NeuroImage 180 (2018) 101–109

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Principles for models of neural information processing

Kendrick N. Kay

Center for Magnetic Resonance Research, Department of Radiology, University of Minnesota, Twin Cities, Minneapolis, MN, USA

A B S T R A C T

The goal of cognitive neuroscience is to understand how mental operations are performed by the brain. Given the complexity of the brain, this is a challenging endeavor that requires the development of formal models. Here, I provide a perspective on models of neural information processing in cognitive neuroscience. I define what these models are, explain why they are useful, and specify criteria for evaluating models. I also highlight the difference between functional and mechanistic models, and call attention to the value that neuroanatomy has for understanding brain function. Based on the principles I propose, I proceed to evaluate the merit of recently touted deep neural network models. I contend that these models are promising, but substantial work is necessary (i) to clarify what type of explanation these models provide, (ii) to determine what specific effects they accurately explain, and (iii) to improve our understanding of how they work.

1. Introduction

There has been a recent surge of excitement in deep neural networks for neuroscience (Kriegeskorte, 2015; Yamins and DiCarlo, 2016). Major advances in training deep neural networks were achieved by the artificial intelligence and computer vision communities, and these networks now achieve unprecedented performance levels on certain computer vision tasks such as visual object recognition (Krizhevsky et al., 2012). Following these developments, neuroscientists studying the visual system have shown that responses of units in deep neural networks correlate strongly with experimentally measured responses in the primate visual system (e.g., Agrawal et al., 2014; Cadieu et al., 2014; Eickenberg et al., 2017; Güçlü and van Gerven, 2015a; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2016; Yamins et al., 2014). Due to these correspondences as well as similarities in architecture between the artificial and biological networks, deep neural networks have been touted as excellent models of biological neural systems.

In this paper, I use the excitement elicited by deep neural networks as an opportunity to think carefully and critically about models of brain function. I step back and consider the broad endeavor of developing models in cognitive neuroscience (Sections 2 and 3) and provide an assessment of why we should develop such models (Sections 4 and 5). I then highlight the important distinction between functional and mechanistic models (Section 6) and propose specific criteria for evaluating models (Section 7). I end by using the principles I propose to evaluate the merit of deep neural network models (Section 8).

While I write this paper as a *Comments and Controversies* article, I acknowledge that many of the proposed ideas (e.g. Sections 2–6) may be introductory and uncontroversial in nature, especially to current practitioners of model-based neuroscience. My intention in this article is to start from first principles and lay out my views clearly and simply, so that the critical, more controversial content (e.g. Sections 7 and 8) comes well justified. I hope that the more basic content will also serve as a useful primer for those interested in understanding modeling or considering engaging in model-based research. Finally, this paper is not a comprehensive review of computational neuroscience, but is rather a personal perspective stemming from my experience developing models of image processing in visual cortex. This perspective is rooted in the traditions of sensory neuroscience, and I hope to spark a dialogue with researchers who hail from other fields of neuroscience.

2. What is cognitive neuroscience?

Before reasoning about models in cognitive neuroscience, we must first define these various terms. Gazzaniga, Ivry, and Magnun define 'cognitive neuroscience' as

> "The question of understanding how the functions of the physical brain can yield the thoughts and ideas of an intangible mind" (Gazzaniga et al., 2014).

It is widely accepted that "thoughts and ideas of an intangible mind," or mental operations more generally, can be viewed as information-processing operations: for example, the brain represents sensory information, stores sensory information, reasons about this information, and uses information to guide motor behavior. Thus, the brain can be viewed as an organ that mediates interactions between an organism and its environment, accepting incoming sensory information and delivering outgoing motor information.

E-mail address: kay@umn.edu.

https://doi.org/10.1016/j.neuroimage.2017.08.016
Received 26 March 2017; Received in revised form 2 August 2017; Accepted 3 August 2017
Available online 6 August 2017

At a coarse level, we already know what the general information-processing operations performed by the brain are. To use an example from visual neuroscience (DiCarlo and Cox, 2007), we know that one information-processing operation performed by the brain is to take complex spatiotemporal patterns of light impinging on the retina and to use this information to decide what the source of these inputs are (e.g. what type of object is present in the environment). Or, to use an example from social neuroscience (Kubota et al., 2012), we know that one information-processing operation performed by the human brain is to form shortcuts (or "stereotypes") about other humans and use this information to influence future behavior. But without further work, we do not know the specific details of how the brain performs these operations. For that, we must carefully measure the components of the brain and identify how specific neurons and neural populations perform information processing.

3. What is a model?

A small but growing number of researchers are using model-based approaches to tackle questions in cognitive neuroscience (e.g., Brouwer and Heeger, 2013; Forstmann et al., 2011; Huth et al., 2012; Kay and Yeatman, 2017; O'Doherty et al., 2007; Santoro et al., 2014; Shadlen and Newsome, 2001; Sprague and Serences, 2013; and others). I propose a simple, general definition of 'model': a model is a description of a system. In neuroscience, a model would describe how the nervous system is physically structured (anatomy) and/or how its activity changes dynamically over time (physiology). In the specific field of cognitive neuroscience, a model would describe how the anatomy and physiology of the nervous system accomplish behaviorally relevant information-processing tasks. The cognitive neuroscientist asks: for a given brain region, what stimulus, cognitive, or motor operations are performed by neurons in that region?[1]

Given the broadness of the proposed definition, nearly any neuroscience result could be viewed as providing a model. However, models vary drastically in how precise and quantitative they are. For example, models can be qualitative, conceptual, and vague about assumptions (e.g., a description in an introductory textbook, or a 'word' model that involves poorly defined jargon), or models can be quantitative, mathematical, and explicit about assumptions (e.g., a formal implementation of a model in computer code). Models can depend on concepts and labels derived from our own cognitive abilities as human observers (e.g., oracle models that require manually labeling complex audiovisual stimuli in order to make predictions (Huth et al., 2012)), or models can provide explicit specification of how concepts and labels are computed independent of an observer (e.g., a computational implementation of stimulus category (Kay and Yeatman, 2017)). Models can describe systems at coarse levels of detail (e.g., overall activity in a brain region) or at fine levels of detail (e.g., ion channels). As cognitive neuroscientists, we all attempt to describe how the brain performs information processing, and so technically we are all 'modelers'. Of course, in practice, when we use the term 'model', we are typically referring to descriptions that have been made precise and quantitative, and I adopt this usage for the rest of this paper.[2]

4. Models make falsifiable claims

Models perform real scientific work, and are not simply *ad hoc* appendages to an experimental study. Rather, models make substantive falsifiable claims and can progressively improve in sophistication and detail. Consider the following simple experiment (Fig. 1, left). We ask a human observer to direct her eyes towards a small dot at the center of a blank display. The small dot changes color periodically and we instruct the observer to press a button when the color changes. Meanwhile, we place a stimulus (e.g. a checkerboard) on the display, and move this stimulus to a variety of different positions. As we manipulate the stimulus, we record neural activity in the observer's occipital cortex using some technology (e.g. fMRI). We discover that there is an increase in activity when the stimulus is present on the display and that there is some variation in activity levels as a function of stimulus position.

