Behavioral and neural constraints on hierarchical representations

Odelia Schwartz
Department of Computer Science, University of Miami, Miami, FL, USA

Luis Gonzalo Sanchez Giraldo
Department of Computer Science, University of Miami, Miami, FL, USA

Central to behavior and cognition is the way that sensory stimuli are represented in neural systems. The distributions over such stimuli enjoy rich structure; however, how the brain captures and exploits these regularities is unclear. Here, we consider different sources of perhaps the most prevalent form of structure, namely hierarchies, in one of its most prevalent cases, namely the representation of images. We review experimental approaches across a range of subfields, spanning inference, memory recall, and visual adaptation, to investigate how these constrain hierarchical representations. We also discuss progress in building hierarchical models of the representation of images—this has the potential to clarify how the structure of the world is reflected in biological systems. We suggest there is a need for a closer embedding of recent advances in machine learning and computer vision into the design and interpretation of experiments, notably by utilizing the understanding of the structure of natural scenes and through the creation of hierarchically structured synthetic stimuli.

Introduction

A central question for both artificial and natural intelligence concerns the ways that complex stimuli of various sorts are represented. Indeed, one can characterize much of the computation performed by putatively intelligent entities as involving transformations between different representations of the same input. Similarly, learning may be viewed as establishing forms of representation and representational transformations and acquiring background knowledge that collectively support such intelligence. One reason for this is conceptually simple, if computationally complex: Intelligence requires appropriate mapping of inputs (possibly over substantial periods of time) to output behavior. Short of being told, or learning, for every possible input what behavior is appropriate, it is necessary to generalize—i.e., permitting correct behaviors for novel inputs when those inputs are suitably closely related to familiar ones. Such proximity may be based on abstract or semantic properties, such as learning to recognize a bird in different poses. The notion of proximity thus governs the quality of behavior. Representations, together with the methods of their exploitation, realize the structure of generalization. They also determine issues of encoding and compression that, in turn, may influence memory and forgetting.

In this review, we concentrate on one important and broad aspect of representations, namely their hierarchical nature. We focus on the domain of static vision, i.e., photographs, although hierarchies are ubiquitous. As an example, consider the images of a single class of objects, say birds, in different poses. Birds generally have wings, a tail, a beak, legs; the wings and tail have feathers, and so forth. In many cases, this hierarchical structure will determine the relevant notion of proximity. Hierarchies are very complicated, since, to adopt the terms that Marr (1982) applied generally to the analysis of complex systems, they pose broad and deep computational, algorithmic and even implementation questions. However, such questions must be answered in order for us to make progress in the understanding of neural processing, since there is ample suggestive evidence for various hierarchical forms as we illustrate in detail below.

Although some answers to questions about hierarchies will arise through conceptual analysis, we suggest there is a clear need to understand what existing empirical work and tractable new experiments can reveal. By testing different sorts of generalization—largely in inference or memory—such experiments can

Citation: Schwartz, O., & Sanchez Giraldo, L. G. (2017). Behavioral and neural constraints on hierarchical representations. *Journal of Vision*, *17*(3):13, 1–25, doi:10.1167/17.3.13.

doi: 10.1167/17.3.13 Received July 16, 2016; published March 29, 2017 ISSN 1534-7362

be interpreted as providing constraints on the nature and form of representations.

One of the sources of inspiration for experimental hypotheses and ideas is machine learning, which has recently enjoyed exciting advances in highly relevant areas. We particularly highlight deep learning in neural networks, which has been recently applied to understanding experiments in visual cortex. We suggest a need for further integration of machine learning with experiments on hierarchy. In the Discussion, we expand on a wider range of modern machine learning and probabilistic approaches, which could be potentially further integrated with experiments.

Some preliminary caveats: First, Marr (1982) described the algorithmic level as including the representation of the input and output of a computation, together with the transformation from input to output. This sits in between the computational level, which characterizes the goals of computations and the logic of the strategy for carrying them out, and the implementation level, which describes how suitable algorithms are executed with hardware or neural machinery. However, the computational level, such as the behavioral tasks and goals of an organism, and ultimately survival priorities, can constrain the representation. Further, what we observe in behavioral experiments might be due to a computation on a representation rather than the representation itself. Finally, we distinguish between a representational *scheme*, which parameterizes the overall characterization of the input, and the actual representation of a particular input in the terms of such a scheme—i.e., the values of the parameters concerned, or perhaps an a posteriori probability distribution over those parameters given the input. To put it another way, an input such as an image lives in a very high-dimensional space of pixel activations. However, actual images only fill up a small portion of that space. The representational scheme provides a parameterization of that portion—a collection of axes; the representation for a particular input is then the coordinate values (or a distribution over those coordinate values) for that input.

Sources of hierarchical structure

We first provide a computational-level analysis of why hierarchical representations might be appropriate. One critical idea is that input images may be synthesized or generated by a process that can be well approximated as being hierarchical. This is known as a top-down generative process. The functional inverse to generation is recognition—an operation that typically works bottom-up, mapping an input into the way that it could have been generated. We discuss top-down and bottom-up hierarchies in the next section. Here we focus on three manifestations of hierarchies. Although we make this distinction, note that these are not mutually exclusive.

Part-whole: Perhaps the minimal hierarchical characteristic of a coherent cause of visual input is the creation of wholes with parts. For instance, a whole (e.g., forest scene) is generated with its parts (e.g., sky, birds, or trees). In turn, each of these wholes (e.g., a visual object such as a bird) is generated with its parts (wings, a beak, legs, etc.); and so on. This example refers to parts in a simplified manner, although the actual parts used by the visual system may be more abstract. Biederman (1987) proposed that primitive parts (termed "geons," from geometrical eons) of blocks, cylinders, spheres, and wedges, are important for object recognition. He further argued that recognizing objects, similar to recognizing speech from phonemes, relies on a modest number of such geon parts, and the arrangement of these parts. Although there are an infinite number of ways that wholes can be constructed from their lower level parts, the allowable constructions respect particular constraints (see discussion on compositionality in Bienenstock & Geman, 1995; Bienenstock, Geman, & Potter, 1997). Statistical constraints arise because of coherence between wholes and parts, a prominent example of which is the geometric arrangement of the parts (Felzenszwalb & Huttenlocher, 2005; Felzenszwalb, Girshick, McAllester, & Ramanan, 2010; Felzenszwalb, McAllester, & Ramanan, 2008; Fergus, Perona, & Zisserman, 2003; Fischler & Elschlager, 1973; Sudderth, Torralba, Freeman, & Willsky, 2005). That generation follows particular rules implies that its recognition inverse will possess certain properties. We describe this as a partnering principle. For the case of part-whole generation, the partnering recognition principle is a form of *binding*. This determines how to align parts with the roles they play in the putative wholes, and thereby represent them appropriately. This process thus identifies the parts (e.g., wings, a beak, or legs) and their coherence to infer the whole (bird), a process which can proceed hierarchically up to the whole forest scene

Also associated with part-whole generation is *reuse*, since many parts can be made the same, or at least have similar constructs (e.g., two related wings for each bird; many organized feathers on each wing).

Part-whole generation licenses the structure of advantageous generalization: Learning or observing something about a new input (e.g., that it has a beak) allows substantial inferences about other likely properties (size, position in the image, presence of other visual structure such as the wings), and very many constraints become automatically relevant. Another way of putting this is that there is less entropy than one might at first think in the way that images are

generated. Take the case of spatial location: There would seem to be total freedom as to where to put an object such as bird in a scene. However, this placement then implies tight constraints on where its parts can be (Fergus et al., 2003; Sudderth et al., 2005), thus implying that the underlying displacements have lower entropy than one might otherwise expect. It is this characteristic that allows part-whole representational schemes to be acquired from inputs—the wholes are seen as hidden or latent causes of the reduction in entropy.

A related idea to part-whole hierarchies and reuse is *recursion*. This is the characteristic of some, but not all, natural input, that parts at one level can be wholes at another. This is of particular relevance for the case of language, where one sentence can include another, or itself, for instance as a quotation, but it can certainly happen in vision too. For instance, a picture of a wall in a house could sport a painting of a picture of the same wall. This is known as the "Droste Effect" (apparently after the Droste cocoa powder; see also Month, 2003). However, it's not clear how much such high-level recursive structure exists in natural images. At a lower level, images exhibit scale invariance (Glasner, Bagon, & Irani, 2009; Ruderman & Bialek, 1994; Zoran & Weiss, 2009), and recursion may be seen in the structure of fractal images (Mandelbrot, 1983; Spehar, Clifford, Newell, & Taylor, 2003). Recursion may theoretically form an infinitely deep tree in a hierarchical representation. In practice, the brain might have to treat recursion in a simplified manner, using similar methods as for part-whole hierarchies. Further, the problem that afflicts reuse of keeping track of instances across the breadth of the tree is closely related to that of keeping track of instances across the depth of the tree, which arises for recursion. For these reasons, and since we are also not aware of experimental work targeting recursion in vision, we will not focus further on recursion hierarchies.

Component: An additional hierarchical facet is that the different underlying components of an object might themselves be separately synthesized (potentially involving wholes and parts of their own), and then be combined to make the final input. An example of this arises in the context of Lambertian imaging. Here, the two separate components are the source of the lighting and the visual objects. The color and intensity of the former arises through one part of the generative scheme. This light then illuminates multiple objects, correlating their appearances or pixel values. These correlations arise from the multiplicative structure of the combination of lighting and form—consider what happens, for instance, if the light is bright or dim.

The recognition partner of component combination is *separation*. This involves pulling apart the distinct aspects of the input. For instance, in the example given above, separation involves teasing apart from the observed brightness, both the surface reflectance and the illuminant (Adelson & Pentland, 1996; Tenenbaum & Witkin, 1983). However, note that some attributes, such as the illuminant, may not be represented explicitly in neurons in the visual system. That is, although a generative model may contain both an illuminant and surface reflectance, the visual system may explicitly represent only the reflectance to achieve lightness or color constancy (e.g., Brainard & Radonjić, 2014; Jin & Shevell, 1996; MacEvoy & Paradiso, 2001; however, see "luxotonic" units in Bartlett & Doty, 1974; Kinoshita & Komatsu, 2001).

In real world scenes, separating the illuminant from surface structure is a difficult problem that is likely to be solved hierarchically (see, e.g., Tang, Salakhutdinov, & Hinton, 2012; Tenenbaum & Witkin, 1983). Furthermore, although we have described the illuminant as a global component acting on multiple objects, component hierarchies can act at more local scales and in complex ways. Take the example of an illuminant reflecting off of one whole or part and onto another (Schrater & Kersten, 2002).

In addition to examples of the illuminant, a visual object itself or an object part, may be composed of distinct feature components at multiple levels (e.g., yellow taxi or red beak). This may be seen as another instantiation of component hierarchies. In this case, the partnering recognition process may either be separation or binding. That is, in recognition, color and form may be separated (consider tables that come in many different colors; the system may represent form and color separately), or bound. This question of whether object features are "bound" has been a target of some experimental work. We distinguish between part-whole binding and components (feature) binding, since the generative and recognition models in these two cases might act differently, and experimental studies have sometimes targeted one or another.

Inheritance: This refers to the idea that the entities in the world that are observed as visual objects enjoy higher order semantic structure that licenses further generalization. The key difference between part-whole and inheritance is that the top-level whole, and also the inherited parts, have a semantic status. Knowing, for instance, that birds are vertebrates but not mammals, or, as in the example of Murphy, Hampton, and Milovanovic (2012), knowing that a friend's new "muffelet" is a dog, has many implications for sensory input associated with such objects, even given very little other knowledge or experience. These implications follow a complex scheme of defeasible inheritance. Such schemes are the conventional target of semantic networks—but have visual consequences too.

Bottom-up and top-down hierarchies

It is because images have hierarchical underpinnings that hierarchical representational schemes are appropriate or even necessary. Indeed, the experiments we describe below mostly test whether statistical structure that we can measure in natural inputs or impose in images taken from artificially constructed collections is reflected in aspects of the representations.

