Using Machine Learning to Study How Brains Represent Language Meaning

DR. TOM MITCHELL

E. Fredkin University Professor, Machine Learning Department,
School of Computer Science, Carnegie Mellon University

The President's Distinguished Lecture Series • Stevens Institute of Technology
January 31, 2018

Well, thank you, President Farvardin, and thank you, Stevens Institute, for the opportunity to come here and spend what's been a very interesting day for me. I've been on the listening side a lot, and I've learned quite a bit about some of the interesting things going on here on campus. What I want to do today is — well, let me introduce it this way. I've always been interested in intelligence. Who isn't?

And how it works — no, not the CIA kind of intelligence. Fundamentally, there are two ways to study it, and I've considered both. One is we could try to figure out what makes people so darn intelligent. And the other is we can try to build computers that are intelligent. And so those seem to be the two obvious approaches to studying it. And today, what I want to talk about is something kind of at the intersection of the two, but it really has more to do with trying to study the brain.

A video of Dr. Tom Mitchell's lecture, which includes his slide presentation, is available at **stevens.edu/lecture.**

The kind of question that I'd like us all to think about for the next 45 minutes is: How does the brain represent meaning of language, and how does it process language to get that meaning? So if you think about it, if I get you to think about the word *clicker* or the word *mother*, it's the same brain; it's just that when you thought about *clicker*, there was one pattern of neural activity, and when you think about *mother*, it's a different pattern of neural activity.

So I want to start there, because if we could understand something about the neural code for representing those different meanings, it would be at least a starting point for trying to understand thinking and language and so forth. So those are the questions we've been looking at for the last decade or so. And when I say we, I mean a number of colleagues here, including my primary colleague, Marcel Just, up at the top there, who is a professor in our psychology department who took me under his wing in the late '90s and started teaching me about functional magnetic resonance imaging, which he knew a lot about at the time and still does.

"How does the brain represent meaning of language, and how does it process language to get that meaning?"

So he's taught me most of what I know about that. But anyway, the work I'm going to present here is due to this collection of people.

So suppose you were interested in studying what patterns of neural activity represent meaning. One way you could do it is put people in an fMRI scanner and show them stimuli like these, which we show people. Sometimes we show them pictures of things or words. Sometimes we show them words plus pictures. But one way or another, we try to get them to think about some concrete concept, some concrete noun.

And when we do that, and we image their brain, we get data that looks like this. This is — it looks like four images — but it's actually four pieces of a three-dimensional brain image, where the top of the screen is the back of the head, the bottom of the screen is the front of the head. And this is the result of one person looking at the stimulus up in the corner there, *bottle*. So this is what one person's brain activity looks like. And the thermometer here on the side shows red means high neural activity, and green means average neural activity.

So that's what it looks like. That's the kind of data we can acquire. You might ask me, Does it look any different if, instead of *bottle*, we show them *clicker*? What I can show you is the mean activity if we average over 60 different stimuli, and that looks like this. So you can see *bottle* kind of like

your average word, but maybe there are some differences. And if I subtract out this mean activity from the activity for *bottle*, then I get the residue as this.

And so you can see, well, there are some differences. Maybe it's noise; maybe it's signal. But unless I say otherwise today, most of the time I'm going to be analyzing images like this, which are the difference between default neural activity and activity for a particular stimulus. So that's what our data look like. So again, if you're interested in the question of, how does neural activity represent meaning, then one of the first things you might try — and we did — is to train a machine learning system to take as input a brain image and make a classification of that brain image according to whether the stimulus that that person is thinking about is A or B, say *hammer* or *bottle*. So one of the first things we did was try to train these kinds of classifiers. We found that in fact this does work; you can train a classifier.

"... if we could understand something about the neural code for representing those different meanings, it would be at least a starting point for trying to understand thinking and language ..."

You show it examples. Here's the brain image when they're thinking about *bottle*. Here's another brain image when again they're thinking about *bottle*. Here's one when they're thinking about *hammer*. Given enough of those labeled training examples, a system can then learn a classifier so that you can give it a new brain image and try to tell you which of the two stimuli a person is looking at.

