Gary Stix: Your research is focused on understanding the neuronal representations and computational algorithms that underlie visual object recognition in primates. You’ve asked some very basic questions recently, such as what is object recognition and why is it challenging?

Dr. DiCarlo: We all have an intuitive feel for what object recognition is. It’s the ability to discriminate your face from other faces, this car from other cars, a dog from a camel. But making progress in understanding how our brains are able to accomplish this task is a very challenging problem.

Part of the reason it’s challenging is defining what it is; and it’s a challenging problem because we’re using our brain machinery to solve the problem. It seems effortless to us. When I try to explain, for instance, to my mom, “Oh, I’m working on object recognition,” she will reply “Well, I just see and I solve the problem,” but of course, she is using the machinery of her brain that we are trying to reverse engineer and to decipher.

What makes the problem challenging is that each object presents an essentially infinite number of images to your retina, so you essentially never see the same object, the exact same image twice. The ability of the brain to deal with all those different images and still know that they’re coming from the same object in the environment is a challenge our brain solved, we think, through evolution. But machines are still struggling to understand how to do it.

Gary Stix: This is one of the really big problems both in neuroscience and in machine learning.

Dr. DiCarlo: It’s a really great problem because it’s at the intersection of neuroscience and machine learning, and of psychology or cognitive science, because objects in the world are what we use to build higher cognition, things like memory or making decisions. Should I reach for this? Should I avoid it? Our brains can’t do higher cognition without these foundational elements that we often take for granted unless you’re a computer scientist and realize how challenging that problem is.

Gary Stix: Your research is focused on finding some of those foundational elements. There are areas of the brain, such as the inferior temporal cortex, where there might be some clues.

Dr. DiCarlo: It’s been known for several decades that there’s a portion of the brain, the temporal lobe, that when lost or damaged in humans and nonhuman primates leads to deficits of recognition. So we had clues that that’s where the algorithms live. But just saying that that part of your brain solves the problem is not much more specific than saying your brain solves the problem. It’s still a very large piece of tissue. But anatomy’s told us more about the various visual areas that exist in the brain, a whole network of areas that exist there.

The tools of neurophysiology, and now more advanced tools, allow us to look more closely at neural activity, especially in nonhuman primates. Then we can begin to decipher the actual computations to the level, for example, that an engineer might need to be able to emulate what’s going on in our head, or we might need if we wanted to replace those circuits or augment those circuits—to really understand the function at a detailed level, not in just the gross “this part of the brain” sense.

Gary Stix: How are you trying to drill down and find one of these foundational elements?

Dr. DiCarlo: The foundation of any science is the ability to have predictive models of a phenomenon. So for object recognition, as an engineer, if you want to emulate that, you first need to define what are we trying to predict. What would success look like? There are various levels that we might set there, but we’ve set one, a goal for ourselves, the ability to emulate what we do in core-object recognition.

Put simply, what that means is that I’m going to show you an image for about 200 milliseconds, which is about the time that your eyes dwell as they explore a scene. It’s a time that’s driven from the biology of how you explore the world with your eyes. We can do a lot with that short time window. We can easily recognize one or more
objects within that one-fifth of a second glimpse. It’s not all of vision, but it’s a defined space where we can start to get some traction on the problem.

**Gary Stix:** So you’ve got a predictive model, and then you want to test that model.

**Dr. DiCarlo:** In core recognition, images come in and are processed by the eyes, and then through a series of visual areas, processed somehow in ways that are murky, but we can record the neural activity along the processing pathway. Others have done that before us, and now we’re doing it at a much larger scale. We can record neural activity, and we’re especially interested in a place in the brain called the inferotemporal (IT) cortex, at the end of the temporal lobe, the highest level of this processing chain.

We’ve found that the population patterns of neural activity there, the firing patterns of all the neurons in that part of the brain, with a very simple model, an algorithm if you will, can predict very accurately the animal’s perception and our own perception—our ability to do recognition in that core domain. So we have a model from inception and our own perception—our ability to do recognition that we can test that model.

**Gary Stix:** So you could predict, say, that a person or a monkey is looking at a tree in the background from examining that neural activity?

**Dr. DiCarlo:** That’s exactly what I mean. The granularity with which we can do that is still being studied but I can tell if you’re looking at a tree versus a dog, or I can tell you you’re looking at a tree versus a car. I can tell if you’re looking at one tree versus another.

We’re now trying to see if we can do this on a trial-by-trial basis rather than on an average basis. Already the models are quite good for the tasks that we’ve tested. Our next step is to look even harder and build better models.

