Language, Cognition and Neuroscience
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/plcp21

Connecting functional brain imaging and Parallel Distributed Processing

Christopher R. Cox[a], Mark S. Seidenberg[a] & Timothy T. Rogers[a]
[a] Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
Published online: 22 Dec 2014.

Language, Cognition and Neuroscience, 2014
http://dx.doi.org/10.1080/23273798.2014.994010

Routledge
Taylor & Francis Group

Connecting functional brain imaging and Parallel Distributed Processing

Christopher R. Cox, Mark S. Seidenberg and Timothy T. Rogers*

Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

Functional neuroimaging and Parallel Distributed Processing (PDP) theory, both introduced to cognitive science in the 1980s, led to influential research programmes that have proceeded in parallel with little mutual influence. The PDP approach advanced specific claims about the nature of neural representations that, perhaps surprisingly, have gone largely untested in functional brain imaging. One reason may be the widespread use of univariate statistical methods for analysing brain imaging data, which typically rely on assumptions that render them unable to detect distributed representations of the kind that PDP predicts. More recent multivariate methods for image analysis may be better suited to detecting such representations. In the current article, we consider why univariate methods have been insufficient to test PDP's representational claims, articulate some of the properties that neural representations ought to have if the PDP view is valid and then survey the recent neuroimaging literature for evidence that neural representations do or do not have these properties. The survey establishes that the PDP view of distributed representations has considerable evidential support. This analysis underscores the importance of understanding how the assumptions underlying methods for analysing functional imaging data constrain the kinds of questions that can be addressed. We then consider the implications for our developing understanding of the neural bases of cognition and for the design of future brain imaging studies.

Keywords: distributed representations; PDP; fMRI; MVPA

Distributed representation is one of the central tenets of the Parallel Distributed Processing (PDP) framework (Rumelhart, McClelland, & Hinton, 1986). The basic notion is that entities such as words, concepts, objects, faces, places and so on are represented by patterns of activity over sets of neural processing units. An individual unit may participate in many different representations, while representations that express similar content will be coded with similar patterns over many units. The utility of such representations has been demonstrated in PDP models of many phenomena in many domains; a recent special issue of *Cognitive Science*, for instance, surveyed the impact of the PDP approach in the domains of learning, perception, language, memory, cognitive control and consciousness (see Rogers & McClelland, 2014 and accompanying articles). Together such models instantiate a theory of cognitive representation and processing that differs from traditional approaches involving rules (Pinker, 1991), 'theories' (Gopnik & Wellman, 1994) and other symbolic representations (Tenenbaum, Griffiths, & Kemp, 2006). The models explain how representations of different types of knowledge develop, how such knowledge is structured and organised, and how it is used in performing different tasks. Models using distributed representations have provided new accounts of important elements of intelligent behavior (e.g., generalisation) and explain detailed aspects of behaviour that other theories miss (e.g., the quasiregular character of language and other types of knowledge; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & Plaut, 2014).

Despite these successes, important questions remain about the epistemic status of distributed representations. The promise of the PDP approach was that the use of 'neurally inspired' constructs such as distributed representations would prepare the way for integrated theories of behaviour and its brain bases. On the cognitive side, the relevance of distributed representations to understanding behaviour is well-established, but the models obviously abstract away many complex properties of neural systems. On the neurobiological side, it is generally accepted that mental representations are instantiated as patterns of activity over large systems of individual neurons, which communicate through synaptic networks with structure at multiple spatial scales. It remains unclear, however, just how such networks represent entities such as words, concepts, objects, etc. (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). The gulf between high-level cognitive and low-level biological understanding of representation thus raises questions about the extent to which the neural representations of cognitive entities are distributed in the PDP sense.

This paper considers the status of distributed representations by examining their relevance at the level of analysis we will term 'neurocognitive' – the level at

*Corresponding author. Email: ttrogers@wisc.edu

which the processing units of neural network models arguably make closest contact with measurements of the neural activity underlying cognitive behaviours. Specifically, we consider whether measurements taken at the scale supported by fMRI and other contemporary functional brain imaging methods reveal neural representations that are distributed in the PDP sense. The grain at which these methods engage cognitive phenomena seems roughly similar: the units in PDP models are not neurons but capture, in simplified and abstract form, the aggregate behaviour of many neurons. Likewise voxels in neuroimaging studies reflect, not the activity of individual neurons, but the aggregate behaviour of many thousands of neurons. Many important phenomena have been explored using both approaches. Our question, then, is whether neural representations of cognitive entities like words, objects, faces and concepts are distributed at this level.

It is perhaps surprising that this question has only recently begun to receive serious attention. Neuroimaging methods and neural network modelling were introduced to the study of human cognition at about the same time (see Posner, Petersen, Fox, & Raichle, 1988; (Rumelhart, McClelland, & PDP Research Group, 1986), and the two methods have developed in parallel. The initial applications of neuroimaging involved questions about the neural structures subserving different types of skills (e.g., decision making) or types of information (words, faces, places) –indeed, many contemporary studies retain this focus. To answer such questions, researchers have often applied univariate statistical methods that, though they may initially seem appropriate for addressing many fundamental questions about brain organisation, tacitly adopt assumptions about representation that run contrary to those developed under PDP. In fact, we will argue, these methods are not merely unsuitable for testing alternative hypotheses about the characteristics of neural representations: the assumptions underlying the analytical procedures themselves limit the kinds of results that can be observed. Although many studies appear to provide evidence for local rather than distributed representations (e.g., of words: Glezer, Jiang, & Riesenhuber, 2009; faces: Kanwisher, McDermott, & Chun, 1997), reasoning from data to conclusion in these cases is often informal, appealing to intuitions rather than strong tests of competing hypotheses. The consequence is that, despite being close contemporaries, the two methods have exerted comparatively little mutual influence and have led, in some cases, to quite different views about how information is organised in the brain.

Recent years have witnessed the elaboration of new multivariate methods for analysing neuroimaging data that seem better suited to assessing the status of neurocognitive distributed representations in the brain (Pereira, Mitchell, & Botvinick, 2009; White & Poldrack, 2013). Indeed, in some cases, the new methods were directly motivated by the theoretical gap between classical functional imaging and the view of cognition offered by neural network models (Kriegeskorte, 2009; Norman, Polyn, Detre, & Haxby, 2006). Although they also adopt important underlying assumptions that affect the kinds of results that can be obtained (Cox & Rogers, submitted), the variety of methods currently on offer, and the general ubiquity of multivariate studies in recent work, together have yielded sufficient evidence to permit us to assess, for the first time, whether neural representations at this scale are distributed in the way that PDP models have long suggested.

Before beginning, it is worth considering in more detail the rough correspondence suggested earlier between units in a PDP model and voxels in a brain imaging study. What motivates this analogy, beyond convenience? Brains are, of course, composed of neurons, and neural network models are sometimes described as assemblies of neuron-like processing units. Thus, it might seem natural to think of a unit in a PDP model as roughly analogous to a single neuron. The analogy is tenuous, however. Whereas individual neurons exhibit all-or-nothing spiking behaviour, units assume continuous activation states. Low-level dynamics such as lateral inhibition, temporal coherence and local extra-cellular conditions are glossed over in most connectionist models despite being critically important for understanding the behaviour of individual neurons. The models also abstract away from morphological differences among neuron types, cytoarchitectonic details such as the organisation of neurons into cortical columns and other facts about brains. PDP units can instead be viewed as capturing, in a modest number of processing elements, the same informational states existing across vast numbers of heterogeneous spiking neurons in real nervous systems (Rogers & McClelland, 2014; Smolensky, 1986). The central assumption is that the representational content and cognitive functions expressed in the coordinated spiking behaviours of hundreds or thousands of neurons can be usefully approximated as a much smaller vector of continuous-valued activations, with individual units corresponding to single elements within the vector and summarising the informational states of large populations of neurons.