So when we tried this, we found, in this case — training a classifier to distinguish whether the person was thinking about *tools*, *hammer*, *screwdriver*, et cetera, versus *buildings*, *palace*, *house*, *castle* — then, in fact, the classifier worked quite well. Each of these black bars represents a different participant in our study, and the height of the black bar indicates how accurately we could classify beyond the training data, if we gave it after training new brain images to classify.

And so you see, for our best subjects, we were getting very high accuracy. Even for our worst subjects, we were getting better than the 50 percent that we would've gotten at chance. By the way, these differences among subjects tend to correlate with how much head motion we see them do in the scanner. Okay. So this is good. What this means, and the way I think of these trained classifiers, is they're a wonderful virtual sensor of information content.

So just like fMRI scanners are a wonderful invention — because for the first time, we can look inside people's heads and see the neural activity — these classifiers are also wonderful, because

they let us see, not just the neural activity, but what's the information coded in the mental state of that neural activity? So that means we can use these to answer very interesting questions.

An interesting neuroscience question is: Are these representations similar across brains? Is your neural representation the same as mine? Or are we all different? And we can turn this into the corresponding machine learning question, which is: Can we train our classifier using the people on this side of the room and then use it to successfully decode the people on that side of the room?

"An interesting neuroscience question is: Are these representations similar across brains? Is your neural representation the same as mine? Or are we all different? And we can turn this into the corresponding machine learning question, which is: Can we train our classifier using the people on this side of the room and then use it to successfully decode the people on that side of the room?"

And in fact, we ran that experiment, and the answer is absolutely, Yes. So here again, you see in black the accuracies we get from these different subjects if we train and test on data from that one subject. But here now you see in white the accuracy that we get, say, for participant 1, if we train using zero data from participant 1, but instead use all the data from all the other participants to train.

And you see that, in fact, on average the classification accuracy is really not very much different in the two cases. This is, to me, an astounding result. When we started this, I had no idea that we would find a result like this. What it really means is that, even though we all have very different backgrounds, very different life experiences, somehow we have a shared physical, spatial pattern of neural activity that encodes these different words.

We can run a similar kind of analysis by asking, What about cross-language? And in fact, we found bilingual Portuguese/English people. We trained them when we presented words in English. Then we tested whether we could decode the corresponding words translated into Portuguese. Again, the answer is yes, which is an indication that what our classifiers are actually learning is the neural activity that really corresponds to the conceptual representation, the real meaning, not the surface form of *casa* versus *house*, but the meaning that it's *house* in both cases.

Similarly, for words and pictures, we did the same thing, showing that we could train using words and then decode by showing pictures of these objects, decode which item was in the picture. We

also found that different kinds of words, concrete nouns, meaning nouns describing things you can physically touch, were what we began with, but then we got interested in emotions. We found that, in fact, the patterns of neural activity for emotion words like *anxiety*, *love*, *hate*, *angst* are equally decodable and shared across people.

We found that abstract nouns like *justice* and *democracy*, however, were much more difficult to decode across people and even within a person. So it turns out that, for some reason, you might speculate which reason, words like *democracy*, we can't even get a stable-enough, repeatable-enough, neural signature to train a classifier within a single person.

For verbs, we thought, Oh now we'll do verbs. And we found we couldn't. We presented individual verbs and tried to classify those and found it didn't work. But then we later found that if we took the same verbs and we put them inside a sentence, then they were as decodable as the nouns. So, for example, we had the verb *cut* and the verb *heal*. Presented in isolation, it was impossible to train our program to reliably distinguish those verbs. But if, instead, we used the stimulus *doctors cut* and *doctors heal*, then the verbs were as distinguishable as nouns.

So there's something about verbs that, in the human brain, requires a context, a grounding, in order to get a stable neurosignature. Interesting. So this was step one, and it kind of gave us some courage to push further, because once you understand that there is in fact a neural code that's visible enough through the noisy images of fMRI scans, despite the noise in fMRI, it's at least a strong-enough signal that we can train these decoders.