**Gary Stix:** The challenge with object recognition is that if I’m looking at a tree and then move slightly to the left or the right, the tree changes or I start to see another tree. Will your model still recognize that that’s a tree or that it’s the same tree?

**Dr. DiCarlo:** That’s the largest thing the model has to deal with. When I say the model deals with that, the neurons up to the IT cortex when we record have dealt with that, so that part of the problem from an engineering perspective is solved.

So once I build a decoder on the IT cortex, a reading of the IT activity, if you will, then a new image of the tree will be properly decoded as a tree. It’s a brand-new image, but the model will still make a prediction of what you will say, and the model will be quite accurate.

**Gary Stix:** What are some of the implications of this for machine learning and perhaps, one day, even for understanding problems that people have with disruptions to neural circuitry?

**Dr. DiCarlo:** From the machine-learning point of view, this neural activity in IT is something that machine-learning folks would call features. Those are features computed on the image. They’re a very powerful set of features for the reason I described. What many people would love to do is to be able to have algorithms that produce those features. Much of machine learning is devoted to finding good features.

The brain, through evolution, has already found some good features, and that’s essentially what we’re reporting. Here are some nice features. Here’s where they are. Here’s our evidence that they’re nice features. Now the machine-learning community is working, and we’re working alongside of them, to help build what are called encoding algorithms that produce those features.

There’s a lot of exciting progress in the field, even in the last few years, driven by what are essentially brain-inspired models that are actually now some of the state-of-the-art computer vision algorithms. They were inspired by work of the type done by those who went before us, and I hope from some of our work, to build those kind of models. They turned out to be very powerful in the computer-vision and machine-learning community. Those are the most active and exciting models in computer vision.

**Gary Stix:** The grand vision of what you’re doing is the ability to model this all the way from encoding to neural activation and then to the decoding and perception in the brain?

**Dr. DiCarlo:** That’s exactly the grand vision. If we can do all of that, then we would have a complete end-to-end understanding of this domain of behavior.

**Gary Stix:** How long until you get there?

**Dr. DiCarlo:** Within basic-level core recognition, it depends on your level of detail, but in the next 10 years we will have a very good understanding of core basic-level object recognition to the degree that many engineers will be satisfied. We won’t know it down to the synapse, but we will know it in the way that the algorithms are very predictive of the neural activity at various levels of the system.

**Gary Stix:** Do you think that this could provide some insight into what goes wrong when the circuitry in the visual cortex or the temporal cortex goes awry?

**Dr. DiCarlo:** The most common deficit that affects recognition is major damage to that part of the brain through stroke or lesion. You’ve taken out those neurons. Maybe our studies could lead to ways that you could sidestep the damages or replace the function.

There are other deficits to the temporal lobe where people have things like deficits in the ability to discriminate among faces or other types of objects. They’re not very common, but this kind of work should bear on those deficits, as well. We hope it will also bear on things like how kids learn to read.

At the end of the day, whenever you’re doing visual tasks, you’re leaning on these kind of representations to do much of your vision. I think this work will help us
understand these higher-level issues of, say, social cognition or things like dyslexia. Things that depend on these foundational representations will be better understood because of our understanding of these neural circuits.

**Gary Stix:** Despite what people see in movies, robots in the real world are still very limited in what they can do. One of their big limitations is the ability to recognize and process information they perceive. Do you think this work could help with that?

**Dr. DiCarlo:** Certainly. The computer-vision community is already using brain-like algorithms. The next frontier though is expanding the domain of task, not just what you can do in 200 milliseconds, but as you explore a scene with many eye movements or navigate a scene, accumulating information over time. There will be more feedback in the system.

Those are things that we’ve not yet begun to touch very much, but that’s the next frontier as we understand this more core domain, this foundational domain. I won’t say that if we do this work we will have robots doing everything you see Data on *Star Trek* doing, but it will be a foundation to enable those next steps.

**Gary Stix:** Is a larger conceptual framework necessary to create kind of a movie of this whole process, to go beyond 200 milliseconds and create something that can process the full variety of our visual environment?

**Dr. DiCarlo:** I think what you’re asking is what goal we are seeking. Sometimes to know that we’ve met it, we have to just emulate the system. For instance, if we could build a robot that behaved as well as us and fooled us like Data on *Star Trek*, I think we would all declare that as evidence of success.

This will certainly require, at some point, really embodied systems to test. Vision is not an isolated sense. It has to interact with the motor system, so as these systems get better, you’re going to see them on moving things like robots and other autonomous devices. This is already happening. Then the work of my lab and others will have to move beyond vision, and integrate with motor action and other senses.

**Gary Stix:** The complexity of that seems overwhelming. Don’t you think it’s going to take many generations to get to these ultimate goals?

