In making sense of the world, brains actively reformat noisy and complex incoming sensory data to better serve the organism’s behavioral needs. In vision, retinal input is transformed into rich object-centric scenes; in audition, sound waves are transformed into words and sentences. As a computational neuroscientist with a background in applied mathematics, my basic goal is to reverse-engineer the algorithmic principles underlying these sensory transformations.

I think of the processes that shape any given sensory system as having three basic components: (i) an architecture class from which the system is built, embodying our knowledge about the area’s anatomical and connectivity structure; (ii) a behavioral task that the system must accomplish, representing what we think the sensory area does; and (iii) a learning rule that selects a specific architecture from the class and refines it to best accomplish the behavioral goal. In my work, I fit these three components together to make detailed computational models that yield testable predictions about neural data.

My program begins by applying this approach to model neural responses in ventral visual cortex, and is grounded in recent algorithmic advances in invariant object recognition (§ 1). These advances help lay a foundation for exploring richer visual tasks beyond object categorization (§ 2), modeling other sensory domains such as audition (§ 3), and studying the learning rules active during visual development (§ 4).

1. Predictive Models of Ventral Visual Cortex  Starting with the seminal ideas of Hubel and Wiesel, work in visual systems neuroscience over the past 50 years has shown that the ventral visual stream generates invariant object recognition behavior via a hierarchically-organized series of cortical areas that encode object identity with increasing selectivity and tolerance [2]. My work in visual cortex seeks to go beyond this powerful but broad-stroke understanding to identify concrete predictive models of ventral cortex, and then use these models to gain insight inaccessible without large-scale computational precision.

Mathematically, much of our knowledge about the ventral stream can be boiled down to a class of computational architectures known as Hierarchical Convolutional Neural Networks (HCNNs), a generalization of Hubel and Wiesel’s simple and complex cells [3]. HCNN models are composed of several retinotopic layers combined in series. Each layer is very simple, but together they produce a deep, complex transformation of the input data — in theory, like the transformation produced in the ventral stream. In this formal language, a key step toward understanding would involve identifying a single HCNN model whose internal layers correspond to the known ventral cortical areas and which accurately predict response patterns in those areas. This has proven extremely difficult to do [4, 5], in part because subtle parameter changes (eg. number of layers, local receptive field sizes, &c) can dramatically affect a model’s match to neural data [6]. Broad stroke understanding is not, by itself, enough.

In high-throughput computational experiments evaluating thousands of HCNN models on both task performance and neural-predictivity metrics, we found that architectures that performed better on high-level object recognition tasks also better explained cortical spiking data (Fig. 1a and [1]). Pushing this idea further, we then combined recent advances from machine learning [7] with novel approaches to parameter optimization [8] to discover a hierarchical neural network model architecture that achieved near-human-level performance on challenging object categorization tasks. It turned out that the top layer of this model is highly predictive of single-site neural responses in inferior temporal (IT) cortex (Fig. 1b, top), yielding the first quantitatively accurate model of this area at the top of the ventral hierarchy. Moreover, intermediate layers were highly predictive of neural responses in V4 cortex (Fig. 1b, bottom), and lower model layers of voxel responses in V1 voxel data [1, 9].

In other words: combining two general biological constraints — the behavioral constraint of recognition performance, and the architectural constraint imposed by the HCNN model class — leads to greatly improved models of multiple layers of the visual sensory cascade. Though the models end up being predictive of neural data, they were not explicitly tuned using this data, since model parameters were independently selected to optimize categorization performance on an unrelated image set. Consequently, each layer is a generative model for its corresponding cortical area, from which large numbers of IT-, V4- or V1-like units can be sampled. A common assumption in visual neuroscience is that understanding the qualitative structure of tuning curves in lower cortical areas (e.g. gabor conjunctions in V2 or curvature
Figure 1: Behavioral Optimization Yields Neurally Predictive Models. a. Across thousands of candidate neural network models, performance on categorization tasks strongly correlates with ability to fit neural spiking data [1]. b. The best of these models predicts IT neural responses with its top layers (“HMO model”, red bars, top panel), and V4 neural responses with its intermediate layers (bottom panel).

in V4 [10]) is a necessary precursor to explaining higher visual cortex. Our results show that higher-level constraints can yield quantitative models even when bottom-up primitives have not yet been identified.

Over the next five years, I plan to leverage this computational approach to explore a spectrum of questions focused on tightening the connections between model components and neurons:

- **Detailed characterization of IT substructures.** IT cortex is not a single monolithic computational mass, but instead contains specialized face, place, body, and color-selective regions [11, 12, 13, 14]. Are these the only regions? If so, why these and not others? How do the regions arise in the first place? I plan to extend my models to yield detailed predictions about IT substructure, first to see if they account for known IT heterogeneities, and if so, to search for new regions that can be confirmed or falsified using primate fMRI and electrophysiology experiments. If the known regions do not emerge in the models, I will focus on figuring out what additional principles are required to build them.

- **Understanding tuning in intermediate visual areas.** Intermediate visual areas such as V2 and V4 have proven especially hard to understand because they are removed both from low-level image properties and from higher-level IT-like semantic structures [10]. Because intermediate layers of our models are predictive of these cortical areas, performing high-throughput “virtual electrophysiology” to characterize the model’s internal structure should yield insight into tuning curves in corresponding cortical areas.

- **Guiding causal (in-)activation studies.** Computational models enable us to efficiently perturb unit activations in a highly selective manner. In the short term, I plan to make testable predictions of the behavioral changes (eg. in facial expression identification ability) that arise from inactivating/stimulating specific subpopulations of units defined by cell type or functionality (eg. high face-selectivity). In the 2-5 year term, I plan to collaborate with experimental colleagues developing cutting-edge optical techniques in non-human primates ([15, 16]) to help design highly targeted real-time perturbation studies.

2. Richer Visual Behaviors Object recognition is only a part of visual cognition. Surprisingly, our ventral stream models achieve high performance in estimating a variety of non-categorical visual properties, including object position, size, and pose — even though the models were only optimized for categorization tasks (Fig. 2, top row). In fact, these “identity-orthogonal variables” are increasingly better captured with each successive network layer, even as tolerance to these same variables is also built up (Fig. 2, middle row). This observation makes a counterintuitive neural prediction: IT cortex should contain encodings for a whole spectrum of non-categorical visual properties. In recent experiments, we have found that this
Figure 2: Non-categorical visual properties in IT cortex. a. The same computational model that predicts IT spiking responses [1] also performs well on a variety of “identity-orthogonal” estimation tasks, e.g. position, size, and pose estimation, even though the model was only optimized for categorization performance. x-axis shows model training timecourse, y-axis shows task-dependent performance metric for model top layer features. b. Performance on all tasks also increases through successive model layers. c. Consistent with model results, an IT neural population sample (solid bars) outperforms a similarly-sized V4 population sample (hatched bars), which in turn usually outperforms a V1 population (light-color bars), across a whole battery of categorical and non-categorical visual tasks [17], even on tasks that are usually associated with low-level visual features (e.g. position estimation). y-axis shows performance of linear classifier or regressor trained on neural population data and assessed on held-out images. “Ret.” refers to the pixel-level (“retinal”) control.

prediction is strongly borne out: the actual IT neural population significantly outperforms lower visual areas such as V1 and V4 in estimating all these properties (Fig. 2, bottom row), including those (e.g. position) that are normally considered either low-level ventral or dorsal visual features [17]. These results suggest a reconceptualization of IT cortex as encoding key aspects of scene understanding, rather than just invariant object recognition.

A key future for me direction lies in further extending computational modeling approaches to these richer visual behaviors. Psychophysics has shown that humans understand scenes not only with single category labels or lists of object properties, but in fact represent core 3-D object geometry [18]. I am particularly interested in extending my computational methods to produce biologically-consistent neural network algorithms that extract 3-D geometries from complex scenes. This is an important artificial intelligence question, because natural-image geometry-extraction is at the forefront of what existing computer vision approaches can do. Collaborating with computer-vision experts, I hope to make progress on the algorithm-design aspect of this question over the next 2-3 years, during which time I believe it will likely be necessary to go beyond feedforward computational architectures. If this work is successful, I will be very curious to see whether networks optimized to solve these richer tasks are more predictive of IT neural response data than our best existing feedforward models.

At the same time, understanding exactly where and how 3-D object geometries are represented neurally is an important open question [19, 13]. As a step in this direction, I have helped design stimuli with challenging but controlled 3-D geometric scenes, on which colleagues in Jim DiCarlo’s group have begun recording macaque V4 and IT neural responses. Our first challenge in analyzing this data will be designing more sophisticated neural decoder algorithms to detect if 3-D geometric information is present in our neural sample at all. Over the longer term, I hope to address the question of 3-D geometry representation more deeply through extended collaborations with primate electrophysiologists.
3. Exploring Cortical Generality: Audition  My work in vision suggests a more general hypothesis about how to model sensory cortex: selecting biologically-plausible neural networks for high performance on an ecologically-relevant sensory task will yield a detailed model of the actual cortical areas that underlie that task. Since this idea has some traction in the ventral visual stream, I’m deeply curious to see whether it also yields insight in other sensory domains.

In recent ongoing work with Josh McDermott and colleagues (MIT), I’ve found that HCNNs trained to solve challenging high-variation word recognition tasks are predictive of voxel patterns in auditory cortex (Fig. 3a). These models are also able to differentiate auditory areas, with lower model layers more predictive of inferior colliculus, intermediate layers more predictive of primary auditory cortex, and higher layers of speech and music-selective areas identified in recent imaging studies (Fig. 3b).

These results open up a variety of computational audition questions that I plan to study over the next five years, including:

- **Detailed characterization of non-primary auditory cortex.** As with higher visual areas, I plan to use models to make detailed testable predictions about poorly-understood auditory cortex subregions, especially in relation to speech and natural sound representations [21, 22]. In the short term, I plan to obtain neural data via ongoing human fMRI experiments, and in the 2-3 year term via collaboration with primate electrophysiologists.

- **What auditory tasks best explain cortical differentiation?** Given the evolutionary history of audition, non-speech tasks (eg. environmental sound differentiation) might be as important for driving auditory cortex structure as speech. I plan to explore this question by training networks on a variety of ecological auditory tasks.

- **How do audition-optimized architectures compare to those optimized for vision?** Are there deep but hidden structural similarities between visual and auditory cortex that arise from underlying similarities in auditory and visual data? I plan to attack this fascinating question both from a purely algorithmic point of view, as well as by comparing auditory neural data to ventral-stream data.
models — which, at heart, are really learning rules in action — can help us think about these key questions in a new way (Fig. 4a).

Over the next few years, I plan to work on developing improved semi- or un-supervised neural network learning algorithms that blend features of existing machine-learning techniques with inferences from neural data. Given the results on non-categorical property estimation described in § 2, an intriguing possibility is that heavily supervised category-label training could be replaced by optimization for properties (e.g. position, size, and pose) that can be more easily estimated from motion heuristics in natural video.

In ongoing work with Nancy Kanwisher and colleagues (MIT), I am also using ventral stream models to characterize how much and what kind of face-related experience is required during training to generate known face-specificity in IT cortex. An intriguing initial result (Fig. 4b) is that even without any training on faces (or any animate objects), models nonetheless exhibit a significant fraction of strongly face-selective units at top layers. Moreover, given a very small amount of additional training with (512 images including 64 faces), the number of face-selective units quickly increases to the experimentally-observed incidence. This is, in effect, a prediction about a controlled-rearing study: face-deprived animals should still exhibit an electrophysiologically-observable population of face-selective units, and this population will quickly grow after deprivation ends. Counterintuitively, this result suggests that much of the machinery needed to support face selectivity is present even without a large amount of face experience (see [30]).

In the immediate term, I plan to perform systematic in silico deprivation studies to help understand the role of differentiated visual experience in shaping visual features at all levels of the ventral hierarchy. Over the somewhat longer term, I hope to build a collaboration to design and execute such studies in animal subjects. As a first step toward this goal, I have started a new collaboration with Jim DiCarlo (MIT) and Marge Livingstone (Harvard), in which we plan to make large-scale multi-array electrode recordings in 6-9 month old macaques. By comparing to comparable data we have already obtained in adult macaques [1], we will be able to detect subtle differences between juvenile and adult visual representations. Comparing the predictions of the model learning rule with this experimental data will help expose key mismatches between our core machine learning technology and real visual learning.
References