30 Population-Based Representations: From Implicit to Explicit

NICOLE C. RUST

ABSTRACT Many of our everyday perceptual and cognitive tasks require our brains to transform information from *implicit representations* in which task-relevant information exists but in a format that is difficult to extract, into *explicit representations* in which this information is accessible. For example, determining the identities of objects that are currently in view across naturally occurring variation, such as changes in an object's position, requires our brains to reformat the pattern of light-based representations encoded by our photoreceptors into representations that explicitly reflect object identity. The brain faces similar challenges for other perceptual tasks, such as identifying words spoken by different voices, as well as more cognitive challenges, such as determining whether a chair belongs to the category of furniture. Insight into these challenges, and the brain's solutions, can be understood using geometrical, population-based coding approaches. Once formulated, these population-based descriptions can be linked to the single- and multi-neuron mechanisms that support a successful task solution as well as provide important insights into the computations that the brain uses to process information.

Many of the computations that our brains perform can be restated as transformations of information from *implicit representations* in which task-relevant information exists but in a format that is difficult to extract, into *explicit representations* in which this information is accessible. Transformations from implicitly to explicitly formatted information are often required when our brains need to "group" the neural responses to different conditions into sets, and this requirement is ubiquitously present in the tasks that we perform every day. For example, consider the perceptual challenge of identifying the objects that are currently in view. Each of our photoreceptors encodes light intensity at a particular position, and thus the photoreceptor population represents the visual environment as patterns of light. These light patterns can differ substantially with natural variation in an object, such as changes in its position on our retina, or its retinal size as the object moves toward us, yet we have no problems identifying objects across these identity-preserving transformations. This is because our brains successfully "group" all the light patterns that contain the same object and differentiate those from the light patterns that contain different objects (DiCarlo & Cox, 2007; Hung, Kreiman, Poggio, & DiCarlo, 2005a).

The need to group neural responses into sets applies to more cognitive challenges as well. For example, signaling whether an item is a member of a particular category (e.g., whether a chair is a piece of furniture) requires our brains to group the sensory representations of different items (e.g., Freedman, Riesenhuber, Poggio, & Miller, 2001). Moreover, we know that our brains can flexibly group the same items in different ways, depending on task demands, as exemplified by the Wisconsin card sorting task in which subjects are instructed to take a deck of cards and switch between groupings of cards with the same shape, color, and number of items (Berg, 1948). In these and similar cases, our brains transform lower-level population representations in which the task solution exists but is hidden (because the neural responses to the members of different groups are intermingled in the population representation), into higher-level population representations in which the appropriate neural responses are grouped together and the task solution can be extracted.

Understanding neural transformations at multiple levels

Explicit representations tend not to emerge until higher stages of neural processing. In these high-level brain areas, neural response properties tend to be diverse, and this diversity is thought to be advantageous insofar as a population that contains a diversity of neural responses is capable of performing a diversity of tasks (Rigotti et al., 2013). However, response heterogeneity also makes these high-level brain areas difficult to understand using classical single-neuron approaches, which inherently rely on identifying regularities in the response properties of individual neurons across a population (e.g., discovering that the majority of V1 neurons are tuned for orientation).

Inspired in large part by the proposals of David Marr (1982), we and many others have converged on using

a multilevel approach to understand how heterogeneous brain areas process information. Specifically, we have found it useful to begin by taking a population-based approach (Level 1), which has proven to be an effective way to investigate heterogeneous brain areas (e.g., Churchland et al., 2012; Hung, Kreiman, Poggio, & DiCarlo, 2005b; Machens, Romo, & Brody, 2010; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Rigotti et al., 2013) as it inherently focuses on how the combined population response reflects a specific type of information. Next, we can use these population-level descriptions to constrain explanations at the response mechanism level (Level 2), where we focus on determining the single and multineuron mechanisms that give rise to the population representation. Finally, we can use both population and response mechanism level descriptions to constrain descriptions at the computational level (Level 3), which seeks to describe the computations that transform signals from one stage to the next. Below we review how we have applied this multilevel approach to understand how the brain reformats task-relevant information to make it explicitly accessible, by first describing each level in more detail, followed by its application to two specific examples.

Level 1—Population representations: Implicit, explicit, and "untangling"

To envision how a population of neurons might represent information, we begin by considering the population response to some condition (e.g., the visual response to a particular image) at one point in time as a *population response vector*, defined as the pattern of spike count responses produced by each neuron in the population on a single trial (figure 30.1A). This vector lies in a space whose dimensionality is defined by the number of neurons in the population but is most easily envisioned in a two-dimensional space that corresponds to the responses of just two neurons. Because neurons are noisy, the response vector for a given condition can

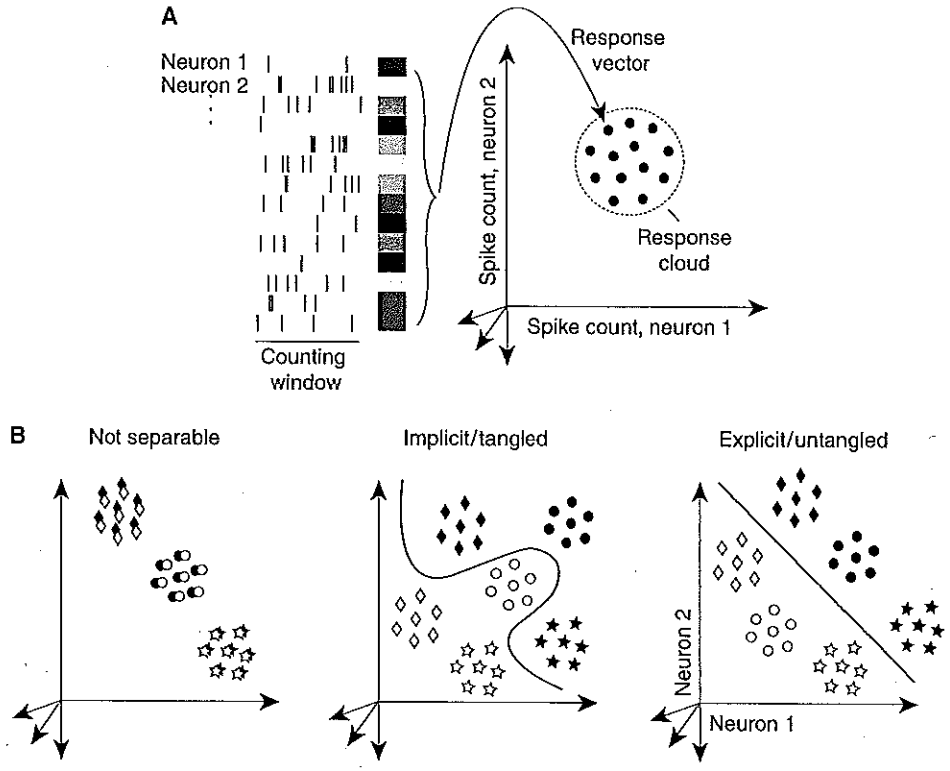

FIGURE 30.1 Population representations. (A) The spike count responses for N neurons combine to form a "response vector" of length N. This response vector exists in an N-dimensional space but is illustrated in the two-dimensional space defined by the responses of two neurons (one dot). Because neurons are noisy, repeated trials of the same condition produce slightly different response vectors, and together all trials form a "response cloud." (B) Shown are the hypothetical responses for a two-way discrimination task that requires parsing the white and black sets of response clouds. Panels show scenarios in which the information required for this task does not exist (*left*); the information required for this task is present but requires a highly nonlinear population readout (i.e., is implicit or tangled; *middle*); the information required for this task can be accessed via a linear readout of the population (i.e., is explicit or untangled; *right*).

fall at slightly shifted positions within the population space, and this distribution is called a *population response cloud* (figure 30.1A).

Tasks that require the brain to group different conditions (e.g., identify a face across changes in position, size, and pose) amount to associating the response clouds that belong to the same set and differentiating those from response clouds that belong to different sets (DiCarlo & Cox, 2007; DiCarlo, Zoccolan, & Rust, 2012). The amount of total information available in the population to perform such a task depends on the degree to which the response clouds corresponding to different sets are nonoverlapping (figure 30.1B). When information exists in a population, we envision that the brain might group responses to different members of a set by placing *decision boundaries* in this space that separate the response clouds for different sets (e.g., a boundary that parses the response clouds corresponding to one face presented at different views from the response clouds for all the views of other faces). The shape of these decision boundaries corresponds to the *population readout rule* required to discriminate these sets based on the population response. These decision boundaries can range from simple rules that correspond to lines or hyperplanes to complex, nonlinear rules that correspond to boundaries that are curved and contorted (compare figure 30.1B, center vs. right). Thus two populations could (in theory) have the same amount of total information to discriminate two sets, but that information could be formatted very differently and require differently shaped decision boundaries to parse them.

The shape of the decision boundary required to extract information determines whether a certain type of information (e.g., object identity) is represented in a manner that is implicit or explicit: "implicit" information is defined as information that requires a complex, highly nonlinear readout, whereas "explicit" information can be extracted using a simple readout rule (e.g., linear). The rationale behind this distinction is the notion that the entities that extract information from a population of neurons are higher-level neurons, and thus our models of neural machinery can be used to guide our determination about whether a representation is implicit or explicit. The simplest decision boundary, a line or a hyperplane, corresponds to the most basic model of a neuron—one that receives weighted input from a population, followed by a threshold, to produce a response that signals when a particular event occurs (e.g., when a particular object is in view). Slightly more complex decision boundaries correspond to slightly more complex models of readout neurons (e.g., a "bent hyperplane" follows from a model neuron with divisive normalization). However, highly complex decision boundaries are likely to be beyond the machinery that can be implemented by individual neurons and instead reflect scenarios in which the information must be reformatted before it becomes accessible to a neurally plausible readout. Because implicit and explicit information loosely map onto nonlinear and linear decision boundaries, respectively, the reformatting process from a nonlinear or "tangled" representation into a linear or "untangled" representation has been coined as "untangling" (DiCarlo & Cox, 2007).

Level 2—The response mechanisms that support explicit representations

As a proxy for measuring how a specific type of information is represented by a population, one can measure how well a population of neurons can solve a particular task (with a particular type of decision boundary). Population performance depends on several factors that are largely nonexclusive. First and foremost, population performance depends on the information conveyed by individual neurons. Notably, population performance also depends on population size. To gain some intuition for this relationship, we can calculate how the distance between two population response clouds ("population discriminability") depends on the discriminability of individual neurons. For a two-way classification (e.g., face 1 versus 2) and a linear decision boundary, one can calculate the commonly used measure of single neuron discriminability, d', as the difference between the mean firing rate responses to each set, divided by the pooled standard deviation of the two distributions (figure 30.2A). This single-neuron measure can be extended to a measure of the distance between two response clouds in the population space (the "normalized Euclidean distance," or NED), computed as the square root of the summed squared d' for all neurons (figure 30.2A). Consequently, the distance between two population response clouds increases as a function of the number of neurons in a population, and even small single-neuron d' can translate into large population discriminability given enough neurons.

Population performance also depends on the number of different sets that need to be parsed. Multiway classification problems (e.g., which of 100 possible objects is currently in view?) are often envisioned as multiple two-way classifications (A/not-A, B/not-B, etc.), followed by a "max" operation to determine a population's final answer (figure 30.2B). Thus one can envision that the solution to a multiway classification problem is computed by parsing the population space with multiple

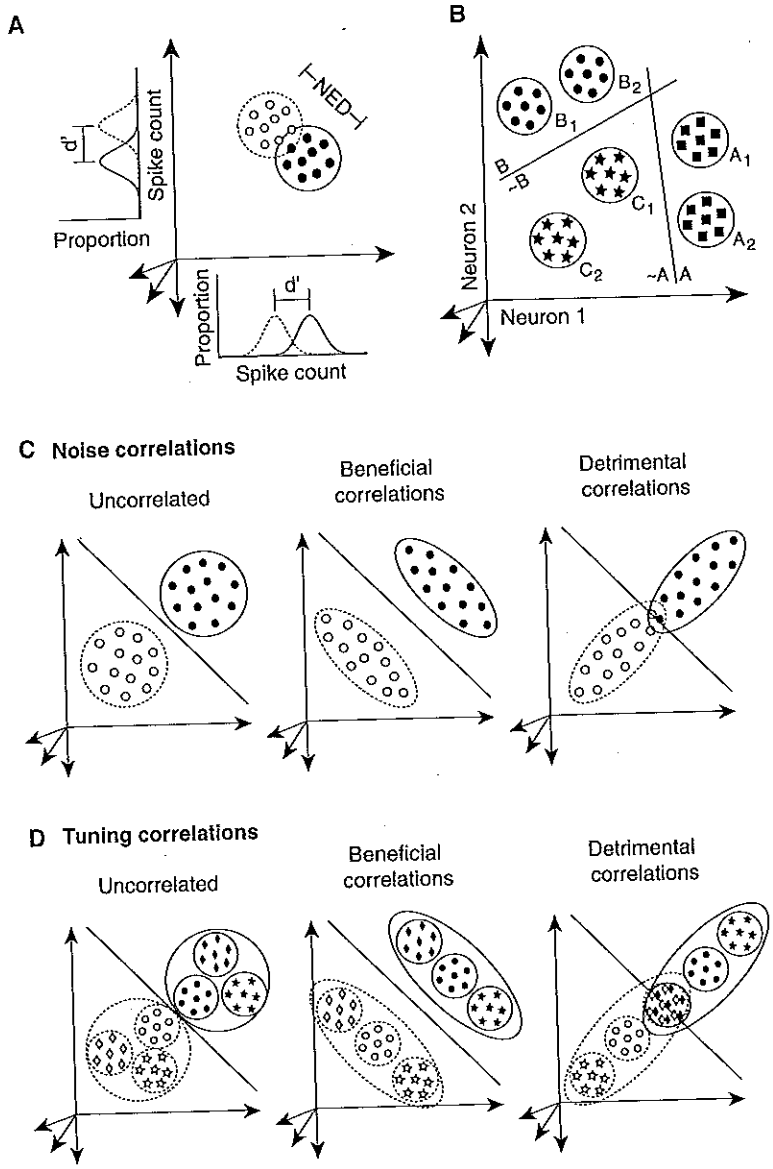

FIGURE 30.2 Factors that determine population performance. (A) The distance between population-response clouds depends on the information conveyed by single neurons and grows as a function of population size. Shown are the responses of two hypothetical neurons, each with a $d' = 3$ for two conditions. The normalized Euclidean distance (NED) for these two conditions in the 2D population space can be calculated as $NED = \sqrt{3^2 + 3^2} = 4.2$. (B) An N-way classification can be envisioned as parsing the population space into N two-way classifications (e.g., A/not A, B/not B, etc.). (C) Noise correlations are determined by the trial-by-trial variability between neurons, and they determine the shape of the individual response clouds. (D) Tuning correlations are determined by the mean firing-rate responses within and across sets, and they determine the relative positions of the different response clouds. For (C) and (D), shown are hypothetical scenarios that include no correlations, correlations that are beneficial because they are aligned to the decision boundary, and correlations that are detrimental because they are perpendicular to the decision boundary.

decision boundaries, each of which separates the response clouds corresponding to one set from all the other sets.

Population performance also depends on "correlations," which come in two varieties. "Noise correlations," defined by the correlated trial-by-trial variability between neurons, determine the shapes of the individual response clouds (i.e., uncorrelated noise has a spherical shape whereas correlated noise is oblong; figure 30.2C). Whether this type of correlation is beneficial or detrimental to population performance depends on how these correlations are aligned relative to the decision boundaries (figure 30.2C). In contrast, "tuning correlations" determine how the response

clouds for the different conditions within a set are positioned relative to one another. This type of correlation can also be beneficial or detrimental to population performance, depending on how the response clouds within a set align relative to the decision boundaries (figure 30.2D). See chapter 29, this volume, for a more extensive description of noise correlations and their impact on population representation.

Finally, population performance depends not only on the shape of the decision boundary selected, but also on the details by which the decision boundary is positioned within the population space. For example, a linear decision boundary might be placed midway between the means of two sets of response clouds, or, alternatively, another position along this vector may be more appropriate (e.g., because one set of response clouds has a larger variance). Returning to the notion that higher-level neurons are responsible for reading out neural populations at an earlier stage, these decision boundaries are presumably positioned via the learning rules by which neurons are wired together (discussed in more detail below). Issues related to how classification performance depends on the manner by which decision boundaries are positioned can be informed by the rich engineering literature focused on machine learning and information processing (e.g., Manning, Raghavan, & Schutze, 2008). In practice, absolute population performance values depend on many factors, including the number of classifications, the number of neurons, the specific details about the classification scheme and its optimization, and thus it is often difficult to interpret absolute levels of performance (i.e., "75% performance" means little by itself). Thus population performance values are almost always studied as comparisons with these parameters fixed—for example, between different populations of neurons or between different conditions within the same population.

Level 3—The computations that produce explicit representations

In comparison to the other two levels, understanding the computations that the brain uses to transform information is probably the most challenging. This is because the other two levels tend to lend themselves more to pure data-analysis approaches (e.g., a comparison of classifier performance for two populations), whereas arriving at a computational description most often involves the more challenging task of finding and fitting a model that is sufficiently simple to be constrained by the data but at the same time is sufficiently complex to provide a good account of most neurons. One popular approach attempts to describe the response properties of neurons within a brain area using variants of the "linear-nonlinear" (LN) model, in which the computations performed by individual neurons are described as a weighted sum of the firing rate responses from an input brain area, followed by the application of an instantaneous nonlinearity (e.g., thresholding) to produce an output firing rate response. Due to the relative simplicity of this type of model, one can often identify regularities across the models fit to different neurons and arrive at an intuitive yet accurate description of "how" the response properties of neurons in a particular brain area are constructed from the responses of the input population (Adelson & Bergen, 1985; Carandini et al., 2005; Heeger, 1993; Rust, Mante, Simoncelli, & Movshon, 2006; Rust, Schwartz, Movshon, & Simoncelli, 2005; Simoncelli & Heeger, 1998). See chapter 29, this volume, for a more extensive description of the LN model concept.

One attractive proposal that can be regarded as an extension of the LN model concept is that each cortical brain implements the same "canonical" computation, albeit with different inputs, and thus achieves different goals (Douglas & Martin, 1991; Fukushima, 1980; Heeger, Simoncelli, & Movshon, 1996; Kouh & Poggio, 2008; Riesenhuber & Poggio, 1999). The appeal of this idea arises in part from the fact that iterative stacks of simple, LN canonical computational elements are known to be capable of powerful computations, including untangling (Fukushima, 1980; Riesenhuber & Poggio, 2000; Serre, Oliva, & Poggio, 2007), and one can envision how these computational elements might arise from relatively simple genetic programming. As a refinement of these ideas, some have proposed that we should focus on identifying whether a particular canonical quantity is optimized at each stage of processing through a learning process, such as the degree to which input responses are locally untangled within subpopulations of neighboring neurons (DiCarlo et al., 2012).

The ideas associated with each of these levels of explanation are elaborated in more detail below, where we describe two examples of how complementary descriptions at the population, response mechanism, and computational levels have been combined to understand how the brain transforms task-relevant information from an implicit format into an explicit representation.

Example 1—Explicit representations of object identity

LEVEL 1: POPULATION REPRESENTATIONS OF OBJECTS

One example in which the brain is thought to untangle information is the case of invariant object recognition.

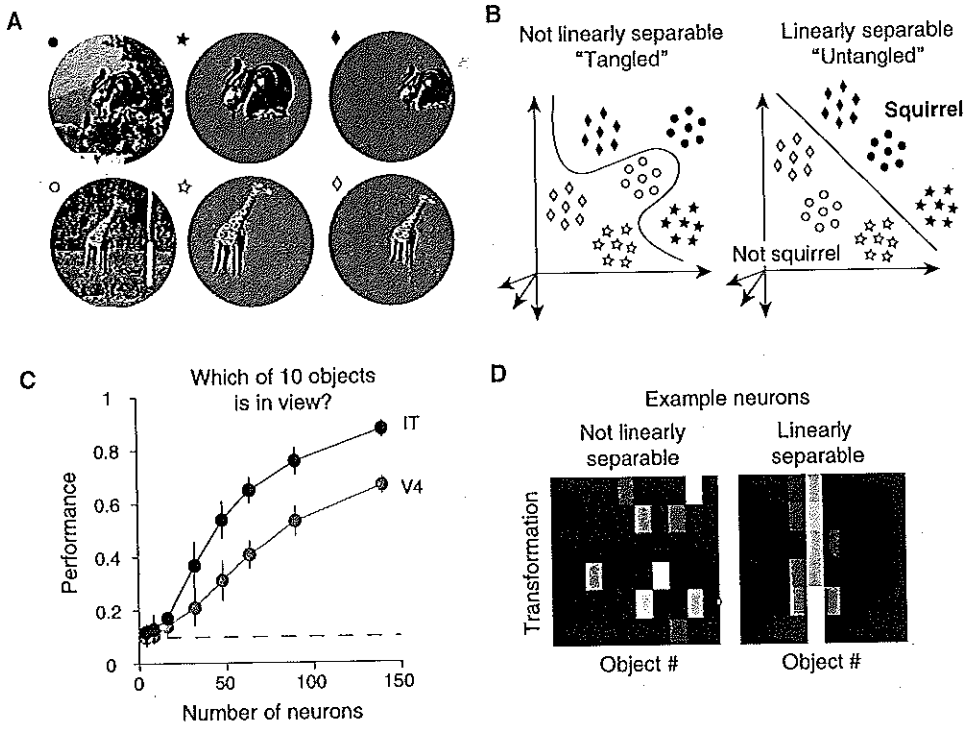

FIGURE 30.3 Explicit representations of object identity. As reported by Rust and DiCarlo (2010). (A) A subset of images used to probe the representation of object identity, across identity-preserving transformations, in V4 and the inferotemporal cortex (IT). In total, 10 objects were presented under six different transformations that included changes in the objects' position, size, and background context. (B) The problem of invariant object recognition can be formulated as multiple, two-way classifications (see also figure 30.2B). (C) Object classification performance in V4 and IT using a linear readout, reprinted from Rust and DiCarlo (2010). (D) Firing-rate response surfaces for 10 objects presented under six different transformations for one example V4 neuron that is not linearly separable, and one example IT neuron that is linearly separable. Single-neuron linear separability translates into the preservation of object preferences across these transformations, or equivalently a separable response matrix.

Information about objects is processed in the primate brain along the ventral visual pathway, a hierarchically arranged collection of neural structures that includes the retina and lateral geniculate nucleus (LGN) as well as cortical brain areas V1, V2, and V4 and inferotemporal cortex (IT; Felleman & Van Essen, 1991; reviewed by DiCarlo et al., 2012). Within each structure, the response vectors corresponding to all possible identity-preserving transformations of the same object (e.g., changes in position, scale, pose, etc.) define an *object manifold* (DiCarlo & Cox, 2007). As described above, the population of photoreceptors represents the visual scene as patterns of light, and because natural object variation produces markedly different light patterns, the manifolds for different objects are "tangled" in the photoreceptor representation. Thus, to extract object identity information, the brain must untangle these signals. To evaluate where and how the brain accomplishes this, we and others have recorded and/or simulated the responses of neurons within different ventral visual pathway brain areas to images of objects presented at different positions and sizes and within different background contexts (figure 30.3A; Hung et al., 2005a; Li, Cox, Zoccolan, & DiCarlo, 2009; Rust & DiCarlo, 2010). We then evaluated the degree to which object identity was explicitly represented at each stage by probing how well a linear decision boundary could separate the responses corresponding to one object from the others using a cross-validated linear classification scheme (figure 30.3B). These results revealed that a linear readout of object identity applied to the IT population performs robustly with only a few hundred neurons and that the IT population performs invariant object recognition tasks much better than real or simulated populations at earlier stages, such as V1 and V4 (figure 30.3C; Hung et al., 2005a; Freiwald & Tsao, 2010; Li et al., 2009; Rust & DiCarlo, 2010; figure 30.3A). These results support the notion that the ventral visual pathway reformats object manifolds to make object identity explicit at its final stage (DiCarlo & Cox, 2007).

LEVEL 2: THE RESPONSE MECHANISMS UNDERLYING OBJECT RECOGNITION What single-neuron responses correspond to "tangled" and "untangled" object manifolds? At the final stage of the pathway (in IT), neural responses across transformations (i.e., for changes in position) tend to change in a manner that maintains their rank-order selectivity preferences for objects across identity-preserving transformations (Ito, Tamura, Fujita, & Tanaka, 1995; Zoccolan, Kouh, Poggio, & DiCarlo, 2007). This property, coined *tolerance,* translates into single-neuron response surfaces that correspond to different object identities presented at different transformations that are linearly separable, or "untangled" (figure 30.3D; Li et al., 2009; Rust & DiCarlo, 2010). In contrast, at early stages of the ventral visual pathway, such as the retina, LGN, and V1, neurons have small receptive fields that are activated by simple light patterns, and the object identity response surfaces at earlier stages of the pathway are nonlinearly separable, or "tangled" (figure 30.3D; Li et al., 2009; Rust & DiCarlo, 2010). These nonlinearly separable responses within individual neurons translate to population object manifolds that are "tangled" together, similar to pieces of paper crumpled into a ball. Similarly, the untangled single-neuron responses observed in IT translate into object manifolds that are more "untangled," in that they are both more flat and are more separated from one another. At intermediate stages of processing, such as V4, neurons appear to have intermediate response properties, consistent with gradual untangling along the pathway.

LEVEL 3: THE COMPUTATION OF OBJECT IDENTITY The gradual transformation from the tangled, light-pattern representation encoded by the eye into the untangled representation of objects found in IT is often envisioned as an iterative cascade of two types of computations: selectivity and invariance. Selectivity computations combine inputs (that each respond to different visual features) with what is often described as an "AND-like" operation to create a neuron that responds more robustly to the conjunction of the features (e.g., features A and B presented together) than the response predicted from the features presented individually (e.g., the response to feature A alone + the response to feature B alone). Selectivity computations are responsible for transforming the representation from its pattern-of-light based form into one based on the conjunctions of visual features that define objects. In contrast, invariance computations act with what is often described as an "OR-like" operation to create neurons that respond whenever any of the features encoded by any of its inputs are in view. Invariance computations are responsible for transforming the small receptive fields found in the eye into the larger receptive fields found at higher stages of the pathway and, similarly, for creating tolerances for the other identity-preserving transformations (e.g., size and background context).

Prototypical examples of selectivity and invariance computations are those used to describe the construction of V1 simple and complex cells (Hubel & Wiesel, 1962). In this description, V1 simple cells implement selectivity operations on the LGN inputs to produce an orientation-selective response. Next, V1 complex cells implement polarity (i.e., light/dark) invariance by combining simple cells tuned for the same orientation. Together, these two operations accomplish a small amount of "untangling" in the visual representation by transforming it from the center-surround representation found in the LGN to one based on local orientation partially invariant to polarity and position. Models of biological object recognition have extended this selectivity/invariance framework to an iterative, hierarchical cascade of these operations that loosely maps onto the processing thought to occur at different stages of the ventral visual pathway (Fukushima, 1980; Riesenhuber & Poggio, 1999; Serre et al., 2007). These relatively simple hierarchical models are largely successful at producing an "untangled" representation of object identity at their highest stage. However, they cannot completely account for human object-recognition performance, and some have speculated that doing so will require a deeper, more systematic exploration of this large class of models (DiCarlo et al., 2012).

For a complementary perspective on the neural computations underlying invariant object recognition, we can consider how the ventral visual pathway might "wire itself up" to untangle information about object identity. The scheme described above implies that neurons learn to appropriately position the decision boundaries at each stage such that object identity is ultimately untangled. How might these boundaries be determined? One proposal is that the brain might exploit the fact that the identities of the objects that are in view tend to change more slowly than the specific light patterns encoded by the eye (Foldiak, 1991; Stryker, 1992; Wallis & Rolls, 1997; Wiskott & Sejnowski, 2002). This notion incorporates both changes due to variation as objects move as well as changes in the light patterns produced by a moving observer. For example, when we move our eyes to scan a scene and across saccades, the same object is likely to fall at different positions on our retina, thus producing different light-pattern representations and adjacent points in time.

These naturally occurring temporal contiguity cues can instruct the building of tolerance to identity-preserving

transformations. In the population-based description presented in figure 30.3A, the notion is that the response vectors that are produced by the retina at points close together in time tend to be those that correspond to identity-preserving image variation. Thus, a learning rule that attempts to produce similar patterns of neural responses at points close in time can achieve the larger goal of creating an untangled object representation. Experiments targeted at testing these ideas have manipulated the statistics of temporal contiguity cues within the sensory input (e.g., by changing object identity across saccades) and then evaluated the impact of these manipulations on human behavioral judgments (Cox, Meier, Oertelt, & DiCarlo, 2005) and neural responses in IT (Li & DiCarlo, 2008, 2011). The results of these experiments support the notion that invariant object recognition mechanisms are shaped via an unsupervised learning process, and they suggest that even in the adult brain, invariant object representations are constantly being recalibrated.

Example 2—Explicit representations during target search

LEVEL 1: THE POPULATION REPRESENTATION OF TARGET MATCHES The need to reformat task-relevant information to make it explicitly available is not confined to the realm of perception (i.e., determining the content of the sensory environment), but naturally extends to more cognitive tasks as well. One example is the case of finding target visual objects. Finding a target object, such as your wallet, requires your brain to not only determine the items you are looking at, but also to compare this visual representation with a working memory representation of what you are looking for. To act on the event of finding your target, your brain must transform this information into an "I found it" signal that reports when a target is in view, no matter what that target is (e.g., a neuron that fires when you view your wallet or your car keys, but only when you are looking for those items and not when you are looking for other things). Creating such a signal requires nonlinear computation (i.e., conjunctions of visual and working memory information) and, before these signals have been appropriately combined, the representation of whether a currently viewed object is a target match or not will be tangled.

While the process is not well understood, visual and target-specific information are thought to be combined at middle to higher stages of the ventral visual pathway (i.e., V4 and IT), where the firing rate responses of neurons are largely visual, but are also modulated by changing the identity of a target (e.g., Eskandar, Richmond, & Optican, 1992; Haenny, Maunsell, & Schiller, 1988; Miller & Desimone, 1994). To evaluate the degree to which target match signals are explicitly represented within and just beyond the ventral visual pathway, we recorded the responses of neurons in IT and one of its projection areas, perirhinal cortex (PRH), as monkeys performed a task in which they had to indicate when a target image appeared within a sequence of distractor images (figure 30.4A; Pagan, Urban, Wohl, & Rust, 2013). We then evaluated the degree to which each population could distinguish the same images presented as targets and as distractors with a linear readout by applying a cross-validated linear classification scheme (figure 30.4B). We found that PRH performed this task better than IT, suggesting that PRH has more untangled target match information (figure 30.4C).

More untangled target match information in PRH as compared to IT could follow from a scenario in which PRH has more total information for this task than IT because it receives information that IT does not (compare figure 30.1B, left vs. right). Alternatively, more untangled target match information in PRH could follow from a scenario in which the two areas have similar total information, but that information is more tangled in IT and more untangled in PRH (compare figure 30.1B, center vs. right). To discriminate between these alternatives, we measured the amount of total target match information using an ideal observer classifier whose performance depended on the distance between the response clouds for target matches and distractors but not on the relative positioning of response clouds within each set. We found total information to be similar in IT and PRH, consistent with a description in which visual and working memory information are combined within or before IT in the ventral visual pathway in a largely tangled fashion, and then this information is sent to PRH, which then untangles it.

LEVEL 2: THE RESPONSE MECHANISMS UNDERLYING TARGET SEARCH As described above, finding a specific target requires the brain to compare visual and working memory signals. Within both IT and PRH, we found example neurons whose responses reflected pure versions of these signals, as illustrated by firing-rate response matrices defined by each image viewed in the context of every image as a target, where visual and working memory signals are represented with vertical and horizontal matrix structures, respectively (figure 30.4D). The solution to the monkeys' task required differentiating target matches, which fall along the diagonal of these response matrices, from distractors, which fall off the diagonal, and we also

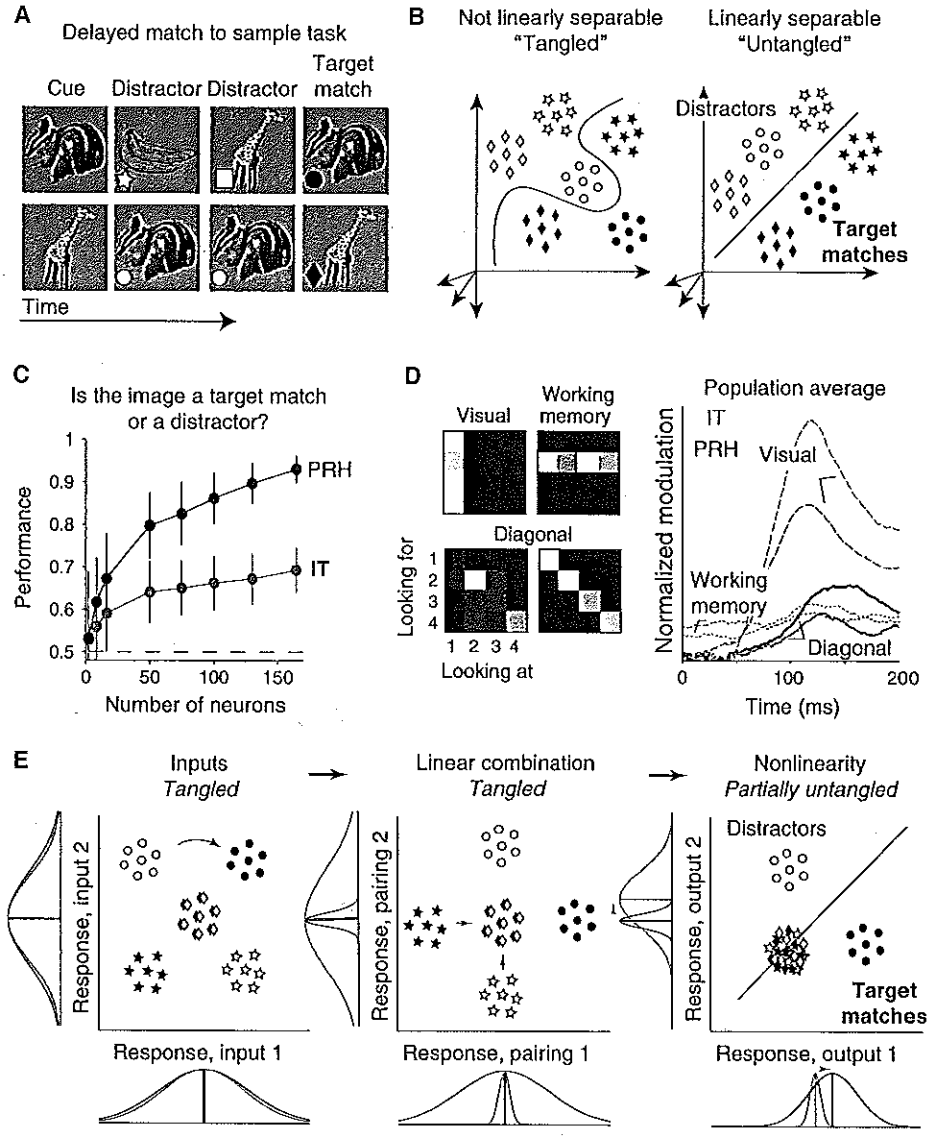

FIGURE 30.4 Explicit representations during target search. As reported by Pagan et al. (2013). (A) Monkeys performed a delayed match to sample task in which a cue image was presented, followed by a random number of distractor images and then a target match. The task required the monkeys to fixate throughout the presentation of the distractors and make a downward saccade in response to the match. The experimental design involved presenting four images in all possible combinations as a visual stimulus and a target. Throughout the figure, the different images are indicated with different shapes, with target match conditions in black and distractors in white. (B) This task can be formulated as a two-way classification of the same images presented as target matches and as distractors. For simplicity, only a subset of the conditions is depicted. (C) Target match/distractor linear classification performance measured in IT and perirhinal cortex (PRH; reprinted from Pagan et al., 2013). (D) Firing-rate response matrices for four example neurons (*left*). Amounts of each type of signal as function of time relative to stimulus onset and averaged across each population, computed via a bias-corrected, ANOVA-like decomposition (*right*; described by Pagan et al., 2013). (E) A linear-nonlinear model provided a good account of the transformation from IT to PRH. This model untangled target match information by combining IT neurons that had tuning correlations for target matches and distractors that were asymmetric (e.g., positive tuning correlations for target matches and negative correlations for distractors are shown in this hypothetical example). Pairing such neurons with appropriate weights resulted in a rotation in the population space that resulted in variance differences between the firing-rate response distributions to target matches and distractors (depicted at the bottom and left of the plot). These variance differences were then exploited to increase linear separability via an instantaneous nonlinearity that included thresholding and/or saturation.

found neurons with diagonal matrix structure. These included neurons that responded to individual images but only when they were presented as targets, which were found in both IT and PRH, as well as a handful of compelling PRH neurons that responded whenever any image was presented as a target (figure 30.4D, "diagonal"). Notably, the intuitive neurons described above were more the exception than the rule in both brain areas, as we found that most neurons were heterogeneous and difficult-to-describe mixtures of different types of signals (not shown). To quantify the magnitudes of different types of signals within each population, we developed a method to parse the responses of each neuron into different types of signals (e.g., visual, working memory, and diagonal; figure 30.4D), and we used these quantifications to constrain our models of how signals were reformatted between IT and PRH, as described below. This analysis confirmed that higher performance in PRH as compared to IT corresponded to an increased amount of diagonal structure within the response matrices of PRH neurons (figure 30.4D), consistent with other types of IT signals that were reformatted into diagonal signals in PRH. We also investigated whether noise correlations might contribute to higher PRH population performance by analyzing the responses of small, simultaneously recorded populations, but we found no evidence for that in our data. This result is likely due to the fact that performance in this task is more limited by the separation of the responses to different target match and different distractor conditions, which must be grouped together (figure 30.2D), than it is by the noise correlations, which determine the shapes of the individual response clouds (figure 30.2C).

LEVEL 3: THE COMPUTATIONS RESPONSIBLE FOR UNTANGLING TARGET MATCH SIGNALS Probing computations in high-level brain areas like PRH can be challenging, particularly in the absence of a complete understanding of all of the computations up to the brain area of interest (e.g., to describe computations in PRH that act on the inputs from IT, a model of processing up to and including IT). In the absence of such a model of IT, we circumvented these challenges by developing novel methods to "leap-frog" into the system and fit a simple linear-nonlinear model of PRH that described how our recorded IT responses were reformatted to produce a more explicit representation of target matches (despite an incomplete understanding of how these IT responses came to be). To constrain the model, we assumed that the untangling process in PRH acted optimally on its inputs arriving from IT, and we searched for the combinations of IT neurons and the model parameters that would maximally untangle information. The resulting model PRH population matched task performance of the actual PRH population for target match–distractor distinctions, and remarkably, this model of PRH also matched the PRH data in many other respects as well, including a decrease in the amount of visual modulation in PRH as compared to IT (figure 30.4D).

An investigation into the mechanism that the model used to untangle target match information revealed that the model worked by combining IT neurons with slightly offset tuning preferences (i.e., offset tuning correlations) for target matches as compared to distractors (figure 30.4E, left). Linear combinations of such neurons acted to rotate the population representation to produce differences in the variance across the responses to target matches as compared to the variance across the responses to distractors (figure 30.4E, middle). Once produced, nonlinearities could act on these response variance differences to produce a more untangled target match representation (figure 30.4E, right). We found that the offset-tuning preferences that this mechanism relied on were ubiquitously present in our data (and, by extension, any population that reflects heterogeneous mixtures of visual and target signals). This suggests that the connectivity responsible for untangling IT inputs within PRH could reasonably be determined via reinforcement learning during the natural experience of searching for targets.

Conclusion

Many tasks require our brains to group neural responses via transformations from implicit representations in which the responses to different sets are intermingled into explicit representations in which these groupings can be easily accessed. While described above for the specific problem of object recognition, the challenge of transforming the elemental representation encoded by our sensory receptors into an explicit representation of the content in our environment is a common perceptual problem. Other examples include identifying motion direction invariant to the moving pattern (Movshon, Adelson, Gizzi, & Newsome, 1985); determining the relative depths (Thomas, Cumming, & Parker, 2002; Umeda, Tanabe, & Fujita, 2007) or tilt (Nguyenkim & DeAngelis, 2003) of two surfaces independent of absolute depth; and independently determining the pitch and location of a sound (Walker, Bizley, King, & Schnupp, 2011). Similarly, target search is one instantiation of a larger class of "cognitive flexibility" problems in which the brain must flexibly switch between different states (in this case between different

visual targets) by modifying its task-relevant working memory signals. To solve these tasks, the brain must combine task-relevant working memory and sensory information in a manner that ultimately produces an explicit representation that indicates when the state of the environment matches the task goals.

A complete understanding of how the brain transforms implicit information into an explicit format requires a complementary approach at multiple levels of explanation. For at least two tasks, identifying objects and flexibly switching between different visual targets, explicit representations are produced in high-level brain areas (IT and PRH), in which neural responses are heterogeneous and difficult-to-understand mixtures of different types of information. To describe how the brain solves these tasks, we have found it useful to first constrain explanations with population-level descriptions, followed by a determination of the specific single-neuron response mechanisms that give rise to those population representations. We have found that both types of data, in turn, provide useful and important constraints for computational-level descriptions of how the brain reformats and processes information.

