Psychological Review
2009, Vol. 116, No. 1, 220–251

© 2009 American Psychological Association
0033-295X/09/$12.00 DOI: 10.1037/a0014462

On the Biological Plausibility of Grandmother Cells: Implications for Neural Network Theories in Psychology and Neuroscience

Jeffrey S. Bowers
University of Bristol

A fundamental claim associated with parallel distributed processing (PDP) theories of cognition is that knowledge is coded in a distributed manner in mind and brain. This approach rejects the claim that knowledge is coded in a localist fashion, with words, objects, and simple concepts (e.g. "dog"), that is, coded with their own dedicated representations. One of the putative advantages of this approach is that the theories are biologically plausible. Indeed, advocates of the PDP approach often highlight the close parallels between distributed representations learned in connectionist models and neural coding in brain and often dismiss localist (grandmother cell) theories as biologically implausible. The author reviews a range a data that strongly challenge this claim and shows that localist models provide a better account of single-cell recording studies. The author also contrast local and alternative distributed coding schemes (sparse and coarse coding) and argues that common rejection of grandmother cell theories in neuroscience is due to a misunderstanding about how localist models behave. The author concludes that the localist representations embedded in theories of perception and cognition are consistent with neuroscience; biology only calls into question the distributed representations often learned in PDP models.

Keywords: grandmother cells, distributed representations, parallel distributed processing, localist representation, connectionism

Of the many different schools of neural network modeling, the parallel distributed processing (PDP) approach is the most influential within psychology (McClelland, Rumelhart, & PDP Research Group, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986). A key claim of this approach is that knowledge is coded in a distributed manner in the mind and the brain. That is, knowledge is coded as a pattern of activation across many processing units, with each unit contributing to many different representations. As a consequence, there is no one unit devoted to coding a given word, object, or person. This contrasts with the classic view, according to which knowledge is coded in a localist fashion. That is, individual words, objects, simple concepts, and the like are coded distinctly, with their own dedicated representation. For example, the words *mother* and *father* would be coded with distinct and nonoverlapping mental representations. On any localist account, the words *mother* and *father* would be linked by virtue of sharing some features (e.g., letters), but the words themselves would be stored explicitly and separately in the mind.

The rejection of localist coding schemes in favor of distributed representations is a core principle of the PDP approach (e.g., Plaut & Shallice, 1993; Seidenberg, 1993). But, it is important to note that it is not a principle of neural network theories in general (cf. Bowers, 2002; Feldman & Ballard, 1982). Indeed, many neural networks rely on local representations for their functioning. In some cases, localist representations are built into a model (e.g., Dell, 1986; Morton, 1979; McClelland & Elman, 1986; McClelland & Rumelhart, 1981), and in other cases, they are learned (e.g., Grossberg, 1980, 1987; Kruschke, 1992; Rumelhart & Zipser, 1985). At present, localist and distributed neural networks have been applied to a wide range of cognitive phenomena, and the goal in both cases is to identify a set of general principles that apply across domains.

Given that neural networks can learn either local or distributed representations, why have so many theorists rejected localist representations in favor of distributed ones? Of course, criticisms levied against localist models differ, case-by-case, but one underlying criticism is often raised against the approach as a whole; namely, the models are thought to be biologically implausible. By contrast, advocates of the PDP approach often highlight the link between the distributed representations learned in PDP models and the neural coding systems employed in the brain. For example, in an article titled "Six Principles for Biologically Based Computational Models of Cortical Cognition," O'Reilly (1998) included distributed representations as Principle 2, writing,

> The cortex is widely believed to use distributed representations to encode information. A distributed representation uses multiple active neuron-like processing units to encode information (as opposed to a single-unit, localist representation), and the same unit can participate in multiple representations . . . . Electrophysiological recordings demonstrate that distributed representations are widely used in the cortex. (p. 456)

I would like to thank Patricia Bowers, Colin Davis, Nina Kazanina, Chad Marsolek, and Sven Mattys for reading and commenting on drafts of this article. Finally, I would like to thank Colin Davis for creating Figure 4 and Markus Damian for creating Figures 5A–5B and 11.

Correspondence concerning this article should be addressed to Jeffrey S. Bowers, Department of Experimental Psychology, University of Bristol, 12A Prior Road, Clifton, Bristol BS8-ITU, England. E-mail: j.bowers@bris.ac.uk

In a general review of cognitive neuroscience, McClelland (2001) wrote,

> Studies relying on microelectrodes to record from neurons in the brains of behaving animals can allow researchers to study the representations that the brain uses to encode information, and the evolution of these representations over time . . . . These studies indicate, among other things, that the brain relies on distributed representations. (p. 2137)

Countless other similar quotes could be provided.

At first blush, these claims regarding the relative biological plausibility of localist (grandmother) and distributed coding schemes seem well founded. As reviewed by Gross (2002), few neuroscientists have taken localist representations seriously. Indeed, the term *grandmother cell* (a term often attributed to Jerry Lettvin; cf. Gross, 2002) is generally used to ridicule the claim that complex and meaningful stimuli are coded by individual cells in the cortex. Finkel (1988) called them "infamous grandmother cells" (p. 787). As noted by Connor (2005), "No one wants to be accused of believing in grandmother cells" (p. 1036). Instead, after 50 years of single-cell recording studies, it is widely claimed (and implicitly assumed) that the brain codes information with some form of distributed coding. As Averbek, Latham, and Pouget (2006) put it, "As in any good democracy, individual neurons count for little; it is the population of activity that matters" (p. 358).

Given all this, what is an advocate of localist representations to do? Grandmother cells in brain are the physiological counterpart to localist representations in cognitive models (e.g., Coltheart, Curtis, Atkins, & Haller, 1993; Davis, 1999; Hummel & Holyoak, 2003; McClelland & Rumelhart, 1981; Morton, 1969; Norris, 1994; Page & Norris, 1998). If localist and PDP models are to be compared (and assessed) in terms of their biological plausibility, grandmother cells had best not be a joke.

One response is to reject the assumption that cognitive models should be evaluated on biological criteria. According to Broadbent (1985), this is justified because neuroscience is only relevant at what Marr (1982) called the implementational level of description. On this view, psychological theory should only concern itself with a computational description in which one considers the goals and the strategies for carrying out mental processes. That is, according to Broadbent, findings from neuroscience do not (and never will) matter when developing theories in psychology. More pragmatically, authors might agree that the brain codes information in a distributed manner but still argue that attempts to link cognition to brain hardware should wait till there is a better conceptual understanding of how the mind works—which is thought of as best achieved through the development of localist models. However, I expect most researchers today would find these responses unsatisfying. With the rapid development of cognitive neuroscience, it is becoming clear that cognitive theories can and should be constrained by biology.

Another (more common) response has been to ignore the neuroscience data taken to support distributed coding schemes, namely, the single-cell electrophysiological recording studies discussed in detail in Part 2. It is not that these theorists reject biological constraints to cognitive theories in general; indeed, localist models are often inspired by neuropsychological findings (e.g., Coltheart, 2004). Rather, I assume that the relevance of this data is not appreciated.

Of course there is one more possible response for an advocate of localist models. That is, the common claim that distributed representations better capture key aspects of brain functioning can be challenged. Although this latter position is rarely adopted, there are a few prominent researchers in neuroscience who either endorse the grandmother neuron hypothesis or who consider it a serious option on the basis of the neurophysiological data (e.g., Barlow, 1972; Parker & Newsome, 1998; Thorpe, 1989, 2002).

This is the position taken here. Indeed, the main goal in this article is to show that the current findings in neuroscience are compatible with localist models in psychology. It is not my claim that current data provide unambiguous support for localist coding schemes. But, I do intend to show that there is no reason to prefer distributed over localist representations on the basis of their relative biological plausibility. Indeed, I argue that the distributed representations learned in PDP models are often inconsistent with much of the relevant neuroscience data.

Although this article is largely directed to cognitive psychologists, I hope the article is of some relevance to neuroscientists as well. Despite the widespread rejection of grandmother cells in neuroscience literature, there is now a large body of evidence highlighting the extent to which single neurons respond selectively to inputs, including words, objects, and faces. So why then do neuroscientists favor distributed coding schemes? Part of the reason is that the term "distributed representation" is used somewhat differently in the two literatures. Indeed, the data taken to support distributed coding schemes in the brain rarely provide any support for the types of coding schemes employed in many PDP networks. More importantly, there seems to be confusion in the neuroscience literature with regard to how local coding schemes work, and what sorts of data support or refute this framework. It is striking how often data entirely consistent with localist coding in the brain is taken to falsify this approach. My secondary goal is therefore to clarify how localist models behave in order to challenge neuroscientists to reinterpret their findings in this light.

The article is organized in three sections. In Part 1, I set the stage by describing in some detail the localist coding schemes employed in cognitive models and highlight how this approach differs from distributed coding schemes. Although the hypothesis that a model (or brain) codes information in a localist manner might at first appear to be a straightforward claim that is easy to assess (and reject), there is in fact a great deal of confusion and disagreement about what a grandmother cell might be—as highlighted by the excellent article by Page (2000) and the associated commentaries. Certainly some versions of grandmother cells are untenable, but these are often caricatures of a serious hypothesis. Similarly, there is some confusion as to what constitutes a distributed coding scheme, and indeed, at least three different versions of the hypothesis have been proposed. I describe three different types of distributed coding, and contrast them with localist coding schemes, so that it is possible to evaluate the various hypotheses in light of the neuroscience data.

In Part 2, I review single-cell electrophysiological studies relevant to the question of grandmother versus distributed coding schemes in the brain. The review is organized around data collected from simple organisms (e.g., sea slugs and flies), simple responses in complex organisms (e.g., motion perception in macaque monkeys), and complex processes in complex organisms (e.g., face recognition in humans).

In Part 3, I evaluate local and distributed theories based on the data reviewed in Part 2. The two main conclusions I draw are that the current data (a) lend some support to the grandmother cell theory of mental representation and (b) strongly challenge the link that is often drawn between the representations learned by PDP models and the neural representations in brain. I also consider various objections that are often raised against grandmother coding schemes and show that these objections are wanting. I hope to convince you that a grandmother cell hypothesis is not only consistent with the data, but also plausible.

Part 1: Some Background, Definitions, and Confusions Associated With the Concepts of Local and Distributed Coding

Hubel and Wiesel (1962, 1968) provided some of the first insights into the neural representations that support early vision. They observed that single cells in the primary visual cortex (V1) were driven by simple but readily interpretable visual inputs (a line projected at a given location and orientation on the retina), and these neurons (so-called simple cells) were organized into topographic maps. That is, V1 is organized into columns, with simple cells in adjacent columns coding for similar but slightly different line orientations at the same retinal position (cells in adjacent columns code for line segments that vary by approximately 10° in rotation). These columns are in turn organized into hypercolumns, such that all simple cells within a given hypercolumn code for lines at the same retinal location (albeit varying in orientation preference), with adjacent hypercolumns coding for adjacent retinal locations. In this way, V1 codes for a range of line orientations in a range of retinal locations.

Most important for present purposes, Hubel and Wiesel (1968) also found that information is coded in a hierarchical fashion, with complex cells in V1 combining the inputs from multiple simple cells in order to code for more complex stimuli. Konorsky (1967), an early advocate of grandmother neurons (what he called "gnostic units") took these findings as strong support for this hypothesis. On his view, simple and complex cells constitute the first levels of a larger hierarchy that includes gnostic units on top. Hubel (1995) also considered the implications of this hierarchal organization within early vision and rejected the idea that the hierarchy could continue to the level of grandmother cells. According to Hubel, it is implausible to imagine that there could be one neuron corresponding to a grandmother smiling, another corresponding to a grandmother weeping, and yet another corresponding to a grandmother sewing. But this betrays a common confusion. It is indeed implausible to suggest that there is a separate cell for each mental state or action of grandma, but this has never been claimed by advocates of local coding in cognitive psychology or neuroscience. Hubel has only rejected a caricature of a grandmother cell. In fact, confusions regarding this hypothesis are widespread, as I detail below.

What Is a Localist Representation?

The defining feature of a localist representation is that it codes for one thing, and accordingly, it is possible to interpret the activation of a single unit (e.g., Bowers, 2002; Page, 2000; Thorpe, 1989). For example, if a simple cell is highly active, then you can infer with a high degree of confidence that something in the world is projecting a line of a given orientation on its receptive field (the identification of the object can only be determined at a later stage of processing, but it is sure to have the relevant feature). Similarly, if the word *cat* is coded by a localist unit, then it is possible to determine that *cat* was presented to the network by monitoring its activation—no other unit need be consulted. Hummel (2000) highlighted the fact that a localist representation involves a relation between a single unit and a meaningful equivalence class of entities in the world. The set of entities that activate the unit might all be instances of the word *cat* (regardless of font, size, position, etc.), an image of a specific person's face in whatever profile (a view-independent face cell), or a specific face presented in a restricted set of orientations (a view-specific face cell) and the like. In each case, the key feature of the cell is that it is possible to provide a meaningful description of the thing or the equivalence class to which it responds.

Although this definition seems straightforward enough, at least two points of potential confusion merit attention here. First, when considering the plausibility of localist (and distributed) coding schemes it is important to restrict the types of things under consideration. It is sometimes claimed that localist representations extend to all sorts of familiar things, from letters to propositions (e.g., Plaut, 2000). But this is not required by advocates of localist coding. Although there must be a local representation for the concept grandmother, there is no commitment to the claim that a single cell codes for the familiar proposition "Have a nice day."

Indeed, advocates of localist coding schemes are quite explicit in restricting the level to which knowledge is coded locally. For example, within traditional psychological (and linguistic) models of language, word knowledge is coded in a localist format up to (but not beyond) the morphological and lexical levels (but see MacKay, 1987, for an exception). A proposition (or complex thought) is only conceived when a collection of localist representations are entered into some syntactic relation to one another. That is what syntax is for—to allow "the infinite use of finite media" (von Humboldt, quoted by Pinker, 1998, p. 118). Accordingly, there is no combinatorial explosion in which infinity of possible thoughts requires infinity of neurons, one for each thought.

A similar point was made by Barlow (1972) with respect to vision. He argued that cardinal cells code for elements of perception, and the whole of perceptual experience would be coded by some combination of active cardinal cells (much like a sentence is composed of a collection of words). For present purposes, there is no distinction between Barlow's (1972) cardinal cells and grandmother cells as long as elements of perception extend to single objects or faces. To summarize then, advocates of grandmother (and cardinal) cells never claimed that a proposition (complex thought) or unique perceptual event is coded by a single neuron, and accordingly, the objection of Hubel (1995) is misguided.

A second possible point of confusion is that it might appear that localist theories are committed to the claim that one and only one unit is devoted to coding each thing. This assumption sometimes leads to a quick dismissal of localist coding schemes, as indeed it is implausible to suggest that the concept "grandmother" could be lost if the one corresponding neuron was damaged (e.g., Eichenbaum, 2001). Although a strict one-to-one correspondence between neuron and knowledge constitutes an example of grandmother theory, the hypothesis also admits the possibility that multiple neurons represent the same thing. Indeed, Konorsky (1967) suggested that the number of grandmother cells (gnostic units) devoted to a stimulus might be proportional to the importance of the stimulus to the individual. Similarly, Barlow (1985), Gross (2002), Page (2000), and Perrett et al. (1989), among others, are all clear that redundant coding would be required on any feasible grandmother coding scheme. Throughout the article, I employ the term g*randmother neuron* or *grandmother cell* for the sake of convenience, but this should not be taken as a commitment to the claim that one and only one neuron codes for a given item. As discussed in Part 3, there may be massive redundancy.

In sum, the key claim of localist coding schemes is that a given unit (neuron) codes for one familiar thing (and does not directly contribute to the representation of anything else), and that it is possible to interpret the output of a single unit in a neural network. Grandmother cell theories are committed to the claim that there are localist representations for words, objects, and faces, but there is no corresponding claim that localist codes extend to propositions or complex visual scenes. I expect (hope) most researchers would be willing to accept these points, but there are a number of more subtle issues that need to be considered in order to avoid some common misunderstandings.

Do Localist and Distributed Representations Always Co-Occur?

It is sometimes claimed that all models that include localist representations at one level of a network (e.g., Layer *n*) include distributed representations of the same items at lower levels as well (Layers $n - 1$, $n - 2$, etc.). For example, Page (2000) considered the interactive activation (IA) model of visual word recognition (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982), the archetypal localist model in psychology. The model includes three levels: An input layer composed of a set of visual-feature detectors that respond selectively to line segments in various orientations; a second layer composed of nodes that respond selectively to letters in specific positions within a word; and a third layer composed of nodes that respond maximally to individual familiar words. According to Page, words are coded locally in Layer 3 of the IA model and in a distributed format at Layers 1–2. For example, *time* is coded locally (by a single unit) in Layer 3 and as a pattern of activation over four letter units (*t-i-m-e*) at Layer 2. On this view, all models include distributed representations, and the relevant question for distinguishing localist versus distributed models is whether there are localist representations of the relevant entities at a given level of the network. The same point has been made by Hummel (2000), Thorpe (2002), Seidenberg and Plaut (2006), and others.

However, it is a conceptual mistake to assume that a model (or neural system) that codes for an item locally at Layer *n* must represent the same item in a distributed manner at Layer $n - 1$. Consider what this hypothesis entails. It must be assumed that there are redundant representations of words, objects, faces, and the like, with localist representations at the top of a hierarchy and with distributed representations of the same items at a lower stage. In the case of the IA model, the claim must be that there are lexical representations of words at Layer *n* and distributed representations of the same words at the letter level at Layer $n - 1$. This claim has a straightforward implication: The pattern of activation across a set of letters at Layer $n - 1$ should support the same (or at least similar) functions as the corresponding localist representations at Layer *n*. For example, the distributed representations at Layer $n - 1$ should support a "yes" response in a lexical decision task, support all the various forms of the word superiority effect, interact with phonology to support the naming of both regular and irregular words, and support access to meaning, and the like. Indeed, distributed representations of words in PDP models can support most of these functions.

However, none of these functions can be performed by a collection of activated letter detectors in the IA model. Consider Figure 1. According to Page (2000), the word *time* is coded locally at the word level but in a distributed format at the letter level. However, if the *time* unit is removed, the word *time* is now unfamiliar to the model, similar to countless other unfamiliar letter strings (e.g., the pseudowords *blap*, *slad*, etc.). When the lesioned model is presented with *time*, the coactivate letter units *t-i-m-e* cannot support the word superiority effect, in which words are better identified than pseudowords (a key experimental phenomenon that the IA model was designed to explain), nor can the letter units support lexical decisions, among many other lexical effects. In short, a collection of localist letters does not constitute a representation of a familiar word (distributed or otherwise). In the same way, if a Chinese character was presented to the IA model, a collection of local feature detectors at Layer $n - 2$ would become activated, but the IA model does not code for Chinese words in a distributed manner at Layer $n - 2$ (even assuming all

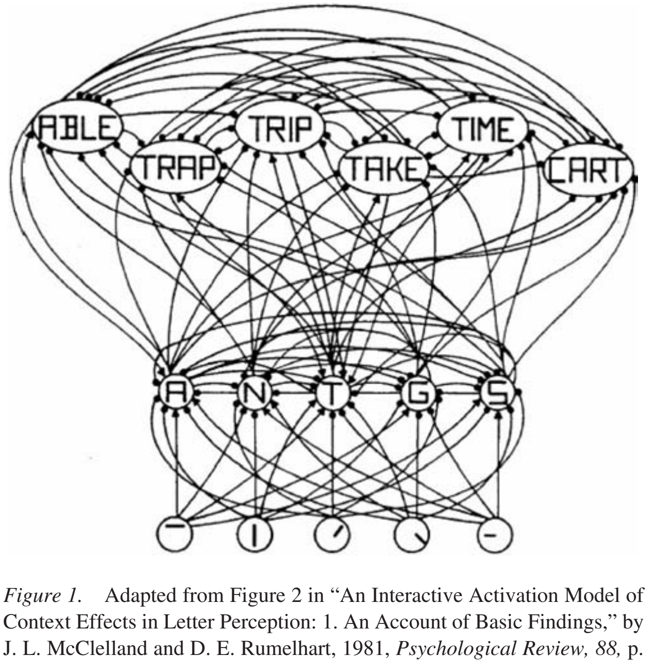

Figure 1. Adapted from Figure 2 in "An Interactive Activation Model of Context Effects in Letter Perception: 1. An Account of Basic Findings," by J. L. McClelland and D. E. Rumelhart, 1981, *Psychological Review, 88,* p. 380. Copyright 1981 by American Psychological Association. Schematic diagram of a small subcomponent of the interactive activation model. Bottom layer codes are for letter features, second layer codes are for letters, and top layer codes are for complete words, all in a localist manner. Arrows depict excitatory connections between units; circles depict inhibitory connections.

the relevant features were present at Layer $n - 2$). The reason why it is tempting to claim that the IA model codes for words in a distributed manner at the letter level (but not Chinese words at Layer $n - 2$) is that we know that *time* is a word—the model does not.

This raises a related issue. If there are no distributed representations of words in the IA model, how should one characterize the representations of unfamiliar words? Again it might be tempting to argue that the IA model includes distributed representations for this purpose. For example, the pseudoword *pime* will activate a collection of localist word representations that share a number of letters (e.g., *time, lime, pine, dime*, etc), and the pattern of activation over these words might constitute its distributed representation. Indeed, pseudowords like *pime* are better identified than are random letter strings (e.g., *xyfk*) in the IA model because nonwords activate more word representations (the so-called pseudoword superiority effect; McClelland & Rumelhart, 1981). Similarly, some models of word processing support pseudoword naming through lexical analogy, in which coactive local word representations specify the pronunciation of a pseudoword (Humphreys & Evett, 1985; Kello, Sibley, & Plaut, 2005). For example, the coactivation of *time, lime, pine,* and *dime* might code for the pronunciation of *pime*. Similar accounts have been applied to faces, with unfamiliar faces identified through the coactivation of familiar ones (e.g., Jiang et al., 2006). In addition, *pime* activates the letters *p-i-m-e*, and this pattern of activation over the letter units might also constitute its distributed representation.

