Invariant visual object recognition and the underlying neural representations are fundamental to higher-level human cognition. To understand these neural underpinnings, we combine human and monkey psychophysics, large-scale neurophysiology, neural perturbation methods, and computational modeling to construct falsifiable, predictive models that aim to fully account for the neural encoding and decoding processes that underlie visual object recognition. A predictive encoding model must minimally describe the transformation of the retinal image to population patterns of neuronal activity along the entire cortical ventral stream of visual processing and must accurately predict the responses to any retinal image. A predictive decoding model must minimally describe the transformation from those population patterns of neuronal activity to observed object recognition behavior (i.e., subject reports), and, given that population pattern of activity, it must accurately predict behavior for any object recognition task. To date, we have focused on core object recognition—a remarkable behavior that is accomplished with image viewing durations of $\sim 200$ msec. Our work thus far reveals that the neural encoding process is reasonably well explained by a largely feed-forward, highly complex, multistaged nonlinear neural network—the current best neuronal simulation models predict approximately one-half of the relevant neuronal response variance across the highest levels of the ventral stream (areas V4 and IT). Remarkably, however, the decoding process from IT to behavior for all object recognition tasks tested thus far is very accurately predicted by simple direct linear conversion of the inferior temporal neuronal population state to behavioral choice. We have recently examined the behavioral consequences of direct suppression of IT neural activity using pharmacological and optogenetic methods and find them to be well-explained by the same linear decoding model.

Humans rapidly and accurately process visual scenes from their environment, an ability that is critical for everyday functioning. The ease with which this task is accomplished belies the fact that it is an enormous computational feat: The retina presents visual cortex with a barrage of incoming photoreceptor activation patterns (images) from which higher-level structures (e.g., objects) must be extracted, even in the face of dramatic image variation.

The core of the problem is that the information that is explicit in the image (i.e., the luminance of every pixel) is misaligned with the needs of behavior. Relevant behavioral distinctions (e.g., telling the face of a friend from that of a foe) are swamped by low-level variations in pixel-space (e.g., the differences between the same friend’s face under variable viewpoints). Thus, the challenge that the brain faces is producing a transform from pixels to a high-level representation that solves this misalignment (DiCarlo et al. 2012).

Extensive research in visual systems neuroscience suggests that this transform is affected by the ventral visual stream, a series of connected cortical areas—V1, V2, V4—situated in the occipital-temporal lobe of humans (Farah 1990; Malach et al. 1995), with homologous areas in nonhuman primates (Desimone et al. 1984; Tanaka 1996; Bell et al. 2009), that culminates in an area called inferior temporal cortex (IT). Somehow, the circuits of each of these areas have constructed a mapping from pixel space to a neuronal representation that supports complex visual behaviors. However, only a decade ago, these mechanisms were mysterious.

In our work on this problem, we find it useful to separate the problem into two basic components (Fig. 1): (1) understanding the transform from pixel stimuli to the neuronal representation in higher visual cortex—that is, the “Image $\rightarrow$ Neural population state” link—and (2) understanding the mapping from neuronal representation to visual behaviors—that is, the “Neural population state $\rightarrow$ Behavioral report” link. Here, what we mean by “understanding,” is quite specific: namely, we seek falsifiable, predictive models of each of those links. Thus, we will consider our effort successful when it has produced two specific quantitative models: (1) a model whose inputs are arbitrary images, and whose outputs accurately describe the responses of each neuron in higher visual cortex to these images, and (2) a model whose inputs are neuronal response patterns, and whose outputs accurately predict behavioral choices across the whole variety of visual tasks supported by the ventral stream. These predictive models should not only be black boxes, but should be mapped to the underlying neural hardware so that they can be explicitly tested (e.g., potentially falsified) and leveraged for future applications such as improved computer vision and visual prosthetics.