However, at least two, at least superficially different, routes to layered representations have been investigated. One is discriminative or bottom-up, concerned directly with the way that input images are represented and rerepresented at levels within the sort of anatomical hierarchy that is apparent in successive areas of visual cortex. The second is more complicated, since it involves the top-down connections that are known to coexist with bottom-up ones.

Bottom-up: The bottom-up approach to hierarchical representation has reached a recent apogee in the popular work on deep learning (Hinton, Osindero, & Teh, 2006; Krizhevsky, Sutskever, & Hinton, 2012; LeCun, Bengio, & Hinton, 2015). Here, the standard idea is to solve a supervised learning task, such as recognizing hand-written digits or faces or scenes, by mapping input through a series of nonlinear transformations. Each such transformation creates a new representation from old, in the service of improving recognition. There is a range of methods for tuning these transformations given labelled examples (of which various forms of the back-propagation learning rule are currently most popular).

These bottom-up labelled approaches are known as discriminative, since they solve the supervised task of discriminating between the labelled images. Other approaches compatible with bottom-up criteria one can adopt, include unsupervised criteria (such as sparseness or efficient coding), that have been a basis for modeling lower level vision (Bell & Sejnowski, 1995; Olshausen et al., 1996; but see also Zhaoping, 2014).

For supervised discriminative networks, the fact that input might come from distributions generated according to the hierarchical principles adduced above is formally irrelevant—all that actually matters is being able to solve the supervised task. Take the example of faces. Faces have a clear hierarchical part-whole structure, containing for instance distinct eyes, a mouth, and a nose. However, as discussed in the experimental section, human face recognition does not appear to require parsing out the parts (Jiang et al., 2006; Tanaka & Farah, 1993). This might be due to the observation that faces are typically encountered as a whole, and we rarely see face parts in isolation—a factor that is likely to influence the supervised learning process, favoring a representation in which parts are bound together.

Nevertheless, solving the supervised task will often require an implicit coding of aspects of generation. This is because conventional supervised tasks respect the nature of generation. Consider, for instance, recognizing a digit hand-written in pens with different colors. Thus (re-)representations that lead to high-quality performance with ready generalization to new examples will be favored by the learning process. In this example, high performance can be achieved invariant of the color. Therefore, color is effectively separated, though not necessarily discarded, from the digit form. Other tasks may still require coding of the color information.

Further, in practice, much use has been made of ways of turning some assumptions about the constraints as to the way that images are generated into features of the bottom-up architecture. Take the case of spatial location. There is obvious freedom as to where objects can appear. One would like to be able to learn how to recognize them in just one location, and generalize this knowledge across the image, rather than having to learn separately in every location. This has been made to work by arranging that at least some of the representational layers perform spatial convolution—effectively discarding particular aspects of the dimension of spatial variation or, equivalently, sharing structure between the representational transformation across different parts of the scene, and thereby allowing explicit generalization (Anselmi et al., 2014; Fukushima, 1980; Hinton, Krizhevsky, & Wang, 2011; Hubel & Wiesel, 1959; Jarrett, Kavukcuoglu, Ranzato, & LeCun, 2009; Riesenhuber & Poggio, 1999). This could lead to an imperfect dual to the workings of a component hierarchy, in which the idea for recognition was to separate out the two characteristics of the image: the object itself, and the location in the scene where that object is placed. The imperfection arises if information about the location is discarded. However, experimental and simulation work by Hong, Yamins, Majaj, and DiCarlo (2016) and by Golomb and Kanwisher (2011) have shown that position information may be *distilled* at a population level as we move up the hierarchy.

One putative characteristic of discriminative part-whole representations is that of microfeatures (Hinton, 1984, 1990; McClelland, Rumelhart, & Hinton, 1986). Microfeatures are intended to allow for partial similarity in distributed representations and can be used to perform inferences that lead to binding. In reuse, microfeatures are meant to address the paradox that a whole (e.g., a wing) when it is the whole of the image should be represented both similarly and differently when it is a part of something more substantial (e.g., a bird). The similarity is necessary to exploit generalization. The difference is necessary because the semantics are quite different.

Top-down: The top-down approach to hierarchical representation is directly tied to the generative structure

of the patterns listed in the previous section. The notion is that successive top-down layers parameterize the various top-down stages of generation, so that the individual generators (often parts or components) are explicitly captured in activity patterns. One set of key roles that such a model plays is as the substrate of imagination, reconstruction from memory, directed top-down attention, higher level hypotheses regarding object or scene properties helping to resolve perceptual ambiguities in the input (Kersten, Mamassian, & Yuille, 2004; Weiss, Simoncelli, & Adelson, 2002; Yuille & Kersten, 2006), and certain cases of iterative recognition in which constraints have to be propagated from one part of an input to another, or cases in which direct sensory and central information (e.g., from memory or other modalities) must be combined for appropriate recognition.

The other key role of the top-down model is providing a set of targets for bottom-up processing. That is, they define the representational scheme that bottom-up processing should optimally realize to take new inputs and represent them in terms of their particular generators. This form of paired top-down and bottom-up processing is often called analysis by synthesis (Hinton & Ghahramani, 1997; Hinton & Zemel, 1994; Hinton, Dayan, Frey, & Neal, 1995; Neisser, 1967; Von Helmholtz, 1867; Yuille & Kersten, 2006). It is consistent with the partnering principles described earlier; for instance, a part-whole generative process could be paired with bottom-up binding. Bottom-up processing could be on-line, with only the top-down model being parameterized, and recognition being its calculated inverse. Alternatively, this inverse could be compiled or distilled in the form of a conventional bottom-up discriminative model (Dayan, 2006; Hinton et al., 1995; Hinton, Vinyals, & Dean, 2015). Developments and elaborations of these are found in purely unsupervised large-scale hierarchical autoencoder networks (e.g. Le, 2013; Ranzato, Huang, Boureau, & LeCun, 2007), which perhaps offer the most direct coupling between generation and recognition.

Top-down computation can help resolve ambiguities for image interpretation, by identifying objects and parts, from bottom-up computations followed by top-down refinements (Epshtein, Lifshitz, & Ullman, 2008). They can also propagate information, for instance, regarding attention that is relevant to bottom-up discrimination (Cao et al., 2015). In addition, hierarchical generative models that learn rich priors over the Lambertian components could then be used for recognition by the partnering principle of separation, achieving good generalization in face recognition under illumination variations (Tang et al., 2012).

In neural terms, it has been conventional to identify the generative characterization with top-down (and perhaps) lateral or horizontal weights in the cortical hierarchy, and identify recognition with bottom-up processing. However, bottom-up hierarchies living within the constraint of a limited anatomy can cope with the size or complexity of a scene by swapping space for time, and exploiting memory (Hochreiter & Schmidhuber, 1997; Mnih, Heess, Graves, & Kavukcuoglu, 2014). That is, they can accumulate information over multiple snapshots to form a more complete picture.

Since generation-based models also often employ a bottom-up recognition network, albeit with a more transparent logic (e.g., Zeiler, Taylor, & Fergus, 2011), they are generally susceptible to the same tests of hierarchical structure as bottom-up ones. However, since these models provide a means of generating stimuli, they differ from purely bottom-up approaches in that they are further susceptible to experimental tests that require some form of reconstruction or imagination.

Experimental approaches

It has been a very general experimental goal to examine the nature and structure of representations. Three major classes of tasks in which this has been done are recall from short-term or long-term memory, adaptation, and on-line inference (i.e., forms of recognition, discrimination, or segmentation).

In turn, four particularly significant measurement modalities have been used to examine hierarchical representations. Three of these are behavioral. The first is co-determination of the fate of parts of inputs in memory and inference. Co-determination is a broad topic that we expand on considerably, including the following: biases that are prevalent both in the adaptation and memory literature, whereby adaptation to or memory of one part of an input biases the perception or recall of another part or of a whole associated with that part; mnestic co-determination, which is focused on how parts of an input are forgotten; and segmentation co-determination, since how we segment hierarchical stimuli may be indicative of parts that make up the whole. The second behavioral measurement focuses on errors of recall of whole inputs, and the third on reaction times for recall and inference.

Such behavioral measures are inevitably incomplete—that is, it will typically be impossible to determine whether the effects observed arise directly because of the representation itself, or just indirectly as a form of inference performed on a different representation. Further, there are many possible hierarchical representations of parts and features that can be used to represent objects and scenes, and fairly limited

understanding of what the visual system actually uses (e.g., beyond primary visual cortex). Nevertheless, the behavioral measures will at least show the constraints that are embodied in the representational computations that are performed.

The fourth form of measurement more directly gets at representation, and involves neurophysiology recordings, or representational neuroimaging (to use a term from Behrens encompassing at least cross-stimulus suppression; Boorman, Rajendran, O'Reilly, & Behrens, in press; Grill-Spector, Henson, & Martin, 2006; Klein-Flügge, Barron, Brodersen, Dolan, & Behrens, 2013; and representational similarity analysis; Kriegeskorte, Mur, & Bandettini, 2008).

These assessments can all be made on the hierarchical structure of an existing representational scheme, or one can look at hierarchies associated with novel input statistics. In both, the assumption is that the hierarchies arise in the light of the structure of the statistics in the world (for the generative, top-down, origin) or the requisite tasks (for the discriminative, bottom-up, origin). Existing hierarchies will presumably be based on the statistics of, and tasks implied by, the normal visual environment, whereas we have more freedom in designing novel hierarchies to ask focused questions and test hypotheses. An example of a novel hierarchy is a set of "parts" (e.g., abstract colored circles, or brush strokes), along with the combination rules and statistics by which more complex hierarchical structure is formed (e.g., colored circles abstractly combine into "wholes," or brush strokes combine into novel letters).

Existing and novel hierarchies have their own advantages and disadvantages. For existing schemes, as its name implies, we have the potential to tap into the existing hierarchical organization of cortex, which is reinforced through evolution and experience with the natural environment. But we lack an independent way of assessing the exact statistics that an individual has encountered, implying that hypotheses might be underconstrained. Novel schemes, on the other hand, provide control over the statistics. The representations may thus be studied in tandem with the learning processes by which they arise. However, learning becomes a factor, potentially making it hard to disentangle what is unique and what was existing.

In the next section, we discuss experimental work across the four measurement modalities, and what they imply (often indirectly) about hierarchical representations.

Co-determination

A range of studies has probed hierarchy experimentally by examining the consequences of the existence of interrelated parts and wholes on various forms of judgment. Such studies span a number of task domains, including memory recall, and other forms of processing such as visual adaptation. Coarsely, if some aspect of a task (e.g., adaptation or forgetting) affects the activity of a representational unit, then it will affect decisions, storage, and recall of all aspects of the input that share this representational unit either directly, of through the hierarchical construction of a representation. Thus, the adaptation to or memory of one part will influence or bias perception or recall of another part, or of a whole associated with that part. Which aspects of the input are jointly affected indicates how assessments of the proximity between stimuli are influenced by different parts of the input—i.e., binding and separation.

Consider, for example, the case of representing images of faces. The part-whole structure appears straightforward—the face contains at least eyes, a nose, a mouth, plus various other structural features. Co-determination might arise at various levels: at a lower one because of characteristics such as shared illumination; and at a higher one from such effects as ethnicity. If the hierarchical representational scheme for faces reflects the fact that the two eyes are usually co-determined by combining them together under one unit, then any influence associated with one eye (such as attractive or repulsive biases, which we describe in the next subsection below), should affect the other eye too. A direct example of this, albeit in a different domain, is the observation that the influence of one part (a face) can transfer to another part (body appearance) (Palumbo, DAscenzo, & Tommasi, 2015). Equally, if monobrows (synophrys) are more common for some groups of people than others, and covary with other facial features within the groups, then these structures might be separated into a set of representational units, and then potentially influence aspects of perception of other characteristics of new faces with monobrows, reflecting the structured influence of this component.

Co-determination can present itself in a number of ways, which we discuss in the next subsections.

Biases

Bias typically refers to nonveridical perception or recall.

Attractive biases are more prominent in the memory and inference paradigms. They typically arise from some sort of reversion to the mean, a manipulation that can often be given a normative Bayesian explanation in terms of priors (Raviv, Ahissar, & Loewenstein, 2012), and could arise from a form of priming.