However, the suggestion that coactive letter or word units constitute a distributed representation of a pseudoword is not correct either. The core claim regarding distributed representations is that individual units are involved in coding multiple familiar things. The localist letter and word units in the IA model are at odds with this assumption, and thus, to call a pattern of activation over these local units distributed is misleading. In any case, the grandmother cell theory is concerned with how familiar knowledge is coded. The hypothesis that unfamiliar knowledge is coded through the coactivation of multiple localist representations does not bear on the issue.

The (mis)use of the term *distributed* occurs in a related context. As discussed above, localist coding schemes assume a hierarchy of progressively more complex representations, at the top of which are single units (neurons) that code for complex stimuli, including individual persons, objects, and words. An obvious question then arises; namely, how should one characterize the representations that underlie more complex thoughts and perceptions? For instance, in order to code for the proposition *the dog chased the cat*, the local representations of *dog*, *chased,* and *cat* need to be coactivated, and further, they need to be activated in such a way as to distinguish this from the related proposition *the cat chased the dog*. One possible solution was proposed by Hummel and Holyoak (1997), who developed a neural network model that includes localist representations for nouns (e.g., unique nodes for *dog*, *cat*, etc.) and predicates (e.g., a node that codes for *x chased y*—with *x* and *y* unspecified) and a mechanisms to bind (transiently) nouns to predicates, such that $x = dog$, $y = cat$ when representing the proposition *the dog chased the cat*, and $x = cat$, $y = dog$ when coding *the cat chased the dog*.

For present purpose, the key question is whether their model includes distributed representations, given that *the dog chased the cat* is coded as a pattern of activation over a collection of localist units, and furthermore, each unit is involved in representing many different propositions (e.g., the same *dog* unit is involved in coding the related proposition *my dog likes ice cream*). Hummel and Holyoak (1997) adopted this term and titled the article that introduced their model, "Distributed Representations of Structure: A Theory of Analogical Access and Mapping."

But again, the use of the term *distributed* in this context risks confusion, given that Hummel and Holyoak (1997) meant something quite different from the distributed representations proposed by PDP modelers. Indeed, their model is fundamentally inconsistent with the PDP approach. That is, their model not only includes local coding at the lexical level but also implements symbolic processing: Computations are performed over context independent representations in order to ensure that complex thoughts are compositional and systematic. By contrast, a key claim of the PDP approach is that cognition (and the brain) computes without recourse to context independent representations and, more generally, that cognition does not rely on symbolic processing (for detailed discussion of the contrast between symbolic and PDP approaches, see Bowers, 2002; Fodor & Pylyshyn, 1988; Marcus, Vijayan, Rao, & Vishton, 1999; Prasada & Pinker, 1993). It can only confuse matters to use the term *distributed* to describe the qualitatively different representations in these alternative (indeed, theoretically opposite) approaches to understanding how the brain implements perception and thought.

To summarize, the grandmother cell hypothesis is concerned with how familiar knowledge is coded, and there is no logical reason that localist and distributed representations must co-occur in a hierarchy of processing steps. A localist model must account for how novel information is coded, but this can be accomplished through the coactivation of multiple localist codes. Of course, the brain may implement localist and distributed coding schemes at different stages of processing, but it is also possible to envisage a strong version of a grandmother theory in which our brain is localist throughout, from photoreceptors to grandmother cells (for more discussion of the possibility that local and distributed coding co-occur, see discussion below). Whatever the connection between local and distributed coding, a grandmother theory must predict that the cells at the top of a visual hierarchy code for words, objects, and faces in a localist manner.

How Do Localist Models Behave?

There is also a common misunderstanding in the neuroscience literature about how localist models behave. In particular, it is often assumed that localist representations are activated by a specific object or face and are entirely unaffected by similar inputs (Földiák, 2002; Gross, 2002; Poggio & Bizzi, 2004). This conceptualization of grandmother cells appears widespread and leads to some unwarranted conclusions. For example, Young and Yamane (1992) presented macaque monkeys images of human faces while recording from 850 temporal lobe neurons. Figure 2 shows the set of 27 test faces, and Figure 3 shows the firing pattern of the most selective cell in response to these faces. The letters and numbers in the two figures correspond to the faces depicted in Figure 2. As can be seen in Figure 3, this neuron fires robustly to 1 face (face E), minimally to 1 other (face R), and not at all to the remaining faces. Despite the striking selectivity of this cell, Young and Yamane

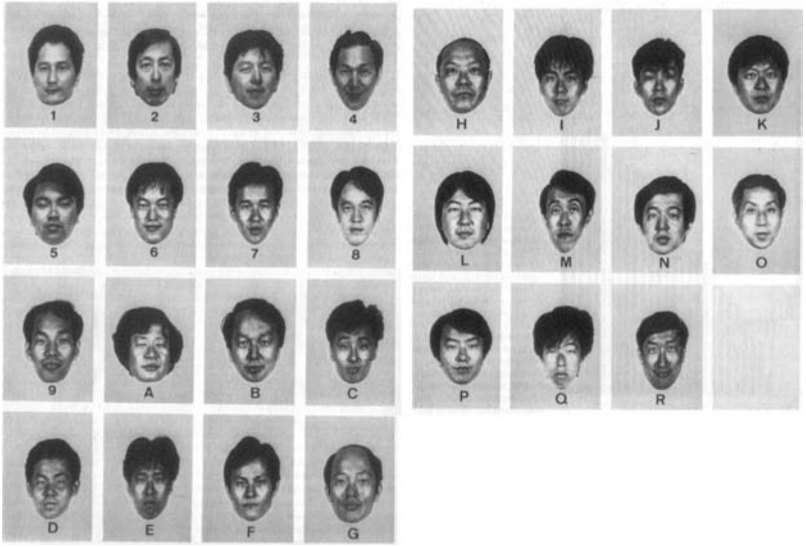

Figure 2. Adapted from Figure 4-2 in "An Analysis at the Population Level of the Processing of Faces in the Inferotemporal Cortex," by M. P. Young and S. Yemane, from *Brain Mechanisms of Perception and Memory: From Neuron to Behavior* (p. 50–51), edited by T. Ono, L. R. Squire, M. E. Ratchle, D. I. Perrett, and M. Fukuda, 1993. Copyright 1993 by Oxford University Press. Reprinted by permission of Oxford University Press. The set of 27 face images that were presented to the macaque monkeys while recording from 850 temporal lobe neurons. Each face was assigned a unique letter or number, as indicated in the figure.

(1992) focused on the fact that most neurons did respond to more than 1 face and concluded that this provides evidence for a distributed code. In order to reconcile the selectivity of this one neuron to the above claim, the logic must be that a grandmother cell should only respond to a single stimulus, and the partial (minimal) activation of this neuron to a second face (and, presumably, other untested faces) falsifies the hypothesis. The problem with this analysis, however, is that the premise is false; in standard localist models, multiple face or word representations are coactivated during the encoding of a stimulus.

To illustrate this, consider Figure 4. The figure shows the relative amount of activation of localist word nodes *blue* and *blur* in the IA model, in response to the familiar word *blur*. The key point to note is that although the word unit *blur* receives the most activation in response to the input *blur*, visually similar words (in this case *blue*) receive activation as well. As can also be inferred from this figure, the localist word unit *blue* is activated by both *blur* and *blue*. So a single input (in this case a word) activates a pattern of activation across a (limited) collection of localist word representations, and a given localist word unit is activated by a (limited) range of inputs (cf., Jacobs, Rey, Ziegler, & Grainger, 1998). It is important to note that activation based on similarity is not some idiosyncratic property of the IA model. It applies to all localist models and, likely, to any networks in the brain.

Two points should be emphasized here. First, if a neuron responds to more than one stimulus, as in the Young and Yamane (1992) study, this does not show that the brain coded for the stimulus in a distributed way. Similarly, if a single stimulus produces a pattern of activation across a set of neurons, this does not, by itself, show that the brain coded the item in a distributed format. The critical question is not whether a given neuron responds to more than one object, person, or word but rather whether the neuron codes for more than one thing. Localist coding is implemented if a stimulus is encoded by a single node (neuron) that passes some threshold of activity, with the activation of other nodes (neurons) contributing nothing to the interpretation of the stimulus. For example, the coactive *blue* unit does not play a role in representing *blur* in Figure 4. Young and Yamane (1992) did not provide any evidence that the second most active neuron contributed to the coding of the face, and accordingly, their conclusion is unfounded. Barlow (1995) made a similar point when he introduced the term *the lower-envelope principle* to describe coding in sensory systems. On this view, sensory thresholds are determined by the neurons that have the lowest threshold for a given stimulus and are not influenced by the responses of less sensitive neurons.

The second point is that the coactivation of knowledge in localist models complicates the task of distinguishing between localist and distributed coding schemes in the brain. But the conceptual distinction is unaffected by these empirical challenges, and the distinction is fundamental to understanding how the brain works.

What Is a Distributed Code?

The fundamental point about all distributed coding schemes is that each unit in a network is involved in coding more than one familiar thing, and as a consequence, the identity of a stimulus cannot be determined by considering the activation of a single unit (neuron). More precisely, a single unit within a distributed system codes for a relation between itself and a set of entities in the world, but unlike localist representations, the entities do not constitute a meaningful equivalence class (Hummel, 2000). For instance, a given unit might contribute to the coding of the words *blue* and *blur*.

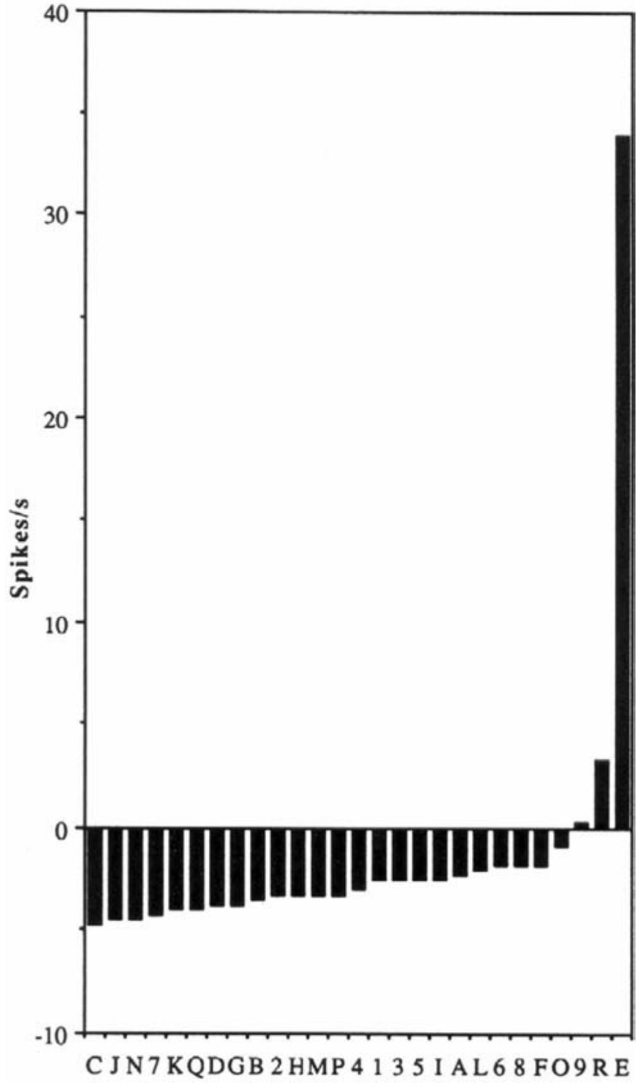

Figure 3. Adapted from Figure 4-4 in "An Analysis at the Population Level of the Processing of Faces in the Inferotemporal Cortex," by M. P. Young and S. Yemane, from *Brain Mechanisms of Perception and Memory: From Neuron to Behavior* (p. 50–51), edited by T. Ono, L. R. Squire, M. E. Ratchle, D. I. Perrett, and M. Fukuda, 1993. Copyright 1993 by Oxford University Press. Reprinted with permission of Oxford University Press. The activation of one inferior temporal cell in a macaque monkey in response to the 27 faces from Figure 2. The cell was inactive (or below baseline) for all but face E and, to a small extent, face R.

When distinguishing between distributed and localist coding schemes, it is important to also distinguish between the representations and the processes within a model. In the localist IA model, the input *blur* will coactivate the localist *blur* and *blue* word representations (as discussed above). Furthermore, coactive representations may impact the identification of the target. For instance, *blur* and *blue* might compete for identification through lateral inhibition, such that the identification of *blur* is delayed by virtue of *blue* being active (cf. Bowers, Davis, & Hanley, 2005). Nevertheless, the key point remains that each unit in a localist system codes for one thing, and each unit in a distributed system is involved in coding multiple things. The fact that the *blue* unit is partially active in response to the input *blur* in a localist system does not compromise the conceptual distinction between local and distributed coding because *blue* does not contribute to the representation of *blur*. However, as noted above, it does make the task of distinguishing between these theories more difficult.

Another challenge in distinguishing between local and distributed coding schemes is that distributed representations come in at least three different forms: what I call dense, coarse, and sparse distributed coding. The coding schemes differ (in part) with regards to how much information can be derived from the activation of a single unit, and they make different predictions concerning the outcomes of single-cell recording studies. Accordingly, when assessing the relative biological plausibility of local versus distributed coding schemes, the contrast is not between local and distributed but between local and dense distributed, local and coarse coding, and local and sparse coding. These three coding schemes are described next.

Dense Distributed Representations

Dense distributed representations contrast most sharply with localist ones. In dense distributed coding schemes, each unit or neuron is involved in coding many different things, and as a consequence, little information can be inferred from the activation of a single neuron. This hypothesis is commonly associated with PDP theories. For example, when discussing this approach, Thorpe (1995) wrote, "with distributed coding individual units cannot be interpreted without knowing the state of other units in the network" (p. 550). Elman (1995) endorsed this view when he wrote, "These representations are distributed, which typically has the consequence that interpretable information cannot be obtained by examining activity of single hidden units" (p. 210). More generally, Smolensky (1988) introduced the term *subsymbolic* to describe the function of individual units in a PDP network and the term *subconceptual* to describe their meaning.

Because of the widespread assumption that individual hidden units in PDP models convey little information, little effort has been devoted to studying hidden units one at a time. Nevertheless there is some evidence consistent with this claim, at least under some conditions. For example, consider the classic Seidenberg and

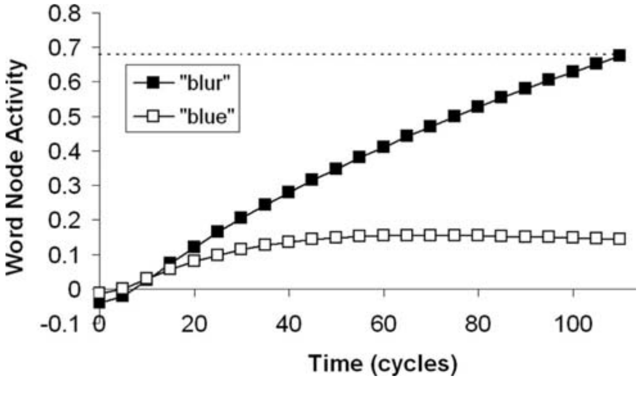

Figure 4. The activation of *blur* and *blue* units within the interactive activation model in response to *blur*. Although only *blur* is activated beyond threshold, *blue* is activated to some extent as well.

McClelland (1989) PDP model of word naming (which was developed as an alternative to the IA model). The model includes 200 hidden units that map between orthographic and phonological representations. After training, the model correctly pronounced 97.3% of the 2,897 words it was trained on, as well as many novel items. In an attempt to gain some understanding of how the hidden units contributed to performance, Seidenberg and McClelland (1989) recorded the activation values of the hidden units in response to the words *pint*, *mint*, and *said*. Because *pint* is an irregular word, in that its pronunciation is not predictable from its spelling, a lexical (localist) representation is required to support its pronunciation on some theories (e.g., Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001). In response to *pint*, 22 hidden units were highly active (>.8 of maximum activation), constituting ~10% of all hidden units. The orthographically related word *mint* activated 11 of the same units, which highlights the fact that similar words are coded by similar patterns of activation (50% overlap in this case). It is interesting to note that the dissimilar word *said* activated 4 overlapping units (18% overlap).

The critical point to take from this is that it is not possible to infer much of anything on the basis of the activation of a given unit. For the sake of argument, imagine that unit 100 is highly active in response to *pint*. If one assumes that the 18% overlap between *pint* and *said* is typical for 2 dissimilar words, then there is also an 18% chance that unit 100 will be activated by another dissimilar word (the 18% chance overlap for unit 100 applies to all other units as well, producing 18% overlap overall). If one further assumes that there are approximately 2,000 words in the trained vocabulary set that were dissimilar to *pint* (there were 2,897 words in the training set in all, but some will share some orthographic overlap), then 18% of these words should also activate unit 100. That is, on the basis of the activation of this single unit it is only possible to infer that 1 of approximately 360 dissimilar words (.18 × 2,000 = ~360) was presented to the network. The same would hold if the overlap between dissimilar items was much smaller. For instance, if dissimilar words produced orthogonal activation patterns, and each pattern activated 10% of the hidden units, then the probability of the same hidden unit being active for 2 dissimilar words would be .01 (rather than .18). Still, if the network codes for 2,000 dissimilar words, each unit will be activated by ~20 dissimilar words (.01 × 2,000). So again, little meaning can be attached to the activation of a given unit. The ambiguity is only enhanced by the fact that each unit is activated by orthographically similar words as well.

A more direct demonstration that PDP models often learn dense distributed representations has been provided by Berkeley, Dawson, Medler, Schopflocher, and Hornsby (1995). They developed an analytical technique in which the activity of each hidden unit was recorded in response to a range of inputs by means of a separate scatter plot for each unit. The unit's response to an input is coded along the *x*-axis, with values on the *y*-axis kept random (in order to prevent points from overlapping). These so-called jittered density plots effectively provide single-cell (or in this case, single-unit) recordings for each hidden unit in response to a large number of inputs.

Berkeley et al. (1995) carried out this analysis on a series of three-layed networks trained by back-propagation to categorize logical inference problems. One of the models included 14 input units (a pattern of activation across these units defined the input problem), 3 output units (a pattern of activation across these units categorized the input problem into one of four different argument types and indicated whether or not the argument was valid), and 15 hidden units. The key point for present purposes is that after training, the model was able to correctly categorize 576 input patterns into six categories. Furthermore, the jittered density plots confirmed the model categorized the inputs based on dense distributed representations. That is, each unit responded to most of the 576 inputs (to varying degrees), so that there is little information about the identity of the input recoverable from the activation of a single hidden unit. Indeed, the plots looked quite similar to those displayed in Figure 5B, as described below.

To further explore the nature of the distributed representations learned in PDP models, Bowers, Damian, and Davis (2008) recently carried out a set of these analyses on a PDP model of short-term memory (STM) and word naming. The simulations were based on Botvinick and Plaut's (2006) model of STM that includes a localist input and output unit for each letter and 200 hidden units. After training, the model could reproduce a series of 6 random letters at ~50% accuracy, which roughly matches human performance in an immediate serial recall task. In one analysis, we successfully trained their model to this criterion and then computed jittered density plots for all the hidden units in response to all 26 letters (the scatter plots were constructed in response to single letters rather than lists of letters). The plots of the first 24 (out of 200) hidden units are presented in Figure 5A. As is clear from these plots (and equally true of the remaining plots), it is not possible to interpret the output of any given hidden unit, as each unit responds to many different letters. For the model to correctly retrieve a single letter (let alone a list of 6 letters in STM), the model must rely on a pattern of activation across a set of units. That is, the model has learned to support STM on the basis of dense distributed representations.

In another analysis, I trained the same network to name a set of 275 monosyllable words presented one at a time. That is, rather than supporting STM, the model learned to name single words. Each input and output was a pattern of activation over three letter units, and the model was trained to reproduce the input pattern at the output layer. After training, it succeeded (~100%) on both trained words and 275 unfamiliar words. Figure 5B presents the jittered density plots of the first 24 hidden units in response to the 275 familiar words. Once again, the model succeeded on the basis of dense distributed representations. These analyses support Seidenberg and McClelland's (1989) earlier conclusion.

In sum, each hidden unit in the above PDP models contributes to the coding of many different things and, as a consequence, each unit conveys little information. That is, these models learn dense distributed representations. I would suggest that many researchers consider this as a core theoretical claim of the PDP approach. Consistent with this view, some PDP models do indeed learn dense distributed representations.