"... even though we all have very different backgrounds, very different life experiences, somehow we have a shared physical, spatial pattern of neural activity ..."

And once you realize that we all have pretty much the same neural code, then it actually makes sense to try to understand more about that code. If it had turned out the other way, that we all had different codes, maybe that would've been kind of the end of the story. But it wasn't. So we got more aggressive. Think about what would it mean really to have a theory of neural representation of meaning?

Well, it's not sufficient to just have a bunch of classifiers. If I had trained — you know, there are roughly 100,000 words in English — if I could train a classifier on all 100,000 words, I'd have, not a *theory* of neural codes, but I'd have a *catalog* of neural codes. So a theory is really some kind of formal system that can make predictions about phenomena that we haven't yet run experiments on, and then we can test that theory by collecting new data and seeing whether it's right.

So we thought what we'd really like is something like this. If we had a theory, it would mean that you could input an arbitrary noun and get a prediction of what the neural code should be for that noun. So we worked on this for a while, and the first method that succeeded looks like this. We built a computer model, again, using machine learning, that we could train.

And once trained, it would predict the neural activity, the fMRI image, for an arbitrary noun. And it predicts it in two steps. The first step is you take an arbitrary noun like *telephone*, in this case. Step one, we look that word up in a trillion words of text from the web, donated by Google, and represent that word as a vector. These days, we'd say *vector* and *vetting* if we're neural nets people.

"... the patterns of neural activity for emotion words like *anxiety*, *love*, *hate*, *angst* are equally decodable and shared across people."

But a vector of statistics about how that word is used in that text. And in our case, each of the green dots in that vector corresponds to how often the word, in this case *telephone*, co-occurs with each of 25 different verbs. How often does *telephone* occur with *hear* or with *eat* and so forth? And then in step two, that vector canonical representation of any noun, now coded by this 25-dimensional vector, is used to — and the machine learning part comes in — to predict the activity in each of 20,000 locations in the fMRI image.

So let me do this in a little more detail so you can see what this looks like. So, for example, here on the left is the result of step one for the noun *celery*. If we're trying to predict the image for *celery*, we find it occurs frequently with the verb *eat* and *taste*, but not very much with the verb *ride*. Whereas *airplane* occurs a lot with *ride*, but not very much with *manipulate*.

So these are the kind of statistics that come out of looking up the nouns and their verb occurrences on the web. And then in step two, now that we know that *celery* occurs this often with *eat* and this often with *taste* and so forth, the model predicts the activity in each location, like the prediction at voxel v-th, say here, this is the sum over those 25 verbs of the observed frequency with which the word *celery* occurs with the i-th verb, like .84 with *eat*, times some coefficient — that's the machine learned part of this system — some coefficient that tells us for the i-th verb, how much does that contribute to the v-th voxel in the image?

And so for each verb, the model learns these different coefficients. For each of the 20,000 locations in the image, it learns how much *eat* would contribute to activity in each location, and similarly learns that for taste. So in the end, it's about half a million parameters that get learned.

And then the linear combination like this gives us a predicted image. So that's the model.

And now you can ask, Well, does it work? And there are two ways I can show you. One is, I could show you, here are *celery* and *airplane* and the predictions made by a model when we did not even include *celery* and *airplane* in the training data. So the model was trained on other words. And here you see the predicted and the observed patterns of neural activity for those two words.

And you can see, it's not a perfect prediction, but it's capturing significant parts of the observed neural activity. And so this is encouraging. We can also ask more quantitatively. We could test the system, by showing it, after it's trained on these words, we can take two new words, collect the fMRI data for those two words, and say, well, here are two words: *celery* and *airplane*; here are two images: Which one is *celery* and which one is *airplane*?

And of course, if it guessed at chance, then we get 50 percent of those correct, just at chance. But in fact, we found that it was correct 79 percent of the time. And so, three times out of four, given words that it had never seen during training, this model is able to distinguish, predict the patterns of neural activity well enough that it can distinguish which of those two words generated which of these two patterns of neural activity.

"... any word you can think of results in a pattern of neural activity which is a linear combination of 25 fundamental components ..."