Repulsive biases commonly arise in aftereffect paradigms (Clifford & Rhodes, 2005). In these, one is typically adapted to a stimulus (such as a diagonal

grating) for a period of time, and this causes subsequent stimuli that are vertical to appear biased away from the adapter (i.e., tilted diagonally the other way). Repulsive biases also occur at higher levels, such as adapting to a sad face, and then observing that a neutral face appears happy. Indeed, repulsive biases are ubiquitous, and occur in a variety of aftereffects spanning low-level visual features such as contrast and orientation, and high-level stimuli such as objects, faces, and even scenes (Clifford & Rhodes, 2005; Greene & Oliva, 2010). It has been hard to account for repulsive biases in terms of a prior (see Schwartz, Hsu, & Dayan, 2007). We next discuss each of repulsive and attractive biases in more detail as they pertain to hierarchy.

Various studies have examined co-determination in the context of repulsive biases for low-level features of color, luminance, and tilt. One main question in these studies has been whether features (such as color and tilt) interact, or are represented independently (i.e., are separated). To get at this question, Clifford, Pearson, Forte, and Spehar (2003) and Clifford, Spehar, Solomon, Martin and Qasim (2003) considered repulsive biases both for adaptation in the tilt aftereffect, and for its spatial counterpart of the tilt illusion. For instance, in the tilt aftereffect, adapting to a grating oriented to the left of vertical, leads to a repulsive bias in which the perception of a vertical test grating appears to the right of vertical. If the tilt aftereffect is selective to chromaticity (i.e., adapting to orientation along one color axis brings about repulsion for an oriented test grating in the same color axis, but not to a test grating in an orthogonal axis), then this might be indicative that color and orientation are represented together. If the tilt aftereffect is invariant to chromaticity (i.e., adapting to orientation along one color axis brings about repulsion for an oriented test grating along any axis), then this suggests that orientation might be represented independent of color.

Clifford, Pearson et al. (2003) and Clifford, Spehar et al. (2003) found that there is most repulsion when adapter and test are matched in chromaticity, particularly in the tilt illusion but also in the tilt aftereffect, suggesting that color and tilt might partly be represented together. This might also reflect the notion that objects are smooth in their statistics, so orientations of similar color in center and surround locations (in the tilt illusion), and possibly across time (in the tilt aftereffect), might be interpreted as parts being bound together as the same whole object (Qiu, Kersten, & Olman, 2013; Schwartz et al., 2007, 2009). However, they also found repulsive biases for orthogonal color axes, suggesting some invariance between color and orientation. This invariance may be interpreted as relating to component hierarchy, and separation between color and form. Earlier work and potential cortical correlates are discussed in Clifford, Pearson et al. (2003) and Clifford, Spehar et al. (2003).

One caveat in the interpretations of being bound together as the same object or separation in terms of color and form is that the experimental design usually presents a range of features (such as color and orientation) at different times or spatial locations, but rarely manipulates or systematically investigates natural or artificial statistical regularities between these features. It is interesting to consider whether one could design a more comprehensive set of test cases based more directly on the analysis of image statistics and parts that are typically bound together. One could also consider teaching new generated artificial stimuli, for which the bottom-up partnering principle of binding and separation can be made more explicit. One could then test for adaptation, although this may be difficult given the strong repulsive biases that already exist.

The discussion thus far has been on transfer of adaptation with parts. Another route has been to study adaptation of a lower level visual feature, and ask if this adaptation transfers to a higher level whole. Such studies aim to address hierarchical representation from a more bottom-up perspective: What parts are potentially transferred from a lower level to a higher level? For instance, Xu, Dayan, Lipkin, and Qian (2008) examined how adaptation to low-level curvature, or to the shape of a cartoon mouth, affects perception of facial expression. They found that the lower level adaptation resulted in a facial expression aftereffect (provided that there was positional specificity). This suggests that adaptation can be inherited from lower levels at higher levels of the cortical hierarchy. Note that here we use the term inheritance as common in this literature, but are not referring to the semantic inheritance described in the *Sources of hierarchical structure* section. Dickinson, Almeida, Bell, and Badcock (2010) found that low-level adaptation to tilt can result in global shape aftereffects. Further studies by Xu, Liu, Dayan, and Qian (2012) manipulated the stimuli in a way that could dissociate low- and high-level effects, and showed that part of the adaptation was inherited at the higher face level, but that part was created de novo.

Along with these powerful repulsive biases arising from adaptation are a set of cases in which attractive biases are apparent. This has been prominent in studies of memory, in which properties of the stimulus ensemble, such as the mean size of an expected category of objects, attractively biases recall (Brady & Alvarez, 2011; Huang & Sekuler, 2010; Wilken & Ma, 2004). Bias to ensemble statistics is not unique to memory recall, and has also been reported in perceptual studies (Konkle & Oliva, 2007; Raviv, Lieder, Loewenstein, & Ahissar, 2014). One study examined the error in judging ensemble statistics of a group of stimuli, across

both low-level visual stimuli such as orientation and color, and high-level stimuli such as facial identity and expressions (Haberman, Brady, & Alvarez, 2015). They found that low-level ensemble representations (e.g., orientation and color) were correlated with each other in terms of individual subject errors, but not low-level with high-level ensembles. This suggested that the relationship between ensemble representations depends on how close they are (qualitatively) along a representational hierarchy.

Of relevance to semantic inheritance are results from Hemmer and Steyvers (2009b), who measured the recall of sizes of objects of fruits and vegetables. These were found to be biased according to the mean ensemble statistics of particular categories in a way that distinguished two levels of a hierarchy: broader categories such as fruit and vegetables; and narrow ones associated with individual objects. Familiar objects were comparatively biased towards their particular object categories, whereas unfamiliar objects were more biased towards the broader (e.g., vegetable) category, thus showing a form of reversion or smoothing associated with inheritance.

Mnestic co-determination

Co-determination may also be present in the way that parts or components of an input are forgotten, or, concomitantly, variability and covariability in the way that they are reconstructed, having been forgotten.

A range of studies has considered recall of multiple stimulus features from memory. These have focused on whether features are represented independently in memory, or whether they are represented dependently (which is sometimes referred to as "bound units" in this literature, although note that this corresponds to binding of features). A common finding has been that features are represented independently, or separated according to the partnering principle for component hierarchies. For instance, Bays, Wu, and Husain (2011) asked subjects to recall both color and orientation features of an artificial object (a bar) in visual working memory. The errors in the feature dimensions were found to be independent, so, for instance, there was no advantage in recalling the orientation of an object whose color had been correctly recalled. Equally, Brady, Konkle, Gill, Oliva, and Alvarez (2013) examined memory performance across time for various object features including color, orientation, object state (e.g., open, or closed states), and exemplars (e.g., "ornate wooden door" or "plain metal door"). Forgetting curves suggested again that features are not represented together, since the quality of memory recall decreased at different rates for different features. In particular, recall of color deteriorated more rapidly over time than other features. Brady et al. (2013) point out that the property of change in state relates to a change in the configuration of the object parts, though they also note that the study does not directly get at whether object parts are forgotten separately. Similar results of independent feature representation have been suggested in other studies (Fougnie & Alvarez, 2011) although there is some evidence for dependence of features in the representation (Quinlan & Cohen, 2011).

It is here that issues with the statistical structure of the input—i.e., the features and objects, are most obviously problematical, motivating work both on natural stimuli (Orhan & Jacobs, 2014) and novel structured representations (e.g., Brady & Tenenbaum, 2013; Brady, Konkle, & Alvarez, 2009, 2011). Color is a good example: For some objects, color and form are strongly linked (e.g., yellow bananas), and so plausibly share generative structure and thus are likely to be represented together. For other object classes, the link is statistically much weaker (tables come in many different colors, for instance), thus more likely to mandate separation. Therefore, short of either analyzing the stimulus statistics of ensembles (such as natural stimuli; see also Orhan & Jacobs, 2014), or teaching novel but controlled statistics, it is difficult to know if and how such results generalize.

There have been some studies asking whether objects in images are represented or bound together. These have mostly been manipulations of pairs or groups of objects within images, chosen such that they are congruent or incongruent in scenes. Such choices are typically justified intuitively or qualitatively, rather than through statistical analysis. Issues of congruency have been of interest both in perceptual and memory studies. In a classical perceptual study, Biederman, Mezzanotte, & Rabinowitz (1982) defined five violations of an object with respect to its background: support in which the object appeared floating; interposition in which the background appeared to pass through the object; probability of the object is unlikely, such as a fire hose in a kitchen; position unlikely; and size unlikely. For most cases, the detection of objects was less accurate and slower when there was a violation in the image. A number of studies have suggested that congruence between objects (e.g., one expects an oven and fridge to appear together in a scene), or between objects and backgrounds in scenes (a tree and a forest), leads to less error (and also faster reaction time) in memory recall of another object or the background (Davenport, 2007; Davenport & Potter, 2004; Joubert, Fize, Rousselet, & Fabre-Thorpe, 2008; Joubert, Rousselet, Fize, & Fabre-Thorpe, 2007; Potter, 2012). This improved recall could be a consequence of the deployment of semantic memory at the time of recall, but it could also result from the representation employed at the time of encoding incorporating knowledge of this congruent structure and acting as a

Bayesian prior for parts that are expected to be grouped (Hemmer & Steyvers, 2009a; Steyvers & Hemmer, 2012). Clever experimental designs can distinguish these possibilities (Steyvers & Hemmer, 2012).

Only few papers have employed artificial stimulus ensembles with controlled statistics. One experiment continued the line of studies noted above for recall, considering whether parts that have statistical regularities are represented as bound units in working memory. Brady et al. (2009) designed an artificial circle composed of two colors, with an inner circle and an outer ring, making up an object whole (like parts of an object); or two circles side by side, each of different color, making up a whole consisting of two objects (like object parts of a scene). Some colors appeared together more frequently, introducing a simple statistical regularity. In the control case, all color appeared randomly, similar to most work in this area. This is one of few papers that introduced a statistical regularity in the inputs that subjects learned to expect. Learning the regularities between the colors improved working memory performance, and was compatible with a Bayesian learning model (see also Brady & Tenenbaum, 2013). This signified that parts may be represented together in working memory as bound units.

In a rather different approach, some studies have generated random displays and considered collective rather than singleton recall. This condition perhaps invites subjects to create one or more chunks on the fly. For instance, Orhan and Jacobs (2013) asked subjects to remember a feature value, such as the horizontal locations of *all* displayed square stimuli. This allowed them to examine whether there are dependencies between feature value estimates across the set of square stimuli in the display. They found correlation between the feature value estimates that was highest when the square stimuli were spatially nearby and when they had similar horizontal location. They also found biases in recall to the mean horizontal location. The results were explained in the context of a probabilistic clustering model, in which representations of items belonging to the same cluster share parameters, and thus are dependent. Brady and Alvarez (2015) generated random sets of colored circles and considered collective recall of the displays. Here the proposal was that some random configurations might by chance be better recalled than others. For instance, Brady and Alvarez (2015) give the examples of warm colors that happen to be on one side of the display and cold colors on the other side. They found that subjects were indeed highly consistent in which displays were hardest or easiest to remember. They suggested that this was captured by a model that includes clustering of groups of colors hierarchically, and keeping track of the mean and variance of clusters. These studies provide examples of low-level (rather than semantic) collective perceptual hierarchical grouping.

A different facet of co-determination is to consider how forgetting of a part depends on its relationship to the whole. In terms of a hierarchical generative model, this might provide some insight into representational relativity—how parts are coded relative to wholes. This has certainly been studied—even at the level of relatively fast encoding (on the order of 35 seconds to a minute) by subjects of the contents of an actual room (Brewer & Treyens, 1981; Pezdek, Whetstone, Reynolds, Askari, & Dougherty, 1989) or the congruence between words and context (Craik & Tulving, 1975; Schulman, 1974). However, there have again been conflicting results. For instance, Pezdek et al. (1989) found that inconsistent objects in rooms actually resulted in better memory recall.

One potential reason for such confounding results is attentional. The input stimuli have rich hierarchical structure. However, this makes it hard to control the allocation of attention when such stimuli are presented. For instance, Loftus and Mackworth (1978) found that more fixations are made to novel objects in scenes. Recent work has tried to explicitly control attention to schema consistent or inconsistent objects via task instructions, while recording eye movements, and then testing for memory (Silva, Groeger, & Bradshaw, 2006). They found that attention (and correspondingly, more eye movements) was important for remembering low-level object properties and recalling schema inconsistent objects, but not for recalling schema consistent objects (see also similar results in Santangelo, 2015, and review paper of Coco, Malcolm, & Keller, 2014). This suggests indirectly a hierarchical organization according to schema consistency.

The relationship between a whole and its parts has also been studied for faces. Although faces have a clear hierarchical structure, face recognition does not appear to rely on actually parsing out the parts. Rather, the parts appear to be bound or represented together. Tanaka and Farah (1993) asked subjects to memorize faces, and showed much less accurate recognition of face parts presented in isolation ("Which is Larry's nose?"), than recognition of the parts when whole faces were presented. Computational models have been able to quantitatively fit such behavioral face data, without relying on explicit part-based representations (Jiang et al., 2006; Riesenhuber & Wolff, 2009). This was in contrast to other object classes such as houses (Tanaka & Farah, 1993), for which the parts recognition was not compromised. As noted in the Introduction, this may be related to the learning process, by which we typically encounter faces as a whole.

Segmentation co-determination

Understanding the rules of how we segment hierarchical stimuli may also be indirectly indicative about the co-determination of parts that make up whole representations. Indeed, Brady et al. (2009) motivate their synthetic correlated visual color stimuli for working memory recall, by referring to literature on segmentation. They give the example from language of the sequence FBICIA, for which FBI and CIA are better recalled because they are associated with and segmented from each other. They contrast this with random chunks (as in HSGABJ), which are more difficult to recall.

Questions about segmentation have been addressed extensively in the statistical learning literature, focusing on how infants and adults learn to segment syllables in language (Aslin & Newport, 2014; Marcus, Vijayan, Rao, & Vishton, 1999) or to segment hierarchical arrangements of visual shapes (Fiser & Aslin, 2002; Orbán, Fiser, Aslin, & Lengyel, 2008). These studies have generated artificial stimuli with controlled statistics and asked whether infants and adults can use statistical cues (such as at its simplest, transition or co-occurence probabilities) to learn which parts are likely to be grouped together hierarchically as words or chunks and segmented from the whole. Here the parts are designed by the experimenter, so this approach can reveal how parts are grouped together, but does not address harder segmentation problems in scenes.

A different approach to studying segmentation has focused on the problem of figure-ground organization in images. Some studies have argued that access to local visual cues such as convexity, and to local component cues such as luminance, provide information for figure-ground discrimination (Fowlkes, Martin, & Malik, 2007). Moreover, Ren, Fowlkes, and Malik (2006) show that enforcing global consistency in this bottom-up approach provides a significant account of performance over local cues alone.

Errors of recall of whole inputs

Some studies have focused on forgetting of whole images and asked indirectly what we can conclude about the proximity of images that are forgotten. These studies are less direct than the work on co-determination we mentioned above, but may still be revealing about representation.

One striking aspect of recall of whole images is the observation that we can apparently recall whether or not we have seen one out of thousands of scenes (Konkle, Brady, Alvarez, & Oliva, 2010a, 2010b; Standing, 1973), or one out of thousands of objects (Brady, Konkle, Alvarez, & Oliva, 2008). These studies have concluded that memory for both scenes and objects is "more detailed than you think" (see title of Konkle et al., 2010a). At question is what we can learn from the remaining failures.

One way that might show promise is to make subjects believe that they have experienced something that they have actually not. Such approaches were pioneered in the case of language, with the Deese-Roediger-McDermott (DRM) paradigm, for which lists of words associated with a word that was not shown, constituted a lure of false memory (Roediger & McDermott, 1995). In vision, a way to induce false memory (known as visual inception) is to ask subjects to remember a collection of patterns that are each close to a pattern that is not, in fact, presented (Khosla, Xiao, Isola, Torralba, & Oliva, 2012). For instance, Khosla et al. (2012) presented scenes that have similar gist and geometry to the actual scene that was previously shown, and suggested, based on informal experiments, that this leads to visual inception. By determining a proximity metric, this approach has the potential to be informative about representation. This could indirectly be indicative about hierarchical representation, as such proximity might imply properties of shared parts (presumably at a lower level of the hierarchy).

Another example of false memory is the boundary extension phenomenon (Intraub & Richardson, 1989). In this case, subjects remember a greater extent of the scene than actually shown, presumably based on imagination of the expected surrounding (i.e., relating to top-down, generation based ideas). This may be thought of as indicative of the parts that make up the scene, beyond the boundaries of what is shown. Amnesic patients fail to exhibit boundary extension, and so are more veridical in their recall (Mullally, Intraub, & Maguire, 2012).

Other studies looking at our remarkable capacity for memorability have also asked related questions. For instance, by introducing foils in the experiments that have similarity to the targets, Konkle et al. (2010a) found that conceptual similarity led to interference as more exemplars were shown from a stimulus category—but not perceptual similarity of color and shape. This suggests that the key representations were appropriately more abstract.

However, in terms of forgetting, the more straightforward suggestion, which is dual to inception, is the idea that there might be a simple relationship between the length of the code in bits needed to describe a given pattern, and forgetting. In this case, the quality of recall could be used to quantify at least the size of the representation. In the context of a discriminative representation, something similar could be true about the total extent of sensitivity of the representational units to the pattern.

Distinctive images are often better remembered than conventional ones (Franken & Rowland, 1979; Levie & Hathaway, 1988; Standing, 1973); something that was carefully quantified by Bylinskii, Isola, Bainbridge, Torralba, and Oliva (2015) in the context of image ensembles. This result suggests indirectly that when the arrangement of parts is unexpected or surprising, this situation may result in better recall. However, along with these advantages for distinctiveness, we are also impaired at remembering *very* unlikely inputs—as evidenced, for instance, by the observation that chess masters' highly superior memory for board positions only extends to those that could plausibly have occurred during a game, rather than random positions of the same number of pieces (Chase & Simon, 1973). Thus, it seems unlikely that there should be a simple, monotonic, relationship between representation size and memorability, making it hard to draw conclusions from forgetting of whole scenes. In addition, the issue about differential attention once again makes such assessments tricky. In the context of multiple items in a collection, problems may also arise from online organization and reorganization in memory.

Reaction times

Even in cases in which subjects do not make substantial biased or unbiased errors, reaction times can be revealing. Binding, for instance, is known to influence the timing characteristics of visual search. Similarly, cross-stimulus priming, reflected in the reaction times for processing a stimulus in the light of preceding stimuli, could be revealing in a similar manner to cross-stimulus transfer in adaptation.

A more particular timing issue that is relevant to a tree-like or other structurally extended hierarchical representation is that if one truly needs to traverse the structure to make inferences or to recall information, then the reaction time might increase. This has been studied more in language and semantic hierarchies where there remain substantial uncertainties (Holyoak, 2007), but is also applicable to vision hierarchies. Consider, for instance, a study by Murphy et al. (2012), who generated a set of artificial stimuli according to a hierarchy, such as artificial bugs with parts and textures. These bugs also belonged to categories and received artificial names. Murphy et al. then measured reaction times for answering questions about properties pertaining to the categories that could be derived from the hierarchy. The idea was that if subjects formed a hierarchical representation of the stimuli, then reaction times would be longer if one had to traverse the entire tree.

Murphy et al. (2012) used several different approaches to introduce the visual stimuli to subjects. In the first experiment, the hierarchy was not explicitly taught and categories not explicitly learned. In the second experiment, they taught subjects the category names for the bugs at the different hierarchy levels, but did not show subjects the hierarchy explicitly. In the third experiment, they used the same artificial bugs, but now subjects were shown a schematic of the hierarchical generating tree structure. Finally, the last experiment taught subjects pairwise associations, following explicitly the hierarchy structure. With only a limited exception in the last experiment, no difference in reaction time relative to location in the hierarchy was found, suggesting that for the most part subjects avoided learning a full tree hierarchy. Murphy et al. (2012) noted that the learning time in their experiment may have been short, and that possibly learning over days would increase the representation of a tree hierarchy—the problem being that this would then also afford ample means and opportunity to learn non-hierarchically dependent answers to the questions employed.

Representational neuroimaging and neurophysiology

More direct measures of hierarchical representation can be obtained with recent approaches in representational neuroimaging, notably representational similarity analysis (RSA; Kriegeskorte et al., 2008) and cross-stimulus repetition suppression (e.g., Boorman et al., 2016; Klein-Flügge et al. 2013), along with neurophysiological measurements.

Cross-stimulus repetition suppression

Cross-stimulus repetition suppression is the fMRI equivalent of cross-stimulus transfer in adaptation. fMRI repetition suppression is the long-observed phenomenon that stimuli elicit lower amplitude BOLD responses across many brain regions when they are repeated than when they are first presented. Cross-stimulus repetition suppression involves presenting one stimulus to induce suppression, but then testing a *different* stimulus. The idea is that the greater the degree of suppression of the BOLD response in some area to the second stimulus, the more it shares a representation with the adapting stimulus at that locus. Thus, it could be possible to work out elements at least of the similarity structure of internal hierarchical representation.

Recent work, reported in scientific abstract, has focused on fMRI cross-stimulus transfer for low-level visual stimuli related to the psychophysical cross-stimulus questions of Clifford, Spehar et al. (2003) and Clifford, Pearson et al. (2003). Kuriki, Xie, Tokunaga,

Matsumiya, & Shioiri (2014) found cross-adaptation effects between color and luminance motions in the BOLD activity of most visual areas tested (and also behaviorally), suggesting perhaps an invariant representation of form from color and luminance. Chang, Hess, Thompson, and Mullen (2014) tested for cross-transfer of achromatic and chromatic contrasts across different neural areas. They found that for only one area they tested (hMT+) but not for other areas, adaptation to achromatic contrast affected test stimuli that are either achromatic or chromatic, suggesting invariance to achromatic or chromatic contrasts. This might again relate to the partnering principle of separation of color from form as in a component hierarchy, although it is not made explicit.

Representational similarity analysis

Representational similarity analysis is described in (Kriegeskorte et al., 2008). It considers the similarity structure at various levels of cortical processing among the activity evoked by a collection of stimuli—up to noise and coarse sampling. These similarities are intended then to be exactly revealing of the representation. For instance, they considered images of faces and objects, and characterized brain regions that exhibit similar activity along some dimension (such as faces), versus areas that do not.

Recent work has started to make interesting links between representational similarity analysis approaches, bottom-up deep neural networks, and hierarchical representations. For instance, Khaligh-Razavi and Kriegeskorte (2014) suggested that supervised but not unsupervised learning in deep neural networks, leads to similarity representations that are closer to neural population representations in object processing areas of inferior temporal cortex. In Khaligh-Razavi, Henriksson, Kay, and Kriegeskorte (2014), they consider the relation between the similarity metrics in the fMRI and feed forward models of the ventral stream. These approaches have intriguing potential to link with hierarchical representations across several levels of the hierarchy.

Recent neurophysiology studies have also made strides in linking between the responses of deep convolutional neural networks and the responses of neurons along the ventral stream (Güçlü & van Gerven, 2015; Yamins et al., 2014). For instance, Yamins et al. (2014) used a high-throughput method to select from a class of deep convolutional networks with different parameters. They found that deep networks that are more optimized for object recognition show more similarity to inferior temporal cortex neurons. They also found that the highest output level of the selected deep networks trained on object recognition and was predictive of neural responses to images in inferior temporal cortex. The mid-levels of the network were more predictive of visual area V4 (see also Pospisil, Pasupathy, & Bair, 2016). Yamins et al. (2014) suggested that top-down constraints on object recognition performance may be important for shaping mid-level area representations.

Changes along the cortical hierarchy

Other neurophysiology studies have also followed a bottom-up hierarchical perspective, asking what aspects of the representation change as one proceeds along the hierarchy. We focus on neural areas along the ventral stream beyond primary visual cortex (V1). We discuss briefly some aspects that have been learned in single neural areas along the ventral stream, and where relevant studies that have explicitly compared across neural areas. A usual caveat is the difficulty in knowing what stimuli are appropriate for probing a given neural area along a hierarchy. Nevertheless, this is a more direct way of studying neural representations and can be suggestive about how "parts" in a given neural area contribute to "whole" representations in higher areas.

Neurophysiology studies in secondary visual cortex (V2) have shown selectivity to combinations of orientations, such as corners and junctions (Ito & Komatsu, 2004). Unsupervised learning approaches have resulted in models of secondary visual cortex that combine V1-like units, leading to such corner selectivity and other phenomena (see, e.g. Coen-Cagli & Schwartz, 2013; Hosoya & Hyvärinen, 2015; Lee, Ekanadham, & Ng, 2008; Malmir & Ghidary, 2009). Studies in humans and macaque have suggested that V2 is sensitive to textures (Freeman et al., 2013; Ziemba, Freeman, Movshon, & Simoncelli, 2016) and more complex features of images (Willmore, Prenger, & Gallant, 2010). Freeman et al. (2013) found via neurophysiology and fMRI that neurons in visual area V2 but not V1 were selective to texture stimuli synthesized according to Portilla and Simoncelli (2000), suggesting that V2 represents texture structure in scenes. Other neurophysiology (Williford & von der Heydt, 2016; Zhou, Friedman, & Von Der Heydt, 2000) and modeling (Zhaoping, 2005) studies have suggested that single units in area V2 contribute to border ownership.

Area V4 is a mid-level area that has been studied considerably (for a review, see for example, Kourtzi & Connor, 2011). We note briefly some aspects that relate to representation. Studies in area V4 have revealed curvature and shape selectivity, even in the face of occlusion (Kourtzi & Connor, 2011; Pasupathy, 2015). Area V4 has shown selectivity to surface features and boundaries, including chromatic boundaries and shapes (see review paper of Roe et al., 2012). Area V4 therefore is thought to contribute to figure ground

segregation and segmentation (Pasupathy, 2015; Roe et al., 2012). There are also suggestions in the experimental literature that some classes of textures are better represented in visual areas V3 and V4 (and not V2 or V1) (Kohler, Clarke, Yakovleva, Liu, & Norcia, 2016; Okazawa, Tajima, & Komatsu, 2015). Studies in V4 have also found evidence for color constancy (Kourtzi & Connor, 2011), relating to component hierarchies. In the context of component hierarchies, note that in area V1, there are studies showing selectivity of some neurons to brightness and to global illuminant (Kinoshita & Komatsu, 2001; Rossi, Rittenhouse, & Paradiso, 1996), and evidence for lightness constancy (MacEvoy & Paradiso, 2001).

Other studies have focused on comparing high- and mid-level visual areas along the ventral stream. Rust and DiCarlo (2010, 2012) asked how well neural populations in two levels of a hierarchy can discriminate and generalize across images and "scrambled" images (Portilla & Simoncelli, 2000) that only retain more local statistics. They found that higher visual areas (inferior temporal cortex) could better discriminate between regular and scrambled images than mid-level areas (visual area V4), suggesting that higher areas along the ventral stream are more sensitive to conjunctions in natural images. Rust and DiCarlo (2010) also found increased tolerance in inferior temporal cortex, in terms of the ability of the population to generalize the same image presented at different positions, scales, or backgrounds. Hong et al. (2016) suggested that "category-orthogonal" object properties (position, size, and pose) in scenes are better represented in inferior temporal cortex than in earlier areas, and more predictive of human performance.

There is neurophysiological evidence for a cortical area consisting of face-selective neurons (Tsao, Freiwald, Tootell, & Livingstone, 2006). This may link back to the observation that faces are typically encountered as a whole. This appears to be in contrast to other modalities such as letters and words, for which there is evidence of a part-whole relationship in cortex (Vinckier et al., 2007).

Other neurophysiology studies parallel the psychophysical investigations of Xu et al. (2008, 2012) and address the influence of adaptation at one level of the hierarchy on the next level of hierarchy (see discussion in the recent review of Solomon & Kohn, 2014). This is bottom-up inheritance, which fits more readily discriminative modeling. As discussed in Solomon and Kohn (2014), some studies suggest that higher levels simply inherit their adaptation (and therefore representation) from lower levels, and are even disrupted by lower level adaptations (as in being "unaware" of the adaptation). But as in the psychophysics, such recordings also offer an opportunity to find aspects of the representation that are set anew at the higher level.

Discussion

The representation of inputs in the brain is a foundational, and yet incompletely addressed, issue. Representation is the target of, and influences, almost all computations, and thus bears on a huge range of subfields, including the ones we discussed here: memory, adaptation, segmentation, and inference. We focused on the representation of visual information. We argued that the representation in the brain is hierarchical—almost trivially in a bottom-up sense, given the multilayered nature of cortical processing (e.g., Felleman & Van Essen, 1991), but also more subtly in the context of generative models.

Hierarchical representation of images has been a focus of extensive study in biological vision and in machine learning. Our review raises issues and directions for future studies from an experimental perspective. We also suggest the need for greater interplay between modern machine learning approaches and experiments. We discuss each of these in turn.

Experimental issues

We have focused on three different aspects of hierarchical representation: part-whole, component, and inheritance. Of these, questions about binding of parts into wholes, as well as binding versus separation of feature components, have perhaps attracted most investigations. Experiments looking at working memory and episodic memory for rather arbitrary objects have indicated distinct limits to the extent of binding, but many questions remain. The same is true for component hierarchies and cross-stimulus adaptation studies. These have shown some hierarchical influences—but mostly in a few special cases addressing the relationship between color and form. Studies in both memory and adaptation have considered part-whole hierarchies for stimulus classes such as faces, or for objects in a scene that are congruent or incongruent. These are suggestive of how parts that are often encountered together may influence binding, but are limited to fairly specialized cases.

Inheritance was a very intense focus of work on semantics some 40 years ago (Collins & Loftus, 1975; Collins & Quillian, 1969), particularly using reaction times as a measure of how structure might be stored and employed. However, the substantial debates about how to interpret these results in terms of localist or distributed networks of knowledge, or various forms of collections of features seem not to have been well resolved (Holyoak, 2007; Rogers & McClelland, 2004); and we could find particularly little work on the role of inheritance in representing visual inputs.

One striking aspect of this area of study is how sparse the various experimental approaches appear to be, leaving much to be investigated. Even though not all combinations are possible, the overall matrix of possibilities of task, measurement, stimulus ensemble, and type of hierarchical representation in question is only sparsely covered by existing experiments. The contrast between our apparently extraordinary capacity to remember huge numbers of scenes (Konkle et al., 2010b; Standing, 1973), and yet to show substantial biases and selective blindness for particular ones (Loftus, 1974) is a seductive target for behavioral work.

One of the goals of this review was to consider experimental approaches to understanding hierarchical representation across a range of fields that are often described separately. Although these very disparate fields study representation in their own way (and often only rather peripherally), we have shown that there are some important potential commonalities in terms of measurement modalities, particularly various types of behavioral co-determination, similarities between wholes, and timing. We have also considered the relatively recent more direct approaches on the problem, with neuroimaging and neurophysiology.

Interesting links between memory and perception have been previously discussed (Palmeri & Tarr, 2008); here we focused on approaches to studying hierarchy. One question is to what extent these systems actually use the same hierarchical representations. For instance, is the hierarchical organization of memory related to that of the ventral stream? Studies of working memory have shown involvement of higher areas such as prefrontal cortex; but studies also show the involvement of high- and mid-level areas of the ventral stream such as inferior temporal cortex and V4 (see references in Pagan, Urban, Wohl, & Rust, 2013), and even suggestion of V1 involvement for some tasks (Super, Spekreijse, & Lamme, 2001). The hippocampus and other medial temporal lobe structures are considered part of the memory system, though there is some controversy about their potential role also in perception (see, for instance, Baxter, 2009; Nadel & Peterson, 2013; Suzuki, 2010; Zeidman & Maguire, 2016). The perirhinal cortex may inform about familiarity versus novelty of parts versus wholes (Nadel & Peterson, 2013). For memory and other tasks, the system may place priority on particular "parts", such as places, people, and actions or functions that one can perform within the scene (see, for example, Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Khosla, Raju, Torralba, & Oliva, 2015; Nadel & Peterson, 2013).

One pressing direction is that of more systematic generation of rich synthetic stimulus hierarchies, for which the statistics are known and controllable. Without this knowledge, formal characterization of the ideal or actual results is very difficult. Two particularly popular classes of rich artificial stimuli, greebles (Gauthier, Williams, Tarr, & Tanaka, 1998; Rezlescu, Barton, Pitcher, & Duchaine, 2014) and ziggerins (Wong, Palmeri, & Gauthier, 2009), have not been particular foci of investigations of hierarchical representations. Work assessing subjects' sensitivity to higher order novel statistical structure in the arrangements of familiar objects (Fiser & Aslin, 2002; Orbán et al., 2008) has concentrated more on statistical normativity than on the sorts of bias errors (or reaction time differences) that would be indicative of the underlying hierarchical representation. Given synthetic stimuli with knowledge of the statistics, one could use the more powerful of both the behavioral methods, of which co-determination in memory and adaptation appear specially promising, and of the representational neuroimaging methods, which are evolving quickly.

Interplay of machine learning and experiments

There is appealing potential for stronger interplay between computational approaches and experiments, given rapid advances in machine learning and computer vision. One important direction is incorporating knowledge about natural scene statistics in both the analysis (Stansbury, Naselaris, & Gallant, 2013) and design of hierarchical experiments. The approach of Portilla and Simoncelli (2000) for generating synthetic textures based on image statistics has been successfully applied in experiments. For instance, this has been useful in studying mid-level visual areas, revealing that area V2 can better discriminate such texture stimuli (Freeman, Ziemba, Heeger, Simoncelli, & Movshon, 2013; Movshon & Simoncelli, 2014).

Furthermore, we can take advantage of the observation that deep convolutional neural networks, which have been trained on a particular supervised learning task on the basis of a huge ensemble of data can capture some aspects of cortical responses to natural stimuli (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Yamins et al., 2014). This means that we can investigate the hierarchical organization of spatial receptive fields through the medium of these models. Work by Gatys, Ecker, and Bethge (2015b) on how content and style are represented in Convolutional Neural Networks can be useful in understanding and manipulating the statistics of natural images. Convolutional Neural Networks together with other recent advances in unsupervised learning and generative models using deep networks and hierarchical approaches in general, offer great potential as methods to generate synthetic stimuli from different layers (Ballé, Laparra, & Simoncelli, 2015; Dosovitskiy & Brox, 2016; Gatys, Ecker, & Bethge,

2015a; Goodfellow et al., 2014; Mahendran & Vedaldi, 2015) and to apply the stimuli in experiments. There is the possibility to use some of the co-determination metrics discussed in this review, along with stimuli generated according to deep networks, to make progress in understanding hierarchical representations.

Other more radically different approaches using artificially generated stimuli or natural stimuli to learn behavioral priors can also be tried. There has been interesting work along these lines applying Markov-chain Monte-Carlo with people (Sanborn & Griffiths, 2007; Sanborn, Griffiths, & Shiffrin, 2010), Representations Envisioned Via Evolutionary ALgorithms (REVEAL) (Greene, Botros, Beck, & Fei-Fei, 2014), and cognitive tomography (Houlsby et al., 2013). These have thus far not been applied to studying hierarchies. There is also a need to gain better understanding of how well current hierarchical models can explain experimental data and when they fail. Studies on transfer learning suggest that a discriminative network optimized for one task might be readily turned to solve other tasks (Donahue et al., 2013; Razavian, Azizpour, Sullivan, & Carlsson, 2014). These results, together with the applicability of the high-level units in deep convolutional networks to questions on distinctiveness in memory (Bylinskii et al., 2015), intriguingly suggest that these networks might also be viable as a basis for simulating hierarchical co-determination behavioral phenomena discussed here for adaptation and memory. One question is how well such models can capture existing data on co-determination (such as biases and forgetting). Understanding when the models break down can lead potentially to model refinement and new experiments.