Functional brain imaging adopts essentially the same central assumption. fMRI does not measure the activity of individual neurons but infers, via changes in blood oxygenation level at the scale of approximately 3 mm^3, the net synaptic input delivered to a population of thousands of individual spiking neurons (Arthurs & Boniface, 2002; Logothetis & Wandell, 2004). That is, each voxel provides approximate summary information about metabolic demands exerted by a large population of individual neurons. The effort to relate such measurements to cognitive representations and processes entails the assumption that there exists, in real brains, an important relationship between neural activity abstracted at this scale

and the representations and processes that underlie cognition. Put differently, if functional brain imaging is to have any validity, it must be the case that the representational content and cognitive functions expressed in the coordinated spiking behaviours of hundreds of thousands of individual heterogeneous neurons can be usefully approximated with a smaller vector of continuous-valued activations. In this case, the elements of the vector are individual voxels and their values summarise a statistical relationship between the BOLD time-series at the voxel and other cognitive events, but the parallel to the central PDP assumption is clear. For this reason, in what follows we take the activation of a single unit to be a model analogue of the mean activity in a population of neighbouring neurons, similar to that estimated from changes in the BOLD response at a single voxel using fMRI. The central question is whether neural representations so measured have the properties that the PDP framework predicts.

In the next section, we make explicit the representational assumptions that underlie standard univariate approaches to image analysis and then contrast them to those inherent in PDP models of cognition. This exercise establishes why standard neuroimaging methods have been insufficient to test PDP's representational claims, and thus the need to consider results from multivariate methods. We end the section by stating four properties that neurocognitive representations are predicted to possess according to the PDP framework. The remainder of the paper then surveys current evidence bearing on each of the four properties.

Contrasting representational assumptions in classical brain imaging and PDP models

We have suggested that the univariate methods that have been standard in functional brain imaging for many years are not suited to assessing whether neurocognitive representations are distributed. Such methods developed from a mainly modular view of neurocognitive organisation: different cognitive functions or representational domains were thought to be supported by different discrete and contiguous regions of cortex; neurons within the region were thought to be strongly active when the region's function was being carried out and relatively inactive otherwise; and the goal was to assess which regions supported which functions or representational domains. To meet this goal, statistical methods were developed that allowed the researcher to identify contiguous regions of cortex that, across groups of participants, showed systematically different levels of activation in different experimental conditions.

The analysis that became standard proceeded as follows: data are first pre-processed to minimise noise and eliminate confounding trends in the data, such as slow oscillations and head motion. The data are then spatially smoothed: each voxel's estimated response is replaced with a weighted average of neighbouring voxel responses, with the weight diminishing as a Gaussian function of spatial (anatomical) distance. Next, data from multiple subjects are aligned to a common atlas on the basis of a few common anatomical landmarks and an affine transform, so that voxels in the same relative anatomical location can be related across participants. The time series of activation from each voxel is then modelled separately; it is now common for this to be done using a mixed-effects regression model, with participants treated as a random effect. The analysis produces a single statistical map of the brain, with the effect of the experimental manipulation estimated at each voxel, independent of the activity estimated at other voxels (aside from the local correlations emphasised by spatial smoothing). To determine the statistical significance of these effects while avoiding punishing corrections for multiple comparisons, steps are typically taken to reduce the number of comparisons assessed, for instance by cluster-thresholding (i.e., only testing voxels whose anatomical neighbours show similar patterns of contrast) or averaging voxel responses over regions of interest.

From this brief description, it is clear that univariate methods adopt particular assumptions about what neural representations must be like. The assumptions are not typically spelled out, however, so we articulate them here, then consider how they differ from the corresponding PDP assumptions:

(1) *Independence of representational elements*. The approach tests statistical associations between cognitive states and the states of individual voxels taken independently. Interactions among voxels are not considered. Thus, the approach assumes that the representational or processing significance of a given voxel's activation does not depend upon the states of other voxels. Each voxel encodes whatever it encodes, regardless of what other voxels are doing at the time.

(2) *Discrete representation and functionality.* The group-level significance tests that are typically the final result of a functional imaging study discriminate voxel groups that show reliable differences between conditions from those that do not. The implication is that such voxel groups are involved in encoding a given kind of representation or carrying out a particular process while other voxels are not, suggesting a discreteness of functionality in which each region contributes to one kind of representation or process and not to others.

(3) *Homogeneity of representation within and across individuals*. A third assumption is that the voxels contributing to a given representation respond to relevant items in essentially the same way, both within and across individuals. For instance, if one

voxel contributes to the visual representation of a face by increasing its activation, other face-representation voxels should also respond by increasing their activation, and this code – increased activation for faces – should be the same across individuals. This assumption is part of what licenses several common statistical steps, including region-of-interest analysis (in which the activations of all voxels thought to contribute to a representation are averaged to yield a single number), cluster thresholding (where voxel activations are discounted if their anatomical neighbours do not respond similarly), spatial smoothing (where activation of a voxel is replaced with a weighted average of its neighbour's activations) and group-level statistical tests of voxel activations – steps that will only improve signal discovery if, in fact, elements of a representation respond to objects of representation in essentially the same way within and across individuals.

(4) *Homogeneity of location within and across individuals*. Finally, the approach assumes that the voxels contributing to a given function or representation are localised similarly both within and across individuals. For instance, if a given voxel encodes the presence of a face, then neighbouring voxels in the same individual should also encode faces, and voxels residing in the same anatomical location in other individuals should likewise encode faces. This assumption is also required to justify several of the data-processing steps noted in (3) above: region-of-interest analysis, cluster-thresholding, spatial smoothing of the BOLD signal and group-level statistical tests at a given anatomical location. Again, these steps will only lead to accurate signal discovery if the elements of a representation are anatomically localised in similar ways within and across individuals.

We do not assume that all researchers who employ univariate methods explicitly endorse all of these assumptions about the nature of neurocognitive representations. Because the methods have become so widespread, they often may be treated as the default method by researchers who might otherwise view the assumptions as we have stated them with some scepticism. Our purpose here is to make the underlying assumptions of the statistical methods very explicit, so as to better illustrate why they can fail to uncover important structure if the central assumptions are invalid.

It is also important to note that there are at least two ways in which neurocognitive representations may be viewed as being distributed while still conforming to these assumptions. First, the word 'distributed' is sometimes used in cases where univariate contrasts reveal reliable differences in BOLD response, not just in one cortical area, but in multiple anatomically well-separated areas. For instance, univariate fMRI studies of visual perception often show elevated BOLD responses for faces relative to other objects in parts of the occipital cortex, the posterior fusiform and the antero-ventral temporal lobe (Behrmann & Plaut, 2013). These regions are sometimes then described as forming a distributed network for face representation. Second, the word 'distributed' sometimes refers to the case in which a representation is encoded over multiple regions, each encoding a different kind of information. For instance, theories of semantic representation often view the meaning of a word as being distributed over cortical regions that each individually encodes a particular kind of sensory, motor or linguistic information. Thus, the colour of item is represented within a colour area, shape is coded within a shape area, characteristic motion is encoded within a motion area and so on (Martin & Chao, 2001). In this scenario, the meaning representation is distributed as a pattern of activation across many potentially widely dispersed cortical regions, with the regions themselves individually behaving according to the assumptions of standard univariate image analysis. Within the colour area, for instance, the voxels are still viewed as always encoding colour without contributing to the representation of other kinds of information; as encoding the colour information independently, so that state of units outside the colour area need not be taken into account; as being anatomically situated within the same contiguous region; and as being homogeneously located across individuals.

