Measuring Sparseness in the Brain: Comment on Bowers (2009)

Rodrigo Quian Quiroga
University of Leicester

Gabriel Kreiman
Children's Hospital Boston, Harvard Medical School,
and Harvard University

Bowers (2009) challenged the common view in favor of distributed representations in psychological modeling and the main arguments given against localist and grandmother cell coding schemes. He revisited the results of several single-cell studies, arguing that they do not support distributed representations. We praise the contribution of Bowers (2009) for joining evidence from psychological modeling and neurophysiological recordings, but we disagree with several of his claims. In this comment, we argue that distinctions between distributed, localist, and grandmother cell coding can be troublesome with real data. Moreover, these distinctions seem to be lying within the same continuum, and we argue that it may be sensible to characterize coding schemes with a sparseness measure. We further argue that there may not be a unique coding scheme implemented in all brain areas and for all possible functions. In particular, current evidence suggests that the brain may use distributed codes in primary sensory areas and sparser and invariant representations in higher areas.

Keywords: grandmother cell, visual perception, memory, sparseness, neural coding

Supplemental materials: http://dx.doi.org/10.1037/a0016917.supp

Understanding the principles of how the brain is capable of different functions arguably constitutes one of the greatest scientific challenges of our times. Such an enterprise requires a combined effort across diverse disciplines, such as neuroscience, biology, computer science, psychology, philosophy, and physics, to name only a few. Along these lines, the recent contribution of Bowers (2009) should be praised for its significant attempt to put together knowledge derived from neurophysiological recordings, computational models, and psychology. In this comment, we discuss a few ideas to clarify some of the neurophysiological concepts addressed in Bowers's (2009) review. In particular, we emphasize the technical difficulties in addressing questions about the coding of information by neurons based on single-cell recordings and discuss how these experimental constraints affect claims of distributed, sparse, and grandmother-cell representations.

One of the most striking facts in visual perception is how, in a fraction of a second, the brain can make sense of very rich sensory inputs and use this information to create complex behaviors. It is perhaps the easiness with which such functions are performed that may make people typically unaware of the exquisite machinery in the brain required for such computations. People may be amazed at realizing that they can solve a Rubik's cube or beat a master in a chess game but are hardly surprised when performing something as complex as recognizing a familiar face in a crowd. A key question to understand how the brain processes information is to determine how many neurons, in a given area, are involved in the representation of a visual percept (or a memory, a motor command, etc.) and what information each of these neurons encodes about the percept. On the one hand, the representation of a percept could be given by the activity of a large population of neurons. In this case, the percept emerges from the ensemble response and cannot be understood by inspecting the responses of individual neurons without considering the whole population. On the other hand, the percept might be represented by very few and more abstract cells, with each of these cells giving explicit information about the stimulus. In neuroscience, the first scenario is usually referred to as *distributed population coding* and the second one is usually referred to as *sparse coding*, its extreme case—of having one neuron coding for one percept—usually referred to as *grandmother cell representation* (but note that the term *grandmother cell* can also be taken as meaning many neurons encoding for one percept or just meaning an abstract representation). We anticipate that these definitions may be imprecise (as noted by Bowers, 2009) and that the same terms may be used with different meanings by different communities of researchers. For example, we already mentioned different uses of the term grandmother cell. Moreover, for Bowers (2009), sparse codes are a form of distributed codes and in neuroscience these two types of coding are taken as the opposite. To avoid confusion, in the following we refer to distributed and localist codes, following Bowers's (2009) notation. It is indeed the vagueness and different meaning of these definitions that give rise to some of the discussions in the field.

A central theme in the discussion by Bowers (2009) concerns the biological plausibility of localist models. To address this question, he referred to evidence from single-cell recordings, the gold

---

standard to elucidate the function of neural circuits. Such parallels between neurophysiology and psychological models could have a major impact in both fields. It is in this spirit that we aim to contribute to this discussion by adding to and commenting on Bowers's (2009) claims from the perspective of neurophysiologists trying to extract this type of information from single-cell recordings.

