Grandmother cells and localist representations: A review of current thinking.

Jeffrey S. Bowers

University of Bristol

Author Note:

Jeffrey S. Bowers, School of Experimental Psychology, University of Bristol.

Acknowledgments: I would like to thank Ella Gale for comments on previous drafts. This project was supported by a grant from The Levehulme Trust RPG-2016-113 awarded to Jeffrey Bowers and Colin Davis.

Correspondence concerning this article should be addressed to Jeffrey S Bowers, School of Experimental Psychology, 12a Priory Road, Bristol, BS8-1TU. Email j.bowers@bristol.ac.uk

Abstract

There is now a large literature in neuroscience highlighting how some neurons respond highly selectively to high-level information (e.g., cells that respond to specific faces) and a growing literature in psychology and computer science showing that artificial neural networks often learn highly selective representations. Nevertheless, the vast majority of neuroscientists reject 'grandmother cell' theories out of hand, and many psychologists reject localist models based on neuroscience. In this review I detail some of the conceptual confusions regarding grandmother cells that have contributed to this state of affairs, and review the literature of single-unit recording studies in artificial neural networks that may provide insights into why some neurons respond in a highly selective manner. I then briefly review the contributions from leading theorists in psychology and neuroscience. My hope this special issue contributes to a more productive debate on an important issue that has often been characterized by misunderstandings between disciplines.

Key words: grandmother cells; localist representations; distributed representations; neural networks

Grandmother cells and localist representations: A review of current thinking.

For over 30 years the distinction between localist *vs.* distributed coding has been central to theorizing in psychology (McClelland, Rumelhart, & PDP Research Group, 1986; Rumelhart, McClelland, & the PDP Research Group, 1986). By contrast, the distinction between grandmother cells *vs*. distributed coding has been less central to theory in neuroscience, and indeed, the term 'grandmother cell' is somewhat pejorative, designed to highlight the absurdity of the hypothesis. Nevertheless, in recent years, a number of high-profile publications have highlighted the extremely selective firing of some neurons in humans, and a growing number of computational studies have shown that artificial neural networks learn highly selective representations as well. This has led to a growing interest in grandmother cells (See Figure 1), but still, it remains the case that neuroscientists overwhelmingly dismiss the grandmother cell hypothesis, and attempts to link theory in psychology and neuroscience is often characterized by misunderstandings that have slowed theoretical progress in both disciplines.

Why is the grandmother cell hypothesis still so widely rejected in neuroscience, and what are the theoretical confusions between disciplines? In an attempt to address these questions I thought it would be useful to first explain why I (a cognitive psychologist) became interested in grandmother cells in the first place, and briefly summarize some points of disagreement between myself and others. I hope this is an effective (rather than self-indulgent) way to explain to neuroscientists what psychologists are talking about when contrasting localist and distributed theories and why this has some bearing on the grandmother *vs.* distributed contrast in neuroscience. At the same time, I hope this helps clarify for psychologists the relevance of neuroscience to the localist *vs* distributed debate.

This then provides the context for a brief review of single-unit recording studies in artificial neural networks. As far as I am aware, no one has reviewed this literature

previously (in contrast to the multiple reviews of single-cell recording studies in brain), but the findings may provide some important insights into why some neurons respond highly selectively. Finally, I conclude by providing a brief summary of the excellent set of contributions to this special issue from leading theorists in psychology and neuroscience. Although some of the contributors reject grandmother cells outright (Rolls) or challenge their functional relevance (Thomas and French), most of the contributors take grandmother cells as a serious hypothesis about how the brain codes for information. Together, the chapters provide much needed discussion regarding how grandmother cells theories should be considered in neuroscience and psychology.

Why (some) cognitive psychologists take localist representations seriously, and why this should matter to neuroscientists.

As a cognitive psychologist I am interested in modeling complex behavior at an algorithmic level that describe the processes that are involved in solving a given task (specifying how a problem is solved), but do not consider in any detail how the processes are implemented in the brain. Bayesian theories of mind within psychology are often further removed from neuroscience, as these theories tend to be developed at a computational level of analysis that specify the goals and problems that people face with little consideration algorithmic, let alone the neural, underpinnings of behavior (Bowers & Davis, 2012). Nevertheless, most cognitive psychologists (including most Bayesian theorists) agree that a cognitive model should be consistent with what we know about the brain, and it is considered an advantage when a model has clear links to neuroscience. Indeed, this is one of the main motivations for connectionist networks that provide an intuitive link between artificial units and connections on the one hand, and neurons and synapses on the other.

It is in this context that I often found conversations with students and colleague on the topic of localist vs. PDP models in psychology quite frustrating. Localist models have a long history in psychology, and they often provide the best account of behavior across a range of domains. But when discussing the relative merits of these two approaches I found the successes of localist models did not seem to carry much weight, and the reason was always the same: Localist model were assumed to be biologically implausible. If a localist model did a better job, it was not a sign that localist models should be taken seriously, but rather, that distributed PDP models need to improve.

To illustrate this perspective, here is a passage from Seidenberg and Plaut (2006) in the context of comparing localist and distributed processing models of visual word identification (although they apply same arguments to all domains of cognition). The authors note that localist models often account for more empirical findings than PDP models, but nevertheless, this is not considered key for evaluating the two approaches. They write:

> "The dual-route and PDP approaches to understanding word reading are both supported by explicit computational simulations, but the role that these simulations play in theory development in the two cases is strikingly different. The DRC model of Coltheart et al. (2001) continues the long tradition of a bottom-up, data-driven approach to modelling: A model is designed to account for specific behavioural findings, and its match to those findings is the sole basis for evaluating it. These models aspire to what Chomsky (1965) called "descriptive adequacy". The PDP approach is different. The models are only a means to an end. The goal is a theory that explains behaviour (such as reading) and its brain bases. The models are a tool for developing and exploring the implications of a set of hypotheses concerning the neural basis of cognitive processing. Models are judged not only with respect to their ability to account

> for robust findings in a particular domain but also with respect to considerations that extend well beyond any single domain. These include the extent to which the same underlying computational principles apply across domains, the extent to which these principles can unify phenomena previously thought to be governed by different principles, the ability of the models to explain how behaviour might arise from a neurophysiological substrate, and so on. The models (and the theories they imperfectly instantiate) aspire to what Chomsky termed "explanatory adequacy". The deeper explanatory force derives from the fact that the architecture, learning, and processing mechanisms are independently motivated (as by facts about the brain) rather than introduced in response to particular phenomena."

Although I agree with the authors that a range of criteria are relevant to assessing the merits of a specific model or modelling approach, there are reasons to questions their conclusions regarding the advantages of the PDP approach. First, although they claim that PDP models provide a more general and principled explanation of a wide variety of phenomena, they ignore powerful localist modelling frameworks. For example, the theories of Grossberg and colleagues provide a general set of learning and computational principles that apply across a wide range of cognitive and behavioural domain in a biologically plausible manner. Localist representations have played a central role in in his models (see Grossberg, this issue). Accordingly, there is little reason to assert that the PDP approach has privileged access to explanatory adequacy. Second, and most importantly for present purposes, the claim (common in the literature) that distributed codes in PDP models are more biologically plausible is asserted without evidence. Nevertheless, this claim, more than anything else, dominated my conversations regarding the relative merits of these two approaches.

This lead to the Bowers (2009) paper where I provided a detailed review of single-cell neurophysiology that characterized the responses of single neurons in a range of species performing a variety of tasks, and where I related the findings to the predictions of localist and distributed models developed in psychology. The striking finding from neuroscience literature is that great deal of much information can be retrieved from the firing of single neurons, as nicely captured in this quote from over 20 years ago:

> 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 (105/mm3 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 single 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)

Based on the review of the data, and a consideration of how localist representations in cognitive models function, I concluded that the neuroscience does not falsify localist models in psychology.

