is, three times out of four, 79%, we could give this trained model two new nouns that it has never seen, two fMRI images for those nouns, and it could tell us three times out of four which was which.

So this model is extrapolating beyond the words on which it was trained. And it's extrapolating, not perfectly, but somewhat successfully to other nouns. Now, why? What's the basis on which it's doing that extrapolation? What are the assumptions built into this model?

Well, for one thing, it's assuming that you can predict the neural representation of any word based on corpus statistics summarizing how that word is used on the web. Furthermore, it's assuming that any noun you can think of has a neural representation which lives in a 25-dimensional vector space, where each dimension corresponds to one of those 25 verbs. And every image is some point in this 25-dimensional vector space. That's what that linear equation is doing when it's combining some weighted combination of these 25 axes to predict the image.

So, I don't actually believe that everything you think lives in a 25-dimensional space where the dimensions are those verbs. But the interesting thing is that the model works. And so it does mean that there is some more primitive set of meaning components out of which these neural patterns are being constructed. It's not just a big hash code where every word gets its own pattern. If that were the case, we wouldn't be able to extrapolate and predict new ones by adding together these different 25 components.

So patterns are being built up out of more primitive semantic components. And this model is crudely, only 79%, capturing some of that substructure that gets combined when you think about an entire word. And the substructure are the different meaning components.

The point here, I think, is, here's a model that's different from training a classifier. This is actually a generative model. It can make predictions that extrapolate beyond the training

words on which it was trained. It is assuming that there is a space of semantic primitives out of which the patterns of neural activity are built. And it is assuming that that space is at least spanned by the corpus statistics of the noun.

And since then, we've extended this work, and we no longer use just that list of 25 verbs. We actually use a very high 100-million-dimensional vector, which is generally very sparse, but where every feature comes from a much more precise parse of text on the web.

And for example, when I say parse, I mean if we have a simple sentence like, he booked a ticket, this would be a dependency parse. It's showing, for example, that booked is a verb whose subject is he and whose direct object is ticket. And now each of these edges in the parse becomes a feature in our new representation of the word. So instead of using verbs, we use dependency parse features.

And this actually increases slightly the accuracy of our former model from 79 up a little bit. But importantly, it also lets us work with all parts of speech. So now we're not restricted to just using nouns. We can use these dependency parse vectors for adjectives and all parts of speech. So in terms of broadening the model to be able to handle different types of words, this is helpful.

So at this point you could say, well, this is kind of interesting, because what have we seen? I think the main points so far are, gee, different people have very similar patterns of neural activity that their brains use to encode meaning. Furthermore, those patterns of neural activity decompose into more primitive semantic components. And we can train models that extrapolate to new words on which they weren't trained by learning those more primitive semantic components and how to combine them for novel words based on corpus statistics.

So that's kind of interesting. But everything that I've said so far is really about the static spatial distribution of neural activity that encodes these things. Now, in truth, your neural activity is not just one little snapshot. When you understand a word-- do you know how long it takes you to understand a word? About 400 milliseconds. It takes about 400 milliseconds to understand a word. Well, it turns out there is interesting brain activity dynamics during those 400 milliseconds. And let me show you.

So up till now, we were looking at fMRI data. But here's some magnetoencephalography data. And this data has a time resolution of one millisecond. So I'll show you this movie which begins 20 milliseconds before a word appears on the screen. In this case, the word is the word hand.

And this brain is about to read the word hand. You'll see 550 milliseconds of brain activity.

I'll read out the numbers so you can just watch the activity over here. So here we go. 20 milliseconds before the word appears on the screen. 0, 100, 200 milliseconds, 300, 400 milliseconds, 500. OK, so it wasn't a static snapshot of activity. Your brain is doing a lot of things. There's a lot of dynamism during that 400 milliseconds that you're reading the word.

fMRI captures an image about once a second, but because of the blood oxygen level dependent mechanism that it uses to capture that, it's kind of smeared out over time. So we can't see this dynamics with fMRI, but with MEG we can.

And so now we can ask all kinds of interesting questions, like well, what was the information encoded in that movie that we just saw? I just showed you a movie of neural activity, but I want a movie of data flow in the brain. I want the movie showing me what information is encoded over time.

Given this data, what could we do? Well, here's one thing we can do. In fact, Gus Sudre did this for his PhD thesis. He said, I want to know what information is flowing around the brain there, so I'm going to train roughly a million different classifiers. I'll train classifiers that look at just 100 milliseconds worth of that movie and look at just one of 70 or so anatomically defined brain regions. And I'll use a set of features-- he wasn't using our verbs anymore. He was using a set of 229 features that we had made up manually and that were inspired by the game 20 questions.

These were features of the word, not like, how often does a court does it co-occur with the verb eat? But instead, features like, would you eat it? Yes or no. Is it bigger than a bread box? Yes or no. And so forth. He had a set of 218 questions like that. And every word could be described by a set of 218 answers to those questions, analogous to the verbs.

