CHAPTER SEVEN

What do neurons really want? The role of semantics in cortical representations

Gabriel Kreiman*
Department of Psychology, Children's Hospital, Harvard Medical School, Boston, MA, United States
*Corresponding author: e-mail address: gabriel.kreiman@tch.harvard.edu

Contents

Abstract

What visual inputs best trigger activity for a given neuron in cortex and what type of semantic information may guide those neuronal responses? We revisit the methodologies used so far to design visual experiments, and what those methodologies have taught us about neural coding in visual cortex. Despite heroic and seminal work in ventral visual cortex, we still do not know what types of visual features are optimal for cortical neurons. We briefly review state-of-the-art standard models of visual recognition and argue that such models should constitute the null hypothesis for any measurement that purports to ascribe semantic meaning to neuronal responses. While it remains unclear when, where, and how abstract semantic information is incorporated in visual neurophysiology, there exists clear evidence of top-down modulation in the form of attention, task-modulation and expectations. Such top-down signals open the doors to some of the most exciting questions today toward elucidating how abstract knowledge can be incorporated into our models of visual processing.

Psychology of Learning and Motivation, Volume 70
ISSN 0079-7421
https://doi.org/10.1016/bs.plm.2019.03.005

In this Chapter, I aim to highlight critical lacuna in our understanding of the tuning properties of visual neurons, especially the role of high-level knowledge in neural coding of visual inputs. First of all, I should clarify the obvious. Neurons do not "want" anything. A neuron emits an action potential when its intracellular voltage exceeds a certain threshold, typically but not exclusively, in the axon hillock (Koch, 1999). This voltage is a weighted sum of the influences received through the neuron's thousands of dendritic inputs, which include bottom-up, horizontal, and top-down connections. It remains experimentally challenging to trace all incoming signals to a given cortical neuron and to propagate those signals all the way back to the sensory inputs, not to mention all other non-sensory inputs. Thus, in the vast majority of cases, we correlate the activity of a cortical neuron with the presentation of sensory signals. It is in this sense that the question in the title should be understood. I ask *what sensory inputs best trigger activity for a given cortical neuron* and what type of semantic information may guide or modulate those neuronal responses.

I start with a succinct description of the classical view on what types of visual stimuli trigger activity in neurons along the ventral visual cortex. I introduce state-of-the-art standard computational models of vision and consider them as a basic null hypothesis to evaluate neuronal tuning properties and potential semantic influences, particularly in the context of visual categorization tasks. Next, I provide a few examples of how top-down signals can modulate responses along the ventral visual cortex while emphasizing that we have a long way to go to understand the role of common knowledge on visual processing. I conclude by discussing critical Hilbert question in the field and a brief glimpse of the ample opportunities and challenges ahead.

1. Assumptions and definitions

I focus here on triggering activity in the sense of firing rates defined as spike counts in short windows spanning tens of milliseconds. This is by no means the only or agreed upon relevant property of cortical neurons, there has been extensive discussion about neural codes (see for example, Kreiman, 2004). A neuron might contribute to representing information by firing only a few spikes at a precise time in concert with other spikes in the network. A neuron may also represent information by *not* firing, in the same way that Sherlock Holmes intuited who the murderer was by attending to the dog that did not bark. Additionally, in non-invasive studies, there are multiple

experimental techniques that measure non-neuronal signals that are less well understood and which are difficult to interpret directly in terms of neuronal firing rates. The general flavor of the discussion here could well be extended to other neural codes, but in the interest of simplicity we understand the question in the title to indicate what type of sensory inputs lead to high firing rates for a given cortical neuron.

A few assumptions and disclaimers are pertinent before proceeding. In order to investigate what neurons want, I restrict the question to vision and solicit inspiration from biologically plausible computational models of vision. The focus on vision is merely a practical one: (i) We know more about the architecture of the visual system than other systems; (ii) We can stand on the shoulders of giants that have paved the way through more than a century of behavioral studies of vision and over half a century of neurophysiological scrutiny of visual cortex; (iii) We have an arsenal of tools to synthesize visual stimuli, to precisely control the timing of presentation, to measure eye movements, and to capitalize on millions of digital images and videos. The emphasis on biologically plausible computational models of vision reflects the need to formalize our assumptions, and to generate a common language that can be used to directly test our ideas across labs and across experiments. Verbal descriptions such as "neurons in V2 respond preferentially to angles" or "neurons in IT respond preferentially to objects" are not sufficient and vague, lack predictive power, are hard to reject or validate, and often get us into trouble. We need mathematical models that are instantiated into computer code where we can use exactly the same conditions and exactly the same images as in behavioral or neurophysiological experiments.

Beyond the pattern of inputs propagated from the retina onto visual cortex, what a neuron wants is likely to be modulated by semantic prior knowledge about the world, probably conveyed through top-down connections. What exactly do we mean by semantics? The Oxford's English Dictionary defines semantics as "… the branch of linguistics and logic concerned with meaning." How this definition applies to interpreting the responses of neurons along ventral visual cortex is not clear. We attempt to provide a more quantitative definition later on in this chapter. For the moment, as an example of semantics in the context of visual recognition, we understand pictures of grapes, oranges, or pineapples to represent different types of fruits, even though they are rather different in terms of their visual features. Similarly, we refer to ants, elephants or goldfish as animals. We will ask what roles this and other types of high-level knowledge about the world plays in visual processing.

2. Neuronal responses in visual cortex, the classical view

The introduction of techniques to record the activity of neurons in the beginning of the last century led to decades of experiments interrogating neuronal responses to visual stimulation. The history of studying neuronal tuning properties in visual cortex is the history of visual stimuli. How do we investigate the feature preferences of a neuron in visual cortex? We need to decide which stimuli to use in the experiments. The central challenge in answering this question is that it is essentially impossible with current (and foreseeable) technology to exhaustively explore the entire space of images: the number of possible images is beyond astronomical. Considering a small image patch of 100×100 pixels, there are $2^{10,000}$–$10^{3,010}$ possible binary images, $\sim 10^{24,082}$ grayscale images with 256 shades of gray per pixel, and $\sim 10^{72,247}$ 8-bit color images. As a consequence, investigators have traditionally used several astute and reasonable strategies to select visual stimuli for experiments:

(i) *Stimuli from previous studies*. Past performance is a strong predictor of current performance for neurons. Stimuli that have excited neurons in previous studies are often a good initial guess to design experiments. For example, ever since the discovery that V1 neurons in cats and monkeys respond vigorously to bars or gratings of specific orientations (Hubel & Wiesel, 1968), investigators have used oriented bars and gratings to probe the responses essentially in every species and every visual area (Chapman, Stryker, & Bonhoeffer, 1996; Coogan & Burkhalter, 1993; Ghose & Maunsell, 2008; Hegde & Van Essen, 2007; Nassi, Gomez-Laberge, Kreiman, & Born, 2014; Niell & Stryker, 2010). A variation of this approach is to start with effective stimuli from previous studies and evaluate neuronal responses to modified versions of those stimuli (Kobatake & Tanaka, 1994; Leopold, Bondar, & Giese, 2006; Tanaka, 2003; Tsao, Freiwald, Tootell, & Livingstone, 2006).