Defining Distributed and Local (or Sparse) Coding

Bowers (2009) defined distributed codes as a representation in which each unit is involved in coding more than one familiar "thing," and consequently, the identity of a stimulus cannot be determined from the activation of a single unit (Bowers, 2009; our emphasis on "thing"). Moreover, he distinguishes between *dense distributed representations*, that is, distributed coding schemes in which each neuron is involved in coding many different things, as commonly associated with parallel distributed processing (PDP) theories in psychological modeling (McClelland, Rumelhart, & Group, 1986; Rumelhart, McClelland, & Group, 1986)—for related ideas in theoretical neuroscience, see Hopfield (1982, 2007)—and *coarse coding schemes*, that is, distributed codes in which single neurons have broad tuning curves, such that a single neuron codes for a range of similar things. Although a broadly tuned neuron may respond most strongly to a preferred stimulus, noise would preclude identifying the stimulus precisely from the single-cell activity (Bowers, 2009). In contrast to distributed codes, according to Bowers (2009), a localist representation is characterized by neurons coding for one thing, for which it is possible to infer a stimulus from the activation of a single unit. In between localist and coarse coding schemes, Bowers (2009) introduced one more term, *sparse distributed coding,* but it seems that these definitions lay within the same continuum, and the distinction between a localist and a sparse distributed code is just given by the number of objects encoded by a neuron.

The above definitions seem at first plausible, but the distinction among them becomes fuzzy when considering neurophysiological recordings. We should first mention that the distinction between distributed and coarse coding appears to be in how similar the neuronal preferences are. Defining *similarity* in a rigorous way is already quite a complex challenge in itself. For example, we can loosely imagine that a front view of a face is similar to a profile view of the same face and is very different from a front view of a different face. However, such a statement is quite arbitrary: At the pixel level, the similarity between front views of two different faces is much larger than the similarity between a front view and a profile view of the same face. This is far from a trivial consideration: Achieving a good definition of what humans consider similar things constitutes a central challenge in computer vision, neuroscience, and psychology. A second problem with these definitions, and perhaps a more fundamental one, is given by the ambiguity of what is meant by *thing*. A *thing* could be a face, a car, or an animal but could also be a pixel, an oriented bar, or an abstract concept. How *thing* is defined may radically alter our conclusions regarding how distributed or local a neural coding is. For example, a neuron in V1 may have a local representation for oriented bars in their receptive fields but, at the same time, a distributed representation for faces. To address this problem, Bowers (2009) argued that we cannot think of a distributed representation of a complex familiar thing (e.g., a face) at a low level of the

system (e.g., the retina or V1) because the retina does not know that there is a face. This dichotomy is usually referred to as implicit versus explicit representation. The retina encodes information about the face in an implicit manner (it seems farfetched to argue that the retina does not encode the visual information at all!). In contrast, the representation of the face at the level of the temporal lobe becomes explicit, in the sense that single-cells can give us reliable information about the presence or absence of a face. To be more precise, an explicit representation can be defined as the one in which the information can be decoded by a single layer network (Koch, 2004).

Given the activity of a single V1 neuron, we can discriminate the presence or absence of an oriented bar within the receptive field well above chance, but we cannot tell whether a particular face is present because this information is not explicit at the level of V1. But even when considering only oriented bars, should an oriented bar at 49° constitute a different thing compared with an oriented bar at 50 degrees? How many degrees of separation is required before an oriented bar becomes a new thing? The continuum nature of orientation makes this distinction difficult. In higher visual cortex, it is also possible that there exists a similar continuum of features to which neurons respond, only that it is in general difficult to assess what those features are (Connor, Brincat, & Pasupathy, 2007; Tanaka, 1996). This distinction is even harder for areas such as the hippocampus, where a neuron could fire preferentially to the different views of the tower of Pisa and the Eiffel Tower, and another one could fire preferentially to different pictures of Jennifer Aniston and Lisa Kudrow (both actresses of the TV series *Friends*), see Figures S6 and S7 in Quian Quiroga, Reddy, Kreiman, Koch, and Fried, 2005. Clearly, these responses are related at some high level of abstraction, which seems plausible given the role of hippocampus in coding associations (Miyashita, 1988; Wirth et al., 2003). However, it is unclear how different these concepts are or whether they should be considered as the same thing (landmarks of Europe in the first case, and the two actresses of *Friends* in the second one).

Another problem with these definitions is that in real life, identifying the stimulus encoded by the neural activity involves setting a responsiveness criterion for defining what is a significant response and what is not, which of course depends on the particular criterion chosen. Alternatively, it is also possible to use decoding algorithms or information theory formalism to extract information about the stimulus from the neural responses (Abbott, 1994; Quian Quiroga & Panzeri, 2009; Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997). But this can be also problematic for the above definitions because, due to trial-to-trial variability, noise, lack of enough number of trials, and so on, decoders or information theory do not provide yes–no answers but do provide estimations of performance or amount of information.

To avoid defining what is a thing and whether two stimuli are similar, it seems to us preferable to simplify the nomenclature by describing a continuum with dense distributed representations at one end and localist representations at the other. Central to this discussion is to determine where neuronal representations reside within this continuum, something that can be quantified with a sparseness measure like the one to be discussed in the following sections. Then, a high degree of sparseness will imply a local coding and, conversely, a low degree of sparseness will be evidence for a distributed representation.

Neural Responses and Neural Codes

Bowers (2009) discussed the interactive activation (IA) model of visual word recognition (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1982) to distinguish between what a neuron responds to and what a neuron codes for. In his example (see Figure 4 in Bowers, 2009), a unit at the top level of the IA model responds to both *blur* and *blue* due to the similarity between these two stimuli: They share the first 3 letters and differ only in the last one. However, he argued that this particular unit only codes for *blur* by construction. According to Bowers (2009), in this case the neuron shows a localist code because, although it responded to two things, the neuron encoded the meaning of only one of them. He therefore claims that responses to multiple objects do not provide evidence of a distributed representation. But this argument has some problems. For example, suppose that the same network is used with a new set of words, containing *blue* but not *blur*. The neuron will consistently fire to *blue*, and in fact, from the firing of this neuron one may accurately predict the presence of this word. Should one then say that in spite of such an explicit representation the neuron does not code for *blue*, given that it was trained to code for a similar word in the first place? Unless we define coding from how a network is particularly trained rather than from the meaningful information that can be obtained from it, to us the answer is no. These distinctions become even more problematic with real neuronal activity because we do not have direct access to what a neuron codes for, but rather to what it responds to. In other words, if a neuron responds to more than one thing, how could one know which response is meaningful and which one is not? Moreover, if we extrapolate Bowers's (2009) argument based on the IA model, we could easily conclude that every single neuron in the brain is only coding for one thing: When the neuron responds to many things, it could simply be stated that the neuron surely prefers only one thing and that the neuron merely responds to the other things due to similarity. What would then constitute evidence for a distributed representation (but a neuron responding to multiple things)? In other words, how can one falsify a localist coding scheme if the evidence from neurons responding to multiple things is not accepted? In fact, it seems implausible to argue that there is no evidence for distributed coding because it is not known whether one should ignore most of the responses. For any definition of localist and distributed coding, it is important to specify what type of evidence would support or falsify each type of code. Below, we propose to characterize neuronal responses quantitatively by a single degree of sparseness. Although many of our caveats described above still remain under this simple approach, this quantitative definition allows us to provide support to verify or falsify both distributed and localist representations.

Measuring Sparseness

Given the problems highlighted in the previous sections, it seems preferable to refer to distributed and local (or sparse) responses, with the understanding that neuronal responses constitute a proxy for neuronal codes. In order to quantify the distinction between localist and distributed responses, we need to be able to measure the degree of sparseness of single-cell activations in a reliable way. Figures 1A and 1D show the responses of 2 single units simultaneously recorded from the same microwire, whose activity could be separated after spike sorting (Quian Quiroga, 2007; Quian Quiroga, Nadasdy, & Ben-Shaul, 2004). Both units are nearly silent during baseline (average < 0.01 spikes/sec) and fired with up to 40 spikes/sec to only a few of the 114 pictures shown in this recording session. The first unit responded to two basketball players, and the second one responded to two landmark buildings. Due to space constraints, only 10 responses are shown. There were no responses to the other pictures not shown.

From Figures 1A and 1D the high degree of selectivity of these neurons is clear, but how can we measure sparseness? There are two notions of sparseness in the literature: (a) *population sparseness* refers to the fraction of neurons of a population that respond in a given time window and (b) *lifetime sparseness* refers to the relative number of stimuli to which a neuron responds to (Olshausen & Field, 2004; Quian Quiroga, Reddy, Koch, & Fried, 2007; Willmore & Tolhurst, 2001). These two notions are related because one expects that if a cell fires to few stimuli then each stimulus will be encoded by a relatively small population of cells. However, it is in principle possible that most cells in a given population respond to one stimulus (or a few stimuli) or that a small subset of neurons is very promiscuous, responding exuberantly to many stimuli. Due to technical difficulties involved in recording simultaneously from large numbers of neurons, most studies usually assess lifetime sparseness, assuming that it will be similar to population sparseness. In this context, lifetime sparseness, also termed selectivity or specificity, means that a given cell responds only to a small subset of the presented stimuli. On the contrary, if a neuron responds to many stimuli, it is said to be broadly tuned, pointing toward a distributed representation.

The notion of sparseness—and any measure to quantify it—depends on the stimulus set. In particular, the units of Figure 1A and 1D have sparse responses because they were activated only by very few of the more than 100 pictures presented. However, it is conceivable that a lower degree of sparseness would have been obtained for the unit in Figure 1D if more views of landmarks (and in particular of the tower of Pisa) had been used. To give a more clear (and extreme) example, if one neuron responds to many different faces, as in the monkey inferior temporal cortex (IT; Gross, 2008; Gross, Rocha-Miranda, & Bender, 1972; Hung, Kreiman, Poggio, & DiCarlo, 2005), it would appear to respond in a highly sparse manner if the stimulus set contains only one face and a large number of other stimuli. This seemingly trivial point makes it difficult to compare the degree of sparseness in different areas because different stimulus sets are typically used.

The simplest measure of sparseness would be to report the relative number of stimuli eliciting significant responses in a neuron. However, this number depends on the criterion used for defining what is a significant response and what it is not. In particular, if a very strict threshold is used, then only the few largest responses will cross this threshold and, consequently, this neuron will appear to be sparse. To overcome this dependence, a novel sparseness index (S) was introduced (Quian Quiroga et al., 2007) by plotting the normalized number of responses as a function of a threshold (Figures 1C, 1F). One hundred threshold values between the minimum and the maximum responses were taken, and for each threshold value, the fraction of responses above the threshold was computed. The area under this curve (A) is close to zero for a sparse neuron and is close to .5 for a uniform distribution of responses (dotted line in Figures 1C and 1F). The sparseness index was defined as $S = 1 - 2A$. It is 0 for a uniform distribution
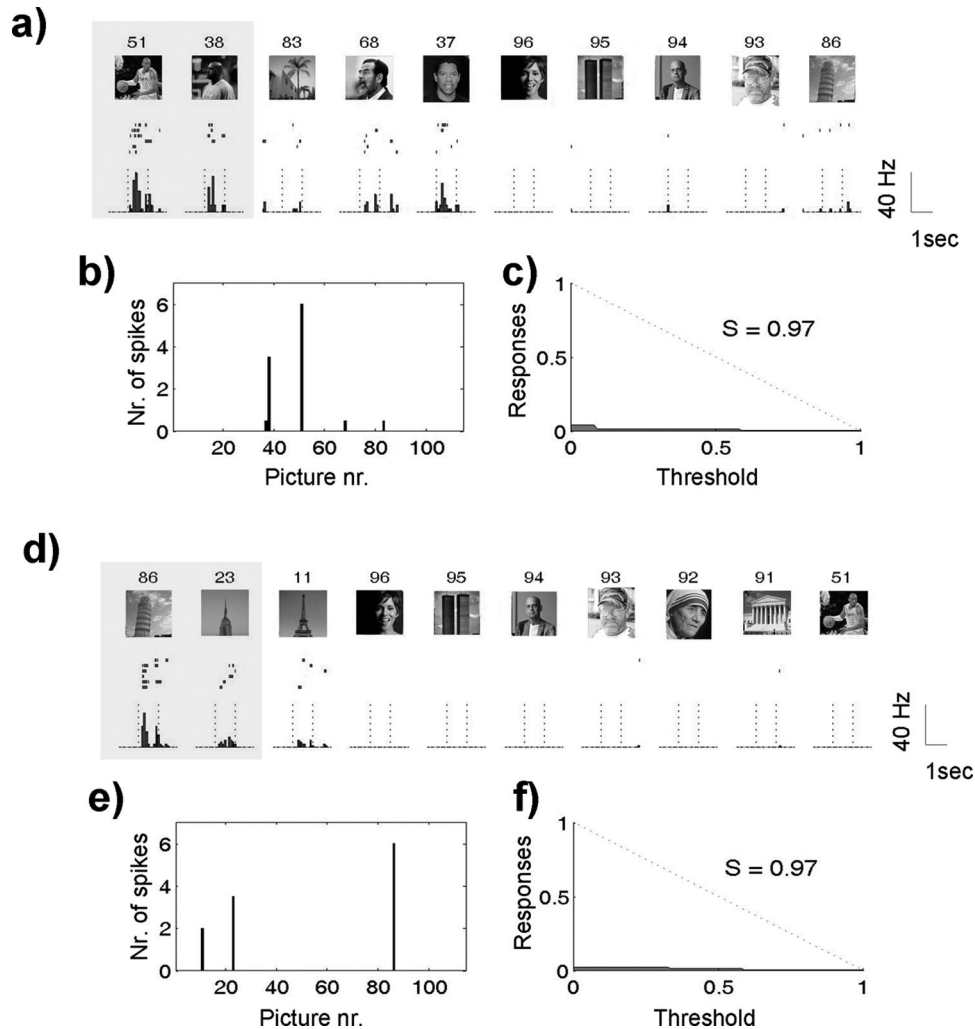
Figure 1. Adapted from "Decoding Visual Inputs From Multiple Neurons in the Human Temporal Lobe," by R. Quian Quiroga, L. Reddy, C. Koch, and I. Fried, 2007, Journal of Neurophysiology, 98, p. 2001. Copyright 2007 by The American Physiological Society. Used with permission. A, D: Ten largest responses of two simultaneously recorded single units in the right posterior hippocampus. There were no responses to the other 104 pictures shown to the patient. For each picture the corresponding raster plots (middle subplots; first trial on top) and poststimulus time histograms with 100 ms bin intervals (lower subplots) are given. Highlighted boxes mark significant responses. The vertical dashed lines indicate the times of image onset and offset, 1 s apart. Note the marked increase in firing rate of these units roughly 300 ms after presentation of the responsive pictures. B, E: Median number of responses (across trials) for all the pictures presented in the session. C, F: Relative number of responses as a function of the variable threshold (see text). Note the high selectivity values for both units (S = 0.97), thus implying a sparse representation. For copyright issues, the thumbnail pictures displayed in the figure corresponds to similar views of the same persons or objects used in the experiment. A color version of this figure is available on the Web at http://dx.doi.org/10.1037/a0016917.supp

(a dense representation), and it approaches 1 the sparser the neuron is (for a localist representation). S values in Figures 1C and 1F (.97 in both cases) confirm that the sparseness of these neurons is not just the consequence of the arbitrary choice of a very large threshold. Sparseness values for a large population of medial temporal lobe (MTL) neurons have been reported in Quian Quiroga et al. (2007).

Related to the discussion of how localist (sparse) or distributed is the representation of neurons in a given area, it should be noted that highly selective neurons, like the ones presented in Figures 1A and 1D, are hard to detect without optimal data processing and recording

conditions. This basically relies on (a) the recording of broad band continuous data allowing off-line analysis, (b) the use of an optimal spike detection and sorting algorithm, and (c) the use of semichronic multiple electrodes in contrast to traditional single electrode recordings. In fact, single electrode recordings are usually carried out with movable probes that tend to miss sparsely firing cells—that are quiet when the electrode passes by their vicinity unless the right stimulus is shown—and are more likely to record the activity of neurons with high spontaneous rates and broadly tuned responses. This introduces a bias toward distributed representations, which is likely to be prevalent in multiple descriptions of apparently distributed representations

in the literature. This issue is becoming quite relevant given recent evidence of highly sparse neurons in different systems (for a review see Olshausen & Field, 2004). For example, Perez-Orive and coworkers (Perez-Orive et al., 2002) using multielectrode recordings found cells in the mushroom body of the locust with a baseline activity of about 0.025 spikes per second, which fired about 2 spikes to very few odors. Hahnloser and coworkers (Hahnloser, Kozhevnikov, & Fee, 2002) found ultra-sparse firing neurons in the songbird using antidromic stimulation. These neurons had less than 0.001 spikes per second baseline activity and elicited bursts of about 4 spikes when the bird sang one particular motive. As shown in Figure 1, neurons in the human MTL can have a baseline firing of less than 0.01 Hz and respond with up to 50 Hz to very few stimuli.

Evidence for Local and Distributed Codes in the Brain

In his overview, Bowers (2009) revisited neurophysiology evidence of sparse and distributed representations and reinterpreted these works as evidence for localist and grandmother cell codes. He particularly referred to the recordings in macaque monkeys by Young and Yamane (1992), saying that these authors claimed to have provided evidence for a distributed code. It seems that Bowers's (2009) criticism of this article is due to the different meanings given to some terms by different communities of researchers. In fact, Young and Yamane argued for a sparse representation (in the sense of being opposite to distributed, as generally taken in neuroscience), which is already in the title of their well known *Science* article "Sparse Population Coding of Faces in the Inferior Temporal Cortex" (Young & Yamane, 1992). What may be confusing is the fact that they also referred to population coding, but this is just reflecting the fact that even with sparse responses, a population of neurons—in contrast to a single-cell—is needed to encode a percept.

Hung and coworkers (Hung et al., 2005) showed that neurons in monkey IT respond to multiple images (see also Kreiman et al., 2006). They used a statistical classifier to decode the activity of an ensemble of hundreds of neurons. Bowers (2009) argued that the classifier units coded for only one object and concluded that the data do not support distributed coding arguments. Here, it is important to distinguish between the experimental data (the recordings of neurons in IT cortex) and the classifier units. The units in IT cortex responded to multiple objects, and it was not possible to decode the presence of individual objects with high accuracy from only one neuron. In contrast to the case of IT neurons (the physiological data), the classifier units that operated on the output of hundreds of IT neurons showed sparser responses. But this does not provide any direct evidence that such a code exists in the brain, as the classifier is a theoretical construct. Further support to the claim of distributed coding by these neurons is given by the fact that decoding performance increased nonlinearly with the number of neurons. For a pure localist code, each neuron contributes to identify one or a few objects, and therefore, the decoding performance or alternatively the capacity—that is, the number of objects that can be identified at a fixed performance level—grows linearly with the number of neurons, as it is the case for recordings in the human MTL (Quian Quiroga et al., 2007). On the contrary, for a distributed code, each neuron contributes to the representation of many objects, and both decoding performance and capacity have a nonlinear growth with the number of neurons, as observed in IT recordings in monkeys (Hung et al., 2005[1]). In fact, it is in

principle possible to encode 2^N objects with a fully distributed network of N binary neurons, but it has to be noted that the exact nonlinear functional dependence with the number of neurons depends on several factors, such as noise levels, trial-to-trial variability, and saturation of decoding performance due to limited sampling of stimuli (Abbott, Rolls, & Tovee, 1996). In this respect, Bowers (2009) claimed that the exponential increase of decoding performance with the number of neurons found by Hung and colleagues (2005; and also by Rolls and colleagues Rolls, Treves, & Tovee, 1997, as described in the next paragraph) does not constitute evidence of distributed representations. This argument brings us back to the previous discussion of how to experimentally establish what a neuron codes for, given what it responds to. In our view, the fact that neurons fire to multiple stimuli (therefore having an exponential increase of performance with the number of neurons) gives strong evidence for distributed coding. Again, the claim that these neurons may encode only one thing and that they fire to the other ones by mere similarity (as in Bowers's, 2009, argument of the IA model) is of limited relevance because it cannot be verified or falsified with the existing data and recording tools.

Further evidence for distributed representations in visual processing areas comes from the recordings of Rolls and coworkers (Rolls et al., 1997), showing also an exponential increase in decoding performance with the number of neurons (see also Abbott et al., 1996). Bowers (2009) criticized these results because (a) the study was carried out on a set of face cells that were not highly selective and (b) the same analysis carried out on our MTL neurons would likely lead to a different conclusion. If Rolls and colleagues had recorded data from different areas, results may have been different because different areas may represent information in a different way. However, we do not see this as a problem with the experiment or the approach taken by these authors, as claimed by Bowers (2009). Rolls and colleagues reported observations based on the area they recorded from and did not generalize their claims to other areas. In fact, they explicitly mentioned that this encoding may be different in other parts of cortex and for other category of visual stimuli (Rolls et al., 1997). It is interesting to note that a similar decoding analysis was indeed carried out with our selective responses in the human MTL (Quian Quiroga et al., 2007). In contrast to the findings of Rolls et al. (Rolls et al., 1997) and Hung et al. (Hung et al., 2005), in this case, the decoding performance increased linearly rather than exponentially, in agreement with a very sparse or localist coding scheme.

An extreme example of sparse coding is given by the single-cell responses to picture presentations in the human MTL, as showed in the examples depicted in Figure 1. In spite of the striking degree of sparseness of these neurons, we argue that they cannot be taken as conclusive evidence of the existence of grandmother cells—understood in the sense that one neuron encodes only one object (Quian Quiroga, Kreiman, Koch, & Fried, 2008; Quian Quiroga et al., 2005). First, if there were one and only one neuron encoding for a given person or object, the chance of finding this neuron would be tiny. Second, the fact that a neuron responds to only one person during an experimental session is not a proof that the neuron encodes only for this person because it may also fire to

_____

[1] See Point 15 in http://klab.tch.harvard.edu/resources/ultrafast/index.html

some other stimuli that we did not happen to show during the recording. It is indeed not uncommon to find very selective MTL neurons firing to more than one person or object, as shown in Figure 1 (for more examples see Quian Quiroga et al., 2005).

From a dataset of 1,425 MTL units recorded in 34 experimental sessions, given the number of responsive units in a recording session, the number of stimuli presented, and the total number of recorded neurons, using probabilistic arguments we estimated that from a total population of about 10^9 neurons in MTL, less than 2×10^6 neurons (not 50–150 as incorrectly reported by Bowers, 2009) are involved in the representation of a given percept (Waydo, Kraskov, Quian Quiroga, Fried, & Koch, 2006). Furthermore, assuming that a typical person can recognize between 10,000 and 30,000 objects (Biederman, 1987), we estimated that each neuron fires in response to 50–150 different objects (Waydo et al., 2006). Bowers (2009) argued that this estimation is flawed because (a) multiple neurons can respond to the same image and (b) these calculations assume that a grandmother cell should only respond to one face or object. Briefly, the fact that multiple neurons can respond to the same image—a possibility that we consider very likely—is not a problem for the above calculations. In fact, it seems highly unlikely that we happened to find the one and only neuron that responds to a particular face. This argument was explicit in Quian Quiroga et al. (2005) and further quantified in Waydo et al. (2006). With regards to the second point, for our calculations we did not assume any properties of how grandmother cells should or should not respond at all, as claimed by Bowers (2009), but rather estimated an upper bound for the number of objects that a neuron may respond to. It should be also stressed that as discussed in Waydo et al. (2006), the estimated number of neurons responding to one concept could be much lower because (a) images known to the participants are more likely to elicit responses than are unfamiliar stimuli and (b) neurons with a higher degree of sparseness are very difficult to detect in our recording sessions lasting, on average, about 30 min.

Evidence from single-cell recordings shows that the brain may go from distributed representations in lower sensory areas to sparse representations in higher areas. We already mentioned the very sparse responses to odors by Kenyon cells in the locust (Perez-Orive et al., 2002). Kenyon cells neurons receive direct inputs from projection neurons in the antennal lobe, which have a largely distributed representation for odors (compare the responses in Figure 1A and Figure 1B in Perez-Orive et al., 2002). Similarly, the ultrasparse responses of the robust nucleus of the archistriatum (RA) neurons in the zebrafinch are driven by high vocal center (HVC) neurons with distributed responses (see Figure 2b in Hahnloser et al., 2002). Further evidence in other species is still scarce because as mentioned in the previous section, to compare selectivity across different areas one should use the same stimulus set. Barnes and coworkers (Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990) showed that neurons in the hippocampus in rats responded more selectively to the rat spatial location than neurons in entorhinal cortex. These results seem to support the hypothesis of complementary learning systems, with higher level of sparseness in the hippocampus than in cortex (Norman & O'Reilly, 2003), an appealing idea that would explain fast learning of new episodic memories and associations in the hippocampus with the use of sparse coding on the one hand and generalization in cortex with the use of a distributed representation on the other. Bowers (2009) criticized the study of Barnes et al. (1990) and its support to the comple-

mentary systems hypothesis by claiming that the entorhinal cortex is not part of neocortex and that a proper comparison of sparseness should be made between hippocampus and neocortex. However, the entorhinal cortex is the main gateway to the hippocampus—that is, most of the information from neocortex is conveyed to the hippocampus through the entorhinal cortex. To us, this gives valuable evidence of how the representation gets sparser when reaching the hippocampus. Moreover, a more recent study with single-cell recordings in the human MTL showed that the selectivity of the single-cell responses in the parahippocampal cortex (one of the main inputs to entorhinal cortex) was significantly lower than the one in the entorhinal cortex, the amygdala, and the hippocampus (Mormann et al., 2008).

It seems also plausible to argue that a distributed representation in IT is transformed into the sparser representation shown in the MTL (compare responses of Hung et al. (2005) in IT with those of Quian Quiroga et al. (2007, 2005) in MTL), given the close anatomical connections between these areas. However, we emphasize that this is still a conjecture due to the different recording techniques, species, and stimuli used in these studies. In this respect, it has been argued that the more distributed representations in IT (compared with MTL) may be necessary to identify the different views of the same person or object with a population code (DiCarlo & Cox, 2007), in contrast to the sparse and invariant responses in the human MTL, where neurons fire to the concept in an abstract manner and the particular view or details of the pictures are irrelevant. It is also possible that very sparse neurons are also present in IT but are hard to find, partially due to the technical difficulties described in the previous sections.

Conclusions

In summary, Bowers (2009) made a commendable effort to link psychological theories and computational models to the firing of individual neurons in the brain. This effort should be praised and hopefully extended through further interactions across these fields. In this commentary, we tried to emphasize the difficulties inherent to neurophysiology and the challenges involved in distinguishing between distributed and local codes. We also attempt to provide a quantitative framework to describe neuronal representations residing in a continuum that ranges from distributed to local representations. Given how poor our understanding of cortex currently is, we hope that this quantitative formulation will avoid semantic discussions and will pave the way to comparisons across areas, laboratories, and experimental conditions, as well as between physiology and computational models. Unraveling the codes used by circuits of neurons to represent information is arguably one of the most fascinating and challenging adventures at the intersection of psychology, computer science, and neuroscience.

References

Abbott, L. F. (1994). Decoding neuronal firing and modeling neural networks. *Quarterly Review of Biophysics, 27,* 291–331.

Abbott, L. F., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex, 6,* 498–505.

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.

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

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251.

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

DiCarlo, J. J., & Cox, D. (2007). Understanding invariant object recognition. *Trends in Cognitive Sciences, 11,* 333–341.

Gross, C. G. (2008). Single neuron studies of inferior temporal cortex. *Neuropsychologia, 46,* 841–852.

Gross, C. G., Rocha-Miranda, C. E., & Bender, D. B. (1972). Visual properties of neurons in inferotemporal cortex of the macaque. *Journal of Physiology (London), 35,* 96–111.

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

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational properties. *Proceedings of the National Academy of Sciences, 7 USA, 9,* 2554–2558.

Hopfield, J. J. (2007). Hopfield network. *Scholarpedia, 2,* 1977.

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

Koch, C. (2004). *The quest for consciousness.* Englewood, NJ: Roberts.

Kreiman, G., Hung, C. P., Kraskov, A., Quian Quiroga, R., Poggio, T., & DiCarlo, J. J. (2006). Object selectivity of local field potentials and spikes in the macaque inferior temporal cortex. *Neuron, 49,* 433–445.

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., & Group, P. R. (1986). *Parallel distributed processing: Psychological and biological models* (Vol. 2). Cambridge, MA: MIT Press.

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

Mormann, F., Kornblith, S., Quian Quiroga, R., Kraskov, A., Cerf, M., Fried, I., & Koch, C. (2008). Latency and selectivity of single neurons indicate hierarchical processing in the human medial temporal lobe. *Journal of Neuroscience, 28,* 8865–8872.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611–646.

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

Perez-Orive, J., Mazor, O., Turner, G. C., Cassenaer, S., Wilson, R. I., &

Laurent, G. (2002, July). Oscillations and sparsening of odor representations in the mushroom body. *Science, 297,* 359–365.

Quian Quiroga, R. (2007). Spike sorting. *Scholarpedia, 2,* 3583.

Quian Quiroga, 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.

Quian Quiroga, R., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation, 16,* 1661–1687.

Quian Quiroga, R., & Panzeri, S. (2009). Extracting information from neural populations: Information theory and decoding approaches. *Nature Reviews Neuroscience, 10,* 173–185.

Quian Quiroga, R., Reddy, L., Koch, C., & Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *Journal of Neurophysiology, 98,* 1997–2007.

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

Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1997). *Spikes: Exploring the neural code.* Cambridge, MA: MIT Press.

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: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89,* 60–94.

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

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19,* 109–139.

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

Willmore, B., & Tolhurst, D. J. (2001). Characterizing the sparseness of neural codes. *Network: Computation in Neural Systems, 12,* 255–270.

Wirth, S., Yanike, M., Frank, L. M., Smith, A. C., Brown, E. N., & Suzuki, W. A. (2003, June). Single neurons in the monkey hippocampus and learning of new associations. *Science, 300,* 1578–1581.

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

Postscript: About Grandmother Cells and Jennifer Aniston Neurons

Rodrigo Quian Quiroga
University of Leicester

Gabriel Kreiman
Children's Hospital Boston, Harvard Medical School, and Harvard University

A typical problem in any discussion about grandmother cells is that there is not a general consensus about what should be called as such. Here, we discuss possible interpretations in turn and contrast them with what we find in our own data (arguably the closest experimental evidence of grandmother cells so far). A first and naïve interpretation of the term grandmother cell is that one and only one neuron encodes for one and only one concept (a face, an object, an animal, etc.). We agree with Bowers (2010) that this is a straw-man version of this idea—although some people still take this view when (incorrectly) arguing that if we would have grandmother cells then the concept of grandma would disappear if her dedicated cell dies—which clearly does not apply to our data. Given that we record from a very tiny fraction of neurons in the medial temporal lobe (MTL), if we do find a neuron firing to a particular concept, there must be more. A more plausible version