And so what Gus did is, for every one of those features, every one of those 218 features like, is it bigger than a breadbox, he trained a classifier to try to decode the value of that for the word that you're reading from just 100 milliseconds worth of this movie, and looking at just one of 70 anatomically defined regions. And so when he did that, he ended up being able to make us a movie of what information is coded, in which part of the brain, when.

And he ran this-- every 50 milliseconds he'd move forward and use a 100 millisecond window starting there. So he found that during the first 50 milliseconds after the word appears on the

screen, none of those classifiers could reliably, in a cross validated way, produce any reliable predictions. Meaning the neural signals seems to not encode any of those semantic features during the first 50 milliseconds. By timing out to 100 milliseconds, there were no semantic features, but you could decode things like the number of letters in the word, the word length.

At 150 milliseconds, at 200 milliseconds, you got the first semantic feature. Is it hairy? I think this is actually a stand-in for, is it alive? But the feature he happened to uncover was, is it hairy? At 200 milliseconds. At 250, now we start to see more semantic features. 300, 350, 400, 450. So literally, these are the semantic features trickling in over time during this 500 milliseconds-- that's the movie-- that corresponds to the neural activity that I showed you in that first movie.

So this is a kind of data flow picture of what information is flowing around in the brain in that neural activity during that 450 milliseconds so far. Here's the set. Out of those 218 questions, here are the 20 most decodable features. So the number one feature that's most decodable, is that bigger than a loaf of bread? But actually, if you look at those questions, you see many of the most incredible ones are really size. And many of the next are manipulability. And many others are animacy. And some are shelter.

In fact, we've across a diverse set of experiments keep seeing these kind of features. Size, manipulability, animacy, shelter, edibility are recurring as features that have their own-- they seem to be kind of naturally some of the primitive components. And they have their corresponding neural signatures, out of which the encoding of the full word is built.

So if you ask me right now, what's my best guess of what are the semantic primitives out of which the neural codes are built, I'd say, I don't really know. But these features plus edibility, for example, keep recurring in what we're seeing. And they have their own spatial regions where the codes seem to live.

OK, so I want to get to the final part, which is, so far we've talked about just single words. And there's plenty of interesting questions we can ask about single words. But really, language is about multiple words. And so I want to show you a couple of examples of some more recent work where we've been looking at semantic composition with the adjective-noun phrases. This is the work of Alona Fyshe.

And what she did is she presented people with just simple adjective-noun sequences. She put an adjective on the screen like tasty, leave it there for half a second, then a noun like tomato.

And she was interested in the question of, well, where and when is the neural encoding of these two words, and what does that encoding look like?

So I'll show you a couple of things. One is, here is a picture of the classifier weights that were learned to decode the adjective. And you have to think of it this way. Here's time. And this is the time, the first 500 milliseconds when the adjectives on the screen. Then there's 300 milliseconds of dead air. Then 500 milliseconds when the noun is on the screen. And then more dead air.

This, the vertical axis, are different locations in the sensor helmet of the MEG scanner. And there are about 306 of those. The intensity here is showing the weight of a trained classifier that was trained to decode the adjective. And, in fact, this is the pattern of activity associated with the adjective gentle. Like gentle bear.

And so what you see here is that there is neural activity out here when the noun is on the screen long after the adjective has disappeared from the screen. That's quite relevant to decoding what the adjective was. And so this is just kind of a quick look. You can see that if I say tasty tomato, even when you're reading the word tomato, there's neural activity here, when you're looking at that noun, that encodes what the adjective had been.

And we can see that, in fact, it's a different pattern of neural activity than was here when the adjective was on. And in fact, one thing that Alona got interested in is, given that you can decode across time what that adjective was, is your brain using the same neural encoding across time? Or is it a different neural encoding, maybe for different purposes across time.

Let me explain what she did. She trained a classifier at one time in this time series of adjective-noun, and then she would test it at some other time point. And if you could train at this time, like let's say, right when the adjective comes on the screen, and use it successfully to decode the adjective way down here when the noun is on the screen, then we can know that it's the same neural encoding, because that's what it's doing.

And then she made a plot, a two-dimensional plot, where you could plot, let's say, the time at which you train the classifier on the vertical axis, and the time at which you test it on the horizontal axis. And then we could show at each training test time whether you could train at this time and then decode at this time. And that'll tell us whether there's a stable neural encoding of the adjective meaning across time.

When she did that, here's what it looks like. OK, so here we have on the vertical axis the time at which she trained. This is when the adjective is on the screen, the first 500 milliseconds, when the noun's on the screen. Here's then using any of these trained classifiers for decoding the adjective. Here's a different time at which she tried to use it. And again, here's when the adjective's on the screen, the noun.

And so what you see-- all this intense stuff means high decoding accuracy-- shows that if you train when the adjective is on the screen, you can use that to decode other times at which the adjective's on the screen. That's good. So we can decode adjectives. But if you try to use it to decode the adjective when the noun's on the screen, it fails. Blue means failure. No statistically significant decoding accuracy.

On the other hand, when the noun is on the screen if you train using the neural patterns when the nouns on the screen, then you can, in fact, decode what the adjective had been while the noun is on the screen. So it's like there are two different encodings of the adjective being used here. One when the adjective's on the screen that lets you successfully decode it when the adjective's on the screen, but it doesn't work when the noun's on the screen. And then the second one that works another neural encoding that you can use to decode what the adjective had been when the noun is on the screen.