So just think about that for a minute and what it means about the systematicity of the neural code in your brain. What this means: the fact that we can predict, extrapolate beyond words that were even in the training set, means that, it's not that you have a hash code assigning randomly different patterns of neural activity to each word in some maximal distance way so that they don't get too overlapping. It's *not* that.

There's a distinct structure to the neural code. It's built up out of more primitive components. And in fact, in this model, what is this model saying? This model is saying that any word you can think of results in a pattern of neural activity which is a linear combination of 25 fundamental components, one for each of those 25 verbs. Now I don't actually believe that's true, but it's true enough, it's close enough to the truth, that it gives us a model that makes predictions with this kind of accuracy.

So it does mean that there's a systematicity in the neural code and we're capturing some approximation to that systematicity in this system. Okay, so that's kind of interesting. Now we got interested in, well, what's the real set of features? Honestly, I just made up those 25 verbs. So we

started experimenting with alternatives to those 25 verbs. We found we couldn't really do much better than the 79 percent accuracy we were getting, until my colleague, Dean Pomerleau, came in one day and he had gone on Amazon Mechanical Turk and had people answer 218 questions like this. Is it heavy? Can it break? Can it swim? If you've ever played Twenty Questions, this will look familiar to you.

He went on Mechanical Turk and he had people answer these questions for each of our nouns and represented them with 218-dimensional code instead of our 25. And this resulted in higher accuracy, so that was encouraging. But then in fact, the best accuracy we encountered of 86 percent was, I'm proud to say, due to turning the job over to a machine learning system.

So this is the work of my former Ph.D. student, Indra Rustandi. And what Indra did was, he built a machine learning system to learn the dimensions of the neural code, as well as the mappings. So how did he do that? Well, first of all, he started with 20 data sets. We had nine human subjects who were in a study where we showed them 60 stimuli. But the stimuli were a combination of a picture, a line drawing, with a word written under it.

"If I put a word on the screen, it takes you 400 milliseconds to understand that word."

We had another 11 subjects where they saw the same 60 stimuli, but this time only words. And so he took those 20 data sets and used them — so in each data set, we have basically 60 fMRI images, one for each of these 60 stimuli — and he ran a technique called *canonical correlation analysis* to learn a latent 20-dimensional representation of each of these data sets.

And when I say that, what I really mean is CCA learns a linear function that maps data set 1 to these 20 dimensions and a different linear function that maps data set 2 to the same 20-dimensional representation. So basically, CCA learns a linear function from each data set to map them into a common representation. And the reason it's called canonical correlation analysis is that the objective of the learning method is to maximize the correlation in these 20 different learned mappings, the correlation in the 20-dimensional vectors that they predict.

So, once we learned that, we don't yet have a model, we just have a 20-dimensional representation that we know can be predicted by 20 different linear functions from these 20 different data sets. But now Indra brought in the 218 Mechanical Turk questions that Dean had given us, and he just learned to predict these 20 latent features from the known 218-dimensional vector that represents

each word. And then finally, he inverted this linear mapping so that we could predict, given a word, we write down the 218 Mechanical Turk features, predict those 20 dimensions and then predict the different brain images for each person.

So I like this model for a couple reasons. One, instead of us pre-committing to what the dimensions are that represent the primitive components of neural activity, we have the program discover what they were. Second, if you look at this model carefully, you notice that, on the left, there's a part of the model that's independent of person. That's really subject-independent. Given a word, we map to this subject-independent 20-dimensional abstract representation of neural activity.

And then on the right side, we have person-specific mappings that take us from that population-wide summary of neural activity to the individual groundings of those neural activity in each brain. And I think this is the kind of model that we need more and more of, going forward. Because there are, in fact, individual differences. Our brains are even different shapes and sizes.

But if we believe that there's a shared representation, it's nice to be able to represent it at this abstract level and then have another part of the model take care of the detailed physical mapping to where exactly at the voxel level is the activity in each brain. Okay. So that's what I want to say about part one, which is really about the question of, What is the spatial distribution of neural activity that represents word meanings?

But now, you're probably thinking, What about timing? I'm thinking, What about timing? Do you know how long it takes you to understand a word? I didn't used to know this, but now I know. 400 milliseconds. If I put a word on the screen, it takes you 400 milliseconds to understand that word. So there's a lot going on — there could be a lot going on in that 400 milliseconds.

Unfortunately, with fMRI, we cannot see it. Because with fMRI, we get an image about once per second, and that's too slow. Furthermore, it's an image of blood oxygen fluctuations, which have a time constant of about five seconds. So there's no chance an fMRI is going to give us the time resolution we need. So, several years ago now, we went off and started studying a second kind of imaging called *magnetoencephalography* (MEG).

Unlike the fMRI, magnetoencephalography is — it's a ridiculous technology. You sit with a helmet around you which has super-cool squid magnetic field sensors, and the machine literally just passively listens to magnetic fields coming out of your brain. They're very, very faint magnetic signals, so that if a bus drives by, it causes more of a signal than your brain did.

But the nice thing about MEG is it gives us much finer time resolution. It gives us 1 millisecond. Remember it took you 400 milliseconds to understand a word. So now we can actually look at the timing of this. We can ask — well, let me show you a movie. Here is a movie showing the activity in the brain. I'm starting at 20 milliseconds before a word appears on the screen.

In this case, it's the word *hand*, plus a line drawing of a hand. And I'm going to play this movie out to 550 milliseconds so that you can see the neural activity that happens while this person is reading that stimulus. As you look at the brain, I will just read out every 100 milliseconds, so you don't have to look at the clock. Okay, here we go. 0 ... 100 milliseconds ... 200 ... 300 ... 400 ... 500. Okay.

So that's what the brain activity actually looks like when you're reading a word. It's not a static picture, and it's not a picture of just faint neural activity getting brighter and brighter with the same spatial distribution. There's interesting dynamics there. Okay. So this is interesting, because with the fMRI data, we get a different view of what's going on. We get some kind of integral over time of neural activity spatially. Now we can see the time.

So now if you're a machine learning person, you're going to look at that movie and you're going to say, Well, that's interesting, but what I want is instead a movie of the information flow coded in that neural activity over time. If you're a machine learning person, you're going to do what my student, Gus Sudre, did for his Ph.D. thesis. Gus said, well, I'm going to train about a million classifiers, and he did; each classifier is going to look at a 50-millisecond subwindow in time.

It's going to look at one of 70 different regions in the brain. And using that 50-millisecond window of neural activity in that one brain location, it's going to try to predict some semantic feature of the stimulus. And he used the 218 Mechanical Turk features, like Is it *hairy*, Does it *breathe*, those kinds of things. So for every possible feature, every 50-millisecond time window, and every location in the brain, he trained a classifier to predict that feature from that activity.

And then he tested on held-out data how accurate those classifiers were. And of course, most of them had accuracy of exactly chance, because that part of the brain at that time was not coding that feature. But some of them succeeded. Some of them actually did successfully predict particular features and particular times in particular places. So he took this movie and turned it into — there's Gus — turned it into the following movie instead.

This goes by frames of 50 milliseconds each. In the first 50 milliseconds, absolutely nothing about the stimulus word could be decoded from neural activity. But at 100 milliseconds, in several

places we could decode the word length, the number of letters in the word. Not a semantic feature of the word, but the number of letters. In the next 50 milliseconds, still perceptual features but nothing semantic. At 200, we get our first semantic feature. Is it *hairy*? Which I think is actually a placeholder for, Is it *animate*?

And then at 250, more, and at 300, more features, and at 350, and at 400. So now we can basically use this idea of training a classifier as a virtual sensor of information content in the neural signal. We can run that across time windows and also run it around in different locations in the brain and end up with the first picture of, What's the sequence of information coded across the brain during the 400 milliseconds that it takes you to understand the word?

"So it's not a confusion of blurring of the image; it's literally their distinct physical regions where magically, simultaneously, this feature is encoded in the neural activity."

