Connection Science

ISSN: 0954-0091 (Print) 1360-0494 (Online) Journal homepage: http://www.tandfonline.com/loi/ccos20

What is a grandmother cell? And how would you know if you found one?

Jeffrey S. Bowers

To cite this article: Jeffrey S. Bowers (2011) What is a grandmother cell? And how would you know if you found one?, Connection Science, 23:2, 91-95, DOI: 10.1080/09540091.2011.568608

To link to this article: http://dx.doi.org/10.1080/09540091.2011.568608

Connection Science
Vol. 23, No. 2, June 2011, 91–95

Taylor & Francis
Taylor & Francis Group

What is a grandmother cell? And how would you know if you found one?

Jeffrey S. Bowers*

Department of Experimental Psychology, University of Bristol, Bristol, UK

(*Received 19 February 2011; final version received 28 February 2011*)

The key claim associated with a grandmother cell theory is that single neurons selectively *represent* one complex 'thing' (e.g. object and face). However, this theory is often mischaracterised in the cognitive and neuroscience literatures. I summarise two common confusions here. First, critics of grandmother cells often fail to distinguish between the selectivity and sparseness of neural firing and, as a result, predict (incorrectly) that one and only one neuron should fire in response to a given input. Second, critics often fail to distinguish between what a neuron responds to and what it represents – as detailed below – and as a result, predict (incorrectly) that a grandmother cell should fire in response to one and only one thing. I argue that these two confusions often lead to the premature rejection of grandmother cell theories.

Keywords: grandmother cell; distributed representation; parallel distributed processing

A number of recent papers highlight the remarkable selectivity of single neurons in the medial temporal lobe (MTL). For instance, a neuron has been found that responds to the image and the name of Saddam Hussein, but to no other stimulus presented in the experiment (Quian Quiroga, Kraskov, Koch, and Fried 2009). Similarly impressive reports of selective neurons in the inferotemporal cortex have been documented (e.g. Logothetis, Pauls, and Poggio 1995). The interpretation of these findings, however, is contentious. I have argued that these findings lend plausibility to 'grandmother cell' theories, according to which, single neurons represent complex perceptual categories such as faces, words, or objects (e.g. Bowers 2009, 2010), whereas others take these findings as evidence for sparse distributed, but not grandmother, coding schemes (e.g. Plaut and McClelland 2010; Quian Quiroga and Kreiman 2010). The usefulness of this exchange, however, has been somewhat compromised by a set of conceptual and terminological confusions. It is hard to reach any strong conclusions if researchers do not agree on the terms of the debate. Below, I highlight two common confusions and describe what I think are the relevant criteria for assessing grandmother cell theories.

The first confusion, highlighted by Foldiak (2009), results from a failure to distinguish between the selectivity and sparseness of neural firing. Selectivity measures the proportion of 'things' (e.g. objects and faces) that single neurons respond to; according to Foldiak, a grandmother cell only

*Email: j.bowers@bris.ac.uk

ISSN 0954-0091 print/ISSN 1360-0494 online

DOI: 10.1080/09540091.2011.568608
http://www.informaworld.com

responds to one thing in a universe of things. By contrast, sparseness measures the proportion of neurons that fire in response to a given thing; according to Foldiak, an object or face is encoded in a *local* format if it activates a single neuron in a population of neurons (i.e. is maximally sparse). Because grandmother cell theories constitute a claim about selectivity, Foldiak notes that it is perfectly sensible to talk about non-local grandmother cells.[1] That is, many redundant grandmother cells might respond to the same object. Indeed, redundancy is necessary for any viable grandmother cell theory in order to guard against noise and cell death (cf. Bowers 2009).

It is important to emphasise that grandmother cell theories have always been about selectivity rather than about sparseness (cf. Gross 2002). Indeed, one of the inspirations for grandmother cells was the discovery of *simple* and *complex* cells in V1 that respond highly selectively to lines of a given orientation in specific retinal locations. In addition to their selectivity, these cells are organised in a hierarchical manner, with complex cells in V1 combining the inputs from multiple simple cells in order to code for information more abstractly (complex cells respond to a line of a given orientation across more retinal locations). Hubel and Wiesel (1979) considered the implications of this hierarchal organisation within early vision and raised the question of whether the same design principles apply throughout, with grandmother cells at the top of a hierarchy:

> What happens beyond the primary visual area, and how is the information on orientation exploited at later stages? Is one to imagine ultimately finding a cell that responds specifically to some very particular item (Usually one's grandmother is selected as the particular item, for reasons that escape us.) Our answer is that we doubt there is such a cell, but we have no good alternative to offer. (p. 96)

Nevertheless, critics of grandmother cell theories often attempt to falsify this approach by providing evidence that multiple neurons fire in response to a given input. For example, when describing the MTL neurons that respond in a highly selective way to faces, Quian Quiroga, Kreiman, Koch, and Fried (2008, p. 88) write

> Although these cells bear some similarities to 'grandmother cells', several arguments make this interpretation unlikely. First, it is implausible that there is one and only one cell responding to a person or concept because the probability of finding this cell, out of a few hundred million neurons in the MTL, would be very small. From the number of responsive units in a recording session, the number of stimuli presented and the total number of recorded neurons, a Bayesian probabilistic argument was used to estimate that out of approximately one billion MTL neurons, fewer than two million represent a given percept. This is a far cry from a grandmother-cell-like. . .

This conclusion, even if true,[2] is irrelevant.

The second conceptual confusion results from a failure to distinguish between what a neuron responds to and what it represents. This contrast is fundamental to 'localist' models in psychology that provide a possible implementation of grandmother cell theories. For illustration, consider the interactive activation (IA) model of word identification (McClelland and Rumelhart 1981). The model includes three levels: an input layer composed of a set of visual-feature detectors that represent line segments in various orientations; a second layer composed of nodes that represent letters; and a third layer composed of nodes that represent individual familiar words. Word identification is achieved when a single word detector is activated beyond some threshold. Although each word unit (e.g. TRAP) selectively represents one word (e.g. *trap*), each word activates multiple word units and each word unit is activated by multiple words. For example, in response to the word *trap*, the TRAP unit will become most active, but the units for WRAP, TRIP, TRAM, TRAY, etc., will become active as well (by virtue of sharing three letters with *trap*). Nevertheless, the WRAP, TRIP, TRAM, and TRAY units play no role in representing the word *trap*. Indeed, removing all the units would have no impact on the representation of the word *trap* (the word *trap* would still be recognised when the activation of the TRAP unit passes some threshold, just as before). In the same way, the unit for TRAP will be activated by the words *trap* and *trip* (among other words), but the TRAP unit is not involved in representing any other words (e.g. remove the TRAP unit,

and the representation for *trip* is unaffected). All implemented localist (grandmother cell) theories work this way. The key claim has always been that single units (neurons) *represent* one and only one thing.

Nevertheless, this distinction is often ignored in neuroscience, or flatly rejected. For example, Quian Quiroga and Kreiman (2010) write: 'Contrary to Bowers (2010), we do not make much distinction between what a neuron codes for and what it responds to.' On the latter view, anytime a neuron fires, it is involved in representing the input. This indeed would be problematic for a grandmother cell theory, as every neuron responds to more than one thing. For example, based on the same Bayesian probabilistic argument used to approximate that millions of neurons respond to a given input, Waydo, Kraskov, Quian Quiroga, Fried, and Koch (2006) estimated that each neuron in MTL responds to between 50 and 150 different faces. This again is taken to rule out grandmother cell theories.

But note that this conclusion is only valid if MLT neurons are involved in *representing* 50–150 faces. These findings are equally consistent with the fact that the neurons represent one thing and are incidentally activated by other similar things (i.e. if neurons work like localist representations in cognitive models). Indeed, if Waydo et al. applied the same logic to the IA model, they would conclude that each node represents many different words (e.g. the TRAP unit in the IA model is involved in representing the words *wrap*, *trip*, *tram*, and *tray* because it responds to all these words). This would be the incorrect characterisation of the IA model, and the same mistake may apply to their analysis of single cells in the MTL. If researchers want to falsify grandmother cell theories, they need to show that there is a fundamental flaw with localist (grandmother cell) theories that do distinguish between what a unit (or neuron) responds to and what it represents.

The focus on what neurons respond to rather than on what they represent can lead to additional confusions. For example, as noted by Foldiak (2009), grandmother cells constitute a theory of neural selectivity rather than sparseness. Nevertheless, Foldiak mischaracterises grandmother cells because he does not define selectivity in terms of what neurons *represent*. This can be seen in Figure 1 (adapted from his paper). The figure is intended to highlight the distinction between sparse and selective responses. But consider the response of Neuron 2. This neuron is described as a grandmother cell because it only fires in response to Stimulus 6. However, this neuron does not represent Stimulus 6 (Stimulus 6 is represented through the co-activation of Neurons 2 and 6). Indeed, the activated Neuron 2 by itself does not correspond to any of the 10 stimuli listed in the first column in Figure 1. In a similar way, the local encoding of Stimulus 4 (only Neuron 5 is active in response to Stimulus 4) does not constitute a grandmother cell representation. The system can only infer that Stimulus 4 is present based on the knowledge that Neuron 5 is on and all other neurons are off. If Neurons 1, 9, and 10 were co-activated with Neuron 5, then the system would be coding Stimulus 6 rather than Stimulus 4. Grandmother cells constitute a theory of how knowledge is represented, and accordingly, it is not sufficient to focus on the response of neurons without considering what they represent.