A follow-up debate between myself and proponents of distributed theories within both psychology (Plaut & McClelland, 2010) and neuroscience (Quian Quiroga & Kreiman, 2010) was useful in highlighting some of the disagreements and confusions regarding the term grandmother cell (mirroring a similar debate in response to Page, 2000). On some definitions

grandmother cells are clearly untenable. For example, if grandmother cell theories are committed to the claim that there is a single neuron associated with each unique experience (e.g., a single neuron coding for my grandmother knitting by the fireplace), with grandmother cells selectively responding to one input and not responding above baseline to anything else, with only one neuron per experience (no redundancy), then indeed, this is an implausible theory that is falsified by the data. This characterization of grandmother cells is widespread in neuroscience (see Rolls, this issue). But on less extreme definitions (e.g., units that code for familiar categories rather than all possible experiences), there is some room for disagreement, and indeed, a small group of theorists have at least considered grandmother cells as a serious hypothesis that should not be dismissed out of hand (e.g., Barlow, 1972; Bowers, 2002; Elliott & Susswein, 2002; Gross, 2002; Newsome, Britten, Movshon, 1989, Page, 2000; Perrett et al., 1989; Thorpe, 1989).

In Bowers (2009, 2010) I suggested that grandmother cells should be defined as localist representations as used in psychology; that is, units that represent one thing but are activated by related things. For example, in the Interactive Activation (IA) model (McClelland and Rumelhart, 1981), a localist representation for the word DOG is activated most strongly by the input DOG, but it is also activated (to a lesser extent) by the visually similar words such as HOG and LOG by virtue of their shared letters (for more detail, see Michele et al., this issue). Apart from having a much more precise definition, the benefit of defining grandmother cells as localist representations is that it adopts a common terminology across disciplines that should help avoid confusions (e.g., making it clear rejecting the extreme version of grandmother cells has no bearing on theory in psychology), it makes the neuroscience relevant to assessing theory in psychology, and it makes the large and sophisticated modelling tradition in psychology relevant to understanding the response properties of neurons. Of course, it is far from clear that brains rely on localist

representations, but it is question worth asking (unlike the extreme grandmother cell theory that no one has ever endorsed and that Rolls, this issue, rightly rejects).

Since 2010 there have been many high-profile reports of neurons responding to high-level information in a highly selective manner in humans (largely in the hippocampus and related structures; e.g., Ison, Mormann, Cerf, Koch, Fried, Quian Quiroga, 2011; Ison, Quian Quiroga., & Fried, 2015; Rey et al., 2015), and multiple review articles on single-cell recording studies (Quian Quiroga, 2012, 2016; Quian Quiroga, Fried., & Koch, 2013; Yuste, 2015; Roy, 2012, 2015). Rather than provide another review of the neuroscience, I thought it would be more useful to provide a brief review studies reporting highly selective units in PDP models and so-called "deep" networks that have recently been the focus of so much attention in computer science. This later work has been carried out with little consideration of how the results relate to theory in psychology and neuroscience, but I would argue that the findings are also relevant to the current issue.

Brief review of single-unit recording in artificial neural networks:

Single-unit recordings in artificial neural networks have been explored in both psychology and computer science, and although similar results have been reported, the amount of interest and attention to the work in the two fields is very different. I briefly review the two literatures next.

Single-unit recordings of PDP models in the psychological literature: Although PDP networks have been popular in psychology since the mid-1980s, few single-unit recordings have been carried out. Given that it is a far easier task to probe units in PDP models compared to real brains (where single-neural coding has been an active field since the 1950s), and given our limited understanding how PDP models work (McCloskey, 1991) an obvious question is why? And I think the answer is simple: most researchers studying connectionist networks in psychology have *assumed* that the learned representations are

distributed with the activations of single units meaningless. As Mozer and Smolensky (1989, p. 3) put it:

> "... one thing that connectionist networks have in common with brains is that if you open them up and peer inside, all you can see is a big pile of goo."

The term "sub-symbolic" (Smolensky, 1988) was commonly used to highlight that individual units cannot be interpreted by themselves, and this was considered a major break from previous theorizing, and key to understanding how cognition is implemented in brain. This continues to be the mainstream view of PDP modelers. For example, in a recent review of PDP modelling, Rogers and McClelland (2014) wrote:

> … a percept of a visual input is assumed to be represented as a pattern of activation distributed over many neurons in several different brain areas, and each neuron is thought to participate in the representation of many different items. This representational scheme is held to apply to essentially all kinds of cognitive content: Words, letters, phonemes, grammatical structures; visual features, colors, structural descriptions of objects; semantic, conceptual, and schema representations; contents of working memory and contextual information affecting processing of current inputs; speech plans, motor plans, and more abstract action plans— all are thought to take the form of distributed patterns of activation over large neural populations.

But despite these strong claims, there have been very few attempts to directly test this assumption by carrying out single-unit recording studies.

As far as I am aware, the first attempt to carry out a single-cell recording study analogous to single-cell recordings in the brain was reported by Berkeley, Dawson, Medler, Schopflocher, and Hornsby (1995). They trained simple three-layered networks via back-propagation on a variety of tasks, including a logical reasoning task that had previously been

simulated by Bechtel and Abrahamsen (1991), and the "kinship problem" problem studied by Hinton (1986). The model they focused on was trained on the logical problem and it included 14 input units (a pattern of activation across these units defined the input problem, with pairs of input units coding for the individual components of a logical problem, such as OR, AND, IF-THEN, etc.), 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 10 hidden units. The key point for present purposes is that after training, the model was able to correctly categorize 576 input patterns (logical statements) into six categories.

After training they recorded the response of each hidden unit to a range of inputs using a scatter plot for each unit. The unit's response to a specific input was coded with a point along the x-axis, with values on the y-axis arbitrary (y-axis is included in order to prevent points from overlapping). These so-called "jittered density plots" are roughly analogous to the raster plots used to measure the firing patterns of neurons to different stimuli (for an example of a jittered density plot, see Vankov and Bowers, this issue). The critical finding was that the scatter plots often took on a "banding" patterns, in which multiple different inputs (different logical problems) drove a hidden unit to the same level of activation. In some cases, the banding was consistent with localist coding. For example, hidden unit 6 in their model responded strongly to all logical problems that included the "OR" feature, and not at all to other inputs. This is analogous to a neuron that responded to all images of Jennifer Aniston but not to other faces. Accordingly, unit 6 appears to constitute a localist representation for the input OR.

In subsequent work, Dawson and Piercy (2001) and Berkeley (2006) carried out lesion studies on the units from the original Berkeley et al. (1995) network, and in some cases, the units functioned like localist units. For example, after removing unit 6, the model

performed well on problems that did not involve OR, and catastrophically failed (0%) on all problems that included the OR feature. Although the authors disagreed somewhat on how to characterize the units, the findings clearly show how single units respond highly selectively to inputs, and that removing single units can have selective impairment on performance (just the opposite of so-called "graceful degradation" in which lesions to single units results in a small overall decrement in performance across many items, a pattern of result predicted from distributed theories).

Similar banding patterns were obtained in other tasks and network designs. Leighton and Dawson (2001) reported banding in a PDP model trained on the Wason's card selection task that involves training a network a conditional rule (of the form of 'If P then Q') using a similar network to Berkeley et al. (1995). Berkeley, and Gunay (2004) found banding patterns when they used a standard sigmoid activation function in their network in contrast with the 'value units' they had used in previous work. Originally, Berkeley et al. (1995) had claimed that this banding pattern was restricted to networks with a specific type of activation function. Further evidence that this pattern of results is quite general was reported by Niklasson and Boden (1997) who reported banding patterns in a different sort of network that used a sigmoid activation function to map a set of inputs into 6 different categories.

It should be noted that in many cases (indeed most cases) the units the networks reported by Berkeley and colleagues learned more than two bands, such that a given unit responded selectively to more than one thing. For example, a unit might activate .5 (out of a maximum of 1.0) to the input feature OR and 1.0 to the AND input feature, and not at all to other inputs (resulting in 3 bands). This is an interesting case in which the unit has properties of both localist and distributed coding. That is, it is possible to interpret what the unit is responding to (if the unit is activated .5 then the OR unit is present), but the unit is involved in coding multiple things. This is perhaps reminiscent of what has been called a "totem pole"

neurons (Malach, 2012). But the most relevant units for present consideration are the units that contain two bands, and that selectively responded to one input, as did hidden unit 6 in Berkeley et al. (1995).

