Localist models are compatible with information measures, sparseness indices and complementary learning systems in the brain

Dr Mike Page

Department of Psychology,

University of Hertfordshire

Hatfield, AL10 9AB, UK

m.2.page@herts.ac.uk

t: +44 1707 286465

Abstract

In this paper, I express continued support for localist modelling in psychology and critically evaluate previous studies that have sought to weaken the localist case in favour of models with thoroughgoing distributed representation. I question claims that information measures and sparseness indices derived from single-cell recording data are supportive of distributed representation and show that the patterns observed in those data can be reproduced from simulations of a model that is known to be localist. I also set out some logical objections to the complementary learning hypothesis, particularly in as much as it is used to justify thoroughgoing-distributed models of the cortex.

Keywords: Localist; neural networks

Localist models in psychology

In Page (2000a), and in the response to the accompanying peer commentaries (Page, 2000b), I set out in some considerable detail my support for localist modelling in psychology. In those papers, I tried to be clear about what a commitment to localist modelling actually comprised and, just as importantly, what it did not comprise. I argued that a commitment to localist modelling did not entail a rejection of distributed representation per se. The commitment was, instead, that including localist representations of certain entities or classes of entity in neural network models of psychological function – such as including localist representations of words in models of spoken word recognition - was both helpful and well-motivated. The opposing position, that I characterized as recommending "thoroughgoing distributed representation", was that localist representations were unnecessary and undesirable in models of various sorts. Very often this position was motivated by the observation that prominent techniques for implementing learning in neural networks, most notably the backpropagation (BP) learning rule and its variants, did not naturally result in localist representations at the critical levels of representation. As a specific example, the use of BP to train a three-layer neural network to map phonology to semantics, would not typically result in units at the hidden layer that were localist representations of particular words. This was the case even when network simulations were set up in such a way that the input layer at which, say, phonology was represented did itself use localist coding of phonemes or phonetic features. Any such coding choice for input patterns was usually explained away as a mere convenience, and everybody agreed to agree that the important thing about a thoroughgoing distributed model in this domain was that it didn't contain localist representations *of words*.

In this paper, I will not reiterate the arguments that I presented in Page (2000a). For what it is worth, there is nothing in that paper from which, sixteen years on, I resile, and the arguments presented there have been supported and extended more recently, notably by Bowers and colleagues in a number of significant articles (Bowers, 2009; Bowers, Vankov, Damian, & Davis, 2014) and by others (e.g., Coltheart, this issue). Instead, I will address several relevant issues that have been raised, or have come to prominence, since the publication of my earlier paper. The first of these, and the one on which I will expend most energy, concerns the claim that certain measurements taken from animal brains in vivo, are inconsistent with localist representation. Given the space available here, it is impossible to review and to discount all such claims individually. I will try to show, though, that at least one technically sophisticated and highly cited study of this kind is just plain wrong in the conclusions that it draws about neural coding. I will suggest that the falsity of those conclusions is likely to extend to many less sophisticated studies too. The second issue that I will discuss concerns the complementary-learning hypothesis, of McClelland, McNaughton and O'Reilly (1995). It seemed to me, even in Page (2000a), that there was a problem with the reasoning behind that hypothesis, a problem that I'm not sure has been resolved in the intervening years. For that reason, I will give the critical issue another airing.

Before doing any of this, it is worth specifying what I mean by a localist representation of a given entity or class of entities, and what, therefore, I mean by a localist model of a given psychological faculty. For the sake of consistency (and, frankly, convenience), I will simply adapt slightly the definition that I provided in Page (2000a): a localist model (e.g., of word processing) is one containing localist representations of a particular type of entity (e.g., words), where the learning or representation of entities of that type is a focus of theoretical interest. A model that is localist with respect to entities of a given type (e.g., words) is characterized by the presence of at least one node that responds maximally to a given familiar (i.e., learned) example of that entity-type (e.g., a given familiar word), all familiar examples of that type (e.g., all familiar words) being so represented (adapted from Page, 2000a, p.448).

This definition came at the end of several pages of build-up, that encompassed familiar (though oft misinterpreted) notions such as grandmother cells (see Coltheart, this issue), and more recondite concepts such as yellow-Volkswagen cells (Harris, 1980). The aim was to provide a workable and specific definition such as would render such terms as "semi-distributed" or "semi-localist" no more meaningful than the term "semi-pregnant". What is important, for current purposes, is that if one is going to posit local representations of members of a certain class of entities – to anticipate what is to come, let us switch from words to using the example of faces - then that will entail that for each face with which the system is properly familiar, there will at least one unit (a unit may be a cell, or a cell-assembly, or a minicolumn in vivo) that achieves its maximum activation on presentation of that familiar face, that maximal activation acting so as to indicate detection of that specific face in whatever sense is relevant to the task at hand. Note that there can be multiple local representations of the same face (i.e., multiple units all of which activate maximally in response to a given face). Indeed, such redundancy in coding would be a helpful way of avoiding what Coltheart, this issue, has called "graceless degradation", that is, the complete loss of response to a given familiar face subsequent to the loss of a single unit. It would not, however, be consistent with my definition to have a localist model of face recognition in which no such unit existed for a face that could properly be described as familiar and known. Armed with this definition, and having cast it in the domain of face recognition, I now turn to an influential study which has claimed, quite wrongly in my opinion, to have found single-cell-recording data contradicting the localist representation of faces in the temporal cortex of monkeys.

Rejecting the Conclusions of Abbott, Rolls and Tovee (1996)

In a paper that has since been cited over one hundred times, Abbott, Rolls and Tovee (1996) applied sophisticated information-theoretic techniques to the results of single-cell recordings taken from the superior temporal sulcus (STS) of rhesus macaque monkeys as they responded to images of monkey and human faces in the context of a visual fixation task. (The number of citations is closer to 300 if one adds citations of the closely related paper by Rolls, Treves & Tovee, 1997, that was based on the same data.) Specifically, recordings were made from 14 STS neurons that responded to faces at "more than twice the maximum rate evoked by any of 48 other nonface images" (Abbott et al., p. 498). After much detailed analysis, the authors concluded that the representation of faces in this area was "truly distributed". To allay any fear (expressed by an anonymous reviewer) that this conclusion has now become dated, I should point out that there are over 50 combined citations of these two papers since 2012, and 13 citations in 2016 alone. The papers' conclusions are cited approvingly by both Rolls (this issue) and Quiroga and Kreiman (2010), among others.

Faced with Abbott et al.'s (1996) conclusion, it would be possible to take Bowers' line regarding these data (Bowers, 2009, p.240) and claim that, although Abbott et al. found a distributed representation in this particular part of the monkey brain, this does not guarantee that some other, probably later, part of the processing hierarchy uses localist representation. The argument is valid, but somewhat unsatisfactory: presumably one could always claim that one's preferred style of coding could be found somewhere else. Besides, Rolls (2007) does consider that possibility and concludes that it is unlikely (see quotation below). Contrary to Bowers, therefore, my argument will assume that Abbott et al. were indeed measuring from exactly the part of the monkey cortex at which faces (including the stimulus faces they used in their experiment) were represented and that, as a consequence, their recordings speak directly to the issue of face representation more generally. Moreover, I will assume, without question, that Abbott et al. were entirely justified in the way in which they used the recordings of responses to 20 actual different stimulus faces to extrapolate to a much larger number of simulated stimuli (Abbott et al., p.500). Even allowing these things, I will claim that they were wrong to infer, from their data and from their analysis, that the representation of faces in macaque STS is distributed rather than localist.