**Dr. DiCarlo:** In that goal, probably. Again, it’s hard to make predictions with confidence. In this ventral visual stream that ends in IT cortex is a series of cortical areas where each local bit of cortex looks anatomically very, very similar. One of the exciting ideas that’s been in our field for a long time is that there’s some interesting local learning algorithm that’s of smaller scale in both size and computational power than the global algorithm I’ve been talking about.

One of the things many of us hope for is that if we could get some insight from this kind of work into that more local algorithm, that could generalize more broadly to things that we’re not even thinking about, that might apply to audition or other senses, maybe to some of these higher-level cognitive things, but that’s hard to see right now. So there’s a hope for a shortcut, if you will, that arises from working on a clearly defined problem.

**Gary Stix:** You mentioned in a recent paper that there are 40,000 neuron building blocks. Is that what you mean?

**Dr. DiCarlo:** That’s exactly what I’m referring to: the idea that you might have sub-modules of something of that scale, say, a millimeter of cortical tissue. Again, this is an idea that’s been around for a while because each piece of cortex looks very similar. What is it doing? What’s that fundamental operation?

We are working on it from the problem of vision, but we hope we will get some understanding of that because we set a clear goal of what success looks like. We’re building models to emulate what the neurons do at all levels, and so we might start to glean some insight that could generalize outside of vision.

**Gary Stix:** So these mini modules might be present in other parts of the cortex that are processing incoming sensory information?

**Dr. DiCarlo:** That’s right. They’re known to be present. This is a hypothesis that’s been in the field for a long time. Like the dream of discovering the DNA code, discovering the cortical processing algorithm, is a long-standing dream.

It will be a big jump and certainly not the whole story. Bits of cortex have to talk over long distances, so everything’s not going to be defined as a local small module, but we might get a very big boost from understanding that kind of module.

**Gary Stix:** When you say an algorithm, is that made up of a kind of code that would be present, say, in the inferior temporal area?

**Dr. DiCarlo:** People use words like algorithm and models in different ways. If we talk about the algorithm of the ventral stream, that is maybe executed by many subalgorithms, that consist of these building blocks, but when put together, give rise to the global algorithm of the ventral stream. The hope is that the global algorithm of the ventral stream built by subalgorithms that are simpler and more primitive and more understandable and that might generalize well. That subalgorithm in a piece of cortex might take some set of inputs, some learning, local processing, and give some set of outputs. It’s essentially a little box that does something clever, and when stacked horizontally across the visual field and then stacked vertically into a deep stack, visual area V1, V2, a series of areas along the temporal lobe, might give rise to the global temporal-lobe or ventral-stream algorithm that I was referring to earlier.

Algorithm is a word computer scientists like, but as neuroscientists, what we’re imagining is not just a set of lines or codes that says, “If this, do that,” in a way that those of us who write code are used to. It’s also going to be a set of neurons that execute that function.

We’re building algorithms that are actually already constrained somewhat by the biology. We’re not just
writing code and hoping that it predicts the responses. We’re building systems that are modeled on the biology so that they will translate both to give constraint and to what this neuron does within the algorithm. That’s a different level of how deeply you want to look at the system, how you want to describe it.

What’s exciting about neuroscience and also challenging is that you have to work at all these levels—what’s the goal of the system, how do we describe the algorithm, what are the mechanisms of the neurons and the synapses? We’d like to say we understand all those levels of the system.

**Gary Stix:** There has been work in the last few years on retinal prosthetics. One approach has been to implant codes into prosthetics so that they can process incoming photons in the same way the retina does. Does the kind of work you’re doing connect in some ways with this?

**Dr. DiCarlo:** This is one of the things we’re most excited about right now. There are visual prosthetics for people who, say, have lost a retina. The dominant approach is to try to bypass the retina and reinject a spatial pattern of activity, say, in an early visual area or the subcortical area that comes right behind the retina, called the LGN.

That makes sense from an engineering point of view. The downside is that if you want to get an image in, it’s a very high-dimensional space, with many, many pixels that you’d want to play to make it feel like your normal vision would feel like.

**Gary Stix:** To sum up, what your work is doing is taking something very basic that all of us can relate to and try to find the physical basis in our brains of how we see, how we recognize an object, and come to a fundamental physical and theoretical understanding of that really huge challenge.

**Dr. DiCarlo:** That’s really been the goal of neuroscience since its formation. We believe the brain is a set of mechanisms that give rise to this amazing phenomenology that we feel. We are studying one example of that phenomenology but it’s one that many of us can relate to and it would be a foundational success if we can create that end-to-end understanding—a large brick in the foundation of building towards understanding cognition.