In the work on "atoms of recognition," Ullman, Assif, Fetaya, and Harari (2016) have developed a hierarchical approach for generating images that are reduced in size or resolution. They test human observers on recognition, until reaching what is called "minimal recognizable images," whereby further reduction has drastic effect on image recognition. Ullman et al. (2016) demonstrate that current models, including deep convolutional neural networks, cannot account for this effect, and propose a possible role for top-down processes.

One of the major fault-lines running through this area concerns the relationship between generative and discrimination hierarchies, acknowledging the asymmetric requirement of the former on the latter. Purely discriminative hierarchies have proven extremely powerful, given an appropriate set of supervised learning tasks. However, it is unclear how one might progress from those, perhaps via a broadening of the transfer learning ideas, to the sort of task-general representations that generative models tantalizingly, though currently incompetently, offer. Also note that the successful purely discriminative solutions involve very implicit solutions to operations such as binding and separation that are explicit in the generative hierarchies.

Another issue regarding supervised deep convolutional neural networks is implementation. The recent intriguing similarities with cortical neural areas raise the question of the plausibility of implementing related networks in the brain. Deep convolutional networks rely on supervised error correction using back propagation, an approach that was put forward in the mid-1980s (Rumelhart, Hinton, & Williams, 1986). Shortly thereafter, Crick (1989) commented on "the recent excitement about neural networks" and raised questions about the biological plausibility of implementing back propagation in the brain, questions that still largely hold and for which there is renewed interest today. For recent discussion on potential mechanisms that might be candidates or alternatives for approximating back propagation, see Bengio, Lee, Bornschein, and Lin (2016), Hinton (2016), and Marblestone, Wayne, and Kording (2016). There is also debate as to whether high-level areas of the brain are more compatible with unsupervised or supervised learning. For instance, studies have suggested that even at high-level visual areas, neural representation of objects according to similarity largely depend on shape or low-level structure rather than semantic class (Baldassi et al., 2013; Freedman, Riesenhuber, Poggio, & Miller, 2003; Jiang et al., 2007), although there have been conflicting results on the importance of semantics (Kiani, Esteky, Mirpour, & Tanaka, 2007; Kriegeskorte et al., 2008).

There are other machine learning approaches that could be applied to understanding hierarchical representations. There has been recent progress in linking hierarchical generative models to perception and cognition. Approaches of inferring generative structure about the environment in nonparametric hierarchical Bayesian models are appealing and have the potential for generalization of new concepts with even only a single example (Tenenbaum, Kemp, Griffiths, & Goodman, 2011; Tervo, Tenenbaum, & Gershman, 2016). These approaches have been successfully applied for structured visual stimuli, such as handwritten characters (Lake, Salakhutdinov, & Tenenbaum, 2015). Bayesian hierarchical models have also been applied in the motion domain to capture the dependencies between objects and parts that are grouped together, and to explain some phenomena in motion perception and biases of moving dots that form complex motion structure (Gershman, Tenenbaum, & Jäkel, 2015). Some of the memory co-determination studies on regularities in working memory for synthetic stimuli have been explained within a Bayesian framework (Brady & Tenenbaum, 2013), although these have not

incorporated learning complex regularities, or scaled up to natural scenes.

Predictive coding has also been an appealing target of computational modeling in vision (MacKay, 1956; Mumford, 1992), mostly applied to early cortical stages (Lochmann, Ernst, & Deneve, 2012; Rao & Ballard, 1999; Spratling, 2012). Predictive coding has been an inspiration for perceptual experiments on higher level coherence suppressing activity at lower levels of the hierarchy (Fang, Kersten, & Murray, 2008; Murray, Kersten, Olshausen, Schrater, & Woods, 2002). Related approaches that have been applied to modeling primary visual cortex data also offer a potential route to modeling higher levels of the hierarchy and capturing biological data via divisive normalization (Coen-Cagli & Schwartz, 2013; Heeger, 1992; Schwartz & Simoncelli, 2001; Schwartz, Sejnowski, & Dayan, 2009), a computation that is already used in various simpler forms in deep convolutional networks (Ba, Kiros, & Hinton, 2016; Ioffe & Szegedy, 2015; Jarrett et al., 2009; Krizhevsky et al., 2012).

Other modeling approaches have focused on issues of coping with a limited anatomy. One powerful approach along these lines is the Plate's (1995) holographic reduced representation (HRR). HRRs provide explicit representational "plumbing" that allows binding and recursion (or reuse) to work (the latter involving a sequence of computational operations that might have detectable consequences in reaction times) in the context of palimpsest-like additive working memory. It has been used for a highly impressive range of demonstrations of neural computation (Eliasmith, 2013; Eliasmith et al., 2012).

Conclusion

Hierarchical representations of sensory input play critical computational roles, since they reflect aspects of the way that inputs are created. More elusively, they also play critical algorithmic roles, structuring the way that information processing and memory are carried out. Yet further from our current understanding is their neural realization (Tervo et al., 2016).

Here, we focused on experimental approaches for understanding hierarchical representations in vision for static images. Experiments in memory recall, adaptation, and inference have often been studied separately by different communities in neuroscience and cognitive science. However, we show that in all these areas, there has been extensive interest in measuring hierarchical representations, either indirectly behavirally, or more directly with neural measurements.

A main point of these experiments has been to provide compelling impetus to the development of theories. However, at present we can at best only be described as having fragments of theories—including algorithmic ideas such as holographic reduced representations, supervised-learning based deep networks, semantic networks, forms of blackboard architecture, and more. These have individually compelling features, and indeed have been embedded in impressive computational architectures. However, they do not amount to complete accounts that would extend appropriately to the size and scale of the full problem.

A reason for optimism is the precipitate innovation in machine learning, from deep convolutional networks onwards, which simultaneously offer substantial improvements in engineering performance and increased fidelity as models of aspects of neural information processing. Exploiting and extending these methods, using both natural scene and novel, hierarchically controlled, inputs, offers an attractive prospect for future investigations.

Keywords: hierarchy, *representation*, *natural scenes*, *deep learning*

Acknowledgments

We are most grateful to Peter Battaglia and David Raposo for comments on the manuscript, to Jon Shlens for pointing out some relevant literature, and to Peter Dayan for advice at the outset of the project. This work was supported by a Google Faculty Research Award to OS.

Commercial relationships: none.
Corresponding author: Odelia Schwartz.
Email: odelia@cs.miami.edu.
Address: Department of Computer Science, University of Miami, Miami, FL, USA.

References

Adelson, E. H., & Pentland, A. P. (1996). The perception of shading and reflectance. In D. Knill and W. Richards (Eds.), *Perception as Baysian inference* (pp. 409–423). New York: Cambridge University Press.

Anselmi, F., Leibo, J. Z., Rosasco, L., Mutch, J., Tacchetti, A., & Poggio, T. (2014). Unsupervised learning of invariant representations with low sample complexity: The magic of sensory cortex or a new framework for machine learning? (Technical Report arXiv:1311.4158). Cambridge, MA: MIT Press.

Aslin, R. N., & Newport, E. L. (2014). Distributional language learning: Mechanisms and models of category formation. *Language Learning*, *64*(s2), 86–105.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. Retrieved from arXiv arXiv: 1607.06450

Baldassi, C., Alemi-Neissi, A., Pagan, M., DiCarlo, J. J., Zecchina, R., & Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. *PLoS Computational Biology*, *9*(8), e1003167.

Ballé, J., Laparra, V., & Simoncelli, E. P. (2015). Density modeling of images using a generalized normalization transformation. Presented at the International Conference on Learning Presentations, 2016, San Juan, Puerto Rico. Retrieved from arXiv:1511.06281

Bartlett, J. R., & Doty, R. S., Sr.. (1974). Response of units in striate cortex of squirrel monkeys to visual and electrid stimuli. *Journal of Neurophysiology*, *37*(4), 621–641.

Baxter, M. G. (2009). Involvement of medial temporal lobe structures in memory and perception. *Neuron*, *61*(5), 667–677.

Bays, P. M., Wu, E. Y., & Husain, M. (2011). Storage and binding of object features in visual working memory. *Neuropsychologia*, *49*(6), 1622–1631.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129–1159.

Bengio, Y., Lee, D.-H., Bornschein, J., & Lin, Z. (2016). Towards biologically plausible deep learning. Retrieved from ArXiv preprint at arXiv: 1502.04156.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177.

Bienenstock, E., & Geman, S. (1995). Compositionality in neural systems. In M. Arbib (Ed.), *The handbook of brain theory and neural networks*. Boston, MA: Bradford Books/MIT Press.

Bienenstock, E., Geman, S., & Potter, D. (1997). Compositionality, mdl priors, and object recognition. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems* (pp.838–844). Cambridge MA: MIT Press.

Boorman, E. D., Rajendran, V. G., O'Reilly, J. X., & Behrens, T. E. (in press). Two anatomically and computationally distinct learning signals predict changes to stimulus-outcome associations in hippocampus. *Neuron*, *89*(6), 1343–1354.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory ensemble statistics bias memory for individual items. *Psychological Science*, *22*(3), 384–392.

Brady, T. F., & Alvarez, G. A. (2015). Contextual effects in visual working memory reveal hierarchically structured memory representations. *Journal of Vision*, *15*(15):6, 1–24, doi:10.1167/15.15.6. [PubMed] [Article]

Brady, T. F., Konkle, T., & Alvarez, G. A. (2009). Compression in visual working memory: using statistical regularities to form more efficient memory representations. *Journal of Experimental Psychology: General*, *138*(4), 487.

Brady, T. F., Konkle, T., & Alvarez, G. A. (2011). A review of visual memory capacity: Beyond individual items and toward structured representations. *Journal of Vision*, *11*(5):4, 1–34, doi:10.1167/11.5.4. [PubMed] [Article]

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, USA*, *105*(38), 14325–14329.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981–990.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*(1), 85–109.

Brainard, D., & Radonjić, A. (2014). Color constancy. In J. S. Werner & L. M. Chalupa (Eds.), *The new visual neurosciences*. Cambridge, MA: MIT Press.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*(2), 207–230.

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., & Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, *116*(Pt. B), 165–178.

Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.

(2015). Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2956–2964). New York: IEEE.

Chang, D. H., Hess, R. F., Thompson, B., & Mullen, K. T. (2014). fmri adaptation of color and achromatic contrast in the human lgn and visual cortex: Evidence for color and luminance selectivity. *Journal of Vision*, *14*(10):983, doi:10.1167/14.10.938. [Abstract]

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. Retrieved from arXiv preprint arXiv:1601.02970

Clifford, C. W., Pearson, J., Forte, J. D., & Spehar, B. (2003). Colour and luminance selectivity of spatial and temporal interactions in orientation perception. *Vision Research*, *43*(27), 2885–2893.

Clifford, C. W., & Rhodes, G. (2005). *Fitting the mind to the world: Adaptation and after-effects in high-level vision* (Vol. 2). New York: Oxford University Press.

Clifford, C. W., Spehar, B., Solomon, S. G., Martin, P. R., & Qasim, Z. (2003). Interactions between color and luminance in the perception of orientation. *Journal of Vision*, *3*(2):1, 106–115, doi:10.1167/3.2.1. [PubMed] [Article]

Coco, M. I., Malcolm, G. L., & Keller, F. (2014). The interplay of bottom-up and top-down mechanisms in visual guidance during object naming. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1096–1120.

Coen-Cagli, R., & Schwartz, O. (2013). The impact on midlevel vision of statistically optimal divisive normalization in V1. *Journal of Vision*, *13*(8):13, 1–20, doi:10.1167/13.8.13. [PubMed] [Article]

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 240–247.

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, *35*(3), 393–401.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15*(8), 559–564.

Dayan, P. (2006). Images, frames, and connectionist hierarchies. *Neural Computation*, *18*(10), 2293–2319.