Coarse Coding

In the neurosciences, the term *distributed* (sometimes called *population*, or *ensemble* coding) tends to mean something different. The claim is not that individual neurons code for many different things (and that little meaning can be attached to the activation of a given unit) but rather that neurons have broad

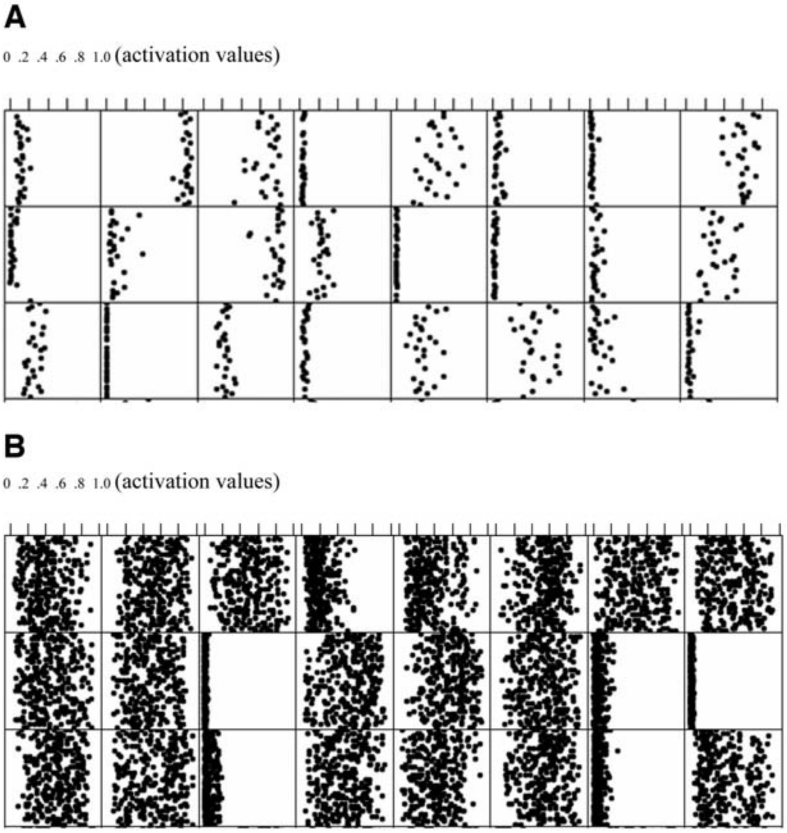

Figure 5. A: Presents 24 jitter density plots depicting the activation values of first 24 hidden units (from a total of 200 units) from Botvinick and Plaut's (2006) model of short-term memory. Each point indicates the activation value for a given hidden unit (ranging from 0 to 1.0) in response to a single letter after the model was trained to code for lists of letters. As is clear from the plots, little meaning can give given to the output of a single hidden unit. The same is true for the remainder of the hidden units. B: Depicts the activation values of the first 24 hidden units in response to 275 trained words. Here the model was trained on words, one at a time, and was effectively a model of word naming. Again, little meaning can be attached to the activation of any given hidden unit.

tuning curves, such that a given neuron plays a role in coding for a range of similar things. Although these neurons may respond most strongly to a preferred stimulus (on average), the noise in the system makes it impossible to reliably identify an object on the basis of the response of a single cell. Accordingly, the core claim is that the unique identification of a given object, word, and the like relies on pooling across a collection of noisy units in order to get a stable and exact measure of an input. Pooling is generally thought to occur across multiple levels of the visual system, with simple cells pooling to complex cells, complex cells pooling to hypercomplex cells, and the like, until pooling converges on cells at the top of a hierarchy that code for complete objects or faces in a coarse manner (e.g., Riesenhuber & Poggio, 1999).

As an illustration, consider Figure 6, which depicts a coarse coding strategy for encoding spatial location. Each cell has a relatively large receptive field (as indicated by the size of each circle) and will fire in response to an object anywhere within its receptive field. Although these cells will tend to fire most strongly to stimuli located in the center of their receptive field, the noise in the system makes it impossible to read off the exact position of an object on the basis of the exact level of activation of any one neuron. Nevertheless, a relatively precise specification of location can be coded by coactivating a set of these noisy neurons. For instance, if Neurons 1–4 in Figure 6 are all coactivated, then the object must be located at the intersection of their receptive fields, and this specifies a small region of space in a reliable way (cf., Hinton, McClelland, & Rumelhart, 1986).

A common feature of coarse coding schemes is that they are implemented in topographic maps; that is, neurons that code for similar things are located in similar areas of the cortex. In the present example, nearby neurons code for similar (overlapping) spatial locations (a so-called retinotopic map). But the same logic applies across dimensions. For instance, similar tones in auditory cortex are coded by spatially proximate neurons (e.g., Formisano et al., 2003; a so-called tonotopic map). In this way, a given unit responds to a limited set of similar inputs, and a given input activates a limited set of units that extend over a small area of cortex.

To give an example from neuroscience, Georgopoulos, Schwartz, and Kettner (1986) studied how neurons in the primary motor cortex of rhesus monkeys encode arm movements. As in the example above, there is a spatial organization to motor neurons,

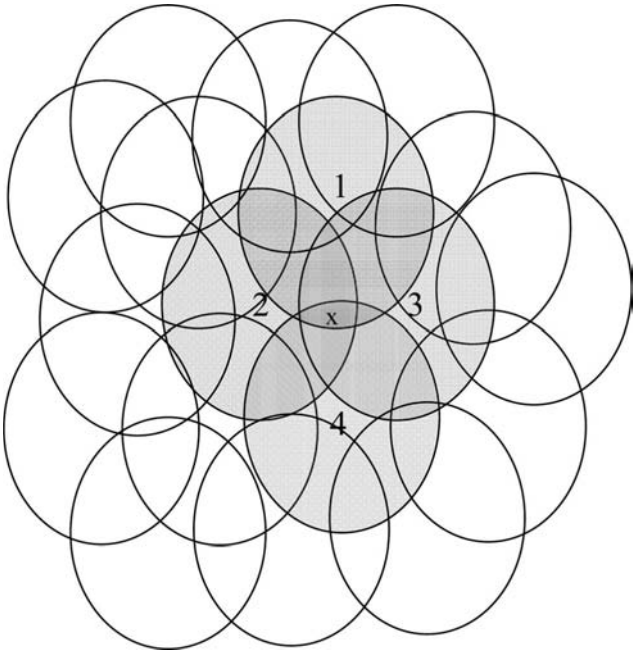

Figure 6. A coarse coding scheme for coding the position of the object X. Each neuron fires in response to an object its receptive field. Given the large size of the receptive fields (as indicated by each circle), the precise location of an object cannot be determined by the activation of a single neuron. However, a more precise estimate of location can be computed through the coactivation of Neurons 1–4, which indicates that the object is located in the small area that intersects all three receptive fields (as marked by dark shading).

with nearby neurons coding for similar directions of motion. The researchers observed that single cortical neurons are broadly tuned to direction of movement, such that they fire robustly to arm movements in one direction but respond (to a lesser extent) to related directions as well. Because of noise in the system, the firing of a single neuron does not encode sufficient information to reliably support precise behavior. Georgopoulos et al. concluded that direction is coded by a pattern of activation across a set of similarly tuned neurons located in close spatial proximity. In their population vector model, each neuron "votes" for their preferred direction, weighted by its firing rate. This vectorial summation of all the votes provided a better measure of the direction of arm movement than did the response of any single neuron. Note, the vector averaging account of Georgopoulos et al. is just one of various proposed algorithms designed to integrate (pool) a pattern of neural activity into a meaningful code for motor control (or perception; for review see Pouget, Dayan, & Zemel, 2000).

With regards to the debate about localist versus distributed coding, the critical issue is how the pooling algorithm is implemented in the brain. Zhang, Ginzburg, McNaughton, and Sejnowski (1998) made the important point that a range of pooling algorithms, including Bayesian methods, can be implemented in winner-take-all networks, that is, networks that implement grandmother coding schemes. Thus, the (often implicit) claim of advocates of distributed coarse coding is that the neural mechanisms that interpret a pattern of activation across a set of neurons do so without pooling to winner-take-all neurons at the top of a hierarchy (otherwise it is a grandmother coding scheme). Similarly, it must be claimed that the brain does not simply code for information on the basis of the most active neuron across a set of coactive and coarsely tuned neurons (otherwise it is operating according to the lower-envelope principle, another version of grandmother coding). Rather, the critical claim must be that a pattern of activation is the final step in the hierarchy of processing steps.

Sparse Distributed Coding

Yet another type of distributed coding that should be distinguished from above is sparse distributed coding as conceptualized in the PDP approach. On this coding scheme, a complex stimulus is coded by the activation of a small number of units, and each individual unit contributes to representation of just a few stimuli.

The obvious contrast between dense and sparse coding schemes is the number of units (neurons) involved in coding a given stimulus, with far fewer in the latter case. But it is also important to distinguish their functions. According to McClelland, McNaughton, and O'Reilly (1995), sparse representations are well suited to support rapid learning without erasing previously stored knowledge, but they are poor at generalization. By contrast, dense distributed representations are better at generalization but suffer from catastrophic interference, in which rapid new learning erases previous knowledge. On the basis of this analysis, it is argued that sparse coding is employed in the hippocampus in order to store new episodic memories following single learning trials, whereas dense distributed representations are learned slowly and reside in cortex in order to support word, object, and face identification (among other functions), all of which require generalization (e.g., to identify an object from a novel orientation).

The contrast between coarse and sparse coding is not as clear with regard to the number of units involved in coding a given stimulus. Presumably, coarse coding requires a greater number of units to be activated in order to represent something, but in practice, it may be difficult to distinguish between the two hypotheses on the basis of this criterion. Nevertheless, the two approaches can be distinguished. For example, unlike sparse representations, coarse codes are thought to support generalization; indeed, this is often described as one of the strengths of this coding scheme (e.g., Poggio and Bizzi, 2004). In addition, coarse coding is assumed to be employed in the cortex in the service of generalization, whereas sparse PDP coding is assumed to be restricted, in large part, to the hippocampus (and, perhaps, other brain areas that support fast learning but that are not involved in generalization). Furthermore, the two coding schemes tend to be embedded in networks with different structures. That is, coarse codes are typically embedded within hierarchical networks, with neurons at each level pooling to progressively more complex perceptual representations. Sparse codes within PDP networks are not generally organized along these lines.

Summary of the Key Features of Local and Distributed Coding Schemes

A grandmother cell is a localist representation at the top of a hierarchy of neural processing steps that codes for individual words, objects, faces, and the like. Although grandmother neurons sit on the top of a hierarchy, localist-coding schemes may operate across all the earlier levels as well. It is not claimed that familiar propositions (e.g., "Have a nice day"), novel propositions (e.g., "My grandmother is a nice Republican"), or complex visual scenes (a weeping grandmother) are coded by a single unit or neuron. Similarly, it is not claimed that novel words or novel objects are coded by a single unit or neuron. Rather, it is claimed that single neurons code familiar words, objects, faces, and the like. Given the confusions that have surrounded the concept of localist coding (cf. Page, 2000, and the corresponding commentaries), it is perhaps worth emphasizing that this description of localist coding is entirely consistent with an archetypal localist model of cognition, namely, the IA model of word identification.

Localist representations can be distinguished from three types of distributed coding schemes described in the literature. In the case of dense distributed representations, there is little meaning that can be assigned to a given unit, as the unit will respond to many different inputs. In the case of coarse coding, each unit responds to multiple similar inputs and is involved in coding multiple similar things. Because of this similarity constraint, it is possible to assign meaning to the firing of a single neuron, but only to a rough approximation. The outputs of multiple neurons need to be considered to get an exact measure of an input. Both of the above codes are assumed to support word, object, and face processing in the cortex. Finally, in the case of sparse distributed codes (as conceptualized by the PDP approach), each neuron contributes to the identification of a few things (more than one), and a given stimulus is defined by the activation of a few neurons. These representations support episodic memory in the hippocampus, but not object, word, and face identification in cortex.

Before reviewing the relevant data, it is perhaps worth emphasizing that the local–distributed distinction is just one of many issues concerning how the brain codes for information. In addition, there are questions about whether neural synchrony plays a role in coding information (e.g., Singer & Gray, 1995), whether information is communicated by the rate of firing or, alternatively, by the timing of the first neural spike (e.g., Thorpe, Delorme, & Van Rullen, 2001), whether information is coded by pooling independent signals, as in an election, or through the coordination (interaction) of signals, as in a symphony (e.g., deCharms, 1998), and the like. These issues are to some extent orthogonal to the question at hand; for instance, synchronous firing may well be implemented in brains that code for information in a localist or distributed manner. But the answer to one question presumably provides constraints to the others, and a full understanding of how the brain computes requires an answer to all the questions, among many others.

Part 2: Review of the Data

In this section, I review key findings from neurophysiology, a subfield of neuroscience that involves recording (or eliciting) the activity of a small number of neurons while an animal performs some task or is exposed to some stimulus. According to the grandmother neuron hypothesis, there should be strong associations between the behavior of single neurons and the ability of an animal to identify a stimulus or perform a task. For example, it should be possible to determine the identity of a face on the basis of the activation of a single neuron.

Distributed theories make different predictions. According to dense distributed coding schemes, there should be little information associated with the activation of single neurons. For example, all neurons involved in coding faces should respond to a wide variety of faces, including some dissimilar ones. That is, just as it is difficult to attach much meaning to the activity of any single hidden unit in a PDP network, it should be difficult to attach meaning to the activity of single neuron in the brain.

According to the coarse coding hypothesis, a given neuron should respond to a restricted set of similar stimuli. For instance, it should be possible to identify single neurons that respond to a set of similar faces and that fail to respond to dissimilar faces. The challenge in distinguishing coarse coding from local coding is that local units also respond to a range of similar things. The difference is that each unit in a coarse coding scheme is involved in representing multiple (related) things, whereas on local coding, each unit represents one thing and is only incidentally activated by similar things (e.g., the *dog* unit in the IA model represents *dog*, and its activation in response to *hog* is incidental). So, to provide direct evidence for coarse coding and to reject grandmother cell coding, it needs to be shown that a collection of broadly tuned neurons (with different response profiles) all play a role in coding a given input.

Finally, on sparse distributed coding, the activation of single neurons should be tightly coupled with a given stimulus or behavior, and accordingly, it might appear difficult to distinguish this from grandmother coding schemes. But, as noted above, the PDP approach makes the prediction that sparse distributed coding is employed in the hippocampus, whereas dense distributed representations are in the cortex. Accordingly, the response profiles of neurons in the two brain areas can provide a test of this claim.

The review starts with an analysis of how a variety of low-level perceptual and motor knowledge is coded in the brains of a variety of simple organisms and concludes with an analysis of how high-level knowledge is coded in complex organisms, including the representations of words, objects, and faces in humans. As noted above, models that include localist units (grandmother cells) at the top of a hierarchy may include localist representations at all levels.

Grandmother Cells in Simple Organisms

Motion perception in flies. Single-cell recordings from the movement-sensitive H1 neuron in the visual system of the blowfly (*Calliphora erythrocephela*) highlight the amount of information encoded in a single neuron. The H1 neuron encodes horizontal movements over the entire visual field, and van Steveninck and Bialek (1995) argued that behavioral decisions are based on just a few spikes from this one neuron. This neuron appears to be an extreme grandmother cell in the sense that there is one and only one H1 neuron, with no others performing a similar function (that is, there may be no redundancy). The reason, according to van Steveninck and Bialek (1995), is that this one neuron performs its task optimally, and there is no functional role for more neurons.

Smell in locusts. In the insect olfactory system, odor is processed in the antennal lobe, the analogue of the vertebrate olfactory bulb. Perez-Orive et al. (2002) observed that the firing of the Kenyon cells within the olfactory system of locust was highly selective to odors, with some neurons responding specifically to 1 out of 16 odorants by firing only one or two spikes. They nevertheless reject a grandmother cell coding scheme because they assume that the system would be too sensitive to damage, a claim I return to later. Nevertheless, the data are consistent with grandmother coding, and indeed, Heisenberg (2003) reviewed the Perez-Orive et al. (2002) study and other studies, reached the opposite conclusion, and suggested that a single synapse might represent the memory trace of an odor (for a related finding, see Keller, Zhuang, Chi, Vosshall, & Matsunami, 2007).

Looming perception in locust. The LGMD neuron is located in the third visual neuropile of the locust optic lobe. The neuron responds to objects approaching on a collision course with the animal (looming), and it is thought to be involved in the generation of escape behaviors. Gabbiani, Krapp, Koch, and Laurent (2002) reported evidence that this single neuron integrates information about angular velocity and object size in order to compute looming.

Feeding in sea slugs. Elliott and Susswein (2002) described how single cells support specific feeding behaviors in gastropods, including *Aplysia* (a sea slug). For example, they point out that the properties of single cells are generally consistent with them being designed for a specific function that is easily described in a few words. They go on to highlight the relevance of single-cell recording studies carried out on simple organisms to grandmother theories in general. That is, if the distributed coding schemes advanced by PDP theorists are correct, this would suggest that there are fundamentally different mechanisms of information processing in vertebrates and invertebrates.

Sensory Thresholds in More Complex Organisms

A detailed review and discussion of the link between sensory thresholds and neural signaling was reported by Parker and Newsome (1998). I have summarized some of the key findings reported in this article, as well as some more recent work.

Somatosensory perception in humans. Mountcastle and colleagues (Mountcastle, Carli, & Lamotte, 1972; Talbot, Darian-Smith, Kornhuber, & Mountcastle, 1976) compared monkey and human ability to detect a vibrating stimulus applied to the skin with the neural signaling of single mechanoreceptive neurons. In both cases, the most sensitive neurons account for the psychophysical performance of the organisms across a range of stimuli. It was these early findings that led to the hypothesis of the lower-envelope principle, according to which psychophysical performance is set by the most sensitive individual neurons, with other active neurons irrelevant.

In subsequent work, Mountcastle and colleagues (Lamotte & Mountcastle, 1975; Mountcastle, Steinmetz, & Romo, 1990) assessed the neural coding at the level of primary somatosensory cortex. They compared psychophysical and neural responses in a discrimination task in which observers compared the frequency of two vibrating stimuli. Once again, the behavior of the most sensitive neurons matched the performance of the organism.

In the above studies, the behavioral and neural recording were carried out in different testing conditions (e.g., recordings are often carried out in anesthetized animals), and accordingly, it is always possible that the similar thresholds were a by-product of this confound. Although this seems unlikely, it is interesting to note that a number of studies have compared behavioral and neural measures simultaneously, which not only eliminates any possible confounds between testing conditions but also allows a trial-by-trial comparison of neural responses of single neurons and psychophysical judgments.

In perhaps the first study of this sort, Vallbo and Hagbarth (1968) compared neural and behavioral sensitivity to indentations of glabrous (smooth) skin of the hand of human participants. Microelectrodes were inserted into the median nerve of a volunteer participant, and the neural response of single, peripheral neural fibers were measured in response to skin indentations that varied in their intensity. These peripheral fibers have extremely low spontaneous firing rates, and a neural response was defined as a single action potential in response to a tactile stimulus on a given trial. The probability of an action potential was 0% in response to an indentation of 5 μm or less, was 100% to indentations of 20 μm or more, and fired approximately 50% in response to indentations of 13 μm (the neural threshold). Vallbo and Hagbarth then assessed the psychophysical function for stimulus detection using the same stimuli and the same skin surface. The participants' task was simply to report whether a stimulus had been presented within a brief interval. The psychophysical function overlapped very closely with the response of the single peripheral neuron.

More impressive is the trial-by-trial correspondence between neural signal and behavioral report. Vallbo and Hagbarth (1968) repeatedly presented a near threshold 10 μm stimulus to the participant. In 30 trials, the participant detected the stimulus 16 times, and the neuron fired 17 times (again showing the similar sensitivity of the neuron and the organism). But the remarkable finding is that the neural response (a single action potential) almost perfectly predicted the behavioral response. This finding highlights the extent to which single action potentials can be highly reliable.

Low-level vision in humans and other animals. The close correspondence between single-cell and psychophysical thresholds has been replicated in the domain of vision. For example, in a spatial discrimination task, Parker and Hawken (1985) found the sensitivity of single neurons in V1 to distinguish different positions of a line was similar to the ability of humans and monkeys to report the direction of a displacement of a line in a binary decision task. Indeed, single V1 neurons could respond within the hyperacuity range, meaning that they could detect displacements smaller than the intercone spacing at the level of the retina. Similarly, Bradley, Skottun, Ohzawa, Sclar, and Freeman (1987) found that best discrimination thresholds of single neurons in V1 for orientation were comparable to the psychophysical performance of cats. The same has been reported with rhesus monkeys (Vogels & Orban, 1990). Single neurons in V1 also match psychophysical performance in stereoscopic depth judgments (e.g., Prince, Pointon, Cumming, & Parker, 2000).

The same close correlation between single neurons and psychophysics extends beyond V1. For example, Newsome, Britten, and Movshon (1989) trained rhesus monkeys to perform a forced-choice direction of motion discrimination task in which a dynamic random dot display was presented and in which a fraction of the dots moved in a coherent single direction. On each trial, the monkey reported whether the coherent dots moved in one direction or another. The difficulty of the task was varied by varying the proportion of dots that moved coherently, and performance was assessed across a range of difficulty levels. Threshold was defined as the level of coherent motion in which performance reached 50%. Motion selective single cells in brain area V5 were identified and recordings were taken simultaneously with the behavioral trials. Figure 7 shows the relative sensitivity of single neurons to behavior for 218 V5 neurons across three monkeys. Ratios less than one indicate that the single neuron outperformed the monkey, and ratios larger than one reflected the opposite. As is clear from this graph, the majority of neurons performed the task at a level similar to that of the monkey, with a similar number of neurons outperforming and underperforming the animal. In a review of this article, Morgan (1989) summarized the Newsome et al. findings as showing that "perceptual decisions may be correlated with the activity of a small number of neurons, perhaps as few as one" (p. 20).

In a subsequent study, Britten, Newsome, Shadlen, Celebrini, and Movshon (1996) assessed the trial-by-trial covariation of neural and behavioral responses for near threshold stimuli, conditions in which the monkeys made an error on a substantial proportion of the trials. They reported a small but significant association between these measures. That is, the monkeys tended to choose the preferred direction of a neuron when it fired at an above average rate. The most sensitive neurons could predict behavior on a trial-by-trial basis at approximately 70%. Similar results have been obtained in the medial superior temporal cortex (Celebrini & Newsome, 1994). Although the psychophysical-neural trial-by-trial covariation in V5 and medial superior temporal cortex is substantially weaker than the correlation between peripheral nerve responses and psychophysical responses reported above (e.g., Vallbo and Hagbarth, 1968), it nevertheless highlights the fact that single neurons carry a high degree of information in cortex. Given that Newsome et al. only recorded from a tiny fraction of the relevant neurons, it is not implausible to suggest that other neurons would show much stronger associations.