The key finding of Abbott et al. (1996) was that the larger the number of STS cells from which they measured neural responses to faces, the larger was the number of stimulus faces that they could discriminate. Specifically, by measuring from a small number of cells (e.g., two or three cells), they could infer, to a given level of accuracy, which of up to around ten possible face stimuli had been presented. However, by looking at the response activity of a larger number of cells (e.g., 11, 12 or 13 cells) they could infer, to the same level of accuracy, which of a much larger number of stimuli (around 100 in their data) had been presented. Their argument - that this pattern of neural responses was indicative of a distributed neural code for faces, as opposed to a localist code - was based on the observation that the number of faces whose presence could be correctly inferred from the pattern of neural response, was exponentially related to the number of neurons from which they measured. Put slightly more technically, the amount of information (measured in bits) about which face stimulus had been presented, was approximately linearly related to the number of STS cells from which responses were measured. To quote from their Discussion:

"Taken together, [the data] provide strong evidence that the coding of face cells by temporal visual neurons in the macaque monkey is truly distributed. This results in an exponential dependence of the representational capacity on the number of cells…we estimate that N neurons can represent about $3(2^{0.4N})$ faces with 50% discrimination accuracy" (Abbott, et al., 1996, p.503)

We can be sure too that Abbott et al. (1996) wanted to contrast this "truly distributed" coding of faces with one that was more localist in flavour. From their introductory paragraph, we have:

"The amount of information that can be represented by the firing of a population of neurons depends on the nature of the neural code…If a stimulus is represented by the firing of a single neuron or "grandmother cell", the number of stimuli that can be represented is proportional to the number of neurons. If the information about each stimulus is distributed across the full population, the number of stimuli that can be represented grows exponentially with the number of coding neurons". (Abbott et al., 1996, p.498)

Given their finding of an exponential relationship between number of neurons and representational capacity (a finding that I will not dispute), it is clear on which side of the localist/distributed debate they alight. Although a full survey has not been possible, I would venture to guess that many of the several hundred papers that have cited this article and its companion (Rolls, et al., 1997) have, like those recent ones specifically cited above, accepted this broad conclusion at something like face value. In what follows, however, I will try to show that that would be a mistake. I will argue that the measurements that Abbott et al. (1996) made were perfectly consistent with a localist coding of faces, even across the very field of cells from which they measured. As a consequence, I will maintain that their data, hard-won and elegantly analysed as they were, give us no reason whatsoever to prefer a distributed coding (still less a "truly distributed" one) over a localist coding.

Careful readers will, perhaps, have already detected some mismatch in the language that I have used to talk about the alternatives to their "truly distributed" coding of faces, as contrasted with the language Abbott et al. (1996) used in the same relation. I have been careful to talk about "localist representation" (as defined above), whereas they gave something of a straw-man characterization of the localist alternative that they had in mind. Specifically, where they say of a "grandmother cell" that a stimulus is represented by the "firing of a single neuron", did they really believe that *that* was the localist alternative to fully distributed coding that others had proposed? Coltheart (this issue) gives a very interesting history of the grandmother-cell idea, but even in the murky history of that term (a term that refers only to a subset of the class of localist representations) it is far from clear that a grandmother-cell was ever conceived as a binary detector, activating *only and then fully* in response to one's grandmother, and *not at all* in response to anyone else (even your great-aunt). If that "extreme grandmother-cell" version of the localist position was the only target upon which Abbott et al. (1996) set their sights, then maybe they did succeed in knocking down their straw man (though they could have done so rather more simply by pointing out that the responses of the cells from which they measured were graded and not binary). I will, however, do them the credit of assuming that they were interested in knocking down more substantial alternatives too. What might the more substantial alternatives be? I suggest that the claims of Abbott et al., arguing for a "truly distributed" representation of faces in monkey STS cortex, could reasonably be taken to have argued against a localist coding of faces more generally, beyond the specific caricature of a binary grandmother-detector. Most readers will, I believe, have taken it to argue just that and Rolls (2007) himself later made this clear when he wrote:

"This direct neurophysiological evidence thus demonstrates that the encoding is distributed, and the responses are sufficiently independent and reliable, that the representational capacity increases exponentially with the number of neurons in the ensemble." (Rolls, 2007, p.128)

It is clear, then, that Rolls (2007) took the Abbott et al. (1996) data to argue against localist representation, and it is this more substantial claim that I dispute here. I will first set out the argument qualitatively.

As is consistent with the definition give above, a localist representation of faces across (let us say) an array of cells in monkey STS cortex, would entail that each cell would respond in a graded fashion to various possible stimuli. For a cell that has learned to represent a particular stimulus, the presentation of that particular stimulus would cause the cell to respond maximally. For other stimuli, that cell would respond less than maximally. As is typical in localist models, the identity of the maximally responding cell might enable identification of the stimulus, perhaps triggering later processes like naming. The crucial point, though, is that the less-than-maximal responses to various face stimuli other than the one that the particular cell represents, are not *random*. Most straightforwardly, these less-than-maximal responses are liable to reflect in some way the degree of similarity between the stimulus being presented and the learned stimulus that the given cell represents. For example, assuming that my grandmother and my great-aunt looked similar (they did), we would expect the monkey STS cell representing my great-aunt (if there were one) to activate significantly, but not maximally, to a picture of my grandmother. We would not, however, expect the cell representing my great-aunt to respond strongly to a picture of the late wrestler Giant Haystacks (I can confirm that any resemblance was mercifully slight). In this way, we might easily use the recorded activation of my great-aunt-cell, to indicate whether a given stimulus was a) a picture of my grandmother or (b) a picture of Giant Haystacks. Even though my great-aunt-cell does not code for either of the stimuli from which we must choose, it carries information about which of those stimuli is present. Moreover, because my great-aunt-cell will respond in a non-random way even to stimuli that it has never encountered before, this argument is not contingent on any previous such encounter.

It should now be clear, in outline at least, why the data of Abbott et. al. (1996) do nothing to rule out localist representation of faces in monkey STS cortex or, therefore, anywhere else. To spell it out, let us assume that it is very unlikely that any of the 14 cells from which Abbott et al. recorded just happened to be one of the cells that represented one of their 20 stimuli (in fact, it doesn't much matter for the purposes of my argument even if it did). We can assume, therefore, that the responses of the observed cells to the presented stimuli are non-maximal responses. They are, however, not *random* responses. Just as above, the response of each cell will reflect any similarity between its learned pattern and the current stimulus pattern and, as such, its activation will convey information (in both the informal and the formal sense) about which stimulus is present. Given that each cell presumably responds optimally to a different stimulus (this is not certain, given possible redundancy, but it is very likely given the sampling), they will each carry somewhat independent information about the stimulus identity. The more such cells we record from, therefore, the more information we will have regarding stimulus identity. This is exactly what Abbott et al. found. Taking a Bayesian stance (of which I'm sure they would approve), if their data are just as predictable from a localist perspective, as they are from a "truly distributed" perspective, then those data offer no evidential grounds for preferring a distributed coding over a localist one.

The argument that I have set out above is a verbal one but, given the mathematical sophistication of Abbott et al.'s (1996) presentation, it is not unreasonable to expect me to cash out this verbal argument in more quantitative terms. Otherwise, it might be suspected that the argument is only qualitative. In fact, notwithstanding their very detailed quantitative analysis of the single-cell recording data, the message that Abbott et al. drew was itself a fairly qualitative one. Specifically, they claimed that linear growth in the stimulus information available as we increase the number of cells from which we record, and a corresponding exponential growth in the number of different stimuli that can be classified to a given degree of accuracy, are, in and of themselves, diagnostic of distributed coding. If I can show, therefore, that these properties are equally true of a model with localist coding and, better still, if my quantitative informational measures turn out to be in the same ball-park as theirs, then Abbott et al.'s argument will fail.

To back up my verbal argument with a quantitative one, therefore, I performed a relatively simple simulation of a localist network for face recognition, subjected it to the sorts of measurements of "cell" activity that Abbott et al. (1996) had performed in vivo, and then derived from these simulated data the same informational measures as did they.

A Simple Simulation

In order to simulate a localist, face-recognition network, I used as stimuli a set of 110 patterns, each comprising a vector of 50 real numbers between -1 and +1. These patterns came from work by Calder, Burton, Miller, Young, and Akamatsu (2001) and were derived from a set of 110 gray-scale pictures, each with 10,000 pixels. The faces in the pictures had been normalized to a common face-shape by a process of morphing, and then the picture-set had been subject to a Principal Components Analysis to identify the fifty 10,000-dimensional components with the largest eigenvalues. Each face could then be represented by a 50-dimensional vector, the value on each dimension representing its coordinate along one of those 50 principal-component (eigenvector) directions. More details can be found in Calder et al.. The face-pictures that they used actually comprised a group of around 16 people posing each of seven canonical expressions and, as such, included inter-subject as well as inter-expression variation. While it is not particularly important for my argument that these 110 50-dimensional vectors are derived from actual faces, it does marry well with the fact that it was face processing that Abbott et al. (1996) themselves studied in vivo.

To model a localist network that could represent and "recognize" each of the 110 faces, I simulated a standard, fully connected, two-layer competitive network. The 50 units in the input layer permitted any one of the stimulus faces to be “presented” to the network, by setting the unit activations to the components of the relevant pattern. Each one of the 110 units in the second layer (the classification layer) was deemed to have learned a particular one of the 110 face-patterns. In the simulation, this simply comprised setting the 50 weights connecting a given classification-layer unit to its input units, to values corresponding to the 50-dimensional vector representing the face-pattern that it had been deemed to have learned. In other words, each of the classification-layer units had an incoming weight vector that was identical to exactly one of the patterns in the 110-item pattern-set. (The learning of such a network is trivially achieved by any of a number of competitive-learning networks, including that described in Page, 2000a.) On clamping a given test pattern to the input of the resulting network, the input to each of the 110 units in the classification layer was set equal to a peak value minus the Euclidean distance between the current input pattern and that unit’s learned pattern. This network therefore instantiates, by definition, a localist model: for each learned input pattern there is a single unit in the classification layer that responds maximally to that input pattern, the distance between that input pattern and the learned pattern encoded in its incoming weight vector being zero. Actually, from trial-to-trial the input to any given classification-layer unit was perturbed by zero-mean Gaussian noise with standard deviation equal to σ_1, to reflect neural noise in vivo. (Any simulation entirely devoid of trial-to-trial noise would not only be unrealistic, but each individual classification-layer activation would contain virtually perfect information regarding stimulus identity, unless, by chance, two or more stimulus patterns were to find themselves at exactly the same distance from a given learned pattern.)

During testing, once a given input pattern was presented to the learned network, and the inputs I_i to each of the classification-layer units had been calculated, the activations of the classification-layer units changed from their initial zero values at *t=0* according to the fairly standard competitive, “leaky integrator” equation:

$$da_i/dt = -0.2a_i - 0.4J_i + I_i + g(0, \sigma_2)$$

where: excitatory input I_i is defined as above as a peak value (here set to 2) minus the Euclidean distance between the current input pattern and the vector of weights connected to the i^{th} classification-layer unit plus zero-mean Gaussian noise of standard deviation equal to σ_1; J_i is an inhibitory (competitive) input to the i^{th} classification-layer unit, equal in magnitude to the maximum activation a_j such that $j \neq i$; and where $g(0, \sigma_2)$ is zero-mean Gaussian noise of standard deviation equal to σ_2. All of the parameter values other than the two noise parameters were inherited from a previous application of a similar network and none was optimized for the specific application here. The two noise parameters $\sigma_1$and σ_2 were set to 0.03 and 0.3 respectively, these values having been arrived at by first setting them to levels at which they both caused the network to misclassify a proportion on the stimulus patterns and then decreasing them until the network was once again capable of perfect classification. The classification of a given input pattern was given by the index of the first classification-level unit whose activation reached a criterion activation of 1. Perfect classification therefore entailed that the i^{th} classification-layer unit was the first of those units to reach that activation-criterion in response to the i^{th} stimulus pattern. The differential equation above was numerically integrated using a simple gradient calculation with time step equal to 0.001.

With the parameters set as above, and for each of ten separate trials, the learned network was presented with all 110 stimulus patterns and classification performance was observed to ensure that it was 100% correct. By design and by definition, therefore, this simple network is a localist network capable of perfect classification of each of 110 face-patterns. For each of 10 presentations of each of the 110 patterns, the mean firing rate for each of the classification-layer units was recorded. As in Abbott et al. (1996), who averaged firing rates between 100ms and 600ms following stimulus presentation, the mean firing rate in the simulations was taken as the mean activation of a given classification-layer unit across all time steps such that $0.1 < t < 0.6$. For any given combination of input stimulus and classification-layer unit, any variation in mean activation across the ten presentations of a given stimulus is a result of the two non-zero noise parameters.

Having simulated the mean firing rates, it then remained to calculate the various information measures that Abbott et al. (1996) themselves calculated for their (real and simulated) single-cell recordings. The details are relatively complex and are contained in their original paper. Nonetheless it was possible to use their procedures to produce two characteristic graphs. The first graph shows the information about the stimulus pattern (measured in bits) that can be derived from the set of cell activations, plotted against the size of the set of measured cells as it varies between one and fourteen (this upper limit being the actual number of cells from which they recorded). (Note that, to simulate the extreme unlikelihood of one of the 14 cells measured in vivo being the optimally tuned cell for one of the small number of stimuli presented, I ensured that the information relating to which one of a given set of stimuli was presented was, in my simulations, always calculated using the activations of cells other than that single cell tuned to the stimulus in question.) For Abbott et al.'s recordings, the graphs had a characteristic shape, starting fairly close to a straight line and tailing off somewhat below that straight line as the number of cells increased above eight or nine. They produced two measures: a raw informational measure and a cross-validated one, and plotted each against the numbers of cells measured. (The reason for including a cross-validated measure is relatively technical and is described in Abbott et al. p.500. Briefly, it is to correct for the fact that in the raw calculation of information, one is using the same probability distribution both to generate sample firing rates and to calculate the likelihood of a given sample stimulus's having produced them.) Their measure of raw information started off fairly close to a line representing an increase in information of 0.5 bits per additional cell, while their measure of cross-validated information started off close to a less steep line with gradient around 0.35 bits per cell. Their second graph illustrated for different numbers of measured cells, the number of different stimuli that could be decoded from those activations at 50% accuracy using a maximum likelihood decoding method. The key finding was that, for both raw and cross-validated measures, this number of decodable stimuli increased exponentially with the number of cells, consistent with the close-to-linear increase in information per cell.

The equivalent graphs from my simulated localist network are shown in Figure 1. As can be clearly seen, both graphs exhibit a startling similarity to those produced by Abbott et al. (1996). The graph of information against number of cells shows exactly the same pattern as did theirs, with information initially growing linearly with small numbers of cells, tailing off slightly below the linear trend line as the number of cells increases beyond about eight or nine. This pattern was seen for both the raw and the cross-validated measures, with estimates for the gradient of the linear portion of about 0.4 and 0.23 bits per cell respectively. Given that the raw and cross-validated values of this gradient ought to bound the true value from above and below, this implies that the true value lies somewhere in the range 0.23-0.4 bits per cell. As indicated above, the equivalent range for Abbott et al.'s data was 0.35-0.5 bits per cell, a correspondence whose closeness is all the more remarkable given the fact that absolutely no attempt was made here to optimize the fit between my simple simulations of a localist model of face recognition and their informational data collected in vivo. The second graph, in which the number of stimuli that can be decoded at 50% accuracy is plotted against the number of cells measured, makes it very clear that the former number is an exponential function of the latter. The graphs show some exponential curves fitted by hand to the data points. These curves are described by the functions $6.5 \times 2^{0.42N}$ for the raw data and $9 \times 2^{0.24N}$ for the cross-validated data, indicating a very similar range to that inferred above, namely, of 0.24-0.42 bits per cell. Ostensibly, according to the Abbott et al.'s logic, these figures would suggest that a field of 100 cells could represent something between 150 million and 10^{13} faces, an observation that would presumably lead them to conclude that the representation in our model is also highly distributed. And yet these data came from what we *know to be a localist model of face recognition*. We can conclude that their logic is flawed.

************* Insert Fig 1 about here, please ***************

To summarize this part, I have shown that a model that is *known* to be a localist model of face classification, with localist representations of a number of individual face-patterns, is capable of reproducing a pattern of data that Abbott et al. (1996) insist is a hallmark of "truly distributed" coding. The logical error that led them to this false conclusion is based on a failure to realize that, for a model in which localist units respond systematically to stimulus input (e.g., with their activations reflecting the similarity of that input to the set of learned patterns), the activations of a set of localist units have many of the qualities of a distributed representation. Specifically, even for a set of localist units, multiple units are activated for any given input stimulus, systematically (i.e., not randomly) and to different degrees for different stimuli. Exactly the same statement might be made of a set of units comprising a distributed representation in, for example, the hidden layer of a learned BP network. It is the presence, in the localist model, of a minimum of one localist representation for each known stimulus pattern, that is critical in defining it as a localist model. But unless one's single-cell recordings are liable routinely to alight upon the cells that locally represent the members of the stimulus-set of interest, there will be nothing in those recordings to distinguish one class of models from another.

One thing is that is very notable about Rolls (2007) description of these and other data, and that relates very directly to the simulation described above, is that although he described representations as distributed, he was very open to the idea that the information present in activations across a field of STS cells could be decoded (e.g., to indicate to the system which face is actually present) by computing the dot-product of that activation pattern with some vectors of learned synaptic weights. What he doesn't seem to acknowledge, though, is that the calculation of a dot-product is exactly what the units in, say, the classification layer of a standard competitive-learning network already effect (Page, 2000, and Bowers, 2009, make similar points). The simulation presented above indicates that by measuring from monkey STS, Abbott et al. (1996) might have *already* been measuring from such a localist classification layer – there is nothing in their data that rules this out or even makes it unlikely. Interestingly, if this conclusion were accepted, then *every* *one* of the supposed problems with localist coding that Rolls lists in sections 6.2-6.4 and section 6.7 of his paper can be disregarded, since his suggested solution is, in each case, to perform dot-product decoding! Moreover, the graceful degradation that he requires in section 6.6 can be achieved by redundancy in a localist-coding model (see above) and pattern completion (section 6.5) is easily achieved in localist models (see Page, 2000), if only by noting that a competitive classification network exhibits the so-called "attractor dynamics" that are a standard approach to pattern completion.

It is worth briefly considering whether *any* single-cell recording data would be able either to rule out or to confirm the existence of localist representations in the brain. The results of the simulations presented above suggest that it would be a brave researcher who stated that their recording data were definitively incompatible with localist representation. That would only seem possible if one were able to record from all of (or a large proportion of) the relevant cells, using all of (or a large proportion of) the possible stimuli represented – this is probably unrealistic. Is it possible, therefore, to take another tack, and to confirm (or, at least, to corroborate) the existence of localist representations? This is also difficult. For example, Quiroga and colleagues have probably gone further than most in providing corroborating evidence for localist representation. In Quiroga, Reddy, Kreiman, Koch, Fried (2005), for example, they measured from cells in the medial temporal lobe of awake human subjects. Some of the cells from which they recorded were extremely selective in their firing, one famously responding only to pictures of Jennifer Aniston (though supplementary materials suggest that that cell also responded to a picture of her Friends co-star Lisa Kudrow), another only responding to either pictures of Halle Berry in a variety of guises or even (most interestingly) to the printed version of her name. And yet Quiroga et al. were very reticent about claiming localist representation:

"We do not mean to imply the existence of single neurons coding uniquely for discrete percepts for several reasons: first, some of these units responded to pictures of more than one individual or object; second, given the limited duration of our recording sessions, we can only explore a tiny portion of stimulus space; and third, the fact that we can discover in this short time some images — such as photographs of Jennifer Aniston — that drive the cells suggests that each cell might represent more than one class of images. Yet, this subset of MTL cells is selectively activated by different views of individuals, landmarks, animals or objects. This is quite distinct from a completely distributed population code and suggests a sparse, explicit and invariant encoding of visual percepts in MTL." (Quiroga et al., 2005, p.1106)

This reticence is perfectly respectable (though Bowers, 2009, questioned some of the calculations and logic that underlay it) but it once again raises the question of whether one could ever find satisfactory evidence of localist coding by looking in the brain. For what it is worth, my favourite piece of positive evidence was supplied by Sakai, Naya and Miyashita (1994), who trained monkeys to associate pairs of visual patterns, whose precise nature could be changed by varying a small number of pattern parameters. Sakai et al. found cells in anterior inferotemporal cortex that responded preferentially to either pattern from a learned pair, compared with their response to other patterns. Crucially, they also found that the responses of those cells were systematically reduced when stimulus patterns were parametrically varied away from the precise patterns learned. In other words, it appeared that the cells responses were centred on the learned patterns. This is exactly what one would expect from localist coding, but is not predicted by a thoroughgoing distributed model. Needless to say, I would be very interested to hear from readers about other data along these lines, because, taken together, they would allow us to be more definitive about the representational styles that are actually found in vivo.

One interesting issue that was later raised by Quiroga (2012) was the extent to which they are different coding schemes in the hippocampus and other parts of the medial temporal lobe, as opposed to in cortex more generally:

"Modelling studies pioneered by Marr have shown that a sparse and explicit coding, as shown by these neurons, is ideal for fast learning and for the creation of new memories and associations. This contrasts with distributed representations in the cortex, which are better suited for the slow learning of shared structures of the stimuli, categorizations and generalizations. In fact, it has been proposed that the brain may use a complementary learning system approach: the fast-learning hippocampal system is used to learn facts of everyday life based on single exposures, and the neocortical system consolidates this information and embeds it within information from past experiences at a much slower pace, thus avoiding interference between different memories." (Quiroga, 2012).

From the point of view of the (or maybe just this) localist modeller, this passage is somewhat frustrating. First, it talks about a "sparse and explicit coding": is this a euphemism for localist coding? "Sparse coding" is a term often used by PDP modellers for a coding that looks and functions for all the world like a localist coding, but about which they can't bring themselves to use the word "localist". Technically, sparse codings (i.e., ones in which only a few cells activate for a given stimulus), can be either localist or thoroughgoing distributed. The same is also true of non-sparse codings. By way of demonstration of this fact (and prompted by a point raised by an anonymous reviewer), I calculated three different sparseness indices for the *known-to-be-localist* representations in my simulations presented above. The sparseness measures were the S measure discussed by Quiroga and Kreiman (2010), and the stimulus sparseness (a^s) and population sparseness (a^p) discussed by Rolls and Treves (2011). The explicit claim of Rolls and Treves (p.463) is that sparseness values a^s and a^p over about 0.60 are indicative of a "very distributed code". But the mean sparseness values from my *known-to-be-localist* model were a^s=a^p=0.85. Similarly, possible values of the S measure vary between +1 (binary grandmother cell) to -1 (every node activates fully to every pattern); the mean value from my localist simulations was S=-0.23. It is thus clear that various measures of sparseness are very poor (in fact, technically useless) guides as to whether representations are localist. Why? Because a localist model does not insist that only a few cells are significantly active to a given stimulus. It only insists that the cell (minicolumn, etc.) representing any given (known) stimulus is more active than others on presentation of that stimulus, and that that activity advantage is detectable and functionally important.

The second, and in some ways more acute, frustration occasioned by the above quote from Quiroga (2012), is that the quoted passage simply assumes that representations in the cortex are distributed, a purported fact that, it is further claimed, makes them better suited for slow learning, categorization and generalization. But there is nowhere any justification for either claim.

It is noteworthy, then, that even Quiroga - one of the group of researchers that has discovered the most localist-like medial-temporal-lobe representations ever identified - is apparently comfortable to assume that cortical representations are distributed, even though this runs contra to the *cortical* recordings of, say, Sakai et al. (1994) and, as it turned out above, finds no evidential support from the *cortical* recordings of Abbott et al. (1996). In the last sentence, we find a clue as to what is driving these assumptions, namely the complementary-learning hypothesis of McClelland, McNaughton and O'Reilly (1995). It is to a brief discussion of this hypothesis that I now turn.

The Complementary-Learning Hypothesis

The complementary-learning hypothesis (CLH) of McClelland, McNaughton and O'Reilly (1995) is one very influential explanation of how memory systems might interact so to avoid one of the well-known problems relating to learning in networks with thoroughgoing distributed representation. This problem is known as catastrophic interference and it is characterized by a network's inability to learn new pattern associations quickly, without risking the overwriting of the synaptic weights that implement previously learned (old) mappings. This is particularly the case for nonsystematic mappings, such as from phonology to semantics or from faces to names, and is a natural consequence of using learning techniques such as BP and other gradient-descent learning rules. To avoid catastrophic interference, such learning rules require that new pattern associations are interleaved with old associations, with the whole pattern set subject to slow learning over (very) many presentations (sometimes called “epochs”). For those committed to thoroughgoing distributed modelling, this raised the question of how such interleaving might be achieved in vivo; McClelland et al.’s CLH was a proposed response. Briefly, their proposal was that there is a fast-learning system in the hippocampus that is able to learn new pattern associations rapidly, even after a single presentation. (Note that, in making this assumption, there is already an implicit retreat from a commitment to thoroughgoing distributed representation in all brain areas, since one-trial learning is not a feature of fully distributed models – if it were, there would be no problem.) This was consistent with a great deal of data suggesting that the hippocampus is involved in episodic memory which, almost by definition, involves rapid, single-trial learning. McClelland et al. further proposed that these new pattern associations could then be transferred gradually to cortical areas, by interleaving their learning with the re-learning of those associations that had already been established cortically.

The CLH has been extremely influential and it is popular among those committed to thoroughgoing distributed representation in their neural networks. This popularity is largely because it appears to offer a solution to the catastrophic interference associated with gradient-descent learning, and in a way that is broadly consistent with a variety of data regarding the specialisms of, and interaction between, different brain areas. In what follows, though, I will question some of the logic that underlies the CLH.

My first point is less about the hypothesis itself than about the (epistemological) uses to which it has been put. Specifically, there is sense in which the CLH has been deployed to bolster the case for using thoroughgoing distributed representation (and usually gradient-descent learning too) in models of cortical function. The logic seems to run thus: gradient-descent learning with distributed representations has a problem with catastrophic interference unless new material is introduced via slow, interleaved learning; fast learning in the hippocampus, as posited by the CLH, permits slow, interleaved learning in the cortex; therefore the cortex must learn distributed representations by gradient-descent learning. This logic is faulty. While the challenges of learning distributed representations might require, say, fast learning in the hippocampus, that does not mean that the reverse is true, namely, that identifying fast learning in the hippocampus requires that learning in cortical areas is with thoroughgoing distributed representations. As noted above, the hippocampus is strongly associated with a form of memory, episodic memory, that must by definition be fast. But that does not rule out the possibility that there is fast learning in the cortex too. (Indeed, McClelland, 2013, himself gives an example of fast cortical learning – see later.) Moreover, even if (hypothetically) the hippocampus were the only part of the brain capable of fast learning, that would still not justify the line of reasoning set out above. It would only do so if one could plausibly maintain that the only possible rationale for slow learning were that it permitted the slow interleaving of new patterns with old, sufficient to avoid catastrophic interference. But one can imagine a number of reasons why slow learning might be advantageous, not least because it permits the learner to average over (perhaps many) individual episodes and to acquire knowledge that is more abstract and generally applicable than that contained in any single such episode. For example, in Page (2001), I discussed how a network employing localist representations of faces (the very same set of faces used in the simulations presented above) might abstract across examples of, say, happy faces, to discover which features/dimensions of that face were important indicators of happiness. This would then permit new, previously unseen happy faces to be recognized as such. The details of the implementation are not important, but it will suffice to say that the process involved pooling knowledge over a number of happy-face examples, so that the happy-face detector could learn to pay more attention to relevant dimensions, and less attention to irrelevant ones. This process necessarily involved slow learning. One can imagine that other similar processes, such as the derivation, by decontextualization, of semantic information from collections of individual episodes, would also require slow learning. The point is, then, that slow learning is independently motivated. Even if all cortical learning were slow (though I doubt it is – see below), one ought not assume that it is slow simply *because* its representations are distributed and would otherwise be subject to catastrophic interference.

But what about distributed modellers who do not adhere to the faulty logic outlined in the previous paragraph, but who would nonetheless claim that, while fast learning in the hippocampus does not logically *require* the slow learning of distributed representations in the cortex, it certainly *permits* the slow learning of distributed representations in the cortex. By showing how new information can be instated in cortex without catastrophic interference, they would claim, the CLH throws a lifeline to the idea that cortical representations are distributed throughout. Well, it turns out that there is another logical problem with CLH that calls even that into question.

The problem involves the means by which interleaved 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. This, after all, is a crucial component of the CLH: without slow, interleaved learning, new knowledge will not get transferred into cortex at all. Imagine, therefore, that we have a set of new pattern associations N, where each new association in the set N comes in the form of an association between an input pattern and an output pattern. For definiteness, let's assume that the new associations in this set N are between new faces and their corresponding names, this being a nonsystematic mapping (i.e., there is no tendency for two people who look similar also to have similar names). Further assume that this new set of pattern associations has been learned (fast) in the hippocampus, so the system's task is to transfer the knowledge of these associations to the cortex, where a much larger set of face-name associations, C, has already been learned by gradient-descent learning to some small and tolerable degree of residual error. Gradient-descent learning of pattern associations requires a supervisor that, for each input pattern in the training set, tells the network what the desired output pattern is. As learning proceeds, it is the difference between this desired output pattern and the actual output pattern produced when the corresponding input pattern is presented to the network, that drives gradient-descent learning. If there is no difference between the actual output pattern and the desired output pattern, then there is no (further) learning for that pattern-pair. The key point is this: According to the CLH, the hippocampus stores a series of input patterns (faces) and a (one-to-one) corresponding series of desired output patterns (names) which together comprise the set of new face-name pattern-associations N. During cortical training, this allows a calculation of the difference between the desired output pattern, for any given input, and the actual output pattern that is produced by forward propagation of that input pattern through the cortical network. But where, during the incorporation of new pattern-pairs to the cortical store, are the paired input and desired output patterns for set C (i.e., the old, known pattern associations) stored? Presumably they are stored in the cortex (where else?), but that raises a new question: Even assuming that we can recover a set of input patterns from the cortical network – normally this is assumed to be the set of input patterns used during previous training episodes, though it is rarely outlined how these are recovered from the bare network – how can we also produce a (one-to-one) corresponding set of desired output patterns, that are *going to differ, as learning proceeds, from the actual pattern that the cortical network currently produces in response to those learning patterns*. The emphasis here is to underline the fact that unless the cortex can produce, for any given input pattern, a desired output pattern that is different from the actual output pattern that it currently generates in response to that input pattern, then there is no error signal to drive learning or, more particularly, to drive relearning of the set of old pattern associations in set C, as and when a set of new pattern associations, N, is introduced. In relation to old patterns, cortex cannot both produce an output pattern and show where that output pattern is in error. Put another way, cortex cannot mark its own homework.

French (1997) at least realized that there was a problem along these lines. He suggested that before the incorporation of new, hippocampally-stored patterns (set N above), one could sample the mapping currently in cortex by clamping a random set of input patterns (he called them pseudopatterns) to the inputs of the cortical network, recording the outputs generated by those pseudopatterns, and then exporting that set of pattern pairs, call it P, back into some system capable of learning them quickly, presumably the hippocampus. There, they could be interleaved with the set of new patterns N with the combined set (N and P) used to train cortex. As I noted previously though (Page, 2000a), even small-scale simulations of this idea did not provide convincing levels of protection to the set of old associations, C, during interleaved learning of sets N and P, and so this is far from being a proven solution to the problem. Moreover, it seems clear that a resort to pseudopatterns can *never* work effectively in the context of highly nonsystematic mappings like that between faces and names: in that context, a small set of randomly generated pseudopattern pairs can never sample the cortical mapping function effectively, not least because there is no well-behaved (i.e., systematic) function that is available to be sampled. Even if it did work, the generation of a representative set of cortical pseudopatterns, P, and their export to the hippocampus every time one wishes to import even a small set N of new associations, seems quite a palaver. And all to avoid using localist representations in the cortex, a self-denial for which neither I nor Bowers (2009, p.239) can see the slightest independent justification.

Maybe the problem outlined here is a problem with gradient-descent learning rather than a problem with the slow, interleaved learning of hippocampal patterns among cortical ones. Possibly. But as football mangers sometimes say when accused of beating a weak team, you can only play the opposition in front of you. Maybe there are efficient ways of learning pattern associations via thoroughgoing distributed representations that are either immune to catastrophic interference or, if not immune, are not prey to the logical problem I describe above. I trust that if there are, readers will let me know. Interestingly and, dare I say it, somewhat hearteningly, McClelland (2013) himself has recently worried about whether gradient-descent learning is really biologically plausible. His corrective was to use a competitive learning network (like the one used above) which comprised, according to him, a “simple, arguably more biologically plausible network model” (McClelland, 2013, p.1203). These simulations resulted in localist representations of his input patterns, as is usual with this learning method. Curiously though, any successes enjoyed by those simulations were rather undermined in the article’s summary when he says rather bluntly, and again without obvious justification, “to me it seems clear that the brain uses distributed representations” (McClelland, 2013, p.1208).

In a very recent review, Kumaran, Hassibis and McClelland (2016) provided an updated perspective on the CLH. There is no indication in their review that the authors (who are at the cutting edge of developments in this area) believe that catastrophic interference in cortex is a problem that has been solved by more recent learning mechanisms. (Indeed, any such advance would entail that the CLH is a solution to some problem other than the one for which it was proposed.) While space does not permit a detailed analysis here, a striking feature of their paper, though, is the extent to which it de-emphasises many of what had previously been taken to be core aspects of the CLH. I shall briefly draw attention to three such aspects.

First, Kumaran et al. (2016) emphasize that fact that hippocampal codes (that they describe as orthogonal, conjunctive, pattern-separated – anything but localist) are capable of generalization. This undermines one of the oft repeated, but seldom justified, canards levelled at localist models, namely that they do not permit generalization (see Page, 2000a, for a fuller discussion). Specifically, Kumaran et al. describe simulations of models designed to account for the results of experiments on Paired Associate Inference (PAI), a particular form of generalization. Because these models must learn fast, and are meant to simulate the operation of the hippocampus, it is not surprising that theirs employed localist representations of the relevant compound stimuli.

Second, it is difficult to find, in Kumaran et al.’s review, much reaffirmation of the idea that the interleaving of new hippocampally stored associations with old cortically stored ones, is a means of learning new material while preserving knowledge of old. To be sure, the authors continue to assert that associations learned quickly in the hippocampus can subsequently be transferred across to cortex for more gradual learning. But that is much less controversial than the original idea, namely that cortical pattern-pairs are interleaved with new pattern-pairs during such learning, as a means of protecting knowledge of the former. The authors do refer to interleaving but admit that “which other memories are selected for interleaving with the new experience remains an open question" (Kumaran, et al. 2016, p.518). My additional question, based on the logical problems outlined above, is: how can any such interleaved presentation of old patterns among new, hippocampally stored ones, help at all in preserving old memories, no matter which specific ones are selected?

And finally, Kumaran et al. (2016) discuss data that suggest that there can be fast learning in the cortex. These data concern the cortical learning, by rats, of new associations between specific odours and designated locations in an "event arena". The authors make clear that this fast learning has been found to be contingent on the same event arena's having been previously exposed, and presumably learned, in the context of a number of different, learned odour-place associations. They characterized this as implying that cortical associations can be learned quickly only when new knowledge is "consistent" with prior knowledge. Unfortunately, though, this simply sidesteps what was hitherto the key question: If the old, event-arena knowledge is represented in a distributed fashion in the cortex, and the arbitrary (i.e., nonsystematically related) new odours are also so represented (both assumptions being integral to the CLH), then by what mechanism can the two be associated in cortex after just a single learning trial, without potentially causing catastrophic interference? If we assume, instead, that prior learning slowly establishes localist cortical representations of event-arena locations, then the problem simply goes away.

So Kumaran et al. (2016) conclude that localist representations can generalize and that cortex can learn quickly. Is there other evidence that cortex can perform fast learning? Some interesting data relate to so-called *fast mapping*, a term used to describe the incidental learning of (nonsystematic) associations between nonwords and novel visual objects. There is considerable controversy regarding precisely what fast mapping tells us about associative learning more generally. Part of the interest in this topic, though, stems from the observation that for some patients with medial temporal lobe (MTL) damage, associative learning can be faster and stronger under fast-mapping conditions than under experimental conditions in which participants are invited to make direct and explicit associations between the very same materials (e.g., Sharon, Moscovitch & Gilboa, 2011). In a particularly interesting paper, Merhav, Karni and Gilboa (2014) tested fast mapping in a small number of patients with amnesia attributable to MTL damage. The data suggested that even for these patients, associations could be formed between nonword labels and novel pictures over as few as three learning trials in a fast-mapping paradigm, with the learning still being evident on memory tests conducted the following day. If we assume that hippocampal learning was not possible for these patients, then we might conclude (as did the authors) that this fast learning was cortically mediated (probably in the anterior temporal lobe). What was notable, however, was that this (presumed cortical) fast learning in patients was very susceptible to subsequent interference occasioned by the fast mapping of a second, different word-label to a previously labelled picture. In the specific experiment performed by Merhav et al., five minutes after associating a given picture with a label A, the same picture was then associated with a different label B, under fast mapping conditions on each occasion. When patients were tested on the following day, no significant learning was evident of an association between the picture and either label A or label B (unlike for control pictures that had been associated with only a single label). In summary, therefore, under conditions of fast mapping, it is possible that there is indeed fast, associative learning possible in cortical regions – an eventuality not specifically envisioned within the CLH - but with the corollary that this fast learning brings with it a (catastrophic) sensitivity to proactive and retroactive interference. This observation of interference in fast cortical learning is, on the face of it, somewhat compatible with the assumptions of the CLH. Note, however, that even if nonsystematic learning were possible over a very small number of learning trials (two or three) in a fully distributed network, it is unclear how that could possibly support learning across a set of around 20 pattern-associations (as in these experiments) without learning becoming unstable. Moreover, on the introduction of a second label for a given picture under fast-mapping conditions, one would expect much more in the way of retroactive interference (RI: the second pattern label takes over from the first) than proactive interference (PI: the first pattern prevents learning of the second). In fact, in the data, there was both PI and RI sufficient that learning of neither label was evident after the interference induced by the second label. This is perhaps more consistent with a “propose-then-verify” account of associative learning (Trueswell, Medina, Hafri, & Gleitman, 2013), that itself presupposes very fast learning of candidate mappings.

Finally in regard to fast cortical learning, I would like to mention briefly some interesting work emerging in my own field, namely, the field of word-learning or, more particularly, the learning of phonological word-forms. In Page and Norris (2009), we suggested that the learning of phonological word-forms is carried out by a localist network that draws, for its information relating to the serial order of sublexical units (e.g., phonemes), on representations in phonological short-term memory. We also suggested that the same learning network can be seen to be in operation in experiments demonstrating the Hebb effect (Hebb, 1961). The Hebb effect is a phenomenon whereby repeated presentation of a particular list over the course of a series of trials involving immediate serial recall, leads to better recall of that list, showing quite rapid learning over the course of only a few (usually 8-10) repetitions. In that paper, we reviewed evidence that the learning seen in the Hebb effect is not just relatively rapid, but also results in learned representations that are stable over a period of months (Page, Cumming, Norris, McNeil & Hitch, 2013). Subsequent experiments with Szmalec, Duyck and colleagues (e.g., Szmalec, Duyck, Vandierendonck, Mata, & Page, 2009; Szmalec, Page & Duyck, 2012) have suggested that the list-learning seen in certain variants of the Hebb experiment is analogous to lexical learning. I mention this work because lexical learning, that is, the learning of new word-forms such that the learned representations are capable of indulging in lexical competition with the representations of previously learned forms (i.e., known words), has been considered, by thoroughgoing distributed modellers, to be the sort of ability that depends on slow, interleaved cortical learning (e.g., Davis & Gaskell, 2009), possibly even the sort of slow, interleaved cortical learning that can only take place in off-line periods such as during sleep. The finding, therefore, that “lexical” competition can result from new “words” presented in the context of a Hebb (serial recall) task, after as few as 12 presentations and without any intervening sleep

(Szmalec, Page, & Duyck, 2012), is potentially a challenge to the idea that cortical learning is necessarily slow. To be sure, Szmalec et al. found that learning of new word-forms sufficient for lexical competition was not immediate, being evident 12 sleepless hours after the corresponding Hebb experiment. Nonetheless, others, including Lindsay, Sedin and Gaskell (2011) and Kapnoula, Packard, Gupta & McMurray (2015) appear to have found evidence for the emergence of lexical competition (or “lexical engagement”) almost immediately. I do not want to give the impression that the matter is yet decided (see e.g., Weighall, Henderson, Barr, Cairney, & Gaskell, 2016), but if a faculty like full lexical learning (by which I mean learning of the word-form and incorporation of that new form into a competitive lexical network of previously acquired forms) is capable of being learned fast, then one might further question the idea that cortical learning is always slow. And if it is not slow, then it is probably not implemented by thoroughgoing distributed representations. There is the possibility, of course, that the full lexical learning of new word-forms is a purely hippocampal process, though that seems less likely when one considers that people with dense hippocampal amnesia are capable of Hebb-effect learning (Gagnon, Foster, Turcotte, & Jongen, 2004 - though whether they can learn phonological word-forms is admittedly less clear) and other people with entirely functional hippocampi (the short-term memory patients - see Baddeley, Gathercole & Papagno, 1998 for a review) are incapable of word-form learning. If it does turn out that fast cortical learning of word-forms is possible, then the CLH, at least as it relates to the issue of thoroughgoing distributed representation in word learning, might lose yet more force.

Summary

In this paper, I have given some reasons to believe why localist modelling in psychology is still a viable approach. I have provided evidence, from computer simulations, that directly contradicts the assertions of those who have claimed that their single-cell recording data, and various measures of information and sparsity derived from them, are inconsistent with localist representation. I have also questioned whether the complementary learning hypothesis (CLH) is really adequate either to rule out localist representation or to give an account of the ongoing learning of thoroughgoing distributed representations. My conclusion is that the CLH achieves neither of these things convincingly, this conclusion being supported somewhat by very recent updates to the CLH formulation. I maintain, then, that there are no brain data incompatible with the sorts of localist representation in which some cognitive psychologists, including myself, have been and continue to be interested. Furthermore, there are implementational and logical problems with the stated alternatives.

References

Abbott, L. F., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex*, *6*(3), 498–505. https://doi.org/10.1093/cercor/6.3.498

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. https://doi.org/10.1037//0033-295x.105.1.158

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220–251. http://doi.org/10.1037/a0014462

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*(2), 248–261. http://doi.org/10.1037/a0035943

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research*, *41*(9), 1179–1208. https://doi.org/10.1016/s0042-6989(01)00002-5

Coltheart, M. (this issue) Grandmother cells and the distinction between local and distributed representation. *Language, Cognition and Neuroscience*.

Davis, M. H., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3773–3800. http://doi.org/10.1098/rstb.2009.0111

French, R. M. (1997). Pseudo-recurrent Connectionist Networks: An Approach to the 'Sensitivity-Stability' Dilemma. *Connection Science*, *9*(4), 353–380. http://doi.org/10.1080/095400997116595

Gagnon, S., Foster, J., Turcotte, J., & Jongenelis, S. (2004). Involvement of the hippocampus in implicit learning of supra-span sequences: The case of sj. *Cognitive Neuropsychology*, *21*(8), 867–882. http://doi.org/10.1080/02643290342000609

Harris, C. S. (1980). Insight or out of sight? Two examples of perceptual plasticity in the human adult. *Visual Coding and Adaptability*, 95–149. https://doi.org/10.4324/9781315803043

Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*, 85–99. http://doi.org/10.1016/j.cognition.2014.09.007

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What Learning Systems do Intelligent Agents Need? Complementary Learning Systems Theory Updated. *Trends in Cognitive Sciences*, *20*(7), 512–534. http://doi.org/10.1016/j.tics.2016.05.004

Lindsay, S., Sedin, L. M., & Gaskell, M. G. (2012). Acquiring novel words and their past tenses: Evidence from lexical effects on phonetic categorisation. *Journal of Memory and Language*, *66*(1), 210–225. http://doi.org/10.1016/j.jml.2011.07.005

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210. http://doi.org/10.1037/a0033812

Merhav, M., Karni, A., & Gilboa, A. (2014). Neocortical catastrophic interference in healthy and amnesic adults: A paradoxical matter of time: Catastrophic Interference in Human Declarative Memory. *Hippocampus*, *24*(12), 1653–1662. http://doi.org/10.1002/hipo.22353

Page, M. (2000a). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, *23*(4), 443–467. http://doi.org/10.1017/S0140525X00003356

Page, M. (2000b). Sticking to the manifesto. *Behavioral and Brain Sciences*, *23*(4), 496–505. http://doi.org/10.1017/S0140525X00533354

Page, M. (2001). Paying Attention to Relevant Dimensions: A Localist Approach. In R. M. French & J. P. Sougné (Eds.), *Connectionist Models of Learning, Development and Evolution: Proceedings of the Sixth Neural Computation and Psychology Workshop, Liège, Belgium, 16–18 September 2000* (pp. 105–112). London: Springer London. Retrieved from http://dx.doi.org/10.1007/978-1-4471-0281-6_11

Page, M., Cumming, N., Norris, D., McNeil, A. M., & Hitch, G. J. (2013). Repetition-spacing and item-overlap effects in the Hebb repetition task. *Journal of Memory and Language*, *69*(4), 506–526. https://doi.org/10.1016/j.jml.2013.07.001

Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall. *Psychological review*, *105*(4), 761-781. https://doi.org/10.1037//0033-295x.105.4.761-781

Page, M. P. A., & Norris, D. (2009). A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Quiroga, R. Q. (2012). Concept cells: the building blocks of declarative memory functions. *Nature Reviews Neuroscience*, 587–597. http://doi.org/10.1038/nrn3251

Quiroga, R. Q., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review*, *117*(1), 291–297. http://doi.org/10.1037/a0016917

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

Rolls, E. T. (2007). The representation of information about faces in the temporal and frontal lobes. *Neuropsychologia*, *45*(1), 124–143. http://doi.org/10.1016/j.neuropsychologia.2006.04.019

Rolls, E.T. (this issue) Cortical coding. *Language, Cognition and Neuroscience.*

Rolls, E. T., & Treves, A. (2011). The neuronal encoding of information in the brain. *Progress in Neurobiology*, *95*(3), 448–490. http://doi.org/10.1016/j.pneurobio.2011.08.002

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*(1), 149–162. https://doi.org/10.1007/pl00005615

Sakai, K., Naya, Y., & Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learning & Memory*, *1*(2), 83–105. http://doi.org/10.1101/lm.1.2.83

Sharon, T., Moscovitch, M., & Gilboa, A. (2011). Rapid neocortical acquisition of long-term arbitrary associations independent of the hippocampus. *Proceedings of the National Academy of Sciences*, *108*(3), 1146–1151. http://doi.org/10.1073/pnas.1005238108

Szmalec, A., Duyck, W., Vandierendonck, A., Mata, A. B., & Page, M. P. A. (2009). The Hebb repetition effect as a laboratory analogue of novel word learning. *Quarterly Journal of Experimental Psychology*, *62*(3), 435–443. http://doi.org/10.1080/17470210802386375

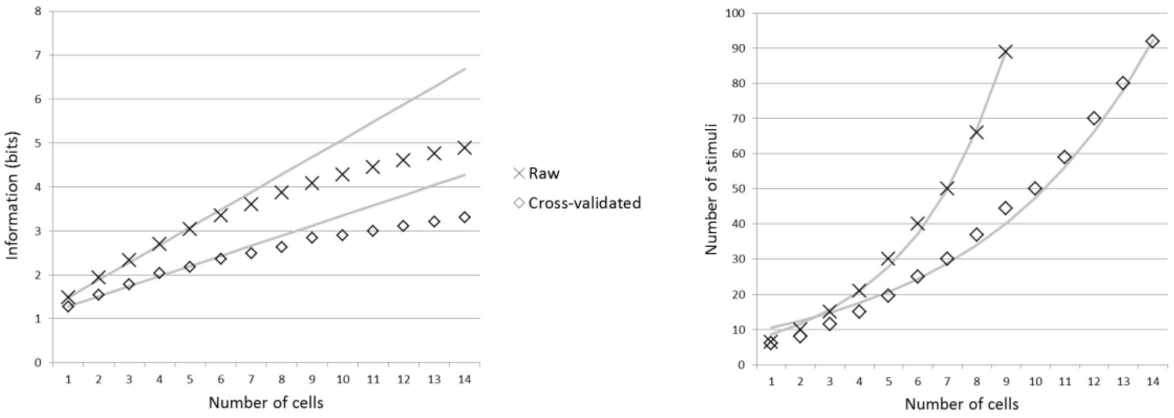
Szmalec, A., Page, M. P. A., & Duyck, W. (2012). The development of long-term lexical representations through Hebb repetition learning. *Journal of Memory and Language*, *67*(3), 342–354. https://doi.org/10.1016/j.jml.2012.07.001

Trueswell, J. C., Medina, T. N., Hafri, A., & Gleitman, L. R. (2013). Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive Psychology*, *66*(1), 126–156. http://doi.org/10.1016/j.cogpsych.2012.10.001

Weighall, A. R., Henderson, L. M., Barr, D. J., Cairney, S. A., & Gaskell, M. G. (2016). Eye-tracking the time-course of novel word learning and lexical competition in adults and children. *Brain and Language*. http://doi.org/10.1016/j.bandl.2016.07.010

Figure captions

Figure 1. The left-hand panel shows the number of bits of stimulus information as a function of the number of recorded cells. The two lines show the initial slopes and have gradients equal to 0.4 bits per cell for the raw information and 0.23 bits per cell for the cross-validated information. The right-hand panel shows the number of stimuli that can be decoded at 50% accuracy using maximum likelihood decoding. The fitted curves are exponential curves with equations $6.5 \times 2^{0.42N}$ and $9 \times 2^{0.24N}$ for the raw and cross-validated data respectively.


Information (bits)
8
7
6
5
4
3
2
1
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of cells
× Raw
◇ Cross-validated
Number of stimuli
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14
Number of cells

Disclosure of Interest

The author reports no conflicts of interest.