Dickinson, J. E., Almeida, R. A., Bell, J., & Badcock, D. R. (2010). Global shape aftereffects have a local substrate: A tilt aftereffect field. *Journal of Vision*, *10*(13):5, 1–12, doi:10.1167/10.13.5. [PubMed] [Article]

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2013). Decaf: A deep convolutional activation feature for generic visual recognition. Presented at the 31st International Conference on Machine Learning, Beijing. China, 2014. Retrieved from *arXiv preprint arXiv:1310.153*

Dosovitskiy, A., & Brox, T. (2016). Generating images with perceptual similarity metrics based on deep networks. Presented at the 30th Conference on Neural Information Processing Sytems (NIPS), Barcelona, Spain, 2016. Retrieved from *arXiv: 1602.02644*

Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*. New York: Oxford University Press.

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012, Nov 30). A large-scale model of the functioning brain. *Science*, *338*(6111), 1202–1205.

Epshtein, B., Lifshitz, I., & Ullman, S. (2008). Image interpretation by a single bottom-up top-down cycle. *Proceedings of the National Academy of Sciences, USA*, *105*(38), 14298–14303.

Fang, F., Kersten, D., & Murray, S. O. (2008). Perceptual grouping and inverse fmri activity patterns in human visual cortex. *Journal of Vision*, *8*(7):2, 2–9, doi:10.1167/8.7.2. [PubMed] [Article]

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(9), 1627–1645.

Felzenszwalb, P. F., & Huttenlocher, D. P. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision*, *61*(1): 55–79.

Felzenszwalb, P. F., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *IEEE Computer Vision and Pattern Recognition, 2008* (pp. 1-). New York: IEEE.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 264–271). New York: IEEE.

Fischler, M. A., & Elschlager, R. A. (1973). The representation and matching of pictorial structures. *IEEE Transactions on Computers, C-22*(1), 67–92.

Fiser, J., & Aslin, R. N. (2002). Statistical learning of new visual feature combinations by infants. *Proceedings of the National Academy of Sciences, USA*, *99*(24), 15822–15826.

Fougnie, D., & Alvarez, G. A. (2011). Object features fail independently in visual working memory: Evidence for a probabilistic feature-store model. *Journal of Vision*, *11*(12):3, 1–12, doi:10.1167/11.12.3. [PubMed] [Article]

Fowlkes, C. C., Martin, D. R., & Malik, J. (2007). Local figure–ground cues are valid for natural images. *Journal of Vision*, *7*(8):2, 1–9, doi:10.1167/7.8.2. [PubMed] [Article]

Franken, R., & Rowland, G. (1979). Nature of the representation for picture-recognition memory. *Perceptual and Motor Skills*, *49*(2), 619–629.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. *The Journal of Neuroscience*, *23*(12), 5235–5246.

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*(7), 974–981.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4): 193–202.

Gatys, L., Ecker, A. S., & Bethge, M. (2015a). Texture synthesis using convolutional neural networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 262–270). Red Hook, NY: Curran Associates, Inc.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015b). A neural algorithm of artistic style. Retrieved from-*CoRR*, abs/1508.06576

Gauthier, I., Williams, P., Tarr, M. J., & Tanaka, J. (1998). Training greebleexperts: A framework for studying expert object recognition processes. *Vision Research*, *38*(15), 2401–2428.

Gershman, S. J., Tenenbaum, J. B., & Jäkel, F. (2015). Discovering hierarchical motion structure. *Vision Research*, *126*, 232–241.

Glasner, D., Bagon, S., & Irani, M. (2009). Super-resolution from a single image. In 12th IEEE international conference on computer vision (pp. 349–356). New York: IEEE.

Golomb, J. D., & Kanwisher, N. (2011). Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cerebral Cortex*, *22*(12), 2794–2810.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology*, *145*(1), 82–94.

Greene, M. R., Botros, A. P., Beck, D. M., & Fei-Fei, L. (2014). Visual noise from natural scene statistics reveals human scene category representations. Retrieved from *arXiv preprint arXiv:1411.5331*

Greene, M. R., & Oliva, A. (2010). High-level aftereffects to global scene properties. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(6), 1430.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23.

Güçlü, U. & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005–10014.

Haberman, J., Brady, T. F., & Alvarez, G. A. (2015). Individual differences in ensemble perception reveal multiple, independent levels of ensemble representation. *Journal of Experimental Psychology: General*, *144*(2), 432–446.

Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*(2), 181–197.

Hemmer, P., & Steyvers, M. (2009a). Integrating episodic and semantic information in memory for natural scenes. In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the*

Cognitive Science Society (pp. 1557–1562). Austin, TX: Cognitive Science Society.

Hemmer, P. & Steyvers, M. (2009b). Integrating episodic memories and prior knowledge at multiple levels of abstraction. *Psychonomic Bulletin & Review*, *16*(1), 80–87.

Hinton, G. E. (1984). *Distributed representations* (Technical Report CMU-CS-84-157). Pittsburgh, PA: CMU Press.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*(1), 47–75.

Hinton, G. E. (2016). Can the brain do back-propagation. Invited talk at Stanford University Colloquium on Computer Systems, Stanford, CA, April 27, 2016.

Hinton, G. E., Dayan, P., Frey, B. J., & Neal, R. M. (1995, May 26). The "wake-sleep" algorithm for unsupervised neural networks. *Science*, *268*(5214), 1158–1161.

Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1177–1190.

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. In *Artificial neural networks and machine learning–ICANN 2011* (pp. 44–51). New York, NY: Springer.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hinton, G. E., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. Retrieved from *arXiv preprint arXiv:1503.02531*

Hinton, G. E. & Zemel, R. S. (1994). Autoencoders, minimum description length and Helmholtz free energy. In J. D. Cowan, G. Tesauro, & J. Alspector (Eds.), *NIPS1993* (pp. 3–10). La Jolla, CA: NIPS.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holyoak, K. J. (2007). Relations in semantic memory: Still puzzling after all these years. In M. A. Gluck, J. R. Anderson, & S. M. Kosslyn (Eds.), *Memory and mind: A festschrift for Gordon H. Bower* (pp. 141–158). Hove, UK: Psychology Press.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613–622.

Hosoya, H., & Hyvärinen, A. (2015). A hierarchical statistical model of natural images explains tuning properties in V2. *The Journal of Neuroscience*, *35*(29), 10412–10428.

Houlsby, N. M. T., Huszár, F., Ghassemi, M. M., Orbán, G., Wolpert, D. M., & Lengyel, M. (2013). Cognitive tomography reveals complex, task-independent mental representations. *Current Biology*, *23*(21), 2169–2175.

Huang, J., & Sekuler, R. (2010). Distortions in recall from visual memory: Two classes of attractors at work. *Journal of Vision*, *10*(2):24, 1–27, doi:10.1167/10.2.24. [PubMed] [Article]

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 179–187.

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In D. Blei & F. Bach (Eds.), *Proceedings of the 32nd international conference on machine learning* (pp. 448–456). JMLR Workshop and Conference Proceedings.

Ito, M., & Komatsu, H. (2004). Representation of angles embedded within contour stimuli in area V2 of macaque monkeys. *The Journal of Neuroscience*, *24*, 3313–3324.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., & LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *Proceedings of the IEEE 12th international conference on computer vision* (pp. 2146–2153). New York: IEEE.

Jiang, X., Bradley, E., Rini, R. A., Zeffiro, T., VanMeter, J., & Riesenhuber, M. (2007). Categorization training results in shape-and category-selective human neural plasticity. *Neuron*, *53*(6), 891–903.

Jiang, X., Rosen, E., Zeffiro, T., VanMeter, J., Blanz, V., & Riesenhuber, M. (2006). Evaluation of a shape-based model of human face discrimination using fmri and behavioral techniques. *Neuron*, *50*(1), 159–172.

Jin, E. W., & Shevell, S. K. (1996). Color memory and color constancy. *JOSA A*, *13*(10), 1981–1991.

Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, *8*(13):11, 1–13, doi:10.1167/8.13.11. [PubMed] [Article]

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*(26), 3286–3297.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304.

Khaligh-Razavi, S.-M., Henriksson, L., Kay, K., & Kriegeskorte, N. (2014). Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models. *bioRxiv*, page 009936.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computer Biology 10*(11): e1003915.

Khosla, A., Raju, A. S., Torralba, A., & Oliva, A. (2015). Understanding and predicting image memorability at a large scale. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2390–2398). New York: IEEE.

Khosla, A., Xiao, J., Isola, P., Torralba, A., & Oliva, A. (2012). Image memorability and visual inception. In *SIGGRAPH Asia 2012 technical briefs* (pp. 35:1–35:4). New York: ACM.

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*(6), 4296–4309.

Kinoshita, M., & Komatsu, H. (2001). Neural representation of the luminance and brightness of a uniform surface in the macaque primary visual cortex. *Journal of Neurophysiology*, *86*(5), 2559–2570.

Klein-Flügge, M. C., Barron, H. C., Brodersen, K. H., Dolan, R. J., & Behrens, T. E. J. (2013). Segregated encoding of reward–identity and stimulus–reward associations in human orbitofrontal cortex. *The Journal of Neuroscience*, *33*(7), 3202–3211.

Kohler, P. J., Clarke, A., Yakovleva, A., Liu, Y., & Norcia, A. M. (2016). Representation of maximally regular textures in human visual cortex. *The Journal of Neuroscience*, *36*(3), 714–729.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010a). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010b). Scene memory is more detailed than you think the role of categories in visual long-term memory. *Psychological Science*, *21*(11), 1551–1556.

Konkle, T., & Oliva, A. (2007). Normative representation of objects: Evidence for an ecological bias in object perception and memory. In *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 407–413). Austin, TX: Cognitive Science Society.

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review of Neuroscience*, *34*, 45–67.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., ... Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In P. Bartlett (Ed.), *Advances in neural information processing systems* (pp.1097–1105). La Jolla, CA: NIPS.

Kuriki, I., Xie, H., Tokunaga, R., Matsumiya, K., & Shioiri, S. (2014). Interaction of color-defined and luminance-defined motion signals in human visual cortex. *Journal of Vision*, *14*(10):291, doi:10.1167/14.10.291. [Abstract]

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015, Dec 11). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. In *Proceedings of the IEEE conference on acoustics, speech and signal processing* (pp. 8595–8598). New York: IEEE.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

Lee, H., Ekanadham, C., & Ng, A. (2008). Sparse deep belief net model for visual area V2. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems* (pp. 873–880). La Jolla, CA: NIPS.

Levie, W. H., & Hathaway, S. (1988). Picture recognition memory: A review of research and theory. *Journal of Visual/Verbal Languaging*, *8*(1), 6–45.

Lochmann, T., Ernst, U. A., & Deneve, S. (2012). Perceptual inference predicts contextual modulations of sensory responses. *The Journal of Neuroscience*, *32*(12), 4179–4195.

Loftus, E. F. (1974). Reconstructing memory: The incredible eyewitness. *Jurimetrics Journal*, *15*(3), 188–193.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565.

MacEvoy, S. P., & Paradiso, M. A. (2001). Lightness constancy in primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, *98*(15), 8827–8831.

MacKay, D. M. (1956). Towards an information-flow model of human behaviour. *British Journal of Psychology*, *47*(1), 30–43.

Mahendran, A., & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5188–5196). New York: IEEE.

Malmir, M., & Ghidary, S. S. (2009). A model of angle selectivity in area V2 with local divisive normalization. In *Proceedings of the IEEE conference on computational intelligence for multimedia signal and vision processing, 2009* (pp. 1–5). New York: IEEE.

Mandelbrot, B. B. (1983). *The fractal geometry of nature* (Vol. 173). London: Macmillan.

Marblestone, A. H., Wayne, G., & Kording, K. P. (2016). Toward an integration of deep learning and neuroscience. *Frontiers in Computational Neuroscience*, *10*, 94.

Marcus, G. F., Vijayan, S., Rao, S. B., & Vishton, P. M. (1999, Jan 1). Rule learning by seven-month-old infants. *Science*, *283*(5398), 77–80.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: Freeman and Company.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). *The appeal of parallel distributed processing*. Cambridge, MA: MIT Press.

Mnih, V., Heess, N., Graves, A., & Kavukcuoglu, K. (2014). Recurrent models of visual attention. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *NIPS '14 proceedings of the 27th international conference on neural information processing systems* (pp. 2204–2212). Cambridge, MA: MIT Press.