What then is a grandmother cell? I am claiming that a grandmother cell theory is consistent with many (perhaps very many) neurons firing in response to a given stimulus (Confusion 1), and that a grandmother cell does not have to respond to stimuli in a completely selective way (Confusion 2). Does this strip grandmother cell theories of any meaning? Not at all. The claim is straightforward and widely rejected (and ridiculed) in psychology; namely, neurons selectively represent information in the same way as localist units selectively represent information in artificial neural networks. That is, grandmother neurons represent one and only one thing.

Note that I am not revising the definition of grandmother cells in order to accommodate the existing data. The question in neuroscience has always been whether the hierarchy of visual processing steps found in V1 continues to the point of having single neurons that represent complex things. Contrary to Hubel and Wiesel (1979), I am arguing that this hypothesis is, in fact, plausible (given current evidence). Within psychology, the widespread assumption has been that

	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	Sparseness
Stim1	1	0	0	0	1	0	0	0	0	0	0.2
Stim2	0	0	0	0	0	0	0	1	1	0	0.2
Stim3	0	0	0	1	0	0	1	0	0	0	0.2
Stim4	0	0	0	0	1	0	0	0	0	0	0.1
Stim5	1	0	0	0	1	0	0	0	1	1	0.4
Stim6	0	0	1	0	0	0	0	0	1	0	0.2
Stim7	0	1	0	0	0	1	0	0	0	0	0.2
Stim8	0	0	0	0	0	0	0	1	1	0	0.2
Stim9	1	0	1	0	1	1	0	0	0	1	0.5
Stim10	0	0	1	0	0	0	0	0	1	0	0.2
Selectivity	0.3	0.1	0.3	0.1	0.4	0.2	0.1	0.2	0.5	0.2	0.24

Figure 1. Sparseness and selectivity measures. Adapted from Foldiak (2009). Ten neurons (N1–N10) are listed in the first row and 10 stimuli (Stim1–Stim10) are listed in the first column. The activation of each neuron in response to each stimulus is depicted as either 1 or 0 (the neurons fire in a binary fashion). Sparseness measures the percentage of neurons across the population of neurons that are activated by a given stimulus (final column), and selectivity measures the percentage of images across the population of images that activate a given neuron (final row). Sparseness varies from a local response (one active neuron in response to a stimulus; e.g. Stim4) to a sparse response (a small percentage of neurons activate in response to a given stimulus, but more than one; e.g. Stim1), to a dense response (a high percentage of neurons respond to a given stimulus; e.g. Stim9). The highlighted row depicts the local coding of Stim4. Selectivity varies in terms of breadth of tuning, from a broad response (many effective stimuli per neuron; e.g. N9) to a narrow response (few effective stimuli per neuron, but more than one; e.g. N6) to a grandmother cell response (only one stimulus drives a given neuron; e.g. N2). The highlighted columns depict grandmother cells. Note that sparseness and selectivity measures only have to be equal on average (0.24 in this example).

local representations are inconsistent with neuroscience. I am arguing that these representations are, in fact, consistent with current findings.

Still, the question remains as how to distinguish between grandmother and (sparse) distributed representations empirically. Here, the answer is not so straightforward, but at least three lines of research are required. First, it is important to characterise the alternative theories correctly so that they can be appropriately evaluated in light of data. I hope that the current paper helps clarify two key confusions. Second, it will be important to compare the responses of real single neurons with the responses of modelled single units in localist and distributed models in order to determine which approach better captures the neurophysiological data. Note that in the case of the distributed representations learned in parallel distributed processing (PDP) models, it is generally assumed that the 'hidden units' respond in a highly unselective manner. Indeed, this assumption has been so widespread that there have been few attempts to study single units in PDP models during the past 25 years. The problem with this view, however, is that it contradicts the key finding that underpins 40–50 years of neurophysiology; that is, single neurons often do respond in a highly selective and interpretable way (Parker and Newsome 1998). More recently, it has been emphasised that PDP models are able to learn sparse and highly selective representations (e.g. Plaut and McClelland 2010), and accordingly, these models may be reconciled with the relevant neuroscience. It will be important to determine under what conditions PDP models learn highly selective representations and then contrast the relative success of PDP and localist models in accounting for the relevant neurophysiology.

Third, it will be important to compare the ability of these models to capture behavioural (as opposed to neural) data. I would argue that localist models have been more successful in this regard, and furthermore, local representations provide a more promising medium to support 'symbolic' cognition and perception (Hummel 2000; Bowers, Damian, and Davis 2009). But whatever be one's view regarding the relative success of these models at the behavioural level, it is striking how models constrained by both single cell neurophysiology and behaviour have inspired localist (grandmother cell) theories. For example, Riesenhuber and Poggio (1999) developed a neural model of face recognition that is similar to the IA theory in many regards. Most critically, the

model includes single units that represent specific faces. Riesenhuber (2005) describes this model as 'an extension of the original model of simple and complex cells of Hubel and Wiesel' (p. 279). That is, although Hubel and Wiesel (1979) rejected the claim that the visual hierarchy in V1 continues up to the level of coding complete faces, Riesnhuber and colleagues have implemented just such a model.

In sum, the key claim of a grandmother cell theory is that single neurons selectively *represent* one complex 'thing' (e.g. object, face, and word), just as local representations in cognitive models selectively represent one thing. Localist models in psychology are often rejected because grandmother cells are said to be biologically implausible, but his line of argument is undermined by the fact that grandmother cells are often mischaracterised. When the above-mentioned confusions are avoided, I would argue that grandmother cell theories constitute a viable account of how knowledge is represented in the brain. Indeed, at present, localist models do a better job than PDP models in accounting for the range of results reported in neuroscience (cf. Bowers 2009, 2010).

Notes

1. Confusingly, the term 'local' is used differently in different contexts. Within psychology, a local representation is selective, whereas in Foldiak's terminology, a local representation is sparse.
2. As Quian Quiroga et al. (2008), and Quian Quiroga and Kreiman (2010) note themselves, the estimates of 2 million active neurons may be far too high.

References

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.

Bowers, J.S. (2010), 'More on Grandmother Cells and the Biological Implausibility of PDP Models of Cognition: A Reply to Plaut and McClelland (2010) and Quian Quirog and Kreiman (2010)', *Psychological Review*, 117, 300–306.

Bowers, J.S., Damian, M.F., and Davis, C.J. (2009), 'A Fundamental Limitation of the Conjunctive Codes Learned in PDP Models of Cognition: Comments on Botvinick and Plaut (2006)', *Psychological Review*, 116, 986–995.

Foldiak, P. (2009), 'Neural Coding: Non-local but Explicit and Conceptual', *Current Biology*, 19, R904–R906.

Gross, C.G. (2002), 'The Genealogy of the "Grandmother Cell"', *The Neuroscientist*, 8, 512–518.

Hubel, D., and Wiesel, T. (1979), 'Brain mechanisms of vision', *Scientific American*, 241, 150–162.

Hummel, J.E. (2000), 'Localism as a First Step Toward Symbolic Representation', *Behavioral and Brain Sciences*, 23, 480–481.

Logothetis, N.K., Pauls, J., and Poggio, T. (1995), 'Shape Representation in the Inferior Temporal Cortex of Monkeys', *Current Biology*, 5, 552–563.

McClelland, J.L., and 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.

Parker, A.J., and Newsome, W.T. (1998), 'Sense and the Single Neuron: Probing the Physiology of Perception', *Annual Review of Neuroscience*, 21, 227–277.

Plaut, D.C., and McClelland, J.L. (2010), 'Locating Object Knowledge in the Brain: A Critique of Bowers's (2009) Attempt to Revive the Grandmother Cell Hypothesis', *Psychological Review*, 117, 284–290.

Quian Quiroga, R., and Kreiman, G. (2010), 'Measuring Sparseness in the Brain: Comment on Bowers', *Psychological Review*, 117, 291–299.

Quian Quiroga, R., Kreiman, G., Koch, C., and Fried, I. (2008), 'Sparse but Not "Grandmother-Cell" Coding in the Medial Temporal Lobe', *Trends in Cognitive Sciences*, 2, 87–91.

Quian Quiroga, R., Kraskov, A., Koch, C., and Fried, I. (2009), 'Explicit Encoding of Multimodal Percepts by Single Neurons in the Human Brain', *Current Biology*, 19, 1308–1313.

Riesenhuber, M. (2005), 'Object Recognition in Cortex: Neural Mechanisms, and Possible Roles for Attention', in *Neurobiology of Attention*, eds. L. Itti, G. Rees, and J. Tsotsos, San Diego, CA: Elsevier, pp. 279–287.

Riesenhuber, M., and Poggio, T. (1999), 'Hierarchical Models of Object Recognition in Cortex', *Nature Neuroscience*, 2, 1019–1025.

Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., and Koch, C. (2006), 'Sparse Representation in the Human Medial Temporal Lobe', *Journal of Neuroscience*, 26, 10232–10234.