In this paper, we discuss our approach and progress to building these two types of models, in the context of what
we call core object recognition behavior. By core object recognition we mean the subdomain of object recognition behavior involving categorization of images containing a single salient object presented briefly (100–200 msec) in the central visual field (within the central 10° of the visual field). In “From neuronal populations activity patterns to behavior” section, we begin by first discussing our work on the neurons-to-behavior (Neural population state → Behavioral report), because it is only by defining this mapping precisely that we will be able to understand which populations of ventral stream neurons we are most interested in modeling, and which properties of these neurons we most care to explain. In other words, a better understanding of the “Neural population state → Behavioral report” link will help us identify those neural events that directly drive behavior, and focus our subsequent modeling efforts on these key events. This section is divided into two subsections: “correlational models” ("Correlative methods to establish a working decoding model" section) and “causal models” ("Causal methods to test our working decoding model" section). In “Correlative methods to establish a working decoding model” section, we bring evidence from monkey electrophysiology and human psychophysics to establish a detailed link between neuronal firing patterns in the inferior temporal cortex and patterns of behavior on challenging object recognition tasks. The work is correliative, but strongly so, in that it uncovers a nearly perfect relationship between neuronal patterns and behavioral patterns over all tested recognition tasks, and falsifies other possible relationships. Although this approach allows us to infer the form of the neural codes and the underlying mechanisms, it does not directly show that the inferred neural code is the cause of the behavior. Thus, in “Causal methods to test our working decoding model” section, we discuss recent experimental techniques for perturbing brain activity in vivo and present evidence generated from these techniques that directly shows the causal link between the neuronal activity in the inferior temporal cortex and object recognition behavior and lays the foundation for future work along this line.

In “From images to neuronal population activity patterns” section, we discuss our work on modeling the encoding of visual images in higher visual cortical areas. The upshot of that work is that we have recently identified a hierarchical neural networks model whose top output layer is highly predictive of IT spiking responses to complex naturalistic images at both the single site and population levels. The model’s intermediate layers are highly predictive of neural responses in the V4 cortex, a midlevel ventral stream visual area that provides the dominant cortical input to IT. Moreover, lower layers are highly predictive of voxel responses from fMRI data in V1 and V2. In addition to substantially advancing our encoding goal (Fig. 1), these results show that our performance-optimization approach—applied in a biologically appropriate model class—can be used to build quantitative predictive models of neural processing.

FROM NEURONAL POPULATION ACTIVITY PATTERNS TO BEHAVIOR

The ventral cortical stream of primates includes a series of hierarchical processing steps that encode image information content (e.g., object category and identity) increasingly explicitly along the ventral stream cortical areas (Gattass et al. 1990; Ungerleider and Haxby 1994; Miyashita and Hayashi 2000; Reddy and Kanwisher 2006). For instance, neurons in the primary visual cortex can be described by Gabor-like filters that extract object contour information (Carandini et al. 2005), although these neurons do not express strong tolerance to geometrical image transformations such as translation and rotation (DiCarlo et al. 2012). In contrast, population patterns of activity in inferior temporal (IT) cortex (highest processing stage of the ventral stream) can directly support, object categorization over a wide range of complex image transformations (Hung et al. 2005; Rust and DiCarlo 2010). Consistently, midlevel ventral stream areas—for example, V4, the main cortical input to IT—express intermediate levels of selectivity for objects under complex image variations (Connor et al. 2007; Rust and DiCarlo 2010). Thus, it is reasonable to hypothesize simple decoding algorithms that drive visual object recognition behavior directly from the neural responses of IT cortex. Although that is a long-standing working paradigm in the field, defining a compact, potentially mechanistic model that explains all object recognition behavior has not yet been accomplished. In what follows, we describe our efforts to accomplish that goal using both strong correlation methods and causal methods.

Correlative Methods to Establish a Working Decoding Model

It has been long known that individual neural responses in higher processing stages of the ventral stream (e.g., IT cortex) encode information about the identity and category of visually presented objects (Tanaka 1996). Such
results motivate a potential link between neuronal responses in the ventral stream and behavioral shape judgments; if ventral stream neurons encode shape of the visual stimuli, they may well be the neural underpinning of behavioral shape judgments. However, this is only a qualitative link, it does not show that ventral stream neural responses support object recognition behavior at high levels of human psychophysical performance, nor does it explain how that behavior can be predicted for any and all object recognition tasks. To advance understanding, the hypothesis that “object recognition behavior is driven by the neural activity in the ventral stream” must be translated into a potentially mechanistic, predictive linking model, and that model must, if correct, consistently predict the relationship between a specific aspect of ventral stream neuronal population activity and behavior over all object recognition tasks.

