Research Statement
Michael H. Coen

1. Introduction

My primary academic interests are developing biologically-inspired approaches to machine learning and reciprocally, to using these approaches to better understand learning in biological systems. My research has been motivated by the observation that animals routinely solve extremely difficult, nonparametric learning problems during development. The goal of my work is to create more sophisticated computational systems by understanding how animals can solve such problems. In my work, these two issues are inexorably linked.

My doctoral thesis presented a new form of *self-supervised* machine learning. This work, which received MIT's Sprowls Award for best dissertation in Computer Science, was inspired by the observation that the sensory information gathered by animals is inherently redundant. This redundancy can enable learning without requiring explicit teaching (as in supervised learning) or statistical modeling (as in unsupervised learning). In Nature, these are frequently unavailable and yet animals learn anyway. In other words, redundancy allows animals to supervise their own learning, hence the designation *self-supervised* learning. It also enables a powerful computational framework to provide this ability to machines.

My current research builds on my doctoral work to enable self-supervised segmentation of human functional MRI (fMRI) data. The framework of my thesis offers a new way to identify functional regions in the brain without knowing of their existence *a priori*. This is an especially attractive approach to this problem, because we lack any detailed functional models of modular brain structure. Not having such models makes it difficult to apply standard machine learning techniques to elucidate functional regions within the brain, which are almost entirely unknown. Among my future research goals is the generation of a distributed, geographic atlas of modular brain functions, using self-supervised learning techniques. Figure 1 presents very recent results demonstrating the automatic detection of the fusiform face area (FFA) using this approach, which

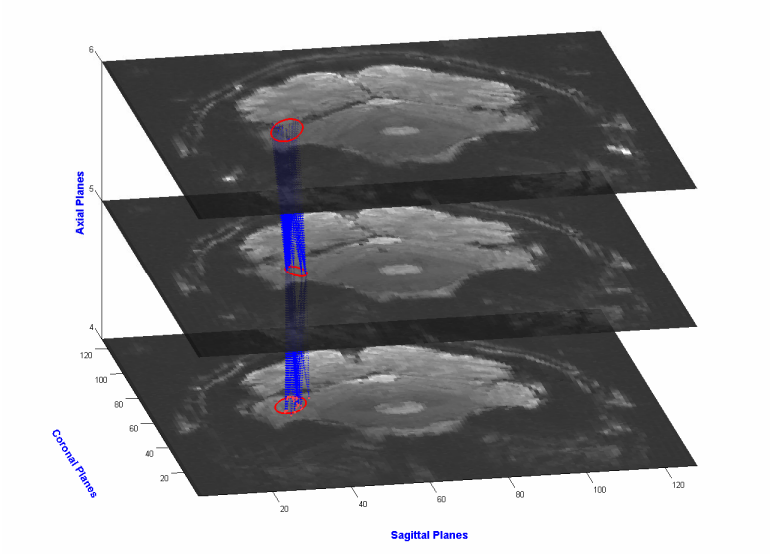

Figure 1 – Self-supervised learning of the fusiform face area (FFA) in functional MRI data. The red ellipses outline the FFA voxels. The blue lines indicate Hebbian co-occurrence linkages between the voxels, which are used by the learning algorithm. This is joint work with N. Kanwisher and L. Reddy, Department of Brain and Cognitive Sciences, MIT, and C. Baker, NIH.

only uses fMRI data and not the actual experimental inputs describing the images or their categories.

Because of its biological motivation, my doctoral work approached machine learning as a perceptual problem, even in non-perceptual domains. In doing so, it makes a number of predictions about the spectrum of sensory integrative disorders in animals and people, commonly grouped under the rubric of *autism*. These address the consequences of information *not* being properly shared, which would interfere with the self-supervision necessary for learning. These hypotheses are testable using the fMRI framework described above and can suggest treatments for overcoming clinical deficits, by relying on alternative strategies for deriving proxy information which would otherwise be unavailable. Connecting my research to real-world, socially valuable problems is a fundamental interest of mine and provides opportunities for obtaining funding outside of traditional computer science channels. These include the NIH, NIMH, PHS, NSF, and especially private foundations.

Although I am primarily a computer scientist, I have a very strong interest in the formulation of biologically plausible theories of learning and development and evaluating their empirical plausibility in close collaboration with neuroscientists.

2. Foundational Work

I briefly outline my doctoral work, which provides the foundation for the research agenda described above. We begin with a simple thought experiment.

Consider how a human infant acquires the phonemes in her native language. This is a classic problem in nonparametric clustering. The child has no knowledge of either the number or identities of the phonemes (which varies from approximately 10 to 140, out of an enormous set of possible alternatives) nor does she know their sensory input distributions, i.e., the frequencies of her exposure to them. Examples of this kind of distribution-free, nonparametric learning are commonplace in the animal world. Many of the classification problems that juvenile animals and children routinely face present severe challenges for traditional machine learning approaches.

My doctoral dissertation presented a new mathematical framework for solving such learning problems called *cross-modal clustering*. This approach is applicable to a broad range of questions that do not fit neatly into the classic supervised vs. unsupervised partitioning of traditional machine learning. For example, infants are not in a position to understand labeled data, and their auditory inputs are limited, unsegmented, and statistically irregular; nonetheless, by approximately three months of age, they start babbling in their native tongues. Thus, neither supervised nor unsupervised learning techniques appear plausible mechanisms for explaining this phonetic acquisition and related developmental phenomena. How then do children learn so readily?

My approach has been to formalize an insight of Aristotle: *differences in the world are only detectable because different senses perceive the same world events differently*. In other words, redundancy can be used in lieu of explicit supervision to autonomously derive supervisory signals. Building on this, my doctoral thesis presented a framework for learning based upon correlations in different sensory modalities. This draws upon an enormous body of neurological and phenomenological evidence gathered in the past half century demonstrating the extraordinary degree of interaction between sensory modalities during the course of ordinary perception. Although it is tempting to view these complexities as artifacts of biological "implementations," I have provided evidence these interactions are critical to solving some of the difficult learning problems animals face. (This approach is conceptually similar to work known as *biclustering*, but

it applies to a structurally broader set of problems and uses an entirely different mathematical framework.)

My doctoral work demonstrated that a biologically-inspired approach can help answer what are historically difficult computational problems, for example, how to cluster nonparametric data corresponding to an unknown number of categories. This is an important problem in computer science, cognitive science, and neuroscience, and the first half of my thesis provided a solution in many perceptual and sensorimotor domains.

3. Applications

The framework described above has been tested on a number of non-trivial applications.

The first of these is a system that learns the formant structure of American English – i.e., the number of vowels and their phonetic structures – simply by watching and listening to someone speak. It is entirely non-parametric, knowing neither the number of categories (vowels) nor their distributions in advance, and it also has no prior linguistic knowledge. This work is the first example of unsupervised phonetic acquisition of which we are aware, outside of that done by human infants, who solve this problem readily. The results of cross-modally clustering auditory and visual inputs are shown in Figure 2.

I intend to extend this result to cover a complete set of phonemes in English to develop a deeper understanding of protolanguage – the poorly studied prelexical states through which infants pass as they acquire word usage.

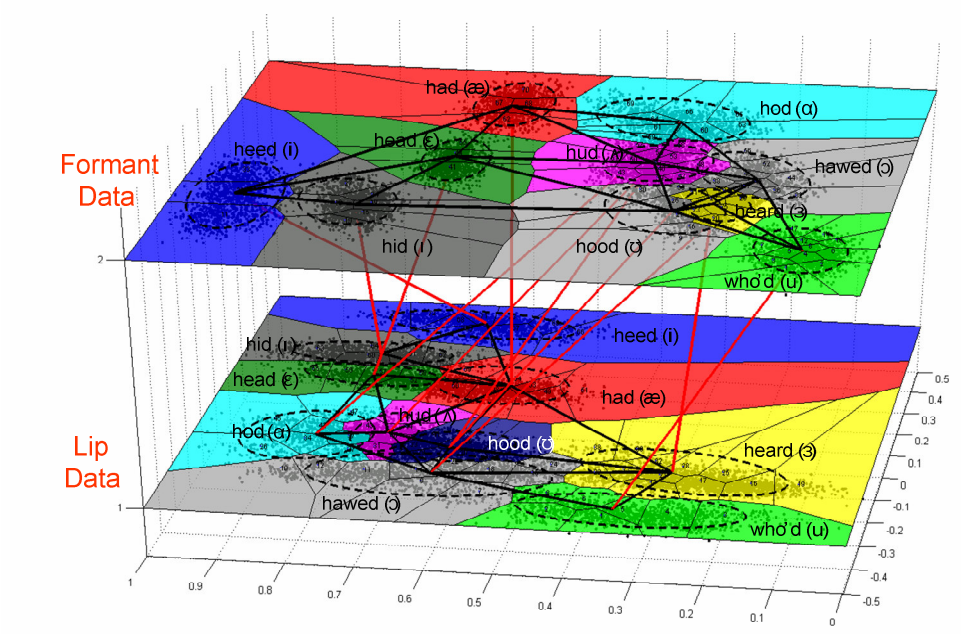

Figure 2 – Acquiring vowels through cross-modal clustering. This figure shows we can learn the number and structure of vowels (i.e., the 10 monopthongs) in American English by simultaneously watching and listening to someone speak. Auditory formant data is displayed on the top place and visual lip data – corresponding to major and minor axes of an ellipse fit on the mouth – are on the bottom. Initially, nothing is known about the events these systems perceive. *Cross-modal clustering* lets them mutually structure their perceptual representations and thereby learn the event categories that generated their sensory inputs. The colors show the region correspondences obtained from cross-modal clustering. Red lines connect corresponding vowels between the two datasets and black lines show neighboring regions within each dataset. The phonetic labels were manually added to show identity. The data are from a real speaker and were normalized.

"Parent"
(Real zebra finch)

"Child"
(Artificial zebra finch)

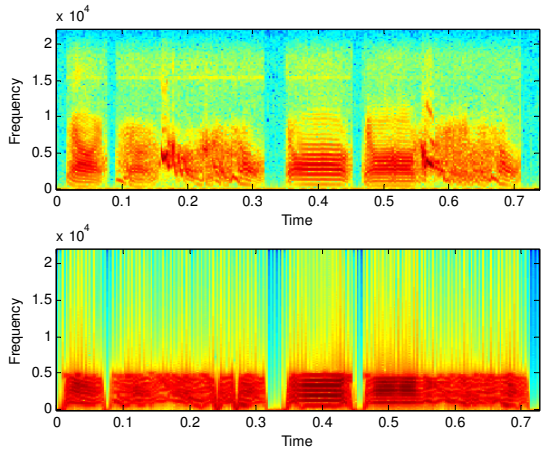

Figure 3 – Birdsong mimicry. On the top is a sample of the zebra finch song used as the "parent" for our system. On the bottom is the system's learned imitation, where the acquired songemes have been fit to the template of the parent's song and smoothed.

To demonstrate sensorimotor learning within this framework, I have also constructed a system that learns to sing like a real zebra finch looking for a mate, following the developmental stages of a fledgling zebra finch. It first learns the song of an adult male corresponding to its "father" and then listens to its own initially nascent attempts at mimicry through an articulatory synthesizer. By recursively reapplying the cross-modal clustering framework described above, the system demonstrates the acquisition of sensorimotor control through what was initially a perceptual framework. Spectrograms displaying the vocalizations of the real bird used to train this system and the resulting learned output of its artificial "son" are shown in Figure 3.

I propose to extend this system for learning vocal mimicry in a range of other species, including dolphins, elephants, and mice, which have all recently been shown to been shown to capable of vocal learning. I am particularly interested in looking for indications of combinatorial, nested syntax in the vocalizations of non-human species, which has long been assumed to be in the exclusive purview of human beings.