In this example, the system consists of the stimulus, task, observer, behavior, measurement device, and recorded activity. Our goal, as cognitive neuroscientists, is to describe this system and, in particular, to describe why the increases in neural activity occur. There are many possible descriptions, or models, that we could propose (Fig. 1, right). For example, let us consider four potential models:

Model 1 There is visual information on the display (it is not blank). That is why occipital cortex shows increased neural activity.

Model 2 There is a point-to-point mapping between positions on the display and positions on cortex (Engel et al., 1997). That is why neural activity at a given cortical position increases for some stimulus positions, but not others.

Model 3 Spatial extent is one property of a visual stimulus. For a given cortical position, this property is represented through a mathematical operation that takes the spatial extent of the stimulus and performs a weighted sum using a Gaussian function to generate the activity level (Dumoulin and Wandell, 2008). Thus, neural activity levels are what they are because occipital cortex performs this operation.

Model 4 Light reflected from the display enters the eye, is refracted by the lens, and is focused onto the retina. Photoreceptors in the retina transduce light energy into electrical voltages. These voltages are communicated by different types of cells to retinal ganglion cells, which send action potentials to the LGN. In turn, neurons in the LGN send action potentials to primary visual cortex. At each stage in this process, sensitivity is local (e.g., photoreceptors are sensitive to light from a restricted region of the visual field, neurons in the LGN receive input from a specific collection of neighboring retinal ganglion cells, etc.). The net result of these processing stages can be summarized by any of the earlier three models.

Although the above models vary widely in sophistication and detail (and we could go into even further detail with respect to molecular mechanisms), all of the models describe the system under consideration and make substantive falsifiable claims. Each model posits certain variables as causally related to the observed neural activity and implicitly excludes other variables. The claim is that the visual stimulus matters to the neural activity, but that for example, the auditory background noise that happened to be present during the experiment, the motor behavior,

[1] This definition is most closely aligned with 'encoding' approaches to cognitive neuroscience in which the experimenter attempts to predict brain activity measurements in terms of specific stimulus, cognitive, or motor features that are present during the experiment. 'Decoding' approaches reverse the directionality, attempting to use brain activity measurements to infer stimulus, cognitive, or motor features. Although there are important technical differences between these approaches (Naselaris and Kay, 2015; Naselaris et al., 2011), these differences are not critical to the issues discussed in this paper.

[2] I briefly comment on the distinction between models, theories, and simulations. Compared to a model, a theory is more expansive in scope and typically more qualitative and less tied to an experimental dataset. For example, one might have a theory for the functional role of feedback connections in sensory processing, whereas one might develop a model that quantitatively accounts for the consequences of feedback connections observed in a particular experiment. There is also the distinction between a model and a simulation. I roughly define 'simulation' as the use of a model to demonstrate an effect. The purpose of a simulation is not so much to account for a specific set of data, but rather to demonstrate an interesting phenomenon or one that is generally observed in experimental data. For example, one might simulate a large number of model neurons under some realistic parameter settings and demonstrate that a surprising network effect emerges.

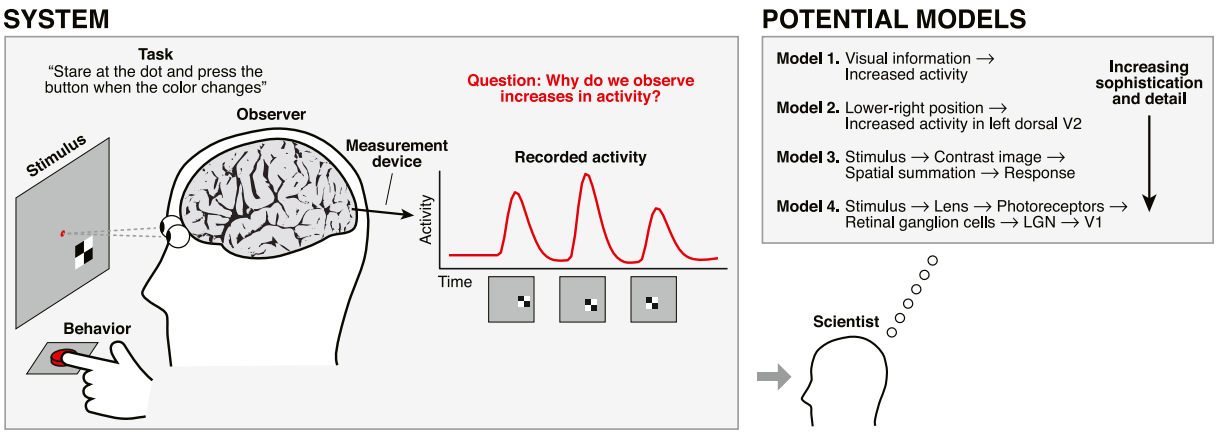

Fig. 1. Models describe systems at various levels of sophistication and detail. A typical cognitive neuroscience experiment consists of a stimulus, task, observer, behavior, measurement device, and recorded activity (*left*). A scientist attempts to develop a model of the system, that is, a description of the events that are occurring in the system (*right*). Of particular interest is to characterize why specific levels of neural activity are observed. A variety of different models can be proposed, ranging in sophistication and detail.

and the internal cognitive state of the observer do not. With additional experimental measurements, we can test whether the models are indeed sufficient or whether modifications to the models are necessary. If we find that variables such as auditory stimulation or cognitive state affect the observed activity, these variables must be included to achieve a complete description of the system.

The examples provided above, like many studies in cognitive neuroscience, characterize neural activity in specific brain regions. This approach assumes we have already accurately identified the relevant brain regions in a given observer. However, this is a non-trivial endeavor that should be performed carefully (Benson et al., 2012; Frost and Goebel, 2012; Glasser et al., 2016; Gordon et al., 2016; Sabuncu et al., 2010; D. Wang et al., 2015; L. Wang et al., 2015; Weiner and Grill-Spector, 2012). To aid localization, an increasing number of researchers are developing quantitative models that describe where distinct regions and networks are located within the brain (Haxby et al., 2011; Huth et al., 2016; Nelson et al., 2010; Yeo et al., 2011). Interestingly, locations of regions and networks in human cortex do not appear to be random and are instead very predictable. Recent research indicates this predictability may stem from several types of neurobiological substrates. For example, cortical folding (Benson et al., 2012; Weiner et al., 2018, 2014), white matter (Saygin et al., 2011; Yeatman et al., 2014), cytoarchitectonics (Rosenke et al., 2018; Weiner et al., 2016a), and myelination (Glasser et al., 2016) can all contribute to predicting the locations of functional regions.

5. Why are models useful?

Developing precise and quantitative descriptions of how the brain performs information processing takes effort. In my view, models provide three main benefits: summary, explanation, and prediction. I provide a general description of these benefits below, and refer the reader to a concrete example taken from previous work (Fig. 2).

5.1. Summary

Neural measurements are complex and noisy, and there is no limit to the number of experimental variations that one could investigate. Models can provide compact summaries of the information processing that a neural system is performing. Thus, a major benefit of a model is that one can make inferences on a focused set of parameters that summarize the data, instead of attempting to interpret a large number of noisy individual data points. Parameters derived in this way can then be compared across brain areas (e.g. Kay et al., 2013b) or subject populations (e.g. Schwarzkopf et al., 2014).