And sure enough, at around 400 is when the greatest number of semantic features are encoded by the neural activity, and it starts dropping off again after that. So this is kind of interesting. In fact, I want to show you a little bit more detail on this one. Because before we did this, we had all kinds of questions, like what happens once your brain finds a feature, does that feature then move around, like it found out, is this *animate*? Does that feature just kind of stay there in that one part of the brain, or does it move around? How long does it stay? Does it stay 100 milliseconds, 500 milliseconds? What does this all look like? So what I can show you, and I will on the next slide, is a picture where I'll show a matrix, and each column is one of 75 different locations in the brain, and each row is time, this would be the time axis, starting from onset of the stimulus and going forward.

And then I can show in color which locations and which times we can successfully decode a particular feature. So here it is for the feature word length, that is the number of characters in the words. So this is a perceptual feature, not semantic. But what you see here is in blue means we're just getting chance accuracy, trying to decode. But here you can see, red means that with relatively high accuracy in the left and right cuneus at about 100 milliseconds to 150, we can decode the number of letters in the word pretty accurately.

But interestingly, there are a whole bunch of — half a dozen different regions — where suddenly, simultaneously you can decode this feature. And these regions are not right next to each other. For example: well, many of them are left/right pairs, like left/right occipital or like back here and back

here. So it's not a confusion of blurring of the image; it's literally their distinct physical regions where magically, simultaneously, this feature is encoded in the neural activity.

And then it goes away. Actually, it lasts for about 100, or at most, 200 milliseconds, but by the time you get to 400 milliseconds, which is when your brain really understands the word, there's no place in the brain where we can reliably decode the letter features anymore. It's like your brain is done with that. It's throwing that away. It's on to the semantic features.

So, and in fact, I could show you these similar plots for semantic features and you see that they occur later in time, like the feature — is it *animate*, or can you *pick it up*, or is it *graspable*, or do you *love it* — are all semantic features. They come later than this 150 milliseconds. Okay. But now back to the point. I mean, this was a total surprise to me that the way it works is that your brain simultaneously in multiple distant regions generates neural activity that codes the same feature.

So this raises the obvious question: Are those different parts of your brain communicating with each other? What's going on? And so, more recently, again, we can ask, Are these collaborating? We can turn that into a data analysis question, which is, Are these different neural activities in these different brain regions synchronized? Are they going up and down together? And are they synchronized particularly at the times when they're coding the same information?

"So the size of the object, even though we're presenting line drawings that are exactly the same size ... the actual physical size of a bumblebee or a truck is coded in your neural activity. So is manipulability. So is animacy. And so is shelter. And these four themes are actually things that we've seen across many experiments now and across both MEG studies and fMRI studies."

We've looked at the neural activity in the different regions. It's not the same activity. But maybe it's phase-locked, maybe it's in synchrony. And so Maria Toneva, here she is, did an analysis and found that in fact, yes, in fact the neural activity — she picked the one region where the activity is most strongly coded right here, the darkest red, left lateral occipital — and she looked at what some people call *functional connectivity* or the *temporal synchronization*, between that region and the other regions that code, in this case word length, so well.

And what she found was, yes, those other regions are correlated in their neural activity with

this one. And furthermore, the correlation peaks about 25 to 100 milliseconds before they simultaneously code that information. So one hypothesis would be that these regions, they're communicating. They're oscillating together. They're phase-locked. And they're somehow working together to figure out what the answer is, to compute whatever this property is that they're encoding.

And then while they're doing that, they're in synchrony. And then once they get the information, that synchrony goes away, and they all just have it. So that's not a proof that that's how the brain works, but this is data that suggests that that kind of coordinated, synchronous brain activity across regions precedes the time when suddenly all of these different regions code that information. So maybe that's a hint about how it's working.

Okay. So that's what I wanted to say about time. Let me just before I leave this tell you that if you ask me, out of those 218 features that represent different aspects of word meaning, which ones were most visible, most decodable from the neural activity, here's the answer. Here are the top 20 in order. And if you look at that list, you notice they fall into interesting groups.