More recently, my colleagues and I have used these scatter jitter plots to characterize the representations learned in larger recurrent neural networks that co-activate multiple items at the same time in short-term memory (STM). Our work was inspired by earlier work of Botvinick and Plaut (2006) who developed a recurrent PDP model of immediate serial recall that was trained to encode a series of letters and then recall them back in the same order (e.g., given the sequence A, F, Q, recall A, F, Q). The authors claimed that the model succeeded by co-activating multiple distributed patterns of letters in the hidden layer. We found this conclusion surprising as it appeared to challenge the claim that distributed codes are poorly suited for co-activating multiple items at the same time, due to the superposition catastrophe (Von der Malsburg, 1986). Indeed, the superposition catastrophe has provided a computational reason for learning localist codes in cortex (Bowers, 2002; Page, 2000), just as catastrophic interference provided a pressure to learn selective and sparse codes in the hippocampus (Marr, 1971). If indeed the Botvinick and Plaut (2006) model supports STM through the co-activation of multiple overlapping distributed patterns, it would undermine this argument.

However, we found that recurrent PDP networks that were successful in co-activating multiple items at the same time learned many localist codes (units with two bands), with the number of local codes increasing when the superposition constraint became more difficult (Bowers, Damian, Vankov, Davis, 2012; Bowers et al., 2014). Furthermore, we found that recurrent PDP models of immediate serial recall could only generalize to novel items (e.g., recalling a sequence of novel words) when they learned localist representations (of letters). That is, we found that localist codes were better able to support generalization, just the

opposite to what is typically claimed. We also found that lesioning learned localist units in these networks often led to highly specific deficits in performance. For example, after deleting unit 152 (out of a total of 200 units) that selectively responding to the letter 'J', we presented the model with 1000 words (all composed of 3 letters), of which 100 contained the letter 'J'. The model was 99% in recalling words that did not contain the letter 'J', and failed 100% of the time on words that did contain the letter 'J'. (See Table 1 from Bowers et al., 2016.) This nicely parallels the results of Berkeley (2006) who found highly selective deficits following the lesioning of single local units.

In Vankov and Bowers (this issue) we explored the impact of arbitrary input-output mappings on the nature of the learned representations in PDP networks. As detailed below, we found that PDP models succeeded on the basis of learned distributed representations in most conditions, but networks did learn localist representations in some conditions even though the model was trained on items one-at-a-time. We concluded that the superposition constraint provided a stronger pressure to learn localist representations that arbitrary input-output mappings, but that the superposition constraint is not the only pressure to learn localist codes.

As far as I am aware, these are the only single-unit recording studies carried out on PDP networks within the psychological literature. Nevertheless, a number of conclusions seem justified, including that PDP models sometimes learn localist codes, and that learned localist representations in PDP models have functional value (such that removing localist units has specific predicted consequences). These findings also suggest hypotheses about when (and why) neurons in cortex (as opposed to hippocampus) respond selectively. For example, selective responding might be expected in cortical systems that generalize and support short-term memory (Bowers et al., 2016).

Single-unit recording of deep networks in computer science: In contrast with the limited number of studies in psychology, there has been an explosion of single-unit recording studies in the computer science literature when applied to "deep" networks. This has followed the extraordinary success of these networks in solving a challenging range of complex tasks, including state-of-the-art speech (Hannun et al., 2014) and image (Krizhevsky, *et al.*, 2012) recognition, and even game playing (Mnih et al., 2015). Deep networks are now being used in a wide range of applied settings, and companies are investing billions of pounds in developing deep networks because their enormous promise.

Two features of these networks are worth noting for present purposes. First, they are not so different from the early PDP networks developed in the 1980s. Although there have been some innovations to improve their performance, by the most important difference is that: a) computers with graphic cards can be trained many thousand times more quickly, and b) there are now much larger datasets of labeled data that are needed for supervised learning (cf., Ciresan, Meier, Gambardella, & Schmidhuber, 2010). This allows massive networks (sometimes up to 1 billion connections over multiple layers) to be trained on massive datasets (e.g., Le et al., 2013).

Second, as is the case with PDP networks, there is relatively little understanding how or why these networks perform as well as they do. Indeed, this has been the prime motivation carrying out single-unit recordings in deep networks. But unlike the single-unit recording studies carried out in psychology or neuroscience, the main goal of these single-unit studies has been to improve the performance of the networks, with little consideration of how the findings relate to theory in psychology or neuroscience. Nevertheless, this work may also be relevant to theory in the same way simple PDP networks are; namely, insights into when and why these large networks learn selective codes may provide hypotheses as to why some neurons respond in a highly selective manner.

One approach to single-unit recording in computer science is broadly similar to the single-unit recordings in psychology (and neuroscience). That is, the activation of single-units is recorded in response to many different meaningful inputs in an attempt to determine whether a consistent set of inputs (e.g., images of specific objects) drive the unit. Yosinski, Clune, Nguyen, Fuchs, and Lipson (2015) call this the "data-centric" approach to characterizing the function of individual units. Once again, localist representations were discovered across a range of different types of networks and across a range of different tasks. For example, Le et al. (2013) observed localist codes in a "deep belief" network that learned a localist representation of a face without supervision, whereas other researchers have reported localist representations in deep convolutional networks trained with supervised learning methods (Agrawal, Girshick, & Malik, 2014; Li, Yosinski, Clune, Pipson, Hopcroft, 2015; Zeiler, & Fergus, 2014; Zhou, Khosla, Lapedriza, Oliva, & Torralba, 2015). And in these later networks, localist coding was observed across a range of training conditions. For example, Zhou et al. (2015) carried out single-unit recordings when the same network was trained on two different tasks. In one condition, a convolutional network was trained to categorize 1.3 million images into 1000 different object categories, and in other, the same model was trained to categorize 2.4 million images into 205 scene categories. They found localist codes in both cases, but surprisingly, found more localist codes for objects in the later case even though the model was not trained on objects (objects were diagnostic of scene categories, which made learning localist codes for the objects relevant to task performance). In other work, Li et al. (2015) reported highly overlapping set of learned localist codes when the same networks were given different random initializations of weights, again showing the generality of the findings.

A second and more popular approach to characterizing single units in deep network is called 'network centric' (Yosinski et al., 2014) and uses a process called activation

maximization. On this method, rather than present a set of meaningful images to a network and measure how individual units respond, the experimenter generates images that best drives the units. For instance, a random pattern (noise) might be presented to a network, and then input is systematically varied (though various algorithms) in order to generate images that drive a unit more strongly. If meaningful images are generated in this way it suggests that the unit selectively codes for this high-level visual information (Erhan, Bengio, Courville, Vincent, 2009; Le et al., 2013; Mahendran & Vedaldi, 2016; Nguyen, Dosovitskiy, Yosinski, Brox, & Clune, 2016; Nguyen, Yosinski, & Clune, 2015, 2016; Simonyan, Vedaldi, & Zisserman, 2014; Wei, Zhou, Torrabla, & Freeman, 2015; Yosinski et al., 2015).

Initial attempts at activation maximization suggested individual units code for information in a distributed manner given that the images that maximally drove individual units did not correspond to interpretable inputs (Simonyan et al., 2013; Nguyen et al., 2015; Szegedy et al., 2013; Goodfellow et al., 2014). Indeed, in some cases, deep networks would confidently categorize noise as a familiar category (e.g. labeling with near certainty that images that look like TV static – to humans – as a robin; Nguyen et al., 2015). This is not what one should expect from a grandmother unit. See Figure 2.