(ii) *Natural stimuli*. It seems reasonable to assume that neurons represent behaviorally relevant stimuli and the types of images encountered in the real world. Thus, multiple studies have probed neural responses to natural images and movies (Isik, Singer, Madsen, Kanwisher, & Kreiman, 2017; Lesica & Stanley, 2004; Okazawa, Tajima, & Komatsu, 2015; Olshausen & Field, 1996; Simoncelli & Olshausen, 2001; Vinje & Gallant, 2000), and also to everyday

objects (Hung, Kreiman, Poggio, & DiCarlo, 2005; Liu, Agam, Madsen, & Kreiman, 2009; Logothetis & Sheinberg, 1996; Sheinberg & Logothetis, 2001), including faces (Allison et al., 1994; Desimone, Albright, Gross, & Bruce, 1984; Kanwisher, McDermott, & Chun, 1997; Tsao et al., 2006).

(iii) *Semi-serendipitous findings*. Hubel and Wiesel claimed that they discovered orientation tuning while they were scrutinizing the activity of primary visual cortex neurons and observed the responses elicited when they inserted a slide in the projector (Hubel, 1981). Gross and Desimone observed that neurons in ITC fired vigorously when one of the investigators passed in front of the monkey, leading to the investigations about neurons that respond to face stimuli (Gross, 1994). Our own descriptions of so-called "Clinton" or "Anniston" cells in the human medial temporal lobe were also fortuitous (Kreiman, 2002; Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). While the role of luck can be debated, rigorous analyses of neural responses to novel stimuli can lead to discovering unexpected feature preferences.

(iv) *Computational methods*. Despite enormous progress in developing computational models to explain and predict neural responses along ventral visual cortex (Connor, Brincat, & Pasupathy, 2007; DiCarlo, Zoccolan, & Rust, 2012; Riesenhuber & Poggio, 1999; Serre et al., 2007; Wu, David, & Gallant, 2006), there have been few efforts to use those models to create stimuli that efficiently drive a visual neuron. One of these approaches is reverse correlation whereby a rapid succession of white noise stimuli is presented followed by averaging the images preceding spikes (Jones, Stepnoski, & Palmer, 1987). This approach has been successful in elucidating the structure of receptive fields in the retina and, to some extent, in primary visual cortex, but it does not seem to work in higher visual areas, due to the accumulation of non-linearities and also the curse of dimensionality dictated by the reduced sampling of stimulus space. Rather than starting from noise, exploiting natural stimulus statistics has been a productive way of synthesizing images and predicting responses in areas V1 (Olshausen & Field, 1996; Olshausen & Field, 2004), V2 (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013), and V4 (Okazawa et al., 2015). An elegant alternative approach is to use a genetic algorithm whereby the neuron under study can itself dictate which stimuli it prefers. A successful implementation of this idea by Connor and

colleagues (Yamane, Carlson, Bowman, Wang, & Connor, 2008) has been used to investigate selectivity in macaque areas V4 and ITC (Carlson, Rasquinha, Zhang, & Connor, 2011; Hung, Carlson, & Connor, 2012; Vaziri & Connor, 2016).

Using a combination of these stimulus selection approaches, seminal studies led to foundational discoveries about visual processing, including center-surround receptive fields (Kuffler, 1953), neurons in primary visual cortex that is tuned to the orientation of a bar placed within their receptive fields (Hubel & Wiesel, 1962), neurons in area MT that discriminate motion direction (Movshon & Newsome, 1992), neurons in area V4 that is sensitive to colors (Zeki, 1983) and curvature (Gallant, Braun, & Van Essen, 1993; Pasupathy & Connor, 2001), selectivity to natural objects (DiCarlo et al., 2012; Logothetis & Sheinberg, 1996) including faces (Desimone et al., 1984; Tsao et al., 2006), among many others. Despite these extensive and heroic efforts, we still do not know that any of those tuning properties are optimal for those neurons – where optimal means triggering high firing rates. It is conceivable that there could be other stimuli that might more strongly drive neurons in all those areas. Mechanistic models can help us understand how neuronal responses arise and thus design better stimuli. The last two decades have seen significant progress in the development of computational models to help us understand neural tuning properties in visual cortex.

3. Computational models of ventral visual cortex

Inspired by neuroanatomy and neurophysiology, many investigators have developed computational models that capture the basic principles that progressively transform a pixel-like representation of inputs into complex features that can be linearly decoded to recognize objects (Deco & Rolls, 2004; DiCarlo et al., 2012; Fukushima, 1980; Mel, 1997; Olshausen, Anderson, & Van Essen, 1993; Riesenhuber & Poggio, 1999; Serre et al., 2007; Wallis & Rolls, 1997). More recently, this family of models has taken over the computer vision community in the form of deep convolutional network architectures that perform quite well in many object labeling and object detection tasks (He, Gkioxari, Dollar, & Girshick, 2018; He, Zhang, Ren, & Sun, 2015; Krizhevsky, Sutskever, & Hinton, 2012; Serre, 2019; Simonyan & Zisserman, 2014).

While there are important variations across different models, they all share basic design principles and we generically refer to them as a family

of models. The models are hierarchical, typically following a sequential path of operations, mimicking the approximately hierarchical nature of ventral visual cortex (Felleman & Van Essen, 1991). The models consist of multiple layers, following a divide-and-conquer strategy breaking the problem of object recognition into multiple smaller and simpler steps. Each of these steps is characterized by a series of biologically plausible canonical computations, typically including a filter implemented by a dot product, a normalization step, and a max pooling operation. In most of the steps, the operations are performed in a convolutional fashion, such that the same computation is repeated throughout the entire visual field. The dot product operation is characterized by sets of weights that are learnt via training. In the computer vision literature, a prominent way of training these weights is via supervised learning algorithms implementing back-propagation. Through the sequence of computations, units show tuning for increasingly more complex features, accompanied by an increasing degree of invariance to transformations of those features such as changes in position, scale, etc.

These models perform quite well in many object labeling tasks. For example, the ResNet architecture achieved a 4% top-5 error rate in the ImageNet dataset consisting of 1000 possible categories (He et al., 2015). These models also provide a reasonable—yet certainly imperfect—first order approximation to characterize human and monkey behavioral performance in rapid object categorization tasks (Rajalingham et al., 2018; Russakovsky et al., 2014). For example, a recent study showed that deep convolutional network architectures performed as well as, and in many cases better than, forensic facial examiner experts, facial reviewers and so-called facial super-recognizers in a face identification task (Phillips et al., 2018). Furthermore, the activity of units in these models can be mapped onto the activity of neurons along the ventral visual cortex (Yamins et al., 2014), even extrapolating across categories when learning the transformation from models to neurons (O'Connell, Chun, & Kreiman, 2018).

Despite the multiple successes of this family of models, it is clear that they only scratch the surface of what we need to understand about visual cortex and there is a large amount of neuroscience and behavioral data that cannot quite be accounted by current instantiations of these algorithms (Kubilius, Bracci, & Op de Beeck, 2016; Linsley, Eberhardt, Sharma, Gupta, & Serre, 2017; Markov et al., 2014; Serre, 2019; Tang et al., 2018; Ullman, Assif, Fetaya, & Harari, 2016). Because such models do not incorporate aspects of common sense cognitive knowledge about the world other than what was used to label images for training, they constitute a suitable standard

benchmark and null hypothesis to contrast against for any study that aims to investigate any type of high-level information encoding (Kreiman, 2017).