5.2. Explanation

Models posit that specific variables relate to neural activity. As such, models provide *explanations* of measurements of the brain. For example, suppose we find that a neuron is highly active when a clip of rock music is played but is only weakly active when a speech clip is played. Why does this occur? One model could be that the neuron computes overall sound intensity, and the reason we observe weak activity for the speech clip is that it has low sound intensity. Alternatively, there are other candidate models that might explain the phenomenon (e.g., selectivity for guitar tones, variations in attentional engagement). With appropriate experimental measurements, we can adjudicate different models and decide which model is most accurate (Naselaris and Kay, 2015).

5.3. Prediction

There are several different senses in which models provide predictive power. One sense comes from cross-validation (Hastie et al., 2001), a procedure that is commonly used in model-based studies. In cross-validation, the researcher sets aside some testing data, fits the parameters of a model on the remaining training data, and then assesses how well the fitted model predicts the testing data. The testing data could reflect distinct trials of the same experimental conditions found in the training data, in which case this demonstrates limited predictive power. Alternatively, the testing data could reflect completely novel experimental conditions, which demonstrates stronger predictive power.

A different sense in which models provide predictive power is if a model developed in one study is able to predict the results of a new study that does not involve exactly the same subjects, stimuli, and task design used in the first study. For example, it has been shown that a model developed using simple artificial stimuli and fMRI measurements successfully generalizes to complex naturalistic stimuli (Kay et al., 2013a) as well as data obtained from a different measurement technique (Winawer et al., 2013). As another example, it has been shown that a model that describes structural-functional relationships in one group of subjects can successfully generalize to a new group of subjects (Rosenke et al., 2018; Weiner et al., 2018, 2016a).

A third and deep sense in which models provide predictive power is if a model can predict the consequences of physical perturbations to the brain. If we had accurate and detailed descriptions of how neural systems coordinate to perform information-processing operations, we should be able—in principle—to predict, for example, the effects of lesions made in specific brain areas (Dricot et al., 2008; Gallant et al., 2000), surgical removal of entire brain areas (Weiner et al., 2016b), the effects of enhancement (Salzman et al., 1990) or disruption (Parvizi et al., 2012;

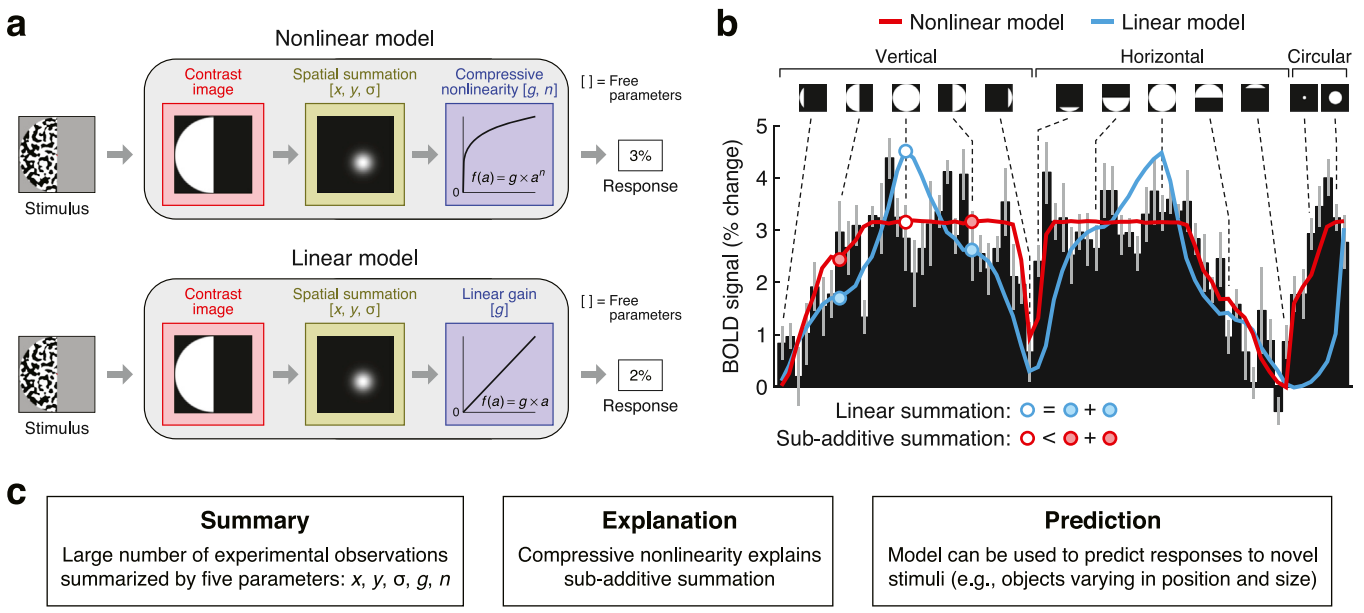

Fig. 2. A concrete example of how models provide summary, explanation, and prediction. Figure adapted from Kay et al., 2013a. *a*, Two potential models of how spatial extent of visual stimuli relates to neural responses. The nonlinear model starts with a contrast image representing stimulus location, computes a weighted sum of this contrast image using a 2D Gaussian, and applies a compressive nonlinearity. The linear model is identical except that the compressive nonlinearity is removed, leaving a linear gain. *b*, Data and model predictions for a voxel in visual area TO-1. Black bars indicate measured BOLD responses to different stimulus locations (depicted by small icons). Leave-one-stimulus-out cross-validation was used to fit the models, and thick lines indicate model predictions. An effect of interest is whether the response to a full stimulus (open dots) is equal to the sum of the responses to two partial stimuli (filled dots). The data support sub-additive summation, which is captured by the nonlinear model. *c*, Three functions performed in this modeling example. (1) The nonlinear model *summarizes* the large set of noisy measurements using just five parameters. (2) The removal of the compressive nonlinearity leads to linear summation, which does not match the data; thus, the compressive nonlinearity is necessary for, and *explains*, sub-additive summation. (3) The nonlinear model *predicts* responses to novel stimuli. For example, the nonlinear model predicts specific levels of tolerance in responses to objects varying in position and size, and experimental measurements have confirmed this prediction (Kay et al., 2013a).

Pascual-Leone and Walsh, 2001; Rangarajan et al., 2014) of neural activity, or the effects of psychoactive drugs (Rokem and Silver, 2010). Note that these are not easy predictions to make, assuming we are careful to avoid the illusion of predictive power that comes from making "predictions" after looking at the data. A model conjured to explain effects that have already been observed generates "postdictions" and should be treated with skepticism.

6. Functional vs. mechanistic models

It is important to distinguish between *functional models* and *mechanistic models* of neural information processing (Albrecht et al., 2002; Carandini, 2012; Carandini and Heeger, 2011). Functional (or 'computational') models characterize the transformation between input and output performed by a neuron or population of neurons (Wu et al., 2006), reminiscent of the concept of functions in mathematics or programming. Mechanistic (or 'biophysical' or 'circuit') models characterize the details of the mechanism by which a neuron or population of neurons carry out such a transformation (Priebe, 2016). Thus, a functional model attempts only to match the outputs of a system given the same inputs provided to the system, whereas a mechanistic model attempts to also use components that parallel the actual physical components of the system.