However, when various constraints on how images are generated are introduced so that the synthetic images share the general structure of natural images, then highly interpretable images emerge (e.g., Nguyen et al., 2015, 2016; Yosinski et al., 2015). For example, Yosinski et al. (2015) introduced a penalty for generating high contrast images (adjacent pixels that had very different values), and penalized images with pixels with large values (such that any new images included less "bright" regions), and penalized pixels with low values (such that any new images tended to have no activation in these regions). Note, none of constraints were designed to produce meaningful patterns, they just insured that the images that were generated shared some basic properties with natural images (e.g., objects in

the world tend to have similar levels of illumination at adjacent locations). Nevertheless strikingly meaningful patterns emerged, as seen in Figure 3. These are just the sort of images that would be expected on a grandmother theory of neural representation. Again, these interpretable images have been observed across a range of networks (convolutional networks as well as deep belief networks without any supervision) and under a range of training conditions. The learned representations revealed through activation maximization have been compared to grandmother cells (Le et al., 2013).

Although activation maximization techniques provide striking demonstrations that units respond selectively to meaningful inputs (e.g., images of rocking chairs) the fact that these same units also respond to very different patterns (that look like TV static) complicate the interpretation of the units. In some ways this parallels Waydo et al. (2006) analysis of the single-cell recording data collected by Quian Quiroga et al. (2005). Waydo et al. argued that even the most selective neurons (e.g., the Jennifer Aniston neuron) would respond to other categories of (untested) objects. Indeed, Waydo et al. estimated that if the experimenter could present all familiar categories of objects to the patient with the Jennifer Aniston neuron (images of all familiar people, places, and things), then between 50-150 other familiar items would drive this neuron. This reasoning lead the Waydo et al. (2006) to reject grandmother cells in favor of what they called sparse distributed coding (also see Quian Quiroga, Kreiman, Koch, & Fried, 2008). In the same way, a "data-centric" analysis of a deep network (analogues to single-cell recording studies) often reveal units that responds highly selectively to one category of object (e.g., images of motorcycles), but the activation maximization method shows that this same unit will respond to other images, just images that are unlikely to be tested (such as a specific image that looks like TV static). An interesting question is whether activation maximization findings support the conclusion that deep networks also learn sparse distributed rather than localist codes.

In fact, there is an important distinction that should be drawn between the Waydo et al. (2006) analysis of neural firing and the activation maximization findings observed in artificial networks. That is, in the Waydo analysis, all the highly selective neurons observed in humans are expected to fire to other familiar objects. It was just a limitation of time that prevented the experimenter from identifying the other relevant familiar objects that drive these neurons. By contrast, both data-centric and activation maximization findings show that selective units in artificial networks often represent only one familiar category of object in the universe of trained objects, or indeed, represent one familiar category of object amongst all possible categories that respect the visual structure of the world. It is only unfamiliar images that do not follow the statistical structure of the visual world that confuse these units. Given that localist representations and grandmother cells are theories about how familiar objects are coded (Bowers, 2009), the observation that these selective units also respond to specific examples of noise does not seem so relevant. In addition, as discussed below, Waydo et al.'s (2006) claim that selective neurons fire to multiple familiar objects is also consistent with localist coding schemes (see Gubian, Davis, Adelman, and Bowers, this issue).

Still, what is to be made regarding the observation that selective units in deep networks often respond strongly to both one meaningful category as well as some specific images that look like meaningless noise? One possible response is that the same may apply to humans. Indeed, some researchers have advanced the argument that mistakenly identifying an unnatural image as familiar objects in a deep networks is analogous to other illusions that humans clearly do experience (e.g., Kriegeskorte, 2015). It is just in practice difficult to find images that would fool humans in the same way that images confuse artificial networks. It is important to note, however, that even if we accept this, it does not challenge the grandmother cell hypothesis. For example, consider a cell that responds to one person amongst all known and unknown possible people (e.g., a unit that only responds to Jennifer

Aniston after testing the individual with all possible human faces) as well as some specific patterns of TV static. It seems reasonable to call this a grandmother cell given that it selectively codes for one familiar person, and it is very unlikely fire to anything else in the person's lifetime.

Another response to this finding is that it reflects a deep disconnect between how deep networks and the human visual system operate. This in turn may lead to the conclusion that single-unit recordings from deep networks (or perhaps all artificial networks) are just too different to be meaningfully related to theory psychology and neuroscience. This indeed is a serious concern (a concern more general than the localist/distributed issue, but of neural network modelling in general). Another possible conclusion, however, is that the important differences between brains and deep networks do not undermine the more general inferences that can be drawn. Indeed, a key feature of the PDP and deep neural networks above is that they rely on learning algorithms that are not designed to learn distributed or selective codes. Rather, the models learn the representations that are best suited for a given task. As Plaut and McClelland (2000) put it, PDP networks "discover representations that are effective in solving tasks…" and this "provides more insight into why cognitive and neural systems are organized the way they are". (p. 489). On this logic (which I accept) the observation of localist coding in PDP networks across some but not all conditions may provide some insight into what sorts of problems are best solved by localist codes, with implications not only for PDP networks, but perhaps for real brains as well.

Single unit recordings of alternative neural networks. I should also note that a number of alternative neural networks also learn highly selective representations, including Adaptive Resonance Theory (ART) networks (e.g., Carpenter & Grossberg, 1986), and spiking neural network models that rely on spike timing dependent plasticity (Masquelier & Thorpe, 2007). I will not consider these findings here because the networks include built-in

processes that contribute to the development of localist representations. As a consequence, the models do no provide independent evidence that localist codes are the best suited for solving the problems they faced.

It is worth noting, however, that the observation that PDP and deep networks often learn localist codes suggest that there are good reasons to build in processes that result in more selective coding. It is also the case that ART and spiking neural network models that rely on spike timing dependent plasticity are more biologically plausible than the PDP and deep networks described earlier. Furthermore, the learned localist representations in these networks are crucial to the functioning of these models, including the capacity of these networks to learn quickly without suffering forgetting – the stability-plasticity dilemma (Grossberg, 1980). So these models again highlight the biologically plausibility of grandmother cells.

Conclusions from simulation studies:

In summary, the single unit recording studies carried out on PDP networks within psychology and on deep networks within computer science have different motivations, but both sets of results demonstrate that single units respond highly selectively to meaningful inputs. Unlike in neuroscience it is possible to more systematically explore a vast number of images of familiar objects, persons, and scenes, and indeed, explore the space of possible images, in order to get a better idea of what single units represent. It is clear that single units sometimes selectively represent meaningful categories when tested against the universe of familiar, as well as unfamiliar but possible items. The ubiquity of localist coding across all these networks and conditions should give pause to cognitive psychologists and neuroscientist who dismiss localist coding (or grandmother cells) as implausible, and suggest reasons as to why brains may adopt similar strategies.

Summary of Chapters

Given the recent excitement regarding the high selectivity of single-cells in hippocampus and cortex, and the recent explosion of research showing selective codes emerge in artificial neural networks, it is a good time to explore the thoughts of leading theorists in psychology and neuroscience on the controversial topic of localist and grandmother cell representations. I'm very pleased that so many authors focused on terminological issues that have led to so much confusion in the literature.

I've organized the contributions as follows. I start with theorists that have emphasized a neuroscience perspective (Krieman; Riesenhuber & Glezer; Grossberg; and Rolls), followed by psychologists associated with the PDP perspective (Rodny, Shea, & Kello; and Thomas & French), followed by psychologists associated with the localist perspective (Coltheart; Hummel; Page; Gubian, Davis, Adelman, & Bowers; and Vankov & Bowers). Some of the assignments of people to research areas are a bit arbitrary and could have been assigned differently, but in any case, together, the articles provide an excellent summary of thinking on this topic from a range of perspectives.