Month, A. (2003). Artful mathematics: The heritage of M. C. Escher. *NOTICES OF THE AMS*, *50*(4), 446–451.

Movshon, J. A., & Simoncelli, E. P. (2014). Representation of naturalistic image structure in the primate visual cortex. In *Cold Spring Harbor symposia on quantitative biology* (Vol. 79, pp. 115–122). Long Island, NY: Cold Spring Harbor Laboratory Press.

Mullally, S. L., Intraub, H., & Maguire, E. A. (2012). Attenuated boundary extension produces a paradoxical memory advantage in amnesic patients. *Current Biology*, *22*(4), 261–268.

Mumford, D. (1992). On the computational architecture of the neocortex. *Biological Cybernetics*, *66*(3), 241–251.

Murphy, G. L., Hampton, J. A., & Milovanovic, G. S. (2012). Semantic memory redux: An experimental test of hierarchical category representation. *Journal of Memory and Language*, *67*(4), 521–539.

Murray, S. O., Kersten, D., Olshausen, B. A., Schrater, P., & Woods, D. L. (2002). Shape perception reduces activity in human primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, *99*(23), 15164–15169.

Nadel, L., & Peterson, M. A. (2013). The hippocampus: Part of an interactive posterior representational system spanning perceptual and memorial systems. *Journal of Experimental Psychology: General*, *142*(4), 1242–1254.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

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, USA*, *112*(4), E351–E360.

Olshausen, B. A., & Field, D.J. (1996) . Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Orbán, G., Fiser, J., Aslin, R. N., & Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences, USA*, *105*(7), 2745–2750.

Orhan, A. E., & Jacobs, R. A. (2013). A probabilistic clustering theory of the organization of visual short-term memory. *Psychological Review*, *120*(2), 297–328.

Orhan, A. E., & Jacobs, R. A. (2014). Toward ecologically realistic theories in visual short-term memory research. *Attention, Perception, & Psychophysics*, *76*(7), 2158–2170.

Pagan, M., Urban, L. S., Wohl, M. P., & Rust, N. C. (2013). Signals in inferotemporal and perirhinal

cortex suggest an untangling of visual target information. *Nature Neuroscience*, *16*(8), 1132–1139.

Palmeri, T. J., & Tarr, M. (2008). Visual object perception and long-term memory. In S. J. Luck & A. Hollingworth (Eds.), *Visual memory* (pp. 163–207), Oxford, UK: Oxford University Press.

Palumbo, R., DAscenzo, S., & Tommasi, L. (2015). Cross-category adaptation: Exposure to faces produces gender aftereffects in body perception. *Psychological Research*, *79*(3), 380–388.

Pasupathy, A. (2015). The neural basis of image segmentation in the primate brain. *Neuroscience*, *296*, 101–109.

Pezdek, K., Whetstone, T., Reynolds, K., Askari, N., & Dougherty, T. (1989). Memory for real-world scenes: The role of consistency with schema expectation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 587–595.

Plate, T. (1995). Holographic reduced representations. *IEEE Transactions on Neural networks*, *6*(3), 623–641.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–70.

Pospisil, D., Pasupathy, A., & Bair, W. (2016). Comparing the brains representation of shape to that of a deep convolutional neural network. In J. Suzuki, T. Nakano, & H. Hess (Eds.), *The first international workshop on computational models of the visual cortex: hierarchies, layers, sparsity, saliency and attention*. New York: ACM.

Potter, M. C. (2012). Recognition and memory for briefly presented scenes. *Frontiers in Psychology*, *3*, 32.

Qiu, C., Kersten, D., & Olman, C. A. (2013). Segmentation decreases the magnitude of the tilt illusion. *Journal of Vision*, *13*(13):19, 1–17, doi:10.1167/13.13.19. [PubMed] [Article]

Quinlan, P. T., & Cohen, D. J. (2011). Object-based representations govern both the storage of information in visual short-term memory and the retrieval of information from it. *Psychonomic Bulletin & Review*, *18*(2), 316–323.

Ranzato, M. A., Huang, F. J., Boureau, Y.-L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). New York: IEEE.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Raviv, O., Ahissar, M., & Loewenstein, Y. (2012). How recent history affects perception: The normative approach and its heuristic approximation. *PLoS Computational Biology*, *8*(10), e1002731.

Raviv, O., Lieder, I., Loewenstein, Y., & Ahissar, M. (2014). Contradictory behavioral biases result from the influence of past stimuli on perception. *PLoS Computational Biology*, *10*(12), e1003948.

Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: An astounding baseline for recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 806–813). New York: IEEE.

Ren, X., Fowlkes, C. C., & Malik, J. (2006). Figure/ground assignment in natural images. In *European Conference on Computer Vision* (pp. 614–627). New York, NY: Springer.

Rezlescu, C., Barton, J. J., Pitcher, D., & Duchaine, B. (2014). Normal acquisition of expertise with greebles in two cases of acquired prosopagnosia. *Proceedings of the National Academy of Sciences, USA*, *111*(14), 5123–5128.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.

Riesenhuber, M., & Wolff, B. S. (2009). Task effects, performance levels, features, configurations, and holistic face processing: A reply to Rossion. *Acta Psychologica*, *132*(3), 286–292.

Roe, A. W., Chelazzi, L., Connor, C. E., Conway, B. R., Fujita, I., Gallant, J. L., ... Vanduffel, W. (2012). Toward a unified theory of visual area V4. *Neuron*, *74*(1), 12–29.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rossi, A. F., Rittenhouse, C. D., & Paradiso, M. A. (1996, Aug 23). The representation of brightness in primary visual cortex. *Science*, *273*(5278), 1104.

Ruderman, D. L., & Bialek, W. (1994). Statistics of natural images: Scaling in the woods. *Physical Review Letters*, *73*(6), 814.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J.

(1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance (invariance) both increase as visual information propagates from cortical area V4 to IT. *The Journal of Neuroscience*, *30*(39), 12978–12995.

Rust, N. C., & DiCarlo, J. J. (2012). Balanced increases in selectivity and tolerance produce constant sparseness along the ventral visual stream. *The Journal of Neuroscience*, *32*(30), 10170–10182.

Sanborn, A., & Griffiths, T. L. (2007). Markov chain monte carlo with people. In J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.), *Advances in neural information processing systems* (pp. 1265–1272). La Jolla, CA: NIPS.

Sanborn, A. N., Griffiths, T. L., & Shiffrin, R. M. (2010). Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, *60*(2), 63–106.

Santangelo, V. (2015). Forced to remember: When memory is biased by salient information. *Behavioural Brain Research*, *283*, 1–10.

Schrater, P., & Kersten, D. (2002). Vision, psychophysics and Bayes. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain* (pp. 37–60). Cambridge, MA: MIT Press.

Schulman, A. I. (1974). Memory for words recently classified. *Memory & Cognition*, *2*(1), 47–52.

Schwartz, O., Hsu, A., & Dayan, P. (2007). Space and time in visual context. *Nature Reviews Neuroscience*, *8*(7), 522–535.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2009). Perceptual organization in the tilt illusion. *Journal of Vision*, *9*(4):19, 1–20, doi:10.1167/9.4.19.

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience*, *4*(8), 819–825.

Silva, M. M., Groeger, J. A., & Bradshaw, M. F. (2006). Attention–memory interactions in scene perception. *Spatial Vision*, *19*(1), 9–19.

Solomon, S. G., & Kohn, A. (2014). Moving sensory adaptation beyond suppressive effects in single neurons. *Current Biology*, *24*(20), R1012–R1022.

Spehar, B., Clifford, C. W., Newell, B. R., & Taylor, R. P. (2003). Universal aesthetic of fractals. *Computers & Graphics*, *27*(5), 813–820.

Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation*, *24*(1), 60–103.

Standing, L. (1973). Learning 10000 pictures. *The Quarterly Journal of Experimental Psychology*, *25*(2), 207–222.

Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, *79*(5), 1025–1034.

Steyvers, M., & Hemmer, P. (2012). Reconstruction from memory in naturalistic environments. In B. H. Ross (Ed.), *The psychology of learning and motivation-advances in research and theory* (Vol. 56, pp. 125–144). New York: Academic Press.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). Learning hierarchical models of scenes, objects, and parts. In *Proceedings of the 10th IEEE conference on computer vision* (Vol. 2, pp. 1331–1338). New York: IEEE.

Super, H., Spekreijse, H., & Lamme, V. A. (2001, Jul 6). A neural correlate of working memory in the monkey primary visual cortex. *Science*, *293*(5527), 120–124.

Suzuki, W. A. (2010). Untangling memory from perception in the medial temporal lobe. *Trends in Cognitive Sciences*, *14*(5), 195–200.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*(2), 225–245.

Tang, Y., Salakhutdinov, R., & Hinton, G. (2012). Deep lambertian networks. In *International conference on machine learning*. Madison, WI: Omnipress.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011, Mar 11). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tenenbaum, J. M., & Witkin, A. (1983). On the role of structure in vision. In J. Beck, B. Hope, & A. Rosenfeld (Eds.), *Human and machine vision* (pp. 481–543). New York: Academic Press.

Tervo, D. G. R., Tenenbaum, J. B., & Gershman, S. J. (2016). Toward the neural implementation of structure learning. *Current Opinion in Neurobiology*, *37*, 99–105.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006, Feb 3). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Ullman, S., Assif, L., Fetaya, E., & Harari, D. (2016). Atoms of recognition in human and computer

vision. *Proceedings of the National Academy of Sciences, USA*, *113*(10), 2744–2749.

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156.

Von Helmholtz, H. (1867). *Handbuch der physiologischen Optik* (vol. 9) [Translation: *Manual of physiological optics*]. Berlin: Voss.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*(6), 598–604.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12):11, 1120–1135, doi:10.1167/4.12.11. [PubMed] [Article]

Williford, J. R., & von der Heydt, R. (2016). Figure-ground organization in visual cortex for natural scenes. *bioRxiv*, 053488.

Willmore, B. D., Prenger, R. J., & Gallant, J. L. (2010). Neural representation of natural images in visual area V2. *The Journal of Neuroscience*, *30*(6), 2102–2114.

Wong, A. C.-N., Palmeri, T. J., & Gauthier, I. (2009). Conditions for facelike expertise with objects becoming a ziggerin expert but which type? *Psychological Science*, *20*(9), 1108–1117.

Xu, H., Dayan, P., Lipkin, R. M., & Qian, N. (2008). Adaptation across the cortical hierarchy: Low-level curve adaptation affects high-level facial-expression judgments. *The Journal of Neuroscience*, *28*(13), 3374–3383.

Xu, H., Liu, P., Dayan, P., & Qian, N. (2012). Multilevel visual adaptation: Dissociating curvature and facial-expression aftereffects produced by the same adapting stimuli. *Vision Research*, *72*, 42–53.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

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, USA*, *111*(23), 8619–8624.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zeidman, P., & Maguire, E. A. (2016). Anterior hippocampus: The anatomy of perception, imagination and episodic memory. *Nature Reviews Neuroscience*, *17*(3), 173–182.

Zeiler, M. D., Taylor, G. W., & Fergus, R. (2011). Adaptive deconvolutional networks for mid and high level feature learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2018–2025). New York: IEEE.

Zhaoping, L. (2005). Border ownership from intracortical interactions in visual area V2. *Neuron*, *47*(1), 143–153.

Zhaoping, L. (2014). *Understanding vision: Theory, models, and data*. Oxford, UK: Oxford University Press.

Zhou, H., Friedman, H. S., & Von Der Heydt, R. (2000). Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, *20*(17), 6594–6611.

Ziemba, C. M., Freeman, J., Movshon, J. A., & Simoncelli, E. P. (2016). Selectivity and tolerance for visual texture in macaque V2. *Proceedings of the National Academy of Sciences, USA*, *113*(22), E3140–E3149.

Zoran, D., & Weiss, Y. (2009). Scale invariance and noise in natural images. In *Proceedings of the IEEE international conference on computer vision, 2009* (pp. 2209–2216). New York: IEEE.