ACKNOWLEDGMENTS NCR was supported by National Eye Institute Grant (R01EY020851), NSF CAREER 1265480 and the Alfred P. Sloan Foundation and the McKnight Endowment Fund for Neuroscience.

REFERENCES

ADELSON, E. H., & BERGEN, J. R. (1985). Spatiotemporal energy models for the perception of motion. *J Opt Soc Am A, 2,* 284–299.

BERG, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *J Gen Psychol, 39,* 15–22.

CARANDINI, M., DEMB, J. B., MANTE,V., TOLHURST, D. J., DAN, Y., OLSHAUSEN, B. A., ... RUST, N. C. (2005). Do we know what the early visual system does? *J Neurosci, 25,* 10577–10597.

CHURCHLAND, M. M., CUNNINGHAM, J. P., KAUFMAN, M. T., FOSTER, J. D., NUYUJUKIAN, P., RYU, S. I., & SHENOY, K. V. (2012). Neural population dynamics during reaching. *Nature, 487,* 51–56.

COX, D. D., MEIER, P., OERTELT, N., & DICARLO, J. J. (2005). "Breaking" position-invariant object recognition. *Nat Neurosci, 8,* 1145–1147.

DICARLO, J. J., & COX, D. D. (2007). Untangling invariant object recognition. *Trends Cogn Sci, 11,* 333–341.

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

DOUGLAS, R. J., & MARTIN, K. A. (1991). A functional microcircuit for cat visual cortex. *J Physiol, 440,* 735–769.

ESKANDAR, E. N., RICHMOND, B. J., & OPTICAN, L. M. (1992). Role of inferior temporal neurons in visual memory. I. Temporal encoding of information about visual images, recalled images, and behavioral context. *J Neurophysiol, 68,* 1277–1295.

FELLEMAN, D. J., & VAN ESSEN, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex, 1,* 1–47.

FOLDIAK, P. (1991). Learning invariance from transformation sequences. *Neural Comput, 3,* 194–200.

FREEDMAN, D. J., RIESENHUBER, M., POGGIO, T., & MILLER, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291,* 312–316.

FREIWALD, W. A., & TSAO, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science, 330,* 845–851.

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

HAENNY, P. E., MAUNSELL, J. H., & SCHILLER, P. H. (1988). State dependent activity in monkey visual cortex. II. Retinal and extraretinal factors in V4. *Exp Brain Res, 69,* 245–259.

HEEGER, D. J. (1993). Modeling simple-cell direction selectivity with normalized, half-squared, linear operators. *J Neurophysiol, 70,* 1885–1898.

HEEGER, D. J., SIMONCELLI, E. P., & MOVSHON, J. A. (1996). Computational models of cortical visual processing. *Proc Natl Acad Sci USA, 93,* 623–627.

HUBEL, D. H., & WIESEL, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol, 160,* 106–154.

HUNG, C. P., KREIMAN, G., POGGIO, T., & DICARLO, J. J. (2005a). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310,* 863–866.

HUNG, C. P., KREIMAN, G., POGGIO, T., & DICARLO, J. J. (2005b). Ultra-fast object recognition from few spikes. MIT AI Memo 2005–022.

ITO, M., TAMURA, H., FUJITA, I., & TANAKA, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J Neurophysiol, 73,* 218–226.

KOUH, M., & POGGIO, T. (2008). A canonical neural circuit for cortical nonlinear operations. *Neural Comput, 20,* 1427–1451.

LI, N., & DICARLO, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science, 321,* 1502–1507.

LI, N., & DICARLO, J. J. (2011). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron, 67,* 1062–1075.

LI, N., COX, D. D., ZOCCOLAN, D., & DICARLO, J. J. (2009). What response properties do individual neurons need to underlie position and clutter "invariant" object recognition? *J Neurophysiol, 102*(1), 360–376.

MACHENS, C. K., ROMO, R., & BRODY, C. D. (2010). Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. *J Neurosci, 30,* 350–360.

MANNING, C. D., RAGHAVAN, P., & SCHUTZE, H. (2008). *Introduction to information retrieval.* Cambridge, UK: Cambridge University Press.

MARR, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt.

MEYERS, E. M., FREEDMAN, D. J., KREIMAN, G., MILLER, E. K., & POGGIO, T. (2008). Dynamic population coding of

category information in inferior temporal and prefrontal cortex. *J Neurophysiol, 100,* 1407–1419.

Miller, E. K., & Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science, 263,* 520–522.

Movshon, J. A., Adelson, E. H., Gizzi, M. S., & Newsome, W. T. (1985). The analysis of moving patterns. *Exp Brain Res, 11*(Suppl.), 117–151.

Nguyenkim, J. D., & DeAngelis, G. C. (2003). Disparity-based coding of three-dimensional surface orientation by macaque middle temporal neurons. *J Neurosci, 23,* 7117–7128.

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. *Nat Neurosci, 16,* 1132–1139.

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

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nat Neurosci, 3*(Suppl.), 1199–1204.

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497,* 585–590.

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

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005). Spatiotemporal elements of macaque V1 receptive fields. *Neuron, 46,* 945–956.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nat Neurosci, 9,* 1421–1431.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc Natl Acad Sci USA, 104,* 6424–6429.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vis Res, 38,* 743–761.

Stryker, M. P. (1992). Neurobiology. Elements of visual perception. *Nature, 360,* 301–302.

Thomas, O. M., Cumming, B. G., & Parker, A. J. (2002). A specialization for relative disparity in V2. *Nat Neurosci, 5,* 472–478.

Umeda, K., Tanabe, S., & Fujita, I. (2007). Representation of stereoscopic depth based on relative disparity in macaque area V4. *J Neurophysiol, 98,* 241–252.

Walker, K. M., Bizley, J. K., King, A. J., & Schnupp, J. W. (2011). Multiplexed and robust representations of sound features in auditory cortex. *J Neurosci, 31,* 14565–14576.

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

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Comput, 14,* 715–770.

Zoccolan, D., Kouh, M., Poggio, T., & DiCarlo, J. J. (2007). Trade-off between object selectivity and tolerance in monkey inferotemporal cortex. *J Neurosci, 27,* 12292–12307.