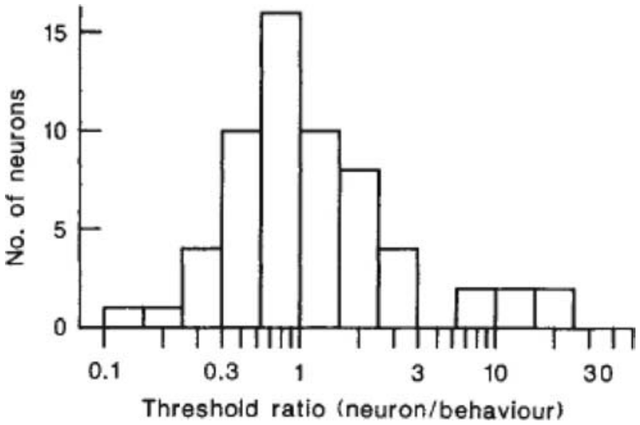

Figure 7. From Figure 2 in "Neuronal Correlates of a Perceptual Decision," by W. T. Newsome, K. H. Britten, and J. A. Movshon, 1989, *Nature, 341,* p. 54. Copyright 2009 by Macmillan Publishers. Reprinted with permission from Macmillan Publishers. A comparison of the relative sensitivity of 216 neurons in the middle temporal area (MT) and the psychophysical performance of rhesus monkey observers in a direction of motion discrimination task. Ratios less than 1 indicate that the single neuron outperformed the monkey, and ratios larger than 1 reflect the opposite. As is clear from this graph, the majority of neurons performed the task at a similar level to the monkeys, with a similar number of neurons outperforming and under performing the animals.

On the other hand, Purushothaman and Bradley (2005) found that when rhesus monkeys discriminated more closely related directions of motion, the sensitivity of single neurons in V5 fell below psychophysical judgments by a factor of 2 to 3. They found that neuron sensitivity levels only matched psychophysical performance when the activity of the most accurate neurons were pooled. One possible interpretation of this finding is to conclude that grandmother cell coding breaks down when fine direction processing is involved and that under these conditions, the activity of multiple neurons needs to be consulted in performing the task. Another possibility, however, is that these V5 neurons pool their outputs onto even more directionally precise neurons in other brain areas, or that the critical neurons in V5 were missed. Indeed, Purushothaman and Bradley (2005) themselves noted that they only recorded from a small subset of the relevant neurons and concluded that they cannot rule out the presence of neurons with even greater precision.

Some More Striking Findings

The observation that the firing of single neurons is correlated with perception and action is widespread. Before considering the association of single neurons with high-level perception, I review a number of additional striking findings.

Escape response in zebra fish. In the domain of motor control there has been much consideration of so-called command neurons, in which the discharge of a single neuron is associated with the execution of a particular behavior. The best-studied example of command cells in vertebrates is the Mauther cell in teleost fish. These neurons are located hindbrain and project down the spinal cord, where they synapse with motor neurons. Gahtan and Baier (2004) reviewed evidence that a single action potential in the Mauther cell leads to a C-shaped bending of the entire body, part of the process of an escape response. Kupfermann and Weiss (1978) rejected the claim that Mauther cells constitute command neurons, as the relevant escape behavior is not eliminated when individual or small groups of Mauther neurons (but not all) were ablated. But the critical point for our purposes is that a single action potential in a single neuron produces a functional behavior. This does provide evidence for grandmother cells by the standard definition of the term that admits redundancy (e.g., Barlow, 1995; Gross, 2002).

Communication in electric fish. Fish from the species *eigenmannia* generate stable sinusoidal electrical discharges that play a role in social communication and in their ability to locate objects through electrolocation. Different electric fish produce slightly different frequencies of electric discharges. When one fish is exposed to an interfering signal of another, it will lower its frequency in response to a slightly higher interfering frequency, and it will raise its frequency in response to a slightly lower frequency—the so-called jamming avoidance response. Heiligenberg (1990) reviewed studies that showed that the performance of the organism could be captured by the behavior of single neurons.

Sound localization in humans and guinea pigs. Humans locate sounds below 1,500 Hz primarily on the basis of small difference in arrive time of sounds in the two ears, so called interaural time difference or (ITD). According to the classic model of sound localization (Jeffress, 1948), ITDs are coded by coincidence detectors that fire maximally when spikes from the two ears arrive at the same time. The way coincident detectors measure a given ITD is that the afferent neurons from the two ears have different propagation delays, and accordingly, simultaneous input to a given detector requires that one ear receives the sound signal slightly before the other (thus, the ITD that a given neuron codes is a function of the difference in the propagation delays from the two ears). A critical finding is that humans can detect ITD as brief as 10–20 μs, and this has been assumed to be too brief a disparity for single-cell coincident detectors to function, given the variability of single neural events. As a consequence, it has been commonly claimed that spatial location was based on a collection of detectors, with responses from many neurons combined. However, Shackleton, Skottun, Arnott, and Palmer (2003) recorded from single neurons in the inferior colliculus of 15 guinea pigs, neurons that are involved in sound localization and that are sensitive to ITD. Shackleton et al. found that there was sufficient information in the firing rates of individual inferior colliculus neurons to match psychophysical performance in humans.

Whisker movements in rats. In the domain of motor control, Brecht, Schneider, Sakmann, and Margrie (2004) reported that a train of action potentials in a single pyramidal cell of rat primary motor cortex can cause whisker movement, demonstrating that the activity of single neurons in cortex can have measurable and adaptive significance. It is critical to note that the whisker responses to the stimulation were prolonged and complex, suggesting that these stimulated neurons coded for motor plans rather than single muscle contractions. As few as 10 action potentials in a single neuron produced robust movement, and fewer still may be sufficient.

Related to this work, Houweling and Brecht (2007) trained rats to respond to microstimulation of somatosensory (barrel) cortex. Microstimulation of somatosensory cortex produces a tactile sensation in humans and animals, but the procedure generally activates multiple neurons. After training with a standard microstimulation procedure that activated multiple neurons, they produced a train of action potentials in a single neuron. The rats continued to show sensitivity to the stimulus. That is, the activity of a single neuron is perceptible, and Houweling and Brecht (2007) took this as evidence for highly sparse cortical coding for sensation.

Birdsong in zebra finch. Hahnloser, Kozhevnikov, and Fee (2002) recorded from single neurons involved in the generation of the vocalizations (song) of the zebra finch. The vocalizations are complex, and are organized into units called song syllables that themselves are broken down into complex sequences of sounds that vary on a 10-ms timescale. They recorded from single neurons in one of the nuclei important in producing songs (the so-called HVC nucleus) and found a short burst of spikes at specific time periods with respect to certain syllables of the song. The association between single neuron activation and song was striking, leading Hahnloser et al. to conclude that neurons within the HCV nucleus may constitute a grandmother cell representation of time in the sequence.

Related to this work, Wang, Narayan, Grana, Shamir, and Sen (2007) recorded from neurons involved in discriminating songs of conspecifics in the primary auditory cortex (field L) of zebra finch. Behaviorally, songbirds can accurately discriminate between songs based on a single presentation. Wang et al. set out to ask the question of whether the activation of a single neuron in L1 could match the behavior of the animal, which is what they found. They took their findings to be consistent with the lower-envelope principle, in which a single neuron was able to distinguish between songs.

Abstract concepts in mice. Lin, Chen, Kuang, Wang, and Tsien (2007) recorded from single neurons in the CA1 region of the hippocampus that responded to the concept of "nest." They recorded from seven mice and identified eight neurons that responded to a wide range different nests. For example, the neurons responded to nests in a variety of locations, independently of physical shape and appearance (e.g., the cells responded to nests of several different geometric shapes, including circular and triangular nests, as well as nests above and below ground), and construction (e.g., nests built of tin caps, plastic bottle caps, and cotton). At the same time, they did not respond to similar shaped objects. This combination of generalization across visual forms but selectivity to nests suggests that these neurons code for the concept nest, rather than some specific visual feature. Lin et al. concluded that these cells constitute grandmother cells for nests.

Single Cell Recording Studies of Objects, Hands, Faces, and Words in Monkey and Man

The review above provides clear-cut evidence that neurons respond in highly selective ways to a variety of stimuli, including visual processing stages beyond V1. But what about neural responses to highly complex visual stimuli? A wide range of neurophysiological research provides evidence that the visual coding of objects and faces is hierarchically organized within the inferior temporal lobes (the so-called ventral visual stream), with downstream neurons responding to progressively more complex visual stimuli. From V1, neurons project to area V2, then V4, and then inferotemporal (IT), with the most anterior section (TE) responding to complex stimuli (and rarely responding to simple ones). For example, IT neurons appear to code object parts in a highly specific manner (e.g, Brincat & Connor, 2004). Furthermore, the general organization of cells in higher visual levels appear to respect the topographic organization of V1, with cells in a given cortical area responding to one complex form and nearby neurons coding for related shapes (e.g., Fujita, 2002; Fujita, Tanaka, Ito, & Cheng, 1992).

The first report that single cells respond selectively to familiar objects was reported by Gross, Bender, and Rocha-Miranda (1969) who identified neurons in the IT cortex of macaque monkey that responded to hands. Gross (1994) reminisced that he was reluctant to use the term *hand cell* for fear of ridicule and noted that there were no attempts to support or deny these general finding for about a decade (when Perrett, Rolls and, Caan, 1979, reported *face* cells). These observations highlight an intuition that many people share; namely, the brain does not work with localist representations. Indeed, Gross (1994) himself did not interpret his initial or subsequent findings as providing evidence for grandmother cells, noting that even the most selective hand or face selective cells responded to more the than one hand or face. Furthermore, he suggested that this level of selectivity might not extend to all categories of knowledge, noting that selectivity was largely restricted to face and hand stimuli. Gross (1994, 2002) took his findings to support coarse coding.

However, there are reasons to question this conclusion on both conceptual and empirical grounds. The first point to (re)emphasize is that the critical question is not whether a given neuron responds to more than one face but rather whether the neuron contributes to the coding of more than one face. The significance of this distinction was recognized by Perrett et al. (1985). They recorded from cells in the temporal lobe of macaque monkey (superior temporal sulcus) and identified some cells that responded to one person more than others. For instance, one cell responded more to a wide range of views of one familiar person (Paul Smith, one of the experimenters), compared to another (David Perrett, another experimenter), such that the identity of the person could be determined by the activation of this one neuron. Perrett, Mistlin, and Chitty (1987) took these findings to be problematic for population encoding models despite the fact that each neuron responded to both Smith and Perrett.

The second point to note is that some neurons appear to respond to only one (or very few) faces (and objects). For example, as discussed above, Young and Yamane (1992) identified 1 neuron among the 850 temporal lobe neurons that was highly selective to 1 of 27 human faces (see Figure 3). This finding strains Gross's claim that face selective neurons always respond to a range of faces. Similar results have been reported a number of times since. For instance, Tamura and Tanaka (2001) recorded from IT neurons while the monkey was presented with 100 photographs of different objects, animals, and faces, and found that half the neurons responded to 7 or fewer pictures, with the most specific responding 5 times more often to the best, compared to second best, picture and only providing a significant response to 3 photographs.

There is also evidence that neural responses to objects can be highly selective (contrary to what Gross, 1994, claimed, above). For example, Logothetis, Pauls, and Poggio (1995) trained two rhesus monkeys to identify a large set of novel computer-generated objects. The novel objects came in two classes—wire-like and amoeboid objects—and varied in similarity within each class. After learning the stimuli, the monkeys performed a visual matching task in which they first fixated at a target from one viewpoint for 2–4 s and then saw a series of test stimuli from the same object class. The monkeys categorized the objects as matching or mismatching, and Logothetis et al. recorded from neurons in the upper bank of the anterior medial temporal sulcus.

The neurons showed a range of selectivity. Some neurons responded more to the wire objects, others responded only to the amoeboid objects. But critical for present purposes, a few (3/796; 0.37 %) responded selectively to one object presented from any viewpoint. Furthermore, some neurons (93/796; 11.6 %) responded selectively to a subset of views of one of the known target objects but less frequently (or not at all) to highly similar objects. See Figure 8 for an example of the types of images and the response of one selective neuron.

In addition, a number of studies have demonstrated the speed at which selective cells are activated. For example, Keysers, Xiao, Földiák, and Perrett (2001) compared the ability of humans to identify naturalistic scenes displayed briefly in rapid sequence with responsiveness of single-cells in the anterior superior temporal sulcus of the macaque. They presented color photographs of faces, familiar and unfamiliar objects, and other naturalistic scenes, using a rapid serial visual presentation procedures, with each picture displayed for the same duration (ranging from 14 ms/image to 111 ms/image) and zero interstimulus interval. In the psychophysical task, human participants were presented the target prior to the sequence and had to indicate whether the target appeared in the subsequent sequence (a detection task where subjects could selectively attend for the target, providing an upper limit of perception in rapid serial visual presentation [RSVP] sequence), and in another version, the sequence was presented first, and then the target was presented. Again, the task was to indicate whether the target was in the sequence (a memory task, in which participants had to identify and remember all the items in the sequence before the target was presented, providing a lower bound of perception in this task).

In the physiological study Keysers et al. (2001) identified 37 neurons that showed selectivity to some of the pictures, responding best to some stimuli and less well to others. The range of the selectivity of the neurons varied, but the most selective neuron fired actively to one stimulus, with its response to the second best

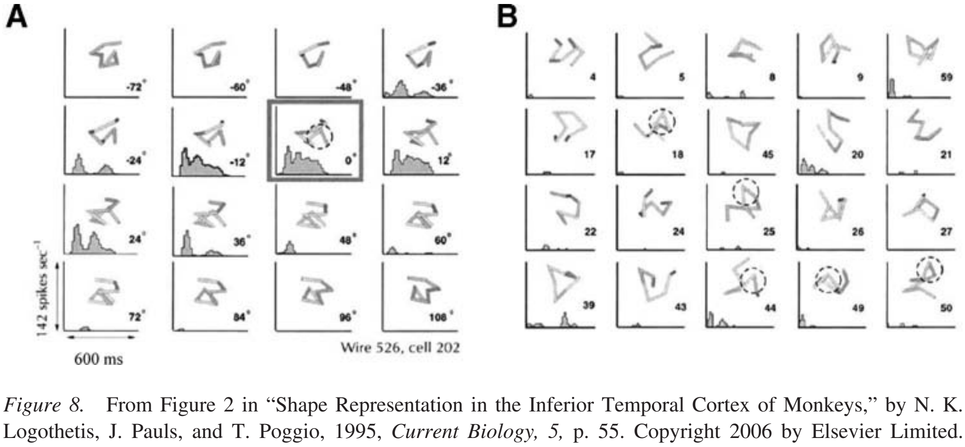

Figure 8. From Figure 2 in "Shape Representation in the Inferior Temporal Cortex of Monkeys," by N. K. Logothetis, J. Pauls, and T. Poggio, 1995, *Current Biology, 5,* p. 55. Copyright 2006 by Elsevier Limited. Monkeys were trained over 3 months to identify images of wirelike images (among other object types). The image depicts the response of one inferotemporal cell (cell 202) in response to one image (wire 526) when depicted in a variety of orientations (A) as well as in response to a variety of other wire objects (B). The neuron responds to a subset of views to wire 256 and very little to other images of trained wires, despite the high similarity between some images.

stimulus barely detectable. Even more striking, neurons showed selectivity when the pictures were presented in RSVP sequence, with 65% of the neurons showing selectivity at the briefest (14 ms) durations. But perhaps most strikingly, performance of individual neurons was comparable to human performance on the task. That is, single neuron performance fell between the detection and the memory conditions. This was the case when neural responses were only considered for 71 ms after the onset of their firing or when the full response of the neurons was considered. Keysers et al. concluded that a very small number of cells in the superior temporal sulcus may support the identification of briefly flashed stimuli.

A few studies have also been carried out on humans (e.g., Fried, Cameron, Yashar, Fong, & Morrow, 2002; Fried, MacDonald, & Wilson, 1997; Kreiman, Koch, & Fried, 2000), and high levels of specificity have been reported here as well. Perhaps the most striking of these finding was reported by Quiroga, Reddy, Kreiman, Koch, and Fried (2005). They studied eight individuals with pharmacologically intractable epilepsy who had depth electrodes implanted in order to localize the source of the seizure onsets. The patients were presented with a range of images on a laptop while recordings were taken from a large number of neurons from the hippocampus, amygdala, entorhinal cortex, and parahippocampal gyrus.

In an initial recording session Quiroga et al. (2005) presented each patient with approximately 100 images of famous persons, landmark buildings, animals, and the like, in search of neurons that responded to a picture. In each patient they identified a number of neurons that responded selectively to one of the images. In order to determine whether the neuron selectively responded to the famous person, object, and the like depicted in the photograph as opposed to some idiosyncratic feature of that specific image, Quiroga et al. (2005) presented the patient with between three and eight distinct photographs of these items in additional sessions on subsequent days. On average, approximately 90 photos were presented in each of these sessions, with a total of 21 sessions across patients. To ensure that participants attended to the photos, participants categorized each picture as a human face or not.

In total, Quiroga et al. (2005) recorded from 343 single neurons, of which 64 responded to at least one of the pictures. The responsive units were all highly selective, responding to ~3% of the pictures. Most of these cells only fired in response to a given person, object, animal, or scene, with 30 of them showing invariance to a particular individual, object, and the like. That is, very different images of the same person or object evoked a strong response. Figures 9–10 show the responses of two highly selective neurons, one that only responded to seven different pictures of the TV star Jennifer Aniston and the other that responded selectively to eight images of the actress Halle Barry. Strikingly, in the case of the Halle Barry cell, the neuron responded not only to a range of quite different images of her face but also to her written name. This highlights the fact that this cell does not code high-level perceptual information, but rather, a memory (or conceptual information) about the person. This is consistent with the finding that the cell was located in the hippocampus, an area of the brain that supports episodic memory, not face perception.

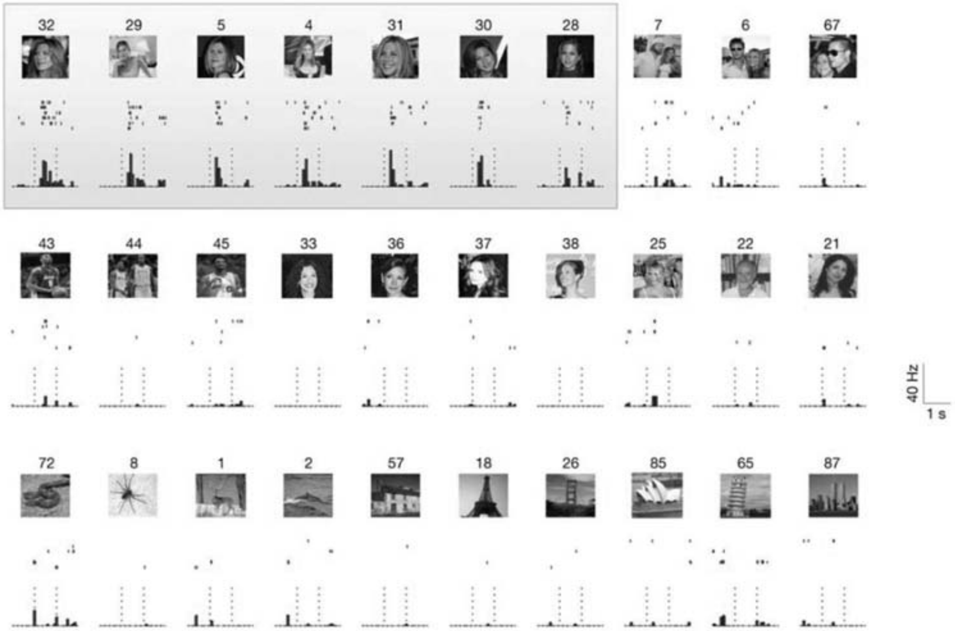

Figure 9. From Figure 1 in "Invariant Visual Representation by Single Neurons in the Human Brain," by R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, 2005, *Nature, 435,* p. 1103. Copyright 2009 by the Macmillan Publishers. Reprinted with permission from Macmillan Publishers. The activation of a single hippocampal neuron in a human in response to a selection of 30 out of the 87 pictures presented. The vertical dashed lines depict the onset and the offset of the image. The cell is highly selective to pictures of the TV star Jennifer Aniston.

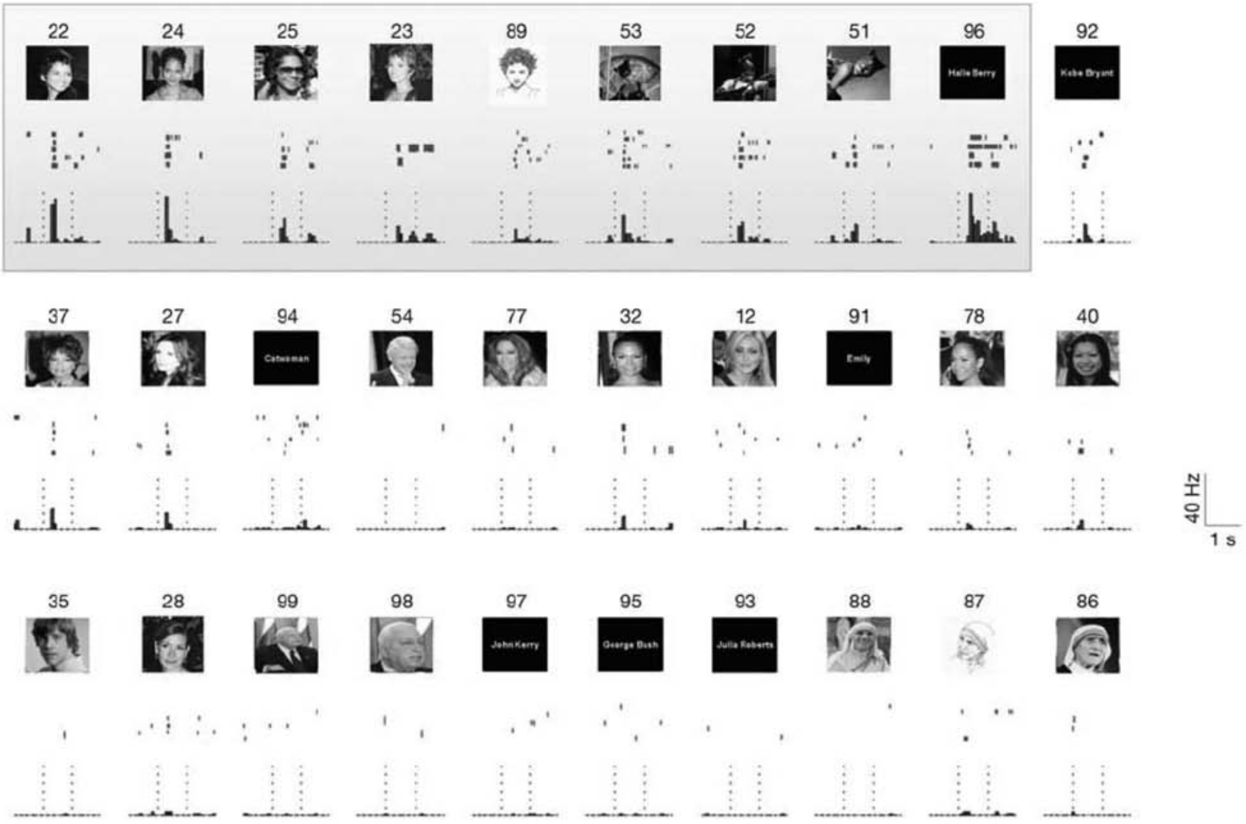

Figure 10. From Figure 2 in “Invariant Visual Representation by Single Neurons in the Human Brain,” by R. Q. Quiroga, L. Reddy, G. Kreiman, C. Koch, and I. Fried, 2005, *Nature, 435,* p. 1104. Copyright 2009 by the Macmillan Publishers. Reprinted with permission from Macmillan Publishers. The activation of a single hippocampal neuron in a human in response to 30 out of 99 images. Again, the vertical dashed lines depict the onset and the offset of the image. The cell is highly selective to pictures of the movie star Halle Berry, as well as her written name.

It is interesting to note that these reports of selective responding are most frequent for highly trained (or familiar) objects and faces. For example, in the Logothetis et al. (1995) study, the monkeys were trained on a large set of objects for months, and selective responses were only observed for the trained images (novel objects failed to support these effects). Various studies have observed increased neural selectivity as a function of amount of training (e.g., Erickson, Jagadeesh, & Desimone, 2000; Kobatake, Wang, & Tanaka, 1998). Furthermore, Freedman, Riesenhuber, Poggio, and Miller (2006) reported a correspondence between the selectivity of single neurons and the ability of rhesus monkeys to categorize stimuli. Two monkeys were trained to categorize images morphed between a prototypical cat and dog. These images were presented in one orientation (with cats and dogs presented in 0° of rotation in the picture plane), and after training, the monkeys could categorize the images appropriately, including the morphed images that fell near the category boundary. At the same time, some neurons in IT and prefrontal cortex responded selectively to these images (with neurons in IT more sensitive to variations in the shape and neurons in prefrontal cortex more sensitive to the category boundary). However, when the monkeys were tested on the same photographs rotated away from the trained orientations, their categorization performance became progressively worse. In the same way, the selectivity of neural responding in IT was also reduced as a function of rotation (prefrontal neurons were not tested in this condition).

The link between training and response selectivity also lends some support to grandmother coding schemes. If distributed representations mediate the identification of objects and faces as commonly claimed, then there is no obvious reason why selectivity should increase with training. By contrast, on grandmother coding schemes, single neurons are only hypothesized to support the identification of familiar things. Accordingly this association between neural selectivity and object familiarity should be expected.

Conscious Perception and Single Cells

Finally, it is worth mentioning that a number of studies have also provided compelling evidence that the activation of single IT neurons is associated with the conscious experience of identifying high-level visual information. Logothetis and Sheinberg (1996) recorded from IT neurons when presenting monkeys with visual stimuli that produce multistable percepts—that is, the same stimulus is alternatively perceived as one stimulus or another, much like a necker cube. These researchers exploited a phenomenon called binocular rivalry, in which the two eyes receive dissimilar images that cannot be fused into a single percept. Under these conditions humans (and monkeys) alternatively perceive one image or another, with the visual image suppressing the percept of the other. In one series of experiments, Logothetis and Sheinberg (1996) presented a sunburst-like pattern to one eye of a monkey, and a range of other images to the other, including pictures of humans, monkeys, and various artifacts. The monkeys were trained to pull and hold one lever whenever they perceived the sunburst-like pattern, and press and hold onto another whenever another image was perceived. Monkeys alternatively pulled one lever or the other despite a constant visual input, in very much the same way that humans perform the task, reflecting the standard binocular rivalry experience of alternatively perceiving one stimulus or the other.

The critical feature of this study is that Logothetis and Sheinberg (1996) recorded from IT neurons that were visually selective to various images including faces but that responded little if at all to the sunburst pattern. During dichoptic viewing, the stimulus selective neurons fired when the monkey indicated awareness of the stimuli (by pulling the lever) but not otherwise. That is, the neuron firing was correlated with the conscious state of the monkey, not the image that was projected onto the retina. Indeed, for 90% of the stimulus specific cells in superior temporal sulcus and TE areas, neural response was contingent on the monkey reporting the perception of the stimulus. (For a similar result in the context of change blindness in humans, see Reddy, Quiroga, Wilken, Koch, & Fried, 2006).

Part 3: How Well Do Localist and Distributed Coding Schemes Account for the Data?

Despite the numerous reports that single neurons respond selectively to both low and high-level perceptual knowledge in all variety of species, the vast majority of cognitive psychologists and neuroscientists still reject the grandmother cell hypothesis. Instead, various versions of distributed coding schemes are endorsed. In the final section, I compare the relative merits of grandmother and various distributed coding schemes in light of the above data. I also consider a number of general objections to grandmother coding schemes that are often used to dismiss this approach. I show that these objections are unfounded.

Grandmother Cells Versus Dense Distributed Representations

On a dense distributed coding scheme, it is difficult to infer much about the identity of a word, object, or face on the basis of the activation of a single unit, as each unit is involved in coding many different things. This coding scheme is often described as a core claim of the PDP approach, and it is widely assumed that these representations are broadly consistent with biology.

Despite this common assumption, there is little evidence for it. Indeed, the assumption appears completely wrong; if it were true, the field of neurophysiology would not have gotten started. The basic insight from single-cell recording studies is that individual neurons code for low- and high-level knowledge in remarkably selective ways. It is striking (and perhaps telling) that the techniques from neurophysiology that are used to characterize the internal representations of the brain are rarely applied to PDP models of cognition.

The speed with which single neurons respond selectively to high-level visual knowledge also challenges the coding schemes used in some PDP models. For example, Plaut, McClelland, Seidenbert, and Patterson (1996) developed a model of single word naming in which units settle into an attractor pattern. The settling times were thought to correspond to the readers' naming latencies. More generally, attractor dynamics were described as a computational principle employed in a wide range of cognitive phenomena. However, Thorpe and Imbert (1989) challenged this claim. They reviewed evidence that IT neurons respond selectively to an image approximately 100 ms poststimulus and that there at minimum 10 synapses separating retina from object and face selective cells in IT. Given that neurons at each stage cannot generate more than 1 or 2 spikes in 10 ms, Thorpe and Imbert argued that neurons at each stage of processing must respond on the basis of one or two spikes from neurons in the previous stage. This in turn limits the amount of feedback that can be involved in identifying an object. They conclude that face identification (and identification in general) cannot be mediated by a process of settling into an attractor pattern (also see Oram & Perrett, 1992).

Although the results from neuroscience appear to rule out dense distributed representations, PDP models are not restricted to learning these codes (nor identifying objects by relaxing into attractors). Indeed, these models can learn to represent knowledge with coarse distributed (e.g., Dawson, Boechler, & Valsangkar-Smyth, 2000), sparse distributed (e.g., McClelland et al., 1995), and perhaps even localist (Berkeley, 2000) representations. For example, consider again the three layered network of Berkeley et al. (1995) that learned to solve a set of logical inference problems through backpropagation. As discussed above, when the model included 15 hidden units, it learned a set of dense distributed representations (similar to Figure 5B). However, when the model included fewer hidden units, most of the jittered density plots developed what Berkeley, Dawson, Medler, Schopflocher, and Hornsby (1995) called banding patterns. That is, each unit responded to a subset of inputs at a similar level of activation (for an illustration of a banding pattern see Figure 11). It is critical to note that all the input patterns that contribute to a band share a common feature. For example, in the Berkeley et al. (1995) model, the sixth hidden unit only responded to input patterns that included the input feature corresponding to the logical operation "OR." In response to these inputs, the unit was maximally active, much like the illustration in Figure 11. Berkeley (2000) called hidden Unit 6 an "OR detector." That is, Unit 6 appears to be a grandmother unit (for a debate regarding the interpretation of this unit see Berkeley, 2006; Dawson, & Piercey, 2001).

What is to be made of the fact that PDP models can learn various forms of internal representations, ranging from dense distributed to localist? The first point to note is that the above analyses show that some (perhaps many) existing PDP models of word naming and memory do learn dense distributed representations. Although these models are generally considered more biologically plausible than are their localist counterparts (e.g., Seidenberg & McClelland, 1989; Botvinick & Plaut, 2006), this claim appears to be unjustified. In future work it will be interesting to

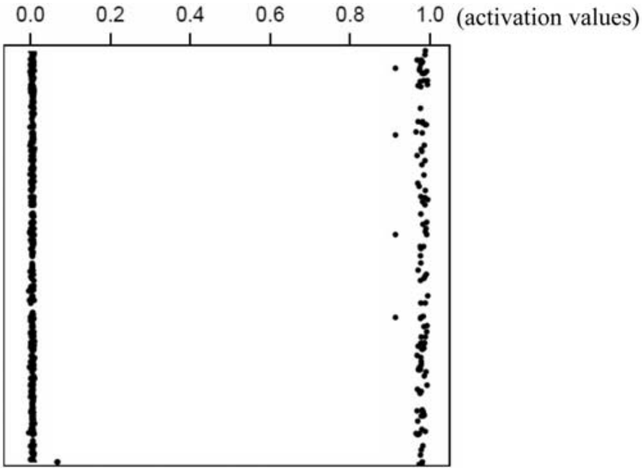

Figure 11. Jittered density plot that displays a banding pattern. In this example, the unit is unresponsive to most items but responds to a subset of items maximally. It is critical to note that inputs that fall within a band often share a common feature. In the case of Berkeley et al.'s (1995) model, which learned to solve logical inference problems with 10 hidden units, hidden Unit 6 of the model responded maximally whenever the input included the OR operation and was off otherwise. That is, the model appeared to learn a grandmother OR unit.

characterize the internal representation in a wider range of PDP models to better specify the conditions in which the various forms of internal representations are acquired.

Second, one of the key theoretical claims of the PDP approach is that the brain does not compute with local representations (there are no grandmother cells). This assumption is challenged by the finding that localist representations can be learned in PDP models through back-propagation and related algorithms. Indeed, the fact that PDP models develop localist representations under some conditions suggests that there are computational advantages to localist coding schemes (otherwise they would not be learned). The obvious implication is that the brain might also use grandmother cells for at least some purposes (cf., Bowers, 2002; Bowers, Damian, & Davis, 2008; Gardner-Medwin & Barlow, 2001; Page, 2000, for some computational reasons why a PDP model might learn localist representations).

However, the most important point is that many researchers consider dense distributed representations a core theoretical claim of the PDP approach (e.g., Bowers, 2002; Elman, 1995; Hummel, 2000; Page, 2000; Smolensky, 1988; Thorpe, 1989). If it turns out that many current PDP models of memory, language, and perception do learn sparse, coarse, or local codes (contrary to the widespread assumption), or if these models are modified so that they learn these types of representations (in order to be consistent with biology), it would amount to a falsification of this theoretical assumption. At minimum, the neuroscience makes it necessary to think about the PDP approach in a fundamentally different way. On the modified approach, the key claim must be that each unit in a biological plausible PDP model should codes for a small number of related and interpretable things.

By contrast, these results are easily reconciled with localist coding schemes. Indeed, single-unit recording studies carried out on localist models mirror the neurophysiological results, with individual units responding in a highly selective manner to stimuli (see Figure 4). Furthermore, the speed of responding to complex images has also inspired localist models in which there is no room for feedback for the sake of identification (e.g., Thorpe, Delorme, & Van Rullen, 2001).[1]

It is also worth emphasizing that the topographic organization of knowledge in the cortex, with spatially proximate neurons coding for related things (tones, arm movements, objects, etc.), is compatible with localist coding. Networks that learn localist representations often rely on lateral inhibition to activate one node in a winner-take-all process, and by representing similar things with nearby units, inhibition can be mediated by short-range lateral connections. Indeed, various competitive neural networks that include winner-take-all units were developed in an attempt to explain the development of the topographic organization of simple cells in V1 (e.g., Kohonen, 1982; Lee & Verleysen, 2002; Williamson, 2001). If biological plausibility is used as a criterion for comparing the coding schemes of current cognitive theories, localist models currently fare better than do PDP models.

Grandmother Cells Versus Sparse Coding

In some respects, sparse coding is similar to grandmother coding schemes; in the limit of sparse coding, an individual neuron is involved in coding two things, and two neurons code for a specific stimulus. Accordingly, it should be possible to infer a great deal about a stimulus by recording from a single neuron, and it is presumably difficult (in practice) to distinguish between sparse and grandmother coding schemes on the basis of assessments of how many neurons are activated by a given stimulus. Nevertheless, there are grounds to contrast these hypotheses.

As noted in Part 1, the sparse representations associated with PDP networks are well suited to support rapid learning, but they are poor at generalization. This computational constraint has led to the complementary learning systems framework, according to which sparse coding is employed in the hippocampus in the service of episodic memory, whereas dense distributed representation are employed in the neocortex to support word, object, and face identification (among other functions).

In a number of places, PDP modelers argue that neuroscience lends support to the hypothesis that coding is sparser in the hippocampus, compared to the neocortex, but there are reasons to

[1] The speed with which single cells respond to faces and objects might appear to be inconsistent with one of the localist models discussed above, namely, the adaptive resonance theory. In this theory, object identification is achieved when localist representations in Layers n and $n - 1$ settle into a stable attractor pattern (a resonance; Grossberg, 1980). However, local representations of familiar stimuli in Layer n are activated in a bottom-up fashion. The resonance was introduced for the sake of learning (to solve the stability–plasticity dilemma in which new learning can erase old memories; Grossberg, 1976), and Grossberg (1980) claimed that resonance is required for the conscious experience of a stimulus but not for its unconscious detection. This leads to the prediction that the selective responding of a face cell (or an object cell) in inferotemporal at ~100 ms poststimulus precedes the conscious identification of the stimulus. At present, there is good evidence that the unconscious detection of a stimulus requires less time than forming a memory trace of a stimulus, with the latter requiring more attention and, perhaps, consciousness (e.g., Subramaniam, Biederman, & Madigan, 2000).

question this claim. For instance, O'Reilly and Norman (2002) cited Boss, Turlejski, Stanfield, and Cowan (1987) as providing evidence for this view. However, Boss et al. did not record from cells in the hippocampus or cortex but rather counted the number of neurons in areas CA1 and CA3 cells in the hippocampus of two species of rats. They did report that the dentate gyrus (not part of neocortex) that feeds into area CA3 includes a higher density of neurons, suggesting a convergence of information, but this does not provide any evidence that neurons have sparser levels of firing in the hippocampus, compared to neocortex. The only other empirical study cited by O'Reilly and Norman (2002) was by Barnes, McNaughton, Mizumori, Leonard, and Lin (1990) who recorded from place cells in regions CA1 and CA3 of the hippocampus in rats, as well as from cells in the entorhinal cortex (not part of the neocortex) and subiculum (again not part of neocortex). The cells in CA1 and CA2 did respond in a more selective (sparse) manner in response to the rat's location in a spatially extended environment, but given that Barnes et al. did not record from the neocortex, the relevance to the complementary learning systems hypothesis is unclear. Similarly, the studies cited in the original McClelland et al. (1995) article do not speak to this issue. They also cited the (irrelevant) Barnes et al. (1990) study, as well as a study by Quirk, Muller, & Kubie (1990). However, this study only recorded from the hippocampus; so again, it does not speak to the relative sparseness of firing in the hippocampus and neocortex.

To provide some evidence for the complementary learning systems hypothesis, it would have to be shown that there is a higher level of sparseness in the hippocampus, compared to the neocortex. Furthermore, the relevant question should be whether the hippocampus codes for information in its domain (e.g., space, or an episodic memory) more sparsely than the neocortex codes for information in its domain (e.g., a perceptual representation of a face, word, or object). This sort of evidence is not currently available.

Indeed, the data reviewed in Part 2 highlight the extent to which sparse coding is used in both the hippocampus and neocortex (cf., Olshausen & Field, 2004; Shoham, O'Connor, & Segev, 2006). There is a good reason that the brain avoids dense distributed representations; namely, the metabolic cost of firing neurons is high. Lennie (2003) estimated that these costs restrict the brain to activate about 1% of neurons concurrently, and he takes these findings as consistent with localist coding. These metabolic costs also undermine a common argument that PDP models are more efficient than localist ones, in that they require fewer units to code a given amount of information (e.g., Hinton et al., 1986). Instead, the opposite is true, with distributed representational schemes maximally inefficient in biological terms.

Could the sparse distributed coding scheme employed in the hippocampus be employed in the cortex as well? Given the review above, it is clear that the cortex does use some form of sparse coding, but there are computational reasons to assume that the cortex does not use the sparse codes learned in PDP networks. These codes do not support generalization, and generalization is a core capacity of the cortex that supports language and perception. What is needed is a form of sparse coding that supports fast learning in the hippocampus and that supports generalization in the cortex. One candidate is localist coding. Indeed, as discussed below, one of the virtues of localist coding schemes is their ability to support widespread generalization. To summarize, the PDP approach is again challenged by biology. It appears that both the hippocampus and the cortex employ highly sparse coding (perhaps grandmother cells), contrary to the complementary learning systems approach.

Grandmother Cells and Coarse Coding

It is commonly assumed within the neuroscience literature that the brain uses some form of coarse coding. On this view, perception (and behavior) involves a hierarchy of processing steps in which individual neurons code for a range of similar things, but without much precision. The core claim of this approach is that the unique identification of a given object, word, and the like, cannot be determined by the activation of one neuron and that it is necessary to pool across a collection of noisy units in order to get a stable and exact measure of an input.

This version of distributed coding does not have the difficulties highlighted above. Unlike dense distributed representations, individual neurons do convey meaningful (interpretable) information, and accordingly, this approach might be reconciled with the neurophysiological evidence reviewed above. And unlike sparse coding, coarse codes are designed to support generalization (as discussed below). Furthermore, coarse coding is often implemented in networks that respect the topographic organization of knowledge. The key question for present concerns, then, is whether coarse coding provides a better account of the data than do grandmother neurons.

The most common argument put forward in support of coarse coding is that single neurons always respond to more than one object or face (e.g., Rolls, Treves, & Tovee, 1997), but as discussed in Part 1, the logic is flawed. Single units in localist models also respond to more than one input (e.g., within the IA model, the *dog* unit responds to the inputs *dog* and *hog*, but it does not contribute to the coding of *hog*). Furthermore, as shown in Part 2, the common claim that single neurons always respond to many different objects or faces is not always borne out; in a number of studies, a single neuron responded to one face or object among many.

Another common way to compare the two approaches has been to determine whether the response of a single neuron can encode enough information to account for high-level vision (e.g., face recognition) or whether it is always necessary to consider the coactivation of multiple neurons. In a series of studies, Rolls and colleagues (Rolls et al., 1997) provide evidence for the latter, and they take this as support for distributed coding. For instance, Rolls et al. (1997) identified a set of 14 face selective neurons in the temporal cortex of the macaque on the basis of the criterion that they responded at least twice as much to an optimal face stimulus, compared to an optimal nonface stimulus (out of a set of 68 images). In the critical analysis, they recorded the responses of these 14 neurons in response to 20 faces. They observed that the response of one neuron was not strongly associated with a given face, but various algorithms that considered the population of neural activity in order to estimate the likelihood that a given face was presented on a given trial did much better. A Bayesian analysis indicated that the percentage correct prediction increased from 14% correct, when one neuron was considered, to 67%

correct, when on all 14 neurons (with 5% being chance) were considered. Furthermore, they estimated that the number of stimuli that could be coded by this population increased approximately exponentially with the number of cells sampled. Rolls et al. took this finding as inconsistent with grandmother cell coding schemes, which the number of stimuli encoded increases linearly with the number of cells.

But there is a problem with this experiment and with the approach in general. First, the study was carried out on a set of face cells that were not highly selective. Indeed, most of the neurons responded to most of the faces presented. If the same analysis were carried out on the set of face cells described by Quiroga et al. (2005), a different conclusion would presumably follow. Second, Rolls et al. (1997)'s conclusion rests on the assumption that they have recorded from the critical face neurons at the highest level of the face identification system. However, if they recorded from neurons earlier in the processing hierarchy (or from neurons that coded for different faces), their results can easily be explained within a localist coding scheme.

To illustrate this point, consider again the IA model of visual word identification. The model includes a set of letter detectors at Layer $n - 1$ that feed into local word representations at Layer n. In this model, the number of words the model can identify is linearly related to the number of word units (one word per unit). But the number of words that the experimenter can infer from the letter level increases much more rapidly. Indeed, the identity of all 4-letter English words (indeed, all four letter words in any language that includes the Roman alphabet) can be determined with reference to only 104 letter units (26 × 4, with each letter coded by position; that is, in the model, the letter A-in-position-1 and A-in-position-2 are coded separately). But the ability of the experimenter to infer the identity of words based on a pattern of activation across the letter level does not provide any evidence that the model does the same; indeed, for the model to succeed, a local (grandmother) word unit needs to be activated beyond some threshold. In the same way, little follows from the observation that Rolls et al. (1997) can infer the identity of faces by pooling across a set of face cell. The critical question is not whether the experimenter can derive information from a pattern of activation, but how the brain does it.

In fact, Rolls (2007) acknowledged that his results can be reconciled with a local coding scheme if cells downstream respond more selectively. However, he dismisses this possibility, noting examples of studies that failed to observe highly selective responding in the amygdala and the orbitofrontal cortex. But this argument is not secure. It rests on further demonstrations that single cells do not encode faces accurately. But it is presumably quite easy not to find grandmother cells. The more impressive (and telling) finding is that a number of labs have identified single neurons that do respond highly selectively to single faces in IT, as reviewed in Part 2. Furthermore, highly selective responding to faces and objects has also been reported in downstream areas, including the prefrontal cortex (e.g., Freedman, Riesenhuber, Poggio, & Miller, 2001, 2003). Indeed, Freedman et al. (2003) described a localist model of object identification that could account for the selective neural firing in both IT and prefrontal cortex. Freedman et al. (2003) themselves rejected grandmother cell theories, but it is not clear why; each unit in their model codes for one specific thing.

A striking example of a confusion relating population and grandmother cells can be found in Hung, Kreiman, Poggio, and DiCarlo (2005). They studied the neural coding scheme of objects in IT by recording from over 300 neurons while monkeys categorized photos of black and white objects. The firing pattern of the neurons was input to a network classifier that pooled all the input signals onto two output units where images were categorized based on which output unit was more active (a winner-take all decision unit).

Hung et al. (2005) noted that the performance of the winner-take-all decision units was highly accurate, with the 94 +/−4% accuracy in categorizing the objects when the recordings from a random set of 256 neurons are considered. This indicates that most of the information required to categorize the objects can be retrieved from these IT neurons. But they conclude that the brain employs population coding based on the finding that performance in their model increased as a function of the logarithm of the number of units that contributed to the winner-take-all decision. Indeed, this logarithmic, as opposed to linear, function is taken to rule out grandmother coding schemes.

However, this is a surprising conclusion given that Hung et al. (2005) have explicitly developed winner-take-all output units that read out this pattern of activation. It is not clear what else a grandmother unit could be but a cell that pools together patterns of activity from a lower level. The fact that performance was a logarithmic function of the number of units provides no evidence that the brain uses a distributed coding scheme, just as their model attests.

The Rolls et al. (1997) and Hung et al. (2005) studies relate to a number of articles in which the optimal methods of decoding the information in populations of coarse coded neurons are considered (e.g., Jazayeri & Movshon, 2006; Oram, 1998). In most cases, it is assumed that the brain computes with population cells rather than grandmother cells. But what is rarely considered in the analysis is how the brain implements these algorithms. The important point to emphasize (again) is that a range of pooling algorithms, including Bayesian methods, can be implemented in winner-take-all networks; that is, networks that implement grandmother coding schemes (Zhang, Ginzburg, McNaughton, & Sejnowski, 1998). Thus, critics of grandmother cell coding need to do more than show how well distributed coding schemes can encode information. They also have to provide evidence that the brain does not implement their theories with winner-take-all neurons. It is striking how authors sometimes argue for some sort of distributed coding schemes but explicitly include winner-take-all units in their models (e.g., Hung et al., 2005; Pouget, Dayan and Zemel, 2000).

A different approach to assessing the relative merits of population and grandmother coding schemes was carried out by Newsome and colleagues (Groh, Born, & Newsome, 1997; Nichols & Newsome, 2002; Salzman & Newsome, 1994). They assessed the impact of artificial microstimulation of motion sensitive neurons in V5 when monkeys viewed coherent motion in dot displays. They presented monkeys with a motion signal while neurons coding for different direction of motion were stimulated, as depicted in Figure 12. The logic of the studies is that on a winner-take-all account, the perceived motion should be determined by the most active motion sensitive neurons, which should either be the neurons stimulated by the motion signal itself or by the microelectrode. On a population account, by contrast, percep-

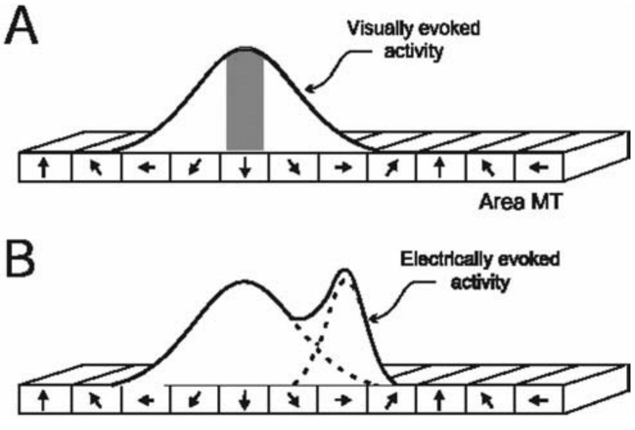

Figure 12. Adapted from Figure 1 in “Middle Temporal Visual Area Microstimulation Influences Veridical Judgments of Motion Direction,” by M. J. Nicholas and W. T. Newsome, 2002, *Journal of Neuroscience, 22,* p. 9531. Copyright 2002 by Journal of Neuroscience. Reprinted with permission from Journal of Neuroscience. In Figure 12A, direction sensitive neurons in inferotemporal are selectively activated in response to a motion signal, and in 12B, another set of directionally sensitive neurons are artificially activated. On a winner-take-all account, the perceived motion should be determined by the most active motion sensitive neurons, which should either be the neurons stimulated by the motion signal itself or by the microelectrode—in this example, perception should be entirely determined by the direction of motion evoked by the electrode. On a population account, by contrast, perception of direction should be some sort of integration of all the active motion signals.

tion of direction should be some sort of integration of all the active motion sensitive neurons, with the perceived motion some sort of compromise between the two distinct signals.

The results of the studies have been somewhat mixed. In the initial studies adopting this logic Salzman and Newsome (1994) provided evidence for a winner-take-all algorithm for computing perceived direction, whereas Groh et al. (1997) provided evidence for a distributed coding scheme. More recently, Nicholas and Newsome (2002) provided evidence for both mechanisms operating, with vector averaging providing a better account of performance when the direction sensitive neurons activated by the motion signal and microelectrode differed by less than 140°. When the visual and microstimulated directions were more disparate (~150°), a winner-take-all algorithm provided a better account of monkeys’ behavioral responses (angles smaller than this was could not provide unambiguous evidence for one account or another for various reasons). They conclude that the visual system may use both local and distributed mechanism to code for motion detection.

The finding that the motion system employs winner-take-all responses (under some conditions) provides strong evidence in support of localist coding schemes beyond V1, but is it the case that these findings also highlight the extent to which distributed coding schemes contribute to perception? Perhaps, but there is another straightforward interpretation. The finding that monkeys sometimes perceived a direction of motion between the coactive directions does indeed show that a collection of these neurons was involved in computing perceived direction, but the manner in which the brain decodes this information is left unspecified. It is possible the pattern of activation across these neurons is pooled at a higher level in a hierarchy, with a winner-take-all process at this level determining perceived direction. Again, the fact that the Nicholas and Newsome (2002) did not identify the winner-take-all neurons under some conditions does not mean that they do not exist.

To summarize it seems clear that the cortex does not code for information with dense distributed representations as often assumed by advocates of the PDP approach. Furthermore, the sparse codes learned in PDP models do not provide a plausible model of processing in cortex, as these representations do not support the forms of generalization necessary for object recognition, among other functions. However, it seems fair to conclude that coarse coding provides a reasonable account of the data reviewed above. Coarse coding can accommodate the single cell recording findings reviewed in Part 2 and can support object recognition. The main point to emphasize, however, is that localist coding schemes also provide a reasonable account for the data. There is no reason to reject grandmother cells in favor of coarse coding based on the above neuroscience.

Might the Brain Exploit Both Distributed and Localist Coding?

On a strong version of the grandmother cell hypothesis, the brain relies on localist coding throughout a hierarchy of visual processing stages, with photoreceptors as the first step and grandmother cells as the last. Similar hierarchies might apply to behavior, with grandmother cells (or command neurons) as the first step specifying a general motor program, and collection of specific (and discrete) motor commands as the final step. Indeed, on an extreme version of this hypothesis, every neuron involved in coding something has a single and precise interpretation.

A weaker version of the grandmother cell hypothesis can accommodate some degree of distributed coding as well. For example, distributed coding might be implemented in a subset of the visual processing steps involved in identifying objects or, alternatively, support visual processes for other functions, such as determining the location of objects. Similarly, distributed coding may play a central role in some (or indeed all) processing steps involved in generating behavior. All that is strictly required for a grandmother cell theory is that the cells at the top of the visual hierarchy code for complete words, objects, and faces in a localist manner.

There is no logical reason as to why the brain is restricted to employing one coding scheme, and indeed Thorpe (1989) argued that low- and high-level vision rely on coarse and localist coding, respectively. On this hypothesis, the simple cells in V1 that compute line orientations exploit coarse coding, and cells in IT involved in identifying objects and faces exploit localist (grandmother) coding. It is interesting to note that this hypothesis turns a classic theory of vision on its head. On a more traditional view, perception of orientation is determined by the most active simple cell in V1 (e.g., Hubel, 1995; Knudsen, du Lac, & Esterly, 1987), whereas high-level vision is supported by some sort of population code. According to Thorpe (1989), however, the opposite might be the case.

Thorpe (1989) outlined both computational and empirical evidence in support of this position. He made the observation that knowledge is structured differently in the two domains. Low-level knowledge must code for continuous variables; for instance, line orientations vary in degree; disparity, by visual angle; movements, by direction in three-dimensional space, and so on. In these cases, specific values along a given dimension (e.g., a given line orientation) might be more effectively coded through a collection of broadly tuned receptors. In this way, there is no need to include a separate detector for each detectible orientation or disparity, and so on. By contrast, high-level knowledge (e.g., faces, objects, letters, words, etc.) is organized into discrete categories (that are limited in number), making local coding schemes more tractable in this context.

The key empirical observation that Thorpe (1989) made in support of this hypothesis is that humans can discriminate line orientations that differ by as little as .5°, whereas the preferred orientation of simple cells differ by ~10°. Clearly, these fine discriminations cannot rely on a comparison of two simple cells that maximally respond to the different orientations. At the same time, there is no evidence that cells higher in the hierarchy of visual processing steps code for orientation more precisely. Somehow, the simple cells in V1 appear to code orientation.

How is it done? According to Thorpe (1989), orientation is coded through the activation of multiple simple cells. However, the distributed coding scheme in V1 appears to be quite different from vector averaging in which neurons vote as a function of their activation level (cf. Georgopoulos et al., 1986). A key observation is that orientation discrimination of single simple cells can approach psychophysical performance (e.g., Bradley et al., 1987). Although this might seem inconsistent with the discussion above, the trick is that orientation discrimination is best performed by cells that only respond weakly to the two contrasting orientations. The excellent discrimination of these cells is due to the fact that generalization in simple cells is Gaussian in form, with a drop off in firing steepest ~10° to either side of the maximal response (the steep slope supports the best discrimination). In other words, discrimination is best performed by weakly active cells that preferentially respond to orientations ~10° to either side (e.g., Parker & Hawken, 1985). Thorpe (1989) reviewed human psychophysical and modeling data (from a PDP model) consistent with the claim that fine discriminations of orientation are supported by partially active simple cells rather by the most active ones.

Critical for present concerns, to read off a specific line orientation from these weak responses, at least two neurons need to be considered. For example, a precise representation of a line oriented just off vertical can be obtained through the relative (and moderate) response of one simple cell that fires maximally to a line oriented 10° to the right of vertical (call this Neuron 1) and another simple cell that fires maximally to a line oriented 10° to the left (call this Neuron 2). A line oriented 1° right of vertical would then be coded by Neuron 1 firing at a slightly higher rate than Neuron 2, whereas a line oriented 1° to the left of vertical would be coded by the Neuron 2 firing at a slightly higher rate (and a vertical line by the two neurons firing at the same rate). This pattern of activation across partially active simple cells provides a more precise measure of orientation than the most active simple cell, and this is required to account for the fine discrimination that can be performed.

Two points need to be emphasized here. First, even if the early visual stages of object identification are implemented in a coarse coding scheme, this only rules out the strong version of the grandmother cell hypothesis, according to which every stage of the hierarchy of visual processing steps is localist. The key claim is not challenged by these observations; that is, the visual representations for words, objects, and faces may still be coded locally.

Second, the finding that the visual system uses coarse coding to support fine orientation discriminations does not rule out a localist coding scheme in V1 for the sake of object identification. That is, the most active simple cells may provide the input to complex cells, which in turn activate more complex localist representations in a hierarchy of processing steps along the ventral visual stream, with grandmother cells in the inferior temporal cortex as the final step. The distributed representations in V1 required for the detailed discrimination of orientations might instead serve as the input to a dorsal visual system that mediates sensorimotor transformations required for visually guided actions directed at objects (Milner & Goodale, 2006). Here, fine spatial and metric knowledge (rather than discrete category knowledge) is critical for skilled performance. By contrast, object identification tends not to rely on fine orientation discrimination (Biederman, 1987).

It is interesting to note that Zipser and Anderson (1988) and Pouget and Sejnowsky (1997) developed PDP models that took, as input, information about the location of an image on the retina as well as the eye position and transformed this information to spatial coordinates that could guide behavior. This is a core function of the dorsal visual stream. The models differed in some important respects, but critical for present purposes, they compared the response profiles of the learned hidden units to the activation of single neurons in the parietal cortex. In both cases, the hidden units developed complex receptive fields that matched quite well with the receptive fields of the neurons.

One possible interpretation of these findings is that the distributed representations learned in PDP models provide a more promising account of the neural coding schemes employed in the dorsal visual stream. Although there are relatively few single-cell recording studies of the dorsal visual stream, and even fewer studies comparing the responses on these neurons to PDP networks, the computational tasks of the ventral and dorsal visual streams make it at least plausible that they code information in qualitatively different ways. For example, the ventral stream needs to address the stability–plasticity dilemma, given that new object representations can be learned without catastrophic loss of prior knowledge. By contrast, the dorsal system only needs to learn sensorimotor mappings for an organism's current body. Indeed, it would be maladaptive to preserve the spatial and motor representations that support behavior across a lifetime, given that our bodies grow. These considerations led Grossberg (2000) to develop qualitatively different theories of learning and processing in these two domains. Relevant for present purposes, these considerations are consistent with the hypothesis that localist coding is more adaptive in the ventral system, given that sparse (or local) representations are more immune to catastrophic interference (e.g., McClelland et al., 1995). In addition, localist representations are well suited for supporting symbolic models of cognition (Bowers, 2002; Hummel, 2000). On symbolic theories, language and perception depend on componential representations, with sentences composed of words (combined by syntax), words composed of letters (combined by rules of morphology and orthotactics), objects composed of object parts (geons), and so on. To the extent that componential representations are a property of the *what* ventral visual system, and not the *where* dorsal system, local representations may be more characteristic of the ventral system. Whether or not these specific arguments hold up, the critical point is that different brain systems engage in different computational tasks, and accordingly, there is no need to assume that all system employ the same coding schemes.

In sum, a grandmother cell theory is committed to the claim that the visual representations of faces, objects and words are coded in a localist format. On a strong version of the hypothesis, the representations at all levels of the visual hierarchy involved in identifying objects are localist, but in principle, earlier visual stages could be distributed (e.g., simple cells; Thorpe, 1989). Similarly, representations in other brain systems, including visual systems not involved in identifying objects (the dorsal visual system) may include some form of distributed coding. In fact, the different computational requirements of different brain systems may favor different representational schemes. Still, it is striking how many systems do code information in a highly sparse manner, from looming detection in locusts, to song production in the zebra finch, to face detection in humans (as reviewed in Part 2). It is obviously critical to characterize the coding schemes in all variety of systems, but in order to falsify a grandmother cell coding theory, it needs to be shown that the visual representations for words, objects, and faces are coded in a distributed format.

Some Objections to Grandmother Coding Schemes

Despite the observation that local and coarse coding are both consistent with neurophysiology, the general consensus in neuroscience and cognitive psychology is that grandmother cells cannot be right. Below, I consider some of the general objections to grandmother cells that motivate this conclusion.

Standard objections that can be quickly discarded. Grandmother coding schemes are often dismissed on the basis of one or more of the following standard claims: They do not degrade gracefully; there are not enough neurons to code all that is known, and they are not efficient. But as already shown, these objections are without merit. Concerns about graceful degradation are easily addressed by devoting multiple grandmother cells to a given stimulus. In this way, a lesion to one neuron does not result in the catastrophic loss of the corresponding concept (e.g., Barlow, 1995; Feldman, 1988; Gross, 2002). Concerns about running out of neurons reflects a basic misunderstanding; advocates of localist models never claimed (nor assumed) that single neurons code for propositions (e.g., “have a nice day”) or complex scenes (a weeping grandmother). Rather, the hypothesis is that individual neurons are devoted to individual objects, faces, words, and so on. Given estimates of 10^{10} neurons in our brain (e.g., Abeles, 1991), there seems little danger of running out of neurons on this hypothesis (even granting that only a small fraction of the 10^{10} neurons are located in the cortical areas responsible for object, face, and word identification). Concerns with regard to efficiency have it exactly backward. Our capacity to express an infinite number of thoughts may be the product of combining localist representations in generative ways. Not only is this efficient in computational terms, it is the most efficient coding scheme from a biological perspective: It is metabolically expensive to fire neurons. Indeed, it is argued that metabolic efficiency constraints alone rule out dense distributed coding schemes (Lennie, 2003).

Single neurons are too noisy and unreliable to support high-level perception. Another common criticism of localist coding schemes is that single neurons are noisy and unreliable, with the same input stimulus evoking a variable number of action potentials on different trials. The implication that is often drawn is that some form of population code is required in order to average out the noise (e.g., Averbek et al., 2006; Jazayeri & Movshon, 2006). But there are both empirical and theoretical problems with this conclusion. First, human performance is also variable, with inconsistent performance in detecting a stimuli presented at threshold and variable latencies to respond to salient stimuli. Furthermore, the reliability of neurons can be surprisingly high. For example, De Weese, Wehr, and Zador (2003) assessed the trial-by-trial response variability of auditory neurons in the cortex of rats in response to tones. They used a cell-attached recording procedure to ensure they were recording from a single neuron. Remarkably, reliability was almost perfect. Although similar reliability has not been observed in visual cortex, demonstrations that the psychophysics of single neurons can sometimes match the behavior of organisms rely on single neurons coping with noise.

But more important, even if it is granted that individual neurons are not sufficiently reliable to code for high-level perceptual tasks, it does not follow that some form of population code is required. Instead, all that is required is (again) redundant grandmother cells that code for the same stimulus. If one neuron fails to respond to the stimulus on a given trial due to noise, another one (or many) equivalent ones will, in what Barlow (1995) called “probability summation.” For example, as discussed above, Georgopoulos et al. (1986) reported evidence that neurons in primary motor cortex of rhesus monkeys encode arm movements through coarse coding. Their key argument was that the responses of single neurons are insufficient to account for precise motor control, given the noise in the system. Thus, in order to code for arm movements reliably and accurately, Georgopoulos et al. proposed a population vector model, where each neuron votes for their preferred direction, weighted by its firing rate. In this way, noise in the system can be averaged out. However, a localist coding scheme might equally account for these data. That is, the noise in the system can also be removed by considering the activation of multiple (redundant) motor neurons all tuned to the same direction. Although any given motor neuron tuned to the correct direction might not respond on a given trial (due to noise), redundant coding would ensure that multiple neurons coding the correct direction were activated. On a redundant grandmother coding scheme, each neuron codes for one thing, and noise is not a problem.

Grandmother cells cannot support the exquisite detail of our perceptual experiences. Localist coding schemes are typically associated with abstract representations, and they provide a natural format for supporting symbolic theories (e.g., Bowers, 2002; Fodor & Pylyshyn, 1988; Holyoak & Hummel, 2000; Pinker & Prince, 1988). For example, in a symbolic model of visual word identification, letter and word representations are coded independently of letter case (e.g., *A/a*, *READ/read*) and independently of their context (e.g., the same *a* detector is activated by *cat* and *act*; the same *cow* detector is activated by both *cow* and *brown cow*; Davis, 1999). These abstract letter representations are thought to support more widespread generalization than do models that include context dependent letter representations (e.g., models that code for the *a* in *cat* and *act* as different distributed patterns of activation; e.g., Seidenberg & McClelland, 1989).

The claim that the brain relies on local and abstract representations might appear to be challenged by the sheer vividness of our visual experience. That is, these representations might not have the fidelity to discriminate among all the relevant states of our consciousness. For example, according to Edelman (2002), local representations may well serve language but not vision, as he calls images ineffable. That is, verbal descriptions will generally fail to convey all the information in an image.

However, not all localist representations are this abstract. For example, on view-based theories of face and object identification, one local representation might represent a given individual from a given orientation, and another would be devoted to coding the same individual from another orientation. A view independent representation of the face is the product of pooling a collection of view specific representations onto a common face unit at the next stage of a visual hierarchy (e.g., Jiang et al., 2006). A similar pooling process is well documented at the earliest stages of vision, where a collection of simple cells that code for a given line orientation in different spatial locations all converge onto a complex cell that codes for the same line orientation across a range of locations. Riesenhuber and Poggio (1999) suggest a MAX computation, in which pooling is achieved by activating the unit in Layer *n*, based on the single most active unit Layer $n - 1$ (e.g., a given complex cell is activated as a function of the most active simple cell). In principle, this or a similar pooling process may occur across all levels of a visual hierarchy in which all units are localist.

Critical for present purposes, the inclusion of a range of abstract and specific local representations across a hierarchy of visual processing steps may help explain our almost limitless capacity to see variation in the visual world. For example, in the adaptive resonance theory, the identification of an object (either at a basic or a subordinate level) reflects the activation of a single localist unit at Layer *n*, and active units in Layer $n - 1$ (and perhaps at Layers $n - 2$, $n - 3$, etc.) code for the visual features of the object. The features are bound to the object through a feedback loop, or resonance, in which active units at different levels reinforce one another. On this theory, consciousness is a by-product of a resonance across multiple levels of a visual hierarchy, and our ability to categorize objects is mediated by active local units at Layer *n*, and our ability to perceive subtle variations in the world is mediated by active local units at lower layers of the network (Grossberg, 2003).

Consistent with this analysis, neurons in inferotemporal cortex respond to faces or objects at various degrees of perceptual specificity. At one extreme, Perrett et al. (1991) identified a cell that responds to a given familiar person from a wide range of viewpoints. Similarly, there are reports of single cells responding to the identity of a face independent of emotional expression (e.g., Gothard, Battaglia, Erickson, Spitler, & Amaral, 2007; Perrett, Smith, Potter, Mistlin, Head, Milner, & Jeeves, 1984), and following dramatic manipulations in lightness, such as contrast reversal (e.g., Sheinberg & Logothetis, 2002). Similarly, as noted above, there are reports of single cells responding to a specific three-dimensional object regardless of its orientation (Logothetis and Pauls, 1995; again, see Figure 8).

In other cases, a given cell only responds to a combination of visual features; for example, the identity and the emotion of a face (Gothard et al., 1984), or the identity and orientation of a given face (e.g., Perrett et al., 1991). In the same way, single neurons sometimes respond to specific 3D objects from a restricted range of orientations (Logothetis & Pauls, 1995). Indeed, on the basis of a greater number of orientation-dependent, compared to independent, object and face representations, Perrett et al. (1991) and Logothetis and Pauls (1995) argued for a view-dependent theories of vision, in which the outputs of many view-specific neurons are combined to support view-independent (object-centered) recognition.

In sum, local representations can code for information at various levels of abstraction, from highly specific perceptual information (e.g., a line of a specific orientation projected on a given location of the retina, or a specific face from a given orientation), to a highly abstract category (e.g., an abstract word embedded in any position within a sentence). This range of representations provides the basis by which detailed visual information can be discriminated.

Grandmother cells cannot generalize. The claim that local coding schemes cannot discriminate between all the relevant states of the visual world is complemented with the opposite criticism, namely, that local coding schemes cannot generalize (e.g., Anderson, 1995; Arbib, 2002; Földiák, 2002; Poggio & Bizzi, 2004; Rolls et al., 1997). But the claim that grandmother cells cannot generalize is false; indeed, as discussed above, localist representations are often included in symbolic systems, and symbolic systems are designed to support widespread generalization (e.g., Hummel & Holyoak, 1997).

The reason that authors often claim that grandmother cell theories cannot generalize is that authors assume a specific version of localist coding; namely, that each unit or neuron should have such a precise tuning curve that it responds only to one item and not at all to related things (e.g., Földiák, 2002). For example, as demonstrated by Jiang et al. (2006), a localist model of face recognition that includes this version of grandmother cell coding does perform poorly in generalizing, as reflected in poor judgments to unfamiliar faces. To make the model succeed with novel faces, they included a set of 180 face units, each tuned to a different (unique) face, but it is critical to note that they designed the network so that a specific input face activated a small number of visually similar face units as well (less than 10). Under these conditions, judgments about unfamiliar faces could be based on the pattern of activation across the familiar face units. Jiang et al. consider this latter version of the model a distributed, rather than a localist (grandmother), theory.

However, there are a number of problems with this conclusion. First, the fact that a model was designed to activate a small number of familiar face units in response to a single input does not make the model distributed. As noted in Part 1, most (if not all) localist models in psychology are designed such that a given input coactivates a collection of localist units. Indeed, as discussed above, the coactive localist units play a role in processing familiar and novel items. Second, when the Jiang et al. (2006) model identifies a familiar face, the other active face units play no role in coding for the face. That is, the only relevant unit for coding a familiar face is the most active one; indeed, this unit was hand-wired to code for this familiar face. As noted in Part 1, the grandmother cell theory is a theory about how familiar items are coded (no one ever claimed there were individual neurons for unfamiliar words, novel faces, or novel thoughts). A localist theory must support generalization to novel forms, but this is achievable by coactivating localist representations of familiar items, as this model nicely demonstrates.

Third, it is important to emphasize that I am not simply making a terminological point, extending the definition of grandmother cell theories to the present case. One of the exciting findings that initially inspired interest (and ridicule) of the grandmother cell hypothesis was the identification of simple and complex cells in V1 that preferentially responded to lines with specific properties, and that were organized into a hierarchy (Hubel & Wiesel, 1962, 1968). The question was whether the hierarchy of selective cells extended beyond V1, and included single neurons that selectively responded to familiar faces (e.g., a grandmother). This is the hypothesis that has been soundly rejected by most theorists (including Hubel, 1995). But this is exactly what Jiang et al. (2006) have implemented. Indeed, Riesenhuber (2005) endorsed what he calls the "standard model" that is an explicit extension of the original Hubel and Wiesel model. That is, Riesenhuber and colleagues (e.g., Jiang et al. 2006; Poggio & Bizzi, 2004) assumed that the ventral visual system is hierarchically organized, with simple and complex cells in the first stages of visual processing, and cells at the final stage (in anterior inferotemporal cortex) tuned to specific familiar faces.

Another point to make about generalization is that these authors are concerned with one specific type of generalization; namely, interpolation, in which a new stimulus is some blend of preexisting familiar representations (localist or otherwise). But blending systems only support interpolation, and other forms of generalization are required as well; that is, when generalizing from novel inputs that are highly dissimilar from past trained items. For instance, speakers can produce the past tense of an unfamiliar verb stem even when the stem is dissimilar to all stored memories (e.g., Prasada & Pinker, 1993). Similarly, babies can learn a new grammar based on one set of words that generalizes to a completely dissimilar set of words (Marcus et al., 1999). Hummel and Holyoak (2003) called this relational generalization, and it is often claimed to require discrete combinatorial systems; that is a system that includes a set of local representations and a syntax for combining these representations in generative ways. Of course there is an active debate as to whether discrete combinatorial systems are necessary to explain human language (and generalization more broadly; e.g., see McClelland & Patterson, 2002; McClelland & Plaut, 1999), but the relevant point for present purposes is that localist systems manifestly do support widespread generalization, contrary to the common claim.

If grandmother cells existed, they could not be found. Critics of grandmother cells often argue that the neurons should be impossible to find in a brain composed of 10^{10} neurons. On this line of reasoning, the frequent identification of neurons that respond to specific stimuli actually challenges the grandmother cell theory. If the identification of neurons that act like grandmother cells actually provided evidence against the hypothesis, then the theory is in trouble.

This argument is commonly advanced. For example, as discussed above, Quiroga et al. (2005) identified neurons in the hippocampus of humans that responded robustly to different photographs of a given person but not to photographs of other (sometimes similar looking) persons. Despite these findings, Quiroga et al. (2005) explicitly rejected the grandmother cell hypothesis and concluded that each cell must represent more than one class of image (otherwise the cells would not be found in the first place).

More recently, Waydo, Kraskov, Quiroga, Fried, and Koch (2006) reanalyzed the Quiroga et al. (2005) data, in an attempt to provide a more precise measure of the number of different things a given neuron codes for. They arrived at this estimate though a Bayesian analysis in which the number of images presented to the participants and the number of neurons recorded from were considered. They concluded "that highly sparse (although not grandmother) coding is present in this brain region" (pp. 10233–10234). Given the number of neurons in the relevant area of the brain, they further speculated that each neuron fires in response to 50–150 different basic level images. Applied to the case of faces, the prediction is that although the Jennifer Aniston cell only responded to images of Jennifer Anniston in the study, there were likely 50–150 other faces that it codes for and responds to (also see Quiroga, Kreiman, Koch, & Fried, 2008; for a similar logic but somewhat different analysis see Valiant, 2006).

But the argument is flawed. It is true that many neurons must respond to a given stimulus; otherwise the neurons would never be found. But multiple neurons could all respond to the same image. Waydo et al.'s (2006) calculations are just as consistent with the conclusion that there are roughly 50–150 redundant Jennifer Aniston cells in the hippocampus. This same point was made earlier by Perrett et al. (1989), who noted that there must be massive redundancy of highly selective cells, given that so many are found with only a limited opportunity to sample cells. Consistent with this general analysis, highly specific neural responses are generally observed for highly trained and well-recognized stimuli (stimuli that are most likely to be redundantly coded).

Indeed, some existing models predict redundant localist coding. For instance, in the adaptive resonance theory of Grossberg (1980), localist units in Layer *n* of the network code for words, objects, and the like, and localist units in Layer $n - 1$ code for features of objects. Learning in the model consists in modifying the connections between multiple (localist) units in Layer $n - 1$ and a single unit in Layer *n*. It is critical to note that to overcome catastrophic interference (what Grossberg, 1976, previously called the stability–plasticity dilemma), the model includes a vigilance parameter that prevents new learning from modifying old knowledge whenever the pattern of activation across Layer $n - 1$ does not match a unit in Layer n sufficiently well. In this situation, the model forms a new localist representation of the input in Layer *n* (rather than modifying preexisting representations). In this way, multiple localist representations develop that all code for similar inputs. If the vigilance parameter is extremely high, the model is effectively an instance theory, with a new dedicated localist representation encoded each time a new stimulus is encoded. The details are beyond this article, but the important point to note is that some degree of redundancy of grandmother cells is predicted by this theory. This redundancy was not proposed post hoc to account for the paradox of identifying grandmother cells, but as a solution to the stability–plasticity dilemma. It is interesting to note that massive replication of grandmother cells may help explain why identification improves with practice according to a power-law function (Page, 2000).

However there is another and more fundamental flaw in Waydo et al.'s (2006) reasoning. Their calculations follow from the assumption that a grandmother cell should only respond to one face or object. However, as discussed above, that is not how localist models work. Consider again the IA model of word identification. If the input *dog* is presented to the model, a subset of form similar words will be coactivated (*hog*, *fog, log*, etc.). If a single-unit recording study was carried out on the model and only a small fraction of the word units were sampled, then there would be a greater probability of recording from one of the form similar word units than the target unit itself (there are more of the former). That is, the experimenter is more likely to find a unit that selectively responds to *dog* when recording from a *fog*, *hog*, or *log* unit. Although a critic of grandmother cell coding schemes might point out that the unit that responds to *dog* might in fact respond to another stimulus better (and be correct), this would not provide evidence against a grandmother cell coding scheme. Indeed, that is exactly what the IA model would predict.

The key point for present purposes is that this undermines Waydo et al.'s (2006) conclusion. It may well be necessary to assume that each neuron responds to multiple different (but similar) things, given the sheer improbability of finding a neuron that responds to one and only one person or object (even allowing for massive redundancy). However, given that this is a property of both localist and sparse coding schemes, their rejection of grandmother cells on the basis of observing grandmother-like neurons is unwarranted. To distinguish between coarse and localist representations, one needs to determined whether neurons that respond to different things are involved in representing multiple things (coarse coding) or whether these neuron code for one thing and are just incidentally activated by form similar things (as when the *hog* unit fires in response to *dog* in the IA model). This is admittedly difficult to determine, but short of this, both hypotheses remain equally viable.

In sum, the multiple reports of grandmother-like neurons may reflect the fact that brains include redundant grandmother cells (perhaps massive redundancy), and these cells become partially active in response to form similar inputs. Of course, the data are also consistent with coarse coding schemes, as argued by Quiroga et al. (2005) and Waydo et al. (2006). However, there is no justification for rejecting grandmother cells on the basis of their findings and analyses.

The current version of grandmother cells is unfalsifiable. A final objection I would like to consider is whether I have rendered the grandmother cell hypothesis unfalsifiable by developing implausible and post hoc assumptions to reconcile any inconvenient data with this hypothesis. I consider two possible examples of this in turn.

First, when researchers fail to identify single neurons that account for behavior as well as a population of recorded neurons (e.g., Rolls et al., 1997), I suggest that the relevant grandmother neuron(s) were missed in the study rather than take the data as evidence for distributed coding. This would indeed render the hypothesis unfalsifiable if it were the case that experimenters always failed to report highly selective neurons. But what is to be made of studies that have identified highly selective neurons? When considering the relative plausibility of coarse coding and grandmother cells, the relevant question should be the following: What pattern of neural responding (highly selective versus less selective) poses a stronger constraint on theories of neural coding? Given that it is presumably easy not to find a grandmother cell, something analogous to a needle in a haystack, I would argue that the reports of highly selective responding should be given more weight.

Second, when researchers identify neurons that selectively respond to a given stimulus, I assume that the neuron codes for this stimulus, as opposed to the countless other stimuli that went untested in the experiment. A critic might find this interpretation of the data unmotivated (and perhaps even implausible). Nevertheless, there are good reasons to take this hypothesis seriously. As a general point, a fundamental feature of coarse coding is that a given neuron codes for a set of similar things. Thus, a neuron that responds to one stimulus in a highly selective way is unlikely to respond to a wide range of dissimilar stimuli (this should only occur in a dense distributed coding scheme). That is, the only untested stimuli likely to drive a selective cell will be similar to the identified stimulus. Given this, it is interesting to note that selective responding has been reported even in the context of highly similar foils (e.g., see Figure 8).

Sakai et al. (1994) attempted to directly address the concern that the selective responding of neurons in an experiment is in fact illusory and that these neurons would inevitably respond to a range of different but untested objects or faces. Sakai et al. (1994) trained monkeys to recognize 12 pairs of computer-generated Fourier patterns and then tested neurons that showed selectivity to these pairs. Sakai et al. identified neurons in IT that responded strongly to one or the other of the newly learned patterns in an associated pair but weakly to the other 22 patterns, despite the high similarity among the patterns. Sakai et al. then compared the responses of neurons that were sensitive to trained Fourier patterns when the neuron was presented with the trained stimulus and a variety of highly similar patterns (by manipulating the parameter set that generated the trained visual patterns in the first place). In the majority of cases, the neurons responded more strongly to the trained pattern compared to the transformed ones, and in no case did the neuron respond more strongly to the transformed pattern. This suggests that the neurons were tuned to the trained visual patterns; that is, the cells were grandmother cells for these patterns. Similar results and conclusions were reached by Logothetis and Pauls (1995).

There is other evidence suggesting that experimenters have, on occasion, identified the object or face that best drives a given neuron. For instance, as described above, Quiroga et al. (2005) reported a neuron that selectively responded to photographs of Halle Barry as well as the name Halle Barry. If the face was coding for someone else (or a collection of other people) and was just incidentally firing because the photographs of Halle Barry shared some feature with the untested target person then it seems unlikely that the neuron would also respond to the name Halle Barry. The same argument applies to an early study by Thorpe, Rolls, and Maddison (1983), who recorded from neurons in the orbitofrontal cortex of alert rhesus monkeys. They reported several neurons that were highly selective to visual stimuli. Out of 494 neurons analyzed, 26 were selective to foodstuff, and 11 of these were selective to one type of food: Four neurons selectively responded to oranges, 4 responded to peanuts, 2 responded to banana, and 1 responded raisins. Critical for present purposes, 5 of these 11 neurons also selectively responded to the taste of the corresponding food. For example, 1 neuron responded to the visual form as well as to the taste of bananas. Again, the cross modal abstraction suggests that the neuron is coding for a given type of food and is not just incidentally activated by something else.

Of course, the above points do not rule out the possibility that neurons are often (or always) mischaracterized, such that a neuron that responds to one object out of many within the context of a study would in fact respond better to another (untested) object as well. But as discussed above, this possibility can be reconciled with both sparse and grandmother coding schemes. Although a critic might conclude that it is difficult to falsify a grandmother cell coding scheme, the criticism applies equally well to sparse coding.

If anything, the unfalsifiable objection applies more strongly to distributed theories. For example, as noted above, Quiroga et al. (2005) rejected grandmother cells after observing cells that respond to one person, object, or face out of many. In the same way, Churchland and Sejnowski (1992) considered the scenario in which a single cell in temporal cortex responds to the face of Grandma Edna among a large set of faces tested. They ask whether this would constitute evidence for a local, as opposed to distributed, coding scheme. They suggest not, noting that there might be some other untested object or person that would also drive the cell to some degree, and even the partial activation of the cell may reflect a functional role in representing another image. That is, it appears that Churchland and Sejnowski would reject grandmother cells even if presented with evidence that a single neuron does in fact respond reliability and selectively to one image out of thousands of related ones. It is hoped that some of the analysis presented above would blunt this skepticism, but if not, it is not clear what sort of evidence is required.

Given the current neuroscience, the only reasonable conclusion is that neither coarse coding nor grandmother cell theories have been falsified and that both provide a viable account of the data. Still, there are fundamental differences between these theories, and accordingly, there is no reason in principle that the theories cannot be distinguished on empirical (or perhaps computational) grounds. But clearly, future work is needed.

Conclusions

It is widely assumed that localist models with grandmother units are biologically implausible. Indeed, this assumption contributes to the widespread popularity of PDP models within psychology. But as reviewed above, neurophysiological recordings from single neurons straightforwardly falsify the distributed coding schemes often learned with PDP models. Furthermore, despite widespread dismissal of grandmother cells in neuroscience, the data are entirely consistent with the hypothesis. The disconnect between data and theory in neuroscience is due, at least in part, to a failure to appreciate how localist (grandmother) models work.

I do not mean to suggest that the evidence provides unambiguous support for grandmother coding schemes, and some versions of distributed coding (that is, coarse coding) might well prove correct. Future empirical and computational work is required to distinguish between these theories. But researchers in both cognitive psychology and neuroscience should think twice before dismissing localist coding schemes. There should be nothing pejorative about grandmothers.

One thing is clear.

> Over the past 50 years, there has been an astonishing change in how we regard cells in the CNS, and especially, in the cortex. At the beginning of this period, it was believed that there was such an incredibly large number of such cells (10^5/mm^3 of cortex, and more than 10^{10} altogether) that it would be absurd and meaningless to consider the role of a single one, and therefore averaging the activity of large numbers of them was the only sensible approach. Now it is possible to record from a singe neuron in the cortex of an awake, behaving monkey, determine how well it performs in its task of pattern recognition, and compare this performance to that revealed by the behavioral responses of the same animal. The fact that thresholds are comparable (Britten et al., 1992) would have astounded the cortical neurophysiologist of 50 years ago. (Barlow, 1995, p. 417)

I expect it will astound most cognitive psychologists today.

References

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex.* Cambridge, England: Cambridge University Press.

Anderson, J. A. (1995). *An introduction to neural networks.* Cambridge, MA: MIT Press.

Arbib, M. A. (2002). *The handbook of brain theory and neural networks.* Cambridge, MA: MIT Press.

Averbek, B. B., Lathan, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience, 7,* 358–366.

Barlow, H. (1972). Single units and sensation: A neuron doctrine for perceptual psychology. *Perception, 1,* 371–394.

Barlow, H. B. (1985). The 12th Bartlett Memorial Lecture: The role of single neurons in the psychology of perception. *Quarterly Journal of Experimental Psychology: Section A. Human Experimental Psychology, 37,* 121–145.

Barlow, H. B. (1995). The neuron doctrine in perception. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 415–436). Cambridge, MA: MIT Press.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L. H. (1990). Comparison of spatial and temporal characteristics of neuronal-activity in sequential stages of hippocampal processing. *Progress in Brain Research, 83,* 287–300.

Berkeley, I. S. N. (2000). What the #$*%! is a subsymbol? *Minds and Machines, 10,* 1–13.

Berkeley, I. S. N. (2006). Moving the goal posts: A reply to Dawson and Piercey. *Minds and Machines, 16,* 471–478.

Berkeley, I. S. N., Dawson, M. R. W., Medler, D. A., Schopflocher, D. P., & Hornsby, L. (1995). Density plots of hidden unit activations reveal interpretable bands. *Connection Science, 7,* 167–186.

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

Boss, B. D., Turlejski, K., Stanfield, B. B., & Cowan, W. M. (1987). On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats. *Brain Research, 406,* 280–287.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113,* 201–233.

Bowers, J. S. (2002). Challenging the widespread assumption that connectionism and distributed representations go hand-in-hand. *Cognitive Psychology, 45,* 413–445.

Bowers, J. S., Damian, M. F., & Davis, C. J. (2008). *A computational reason to develop sparse (perhaps local) representations in PDP networks: The superposition constraint.* Unpublished manuscript.

Bowers, J. S., Davis, C. J., & Hanley, D. (2005). Interfering neighbors: The impact of novel word learning on the identification of visually similar words. *Cognition, 97,* B45–B54.

Bradley, A., Skottun, B. C., Ohzawa, I., Sclar, G., & Freeman, R. D. (1987). Visual orientation and spatial-frequency discrimination: A comparison of single neurons and behavior. *Journal of Neurophysiology, 57,* 755–772.

Brecht, M., Schneider, M., Sakmann, B., & Margrie, T. W. (2004, February 19). Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature, 427,* 704–710.

Brincat, S. L. & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience, 7,* 880–886.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Visual Neuroscience, 13,* 87–100.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (1992). The analysis of visual motion: a comparison of neuronal and psychophysical performance. *Journal of Neuroscience, 12,* 4745–4765.