To illustrate, recall Models 1–3 from the previous example. These models are all stimulus-referred (Heeger et al., 1996; Wandell et al., 2015) in the sense that they specify how the stimulus relates to activity in occipital cortex. Thus, the models can be viewed as functional models that characterize the transformation between input (stimulus) and output (neural activity). In contrast, Model 4 concerns not only the stimulus, but also the series of physical events that intervene between the stimulus and activity in occipital cortex. This model can therefore be viewed as a mechanistic model that characterizes how the brain carries out the transformation described by Models 1–3. There may be multiple possible mechanistic models that are all consistent with a given functional model. Functional models can be rigorously established for a system, even if the underlying mechanisms are not known.

Functional and mechanistic models are complementary to one another and should be judged on their own merits. The value of functional models is that they emphasize the outcomes and meaning of neural information processing. The significance of a signal carried by a neuron or population of neurons ultimately lies in what that signal conveys about sensory or motor information for the observer. For example, if an organism encounters a predator, what matters is successful detection of the predator so that motor behavior can be appropriately guided; how that detection is accomplished is of secondary importance. Focusing on mechanisms without addressing sensory or motor significance would produce an incomplete picture of neural information processing.

These points are directly related to David Marr's well-known levels of analysis where distinctions are made among computational, algorithmic, and implementation levels (Marr, 1982). Slightly generalizing the definition of 'mechanistic' to refer to the details of how something is accomplished, we see the algorithmic level serves as a mechanistic model for the computational level and the implementational level serves as a mechanistic model for the algorithmic level. For example, imagine a situation where an organism is attempting to determine the location of a predator from auditory inputs. We can describe the system at a computational level by characterizing the problem that the organism is trying to solve: given auditory inputs, detect the predator and determine the direction of the predator. We can describe the system at an algorithmic level by identifying the specific set of auditory and decision-making algorithms that the brain uses to solve the problem. Or we can describe the system at an implementational level by identifying the specific configurations of neurons and connections that implement those algorithms. Each level provides details as to how the level above is accomplished.

Many studies in cognitive neuroscience develop functional models

and ignore anatomical implementation. For example, a researcher might use fMRI to investigate how patterns of population activity represent a stimulus, irrespective of details of how this activity is spatially organized across cortex. Or, a researcher might use electrophysiology to study how individual neurons respond to experimental conditions, irrespective of details of cell types or the circuit that a neuron participates in. However, anatomical mechanisms may hold valuable clues to function (Amunts and Zilles, 2015; Kennedy et al., 2016). If the brain spends so much neurobiological energy organizing structure and function across spatial scales, presumably this orderliness is useful for something. For example, perhaps the specific way that functional properties are clustered in the brain enables faster and more efficient readout of information for a particular task (Grill-Spector and Weiner, 2014).

The path to understanding anatomical implementation will be difficult. Every measurement technique has limitations on resolution and coverage, and no single technique provides all of the necessary information. Population techniques that aggregate over multiple neurons (e.g. fMRI) tell us very little about individual neurons. Thus, models of population neural activity are, in a sense, functional models that do not provide implementational details of how the brain creates the population activity. Conversely, fine-scale techniques (e.g. electrophysiology) typically do not sample all types of neurons nor all regions of the brain, and therefore risk missing neurons or brain regions relevant to the cognitive phenomenon under investigation. Thus, models developed using such techniques may lead to incomplete descriptions of neural information processing. My working view is that all levels of analysis are useful and should be pursued: we should strive to build accurate functional models that abstract from implementational details as well as accurate mechanistic models that reveal how function is achieved by the physical components of the brain. Hopefully, with sufficiently developed models, we will one day be able to bridge the vast differences in scales of measurement in neuroscience (Sejnowski et al., 2014).

7. What makes a good model?

Thus far, I have addressed what models of neural information processing are, why they are useful, and the distinction between functional and mechanistic models. Now suppose in our daily work, we come across a model put forth by a researcher in the field. How should we evaluate the merit of the model? I propose the use of two criteria, accuracy and understanding.

7.1. Accuracy

Accuracy refers to how well a given model performs in matching the system under investigation (for example, see Fig. 2b). It is sometimes disparagingly remarked that a model is 'just fitting the data'—on the contrary, quantitatively matching experimental measurements is exactly what a model ought to do. To assess accuracy, we collect experimental data at some spatial and temporal scale, perform proper preparation and binning of those data, and then quantify whether the predictions of a model match the data. Typically, models have free parameters whose values are not known *a priori* and must be set to obtain quantitative predictions. These parameters are usually tuned to fit experimental data, and in such cases, it is crucial to control for overfitting. This can be done by evaluating predictive performance on left-out data (i.e. cross-validation) or by using techniques that penalize goodness-of-fit based on number of free parameters (e.g. Akaike Information Criterion).

Beyond quantifying predictive power for a given set of data, we should also consider the range and diversity of the experimental manipulations represented by those data. A model should describe how the brain carries out information processing in a broad range of situations, not just the specific situations used in one or a few particular studies (Felsen and Dan, 2005; Kay et al., 2013b). For example, suppose a model that operates on images is developed for an experiment in which a fixed image duration is used (e.g. 100 ms). If we obtain new measurements using different image durations (e.g. continuous time-varying image sequences), does the model still accurately account for the data? As another example, suppose visual sinusoidal gratings are presented to an observer and a model is proposed in which the activity of a neuron is calculated as a weighted sum of the luminance values of the stimulus (Carandini et al., 2005). This model posits that neural activity reflects a specific visual attribute and, by implication, does not reflect other visual, cognitive, or motor attributes. Evidence for the accuracy of the model would be greatly strengthened if we performed a diverse range of experimental tests—for example, using naturalistic visual scenes to deliver luminance stimulation (David et al., 2004), manipulating the internal cognitive state of the observer (McAdams and Reid, 2005), or allowing visual stimulation to occur simultaneously with motor responses—and still found that the same model (with exactly the same parameters) accurately predicts neural activity. By performing stringent tests of a model, we gain confidence in its accuracy.

I comment briefly on the topic of phenomenological models. It is possible to have a model that accurately matches a set of data, but performs no actual explanatory work. Such models (which can also be termed 'purely descriptive models') may be useful for comparison purposes, but do not provide neuroscientific insight (Albrecht et al., 2002). For example, suppose we are investigating how neural responses to stimuli change as a function of the cognitive task that a subject is performing (Kay and Yeatman, 2017). We could propose a model that allows each task to induce an additive offset to neural responses, and this model could be fit and evaluated like any other model. However, the model does not make a substantive claim about the specific property of the tasks that is responsible for the additive offsets, and therefore has limited neuroscientific value. For instance, imagine trying to predict responses for a novel cognitive task—the model would be incapable of doing so because it does not provide any insight into the nature of cognitive tasks.

7.2. Understanding

The second criterion for the merit of a model is understanding, which refers to how well we, as scientists, grasp the relationship between the components of a given model and the outcomes that the model predicts. Or, in simpler terms, *do we know how the model works?* To illustrate, suppose we observe neural activity is higher in one experimental condition compared to another. A model that describes this system should indicate what property of the first condition leads to increased neural activity. If the model successfully conveys what this property is, we will have *understood* why the effect occurs. In practice, models can be mathematically or algorithmically complex, and it may take effort to determine which specific model component is responsible for a given effect (for an example of how this can be done, see Fig. 2).

It is helpful to consider examples where model understanding is poor. Suppose we wish to characterize the relationship between two continuous variables, x and y. One approach is to characterize y as a weighted sum of the outputs of nonlinear basis functions defined on x (e.g., the weighted sum of a large number of Gaussian functions). Another approach is to simply characterize y as a linear function of x. Now suppose the relationship between x and y is, in fact, linear. Both the complex nonlinear model and the simple linear model are identical in their behavior and equally accurate in matching the data. However, the complex model has less value because it provides less understanding: to understand the model, we have to expend additional effort analyzing the tuning properties of the basis functions and the weights associated with the basis functions.

As another example, suppose we have two code implementations of a functional model of neural information processing, one set of code being short, concise, and well-documented, the other set of code being long, convoluted, and undocumented. Both sets of code behave identically in their input-output behavior and achieve the same accuracy in matching experimental data. However, the longer code has less value because it provides less understanding: to figure out what model is implemented by

the code, we have to pore through and digest computer code. This example highlights the importance of clarity in model-based research. Clarity should ideally exist at all levels: the verbal or conceptual level (scientific prose), the mathematical level (equations), and for models that are algorithmically complex, the computational level (code).

What are some practical methods for improving understanding of a model? One is to simply observe the model's behavior. Observing how a model behaves across different experimental manipulations is useful, even if empirical measurements of those manipulations are not available. For example, a functional model of visual processing could be probed using a variety of different stimulus manipulations, such as changing the orientation of a bar, changing the semantic category of an object, etc. Carefully controlled experimental manipulations help isolate and identify what effects are explained by a given model (Rust and Movshon, 2005). A second method for improving understanding is to manipulate the model and examine the effect on the model's behavior (Kay et al., 2013b; Nishimoto and Gallant, 2011). If we remove a certain model component or change a certain model parameter, does the model fail to account for the effect of interest? If the model fails, we have learned that the identified component or parameter is critical (for an example, see Fig. 2). If the model still works, we have learned that the identified component or parameter is not critical, and we could remove it to obtain a simpler and easier-to-understand model. A third method is to model the model, that is, perform simulations of the model's behavior and attempt to develop a simpler model that accounts for the observed behavior. For instance, in the previously described example involving variables x and y, we could take the complex nonlinear model, perform simulations, and eventually realize that a simple linear model reproduces the model's behavior.

7.3. Trade-offs between accuracy and understanding

Ideally, we achieve models that are both highly accurate and well understood. But what happens when these criteria come into conflict? For example, how do we choose between a complex model that is accurate but difficult to understand and a simple model that is less accurate but easier to understand?

I acknowledge that unlike accuracy, understanding is difficult to quantify, and I do not think there is a general method for weighing accuracy against understanding. My view is that model assessment is a subjective decision that must be made on a case-by-case basis. Moreover, it is not clear that there needs to be a single "best" model for a given neural system. For example, in some situations, a model is simply used as a tool to summarize a set of experimental data. In these cases, we should choose a model that is good at summarizing and that can be practically estimated from the available data, even though this model might not be the most accurate model available.

I also acknowledge that it is not easy to know when we have achieved 'sufficient' understanding. One possibility is that we sufficiently understand a model if we can simulate and perform predictions of the model in our minds without having to resort to paper or a computer. Admittedly, this sets the bar very high. It may very well turn out that certain neural systems are intrinsically complex and require a very large number of parameters to describe their behavior, and in these situations, it might be impossible to develop simple models that admit understanding. Whether or not this is the case is an empirical question.

8. The case of deep neural networks

Now that I have covered principles for assessing models of neural information processing, I turn to the specific case of deep neural networks (DNNs). These networks, inspired by properties of biological visual systems (Fukushima, 1980; Serre et al., 2007), consist of multiple layers of processing, where each layer is composed of units that perform relatively simple linear and nonlinear operations on the outputs of previous layers. Connections between units are typically designed such that a convolution is performed in the linear weighting step (same weights are applied at different positions), which parallels the visual system. Parameters of the networks are typically set using supervised learning techniques, optimizing performance on specific tasks such as predicting the object category associated with the visual input (Yamins et al., 2014). Researchers have demonstrated high levels of correlation between activity exhibited by DNN units and measurements of activity in visual cortex in response to naturalistic objects and scenes (Eickenberg et al., 2017; Güçlü and van Gerven, 2015b; Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014).

Do DNNs have merit as models of biological visual systems? The answer depends on the specific claim that is being made. Suppose the claim is simply that activity in visual cortex reflects a series of processing operations that are performed on visual inputs provided to an observer. This minimal interpretation, that 'visual cortex is a multi-layer neural network that processes visual inputs', is a simplistic qualitative model that is neither exciting nor objectionable, but nevertheless counts as a valid model (see Section 3). Presumably, there is a deeper, more substantive claim that one wants to make regarding DNNs, and the merit of this claim will depend heavily on how seriously we want to take the details of the architecture and parameters used in a DNN. Do we wish to adopt the extreme view that every parameter value in a DNN is critical and every DNN unit corresponds to a specific neuron or neural population in the brain? If not, what is the proposed interpretation?

An important distinction that affects the interpretation of DNNs is whether they are intended as functional or mechanistic models (see Section 6). Suppose DNNs are intended only as functional models of how stimuli (inputs) relate to neural responses (outputs). In this case, there are a number of open questions regarding the accuracy of DNNs. Thus far, researchers have examined large-scale datasets involving a diversity of complex naturalistic stimuli and demonstrated general correspondence between artificial and biological responses. However, much work in visual neuroscience has characterized in detail how specific visual areas represent specific stimulus dimensions, such as contrast (Albrecht et al., 2002), spatial extent (Kay et al., 2015), curvature (Brincat and Connor, 2004), color (Horwitz and Hass, 2012), and spatial frequency (Lennie and Movshon, 2005), just to name a few. Do DNNs accurately account for these effects? Furthermore, we should scrutinize the range and diversity of the experimental datasets that have been examined thus far. DNNs provide potential explanations of stimulus-driven activity, but these are incomplete descriptions of the brain given that visual activity is affected by non-stimulus factors such as attention (Luck et al., 1997), imagery (O'Craven and Kanwisher, 2000), and working memory (Harrison and Tong, 2009).

Suppose instead that DNNs are intended as mechanistic models that not only characterize stimulus-response transformations, but also the way in which the brain accomplishes those transformations. If this is the intention, we again are faced with a number of open questions. What is the proposed mapping between individual units in a layer of a DNN and the neurons in a given brain area? Are DNNs attempting to account for variations in the physical sizes of different visual areas (Dougherty et al., 2003)? Do layer-to-layer connections in a DNN accurately reflect physical connections in biological visual systems, e.g., the spatial extent of V1 neurons that project to a V2 neuron (Sincich et al., 2003)? How can we reconcile DNNs with the existence of bypass routes in corticocortical connections (Felleman and Van Essen, 1991) which violate a strictly hierarchical organization? Can DNNs account for the role of different cell types, the laminar organization of cortex, and the existence of extensive feedback projections?

In addition to raising questions about accuracy, I also raise questions about our understanding of DNNs. The computational capabilities of DNNs depend critically on the specific parameters used in the models (Coates et al., 2011; Pinto et al., 2009). However, DNNs have many thousands (or even millions) of free parameters, and so understanding DNNs is not a trivial task. If we do not take steps to understand DNNs and treat these models as 'black boxes', they provide the benefit of prediction

but do not provide much benefit with respect to summary and explanation (see Section 5). They do not summarize well because, without further work aimed towards simplifying the models, there are a large number of potential parameters contributing to the behavior of the system; they do not explain well because, without further work, it is not clear which specific parts of the models are necessary to explain specific effects.

Overall, I am not denying that DNNs have merit, but I am highlighting open questions and limitations that apply to DNNs if we want to take them seriously as models of neural information processing. Worries and concerns about the biological accuracy of neural networks are not new (Crick, 1989), but the recent explosion in size and complexity of neural networks warrants additional worries about our understanding of these networks. I believe there is substantial work to be done towards clarifying what type of explanation these models are supposed to provide and determining what specific experimental effects they accurately explain. We also need to improve our understanding of how these models work. Fortunately, there are concrete steps we can take towards improving understanding (as discussed in Section 7). We can observe the models (e.g., inspect responses to controlled stimuli (Eickenberg et al., 2017)), we can manipulate the models (e.g., perturb parameters and examine the consequences (Cichy et al., 2016)), and we can model the models (e.g., perform simulated experiments 'in silico' and derive simpler models that achieve the same behavior).

The concerns voiced here have shaped my own research approach which has produced models that explain specific effects and have components that are well understood. In particular, I have developed a multi-stage model of visual processing (Kay et al., 2013b; Kay and Yeatman, 2017) that shares architectural similarities to DNNs but is much simpler and is linked to specific effects observed in fMRI measurements. In a recent Vision Sciences Society talk, a researcher made a valiant effort to explain how the model works, spelling out the specific parameters and operations that compose the model (Benson et al., 2017). Ironically, even though the model is much more tractable than a DNN, an audience member complained that the model seemed too complicated. Understanding how models work may be a challenging endeavor, but we should applaud such efforts and we should apply the same criterion of 'understanding' to DNNs that we might apply to other models. There are researchers on the outskirts of computational neuroscience and members of the public far outside neuroscience who are viewing the recent rise of deep neural networks with a mixture of interest and wonder. One goal of this paper is to remind ourselves to approach these models with equal balance of excitement and skepticism.

9. Conclusion

I wrote this perspective at a broad, non-technical level to speak to a general audience and to remove us from the messy, often confusing, details of different measurement methods (e.g., fMRI, EEG/MEG, electrophysiology), different data analysis approaches (e.g., multivariate pattern analysis, representational similarity analysis, voxelwise modeling, functional connectivity), and jargon (e.g., encoding, decoding). Although technical details matter (Naselaris et al., 2011), the goal of this paper is to emphasize the larger point that we should use measurements of the brain to build models of how neurons and neural populations perform complex information-processing operations. Although this statement may seem like a platitude, a quick survey of current cognitive neuroscience research reveals that most work is not directed towards the development of quantitative models. Models should accurately predict what happens under a broad range of experimental manipulations, and we should understand these models through clear description, observation, and manipulation. When we encounter a model in the literature, consider questions such as: How well does the model account for the data? How extensive are the experimental manipulations? How clear is the link between the components of the proposed model and the observed effects? Is the model attempting to provide a functional or mechanistic explanation? Is the model clearly described and reproducible?

It is useful to draw inspiration from other domains of science. In chemistry, we know if we mix a certain amount of one chemical with another, we will observe certain outcomes, such as emission of heat. This is because we have working models of the relevant variables (e.g., molecular composition of the chemicals) and how these variables interact. In astronomy, we know if we observe two celestial bodies headed towards each other, we will observe certain outcomes, such as collision or trajectory deviation. This is because we have working models of the relevant variables (e.g., mass, velocity, presence of other nearby bodies) and how these variables interact. In cognitive neuroscience, suppose we developed models that could predict the behavioral and neural outcomes of an arbitrary experiment involving stimuli and task instructions. Such models would predict how fast and accurate an observer will be at the task, what levels of neural activity will be found in different brain areas, and how these neural activity levels relate to the sensory, cognitive, and motor processes involved. Once we achieve such models, we might be able to claim to know how the brain works.

Acknowledgments

I thank K. Weiner for extensive discussions regarding the relevance of neuroanatomy to modeling efforts as well as for edits to previous versions of this manuscript. I also thank S. Engel, B. Hutchinson, M. Moerel, B. Rokers, N. Rust, and J. Winawer for comments on the manuscript. Portions of this work were presented at a symposium held at Vision Sciences Society 2016.

References

Agrawal, P., Stansbury, D., Malik, J., Gallant, J.L., 2014. Pixels to voxels: Modeling Visual Representation in the Human Brain arXiv preprint arXiv:1407.5104.
Albrecht, D.G., Geisler, W.S., Frazor, R.A., Crane, A.M., 2002. Visual cortex neurons of monkeys and cats: temporal dynamics of the contrast response function. J. Neurophysiol. 88, 888–913.
Amunts, K., Zilles, K., 2015. Architectonic mapping of the human brain beyond brodmann. Neuron 88, 1086–1107. https://doi.org/10.1016/j.neuron.2015.12.001.
Benson, N., Broderick, W., Müller, H., Winawer, J., 2017. An anatomically-defined template of BOLD response in V1-V3. J. Vis. 17 (10), 585. https://doi.org/10.1167/17.10.585.
Benson, N.C., Butt, O.H., Datta, R., Radoeva, P.D., Brainard, D.H., Aguirre, G.K., 2012. The retinotopic organization of striate cortex is well predicted by surface topology. Curr. Biol. 22, 2081–2085. https://doi.org/10.1016/j.cub.2012.09.014.
Brincat, S.L., Connor, C.E., 2004. Underlying principles of visual shape selectivity in posterior inferotemporal cortex. Nat. Neurosci. 7, 880–886.
Brouwer, G.J., Heeger, D.J., 2013. Categorical clustering of the neural representation of color. J. Neurosci. 33, 15454–15465. https://doi.org/10.1523/JNEUROSCI.2472-13.2013.
Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J., 2014. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput. Biol. 10, e1003963. https://doi.org/10.1371/journal.pcbi.1003963.
Carandini, M., 2012. From circuits to behavior: a bridge too far? Nat. Neurosci. 15, 507–509. https://doi.org/10.1038/nn.3043.
Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., Rust, N.C., 2005. Do we know what the early visual system does? J. Neurosci. 25, 10577–10597.
Carandini, M., Heeger, D.J., 2011. Normalization as a canonical neural computation. Nat. Rev. Neurosci. 13, 51–62. https://doi.org/10.1038/nrn3136.
Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci. Rep. 6, 1. https://doi.org/10.1038/srep27755.
Coates, A., Lee, H., Ng, A.Y., 2011. An analysis of single-layer networks in unsupervised feature learning. In: Presented at the Journal of Machine Learning Research, pp. 215–223.
Crick, F., 1989. The recent excitement about neural networks. Nature 337, 129–132. https://doi.org/10.1038/337129a0.
David, S.V., Vinje, W.E., Gallant, J.L., 2004. Natural stimulus statistics alter the receptive field structure of V1 neurons. J. Neurosci. 24, 6991–7006.
DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends Cognit. Sci. 11, 333–341.
Dougherty, R.F., Koch, V.M., Brewer, A.A., Fischer, B., Modersitzki, J., Wandell, B., 2003. Visual field representations and locations of visual areas V1/2/3 in human visual cortex. J. Vis. 3, 586–598.

Dricot, L., Sorger, B., Schiltz, C., Goebel, R., Rossion, B., 2008. The roles of "face" and "non-face" areas during individual face perception: evidence by fMRI adaptation in a brain-damaged prosopagnosic patient. NeuroImage 40, 318–332. https://doi.org/10.1016/j.neuroimage.2007.11.012.

Dumoulin, S.O., Wandell, B., 2008. Population receptive field estimates in human visual cortex. NeuroImage 39, 647–660. https://doi.org/10.1016/j.neuroimage.2007.09.034.

Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: convolutional network layers map the function of the human visual system. NeuroImage 152, 184–194.

Engel, S.A., Glover, G.H., Wandell, B., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cerebr. Cortex 7, 181–192.

Felleman, D.J., Van Essen, D.C., 1991. Distributed hierarchical processing in the primate cerebral cortex. Cerebr. Cortex 1, 1–47.

Felsen, G., Dan, Y., 2005. A natural approach to studying vision. Nat. Neurosci. 8, 1643–1646.

Forstmann, B.U., Wagenmakers, E.-J., Eichele, T., Brown, S., Serences, J.T., 2011. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? Trends Cognit. Sci. 15, 272–279. https://doi.org/10.1016/j.tics.2011.04.002.

Frost, M.A., Goebel, R., 2012. Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59, 1369–1381. https://doi.org/10.1016/j.neuroimage.2011.08.035.

Fukushima, K., 1980. Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybern. 36, 193–202.

Gallant, J.L., Shoup, R.E., Mazer, J.A., 2000. A human extrastriate area functionally homologous to macaque V4. Neuron 27, 227–235.

Gazzaniga, M.S., Ivry, R.B., Mangun, G.R., 2014. Cognitive Neuroscience: the Biology of the Mind, 4th ed. W. W. Norton & Company, New York.

Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., Van Essen, D.C., 2016. A multi-modal parcellation of human cerebral cortex. Nature 536, 171–178. https://doi.org/10.1038/nature18933.

Gordon, E.M., Laumann, T.O., Adeyemo, B., Huckins, J.F., Kelley, W.M., Petersen, S.E., 2016. Generation and evaluation of a cortical area parcellation from resting-state correlations. Cerebr. Cortex 26, 288–303. https://doi.org/10.1093/cercor/bhu239.

Grill-Spector, K., Weiner, K.S., 2014. The functional architecture of the ventral temporal cortex and its role in categorization. Nat. Rev. Neurosci. 15, 536–548. https://doi.org/10.1038/nrn3747.

Güçlü, U., van Gerven, M.A.J., 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. J. Neurosci. 35, 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015.

Güçlü, U., van Gerven, M.A.J., 2015. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. NeuroImage 145, 329–336. https://doi.org/10.1016/j.neuroimage.2015.12.036.

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458, 632–635. https://doi.org/10.1038/nature07832.

Hastie, T., Tibshirani, R., Friedman, J.H., 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In: Springer series in Statistics. Springer, New York.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416. https://doi.org/10.1016/j.neuron.2011.08.026.

Heeger, D.J., Simoncelli, E.P., Movshon, J.A., 1996. Computational models of cortical visual processing. Proc. Natl. Acad. Sci. U. S. A. 93, 623–627.

Horwitz, G.D., Hass, C.A., 2012. Nonlinear analysis of macaque V1 color tuning reveals cardinal directions for cortical color processing. Nat. Neurosci. 15, 913–919. https://doi.org/10.1038/nn.3105.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458. https://doi.org/10.1038/nature17637.

Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014.

Kay, K.N., Weiner, K.S., Grill-Spector, K., 2015. Attention reduces spatial uncertainty in human ventral temporal cortex. Curr. Biol. 25, 595–600. https://doi.org/10.1016/j.cub.2014.12.050.

Kay, K.N., Winawer, J., Mezer, A., Wandell, B., 2013. Compressive spatial summation in human visual cortex. J. Neurophysiol. 110, 481–494. https://doi.org/10.1152/jn.00105.2013.

Kay, K.N., Winawer, J., Rokem, A., Mezer, A., Wandell, B., 2013. A two-stage cascade model of BOLD responses in human visual cortex. PLoS Comput. Biol. 9, e1003079. https://doi.org/10.1371/journal.pcbi.1003079.

Kay, K.N., Yeatman, J.D., 2017. Bottom-up and top-down computations in word- and face-selective cortex. Elife 6, e22341. https://doi.org/10.7554/eLife.22341.

Kennedy, H., Van Essen, D.C., Christen, Y., 2016. Micro-, Meso- and Macro-connectomics of the Brain. Springer.

Khaligh-Razavi, S.-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput. Biol. 10, e1003915. https://doi.org/10.1371/journal.pcbi.1003915.

Kriegeskorte, N., 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu. Rev. Vis. Sci. 1, 417–446.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet Classification with Deep Convolutional Neural Networks, pp. 1097–1105.

Kubilius, J., Bracci, S., Op de Beeck, H.P., 2016. Deep neural networks as a computational model for human shape sensitivity. PLoS Comput. Biol. 12, e1004896. https://doi.org/10.1371/journal.pcbi.1004896.

Kubota, J.T., Banaji, M.R., Phelps, E.A., 2012. The neuroscience of race. Nat. Neurosci. 15, 940–948. https://doi.org/10.1038/nn.3136.

Lennie, P., Movshon, J.A., 2005. Coding of color and form in the geniculostriate visual pathway (invited review). J. Opt. Soc. Am. 22, 2013–2033.

Luck, S.J., Chelazzi, L., Hillyard, S.A., Desimone, R., 1997. Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol. 77, 24–42.

Marr, D., 1982. Vision: a Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co. Inc., New York, NY 2, 4.2.

McAdams, C.J., Reid, R.C., 2005. Attention modulates the responses of simple cells in monkey primary visual cortex. J. Neurosci. 25, 11023–11033. https://doi.org/10.1523/JNEUROSCI.2904-05.2005.

Naselaris, T., Kay, K.N., 2015. Resolving ambiguities of MVPA using explicit models of representation. Trends Cognit. Sci. 19, 551–554. https://doi.org/10.1016/j.tics.2015.07.005.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073.

Nelson, S.M., Cohen, A.L., Power, J.D., Wig, G.S., Miezin, F.M., Wheeler, M.E., Velanova, K., Donaldson, D.I., Phillips, J.S., Schlaggar, B.L., Petersen, S.E., 2010. A parcellation scheme for human left lateral parietal cortex. Neuron 67, 156–170. https://doi.org/10.1016/j.neuron.2010.05.025.

Nishimoto, S., Gallant, J.L., 2011. A three-dimensional spatiotemporal receptive field model explains responses of area MT neurons to naturalistic movies. J. Neurosci. 31, 14551–14564. https://doi.org/10.1523/JNEUROSCI.6801-10.2011.

O'Craven, K.M., Kanwisher, N., 2000. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J. Cognit. Neurosci. 12, 1013–1023.

O'Doherty, J.P., Hampton, A., Kim, H., 2007. Model-based fMRI and its application to reward learning and decision making. Ann. N. Y. Acad. Sci. 1104, 35–53. https://doi.org/10.1196/annals.1390.022.

Parvizi, J., Jacques, C., Foster, B.L., Witthoft, N., Withoft, N., Rangarajan, V., Weiner, K.S., Grill-Spector, K., 2012. Electrical stimulation of human fusiform face-selective regions distorts face perception. J. Neurosci. 32, 14915–14920. https://doi.org/10.1523/JNEUROSCI.2609-12.2012.

Pascual-Leone, A., Walsh, V., 2001. Fast backprojections from the motion to the primary visual area necessary for visual awareness. Science 292, 510–512. https://doi.org/10.1126/science.1057099.

Pinto, N., Doukhan, D., DiCarlo, J.J., Cox, D.D., 2009. A high-throughput screening approach to discovering good forms of biologically inspired visual representation. PLoS Comput. Biol. 5, e1000579.

Priebe, N.J., 2016. Mechanisms of orientation selectivity in the primary visual cortex. Annu. Rev. Vis. Sci. 2, 85–107.

Rangarajan, V., Hermes, D., Foster, B.L., Weiner, K.S., Jacques, C., Grill-Spector, K., Parvizi, J., 2014. Electrical stimulation of the left and right human fusiform gyrus causes different effects in conscious face perception. J. Neurosci. 34, 12828–12836. https://doi.org/10.1523/JNEUROSCI.0527-14.2014.

Rokem, A., Silver, M.A., 2010. Cholinergic enhancement augments magnitude and specificity of visual perceptual learning in healthy humans. Curr. Biol. 20, 1723–1728. https://doi.org/10.1016/j.cub.2010.08.027.

Rosenke, M., Weiner, K.S., Barnett, M.A., Zilles, K., Amunts, K., Goebel, R., Grill-Spector, K., 2018. A cross-validated cytoarchitectonic atlas of the human ventral visual stream. NeuroImage 170, 257–270.

Rust, N.C., Movshon, J.A., 2005. In praise of artifice. Nat. Neurosci. 8, 1647–1650.

Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., Haxby, J.V., 2010. Function-based intersubject alignment of human cortical anatomy. Cerebr. Cortex 20, 130–140. https://doi.org/10.1093/cercor/bhp085.

Salzman, C.D., Britten, K.H., Newsome, W.T., 1990. Cortical microstimulation influences perceptual judgements of motion direction. Nature 346, 174–177. https://doi.org/10.1038/346174a0.

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., Formisano, E., 2014. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput. Biol. 10, e1003412. https://doi.org/10.1371/journal.pcbi.1003412.

Saygin, Z.M., Osher, D.E., Koldewyn, K., Reynolds, G., Gabrieli, J.D.E., Saxe, R.R., 2011. Anatomical connectivity patterns predict face selectivity in the fusiform gyrus. Nat. Neurosci. 15, 321–327. https://doi.org/10.1038/nn.3001.

Schwarzkopf, D.S., Anderson, E.J., de Haas, B., White, S.J., Rees, G., 2014. Larger extrastriate population receptive fields in autism spectrum disorders. J. Neurosci. 34, 2713–2724. https://doi.org/10.1523/JNEUROSCI.4416-13.2014.

Sejnowski, T.J., Churchland, P.S., Movshon, J.A., 2014. Putting big data to good use in neuroscience. Nat. Neurosci. 17, 1440–1441. https://doi.org/10.1038/nn.3839.

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., Poggio, T., 2007. A quantitative theory of immediate visual recognition. Prog. Brain Res. 165, 33–56.

Shadlen, M.N., Newsome, W.T., 2001. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. J. Neurophysiol. 86, 1916–1936.

Sincich, L.C., Adams, D.L., Horton, J.C., 2003. Complete flatmounting of the macaque cerebral cortex. Vis. Neurosci. 20, 663–686.

Sprague, T.C., Serences, J.T., 2013. Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. 16, 1879–1887. https://doi.org/10.1038/nn.3574.

Wandell, B., Winawer, J., Kay, K.N., 2015. Computational modeling of responses in human visual cortex. Brain Mapp. 651–659.

Wang, D., Buckner, R.L., Fox, M.D., Holt, D.J., Holmes, A.J., Stoecklein, S., Langs, G., Pan, R., Qian, T., Li, K., Baker, J.T., Stufflebeam, S.M., Wang, K., Wang, X., Hong, B., Liu, H., 2015. Parcellating cortical functional networks in individuals. Nat. Neurosci. 18, 1853–1860. https://doi.org/10.1038/nn.4164.

Wang, L., Mruczek, R.E.B., Arcaro, M.J., Kastner, S., 1 October 2015. Probabilistic maps of visual topography in human cortex. Cerebr. Cortex 25 (10), 3911–3931. https://doi.org/10.1093/cercor/bhu277.

Weiner, K.S., Barnett, M.A., Lorenz, S., Caspers, J., Stigliani, A., Amunts, K., Zilles, K., Fischl, B., Grill-Spector, K., 2016. The cytoarchitecture of domain-specific regions in human high-level visual cortex. Cerebr. Cortex 27, 146–161. https://doi.org/10.1093/cercor/bhw361.

Weiner, K.S., Barnett, M.A., Witthoft, N., Golarai, G., Stigliani, A., Kay, K.N., Gomez, J., Natu, V.S., Amunts, K., Zilles, K., Grill-Spector, K., 2018. Defining the most probable location of the parahippocampal place area using cortex-based alignment and cross-validation. NeuroImage 170, 373–384.

Weiner, K.S., Golarai, G., Caspers, J., Chuapoco, M.R., Mohlberg, H., Zilles, K., Amunts, K., Grill-Spector, K., 2014. The mid-fusiform sulcus: a landmark identifying both cytoarchitectonic and functional divisions of human ventral temporal cortex. NeuroImage 84, 453–465. https://doi.org/10.1016/j.neuroimage.2013.08.068.

Weiner, K.S., Grill-Spector, K., 2012. The improbable simplicity of the fusiform face area. Trends Cognit. Sci. 16, 251–254. https://doi.org/10.1016/j.tics.2012.03.003.

Weiner, K.S., Jonas, J., Gomez, J., Maillard, L., Brissart, H., Hossu, G., Jacques, C., Loftus, D., Colnat-Coulbois, S., Stigliani, A., Barnett, M.A., Grill-Spector, K., Rossion, B., 2016. The face-processing network is resilient to focal resection of human visual cortex. J. Neurosci. 36, 8425–8440. https://doi.org/10.1523/JNEUROSCI.4509-15.2016.

Winawer, J., Kay, K.N., Foster, B.L., Rauschecker, A.M., Parvizi, J., Wandell, B., 2013. Asynchronous broadband signals are the principal source of the BOLD response in human visual cortex. Curr. Biol. 23, 1145–1153. https://doi.org/10.1016/j.cub.2013.05.001.

Wu, M.C., David, S.V., Gallant, J.L., 2006. Complete functional characterization of sensory neurons by system identification. Annu. Rev. Neurosci. 29, 477–505.

Yamins, D.L.K., DiCarlo, J.J., 2016. Using goal-driven deep learning models to understand sensory cortex. Nat. Neurosci. 19, 356–365. https://doi.org/10.1038/nn.4244.

Yamins, D.L.K., 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. Proc. Natl. Acad. Sci. U. S. A. 111, 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yeatman, J.D., Wandell, B., Mezer, A.A., 2014. Lifespan maturation and degeneration of human brain white matter. Nat. Comm. 5, 4932. https://doi.org/10.1038/ncomms5932.

Yeo, B.T.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165. https://doi.org/10.1152/jn.00338.2011.