Neuroscience perspective:

Krieman: Krieman describes 10 characteristics of neural representations (whether localist or distributed), and then summarizes a range of findings in visual and non-visual systems in order characterize representations as localist or distributed. His conclusion is that grandmother representations are found throughout the brain, from low-level visual systems (retinal ganglion cells) to high-level visual system (inferotemporal cortex), as well as memory systems (hippocampus) and systems involved in interpreting inputs and decision making (frontal cortex). In his words, "grandmothers galore". For example, Krieman characterizes simple cells that respond maximally to line segments at a given angle at different locations as grandmother cells. He also emphasizes that these grandmother simple

cells pool to grandmother complex cells, and argues that a collection simple cells constitute a distributed representations of the more complex information coded in complex cells. The claim that co-active local representations at one level of a system are also part of a distributed representation at another level is common (Hummel, this issue; Page, this issue; Plaut & McClelland, 2010).[1]

One point I would like to emphasize here is that Kreiman is adopting the view that grandmother cells in neuroscience are similar to localist codes in psychology. It is striking how this helps clarify issues between psychology and neuroscience. In their commentary of Bowers (2009) paper, Quian Quiroga and Kreiman (2010) were critical of my claim that grandmother cells are biologically plausible, but the disagreements between Kreiman and myself are largely resolved in this contribution, and I think this largely reflects using the same terminology rather than fundamental change of position.

Riesenhuber and Glezer: Riesenhumber and Glezer describes new analyses from a previous study (Glezer, Jiang, & Riesenhuber, 2009) that employed the fMRI rapid

[1] By contrast, I have argued that a collection of local codes that map onto more complex (local) code does not constitute a distributed code. For example, I disagree that co-activating the letter representations D-O-G in the IA model constitutes a distributed representation of the word DOG given that the co-active letter codes do not encode the fact that DOG is a word (it only encodes the fact that three letters are co-active; for more detail, see Bowers, 2009, 2010). Although there is clearly disagreement about what constitutes a localist vs. distributed representation (even amongst advocates of localist coding), one thing should be clear; namely, the co-activation of multiple localist codes is very different from the distributed representations proposed by proponents of PDP models. According to PDP theorists, all cognitive content, including letters, is coded in a distributed manner in which each unit is involved in coding multiple things (see above quote by Rogers and McClelland, 2014). So if the a collection of simple grandmother cells in V1 or localist letter detectors in the IA model are described as part of a distributed representation that map onto higher levels of representations (e.g., the collection of localist letter codes D-O-G constitute a distributed representation of the word DOG), it is important to distinguish between distributed representation composed of meaningful localist units and distributed representations composed in meaningless units. Hummel (this issue) calls the latter form of representation "deep distributed".

adaptation technique (fMRI-RA) to examine the nature word representations in the visual word form area (VWFA). In the previous study the authors reported data suggesting that the VWFA contained neurons tuned to whole words (localist representation) rather than pre-lexical letter combinations, and they took their findings to be consistent with localist models of reading. In the present study, the Riesenhumber and Glezer reanalyzed the Glezer et al. (2009) results to see how quickly localist representations develop for newly learned words (novel words were repeated in this study). The striking finding was that they obtained evidence for newly acquired localist word representations in the VWFA following just 5-6 exposures in a single day.

The evidence for newly acquired localist representations in the VWFA following a few representations is theoretically significant because it is commonly argued that consolidation, a time consuming processes, is necessary before new information is added to the cortex. Indeed, according the complementary learning systems hypothesis, information is coded in a highly selective (although not localist) manner in the hippocampus, and that this information is slowly transferred to the cortex in a more distributed manner (McClelland et al., 1995). This novel finding poses a challenge to this theatrical approach, and indeed, adds to the arguments put forward Page (this issue) against the complementary learning systems hypothesis.

Grossberg: Grossberg provides a summary of his modelling approach, with specific emphasis on visual object identification and how sequences of items can be stored in short-term memory. Grossberg also supports the hypothesis that there are multiple levels of grandmother cells, and that co-active grandmother cells at a lower level of a hierarchy constitute a distributed representation for higher level knowledge (like Krieman and others). However he also introduces another term that I think is quite apt, namely, a “grandmother cohort”. As I detailed in Footnote 1, it is important to distinguish between the view that all

units are meaningless when considered in isolation (the standard definition of distributed coding in PDP modelling; see Rogers & McClelland, 2014) and a collection of meaningful units that map onto more complex localist representation. The term Grandmother cohort captures the former position quite elegantly I think.

Grossberg also makes an important distinction between localist representations that code for specific categories (Grandma Leitner!) and more abstract ones (e.g., a generic person). He notes that ART can accommodate both forms of localist coding, and that a vigilance parameter in his model determines the granularity of learned localist representation (the issue of specificity of grandmother cells is a central point of the Coltheart contribution, discussed briefly below). In addition, he notes that the dynamics of the competitive processes in his network can vary, with winner-take-all dynamics (typical in most of his modelling) or less severe competition that lead to distributed coding (e.g., Carpenter, 1997). Accordingly, in addition to grandmother cells and grandmother cohorts, more traditional forms of distributed coding can also develop in his networks.

Rolls. In contrast with the above authors, Rolls is critical of grandmother cell theories. According to Rolls, grandmother cell theory is committed to the view that each visual experience is coded by a single neuron that does not fire to anything else. He dismisses this view because there are not enough neurons (or synapses) to code for all the possible visual experiences, because it is implausible to suggest that a lesion of a single neuron would lead to the selective loss of knowledge, and because grandmother cells cannot generalize. Rolls also reviews a number of single-cell recording studies that he takes as inconsistent with grandmother cells. For example, he describes a study by Rolls and Tovee (1995) in which the authors reported the responses of 14 neurons in the superior temporal sulcus in 2 monkeys in response to 68 stimuli (23 face and 45 nonface). The critical finding was that the average sparseness was 0.65 (meaning that the average neuron responded to 65%

of the images[2]). Rolls also notes that neuron firing is probably much more selective than this once correcting the spontaneous firing of the 14 neurons and given that the .65 value is based only on the neurons that responded to one or more of the images. As Rolls (this issue) write:

> "There were many more neurons that had no response to the stimuli. At least 10 times the number of inferior temporal cortex neurons had no responses to this set of 68 stimuli. So the true sparseness would be much lower…".

Nevertheless, Rolls considers the results inconsistent with grandmother cell theories that predict much more selective responding (according to Rolls, a grandmother cell coding scheme would predict a selectivity value of 1/68 for these images).

Another finding that Rolls takes to be inconsistent with grandmother cells is based on an analysis of the encoding of information by multiple cells (Rolls, Treves, & Tovee, 1997). On this analysis, grandmother cell theories predict that the number of stimuli that can be represented by a population of neurons rises approximately linearly with the number of neurons, whereas with distributed encoding, the number of stimuli that can be represented should rise exponentially. The results of the Rolls et al. (1997) study is claimed to support the predictions of distributed coding and to be inconsistent with grandmother cell theories.

However, it is important to note that these analysis are only relevant to falsifying the extreme version of grandmother cells that Rolls describes. As detailed by Gubian et al. (this issue) and Page (this issue), localist models can account for the levels of selectivity reported Waydo et al. (2006), and can explain the number of stimuli that can be represented and identified by a population of neurons.

[2] As is common in the neuroscience literature, Rolls uses the term "sparseness" to measure the selectivity of a single neuron rather than as a measure of the proportion of neurons that fire in response to a single image. One of the common confusions in translating neuroscience to psychology is that the terms selectivity and sparseness are used differently in the different literatures (see Bowers 2011).

Theorists from psychology associated with the PDP perspective:

It is unfortunate that many theorists associated with the PDP perspective declined to contribute, but I am pleased that Rodny, Shea, and Kello (this issue) as well as Thomas and French (this issue) have contributed. Interestingly, the authors take very different views from one another as well as from all other contributors.

Rodny, Shea, and Kello: Rodny *et al.* challenge an assumption shared by theorists from both the localist and distributed perspectives, namely, that knowledge representations are stable over time (indeed, stability over time is one of the definitions of a representation provided by Kreiman, this issue). The authors summarize evidence that "…representations continually shift and change, even on relatively fast timescales, and even after learning has stabilized", and describe a spiking neural network that learns localist representations that change over a wide range of timescales. This raises a new potential challenge for understanding the representations that support perception and cognition (a complication that applies to both localist and distributed theories), and it will be interesting to see future developments on this fundamental claim. Most relevant to the current topic, it is interesting that the authors are endorsing a form of (transient) localist coding. It is perhaps worth noting that although Kello is associated with the PDP framework, he has for some time been sympathetic to the view that knowledge is coded in a localist format (e.g., Kello, 2006).