Broadbent, D. (1985). A question of levels—Comment. *Journal of Experimental Psychology: General, 114,* 189–192.

Celebrini, S., & Newsome, W. T. (1994). Neuronal and psychophysical sensitivity to motion signals in extrastriate area MST of the macaque monkey. *Journal of Neuroscience, 14,* 4109–4124.

Churchland, P., & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Coltheart, M. (2004). Are there lexicons? *Quarterly Journal of Experimental Psychology: Section A. Human Experimental Psychology, 57,* 1153–1171.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100,* 589–608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review, 108,* 204–256.

Connor, C. E. (2005, June 23). Friends and grandmothers. *Nature, 435,* 1036–1037.

Davis, C. J. (1999). The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition (Doctoral dissertation, University of New South Wales, Sydney, New South Wales, Australia, 1999). *Dissertation Abstracts International, 62,* 594.

Dawson, M. R. W., Boechler, P. M., & Valsangkar-Smyth, M. (2000). Representing space in a PDP network: Coarse allocentric coding can mediate metric and nonmetric spatial judgements. *Spatial Cognition and Computation, 2,* 181–218.

Dawson, M. R. W., & Piercey, C. D. (2001). On the subsymbolic nature of a PDP architecture that uses a nonmonotonic activation function. *Minds and Machines, 11,* 197–218.

deCharms, R. C. (1998). Information coding in the cortex by independent or coordinated populations. *Proceedings of the National Academy of Sciences of the United States of America, 95,* 15166–15168.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93,* 283–321.

DeWeese, M. R., Wehr, M., & Zador, A. M. (2003). Binary spiking in auditory cortex. *The Journal of Neuroscience, 23,* 7940–7949.

Edelman, H. (2002). Constraining the neural representation of the visual world. *Trends in Cognitive Sciences, 6,* 125–131.

Eichenbaum, H. (2001). Engram. In P. Winn (Ed.), *Dictionary of biological psychology* (p. 558). New York: Routledge.

Elliott, C. J. H., & Susswein, A. J. (2002). Comparative neuroethology of feeding control in molluscs. *Journal of Experimental Biology, 205,* 877–896.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (195–223). Cambridge, MA: MIT Press.

Erickson, C. A., Jagadeesh, B., & Desimone, R. (2000). Clustering of perirhinal neurons with similar properties following visual experience in adult monkeys. *Nature Neuroscience, 3,* 1066–1068.

Feldman, J. A. (1988). Connectionist representation of concepts. In D. Waltz and J. A. Feldman (Eds.), *Connectionist models and their implications* (pp. 341–363). New York: Ablex Publications.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties, *Cognitive Science, 6,* 205–254.

Finkel, L. H. (1988, April 28). Groups and grandmothers in neuroscience. *Nature, 332,* 787.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3–71.

Földiák, P. (2002). Sparse coding in the primate cortex. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., pp. 1064–1068). Cambridge, MA: MIT Press.

Formisano, E., Kim, D. S., Di Salle, F., van de Moortele, P. F., Ugurbil, K., & Goebel, R. (2003). Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron, 40,* 859–869.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001, January 12). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science, 291,* 312–316.

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

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2006). Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cerebral Cortex, 16,* 1631–1644.

Fried, I., Cameron, K. A., Yashar, S., Fong, R., & Morrow, J. W. (2002). Inhibitory and excitatory responses of single neurons in the human medial temporal lobe during recognition of faces and objects. *Cerebral Cortex, 12,* 575–584.

Fried, I., MacDonald, K. A., & Wilson, C. L. (1997). Single neuron activity in human hippocampus and amygdala during recognition of faces and objects. *Neuron, 18,* 753–765.

Fujita, I. (2002). The inferior temporal cortex: Architecture, computation, and representation. *Journal of Neurocytology, 31, 359*–371.

Fujita, I., Tanaka, K., Ito, M., & Cheng, K. (1992, November 26). Columns for visual features of objects in monkey inferotemporal cortex. *Nature, 360,* 343–346.

Gabbiani, F., Krapp, H. G., Koch, C., & Laurent, G. (2002, November 21). Multiplicative computation in a visual neuron sensitive to looming. *Nature, 420,* 320–324.

Gahtan, E., & Baier, H. (2004). Of lasers, mutants, and see-through brains: Functional neuroanatomy in zebrafish. *Journal of Neurobiology, 59,* 147–161.

Gardner-Medwin, A. R., & Barlow, H. B. (2001). The limits of counting accuracy in distributed neural representations. *Neural Computation, 13,* 477–504.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986, September 26). Neuronal population coding of movement direction. *Science, 233,* 1416–1419.

Gothard, M., Battaglia, F. P., Erickson, C. A., Spitler, K. M., & Amaral, D. G. (2007). Neural responses to facial expression and face identity in the monkey amygdala. *Journal of Neurophysiology, 97,* 1671–1683.

Groh, J. M., Born, R. T., & Newsome, W. T. (1997). How is a sensory map read out? Effects of microstimulation in visual area MT on saccades and smooth pursuit eye movements. *The Journal of Neuroscience, 17,* 4312–4330.

Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cerebral Cortex, 5,* 455–469.

Gross, C. G. (2002). The genealogy of the "grandmother cell." *The Neuroscientist, 8,* 512–518.

Gross, C. G., Bender, D. B., & Roch-Miranda, C. E. (1969, December 5). Visual receptive fields of neurons in inferotemporal cortex of monkey. *Science, 166,* 1303–1306.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding II: Feedback, expectation, olfaction, and illusions. *Biological Cybernetics, 23,* 187–203.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review, 87,* 1–51.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63.

Grossberg, S. (2000). The complementary brain: Unifying brain dynamics and modularity. *Trends in Cognitive Sciences, 4,* 233–246.

Grossberg, S. (2003). Bring ART into the ACT. *Behavioral and Brain Sciences, 26,* 610–611.

Hahnloser, R. H. R., Kozhevnikov, A. A., & Fee, M. S. (2002, September 5). An ultra-sparse code underlies the generation of neural sequences in a songbird. *Nature, 419,* 65–70.

Heiligenberg, W. (1990). Electrosensory systems in fish. *Synapse, 6,* 196–206.

Heisenberg, M. (2003). Mushroom body memoir: From maps to models. *Nature Reviews Neuroscience, 4,* 266–275.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Foundations* (Vol. 1, pp. 77–109). Cambridge, MA: MIT Press.

Holyoak, K. J., & Hummel, J. E. (2000). The proper treatment of symbols in a connectionist architecture. In E. Deitrich & A. Markman (Eds.), *Cognitive dynamics: Conceptual change in humans and machines* (pp. 229–263). Mahwah, NJ: Erlbaum.

Houweling, A. R., & Brecht, M. (2007, December 19). Behavioral report of single neuron stimulation in somatosensory cortex. *Nature, 451,* 65–68.

Hubel, D. (1995). *Eye, brain, and vision.* New York: Scientific American Library.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in cats visual cortex. *Journal of Physiology: London, 160,* 106–154.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology: London, 195,* 215–243.

Hummel, J. E. (2000). Localism as a first step toward symbolic representation. *Behavioral and Brain Sciences, 23,* 480–481.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review, 104,* 427–466.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264.

Humphreys, G. W., & Evett, L. J. (1985). Are there independent lexical and monlexical routes in word-processing: An evaluation of the dual-route theory of reading. *Behavioral and Brain Sciences, 8,* 689–705.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005, November 4). Fast readout of object identity from macaque inferior temporal cortex. *Science, 310,* 863–866.

Jacobs, A. M., Rey, A., Ziegler, J. C., & Grainger, J. (1998). MROM-P: An interactive-activation, multiple read-out model of orthographic and phonological processes in visual word recognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 147–188). Mahwah, NJ.: Erlbaum.

Jazayeri, M., & Movshon, J. A. (2006). Optimal representation of sensory information by neural populations. *Nature Neuroscience, 9,* 690–696.

Jeffress, L. A. (1948). A place theory of sound localization. *Journal of Comparative Psychology, 44,* 35–39.

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,* 159–172.

Keller, A., Zhuang, H., Chi, Q., Vosshall, L. B., & Matsunami, H. (2007, September 16). Genetic variation in a human odorant receptor alters odor perception. *Nature, 449,* 468–472.

Kello, C. T., Sibley, D. E., & Plaut, D. C. (2005). Dissociations in performance on novel versus irregular items: Single-route demonstrations with input gain in localist and distributed models. *Cognitive Science, 29,* 627–654.

Keysers, C., Xiao, D. K., Földiák, P., & Perrett, D. I. (2001). The speed of sight. *Journal of Cognitive Neuroscience, 13,* 90–101.

Knudsen, E., du Lac, S., & Esterly, S. (1987). Computational maps in the brain. *Annual Review of Neuroscience, 10,* 41–65.

Kobatake, E., Wang, G., & Tanaka, K. (1998). Effects of shape discrimination training on the selectivity of inferotemporal cells in adult monkeys. *Journal of Neurophysiology, 80,* 324–330.

Kohonen, T. (1982). Self-organization of topologically correct feature maps. *Biological Cybernetics, 43,* 59–69.

Konorsky, J. (1967). *Integrative activity of the brain: An interdisciplinary approach.* Chicago: University of Chicago Press.

Kreiman, G., Koch, C., & Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience, 3,* 946–953.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Kupfermann, I., & Weiss, K. R. (1978). Command neuron concept. *Behavioral and Brain Sciences, 1,* 3–10.

Lamotte, R. H., & Mountcastle, V. B. (1975). Capacities of humans and monkeys to discriminate between vibratory stimuli of different frequency and amplitude: Correlation between neural events and psychophysical measurements. *Journal of Neurophysiology, 38,* 539–559.

Lee, J. A., & Verleysen, M. (2002). Self-organizing maps with recursive neighborhood adaptation. *Neural Networks, 15*(8–9), 993–1003.

Lennie, P. (2003). The cost of cortical computation. *Current Biology, 13,* 493–497.

Lin, L. N., Chen, G. F., Kuang, H., Wang, D., & Tsien, J. Z. (2007). Neural encoding of the concept of nest in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America, 104,* 6066–6071.

Logothetis, N. K., & Pauls, J. (1995). Psychophysical and physiological evidence for viewer-centered object representations in the primate. *Cerebral Cortex, 5*(3), 270–288.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5,* 552–563.

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience, 19,* 577–621.

MacKay, D. G. (1987). *The organization of perception and action: A theory for language and other cognitive skills*. Berlin, Germany: Springer-Verlag.

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

Marr, D. (1982). *Vision.* San Francisco: Freeman.

McClelland, J. L. (2001). Cognitive neuroscience. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social & behavioral sciences* (pp. 2133–2139). Oxford, England: Pergamon, 2133–2139.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Patterson, K. (2002). "Words or rules" cannot exploit the regularity in exceptions: Reply to Pinker and Ullman. *Trends in Cognitive Sciences, 6,* 464–465.

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3,* 166–168.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: 1. An account of basic findings. *Psychological Review, 88,* 375–407.

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing: Psychological and biological models* (Vol. 2). Cambridge, MA: MIT Press.

Milner, A. D., & Goodale, M. A. (2006). *The visual brain in action* (2nd ed.). Oxford, England: Oxford University Press.

Morgan, M. J. (1989, September 7). Perceptual decision-making: Watching neurons discriminate. *Nature, 341,* 20–21.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review 76,* 165–178.

Morton, J. (1979). Facilitation in word recognition: Experiments causing change in the logogen model. In P. A. Kolers, M. E. Wrolstad, & H. Bouma (Eds.), *Processing models of visible language* (pp. 259–268). New York: Plenum.

Mountcastle, V. B., Carli, G., & Lamotte, R. H. (1972). Detection thresholds for stimuli in humans and monkeys: Comparison with threshold events in mechanoreceptive afferent nerve fibers innervating monkey hand. *Journal of Neurophysiology, 35,* 122–136.

Mountcastle, V. B., Steinmetz, M. A., & Romo, R. (1990). Frequency discrimination in the sense of flutter: Psychophysical measurements correlated with postcentral events in behaving monkeys. *Journal of Neuroscience, 10,* 3032–3044.

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989, September 7). Neuronal correlates of a perceptual decision. *Nature, 341,* 52–54.

Nicholas, M. J., & Newsome, W. T. (2002). Middle temporal visual area microstimulation influences veridical judgments of motion direction. *Journal of Neuroscience, 22,* 9530–9540.

Norris, D. (1994). SHORTLIST: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology, 14,* 481–487.

Oram, M. W. (1998). The "ideal homunculus": Decoding neural population signals (Vol. 21, pp. 259). *Trends in Neurosciences, 21,* 259–265.

Oram, M. W., & Perrett, D. I. (1992). Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology, 68,* 70–84.

O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences, 2,* 455–462.

O'Reilly, R. C., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences, 6,* 505–510.

Page, M. P. A. (2000). Connectionist modeling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–512.

Page, M. P. A., & Norris, D. (1998). The primacy model: A new model of immediate serial recall. *Psychological Review, 105,* 761–781.

Parker, A., & Hawken, M. (1985). Capabilities of monkey cortical-cells in spatial-resolution tasks. *Journal of the Optical Society of America: A. Optics Image Science and Vision, 2,* 1101–1114.

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience, 21,* 227–277.

Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., & Laurent, G. (2002, July 19). Oscillations and sparsening of odor representations in the mushroom body. *Science, 297,* 359–365.

Perrett, D. I., Harries, M. H., Bevan, R., Thomas, S., Benson, P. J., Mistlin, A. J., et al. (1989). Frameworks of analysis for the neural representation of animate objects and actions. *Journal of Experimental Biology, 146,* 87–113.

Perrett, D. I., Mistlin, A. J., & Chitty, A. J. (1987). Visual neurons responsive to faces. *Trends in Neurosciences, 10,* 358–364.

Perrett, D. I., Oram, M. W., Harries, M. H., Bevan, R., Hietanen, J. K., Benson, P. J., & Thomas, S. (1991). Viewer-centered and object-centered coding of heads in the macaque temporal cortex. *Experimental Brain Research, 86,* 159–173.

Perrett, D. I., Rolls, E. T., & Caan W. (1979). Temporal lobe cells of the monkey with visual responses selective for faces. *Neuroscience Letters* (Suppl. 3), S358.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1984). Neurons responsive to faces in the temporal cortex: Studies of functional organization, sensitivity to identity, and relation to perception. *Human Neurobiology, 3,* 197–208.

Perrett, D. I., Smith, P. A. J., Potter, D. D., Mistlin, A. J., Head, A. S., Milner, A. D., & Jeeves, M. A. (1985). Visual cells in the temporal cortex sensitive to face view and gaze direction. *Proceedings of the Royal Society of London, Series B: Biological Sciences, 223,* 293–317.

Pinker, S. (1998). The evolution of the human language faculty. In N. G. Jablonski & L. C. Aiello (Eds.), The origin and diversification of language (pp. 117–126). San Francisco: California Academy of Sciences.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed-processing model of language-acquisition. *Cognition, 28,* 73–193.

Plaut, D. C. (2000). Connectionist modeling. In A. E. Kazdin, (Ed.), *Encyclopedia of psychology.* Washington, DC: American Psychological Association.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case-study of connectionist neuropsychology. *Cognitive Neuropsychology, 10,* 377–500.

Poggio, T., & Bizzi, E. (2004, October 13). Generalization in vision and motor control. *Nature, 431,* 768–774.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience, 1,* 125–132.

Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *The Journal of Cognitive Neuroscience, 9,* 222–237.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes, 8,* 1–56.

Prince, S. J. D., Pointon, A. D., Cumming, B. G., & Parker, A. J. (2000). The precision of single neuron responses in cortical area V1 during stereoscopic depth judgments. *Journal of Neuroscience, 20,* 3387–3400.

Purushothaman, G., & Bradley, D. C. (2005). Neural population code for fine perceptual decisions in area MT. *Nature Neuroscience, 8,* 99–106.

Quirk, G. J., Muller, R., & Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rat's recent experience. *Journal of Neuroscience, 10,* 2008–2017.

Quiroga, Q. R., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not "grandmother-cell" coding in the medial temporal lobe. *Trends in Cognitive Sciences, 12,* 87–91.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June 23). Invariant visual representation by single neurons in the human brain. *Nature, 435,* 1102–1107.

Reddy, L., Quiroga, R. Q., Wilken, P., Koch, C., & Fried, I. (2006). A single-neuron correlate of change detection and change blindness in the human medial temporal lobe, *Current Biology, 16,* 2066–2072.

Riesenhuber, M. (2005). Object recognition in cortex: Neural mechanisms, and possible roles for attention. In L. Itti, G. Rees, & J. Tsotsos (Eds.), *Neurobiology of attention* (pp. 279–287). San Diego, CA: Elsevier.

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

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia, 45,* 124–143.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Experimental Brain Research, 114,* 149–162.

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60–94.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations.* MIT Press.

Rumelhart, D. E., & Zipser, D. (1985). Feature discovery by competitive learning. *Cognitive Science, 9,* 75–112.

Sakai, K., Naya, Y., & Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learning and Memory, 1,* 83–105.

Salzman, C. D., & Newsome, W. T. (1994, April 8). Neural mechanisms for forming a perceptual decision. *Science, 264,* 231–237.

Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4,* 228–235.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96,* 523–568.

Seidenberg, M. S., & Plaut, D. C. (2006). Progress in understanding word reading: Data fitting versus theory building. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 25–49). Hove, England: Psychology Press.

Shackleton, T. M., Skottun, B. C., Arnott, R. H., & Palmer, A. R. (2003). Interaural time difference discrimination thresholds for single neurons in the inferior colliculus of guinea pigs. *Journal of Neuroscience, 23,* 716–724.

Sheinberg, D. L., & Logothetis, N. L. (2002). Perceptual learning and the development of complex visual representation in temporal cortical neurons. In M. Fahle & T. Poggio (Eds.) *Perceptional learning* (pp. 95–124). Cambridge, MA: MIT Press.

Shoham, S., O'Connor, D. H., & Segev, R. (2006). How silent is the brain: Is there a "dark matter" problem in neuroscience? *Journal of Comparative Physiology: A. Neuroethology Sensory Neural and Behavioral Physiology, 192,* 777–784.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience, 18,* 555–586.

Smolensky, P. (1988). Putting together connectionism: Again. *Behavioral and Brain Sciences, 11,* 59–70.

Subramaniam, S., Biederman, I., & Madigan, S. A. (2000). Accurate identification but no priming and chance recognition memory for pictures in RSVP sequences. *Visual Cognition, 7,* 511–535.

Talbot, W. H., Darian-Smith, I., Kornhuber, H. H., & Mountcastle, V. B. (1968). The sense of fluttervibration: Comparison of the human capacity with response patters of mechanoreceptive afferents from monkey hand. *Journal of Neurophysiology, 31,* 301–334.

Tamura, H., & Tanaka, K. (2001). Visual response properties of cells in the ventral and dorsal parts of the macaque inferotemporal cortex. *Cerebral Cortex, 11,* 384–399.

Thorpe, S. (1989). Local vs. distributed coding. *Intelletica, 8,* 3–40.

Thorpe, S. (1995). Localized versus distributed representations. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks*. Cambridge, MA: MIT Press.

Thorpe, S. (2002). Localized versus distributed representations. In M. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd ed., PP. 643–645) . Cambridge MA: MIT Press.

Thorpe, S., Delorme, A., & Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural Networks, 14,* 715–725.

Thorpe S. &, Imbert, M. (1989). Biological constraints on connectionist modelling. In R. Pfeifer, F. Fogeiman-Soule, S. Steels, & Z. Schreter (Eds.), *Connectionism in perspective* (pp. 63–93). Amsterdam, Holland: Elsevier/North-Holland.

Thorpe S. J., Rolls E. T., & Maddison S. (1983). The orbitofrontal cortex: Neuronal activity in the behaving monkey. *Experimental Brain Research, 49,* 93–115.

Valiant, L. G. (2006). A quantitative theory of neural computation. *Biological Cybernetics, 95,* 205–211.

Vallbo, A. B., & Hagbarth, K. E. (1968). Activity from skin mechanoreceptors recorded percutaneously in awake human subjects. *Experimental Neurology, 21,* 270–289.

van Steveninck, R. D., & Bialek, W. (1995). Reliability and statistical efficiency of a blowfly movement-sensitive neuron. *Philosophical Transactions of the Royal Society of London: Series B. Biological Sciences, 348,* 321–340.

Vogels, R., & Orban, G. A. (1990). How well do response changes of striate neurons signal differences in orientation: A study in the discriminating monkey. *Journal of Neuroscience, 10,* 3543–3558.

Wang, L., Narayan, R., Grana, G., Shamir, M., & Sen, K. (2007). Cortical discrimination of complex natural stimuli: Can single neurons match behavior? *Journal of Neuroscience, 27,* 582–589.

Waydo, S., Kraskov, A., Quiroga, R. Q., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *Journal of Neuroscience, 26,* 10232–10234.

Williamson, J. R. (2001). Self-organization of topographic mixture networks using attentional feedback. *Neural Computation, 13,* 563–593.

Young, M. P., & Yamane, S. (1992, May 29). Sparse population coding of faces in the inferotemporal cortex. *Science, 256,* 1327–1331.

Young, M. P., & Yamane, S. (1993). An analysis at the population level of the processing of faces in the inferotemporal cortex. In T. Ono, L. R. Squire, M. E. Raichle, D. I. Perrett, & M. Fukuda (Eds.), *Brain mechanisms of perception and memory: From neuron to behavior* (pp. 47–70). Oxford, England: Oxford University Press.

Zhang, K. C., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology, 79,* 1017–1044.

Zipser, D., & Anderson, R. A. (1988, February 25). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature 331,* 679–684.

Received February 18, 2008
Revision received September 15, 2008
Accepted September 16, 2008 ■