Although it was not a priori obvious that such a linking model existed, we made it our task to try to find one. That is, we sought a single linking model (a.k.a. decoding model; see Fig. 1) that accurately predicts behavioral difficulty for a wide range of object recognition tasks based on IT neural population activity evoked by each visual image. The first step to answer this question is to operationally and quantitatively define a set of “object recognition” tasks. This is a particularly important step because the choice of psychophysical tasks can potentially change the outcome. We reasoned that the psychophysical tasks that operationally define “object recognition” should be hard enough to push human and monkey observers to the limits of errors, because without such errors it is not possible to evaluate the consistency of the failure patterns of the human observers with that of the ventral stream neurons. More importantly, because we and others assume that the ventral stream is critically important for “invariant object recognition,” we wanted the psychophysical tasks to engage that problem. For instance, although the behavioral distinction between a white dove and a black bear may be called “object recognition” in common language, such a behavior can be easily supported by color or simple low-level cues—it does not require tolerance to position, scale, viewpoint, or background clutter and thus does not represent the challenging part of object recognition.

To satisfy these constraints, we operationally defined object recognition by measuring human psychophysical performance in 64 object recognition tests that included thousands of images that explore shape similarity under a range of identity-preserving image variations (Fig. 2A) such as translation, rotation, and variation of the background. Not surprisingly, this showed that human object recognition performance measurements span a wide range— that is, some object recognition tests are more difficult than others (Fig. 2B), and this pattern of difficulty is highly consistent across subjects. Using the same set of visual images and viewing durations, we then measured neuronal responses in V4 and IT of two macaque monkeys in a passive viewing paradigm. We were aware of the numerous potential differences between active psychophysical task performance in human observers and passive, visually evoked neural responses in monkeys who do not have learned behaviors with respect to the stimulus set (e.g., images of cars). However, our hypothesis was that the ventral stream transforms each visual image into a new neuronal population pattern at high levels of the ventral stream that can simply predict the

Figure 2. A unified neuronal population code can explain human object recognition behavior. (A) We started with 3D models of real world objects (e.g., faces, cars) with multiple exemplars for each category. Given parameters (size, position, pose), we used ray tracing to render the object and place it on an uncorrelated background producing a test image. (B) We measured human performance for one eight-way basic-level object categorization (face, cars, planes, animals, tables, chairs, boats, fruits), and two subordinate eight-way identifications (cars: Beetle, BMW z3, Ford, etc.; faces: face1, face2, face3, etc.) and determined d's for 384 tasks (24 labels × 6 levels of identity-preserving view variation). Heat map shows range of performance (high d' = red). (C) We find that a single linking hypothesis (aka “unified code” or “decoding model”) is sufficient to explain human core object recognition behavior, as showed by its statistically perfect consistency with the pattern of human performance. Only codes marked with asterisks are sufficient. (Adapted from Majaj NJ, Hong H, Solomon EA, DiCarlo JJ, unpubl.)
animal’s potential object recognition performance over any object recognition test. More specifically, based on previous work (Miyashita 1993; Freedman et al. 2003) we assumed that that underlying mechanism of learning any new object recognition task (e.g., detecting cars) is implemented by downstream neurons (e.g., prefrontal cortex, perirhinal cortex) modifying the strength of their synaptic inputs received from ventral stream neurons. We tested this linking hypothesis by implementing simple linear classifiers that approximate those downstream learning mechanisms and then explored which aspects of ventral stream neural activity, if any, could predict the pattern of behavioral performance over all object recognition tests.

We refer to those aspects of neural activity as “neural codes,” and based on the neural recordings from V4 and IT cortex, we implemented dozens of putative neural codes (results from 14 putative neural codes are shown in Fig. 2). We then compared each code’s predicted behavioral performance (neural predicted d’s) with the actual (human) pattern of behavioral performance (psychophysical discriminability d) for all of the 64 object recognition tests. The correct neural code and linking hypothesis was expected to predict the entire pattern of human performance across all 64 tests. Remarkably, we found a few neural codes that meet this high bar (Fig. 2C). These few neural codes were all based on simple learned linear sums of mean firing rates over the IT population. Many other IT codes and all V4 neural codes and all simulated neural codes from a V1-like model failed to accurately predict behavior.

These results show that human psychophysical performance on challenging object recognition tasks can be fully predicted by a simple, learned linear sums of IT neural populations firing rates, and strongly suggest that this working model will generalize to a large number (perhaps all) of object recognition tasks. Because this work specifies the parameters of a specific proposed mechanisms for linking neuronal activity to any object recognition tasks, it allows us to make quantitative predictions of how perturbation of IT neural activity population patterns will lead to specific changes in behavioral performance over all object recognition tasks. Testing such predictions requires the validation and application of new tools, which is described in the next section.

Causal Methods to Test Our Working Decoding Model

A visual stimulus such as the image of a face activates a cascade of population neural responses in the visual system at different temporal and spatial scales. At least some parts of this stimulus-driven activity must causally support object recognition behavior, whereas other parts might be epiphenomenal in that they play no role in a particular object recognition task. Indeed, this might even be true for neural activity which has high information for that recognition task and high predictive power for the behavioral pattern of performance (e.g., the working IT-based decoding model outlined above). One must thus use tools that directly interact with neural activity to distinguish among these alternatives. Once the causally relevant neural events are identified, the next critical step is to make quantitative links between them and behavior. For example, imagine a single neuron in IT cortex that can discriminate between male and female faces with 90% accuracy (encoding). Does this neuron play any role in gender discrimination behavior? That is, would inactivation of this neuron take a toll on gender discrimination behavior? If so, how much behavioral deficit would result from inactivation of this neuron? 10%? 0.1%? 0.01%? And which temporal aspects of the neural activity are most critical to the behavior? Would the inactivation of this neuron also affect other object recognition tasks? If so, by how much? Answering such questions is only possible by directly perturbing the neuronal activity and measuring its effect on behavior.

The potential causal link between neuronal activity and behavior has been studied by multiple research programs for different brain regions. This includes frontal eye fields (Bruce and Goldberg 1985), somatosensory cortex (Romo and Salinas 2001), superior colliculi (Stryker and Schiller 1975), primary visual cortex (Schiller and Tehovnik 2008), and middle temporal cortex (Salzman et al. 1990). However, for the case of IT cortex there are only two neurophysiological studies of the causal relation between the IT neural response properties and object recognition; a study of the effect of microstimulation of IT face-selective cells on face detection behavior (Afraz et al. 2006) and a recent study about the causal role of 3D-structure-selective IT cells in three-dimensional structure categorization (Verhoef et al. 2012). In our work, we seek a comprehensive understanding of the causal role of the IT neural activity in all object recognition tasks. At the current stage of the work, we have selected a specific task—psychophysically demanding face gender discrimination task—to explore the feasibility of this idea.

We trained two macaque monkeys to report the gender of a briefly presented face by looking at one of the two response targets on the screen. Then in separate experiments, we used optogenetic and pharmacological methods to artificially suppress the activity of various subregions of IT cortex while the animals performed the face gender discrimination task. To determine the neural selectivity profile of the targeted IT subregions, we recorded extensively from the IT cortex while the animals simply fixated visually presented images including faces and other objects. We found that (Afraz et al. 2015) both pharmacological and optogenetic inactivation of IT subregions that are traditionally defined as face selective (neural d’ for discriminating faces from other objects >1) induced a deficit in face discrimination. The deficit was limited only to the contralateral visual hemifield and it was not observed for other IT subregions where the neural responses were not face selective. Specifically, temporally delimited (200 msec), optogenetic inactivation of face-selective subregions led to a small but significant drop in gender discrimination
performance (median = 1.8%, mean = 2.02%, $t_{(16)} = 5.99$, $P < 0.0001$) for faces presented to the contralateral visual field. Similarly, pharmacological inactivation of the same cortical subregions (induced by microinjection of muscimol) lead to a 5.5% drop in gender discrimination performance only for contralaterally presented stimuli, whereas performance remained unchanged in the ipsilateral hemifield (repeated-measures ANOVA interaction test, $F_{(4,20)} = 7.5$, $P < 0.001$). To begin to connect these results to our working decoding model derived from correlational methods (“Correlative methods to establish a working decoding model” section) we sought to use that model to predict the causal importance of each subregion of IT to the gender discrimination task. As a first approximation, we computed the predicted causal importance in each IT subregion as the facial gender task performance of a linear classifier that could read only the neural population activity in that subregion (“local linear classification”), based on our recording measurements of about nine (median) sites taken within the subregion. Consistent with our working model, we found that this local linear classification measure was highly predictive of the impact of direct optical inactivation of that subregion on behavioral gender discrimination performance ($r = -0.6$, $P = 0.01$) (Fig. 3). Linear regression suggests that inactivation of a small IT subregion ($\sim$1 mm in diameter) with our highest observed local linear classification for facial gender induces an impact of (2.5%) on psychophysical performance (see Fig. 3).

In summary, these results establish a causal link between the neural activity in IT cortex and at least one type of object discrimination behavior (face gender discrimination). In a hierarchical processing system, it is possible to imagine a situation where inactivation of an upstream area that does not show linear separability for a given feature influences behavioral discrimination of that feature. The correlation between the behavioral impact of inactivation and linear separability (Fig. 3) will, we think, not be found in such an area. That is, we predict that such correlation will be much lower for inactivation of lower-level areas such as V1 or even V4. The fact that we observe a correlation between local neural linear separability of the visual information in each IT cortical subregions and the behavioral impact of inactivation of those subregions suggests that IT representations are relatively close to the read-out mechanisms that drive the behavior, consistent with the working model outlined above. Our next goal is to extend these causal methods to a battery of object recognition tests and to build stronger quantitative links between neural activity and behavior, guided by our working decoding model.

**FROM IMAGES TO NEURONAL POPULATION ACTIVITY PATTERNS**

Starting with the seminal ideas of Hubel and Wiesel, work in visual systems neuroscience over the past 50 years has strongly suggested 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 (DiCarlo et al. 2012). Our work in visual cortex seeks to go beyond this powerful but broad-stroke understanding to identify concrete predictive models of the underlying neuronal population representation, 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 (Serre et al. 2007; Pinto et al. 2009). 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 along the entire depth of 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 (*neural predictivity*). This has proven extremely difficult to do (DiCarlo and Cox 2007; Pinto et al. 2008), in part because subtle parameter changes (e.g., number of layers, local receptive field sizes) can dramatically affect a model’s match to neural data (Yamins et al. 2013). 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 architectural parameters that performed better on high-level object recognition tasks also tended to have higher *neural predictivity*—that is, they better explained cortical spiking data (Fig. 4A; Yamins et al. 2014). Model param-
eters were selected for evaluation by one of three rules: (1) random draws from the uniform distribution (Fig. 4A, green dots); (2) optimization for performance (blue dots); and (3) optimization directly for architectures that best fit IT neural data (orange dots). All three conditions show large variation in both performance and IT neural predictivity across the parameter space, demonstrating that choices for these architectural parameters have a large effect on a specific model’s ability either to perform object recognition effectively or match neural data. Moreover, performance was significantly correlated with neural predictivity in all three selection regimes, with models that performed better on the categorization task also being more likely to produce outputs more closely aligned to IT neural responses. Critically, optimizing for only task performance significantly increased the correlation with IT neural predictivity even though neural data were not used in the optimization. Moreover, the best performing models identified in experiment (2) predicted neural output equally well as the most predictive models identified in experiment (3). Together, these results imply that performance optimization is an efficient means to identify the parameters of HCNN models that best predict the neural population response data—that is, to identify specific encoding models that respond most like the actual ventral visual stream.

Pushing this idea further, we then used Hierarchical Modular Optimization (HMO) to discover a hierarchical neural network model architecture that achieved near-human-level performance on challenging object categorization tasks. The HMO procedure embodies a conceptually simple hypothesis for how high-performing combinations of functionally specialized hierarchical architectures can be efficiently discovered and hierarchically combined. Applying the HMO procedure, we identified a high-performing four-layer HCNN model with 1250 top-level outputs. This model matched the object recognition performance of the IT neural sample, even when...
faced with large amounts of variation—a hallmark of human object recognition ability (DiCarlo et al. 2012). As a control, we also evaluated a variety of existing models targeting several levels of the ventral hierarchy, including SIFT, a simple baseline computer vision model (Lowe 2004), a V1-like Gabor-based model (Pinto et al. 2008), a V2-like conjunction-of-Gabors model (Freeman and Simoncelli 2011), and HMAX (Serre et al. 2007), a model targeted at explaining higher ventral cortex and that has receptive field sizes similar to those observed in IT. These control models approach IT and human performance levels in low-variation circumstances, but in high-variation conditions all of them except for the HMO model fail to match the performance of human subjects and the IT neural population by a large margin.

Given the HMO model’s plausible performance characteristics, we then measured its ability to predict IT neural data, both for the top-level output, as well as for each of the model’s three intermediate layers (Figs. 4C, 5, red lines/bars). We found that each successive layer predicted the responses of IT units increasingly well, with category selectivity and tolerance to more drastic image transformations emerging gradually through the multiple layers of the model hierarchy (Fig. 5A). Critically, we found that the top layer of the high-performing HMO model achieves remarkably high IT neural predictivity, predicting $48.5 \pm 1.3\%$ of the explainable IT neuronal variance (single site predictivity, median of 168 IT sites; Fig. 5B, top). This represents a nearly 100% improvement over the previous best comparison models. Ideal observer models, with perfect access to all category labels and other image parameters, predict IT units above chance level, but significantly less well than the HMO model, showing that high IT neural productivity does not automatically follow from category selectivity. Put another way, IT neurons are not simple categorizers—there is significant noncategorical structure in IT responses that is likely attributable to hierarchical network architecture of the actual ventral visual stream—but the structure in those responses is quite well captured by the HMO network architecture.