Thomas and French: Unlike Kello, Thomas and French have been long-term critics of localist/grandmother schemes, and accordingly, it is interesting to note that Thomas and French do not reject the hypothesis that grandmother cells exist. However, they do question whether these cells are functionally significant. On their view, neuron that selectively responds to an image of a grandmother "have very little, or no impact on the actual recognition of my grandmother". They detail two scenarios in which grandmother cells might develop, but where distributed representations in fact do all the work. On this view,

grandmother cells are not important topic to explore, and thus the title of their paper: "Grandmother cells: Much ado about nothing".

In support of this view, the authors note that there are no neuropsychological syndromes I which a patient fails to identify a particular person, and describe a study that found non-selective neurons played an important role in categorizing visual stimuli (Thomas, Van Hulle, & Vogels, 2002). With regards to this later point, Thomas et al. (2002) used a Kohonen network to categorize the response of 219 neurons in the inferior temporal cortex of a monkey trained to categorize photographs as tree vs. non-trees (the neural responses were taken from Vogels, 1999). The model was able use these signals as inputs in order to categorize the in inputs quite accurately (83%). They then removed the inputs from neurons that where more or less selective (none of the neurons were completely selective), and found that the more selective neurons did not contribute more to the performance of the network. Indeed, they found that that the input from the less selective neurons was more critical in supporting performance. They took these findings as evidence that the selective responses of neurons often reported in the literature are not functionally relevant, and the important computations are performed by distributed codes.

Theorists from psychology supportive of localist coding:

Coltheart: Coltheart starts with a brief historical review of the concept of grandmother cells, dating back to a course taught at MIT by Jerome Lettvin in 1969, and then highlights many of the conceptual confusions regarding this hypothesis over the years. Like other contributors to this issue, Coltheart emphasizes that the grandmother cell hypothesis is not committed to the claim neurons fire to one thing and nothing else, nor the idea that there is a single neuron with no redundancy. But a novel point that Coltheart makes is that it is important to distinguish between grandmother cells and gnostic units. On this view, grandmother cells selectively code for specific items (my grandmother's face, my hand, that

dog), whereas gnostic units selectively code for general categories (grandmother faces in general, cars in general, dogs in general). According to Coltheart, both types of representations are localist, but a failure to distinguish between levels of abstraction lead to confusions. Indeed, according to Coltheart, the evidence for gnostic units is strong, but the evidence for grandmother cells within the visual system is weak.

Note, there is no reason to think that qualitatively different processes need to be involved in learning and representing grandmother and gnostic units. For example, Grossberg (this issue) also contrasted localist units that code for quite specific inputs (Grandma Leitner!) and more abstract representations (general or prototypical grandmother face), and argued that the different levels of abstraction can be explained with different setting of a vigilance parameter in ART models (so that both grandmother and gnostic units might be coded at the same level of the hierarchy of network). Another possibility is that grandmother units are at level *n*-1 of a hierarchy that pool onto gnostic units at level *n*, such that grandmother units of JOHN, BILL, JANE, SUE, etc. all map onto a common generic representation of person.

I agree with Coltheart that many single-cell recording studies in the visual system provide evidence for gnostic rather than grandmother units (e.g., neurons that respond to hands in general; e.g., Gross, Bender, & Roch-Miranda, 1969), but it should also be noted that highly selective responses to specific items have also been observed in cortex. Perhaps one of the more striking examples was reported by Logothetis, Pauls, and Poggio (1995) who trained two rhesus monkeys to identify a large set of novel computer-generated objects that were highly similar to one another. The most common selective response (by far) was to a specific object in a given orientation, with a smaller number of neurons responding to a given object across orientations (both types of representations would appear to be grandmother unit by Coltheart's definition). Bowers (2009) reviews a variety of additional results that would

seem more consistent with grandmother rather than gnostic units. Nevertheless, it is a distinction worth making, and the distinction becomes all the more important if it turns out that future work shows that neurons in the cortex tend to respond selectively at the category level but not the item level.

Hummel: Hummel provides a detailed analysis of the defining attributes localist and distributed coding and their relative advantages. He notes the terms localist and distributed coding terms are only meaningful with respect to that which is being represented. Every hidden unit in a distributed coding scheme will maximally fire to an input (or a set of inputs), but if the input(s) is meaningless, it does not constitute a localist code (just like you would not argue that a unit in a deep network that responds strongly to a noise that looks like TV static is a localist representation for this specific form of noise). Localist codes represent meaningful things. Hummel also contrasts "deeply distributed" representations in which all units are uninterpretable (the standard theoretical position of PDP modelers; see Rogers and McClelland, 2014, quote above), with distributed representations composed of a pattern of activation over localist units. Again, I think it is a mistake to call co-active local codes a distributed representation, and much prefer Grossberg's term 'grandmother cohort'. But "deeply distributed" seems an excellent term to describe the view that all cognitive content (e.g., letters, phonemes, words, objects, etc.) is coded in a distributed format.

Hummel argues that both localist and distributed representations have their place, and their relative merits depend on the goals of the computation to be performed. A key point that is not raised by any of the other contributors is that localist representations may be best suited for supporting symbolic computations. The long-standing localist/distributed debate needs to be distinguished from the long-standing symbolic/non-symbolic debate, and indeed, localist models can be either be symbolic or non-symbolic. For example, the Spatial Coding Model includes local symbolic letter codes (Davis, 2010) whereas the IA model of word

identification includes local but non-symbolic letter codes (McClelland & Rumelhart, 1981). Although most models with localist representations do not support symbolic computation, Hummel argues that symbolic models require localist codes.

Page: Page makes two quite different points. First, he challenges the common conclusion drawn from an influential study by Rolls and colleagues (Rolls, Treves, & Tovee, 1997; repeated by Rolls, this issue). As noted above, Rolls et al. (1997) recorded from 14 neurons in superior temporal sulcus that responded more strongly to face compared to non-face stimuli. The critical observation was that the ability to identify a specific face from the pattern of neural responses was exponentially related to the number of neurons from which they measured, with more neural responses associated with better accuracy (single neurons did a poor job). This was taken to be a signature of districted coding. However, Page shows that a localist model of face identification can capture these data as well, and concludes that the findings cannot be used to support distributed compared to localist coding. Whether the findings are inconsistent with grandmother cells (as claimed by Rolls, this issue) depends on how grandmother cells are defined.

Second, Page describes some objections to the complementary learning hypothesis that is frequently used to explain why distributed representations are found in cortex (a topic also considered by Riesenhuber and Glezer this issue; Grossberg, this issue). As noted by Page (also see Bowers et al., 2016) many of the original claims motivating the complementary learning systems no longer hold, including the claim that localist codes are poorly suited for generalization. But Page raises a more fundamental objection, claiming there is a logical problem regarding how the interleaving learning of new patterns (from the hippocampus) and old patterns (already stored in the cortex) can be achieved by a system employing gradient-descent learning. I look forward to future discussions on this topic.

Gubian, Davis, Adelman, and Bowers: Gubian *et al*. adopt a similar approach to Page (this issue) in that they challenge the interpretation of an influential finding taken to falsify grandmother cells. Quan Quiroga et al. (2005) reported neurons in the hippocampus and related structures that responded highly selectively to images of objects, persons, and places (e.g., the Jennifer Aniston neuron), and in a subsequent analysis of these results, Waydo et al. (2006) concluded that the average selectivity of these neurons was approximately .5%, meaning that these neurons responds to about .5% of presented images, and that each neuron responds to between 50-150 different categories of image. These findings were taken as evidence in support of sparse distributed coding and inconsistent with grandmother cell coding. The key finding by Gubian et al. is that a localist model of visual word identification can also explain this level of selectivity, and accordingly, the findings should not be used to reject localist representations. Again, whether these findings are inconsistent with grandmother cells depends, on how grandmother cells are defined.