With some exceptions, this family of models has been less concerned with the roles of top-down influences on ventral visual cortex responses. Yet, there is extensive data documenting how top-down signals can modulate neuronal activity in vision. For example, spatial attention can enhance neuronal responses throughout visual cortex (Desimone & Duncan, 1995; Reynolds & Heeger, 2009). Top-down influences are also manifested in the form of modulation by task demands and expectations (Gilbert & Li, 2013).

Of note, this family of models does not explicitly incorporate any type of linguistic or semantic encoding in their design. The models are typically trained to learn to separate images that were labeled as belonging to different classes. For example, an investigator may label 1000 images as pineapples, and label another set of 1000 images as elephants. The model may be trained via supervised learning to separate those two groups of images and the algorithms cited above can do a remarkable job in labeling images, including extrapolating to novel pictures of pineapples and elephants. We next turn our attention to ask whether this ability to assign category labels indicates any type of semantic representation.

4. Category-selective responses do not imply semantic encoding

In many typical neuroscience experiments, investigators may present images containing objects belonging to different categories (Desimone et al., 1984; Freedman, Riesenhuber, Poggio, & Miller, 2001; Hung et al., 2005; Kiani, Esteky, Mirpour, & Tanaka, 2007; Kourtzi & Connor, 2011; Kreiman, Koch, & Fried, 2000b; Liu et al., 2009; Logothetis & Sheinberg, 1996; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Mormann et al., 2011; Quian Quiroga et al., 2005; Sigala & Logothetis, 2002; Sugase, Yamane, Ueno, & Kawano, 1999; Thomas, van Hulle, & Vogels, 2001; Tsao et al., 2006; Vogels, 1999). Throughout inferior temporal cortex, and even in areas of the medial temporal lobe and pre-frontal cortex, investigators have reported selective neuronal responses with higher firing rates elicited by some groups of stimuli compared to others. Do these differential responses indicate any type of semantic encoding?

To be clear about what this question means, we return to the definition of semantics as a linguist representation concerned with meaning. We understand this definition to imply that meaning indicates an abstract

representation, *beyond* what is purely captured by the stimulus features. A system or algorithm that comprehends semantic information should be able to capture the link between lemons and pineapples, and it should be able to discern that a tennis ball is functionally closer to a tennis racquet, even though it looks more similar to a lemon.

To investigate whether distinct neuronal responses to different groups of stimuli reflect semantic encoding, we turn to the null hypothesis for visual representations outlined in the previous section, namely, computational models of object recognition. Consider the model architecture shown in Fig. 1A, consisting of an input image conveyed to a cascade of three convolutional layers (Conv1-Conv3) and a fully connected (fc) layer that classifies input images into one of six possible categories. There are 6 fc units that indicate the probability that the image belongs to each of the six categories. This is clearly a far cry from state-of-the-art models that include hundreds of layers and categorize hundreds of images. The details of the architecture are not too relevant; other architectures including state-of-the-art computer vision models would produce similar results to the ones shown below. We deliberately keep it simple for illustration purposes, and to provide source code that can easily be ran on any machine (see links at the end of the Chapter). This model was trained via back-propagation using images from six categories in the ImageNet dataset (Russakovsky et al., 2014): biological cells (synset number n00006484), Labrador dogs (synset number n02099712), fire ants (synset number n02221083), sports cars (synset number n04285008), roses (synset number n04971313), and ice (synset number n14915184). Examples of these images are shown in the top part of Fig. 1B. The model was able to separate the stimulus categories: top-1 performance in a cross-validated set was 78% (where chance is 16.7%). A 2D representation of the activation strength of the 6 fc units at the top of the model in response to each of the images is shown in Fig. 1B, using a dimensionality reduction technique called tSNE which maps the six dimensional output vector onto two dimensions for visualization purposes (van der Maaten & Hinton, 2008). The colors represent the six different categories, which cluster into overlapping yet distinct groups. For example, the images belonging to the "ice" category (pink) mostly clustered on the bottom left while images belonging to the "rose" category (blue) mostly clustered on the top in Fig. 1B.

We further examined the responses of each of the 6 fc units to all the ~8000 images (Fig. 1C). For example, in the leftmost column, each circle corresponds to the activation of fc unit 1 in response to one of the images.

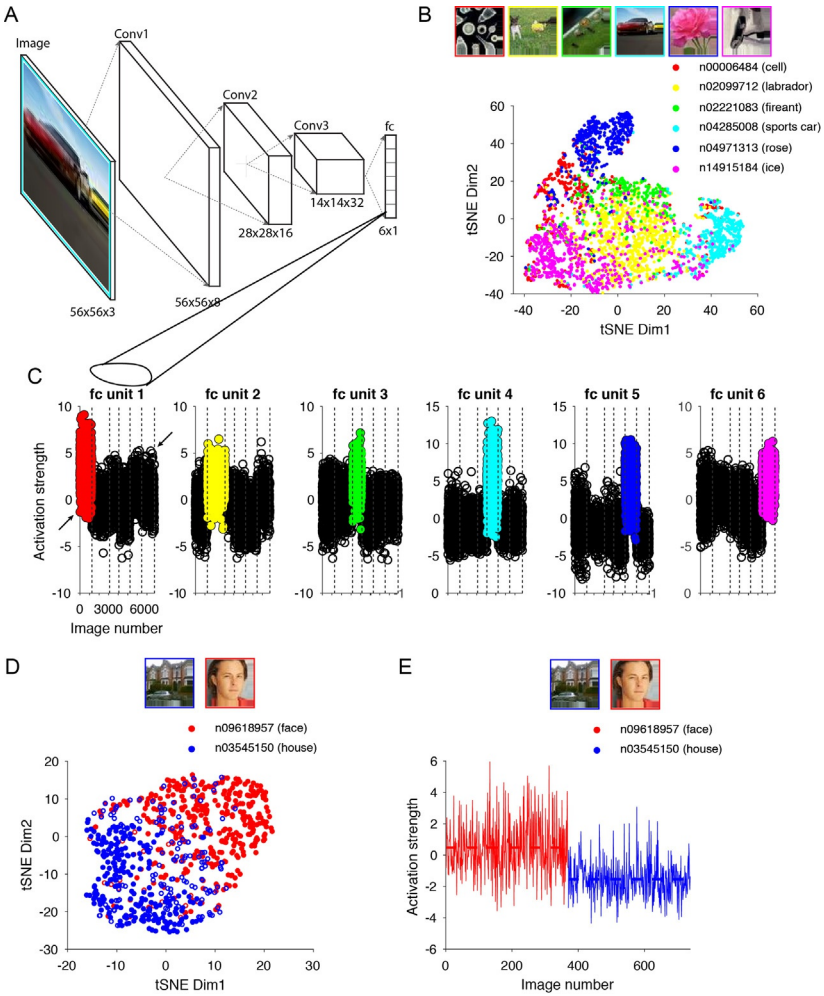

Fig. 1 (A) Simple multi-layer convolutional network consisting of an input layer, three convolutional layers and a fully connected classification layer that classifies images into one of six possible categories: cells, Labradors, fire ants, sports cars, roses and ice (example images from those categories are shown in part B). The network was trained via backpropagation to optimize classification of images belonging to those six categories. (B) Dimensionality reduction using stochastic embedding (van der Maaten & Hinton, 2008) of the activation pattern for the 6 fc layer units from part A in response to each of the images. The color of each dot reflects the image category. (C) Activation strength for each of the 6 fc units in response to all the images. The image categories are separated by vertical dotted lines. The images from the category eliciting the strongest activation for each of the fc units is shown in color, with the colors matching the ones in part B (e.g., fc unit 1 showed stronger activation to images corresponding to the cell category).