So the size of the object, even though we're presenting line drawings that are exactly the same size, if it's a line drawing of a *truck*, it's the same size as our line drawing of a *bumblebee*, but the actual physical size of a *bumblebee* or a *truck* is coded in your neural activity. So is manipulability. So is animacy. And so is shelter. And these four themes are actually things that we've seen across many experiments now and across both MEG studies and fMRI studies.

They seem to be emerging as key primitive components of meaning representation across at least the nouns that we've been dealing with. Okay. So I want to finish up by looking at the third of these three dimensions. So we started by looking at the fMRI, spatial distribution of neural activity coding words. Then we looked at the timing, the 1 millisecond resolution timing for understanding a single word.

But more recently, we've been looking at, How does the brain combine multiple words? Language is really about multiple words. So how does the brain combine, say, a noun and an adjective, *hungry cat* versus *cuddly cat*, into the meaning of the thing? Or how does it interpret sentences? Or in this case, the example I chose for today is a story-reading experiment that Leila Wehbe, one of my former students, did for her Ph.D. recently.

And she put people in a scanner and she showed them the following stimuli. I want you to pay attention here. Imagine you're lying still in the scanner. This is what you're going to see … Okay. So that's a story. In fact, the story continues for about 35 minutes like that. It's one chapter of a

Harry Potter book. And you're seeing one word every 500 milliseconds. Remember it takes you 400 milliseconds to understand a word, so no problem there. And you could see that even in that short little sample, you get used to it pretty quickly.

So you're reading in a — it's not exactly natural reading, but it's pretty close. And it has the advantage that we know down to the millisecond when you saw each word. So Leila put people in the scanner, both fMRI and MEG, showed them 35 minutes of a Harry Potter story, and then her job was to figure out, what can you learn from this? So she did what I think are two very interesting things.

First, in the fMRI data, this is the kind of thing she would get out. So presenting a word every 500 milliseconds, she's getting one fMRI image every 2 seconds. So this is the kind of thing she got. From this, she trained a model reminiscent of the very first one that I told you about, where we had a single word, we would represent it by a vector embedding of verb frequencies, and then predict the fMRI activity.

So she used the same idea to predict — so for each word in the story, we're going to construct a vector and then use that to predict the sequence of fMRI and MEG data. But now, because it's a story, Leila and I sat down and made up a bunch of features and labeled every word in the Harry Potter chapter with what ended up being 199 features. And some of these were low-level perceptual features, like How many letters are in the word? Others were word-level features, like the current word, Is it a noun or a verb or a determiner?

But some of them were about discourse, like Are we in the middle of a conversation between two characters? Is one of the characters doing some physical motion? Did we just move to a new scene in the story? So we tried to capture features all the way from perceptual to individual word semantics, syntax and discourse-level properties. And then she trained a model to predict the sequence of neural activity from the sequence of these 199-dimensional features per word in the story.

She tested that model then, again, in an analogous way to before; she would train it, then she would show it segments of the story that were left out of training, two new text segments, like the two on the left. The model would have to predict a sequence of fMRI images. And then she would give it the observed fMRI image sequence for one of those two passages. And basically, she tested whether the model made good-enough predictions of the neural activity that it could distinguish which of those two passages was being read during this brain image sequence.

And this worked, again, with three times out of four. But what's interesting about it, and this is what I want to show you, is what the model leads to is a refinement of things that had been in the literature before. So here's the 2012 paper by Ev Fedorenko at MIT, who had mapped out regions of the brain that she said were related to reading. These are regions of the brain she published as related to reading. But because of the kind of model Leila had built, she could say not only what regions were related to reading, but what information were they coding.

"... understanding a sentence or a story is much more than independently understanding each word and then forgetting the context. It's all about context in some sense. And so we're interested in how we could study this and came up with the following idea: We trained a neural network to predict the next word in the story, based on the history of words up until time T."

So what are the different regions — so in Leila's work, she could distinguish what parts of those brain regions that were active were, for example, coding dialogue between characters versus motion of one of the characters versus syntax of the current word that you're reading. So again, taking advantage of the idea that these classifiers really can be treated as virtual sensors of the information content.

All right. I have one final thing that I'll wrap up with. It has to do with context. So understanding a sentence or a story is much more than independently understanding each word and then forgetting the context. It's all about context in some sense. And so we're interested in how we could study this and came up with the following idea: We trained a neural network to predict the next word in the story, based on the history of words up until time T.

This is an idea stolen from the natural language community. So we trained a recurrent neural network where the input is the word at time T, plus some context, which I'll call S. We'll see in a minute where that comes from. And the network is trained to predict the next word, the probability distribution, over the next word. So if I say to you, for example, *Harry jumped on his*, you might think the next word is *bed* or *broom*, but probably not *yellow*, or maybe *yellow*, probably not *cloud*.

So you have some kind of probability distribution in your head about what the next word could be. So we trained a neural network to predict the probability distribution over the next word, given the current word and a summary of context. But it was the neural network that had to learn the summary of context by feeding back the hidden layer of the network from the previous time. This is a standard kind of recurrent neural network.

But the point is that the network has to, somehow, in order to succeed, has to learn an internal coding of the history up until time T. And so we thought if we can train that network, and we did, then we could use that learned code of the context, of the history, and see what neural activity in the brain, whether there was any neural activity in the brain, that was actually encoding that context.

And the answer turned out to be yes. So one way to think of it is this: so we have the sequence, *Harry had never*. Then when we're looking at the word *never*, we have the code for *never*. We also have the context, which is based on combining the previous word with the context before that. And we also have the probability distribution over the next word. If you look at the MEG activity and you try to decode where we are in the story using these three different types of information, you find that in fact the context is the best information. If you're trying to see where you are in the story, it's the context information that's most vibrant in the neural activity. And the current word is secondary to that. And in fact, we can map out where in the brain these different features are encoded and have. So the point of this one is that, again, this is, I think, a direction that we want to go more of in the future, which is to study language processing in the brain, you have to ask the fundamental question.

"Once we have an answer to how the brain processes language, what will that answer look like? Will it be an equation? Will it be a schematic of the neural connectivity in the brain? For me, the answer will look like a computer program, because the brain is fundamentally an information processor. "

Once we have an answer to how the brain processes language, what will that answer look like? Will it be an equation? Will it be a schematic of the neural connectivity in the brain? For me, the answer will look like a computer program, because the brain is fundamentally an information processor. And if you want to tell me, once you understand how the brain processes language, I think the most natural way for you to tell me is to give me a computer program that processes language.

Show me the mapping between, if I give that computer program a particular word sequence to read, you can prove to me that your program is a good model if you can show me that, when it reads that sequence of words, it can predict the sequence of neural activity when a person reads

“Brain imaging is a revolution. It’s a 25-year-old revolution, but it’s still young. There’s a lot that we’re going to learn, including how to use machine learning in similar methods to analyze and interpret that data. To me, the field is moving happily toward a point where AI systems become models, become our way of studying and proposing models of how the brain performs similar functions.”

the same sequence. And so this is a step in that direction. It’s not the solution, but it suggests, I think, the kind of thing that we need to see more of and will see more of over the coming decade: that it’s more of a convergence of AI and neuroscience, at least cognitive neuroscience, to studies where we have computer programs that perform a function and we use those literally as models of the information processing steps the brain must also be doing. And the way we prove whether we’re right or wrong is by using those models to predict literally the sequence of neural activity. And if you have a better model than I have, then presumably, your model will predict the neural activity better.

So what we will be arguing about is not whether your program understands the sentence, but whether your program which understands the sentence is a better predictor of the neural activity than program 2, which also understands the sentence. Okay.

So let me end there. Thank you for your patience. So here’s my two-sentence summary of the talk. Brain imaging is a revolution. It’s a 25-year-old revolution, but it’s still young. There’s a lot that we’re going to learn, including how to use machine learning in similar methods to analyze and interpret that data. To me, the field is moving happily toward a point where AI systems become models, become our way of studying and proposing models of how the brain performs similar functions. So we’re still in the early days, but anybody here looking for a thesis topic, this is a great area. Thank you.