Neither of these cases is at odds with the PDP view of distributed representation, but they are of less interest here because on these views, the components of the distributed representation each individually conform to the univariate assumptions and so can be discovered through standard methods with the appropriate contrasts. The PDP framework, however, allows for the possibility that representations may also be distributed in other ways that violate the univariate assumptions and so cannot be discovered by these methods. To see this, it is useful to consider the behaviour of a very simple model, such as the model that learns the XOR ('exclusive or') mapping shown in Figure 1A. In this problem, the model must activate an output unit if the two input units are in different states ([0,1] or [1,0]), but not if they are in the same state ([0,0] or [1,1]). To learn this mapping, the model is first given weights with small random values centred on 0 and is then trained with backpropagation to generate the correct output state for each input pattern. Regardless of the initial weights, the model eventually learns to generate the correct output for each input – thus the model's overt behaviour is invariant with respect to the initial weight state. Nevertheless, the model's internal representations – the patterns of activation generated by a given input over the hidden units – can vary substantially from training run to training run.

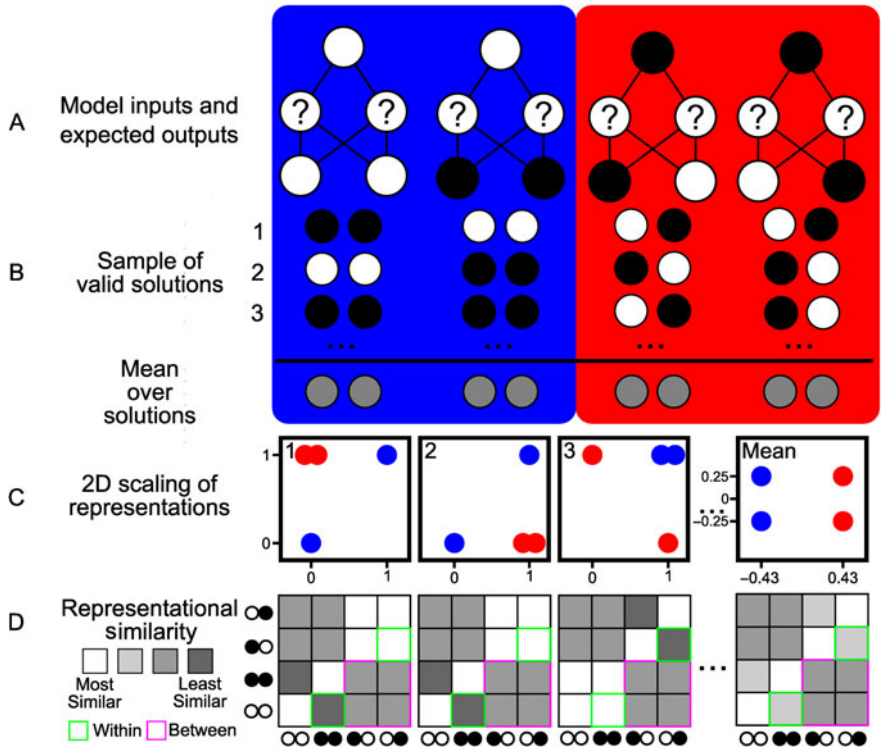

Figure 1. (A) The architecture of the XOR model, a simple feed-forward model that illustrates several properties central to the PDP approach. The model is composed of five units. Two input units send projections to two hidden units, which both project to one output unit; each unit is also given a trainable bias. The model, given two inputs, must produce their exclusive-or (XOR). The four panels show these inputs and outputs; circle shading indicates unit activation, ranging from zero (white) to one (black). (B) Hidden unit activations generated by the four different inputs (shown above in panel A) after training in three different runs of the network initialised with different random weights. The bottom row shows the mean activation of each unit for a given input pattern averaged over a hundred runs of the network. (C) Plots of the hidden unit activations for each of the three individual solutions shown in (B), and a 2D scaling of the average Euclidean distances amongst hidden patterns across a hundred network runs. Though the hidden units take individually do not appear to encode consistent structure, the similarities of the patterns across both units are highly consistent across runs. (D) Representational similarity matrices for the hidden unit solutions and the mean similarities across 100 solutions. Light squares indicate a smaller Euclidean distance between patterns.

Figure 1B show these patterns for three different training runs, initialised with different random weights. Across runs, each hidden unit learns a different response to the input. The bottom row shows the activation of each hidden unit to each pattern averaged over 100 training runs initialised with different random weights. From the mean activations over runs, there appears to be no systematic structure in the representations acquired: the average activation of both hidden units is about 0.5 over all four input patterns. There are, however, isomorphisms in the learned representations across network runs. Figure 1C shows the same hidden unit activations as two-dimensional plots for the three individual runs (panels 1–3), and a 2D scaling of the Euclidean distances between inputs patterns averaged over many runs (panel 4). Though each hidden unit behaves quite differently across network runs, nevertheless the representational structure extracted over *both* units across runs is quite similar: in each solution, two input patterns that generate the same output are represented in one corner of the space while the other two are on the opposite side, with a linear plane separating these. On average across runs, items that generate the same output are represented internally as more similar to one another than they are to items that generate different outputs (Figure 1C, panel 4). In other words, across different training runs of the same network, a given internal unit behaves very differently, but the representational structure coded across both hidden units is quite similar. The same data are presented in Figure 1D as representational similarity matrices to further elucidate these qualities.

Though very simple, the XOR model illustrates four characteristics of the distributed representations that emerge through learning in PDP models and that differ from those assumed by univariate analyses of brain imaging data:

(1) *Interdependence of representational elements*. In PDP models, interesting representational structure – phonological, morphological, conceptual, visual, etc. –is encoded in the patterns of activation evoked

across whole ensembles of units, but may not be apparent in the individual activity of single units within the ensemble. This characteristic is apparent in the XOR model: though there are only two units coding internal representations, both are clearly involved in expressing the interesting representational structure. Neither the mean activity of individual units across models nor the raw activity of individual units within a given model is sufficient to determine which inputs produce similar outputs. When the similarity structure of the patterns evoked by different inputs across both units is taken into account, the underlying representational structure is apparent both in individual models and on average across models.

(2) *Graded representation and functionality.* In PDP models, a given unit can participate in the representation of many different items. Within a distributed representation that robustly distinguishes, say, two different domains, a single unit may activate for subsets of items from both domains, or for all items in one domain and a few items in the other, or for only a subset of items within a single domain and so on. For instance, the XOR model can be viewed as differentiating two kinds of input patterns: those where both units are in identical states from those where the units are in different states. How then do the internal units participate in the representation of these two categories? In the first solution shown in Figure 1, the left-most internal unit activates for the inputs [0,1], [1,0] and [1,1], and thus in some sense contributes to the representation of all items in the 'different' domain and one item in the 'same' domain. The other unit activates only for the [1,1] pattern and so in one sense contributes to the representation of only half the items in the 'same' domain. What really matters to the model is the degree to which different objects of representation elicit similar or different patterns of activation across unit ensembles; because this is true, any single element of the representation may appear to contribute to many different representations, in virtue of showing systematically increased or decreased activation.

(3) *Heterogeneity of representation within and across individuals.* In PDP models, the units that jointly encode a distributed representation – typically units within the same layer, which are connected in similar ways to other units in the network – can nevertheless exhibit very different responses to their inputs. Indeed, because the patterns of activation across units in a layer express representational structure suited to the task at hand, units within a layer must respond at least somewhat differently to different inputs. We see this in the XOR network, where the representational role adopted by one hidden unit always complements and never mimics the role adopted by the other. Within a single network, then, the components of the internal representation are heterogeneous in their responses to inputs. Across different runs of the network, the two hidden units can learn quite different representational codes, illustrating that internal representations can be heterogeneous across individuals as well in this framework. With increased representational capacity in the form of more hidden units, it is likely that some subsets of units would come to exhibit similar responses across inputs, but other unit subsets would still have to adopt different response profiles so that, across hidden units, quite different patterns of responses are observed.

(4) *Heterogeneity of location within and across individuals.* Finally, even where different networks adopt the same representational code for various inputs, the localisation of the code over units – the particular way that a given unit in a given layer responds to various stimuli – can vary arbitrarily. For instance, the XOR network often discovers a solution in which one internal unit acts as an AND operator (activating only when both inputs are active) while the other acts as an OR operator (activating whenever at least one input is active). This set of networks can be said to have discovered essentially the same internal code for the inputs, but the localisation of the code – exactly which hidden unit adopts which functional role – varies at random across the set. The same can be true of hidden unit representations within larger networks. For instance, consider training the XOR model in an architecture containing four hidden units instead of two. In such a model, one solution can involve having two hidden units both function as AND units and the other two both function as OR units – yet there is no need for the two AND units to be anatomically adjacent to one another. Again, exactly which unit adopts which response profile can vary arbitrarily even within the network.[1]

We are now in a position to see why univariate methods for analysing brain imaging data are insufficient to test whether neural representations are distributed at this scale. The PDP view suggests that each of the representational assumptions underlying the standard approach may be invalid, in which case the method cannot detect the stipulated properties if they exist. If representational structure is encoded in patterns over sets of units but not in the states of units taken independently then univariate

statistics will produce null results. Univariate contrasts will fail to identify individual representational elements that happen to contribute to representations of items from both of the contrasting domains. If the elements of a representation encode information in heterogeneous ways, then averaging voxel activations within or across individuals will destroy signal, potentially leading to null results. If the elements encode information in the same way but are differentially localised either within or across individuals, then averaging across voxels at a given anatomical location will likewise destroy signal. Thus, if neural representations do have the properties predicted by PDP, standard univariate methods will fail to discover them. Univariate studies that potentially offer evidence for discrete, local representations are therefore difficult to interpret: they may fail to find evidence for distributed representation either because such representations do not exist at the level probed by functional brain imaging, or because the methods are not capable of detecting them.

Multivariate approaches hold the promise of remediating these limitations. As in PDP, such approaches focus on understanding the nature of the information contained in patterns of activation across sets of representational units taken together. Thus, they do not assume independence of representational elements, discreteness of representation or homogeneity of representation. Some approaches do assume homogeneity of location within and across individuals, but newer methods are beginning to relax this assumption as well. Moreover, in some cases, the contrast between results from univariate versus multivariate analyses and, in other cases, a close inspection of the multivariate results alone can provide direct evidence testing whether neural representations are in fact distributed.

In the remainder of the paper, we survey recent work in functional neuroimaging to assess the status of each of the four properties of representations predicted by PDP. Our principal aim is to assess the face validity of the four properties. Our goal is not to conclusively determine whether representations in some domain are distributed (in the PDP sense) because the current state of the evidence does not allow this; rather it is to determine whether neural representations are even *plausibly* distributed in this sense. In each case, we will first consider what evidence for distributed representation would look like, and then review studies that report relevant evidence. Following this survey, we conclude by briefly considering how brain imaging might best be approached if the PDP representational claims are valid.

Independence versus interdependence of representational elements

As we have seen, one point of contrast in the representational assumptions of univariate brain image analysis and PDP concerns the degree to which elements of a representation – units in a model or voxels in the brain – individually express important cognitive content. Standard brain imaging approaches assume that the important elements of representation can be discovered through univariate analyses, and hence that the elements contribute independently to representations. The fact that univariate methods often succeed in finding such elements indicates that clusters of voxels do indeed sometimes behave in ways amenable to discovery via univariate analysis. PDP, however, posits that information can sometimes exist in the pattern of activation across multiple units, without being reflected in the individual activations of the components. Is there any evidence supporting this hypothesis?

What might such evidence look like? As a start, consider that, if the univariate assumptions are always true, then the information encoded across all units in a representation will also be reflected in the individual elements of the representation. That is, there should be little or nothing gained in analysing sets of voxels all together compared to analysing individual voxels separately, since each element contributes to the representation independently. If the PDP assumptions are valid, however, there should be information contained in the patterns of activation across units that cannot be decoded from individual voxels taken separately. Thus, if multivariate methods and univariate methods, applied to the same dataset in search of the same information, identify different voxel sets, this would suggest that the PDP assumption is valid.

Jimura and Poldrack (2012) conducted just such an analysis in a study of how the brain processes gain and loss in a gambling task. Many cortical regions were identified using a multivariate searchlight analysis (Kriegeskorte, Goebel, & Bandettini, 2006) that were not detected by the univariate method. If the searchlight method had simply identified a superset of the regions identified by the univariate method, the result might not be compelling – it may simply be that multivariate methods are 'highly opportunistic' (Kriegeskorte et al., 2006, p. 550), identifying regions with very weak signal just as might happen by relaxing the significance criterion in a univariate statistical test. What makes the result particularly interesting in the current context is that the voxels identified by the univariate analysis were not simply a subset of those identified in the multivariate analyses. In addition to flagging regions that seemed irrelevant from the univariate analysis, the multivariate analysis did not flag several regions implicated by the univariate analysis.

Other work has demonstrated that the results of univariate and multivariate methods can actually doubly dissociate. Riggall and Postle (2012) noted that regions in frontal and parietal cortex displayed sustained activation

during the delay period of a working memory task in which participants were required to hold in mind the speed and direction of an array of moving dots. The authors trained a multivariate pattern classifier to determine which direction of motion was being held in memory, using the activations of voxels in these fronto-parietal regions as inputs. The classifier was unable to decode patterns at a level greater than chance, indicating that, despite the systematically elevated delay-period activity in these regions, the patterns did not encode the contents of working memory. Decoding was possible, however, from classifiers trained on voxels in the occipital cortex, even though this area did not show elevated delay-period activity according to the univariate analysis.

Such result might initially seem counterintuitive – surely an effect that can be detected by univariate methods must also be picked up by a multivariate approach. To see why this intuition is incorrect, consider the patterns shown in the schematic Figure 2. Each row of large squares corresponds to a different searchlight that fixates on separate sets of voxels (small squares); assume that these are exactly the same voxels across three subjects. The fourth large square in each row corresponds to a blurred average of the voxels in that searchlight over the three subjects. The colouring of the smaller squares indicate the degree to which each voxel's activations reliably predicts an experimental factor of interest: for instance, bright blue voxels might reliably predict that a stimulus item was from category A, while bright red squares reliably predict category B. Pale colours indicate voxels whose activity is only weakly correlated with the contrast of interest, while grey squares show uncorrelated voxels.

The first searchlight exemplifies a case where a searchlight MVPA will identify signal missed by univariate analysis. Within each individual searchlight there are 2 or 3 voxels that reliably carry useful information about the stimulus class so that a trained classifier will successfully generalise to a hold-out cross-validation set. Such a classifier will therefore perform well for each individual subject, and the searchlight method will flag this searchlight location as encoding information relevant to the discrimination. Yet the particular way the information is encoded is highly variable across voxels in the searching for each individual, as is the exact anatomical location of the signal-carrying voxels. Blurring within subjects will thus dilute signal with noise, and averaging at a given location across subjects will further eliminate signal. As a result, the mean difference over subjects within the first searchlight will be near zero for all voxels.

The second searchlight shows the reverse case: here, most of the voxels in each subject are uncorrelated with the distinction between A and B, and only a small subset is weakly correlated with the distinction. A classifier trained on each subject individually has a high likelihood of failing a cross-validation assessment. If the classifier fails in many subjects, the searchlight centre will not be identified as reliably encoding information relevant to the distinction, meaning that this region will not be identified by a searchlight MVPA. The weakly covarying units, however, happen to encode the distinction of interest in the same manner, and to reside near one another within

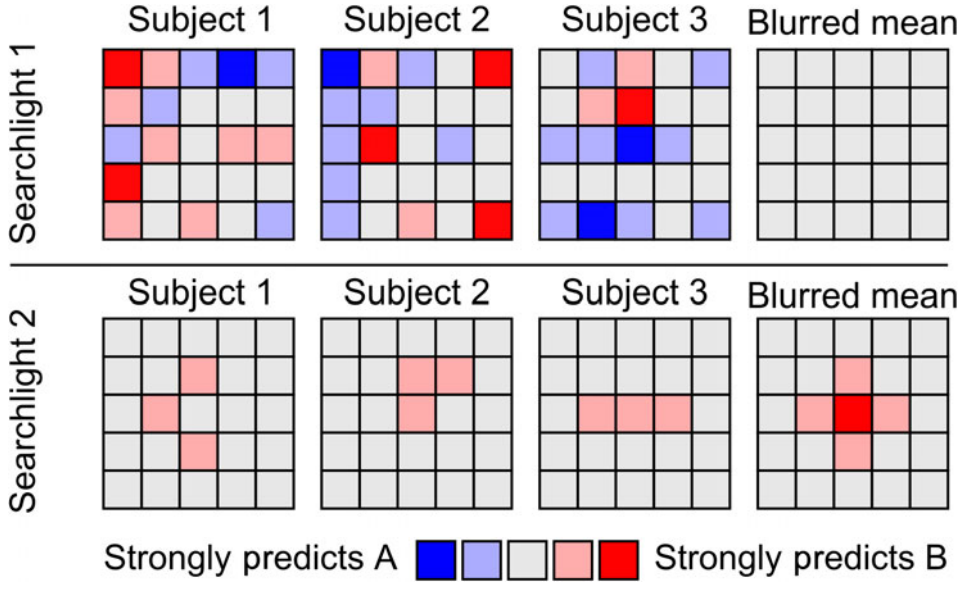

Figure 2. Each row corresponds to a searchlight that contains a set of 25 voxels. These voxels are the same across subjects, but different across the two searchlights. Searchlight 1 exemplifies a case where a searchlight MVPA will succeed but a univariate analysis, employing blurring within and averaging across subjects, will fail. Searchlight 2 exemplifies a case where a searchlight multivoxel pattern analysis (MVPA) is likely to fail but a univariate analysis will succeed. See the text for a more full discussion.

and across individuals. The univariate assumptions are met, so smoothing within and averaging across individuals reduces noise and allows detection of the voxel with univariate tests. Thus, the searchlight MVPA can fail to find signal in the very cases the univariate approach was designed to address – that is, when the signal is buried in noise within individuals, but is coded independently in the same way and in the same location within and across individuals (see Cox and Rogers, submitted, for simulation examples of this case).

Graded versus discrete contributions to representation

The second point of contrast concerns the degree to which a given representational element can participate in many different representations. The classical view posits a discrete functional specialisation, in which each element contributes only to a particular kind of representation – with, for instance, a given voxel activating only for faces (or a subset of faces), or for animals (or a subset of animals) and so on. Distributed representations, in contrast, are useful because they allow representational structure to be expressed as graded similarities across many representational elements. In such a scheme, any individual element will contribute, in graded fashion, to the representation of many different items or even to different representational domains. Thus, a second important question for the literature is whether it contains evidence that individual voxels contribute in a graded fashion to different representations.

A clever indirect method for answering this question leverages neuronal adaptation (Grill-Spector, Henson, & Martin, 2006). Typically, the neural response to a stimulus will decrease over repeated presentations of the same item as active neurons deplete their resources with repeated firing. If representations are distributed so that two stimuli evoke overlapping patterns, the overlapping portions of the patterns would also be expected to adapt. Though the adaptation is happening at a scale much smaller than a functional voxel, if there is sufficient overlap across the representations, the net effect will be to diminish the voxel's response relative to an appropriate control condition. The method naturally extends to any domain where one is interested in testing whether neural representations overlap.

One particularly interesting domain to which fMRI adaptation analysis has been applied is lexical semantics evoked by word reading. Printed words are highly controlled stimuli, and orthography and phonology are both relatively uncorrelated with semantics, so it is possible to dissociate semantic from perceptual similarity (e.g., BIG and LARGE are semantically similar but formally dissimilar; HAIR and PAIR are semantically dissimilar but formally similar). Also, there is a deep psycholinguistic literature that has set a high bar for stimulus set composition; it is standard practice to control for word frequency and other potentially psycholinguistic dimensions, further isolating effects of interest.

With such stimulus sets, it is possible to use the adaptation procedure to assess the extent to which representations of different word meanings overlap. If such meanings are expressed as distributed patterns of activation, with similar meanings evoking similar and therefore overlapping patterns, the predictions for such a study are clear: the adaptation arising from successive presentations of semantically related words should be larger than that produced by successive unrelated words. That words from the same semantic category (e.g., two vehicles) result in more adaptation than words from different categories (e.g. a vehicle and an animal) is a widely replicated effect (Henson & Rugg, 2003; Rissman, Eliassen, & Blumstein, 2003; Wheatley, Weisberg, Beauchamp, & Martin, 2005). However, to address the PDP prediction that representations express graded similarity, more than two points are needed. With only two conditions, whether there is representational similarity among items from the same category cannot be assessed. To our knowledge, graded similarity structure for the meanings of words refering to objects has not been explored using this method. It has, however, been explored in the domain of numbers and numeric magnitude. For example, Piazza, Pinel, Le Bihan, and Dehaene (2007) found that the degree of dishabituation between a habituated numeric quantity or numeral and a deviant stimulus was a function of the difference in magnitude. This suggests that there is graded similarity among number concepts, even when presented as Arabic numerals. This outcome would not be expected if the representations of meanings were discrete and non-overlapping.

Representational overlap can also be assessed by comparing the solutions found by two or more multivariate pattern classifiers within the same subject. Many such classifiers assign real-valued weights to each voxel that indicate the degree to which the voxel contributes to the relevant discrimination. When two or more classifiers are trained to perform different discriminations, the weights assigned by each classifier to each voxel can be compared. Voxels that receive large weights in both solutions can then be identified as important for both representational distinctions.

Studies of this kind are far less abundant, but do exist. One such study performed a three-way linear discriminant analysis of evoked brain activity measured by fMRI to distinguish trials in which subjects were presented with pictures of either faces, houses or chairs (Carlson, Schrater, & He, 2003). After demonstrating above-chance pattern classification, the authors projected the solutions associated with each discrimination onto the brain, producing three maps of weights. The magnitude of each weight indicated how much the activation of a given voxel

'pushes' the distributed representation away from the decision boundary, while the sign of the weight indicated to which side of the boundary the representation is being 'pushed'. The authors found that these solutions did overlap somewhat, meaning that some of the same voxels that were very indicative of 'chairs' were also very indicative of 'faces', and so on. Such results are particularly compelling given that the goal of linear discriminant analysis is to find the voxels that maximally discriminate between the three stimuli types. In principle this means that, so long as there are sufficient voxels responding uniquely to each category, other voxels showing similar responses across two categories should be ignored – yet the analysis nevertheless identified voxels that appear to contribute simultaneously to two different object domains.

Heterogeneity versus homogeneity of representation

The third point of contrast concerns the degree to which elements of a distributed representation respond in the same way to objects of representation. By averaging neural responses across voxels in an individual, and again across individuals in group analysis, the standard approach appears to assume that, by and large, all elements respond to the objects of representation in the same way. For instance, it may seem reasonable to suppose that the voxels involved in coding perceptual representations of faces do so by showing consistently higher activation in response to visually presented faces than to other objects. If this assumption is valid, and given the inherently noisy nature of the measurements in functional brain imaging, voxel averaging is the correct thing to do: the noise at each voxel will cancel out across voxels, revealing the true underlying signal. The PDP view of representation, however, suggests the alternative possibility that the elements of a representation may respond to the objects of representation in quite different ways, both within and across individuals. For instance, one face-relevant voxel might show elevated activation for one subset of faces and decreased activation of another subset; another voxel might show a different pattern of increased and decreased activation across various faces; and the ensemble together might express the degree to which different faces are perceptually similar. Since what matters is the similarity structure taken across elements, the responses of a single element within the representation can vary almost arbitrarily on this view, both in an individual subject and across different subjects.

At first blush, there seems to be a substantial body of evidence in favour of representational homogeneity, both within and across subjects. After all, univariate methods that rely heavily on homogeneity have been applied effectively to fMRI data and yield consistent results, which would not be possible if neural responses to stimuli were purely heterogeneous and arbitrary across individuals. And, indeed, it has been demonstrated that cross-subject classification using multivariate classifiers is possible, albeit on coarse distinctions such as discriminating different tasks (Poldrack, Halchenko, & Hanson, 2009), sentences vs. pictures (Wang, Hutchinson, & Mitchell, 2004) or line drawings of tools vs. dwellings (Shinkareva et al., 2008).

Although these findings demonstrate that individual brains share important structure, they do not demonstrate representational homogeneity *per se*. To see this, consider the study of Wang et al. (2004) and its later reanalysis (Rao, Cox, Nowak, & Rogers, 2013). The data were acquired while participants completed a cross modality match-to-sample task: one of the stimuli was a simple configuration of two symbols, and the other was a sentence which either did or did not correctly describe the configuration of symbols. Stimulus order was counterbalanced, and the goal was to determine, from the evoked BOLD response at a given time, whether the participant was reading a sentence of viewing an image. In the original study, classification across individuals was achieved by averaging voxel BOLD response within a small number of anatomically defined ROIs and training a classifier using the averaged time series data from all but one participant. The solution was then used to classify each time-point in the functional data from the hold-out individual, and the results showed reliable above-chance performance. The analysis thus indicates a degree of consistency across individuals in the mean response to these different stimuli across coarse brain regions.

Still, the averaging at a broad grain ends up revealing little about the nature of the representations within and across individuals beyond this general consistency. Rao et al. (2013) looked for representational structure at a finer grain within and across subjects, using a whole-brain multivariate pattern classification method in which the responses of every individual voxel were provided as input, rather than the mean response averaged over preselected ROIs. To avoid over-fitting, the classifier employed a regularisation penalty that preferred sparse solutions (i.e., most voxels receive weights of zero) in which selected voxels were located in roughly similar anatomical regions across participants (the SOS Lasso; see Rao et al., 2013 for a more detailed explanation). In one sense, the analysis replicated the original study: the majority of voxels that the classifier selected fell within the ROIs determined to be most informative by Wang et al. (2004). The classifier solution also differed from that implied by the original analysis in important respects, however. Specifically, it did not identify some regions where all the weights were positive for all subjects (indicating, for instance, increased activation for sentences relative to pictures) and other regions where all the weights were negative (indicating the reverse). Instead, all regions identified included a mix of both positive and

negative weights, consistent with the view that the representational code – whether high activation is observed for pictures or for words – can be heterogeneous even within a given circumscribed region, both within and between subjects.

Within this general mix, some regions showed a generally higher proportion of positive weights and others a generally higher proportion of negative weights, suggesting one explanation of the original result: when averaging across voxels within an region of interest (ROI), the mean activity may carry signal because a majority of the underlying voxels code the information of interest in a particular way. But the analysis shows that such averaging can mask considerable underlying heterogeneity in the representational code.

Heterogeneity versus homogeneity of location

The final point of contrast concerns the degree to which representations are localised homogeneously within and across individuals. The adoption of ROI averaging, cluster-thresholding and spatial smoothing require the underlying assumption that the elements contributing to a given representation will be located near one another anatomically within subjects, whereas the anatomical alignment and averaging across participants require the additional assumption that localisation will be largely consistent across individuals. The PDP view of representation, in contrast, suggests that the elements of a distributed representation may in fact vary substantially in their anatomical location both within and across individuals.

Several studies have now suggested that, in a variety of cognitive domains, neural representations are not confined to a small number of discrete and homogenous cortical regions but can be quite widely anatomically distributed. Recall that, in the study by Riggall and Postle (2012) discussed earlier, the authors were able to decode the direction of motion being held in working memory from activation patterns measured in occipital cortex. The same study further showed, however, that classification accuracy improved significantly when the logistic ridge-regression classifier was trained on data from the whole brain. The information maps generated from this analysis suggested the direction-of-motion signal was encoded in a very widely distributed cortical network and not solely within a discrete region of visual cortex. Moreover, separate classifiers were trained and tested for each individual participant, so that the result did not arise from variability across subjects but illustrated heterogeneity of location within individual participants.

A similar result was obtained in a different domain in an interesting study by Bulthé, De Smedt, and Op de Beeck (2014). These authors applied multi-voxel searchlight, region of interest and whole-brain classifiers to the same fMRI dataset, where the task was to decode numeric magnitude either from trials where Arabic numerals were presented (symbolic) or from trials where arrays of dots were presented (non-symbolic). This is a particularly interesting case, because prior univariate analyses implicated the intraparietal sulcus (IPS) as functionally specific for numerical magnitude, regardless of the stimulus modality (e.g., Dehaene & Cohen, 1997). The results indicated that both symbolic and non-symbolic magnitudes could be decoded from all lobes of the brain, and that whole brain decoding was on par with, if not better than, decoding from any individual lobe. The ROI analysis indicated that numeric magnitude could be decoded from nearly all ROIs during the non-symbolic trials (the visual word form area being the only exception), while only the IPS, fusiform, inferior occipital, left superior parietal and the right superior frontal gyrus supported the decoding of magnitude during the symbolic trials. Finally, the searchlight analysis revealed that while non-symbolic magnitude could be decoded locally almost everywhere in the brain, symbolic magnitude could not be decoded anywhere from such local information. Thus, in this case, there appears to be information distributed across very widely situated voxel sets that cannot be extracted at more local scales, even by multivariate methods.