(Continued)

The vertical dotted lines separate images from the six different categories. As expected, based on the way the model was trained, each of the fc units showed specialization and responded most strongly to one of the image categories. For example, fc unit 1 showed higher activation on average to the images from the "cell" category (red) compared to all the other categories. The responses were not all-or-none and showed a considerable degree of overlap between categories. For example, certain images of ice (last set of images) yielded stronger activation for fc unit 1 than some of the images of cells (first set of images, compared the two circles highlighted by the two arrows for fc unit 1 in Fig. 1C). The fc units are category units par excellence: by construction, their activation dictates how the model will label a particular image. Yet, the distribution of their activation patterns shows considerable overlap across categorical borders. Even though the model does a decent job at separating the six image categories, the model does *not* seem to have any notion of semantics. A zoomed in picture of a pink car may well be misclassified as a rose. And the diverse and strange patterns of cell shapes can often be misconstrued to indicate ice or ants. The problem in terms of semantics is not with the model performance itself. Deeper models and more extensive training can lead to higher performance. To err is algorithmic, after all. The point here is that the model has no sense of abstract meaning, beyond the *similarity of shape features within a category* represented by its units.

We can still refer to fc unit 5 as a "rose unit" for simplicity. What we mean by a "rose unit" is a unit that is more strongly—but not exclusively—activated by images that contain visual shape features that are common in the set of roses in ImageNet. The unit does not know anything semantic about roses and can show high activation for images from other categories and also low activation for images containing roses, depending on the visual shape features present in the image.

Fig. 1—Cont'd (D) Dimensionality reduction using stochastic embedding of the activation pattern for the 6 fc layer units from part A in response to images of faces (red) or houses (blue). The network was *not* trained to recognize either faces or houses. Yet, a support vector machine classifier with a linear kernel could separate the two categories (empty circles represent wrongly classified images and filled circles represent correctly classified images). (E) Activation pattern of fc unit number 4 (the one showing strongest responses to sports cars) in response to all the images containing faces (red) or houses (blue). The horizontal dashed line indicates the average responses. All the parameters and source code to generate these images are available in http://klab.tch.harvard.edu/resources/Categorization_Semantics.html.

A comparison that pervades the literature is the distinction between images labeled as "human faces" and images labeled as "houses". Would the model in Fig. 1A be able to discriminate human faces versus houses? One might imagine that the model should *not* be able to distinguish human faces from houses because the model was never trained with such images. Even if one were to try to argue that the model has some sort of concrete, as opposed to abstract, understanding of the meaning of cells, sports cars, roses, etc., the model should have no knowledge whatsoever about human faces or houses. In other words, by construction, the model has no semantic information about faces or houses. If the model can still separate faces from houses, then any such separation cannot be based on semantic knowledge. To evaluate whether the model in Fig. 1A can separate pictures of faces versus houses, we considered two additional categories of images: faces (synset number n09618957), and houses (synset number n03545150). We extracted the activation patterns of the 6 fc units of the model in response to each of those human face and house images *without any re-training* (i.e., the model was trained to label the six categories in Fig. 1B and we merely measured the activation in response to these two new categories). We used an SVM classifier with a linear kernel to discriminate pictures of human faces versus houses based on the activity of the 6 fc units. In other words, we asked whether the representation given by the "cell unit," the "Labrador unit," the "fire ant unit," the "sports car unit," the "rose unit," and the "ice unit" was sufficient to separate images of human faces and houses. The classifier achieved a performance of 86% (where chance is 50%). That is, the pattern of activation of the 6 fc units—which are specialized to discriminate cells, Labrador dogs, fire ants, sports cars, roses, and ice—can well separate pictures of human faces from houses. A 2D rendering of the activation patterns of those 6 fc units by the human faces and houses is shown in Fig. 1D, depicting again a clear but certainly not perfect separation of the two categories.

A system that has no semantic knowledge about faces or houses can still separate the two categories quite well. Given the abundant literature on studies about faces versus houses, it is worth further scrutinizing this result. The photographs in the ImageNet dataset are taken from the web and there are a handful of human faces and houses included in the six categories chosen here. The small number of human faces and houses are not uniformly distributed among those six categories and could introduce a small bias. Yet, removing those few human faces and houses does not change the results. Aficionados to the idea that human faces constitute a special group might argue that the images of Labrador dogs do contain animal faces and therefore

the "Labrador" fc unit may help the classifier separate faces from houses. To evaluate this possibility, we computed the signal to noise ratio for each of the 6 fc units in discriminating faces versus houses. The best fc unit was unit number 4 (the one that showed stronger activation by images of sports cars), closely followed by unit number 5 (roses). The worst fc unit was unit number 3 (fire ants), followed by unit number 1 (cells). In other words, the Labrador fc unit is *not* the one that contributes most to the separation of human faces versus houses. The activation pattern of fc unit number 4 (sports cars) in responses to human faces and houses is shown in Fig. 1E. This fc unit showed a clear separation of the two image categories, responding stronger to images of human faces (mean activation $= 0.47 \pm 1.72$) compared to houses (mean activation $= -1.54 \pm 1.18$). As pointed out earlier in connection with Fig. 1C, the distribution of responses for the two categories clearly overlapped.

Now consider an experiment with actual neurons studying the responses to images of faces versus houses. Recording the activity of a neuron that behaved like fc unit 4, in an experiment similar to the one in Fig. 1E, an investigator might be tempted to argue that the neuron represents the semantic concept of faces. Yet fc unit 4 is clearly more strongly tuned to images of sports cars (Fig. 1C, fourth subplot): the mean activation in response to sports cars was 4.59 ± 2.27, which is about 10 times larger than the mean activation in response to human faces (0.47 ± 1.72). There is nothing particularly special about this unit; in fact, all fc units except unit number 3 (fire ants) showed a statistically significant differentiation between images of human faces versus houses. To further dispel any doubts that the Labrador images are playing any role in here, we ran a separate simulation where we trained the same architecture in Fig. 1A from scratch with only 2 fc output units to discriminate images of desks (synset number n03179701) versus images of fried rice (synset number n07868340). The algorithm achieved an accuracy of 98% (chance = 50%). These 2 fc units could be described as a "desk unit" and a "fried rice unit". The pattern of activation of those 2 fc units in response to images of human faces and houses (without any retraining of the network) was able to distinguish them with 73% accuracy. The desk unit showed an activation of 2.49 ± 1.55 in response to images of human faces and an activation of 0.98 ± 1.11 in response to images of houses, clearly differentiating the two categories. The fried rice unit showed an activation of -2.32 ± 1.37 in response to images of human faces versus an activation of -1.13 ± 1.11 for images of houses, clearly differentiating between the two categories. In sum, measuring higher activation

for pictures of one category versus others (e.g., sports cars versus roses or faces versus houses), in and of itself, should *not* be taken to imply any type of semantic representation.

One may still want to maintain that the fc units in Fig. 1A encode some flavor of semantics. After all, a thresholded version of the activity of those units is sufficient to provide a categorical image label. Furthermore, those units are capable of a certain degree of abstraction in the sense that they can label *novel* images that the model has never seen before into those six categories. Such a version of semantics could perhaps be best described as concrete visual shape semantics, as opposed to some abstract version of semantics that transcends visual features.

5. What are the preferred stimuli for visual neurons?

What do those fc units in Fig. 1A actually want? That is, what types of images would trigger high activation in those fc units? We know already from Fig. 1C that images of cells lead to high activation in fc unit 1, images of Labradors lead to high activation in fc unit 2, etc. Therefore, it seems reasonable to argue that fc unit 1 "wants" images of cells, fc unit 2 "wants" images of Labradors and so on. One might even go on to describe fc unit 2 as a "Labrador unit," as we have been doing. But is it possible that there exist other images that lead to even higher activation of those fc units? To investigate this question, we used the Alexnet model (Krizhevsky et al., 2012), pre-trained on the ImageNet dataset (Russakovsky et al., 2014). We considered two of the output units (layer labeled fc 8 in Alexnet). The same analyses can be performed for any other layer but we focus on the classification layer because this is the stage that would presumably contain the highest degree of categorical information. For illustration purposes, we show the activation of fc 8 unit number 209 (Fig. 2A) and fc 8 unit number 527 (Fig. 2B) in response to four categories of stimuli: Labradors, fire ants, desks and sports cars. As expected based on the way that the model was trained, the "Labrador unit" (unit 209) showed larger activation for images containing Labradors compared to the other images (Fig. 2A). Similarly, the "Desk unit" (unit 527) showed larger activation for images containing desks compared to the other images (Fig. 2B). This is the equivalent of the results presented in Fig. 1C. Next, we used the "DeepDream" algorithm to generate images that lead to high activation for those fc units (Mordvintsev et al., 2015). Essentially, the DeepDream algorithm uses the network in reverse mode. Instead of going from pixels to the feature

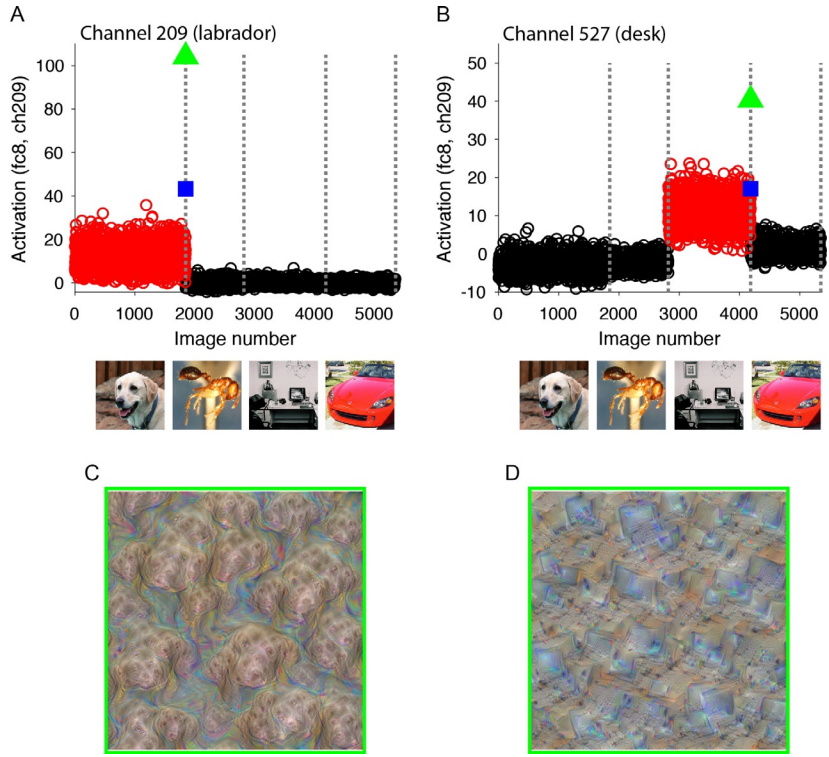

Fig. 2 (A) Activation of unit corresponding to channel 209 in layer fc 8 in Alexnet (Krizhevsky et al., 2012) in response to 1846 images of Labrador dogs (red circles), 972 images of ants, 1366 images of desks, and 1165 images of sports cars (black circles). The vertical dotted lines separate the different image categories. This neural network was trained via backpropagation using 1000 image categories, including the four categories shown here. The channel shown here corresponds to the classification unit for the label "Labrador dog"; as expected, activation for those images was generally larger than activation for other images. (B) Same as A for unit corresponding to channel 527 (Desk). (C) Image generated using DeepDream for Alexnet channel 209 in layer fc 8 (Mordvintsev, Olah, & Tyka, 2015). (D) Image generated using Deep Dream for Alexnet channel 527 in layer fc 8. The images in (C) and (D) led to the activation denoted by the green triangles in (A) and (B). Upon resizing the images in (C) and (D) to be the same size as all the other images in parts (A) and (B), the corresponding activations are the ones shown by the blue squares in (A) and (B). All the parameters and source code to generate these images are available in http://klab.tch.harvard.edu/resources/Categorization_Semantics.html.

representation in a given unit in the network, DeepDream goes from the feature representation in a given unit back to pixels, generating images as its output, and optimizing those images in each iteration to elicit a high activation in the chosen unit. Using DeepDream to generate images that lead to

high activation for the "Labrador unit" produced the image shown in Fig. 2C. The activation strength of the "Labrador unit" in response to the image in Fig. 2C yielded the activation strength shown by the green and blue symbols in Fig. 2A, depending on the size (the blue symbol corresponds to the same exact size as all the other images). The image in Fig. 2C thus triggered higher activation than any of the 1846 photographs of Labradors (even though those photographs were used to train the network). The image in Fig. 2C could well be described using words by a human observer as containing multiple renderings of distorted, sketchy, blurred, Labrador-like patches. Similar results are shown for the "Desk unit" in Fig. 2B and D. After some squinting, it is also possible to discern some resemblance to desk-like features in Fig. 2D, but it is less obvious. In sum, what fc units want is an image rendering complex features, features that are not easily mapped onto English words, though they certainly resemble aspects of the actual photographs used to train the algorithms. Those fc units respond most strongly to images that cannot be obviously predicted by the labels assigned to them. While one may still want to refer to those units as "Labrador units" and "Desk units," it is clear that there are many images that would not be labeled as Labradors or desks by any human observer, and yet they trigger higher activation in those units, even higher than real-world photographs containing those categories.

To summarize, typical Neuroscience experiments are limited by how long it is possible to record from a neuron. Investigators must make hard choices about which stimuli to present. There is a rich and exciting literature with many experiments showing that neuronal responses can discriminate among different categories of stimuli. As illustrated here by the computational models in Figs. 1 and 2, these types of responses do *not* imply any type of semantic encoding. Simple computational models can also yield responses that distinguish different categories (Fig. 1B), those responses are not all-or-none (Fig. 1C), category differentiation can be demonstrated using units that are known to be clearly semantically unrelated to those categories (Figs. 1D and E), and complex images that do not directly map onto any semantic meaning can trigger higher activation in putative categorical units (Fig. 2).

6. Models versus real brains

These deep convolutional bottom-up computational models cast a doubt on claims about semantic encoding based on category-selective responses and provide a null hypothesis to compare against. Yet, these

computational models are a far cry from real biological systems in all sorts of ways and therefore it is fair to question to what extent we can extrapolate conclusions from these computational models to the types of representations manifested by real neurons. Advocates of semantics would rightly argue that the exercises in the previous section merely reflect toy models and that it remains unclear whether the same observations apply to actual neuronal recordings from real brains. The observation that these models can reproduce certain aspects of selectivity in neurophysiological recordings does not imply that one can rule out the presence of semantic information in neural data.

Although deep convolutional models are still rather primitive and fail to incorporate much of the architecture and function of biological circuits, recent studies have shown that these models can explain a relatively large fraction of the variance in neuronal responses (Maheswaranathan, Kastner, Baccus, & Ganguli, 2018; Yamins et al., 2014). In fact, category-selective responses from biological neurons also show the type of properties illustrated in Fig. 1 (e.g., Hung et al., 2005; Kreiman et al., 2000b; Sigala & Logothetis, 2002; Vogels, 1999) and therefore the same cautionary notes should be used in interpreting neuronal selectivity. Furthermore, recent work has shown that it is possible to generate effective stimuli for biological neurons in a fashion similar to the procedure illustrated in Fig. 2 (Ponce et al., 2019). The authors used a procedure similar to the DeepDream algorithm discussed earlier to generate images while recording neuronal responses used to guide the evolution of images triggering high firing rates. The resulting set of synthetic images triggered activation in biological neurons that was as strong as or in many cases even stronger than natural stimuli, similar to the synthetic images created in Fig. 2. In other words, for biological neurons along the ventral visual cortex, the type of stimuli that trigger strongest activation are not real world objects with semantic meanings, but rather complex shapes with features shared with real world objects but distinct and abstract and without any obvious semantic meaning (Ponce et al., 2019).

Absence of compelling evidence for semantic encoding does not constitute evidence of absence of semantics. The fact that we cannot conclude that there is abstract semantic information by observing the responses to a given category versus others in this type of experiments certainly should not be interpreted to imply that semantic information does not exist. The point in the previous section is that merely showing differential patterns of activity between two (or more) categories of stimuli is more of a reflection about the choice of stimuli and about the way the images were gathered

rather than any mysterious notion of abstract meaning. The family of deep convolutional network models should be used as a null hypothesis for any statement concerning the representation of abstract meaning in experiments on visual images. We can thus define abstract semantic encoding as visual discriminations that *cannot* be accounted for by the family of null computational models of visual recognition.

Rather than discussing presence or absence of semantic information in a binary fashion, it is probably more useful to consider different levels of abstraction and invariance. At the bottom level is the notion of template matching, i.e., a neuron that responds when a specific combination of pixels is shown within its receptive field. Increasing the degree of invariance, we can consider a neuron that responds with approximately the same intensity when the stimulus shows small changes such as a complex cell in primary visual cortex and its responses to an optimally oriented bar at different positions within the receptive field. Increasing the degree of abstraction, we can consider neurons in inferior temporal cortex that show tolerance to some amount of 2D rotation of their preferred stimuli and neurons that respond to visually similar exemplars from a given category such as the ones modeled in the previous section. A significant step upwards in invariance would be to find neurons that show a similar response to a tennis ball, a tennis racquet, a tennis court, a tennis skirt and the word Wimbldon. To the best of my knowledge, there is no evidence yet for such a representation.

7. In search of abstraction in the brain

What type of experimental data would provide evidence in favor of abstract semantic information? Returning to the examples used in the definition of semantics, it would be nice to show neuronal responses that are similar for a tennis ball and a tennis racket and yet very different between a tennis ball and a lemon. In other words, it would be nice to show (i) images that have a similar visual appearance (e.g., a tennis ball and a lemon) and yet they trigger very different responses, and (ii) images that are visually dissimilar (e.g., a tennis ball and a tennis racket) and yet they trigger very similar responses.

An elegant step in this direction was carried out by generating morphs between synthetic images of cats and dogs and training monkeys to behaviorally separate them (Freedman et al., 2001). The authors could titrate the

visual similarity of the stimuli and separate purely visual shape features from the task-relevant categorical differentiation between them. The authors described the activity of neurons in pre-frontal cortex that correlated with the categorical distinctions rather than the visual appearance distinctions between stimuli. While pre-frontal cortex neurons better reflected such task-dependent abstract information, neurons in inferior temporal cortex also showed evidence for encoding the categorical boundaries (Meyers et al., 2008). Furthermore, monkeys could be retrained to change their definition of the categorical boundaries and pre-frontal cortex neurons altered their tuning to reflect the new categorical definitions imposed by the task demands (Cromer, Roy, & Miller, 2010).

Another set of intriguing results comes from neural recordings in human epilepsy patients. Some patients suffering from pharmacologically intractable epilepsy are implanted with electrodes as part of the clinical procedure for potential surgical resection of the seizure focus. This clinical situation provides a rather unique opportunity to record the spiking activity of neurons in the human brain, particularly in areas of the medial temporal lobe including the hippocampus, entorhinal cortex, parahippocampal gyrus and the amygdala (Engel, Moll, Fried, & Ojemann, 2005; Fried, Cerf, Rutishauser, & Kreiman, 2014; Kreiman, 2007; Mukamel & Fried, 2012). This line of research has generated observations leading to claims about categorical invariance (Kreiman et al., 2000b; Mormann et al., 2011). There have also been responses to specific individuals or to specific landmarks (Quian Quiroga et al., 2005). These studies are subject to the same type of caveats highlighted in the previous section. However, in several of those cases, the invariant responses were triggered by sets of images that were very different from each other based on visual inspection (there was no quantitative documentation of visual shape similarity based on computational models). The subjective visual dissimilarity of those stimuli suggests that it would be difficult to account for those responses purely based on the type of visual similarity described by the null family of standard models. Particularly striking are the cases where the neurons responded to the image of a particular individual as well as text version of their name (Quian Quiroga et al., 2005), and cases where the neurons responded in a selective fashion during visual imagery in the absence of any visual input (Gelbard-Sagiv, Mukamel, Harel, Malach, & Fried, 2008; Kreiman, Koch, & Fried, 2000a). The activity of human medial temporal lobe neurons taken as a whole therefore reflects a high degree of abstraction. Interestingly, these responses

tend to occur rather late in the game, arising somewhere between 200 and 300 milliseconds after stimulus onset depending on the specific area, which is at the very least 50–150 milliseconds after the selective visual responses described in both monkey (Eskandar, Richmond, & Optican, 1992; Hung et al., 2005; Keysers, Xiao, Foldiak, & Perret, 2001) and human (Liu et al., 2009) inferior temporal cortex. Additionally, both humans (Thorpe, Fize, & Marlot, 1996) and monkeys (Fabre-Thorpe, Richard, & Thorpe, 1998) can behaviorally categorize images well before the onset of those responses. Thus, these medial temporal lobe responses are more likely to reflect the encoding of emotional information and the formation of episodic memories (both of which are likely to depend on semantic encoding of information), rather than the visual categorization per se.

According to the broad definition of semantics as aspects of neuronal responses that cannot be accounted by the null standard models of visual recognition, multiple studies have shown task dependent modulation of neurophysiological responses throughout visual cortex. For example, in an elegant study, neurons in primary cortex responded differently to the same stimulus within their receptive field depending on whether the information was relevant or not for the current task (Li, Piech, & Gilbert, 2004). Task-dependent expectations can also modulate responses all the way down to V1 neurons (Gilbert & Li, 2013). Satisfying such task demands can be considered an important aspect of abstraction in the sense of considering the incoming inputs in the context of current goals.

8. Semantics and the least common sense

Common sense, or general semantic knowledge about the world, is hard to find. The definition of semantics including linguistic-like information, at least taken literally, suggests that we should be looking for a high level of abstraction, beyond what can be described by current visual object recognition models. One practical issue to tackle semantics is that it is difficult to study language in non-human animals. Strangely, there are even investigators that have claimed that language is unique to humans (Berwick & Chomsky, 2015). Additionally, there is minimal data on single neuron responses in language areas in the human brain. One may imagine that any linguistic information from medial temporal lobe structures, from task-dependent representations in pre-frontal cortex, or from language areas, may very well propagate back to ventral visual cortical areas and it might be possible to discern those top-down semantic influences in visual cortex.

In the spirit of stimulating further discussions and future work, we conclude with a brief desiderata of experiments and models to further our understanding of what visual cortical neurons really want and the role of semantic information.

[1] Computational models should play an integral part in the design of visual experiments to elucidate what neurons want. The family of deep convolutional network models provides a reasonable null hypothesis to start with. The models can be used to quantify what fraction of neuronal response variability can be explained, but also to generate images and design the experiments themselves. As an example of this approach, Fig. 2 illustrates a way in which a model can generate images that trigger high activation in its units and it will be interesting to further evaluate this line of reasoning in neuronal recordings.

[2] Given the limited amount of data that we can acquire for a given neuron despite laborious and heroic experiments, we should be open to the idea that we have yet to uncover what neurons truly want. Our understanding and description of the tuning properties of neurons along ventral visual cortex may have to be significantly revisited. Two important recent developments may accelerate progress: the advent of sophisticated computational models that can provide quantitative hypothesis for testing beyond classical experimental designs, and the experimental possibility of holding neuronal recordings for prolonged periods of time (McMahon, Jones, Bondar, & Leopold, 2014).

[3] Task demands seem to play a critical role in dynamically shaping neuronal responses beyond the dimensions that are purely dictated by sensory inputs. As one example of a recent surprising finding in this direction, neurons in rodent primary visual cortex are strongly modulated not only by the visual inputs but also by the speed at which the animal is moving (Niell & Stryker, 2010). There are plenty of opportunities to further investigate how top-down modulation can dynamically route information according to the current behavioral goals.

[4] To uncover semantic encoding, we would like to ensure that the neuronal responses cannot be explained by the null family of models. A neuron encoding semantic information should show a similar response to images that share meaning but which have no similarity in their appearance. Additionally, such a neuron should show a different response to images that are visually similar but do not share the same meaning.

[5] Another important question for future research is to elucidate the neuronal mechanisms of how abstraction can be learnt. There has been extensive work showing that visual cortical neurons can change their responses as a consequence of associations formed by different stimuli (Higuchi & Miyashita, 1996; Messinger, Squire, Zola, & Albright, 2001; Miyashita, 1988; Suzuki, 2007). Extending such mechanisms might lead to the formation of semantic links such as those established by the statistical co-occurrences of tennis balls, racquets, courts, and skirts.

9. Data availability

All the code used to generate Figs. 1 and 2 is available for download from: http://klab.tch.harvard.edu/resources/Categorization_Semantics.html.

We cannot provide the images used in the experiments in Figs. 1 and 2. However, we provide the synset identification numbers, which can be used to freely download all the images from the following site: http://image-net.org/.

References

Allison, T., Ginter, H., McCarthy, G., Nobre, A. C., Puce, A., Luby, M., et al. (1994). Face recognition in human extrastriate cortex. *Journal of Neurophysiology*, *71*, 821–825.

Berwick, R., & Chomsky, N. (2015). *Why only us: Language and evolution*. Cambridge, MA: MIT Press.

Carlson, E. T., Rasquinha, R. J., Zhang, K., & Connor, C. E. (2011). A sparse object coding scheme in area V4. *Current Biology*, *21*, 288–293.

Chapman, B., Stryker, M., & Bonhoeffer, T. (1996). Development of orientation preference maps in ferret primary visual cortex. *Journal of Neuroscience*, *16*, 6443–6453.

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, *17*, 140–147.

Coogan, T., & Burkhalter, A. (1993). Hierarchical organization of areas in rat visual cortex. *The Journal of Neuroscience*, *13*, 3749–3772.

Cromer, J. A., Roy, J. E., & Miller, E. K. (2010). Representation of multiple, independent categories in the primate prefrontal cortex. *Neuron*, *66*, 796–807.

Deco, G., & Rolls, E. T. (2004). *Computational neuroscience of vision*. Oxford Oxford University Press.

Desimone, R., Albright, T., Gross, C., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*, 2051–2062.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*, 415–434.

Engel, A. K., Moll, C. K., Fried, I., & Ojemann, G. A. (2005). Invasive recordings from the human brain: Clinical insights and beyond. *Nature Reviews. Neuroscience*, *6*, 35–47.

Eskandar, E. N., Richmond, B. J., & Optican, L. M. (1992). Role of inferior temporal neurons in visual memory. I. Temporal encoding of information about visual images, recalled images, and behavioral context. *Journal of Neurophysiology*, *68*, 1277–1295.