And then interestingly, there's also this other region here, which says if you train when the adjective was on the screen, you can't use that to successfully decode it when the noun's on the screen. But later on, when nothing is on the screen, the phrase is gone, your brain is still thinking about the adjective in a way that's using this neural encoding, the very first of those neural encodings.

This is evidence that the neural encoding of the adjective that was present when you saw the adjective is re-emerging now a couple seconds later, after that thing is off the screen. But the neural encoding of the adjective when the noun was on the screen doesn't seem to get used again.

Most recently, we've also been looking at stories and passages. And much of this, not all of it, is the work of Leila Wehbe, another PhD student. And here's what she did. She put people in fMRI and in MEG scanners, and she showed them the following kind of stimulus.

So this goes on for about 40 minutes. One chapter of a Harry Potter story. And word by word, every 500 milliseconds, we know exactly when you've seen every word. So she collected this

data in fMRI and in MEG to try to study the jumble of activity that goes on in your brain when you're reading not an isolated word, but a whole story.

And so for her, with the fMRI we get an image every two seconds. So four words go by and we get an fMRI image. So here's the kind of data that she had. She trained a model that's very analogous to the very first generative model I talked about where we would input a word, code it with verbs, and then use that to predict neural activity.

In her case, she took an approach where for every word, she would encode that word with a big feature vector. And that vector could summarize both the meaning of the individual word, but it also could have other features that capture the context or the various properties of the story at that point in time. But the general framework was to convert the time series of words into a time series of feature vectors that capture individual word meanings plus story content at that time, and then to use that to predict the fMRI and MEG activity.

So when she did this, here are some of the kind of features that we ended up using. So some of there were like motions of the characters, like was there somebody flying-- this was the Harry Potter story. Somebody manipulating, or moving, or physically colliding. What were the emotions being experienced by the characters in the story that you're focused on at this point in time? What were the parts of speech of the different words and other syntactic features. What were the semantic content? We also used the dependency parse statistics that I mentioned that capture semantics of individual words.

So altogether, she had a feature vector with about 200 features. Some manually annotated, some captured by corpus statistics. And for every word in the story we then had this feature vector. Then she trained this model that literally would take as input a sequence of words, convert that into the feature sequence, and then, using the trained regression, predict the time series of brain activity from those feature vectors.

So this allowed her to then test, analogous to what we did with our single word noun generative model, to test to see, did the model learn well enough that we could give it to different passages, and then one real time series of observed data, and ask it to tell us which passage this person was reading. And these would be novel passages that were not part of the training data.

And she found that it was, in fact, possible, imperfectly, but three times out of four, to take a

passage which was not part of-- two passages which had never been seen in training, and a time series of neural activity never seen during training, and three times out of four, tell us which of those two passages they correspond to. So capturing some of the structure here.

Interestingly, as a side effect of that, you end up with a map of different cortical regions and which of these 200 features are encoded in different cortical regions. So from one analysis of people reading this very complicated, complex story, in this analysis, we end up-- you can go [AUDIO OUT] features and color code. Some of them have to do with syntax, like part of speech and sentence length. Some have to do with dialogue, some have to do visual properties or characters in the stories.

And you can see here is a map of where those different types of information were decodable from the neural activity. Interestingly, here is a slightly earlier piece of work, from Ev Fedorenko showing where there is neural activity that's selectively associated with language processing. The difference here is that in Leila's work, she was also able to indicate not just where the activity was, but what information is encoded there.

And then again, you can drill down on some of these. If you want know more about syntax, we could actually look at the different syntax features and see, well, where's the part of speech encoded? What about the length of the sentence? What about the specific dependency role in the parse of the word that we're reading right now, and so forth.

So this gives us a way then of starting to look simultaneously at very complex cognitive function, right? You're reading a story, you're perceiving the words, you're figuring out what part of speech, you're parsing the sentence. You're thinking about the plot, you're fitting this into the plot. You're feeling sorry for the hero who just had their brooms stolen, and all kinds of stuff going on in your head.

Here's the analysis that attempts to simultaneously analyze a diverse range of these features, and I think with some success. There still remain problems of correlations between different features. And so it might be hard to know whether we're decoding the fact that somebody is being shouted at, versus the fact that their ears are hurting, so to speak. But there could be two different properties we thinking of that are highly correlated. And so it can still be hard to tease those apart.

But I think that, to me, the interesting thing about Leila's analysis here is that it flips from a style that I would call reductionist. One way that people often study language in the brain is

they pick one phenomena, and then run a carefully controlled experiment to just vary that one dimension. Like we'll use words, and we'll use letter strings that are pronounceable that are not words, and then words. And we'll just look at what's different in those two almost identical situations.

Here, instead, we have people doing natural reading, doing a complex cognitive function, and try to use a multivariate analysis to simultaneously model all of those different functions. And so I think this is an interesting, methodologically, position to take. And it also gives us a chance to start looking at some of these phenomena in story reading.