Vankov and Bowers: Finally, Vankov and Bowers explored some of the factors that contribute to localist coding in simple feed-forward PDP models when items are trained one-at-a-time. As reviewed above, Bowers et al., (2014, 2016) found that recurrent PDP models learn localist representations when trained to co-activate multiple items at the same time in short-term memory, but learned distributed representations when trained on the same items one-at-time. At the same time, a number of PDP and deep networks learn localist representations when trained on images one-at-a-time. Why the different results? We looked into whether learning arbitrary input-output mappings (characteristic of mappings learned the deep networks) provides a pressure to learn localist representations.

Our main finding is the PDP models succeeded with arbitrary input-output mappings using distributed codes in many of the conditions we tested, but we did find localist codes under the conditions in which deep networks learn localist codes, namely, when learning to

map multiple images of an object (in this case faces) to a single output when the input units took on continuous rather than binary values. Clearly these findings highlight that localist codes are adaptive in PDP networks under some conditions and not others, and future work is required to better characterize what exactly is the pressure to learn localist codes when processing items one-item-at-a-time. Based on our results, it is clear that arbitrary input-output mappings do not provide as strong a strong pressure to learn localist codes as does co-activating multiple items at the same time

Overall Summary:

The term 'grandmother cell' is often defined in different ways, and accordingly, it is not always clear what theories are challenged when a researcher rejects grandmother cells. What is clear is that there is a large literature in neuroscience highlighting the extreme selectivity of some neurons in cortex and hippocampus, and a growing literature of single-unit recording studies in artificial neural networks that also report highly selective (localist) representations. If grandmother cells are defined as localist representation in psychological models, then grandmother cells cannot be dismissed out of hand, and indeed, good arguments can be put forward in support of grandmother cells of this sort. The articles in this special issue show that many researchers consider grandmother cells a serious hypothesis about how knowledge is coded, as well as highlight key disagreements and issues that need to be addressed in future work. If nothing else, I hope this special issue contributes to a more productive debate on an important issue that has often been characterized by misunderstandings between disciplines.

References:

Agrawal, P., Girshick, R., & Malik, J. (2014, September). Analyzing the performance of multilayer neural networks for object recognition. In *European Conference on Computer Vision* (pp. 329-344). Springer International Publishing. doi: 10.1007/978-3-319-10584-0_22

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

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

Bechtel, W., & Abrahamsen, A. (1991). Connectionism and the mind*: An introduction to parallel processing in networks.* Basil Blackwell.

Berkeley, I. S. N. (2006). Moving the goal posts: A reply to Dawson and Piercey. Minds and Machines, 16, 471–478. Doi: 10.1007/s11023-006-9048-9

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

Berkeley, I. S., & Gunay, C. (2004). Conducting banding analysis with trained networks of sigmoid units. *Connection Science, 16,* 119-128. doi: 10.1080/09540090412331282278

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113,* 201–233. doi: 10.1037/0033-295X.113.2.201

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. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251. doi: 10.1037/a0014462

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 Quiroga and Kreiman (2010). *Psychological Review, 117,* 300–308. Doi:10.1037/a0018047

Bowers, J. S. (2011). What is a grandmother cell? And how would you know if you found one? *Connection Science, 23,* 91-95. Doi: 10.1080/09540091.2011.568608

Bowers, J.S. and Davis, C.J. (2012) Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin, 138,* 389-414. doi: 10.1037/a0026450

Bowers, J.S. Damian, M.F., Vankov, I., & Davis, C.J. (2012) Why Do Neurons in Cortex Respond to Information in Such a Selective Way? Talk presented at the Psychonomic Society, Minneapolis, Minnesota.

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2014). Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychological Review, 121,* 248-261. doi: 10.1037/a0035943

Bowers, J. S., Vankov, I. I., Damian, M. F., & Davis, C. J. (2016). Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. *Cognition, 148,* 47-63. doi: 10.1016/j.cognition.2015.12.009

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.

Carpenter, G. A. (1997). Distributed learning, recognition, and prediction by ART and ARTMAP neural networks. *Neural Networks, 10*, 1473–1494. doi: 10.1016/S0893-6080(97)00004-X

Carpenter, G. A., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer vision, graphics, and image processing, 37,* 54-115.

Chomsky, N. (1965). *Aspects of the theory of syntax. Cambridge*, MA: MIT Press.

Ciresan, D. C., Meier, U., Gambardella, L. M., & Schmidhuber, J. (2010). Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, *22*, 3207-3220. doi: 10.1162/NECO_a_00052

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. doi: 10.1037/0033-295X.108.1.204

Davis, C. J. (2010). The spatial coding model of visual word identification. *Psychological Review, 117*, 713-758. Doi: 10.1037/a0019738

Dawson, M. R., & Piercey, C. D. (2001). On the subsymbolic nature of a PDP architecture that uses a nonmonotonic activation function. *Minds and Machines, 11(2*), 197-218. doi:10.1023/A:1011237306312

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

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network*. University of Montreal, 1341.

Glezer, L. S., Jiang, X., & Riesenhuber, M. (2009). Evidence for highly selective neuronal tuning to whole words in the "visual word form area". Neuron, 62(2), 199-204. doi: 10.1016/j.neuron.2009.03.017

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

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.

Gross, C. G. (2002). The genealogy of the "grandmother cell." *The Neuroscientist, 8*, 512–518. DOI: 10.1177/107385802237175

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

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... & Ng, A. Y. (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Ison, M., Mormann, F., Cerf, M., Koch, C., Fried, I., Quian Quiroga, R., 2011. Selectivity of pyramidal cells and interneurons in the Human Medial Temporal Lobe. *Journal of Neurophysiology,* 106, 1713–1721. DOI: 10.1152/jn.00576.2010

Ison, M. J., Quian Quiroga, R., & Fried, I. (2015). Rapid encoding of new memories by individual neurons in the human brain. *Neuron, 87(1),* 220-230. doi: 10.1016/j.neuron.2015.06.016

Leighton, J. P., & Dawson, M. R. (2001). A parallel distributed processing model of Wason's selection task. *Cognitive Systems Research, 2(3),* 207-231. doi: 10.1016/S1389-0417(01)00035-3

Kello, C. T. (2006). Considering the junction model of lexical processing. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in lexical processing* (pp. 50–75). New York: Psychology Press.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science, 1*, 417–446. doi: 10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

Le, Q. V. (2013, May). Building high-level features using large scale unsupervised learning. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595-8598). IEEE.

Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2015). Convergent Learning: Do different neural networks learn the same representations?. *arXiv preprint arXiv:1511.07543*.

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology, 5(5),* 552-563. doi: 10.1016/S0960-9822(95)00108-4

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

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science, 2(6),* 387-395.

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.

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision,* 1-23. doi: 10.1007/s11263-016-0911-8

Malach, R. (2012). Neuronal reflections and subjective awareness. In S Edelman, T Fekete, and N Zach (Eds). *Being in Time: Dynamical models of phenomenal experience* (p. 21-36). John Benjamins Publishing Company, Amsterdam/Philadelphia.

Masquelier, T., R. Guyonneau and S. J. Thorpe (2009) Competitive STDP-Based Spike Pattern Learning. *Neural Computation 21 (5*) 1259-1276

Marr, D. (1971). Simple memory: A theory for archicortex. Philosophical Transactions of the *Royal Society of London, Series B: Biological Sciences, 262,* 23–81.

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

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature, 518(7540),* 529-533. doi:10.1038/nature14236

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

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *arXiv preprint arXiv:1605.09304.*

Nguyen, A., Yosinski, J., & Clune, J. (2015, June). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 427-436). IEEE.

Nguyen, A., Yosinski, J., & Clune, J. (2016). Multifaceted Feature Visualization: Uncovering the Different Types of Features Learned By Each Neuron in Deep Neural Networks. *arXiv preprint arXiv:1602.03616.*

Niklasson, L. F., & Boden, M. (1997). Representing structure and structured representations. *Neural Network Perspectives on Cognition and Adaptive Robotics*, Institute of Physics Press, Bristol, UK.

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

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.