Cortical area V4 is the dominant cortical input to IT, and the neural population representation in V4 is known to be even less categorical than that of IT (Rust and DiCarlo 2010). We found that the HMO model’s intermediate layers are highly predictive of V4 neural responses ($51.7 \pm 2.3\%$ explained V4 variance), providing a significantly better match to V4 than either the model’s top or bottom layers (Fig. 5B, bottom). These results suggest that V4 corresponds to an intermediate layer in a hierarchical model whose top layer is an effective model of IT. Unlike the case of IT, ideal observer semantic models explain effectively no response variance in V4, consistent with V4’s lack of explicit category selectivity in single neurons. Together these results suggest that performance optimization not only drives top-level output model layers to resemble IT, it also imposes strong biologically consistent constraints that predict the “internal” structure of the ventral stream as well.

In summary, 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. Although the models end up being predictive of neural data, they were not explicitly tuned using these data, because 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- or V4-like units can be sampled. A common assumption in visual neuroscience is that understanding the
yield quantitative models even when bottom-up primitives have not yet been identified.

CONCLUSION

In our work, we have built two basic types of models. One of these models (encoding) maps the retinal input to reasonably accurate predictions about neural population responses in V4 and IT cortex. The other one (decoding) maps IT neural population activity to very accurate predictions about behavioral object choice. The encoding model is a hierarchical neuronal network that computes a complex nonlinear transform of the retinal input. This nonlinearity takes the information that is implicit in these inputs (e.g., the category of an object independent of where and how it appears in the visual field) and renders it explicit and easily accessible to downstream decode mechanisms. Our decoding model is a candidate for exactly this type of downstream read-out mechanism, a simple linear operation that projects a desired behavioral output (e.g., category identity) out of the IT neural population code. Combining these two models provide a working, quantitative end-to-end mechanistic model that begins to capture the neural mechanisms underlying core visual object recognition.

These models are largely feed-forward, suggesting that the core object recognition subdomain of visual behavior (rapid categorization of single centrally presented objects) is mainly supported by a single unidirectional cascade of neural activity along the ventral visual hierarchy. These results also suggest that encoding for core visual object recognition (i.e., the generation of IT population patterns) may be a largely automatic process that is constantly in operation, independent of behavioral state.

Our work is circumscribed by a number of key limitations within the subdomain of core object recognition. Both the encoding and decoding models are imperfect in that they are capturing a significant fraction but not all of the variance in each of these processes. We have explored only a subset of possible behavioral tasks, both in correlational and perturbative studies, leaving open the possibilities that there are images and recognition tasks on which these models would fail. We have not explained the origin of cortical heterogeneity (e.g., spatial clustering of face-selective neurons in multiple interconnected patches) or the functional significance of having multiple cortical layers in each visual area. Although we have provided a quantitative predictive model, we have not yet provided a qualitative characterization of neural tuning functions in intermediate visual areas (e.g., V2 and V4).

Beyond the core object recognition subdomain on which we have established this new foothold, a number of significant advances must be made before neural mechanisms of object recognition can be completely understood. One of the largest remaining gaps is in understanding the role of feedback circuits in complex visual behaviors. These behaviors will likely involve visual memory, visual learning, top-down inference and sensorimotor integration over extended time periods. We believe that a tight loop between quantitative models and large-scale neural and behavioral data collection is the only way to reach that deeper understanding, and that the work described above is an example of the power of that approach.

REFERENCES

Pinto N, Doukhan D, DiCarlo JJ, Cox DD. 2009. A high-throughput screening approach to discovering good forms of


Rust NC, Dicarlo JJ. 2010. Selectivity and tolerance (“invariance”) both increase as visual information propagates from cortical area V4 to IT. *J Neurosci* 30: 12978–12995.


Neural Mechanisms Underlying Visual Object Recognition

Arash Afraz, Daniel L.K. Yamins and James J. DiCarlo

Access the most recent version at doi:10.1101/sqb.2014.79.024729

References
This article cites 30 articles, 11 of which can be accessed free at:
http://symposium.cshlp.org/content/79/99.full.html#ref-list-1

Email alerting service
Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here

To subscribe to Cold Spring Harbor Symposia on Quantitative Biology go to:
http://symposium.cshlp.org/subscriptions

Copyright © 2014 Cold Spring Harbor Laboratory Press; all rights reserved