As a third example, Rish, Cecchi, Heuton, Baliki, and Apkarian (2012) used elastic-net classifiers to predict judgements about the magnitude of a perceived stimulus in three quite different tasks, including magnitude judgements for visual object size, velocity of motion and pain intensity. In each task, an elastic net regression was run to select the 1000 most predictive voxels, a procedure that identified widely distributed sets of voxels that reliably predicted the magnitude of the pain. The authors then reran the analysis after excluding the 1000 voxels identified on the first run and found to their surprise that the predictive accuracy of the new solution declined only negligibly relative to the original one. This process was repeated until performance reached floor. Remarkably, predictive accuracy in all three tasks declined very slowly. The authors interpreted this result as indicating that some kinds of information, such as stimulus magnitude, may be very broadly represented in the brain.[2]

Each example suggests that, at least in these particular cases, voxels that contribute to the discrimination of different cognitive states need not be situated near one another within a small set of cortical regions. What about localisation across individuals? Is it possible that neural representations, even if they are widely dispersed anatomically within individuals, are nevertheless anatomically situated in similar ways across individuals?

The question can be very directly and elegantly addressed by leveraging a simple insight: if a representation is localised in the same way across a sample of

subjects, the alignment of functional data should improve as the anatomical alignment improves. In turn, improving the functional alignment should increase the effect size in a univariate analysis. Tahmasebi et al. (2012) systematically varied how well participants' brains were anatomically aligned within a common space by applying a series of increasingly precise methods. He then assessed whether better anatomical alignment subsequently led to stronger effects in the analysis of functional data. In the experimental paradigm, subjects listened to sentences with ambiguous words ('His new post was in China'), matched unambiguous sentences ('The old tree was in danger'), signal-correlated noise (unintelligible noise matched to the intelligible sentences with respect to their length, spectral profile, and amplitude envelope) and silence in equal measures. Prior work had established where different univariate contrasts should produce reliable effects: the contrast of sound to silence should activate the auditory thalamus, for instance, whereas the contrast of sentences to noise should activate primary auditory cortex, and the contrast of ambiguous to unambiguous sentences should activation the left posterior inferior temporal gyrus and the left inferior frontal gyrus. With these predicted effects, the central question was whether improved anatomical alignment would increase the functional effect size in the relevant regions for each contrast of interest. The authors found that such an increase was indeed observed in the auditory thalamus, where auditory codes are presumably highly localised and consistent across subjects. A similar but weaker influence of alignment was also observed in primary auditory cortex, again consistent with the view that representations in this region should be relatively consistent across participants. This result was not obtained, however, for the contrast of ambiguous to unambiguous sentences. The size of the ambiguity effect was independent of the quality of the anatomical alignment, suggesting that the processes underlying ambiguity resolution are not anatomically localised in precisely the same way across subjects.

Other work has very directly assessed the degree to which the location of representational and processing structure varies across individuals. In one particularly compelling study, Feredoes, Tononi, and Postle (2007) considered a discrepancy in the neuroimaging literature related to working memory: group-level analyses tend to yield data consistent with the hypothesis that the prefrontal cortex (PFC) serves as a working memory buffer, evidenced by a delay-period sensitivity to memory load, whereas single-subject case study analyses tended to not show this effect. Instead, single-subject analyses implicated quite different regions in different people. These single-subject effects tended to be of greater magnitude than the group level effect in the PFC, leading to the hypothesis that working memory is supported by different regions in different people, with only weak involvement of the PFC in any individual. Because the weak PFC activity is more consistently localised across individuals, however, this is the region that emerges in the group-level univariate analysis. An alternative hypothesis, and the one typically adopted in standard image analysis, is that the single subject effects are just noise. To adjudicate these interpretations, the authors applied transcranial magnetic stimulation (TMS) in each participant to either the PFC region identified in the group analysis or to an individual-specific location corresponding to the region of greatest activation during delay-period in the fMRI session. Larger effects were observed when TMS was applied to the individual hotspots than to the shared PFC region – suggesting that these regions, which were heterogeneous in location across individuals, nevertheless were playing a more important role in supporting the working memory task.

General discussion

The preceding review supports the face validity of the representational claims staked by PDP. There is at least some evidence in at least some cognitive domains that neural representations measured at the scale of fMRI possess each of the four properties of distributed representations articulated earlier. What are the implications of these observations for our developing understanding of representation in the mind and brain? In particular, what can be concluded from the body of neuroimaging literature that has relied on univariate analyses, and how should future efforts proceed?

First, it seems clear that claims of regional specialisation based on classical functional brain imaging – a face area, a visual word-form area, a magnitude-estimation area and so on – need to be revisited. The primary evidence for such areas was obtained using univariate methods that are capable of revealing where sets of anatomically contiguous voxels individually show similar patterns of contrast across groups of participants. The concern is not that such findings are literally wrong or that they do not address important issues; rather, it is that their significance for understanding the neural bases of cognition is unclear because the methods that were used are not also capable of uncovering other types of neural structures, in particular distributed representations of the kind for which the PDP framework provides a computational rationale. A question for future research is whether the same results will be obtained using methods that are capable of detecting other types of representations as well. The literature just reviewed indicates that methods suited to finding distributed representations can yield quite different answers about both the location and the nature of neural representations, but much more research is required.

In general, perspectives on neural representation and how it ought to be studied are shifting. As a field, we are

scrutinising not only our methodological rigour (Vul et al., 2009) but the capabilities of our techniques and what inferences can and cannot be safely drawn from them (Poldrack, 2006, 2011). A recent publication by Davis et al. (2014) scrutinised the natural assumption that the difference between univariate analysis and MVPA, and lies primarily in MVPA's sensitivity to high-dimensional information. They clearly demonstrate via simulation what we have argued in this review: univariate methods are only sensitive to signal that meets the set of representational assumptions outlined in our introduction. Differences between univariate and multivariate analyses can thus arise for many reasons, and not solely because MVPAs are able to exploit high-dimensional information. Consequently, after observing a difference between a univariate and multivariate analysis, one's data may remain consistent with many representational hypotheses – such differences do not, in themselves, indicate what the 'right' representational structure is.

A second important observation is that the multivariate methods to which we have alluded are not themselves all of a piece. Many different methods are currently being explored in the literature, and each carries its own underlying assumptions about what neural representations might be like. As just one example, consider the popular searchlight approach to multivariate pattern classification discussed earlier (Kriegeskorte et al., 2006). In this method, a 'searchlight' of fixed radius is centred on each voxel in the brain; voxels within the radius are treated as inputs to a multivariate pattern classifier; for each such searchlight in the brain the associated classifier is trained to discriminate different cognitive states. The classifier's cross-validation accuracy is stored in the voxel corresponding to the searchlight's centre, producing, for each subject, an 'information map' that indicates where classification accuracy from associated searchlights exceeds chance. These information maps are then typically subjected to the standard cross-subject univariate significance tests, with the goal of identifying contiguous sets of voxels whose searchlights can be reliably decoded by the classifier across subjects. As was illustrated in Figure 2, this approach relaxes some assumptions of the standard univariate approach – for instance, voxels are not treated independently, but are considered in sets, so that each classifier can decode information reliably distributed over a searchlight even if the information is not reliably expressed in each individual voxel. Likewise, the searchlight approach discards the assumption that elements of a representation must all respond to their preferred stimuli in similar ways. But the approach holds other assumptions in common with univariate methods: for instance, it assumes that the information of interest is distributed locally at the scale of individual searchlights, and that it is localised consistently across subjects. Other multivariate methods, such as whole-brain ridge regression (e.g., Riggall & Postle, 2012), discard these assumptions but adopt still others. Each such method, by virtue of its underlying assumptions, is constrained in the kinds of signal it is capable of detecting, just like the univariate approach (Cox & Rogers, submitted). Consequently, each method can potentially yield substantially different results when applied to the same data. Indeed, the study by Bulthé et al. (2014) and colleagues cited earlier has already shown that the different methods do indeed lead to quite different results in the search for neural representations of numerical magnitudes. What then is the researcher to do? Which method is 'correct'?

One approach is to compare and contrast the results yielded by different methods, as in several studies we have discussed. Aided by a good understanding of each method's underlying assumptions, and hence of its blind spots, such an approach can at least clarify what the data actually show. For instance, in the study by Riggall and Postle (2012), the contrast of multivariate and univariate results makes it clear that fronto-parietal regions show elevated delay-period activity without containing information about the contents of working memory, whereas occipital regions contain such information without showing elevated delay-period activity. The contrast makes it quite clear that the different regions are playing different roles in the memory task, and it is not difficult to see how such differences contribute to advancing theories about the neural bases of working memory. It seems possible, then, that simply laying the results of different analyses side-by-side and understanding how and why they differ might be an important advance.

Still, something about this strategy – always assess the data using all available methods – seems unsatisfying. The alternative is to tie the choice of methods to the question being addressed, in this case, the nature of neural representations. Given a theory of what neural representations are like, it can then be determined which methods are suitable for testing the theory and alternatives to it. This perspective is quite the opposite of that adopted in much of the literature, where one typically begins with a standard method of analysis and applies it to determine what neural representations are like. If the standard method is predicated on the wrong representational assumptions, such an approach is likely to be misleading.

This is why we feel it is useful to begin to connect methods for functional brain imaging more directly to the PDP framework. The PDP assumptions about representation articulated earlier were not drawn from thin air – they arise from a coherent theoretical framework for cognition that has already proven highly successful at explaining how cognitive phenomena can arise from neural processes abstracted at a level comparable to that probed by brain imaging. That is, the representational assumptions of PDP have already been validated at the cognitive level (see Rogers & McClelland, 2014 for a current review), so there

is good reason to develop statistical methodologies that begin with these assumptions. The new multivariate methods that have emerged in the last 10 years are a clear advance in this direction; the field awaits a deeper understanding of their properties and the kinds of questions they can and cannot address.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes

1. There are of course learning algorithms, such as the self-organizing map, that constrain anatomically neighbouring units to respond to similar inputs, but neural networks function just as well without such constraints.
2. Another possibility is that the tasks induce whole-brain metabolic changes that are correlated with stimulus magnitude but not involved in the cognitive representation of magnitude, an explanation that seems likely especially in the case of pain perception.

References

Arthurs, O. J., & Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? *Trends in Neurosciences*, *25*(1), 27–31. doi:10.1016/S0166-2236(00)01995-0

Behrmann, M., & Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, *17*, 210–219. doi:10.1016/j.tics.2013.03.007

Bulthé, J., De Smedt, B., & Op de Beeck, H. P. (2014). Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. *NeuroImage*, *87*, 311–322. doi:10.1016/j.neuroimage.2013.10.049

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704–717. doi:10.1162/jocn.2003.15.5.704

Cox, C. R., & Rogers, T. T. Taking distributed representations seriously. Manuscript submitted for publication

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, *97*, 271–283. doi:10.1016/j.neuroimage.2014.04.037

Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, *33*, 219–250. doi:10.1016/S0010-9452(08)70002-9

Feredoes, E., Tononi, G., & Postle, B. R. (2007). The neural bases of the short-term storage of verbal information are anatomically variable across individuals. *The Journal of Neuroscience*, *27*, 11003–11008. doi:10.1523/JNEUROSCI.1573-07.2007

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*, 199–204. doi:10.1016/j.neuron.2009.03.017

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York, NY: Cambridge University Press.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. doi:10.1016/j.tics.2005.11.006

Henson, R. N. A., & Rugg, M. D. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, *41*, 263–270. doi:10.1016/S0028-3932(02)00159-8

Jimura, K., & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*, 544–552. doi:10.1016/j.neuropsychologia.2011.11.007

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience*, *17*, 4302–4311.

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, *3*, 363–373. doi:10.3389/neuro.01.035.2009

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 3863–3868. doi:10.1073/pnas.0600244103

Logothetis, N. K., & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, *66*, 735–769. doi:10.1146/annurev.physiol.66.082602.092845

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: Structure and processes. *Current Opinion in Neurobiology*, *11*, 194–201. doi:10.1016/S0959-4388(00)00196-3

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*, 424–430. doi:10.1016/j.tics.2006.07.005

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1 Suppl 1), S199–S209. doi:10.1016/j.neuroimage.2008.11.007

Piazza, M., Pinel, P., Le Bihan, D., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*, 293–305. doi:10.1016/j.neuron.2006.11.022

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535. doi:10.1126/science.1857983

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. doi:10.1037/0033-295X.103.1.56

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data?. *Trends in Cognitive Sciences*, *10*, 59–63.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, *72*, 692–697.

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, *20*, 1364–1372. doi:10.1111/j.1467-9280.2009.02460.x

Posner, M. I., Petersen, S. E., Fox, P. T., & Raichle, M. E. (1988). Localization of cognitive operations in the human brain. *Science*, *240*, 1627–1631. doi:10.1126/science.3289116

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in

the human brain. *Nature*, *435*, 1102–1107. doi:10.1038/nature03687

Rao, N., Cox, C., Nowak, R., & Rogers, T. T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 2202–2210). Curran Associates. Retrieved from http://papers.nips.cc/paper/4891-sparse-overlapping-sets-lasso-for-multitask-learning-and-its-application-to-fmri-analysis.pdf

Riggall, A. C., & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *The Journal of Neuroscience*, *32*, 12990–12998. doi:10.1523/JNEUROSCI.1892-12.2012

Rish, I., Cecchi, G. A., Heuton, K., Baliki, M. N., & Apkarian, A. V. (2012). Sparse regression analysis of task-relevant information distribution in the brain. *SPIE*, *8314*, 831412–831418. doi:10.1117/12.911318

Rissman, J., Eliassen, J. C., & Blumstein, S. E. (2003). An event-related fMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience*, *15*, 1160–1175.

Rogers, T. T., & 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

Rumelhart, D. E., McClelland, J. L., & Hinton, G. E. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition* (Vols. 1–2, Vol. 1). Cambridge, MA: MIT Press.

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*, 1190–1228. doi:10.1111/cogs.12147

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE*, *3*(1), e1394. doi:10.1371/journal.pone.0001394

Smolensky, P. (1986). *Neural and conceptual interpretations of parallel distributed processing models*. Boulder, CO: Colorado University at Boulder Department of Computer Science.

Tahmasebi, A. M., Davis, M. H., Wild, C. J., Rodd, J. M., Hakyemez, H., Abolmaesumi, P., & Johnsrude, I. S. (2012). Is the link between anatomical structure and function equally strong at all cognitive levels of processing? *Cerebral Cortex*, *22*, 1593–1603. doi:10.1093/cercor/bhr205

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*, 309–318. doi:10.1016/j.tics.2006.05.009

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Wang, X., Hutchinson, R., & Mitchell, T. M. (2004). Training fMRI classifiers to detect cognitive states across multiple human subjects. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 709–716). Cambridge, MA: MIT Press. Retrieved from http://papers.nips.cc/paper/2449-training-fmri-classifiers-to-detect-cognitive-states-across-multiple-human-subjects.pdf

Wheatley, T., Weisberg, J., Beauchamp, M. S., & Martin, A. (2005). Automatic priming of semantically related words reduces activity in the Fusiform Gyrus. *Journal of Cognitive Neuroscience*, *17*, 1871–1885.

White, C. N., & Poldrack, R. A. (2013). Using fMRI to constrain theories of cognition. *Perspectives on Psychological Science*, *8*(1), 79–83. doi:10.1177/1745691612469029