Fabre-Thorpe, M., Richard, G., & Thorpe, S. J. (1998). Rapid categorization of natural images by rhesus monkeys. *Neuroreport*, *9*, 303–308.
Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47.
Freedman, D., Riesenhuber, M., Poggio, T., & Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312–316.
Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience*, *16*, 974–981.
Fried, I., Cerf, M., Rutishauser, U., & Kreiman, G. (2014). *Single neuron studies of the human brain. Probing cognition*. Cambridge, MA: MIT Press.
Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*, 193–202.
Gallant, J. L., Braun, J., & Van Essen, D. C. (1993). Selectivity for polar, hyperbolic, and Cartesian gratings in macaque visual cortex. *Science*, *259*, 100–103.
Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2008). Internally generated reactivation of single neurons in human Hippocampus during free recall. *Science*.
Ghose, G. M., & Maunsell, J. H. (2008). Spatial summation can explain the attentional modulation of neuronal responses to multiple stimuli in area V4. *Journal of Neuroscience*, *28*, 5115–5126.
Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews. Neuroscience*, *14*, 350–363.
Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cerebral Cortex*, (5), 455–469.
He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2018). Mask R-CNN. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://arxiv.org/abs/1703.06870.
He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. *arXiv*. 1512.03385.
Hegde, J., & Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas v2 and v4. *Cerebral Cortex*, *17*, 1100–1116.
Higuchi, S., & Miyashita, Y. (1996). Formation of mnemonic neuronal responses to visual paired associates in inferotemporal cortex is impaired by perirhinal and entorhinal lesions. *PNAS*, *93*, 739–743.
Hubel, D. (1981). Evolution of ideas on the primary visual cortex, 1955–1978: A biased historical account. In *Nobel lectures*. https://www.nobelprize.org/uploads/2018/06/hubel-lecture.pdf.
Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–154.
Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*, 215–243.
Hung, C. C., Carlson, E. T., & Connor, C. E. (2012). Medial axis shape coding in macaque inferotemporal cortex. *Neuron*, *74*, 1099–1113.
Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast read-out of object identity from macaque inferior temporal cortex. *Science*, *310*, 863–866.
Isik, L., Singer, J., Madsen, J. R., Kanwisher, N., & Kreiman, G. (2017). What is changing when: Decoding visual information in movies from human intracranial recordings. *NeuroImage*, *180*, 147–159.
Jones, J. P., Stepnoski, A., & Palmer, L. A. (1987). The two-dimensional spectral structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*, 1212–1232.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, *17*, 4302–4311.
Keysers, C., Xiao, D. K., Foldiak, P., & Perret, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience*, *13*, 90–101.
Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*, 4296–4309.
Kobatake, E., & Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, *71*, 856–867.
Koch, C. (1999). *Biophysics of computation*. New York: Oxford University Press.
Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review of Neuroscience*, *34*, 45–67.
Kreiman, G. (2002). On the neuronal activity in the human brain during visual recognition, imagery and binocular rivalry. In *Biology*. Pasadena: California Institute of Technology.
Kreiman, G. (2004). Neural coding: Computational and biophysical perspectives. *Physics of Life Reviews*, *1*, 71–102.
Kreiman, G. (2007). Single neuron approaches to human vision and memories. *Current Opinion in Neurobiology*, *17*, 471–475.
Kreiman, G. (2017). A null model for cortical representations with grandmothers galore. *Language, Cognition and Neuroscience*, *32*, 274–285.
Kreiman, G., Koch, C., & Fried, I. (2000a). Imagery neurons in the human brain. *Nature*, *408*, 357–361.
Kreiman, G., Koch, C., & Fried, I. (2000b). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, *3*(9), 946–953.
Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS*. *Montreal*. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.
Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, *12*, e1004896.
Kuffler, S. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neurophysiology*, *16*, 37–68.
Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*, 572–575.
Lesica, N. A., & Stanley, G. B. (2004). Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *Journal of Neuroscience*, *24*, 10731–10740.
Li, W., Piech, V., & Gilbert, C. D. (2004). Perceptual learning and top-down influences in primary visual cortex. *Nature Neuroscience*, *7*, 651–657.
Linsley, D., Eberhardt, S., Sharma, T., Gupta, P., & Serre, T. (2017). What are the visual features underlying human versus machine vision. In *IEEE ICCV workshop on the mutual benefit of cognitive and computer vision*. https://ieeexplore.ieee.org/document/8265530.
Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: Fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, *62*, 281–290.
Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621.
Maheswaranathan, N., Kastner, D. B., Baccus, S. A., & Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. *PLoS Computational Biology*, *14*, e1006291.
Markov, N. T., et al. (2014). A weighted and directed interareal connectivity matrix for macaque cerebral cortex. *Cerebral Cortex*, *24*, 17–36.

McMahon, D. B., Jones, A. P., Bondar, I. V., & Leopold, D. A. (2014). Face-selective neurons maintain consistent visual responses across months. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8251–8256.

Mel, B. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, *9*, 777.

Messinger, A., Squire, L. R., Zola, S. M., & Albright, T. D. (2001). Neuronal representations of stimulus associations develop in the temporal lobe during learning. *Proceedings of the National Academy of Sciences of the United States of America*, *98*, 12239–12244 [Epub 12001 Sep 12225].

Meyers, E., Freedman, D., Kreiman, G., Miller, E., & Poggio, T. (2008). Dynamic population coding of category information in ITC and PFC. *Journal of Neurophysiology*, *100*, 1407–1419.

Miyashita, Y. (1988). Neuronal correlate of visual associative long-term memory in the primate temporal cortex. *Nature*, *335*, 817–820.

Mordvintsev, A., Olah, C., & Tyka, M. (2015). DeepDream—A code example for visualizing neural networks. In *Google research*. Mountain View: Google.

Mormann, F., Dubois, J., Kornblith, S., Milosavljevic, M., Cerf, M., Ison, M., et al. (2011). A category-specific response to animals in the right human amygdala. *Nature Neuroscience*, *14*, 1247–1249.

Movshon, J. A., & Newsome, W. T. (1992). Neural foundations of visual motion perception. *Current Directions in Psychological Science*, *1*, 35–39.

Mukamel, R., & Fried, I. (2012). Human intracranial recordings and cognitive neuroscience. *Annual Review of Psychology*, *63*, 511–537.

Nassi, J., Gomez-Laberge, C., Kreiman, G., & Born, R. (2014). Corticocortical feedback increases the spatial extent of normalization. *Frontiers in Systems Neuroscience*, *8*, 105.

Niell, C. M., & Stryker, M. P. (2010). Modulation of visual responses by behavioral state in mouse visual cortex. *Neuron*, *65*, 472–479.

O'Connell, T., Chun, M. M., & Kreiman, G. (2018). *Zero-shot neural decoding of basic-level object category*. Denver: Cosyne.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences of the United States of America*, *112*, E351–E360.

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, *13*, 4700–4719.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Olshausen, B., & Field, D. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*, 481–487.

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area V4: Position-specific tuning for boundary conformation. *Journal of Neurophysiology*, *86*, 2505–2519.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*, *115*, 6171–6176.

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. (2019). Evolving super stimuli for real neurons using deep generative networks. *Biorxiv*. https://doi.org/10.1101/516484.

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*, 7255–7269.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*, 168.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019–1025.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). ImageNet large scale visual recognition challenge. *CVPR: arXiv*. 1409.0575, 02014.

Serre, T. (2019). Deep learning: The good, the bad and the ugly. In *Annual Review of Vision*. in press.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research*, *165C*, 33–56.

Sheinberg, D. L., & Logothetis, N. K. (2001). Noticing familiar objects in real world scenes: The role of temporal cortical neurons in natural vision. *Journal of Neuroscience*, *21*, 1340–1350.

Sigala, N., & Logothetis, N. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. *Nature*, *415*, 318–320.

Simoncelli, E., & Olshausen, B. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 193–216.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv*. 1409.1556.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature*, *400*, 869–873.

Suzuki, W. A. (2007). Making new memories: The role of the hippocampus in new associative learning. *Annals of the New York Academy of Sciences*, *1097*, 1–11.

Tanaka, K. (2003). Columns for complex visual object features in the inferotemporal cortex: Clustering of cells with similar but slightly different stimulus selectivities. *Cerebral Cortex*, *13*, 90–99.

Tang H, Lotter W, Schrimpf M, Paredes A, Ortega J, Hardesty W, Cox D, Kreiman G (2018) Recurrent computations for visual pattern completion. PNAS, 115 (35) 8835-8840.

Thomas, E., van Hulle, M., & Vogels, R. (2001). Encoding of categories by noncategory-specific neurosn in the inferior temporal cortex. *Journal of Cognitive Neuroscience*, *13*, 190–200.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*, 670–674.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer vision. *PNAS*, *113*(10), 2744–2749.

van der Maaten, L., & Hinton, G. (2008). Visualizing high-dimensional data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Vaziri, S., & Connor, C. E. (2016). Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Current Biology*, *26*, 766–774.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*, 1273–1276.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys: Part 2: Single-cell study. *European Journal of Neuroscience*, *11*, 1239–1255.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology*, *51*, 167–194.

Wu, M. C., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505.

Yamane, Y., Carlson, E. T., Bowman, K. C., Wang, Z., & Connor, C. E. (2008). A neural code for three-dimensional object shape in macaque inferotemporal cortex. *Nature Neuroscience*, *11*, 1352–1360.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, 8619–8624.

Zeki, S. (1983). Color coding in the cerebral cortex—The reaction of cells in monkey visual cortex to wavelengths and colors. *Neuroscience*, *9*, 741–765.