Plaut, D. C., & McClelland, J. L. (2000). Stipulating versus discovering representations. *Behavioral and Brain Sciences, 23,* 489–491. doi: 10.1017/S0140525X00473358

Plaut, D.C., McClelland, J.L., (2010). Locating object knowledge in the brain: comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review 117*, 284–290. doi: 10.1037/a0017101

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. doi: 10.1016/j.tics.2007.12.003

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature, 435(7045),* 1102-1107. doi:10.1038/nature03687

Quian Quiroga, R. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience, 13(8),* 587-597. doi:10.1038/nrn325

Quian Quiroga, R. (2016). Neuronal codes for visual perception and memory. *Neuropsychologia, 83,* 227-241. doi: 10.1016/j.neuropsychologia.2015.12.016

Quian Quiroga, R., Fried, I., & Koch, C. (2013). Brain cells for grandmother. *Scientific American, 308(2),* 30-35. doi:10.1038/scientificamerican0213-30

Quian Quiroga, R., Kreiman, G., 2010. Measuring sparseness in the brain: comment on Bowers (2009). *Psychological Review, 117*, 291–297. doi: 10.1037/a0016917

Rey, H., Ison, M., Pedreira, C., Valentin, A., Alarcon, G., Selway, R., Richardson, M., Quian Quiroga, R., 2015. Single cell recordings in the human medial temporal lobe. *Journal of Anatomy, 227,* 394–408. doi: 10.1111/joa.12228

Rogers, T.T. and McClelland, J.L. (2014) Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science 38,* 1024-1077 doi: 10.1111/cogs.12148

Rolls, E. T., & Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *Journal of Neurophysiology, 73(2),* 713-726

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.

Roy, A. (2012). A theory of the brain: localist representation is used widely in the brain. *Front Psychology, 551*. doi:10.3389/fpsyg.2012.00551.

Roy, A. (2015). On findings of category and other concept cells in the brain: Some theoretical perspectives on mental representation. *Cognitive Computation, 7(3),* 279-284. doi: 10.1007/s12559-014-9307-7

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

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). New York: Psychology Press.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and brain sciences, 11(01),* 1-23.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199.*

Thomas, E., Van Hulle, M., & Vogels, R. (2002). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience, 13*, 190–200. doi:10.1162/089892901564252

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

Thomas, E., Van Hulle, M., & Vogels, R. (2002). Encoding of categories by noncategory-specific neurons in the inferior temporal cortex. *Journal of Cognitive Neuroscience, 13*, 190–200. doi:10.1162/089892901564252

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 1: Behavioral study. *European Journal of Neuroscience, 11*, 1223–1238. doi:10.1046/j.1460-9568.1999. 00530.x

Von Der Malsburg, C. (1986). Am I thinking assemblies?. *In Brain theory* (pp. 161-176). Springer Berlin Heidelberg. Chicago.

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. doi: http://dx.doi.org/10.1523/JNEUROSCI.2101-06.2006

Wei, D., Zhou, B., Torrabla, A., & Freeman, W. (2015). Understanding Intra-Class Knowledge Inside CNN. *arXiv preprint arXiv:1507.02379.*

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579.*

Yuste, R. (2015). From the neuron doctrine to neural networks. *Nature Reviews Neuroscience, 16(8),* 487-497. doi:10.1038/nrn3962

Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818-833). Springer International Publishing.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2014). Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856.*

Figure Captions:

Figure 1: Number of citations to "grandmother cell" or "grandmother cells" per year averaged over decades (according to Google Scholar).

Figure 2: Eight examples of noisy images that are confidently categorized as a familiar object. Taken from Nguyen et al. (2015).

Figure 3: Synthesized images that best activate three different units from layer 8 of a deep convolutional network. Four different examples of the best synthesized images for each unit is presented. Clearly, these units are most activated by meaningful objects. Taken from Yosinski et al. (2015).

Figure 1.

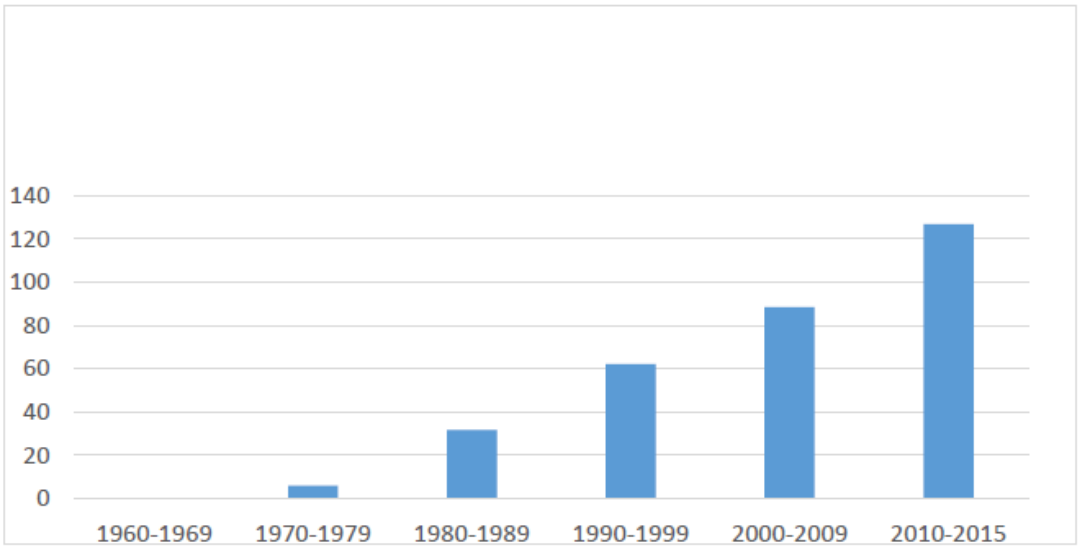
140
120
100
80
60
40
20
0
1960-1969
1970-1979
1980-1989
1990-1999
2000-2009
2010-2015

Figure 2:

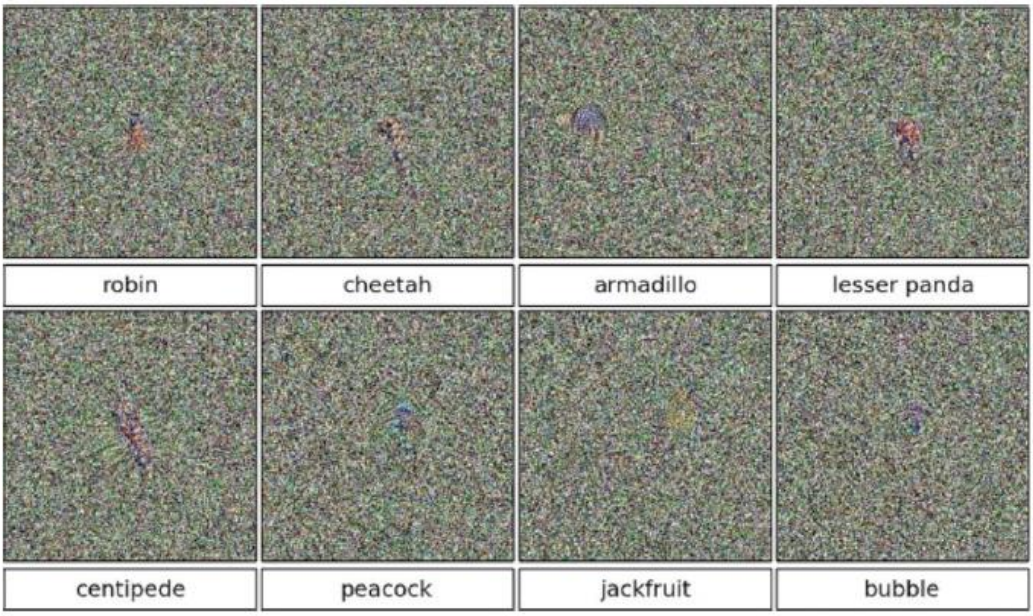
robin
cheetah
armadillo
lesser panda
centipede
peacock
jackfruit
bubble

Figure 3: