Pattern-information fMRI: new questions which it opens up, and challenges which face it

Rajeev D. S. Raizada[1] and Nikolaus Kriegeskorte[2]

[1]Neukom Institute for Computational Science, HB 6255, Dartmouth College, Hanover NH 03755.
[2]Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge, CB2 7EF, UK.

Abstract

Recent years have seen a strong growth of interest in multivariate approaches for analysing brain activity patterns. The primary goal of these approaches is to reveal the information represented in neuronal population codes. Here we review how these methods have been used to relate neural activity patterns both to stimulus input and to behavioural output, and how they might help to explain individual differences in behavioural performance. We examine the neuroscientific interpretation of different types of pattern-information analysis, and highlight current challenges and promising future directions for this emerging field. The open challenges that we discuss are as follows: inferring the causal role of pattern information, seeking diagnostic power for fMRI at the level of individuals, determining whether observed patterns have real functional significance, finding the structure underlying high-dimensional activation spaces, and relating one person's neural patterns to another's.

1 Introduction: brain representations are inherently multivariate

Nothing is more multivariate than the brain. Populations of neurons, densely interconnected at both short-distance and long-distance scales, work in concert to process richly structured information. From this perspective, it may seem odd that the dominant approaches for analysing the brain have been univariate in nature. However, simple models have many virtues: they are directly testable, easily interpretable, and computationally tractable. Although they only reveal a part of the picture, that which they do reveal can be useful and important. For example, in neurophysiology, information about a great many stimuli and tasks has been found in the responses of single neurons (Parker & Newsome, 1998).

Functional neuroimaging may not at first sight appear to have taken a univariate approach. After all, a whole volume of voxels is acquired at once, and a General Linear Model (or GLM) is used to statistically analyse *all* of these voxels. However, each voxel is analysed individually: the model relates the experimental design to the time-course from that voxel alone. For this reason, such analyses can be described as "massively univariate" (Luo & Nichols, 2003). Note that the term "univariate" here refers to the fact that such analyses model each dependent variable (i.e. each voxel) individually. Multiple predictor variables may still be used to model different aspects of the experimental design (hence, the term "multiple linear regression").

In classical brain mapping analyses, the data are typically spatially smoothed so as to focus sensitivity on overall activations of functional regions. As a consequence, population-code information reflected in subtle differences between nearby voxels may be lost, as is schematically illustrated in Fig. 1A (but see Op de Beeck (2009) and the recent debate in NeuroImage: Gardner (2009); Kamitani & Sawahata (2009); Kriegeskorte et al. (2009); Shmuel et al. (2009)). Importantly, univariate statistical models do not encode relationships between voxels. Instead each voxel's activation value is modeled separately, so as to detect brain regions that respond more strongly during one experimental condition than during another.

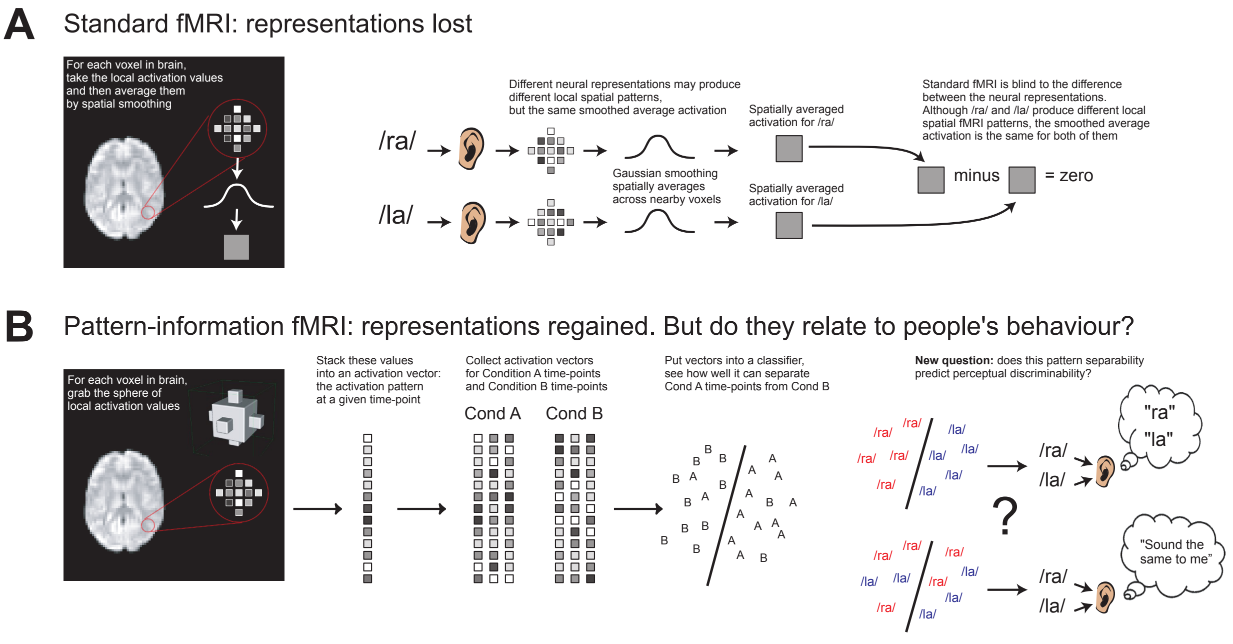
A
Standard fMRI: representations lost
For each voxel in brain, take the local activation values and then average them by spatial smoothing
/ra/
/la/
Different neural representations may produce different local spatial patterns, but the same smoothed average activation
Gaussian smoothing spatially averages across nearby voxels
Spatially averaged activation for /ra/
Spatially averaged activation for /la/
Standard fMRI is blind to the difference between the neural representations. Although /ra/ and /la/ produce different local spatial fMRI patterns, the smoothed average activation is the same for both of them
minus
= zero
B
Pattern-information fMRI: representations regained. But do they relate to people's behaviour?
For each voxel in brain, grab the sphere of local activation values
Stack these values into an activation vector: the activation pattern at a given time-point
Collect activation vectors for Condition A time-points and Condition B time-points
Cond A
Cond B
Put vectors into a classifier, see how well it can separate Cond A time-points from Cond B
A
B
New question: does this pattern separability predict perceptual discriminability?
/ra/
/la/
?
"ra"
"la"
"Sound the same to me"

Figure 1: Schematic illustration of how conventional fMRI analysis **(A)** is unable to distinguish between activation arising from distinct neural representations within the same brain area. In contrast, pattern-information fMRI (Kriegeskorte et al., 2006) is able to distinguish between such representations by examining spatial fMRI patterns, as opposed to computing statistics for each voxel considered individually **(B)**. This potentially opens up many new questions that were previously inaccessible to fMRI. One such question is whether these distributed fMRI patterns can be related to people's behavioural performance. An example, from (Raizada et al., 2009a), is illustrated at the right-hand end of panel B: the hypothesis was that if the fMRI patterns elicited by /ra/ and /la/ stimuli were separable from each other, then the listener would be able perceptually to tell /ra/ and /la/ apart. Conversely, if the patterns were too intermingled to be statistically separable, then /ra/ and /la/ would not be discriminable by the listener. Raizada et al. (2009a) tested this hypothesis in native English speakers and Japanese speakers, and found that neural pattern separability did indeed correlate with behavioural performance, not only across groups but also across individuals.

In a multivariate analysis, multiple responses are *jointly* tested for differences between experimental conditions. In fMRI, the responses are from a collection of voxels, with the number of voxels ranging over anything from a handful to tens of thousands. In neurophysiological studies, the responses are from a collection of electrode measurements of single-neuron or multi-unit activity, recorded in animals (Hung et al., 2005; Kiani et al., 2007; Meyers et al., 2008; Mesgarani et al., 2008) or even in humans (Quiroga et al., 2007; Liu et al., 2009). The common factor linking all of these studies is that by jointly analysing multiple neural responses at once they try to probe the brain's distributed population codes. This approach opens up new questions and ways of thinking about neural representations which are difficult or even impossible to formulate from a traditional univariate point of view. In the present review, we highlight some of these novel questions and the insights that they have yielded, and discuss some of the difficult but exciting challenges which face this new area of research.

2 Recent developments in the literature, and the aims of this review

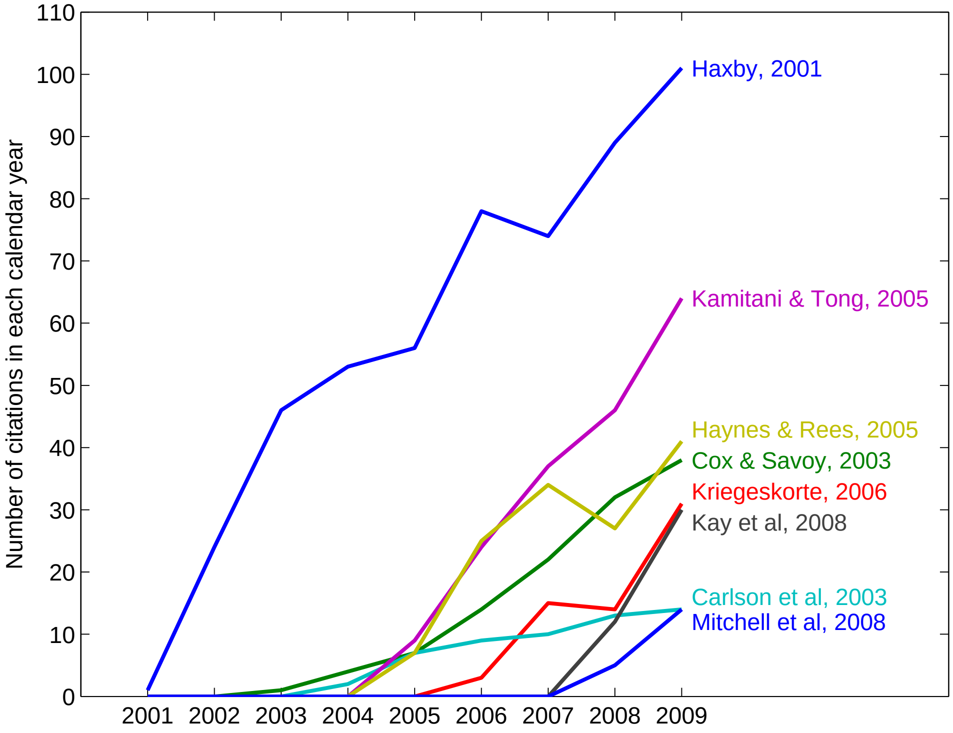

Figure 2: Pattern-information fMRI is a rapidly growing field. This graph shows the citation counts of some key papers (Haxby et al., 2001; Cox & Savoy, 2003; Carlson et al., 2003; Kamitani & Tong, 2005; Haynes & Rees, 2005; Kriegeskorte et al., 2006; Kay et al., 2008; Mitchell et al., 2008) , using data from ISI Web of Knowledge, as of mid-November 2009. Values for the full year 2009 are projected from the data up to mid-November, by multiplying that value by 12 / 10.5. Note that the y-axis shows the number of citations in each calendar year, not the cumulative total number of citations.

Population analyses have a long history in electrophysiology (e.g., Georgopoulos et al., 1986). In functional imaging as well, multivariate analyses have been explored early on (e.g., Worsley et al., 1997). However, the current dynamic started with Haxby et al. (2001), who used pattern-information analyses to investigate the object-category information contained in multivoxel fMRI activation patterns in the human ventral-stream. Following this seminal paper, interest in this area has grown rapidly. Figure 2 shows the growth of citation counts for some key fMRI pattern-information papers. A number of papers using this new approach have been published in high-profile journals, including Nature (Kay et al., 2008; Peelen et al., 2009), Science (Haxby et al., 2001; Polyn et al., 2005; Mitchell et al., 2008; Formisano et al., 2008; Li et al., 2008; Knops et al., 2009; Schurger et al., 2009) and Nature Neuroscience (Kamitani & Tong, 2005; Haynes & Rees, 2005;

Williams et al., 2007; Soon et al., 2008; Howard et al., 2009).

In response to this growing interest, several review articles have appeared. Conceptual reviews on multivariate "decoding" or "mind reading" can be found in Norman et al. (2006), Haynes & Rees (2006) and and O'Toole et al. (2007). The synergies between pattern-information analysis and high-resolution fMRI have been explored in Kriegeskorte & Bandettini (2007). A gentle methodological tutorial is provided by Mur et al. (2009). More technical aspects of the machine-learning algorithms which are used to analyse such data are reviewed for fMRI in Pereira et al. (2009), and for EEG in Lotte et al. (2007) and Muller et al. (2008). A review of clinical and pediatric applications of pattern-information analyses can be found in Bray et al. (2009), and an overview of the potential of fMRI for implementing brain-computer interfaces is given by Sitaram et al. (2009).

The present review does not seek to duplicate those efforts. Instead, we wish to describe what we believe to be some of the most promising new research directions, important under-explored issues, and key future challenges which face the field. We will highlight two topics in particular: the neuroscientific interpretation of pattern-information results, and how the approach is moving beyond looking only at brain activity patterns by seeking to relate them to individual differences in perception and behaviour.

3 Activation versus pattern information

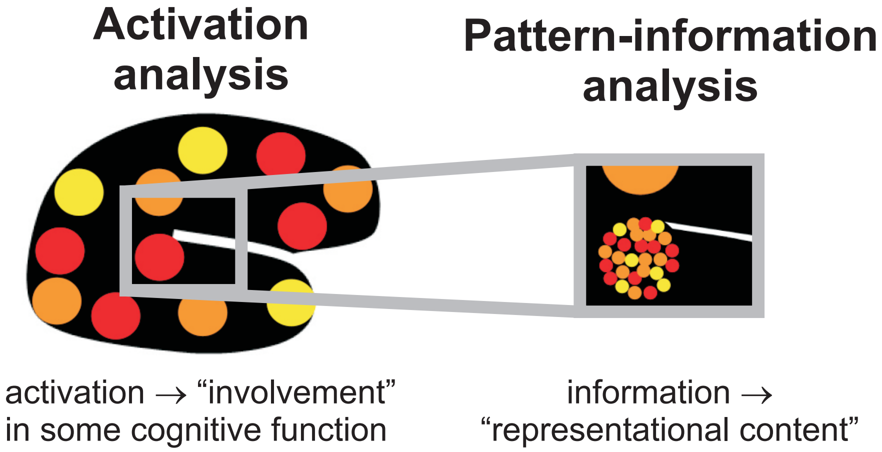
Activation analysis
Pattern-information analysis
activation → "involvement" in some cognitive function
information → "representational content"

Figure 3: Looking into brain regions to reveal representational content. The neuroscientific rationale for pattern-information analysis is to infer a brain region's "representational content" from the presence of information (about stimuli or responses) in the activity patterns. This is in contrast to the classical brain-mapping paradigm, whose neuroscientific rationale is to infer a region's "involvement" in some cognitive function from its (spatially averaged) overall level of activation.

Classical activation analysis aims to reveal a region's "involvement" in some cognitive function. Pattern-information analysis, by contrast, aims to look into each region and reveal its "representational content" by testing for combinatorial effects (Hanson et al., 2004), as is illustrated in Figure 3. Consider the fMRI activation of a single voxel. When different experimental conditions come and go, this voxel may become either more or less active, and we can calculate the activation difference between the various states along this one-dimensional continuum. However, consider the joint state of the activity of two voxels. Different experimental conditions now correspond to positions in a two-dimensional activation space. This space has a far richer structure: a great many different states are possible, and they can be similar to each other, i.e. bunched together in the space, or dissimilar, i.e. dispersed far apart. Multivariate analysis typically deals with much higher-dimensional spaces (i.e. greater numbers of response variables). The multivariate approach is sensitive to the combinatorial effects that lend a neuronal population code its representational power. Population codes can thus be quantitatively investigated and related to those most richly structured sets of phenomena: perceptual and cognitive content, and behaviour.

By using fMRI to look for information rather than activation, new types of questions can be formulated and addressed. Pattern-information analysis is sensitive to subtle distributed effects. However, it would be a mistake to view multivoxel analyses as simply asking the same questions as standard massively univariate analyses but with greater sensitivity. Pattern- information analyses allows us to ask different questions of the data. One example is that they allow the study of spatially overlapping neural representations, in a way which standard univariate analyses do not. Standard analyses smooth away the distinctions between distinct patterns of fMRI activation which are colocalised in the same region, as illustrated in Fig. 1A, with the result that they will be blind to any pattern differences which do not also happen to correspond to differences in average local activation intensity. Many pattern-information fMRI studies have investigated overlapping representations of precisely this sort, starting with Haxby et al. (2001). Examples include distinguishing between feature-based attentional signals which both spread globally across the visual field but which are directed to different motion orientations (Serences & Boynton, 2007), different phonemes whose representations overlap in auditory cortex (Formisano et al., 2008; Raizada et al., 2009a), spatially overlapping processing of different aspects of visual form object categories (Hanson et al., 2004; Downing et al., 2007; Peelen & Downing, 2007; Kriegeskorte et al., 2008a) and of faces (Kriegeskorte et al., 2007), and the representations of different odours all colocalised within posterior piriform cortex (Howard et al., 2009).

However, a greater advantage that arises from investigating multivariate population codes is that it opens up a rich new way of thinking about neural representations: we can consider them in terms of the structure of similarity space. We look at some specific examples of this in the next section.

4 Multivariate approaches open up new questions: the structure of similarity space

Univariate measures, which look at just a single voxel or a single neuron at a time, can only go up or down. However, multivariate measures correspond to positions in a multidimensional activation space, which has a far richer structure. This allows new questions to be asked: for example, how do the representational similarity structures of human and macaque inferior temporal cortex compare to each other? A recent study (Kriegeskorte et al., 2008a) found that the similarity structures from human and monkey are remarkably alike. This could be demonstrated at the level of pattern-similarity structure despite the fact that humans were measured with fMRI and macaques with invasive cell recording. Two visualisations of the similarity space found in monkeys are shown in Figure 4 (Kiani et al., 2007). Univariate analysis has the advantage of simplicity. However, multivariate approaches are needed to relate neural activity patterns to complex mental rep-

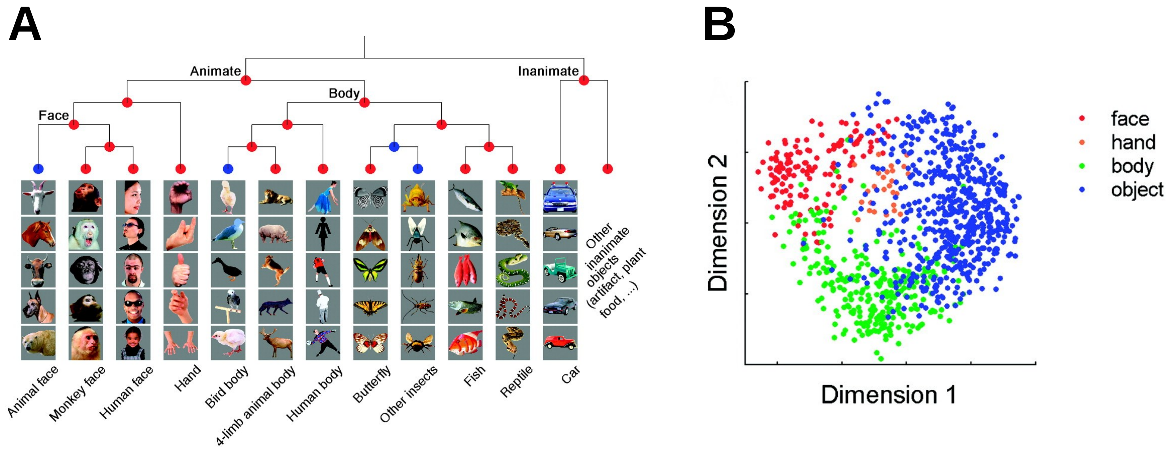

Figure 4: Kiani et al. (2007) investigated the similarity space of responses across more than 600 neurons in monkey inferotemporal cortex to a set of visual object stimuli. Different aspects of resulting similarity space can be visualised using **(A)** a cluster diagram, and **(B)** multidimensional scaling. Reproduced with permission from Kiani et al. (2007).

resentations. Cognitive psychology has a rich tradition of using similarity to investigate the structure of mental representations across multiple perceptual, linguistic and semantic domains (Tversky, 1977; Shepard, 1987; Nosofsky, 1988; Edelman, 1998; Medin, 1989). Moreover, psychologists developed the tool of multidimensional scaling in order to project complex high-dimensional similarity spaces onto a more visualisable lower-dimensional representation (Shepard, 1962; Kruskal, 1964), and also methods for combining similarity spaces across multiple subjects (e.g. INDSCAL: Carroll & Chang, 1970). Functional MRI investigations of similarity space can build upon the insights from that body of work. Comparing representational dissimilarity matrices from different sources provides a general framework (Kriegeskorte et al., 2008b) for relating not only species and measurement modalities, but also to brain representations to behaviour and to computational models.

5 "Prediction," "decoding," inferential statistics, and classifiers

A popular basic variant of pattern-information analysis is response pattern classification. In this approach, the experimental conditions (e.g. the stimuli) are "predicted" from the activity patterns they elicit. We put "prediction" in quotes here, because it does not refer to the prediction of future events or of subsequent brain-function dynamics. We can interpret the term in the context of an imaginary game of "Give me the response patterns, and I will tell you the stimuli." This paradigm is also referred to as "decoding" (Mitchell et al., 2004; Haynes & Rees, 2006; Friston et al., 2008). The rationale for this approach is that if "prediction" works better than chance, then there must be information about the stimuli in the response patterns.

The sense in which a classifier performs prediction is that it can be used to make inferences beyond the data that is presented to it, in other words it generalises from a training set to a testing set. However, such generalisation is not unique to classifiers: the whole purpose of standard inferential statistics is to reason from a sample to the broader population from which that sample is drawn. Classifiers are often used for handling multivariate data, but standard tools from inferential statistics can also be multivariate: examples include multivariate analysis of covariance (MANCOVA), Hotelling's T^2 and Wilks' lambda (Rencher, 2002). Such methods, in principle, can offer elegant and computationally inexpensive ways to model data. However, they rely on distributional assumptions (i.e. multivariate normality), which may not always hold. Non-parametric multivariate methods (e.g., Racine & Li, 2004) do not require such assumptions, but they typically involve computationally intensive permutation operations, not unlike the computation required when using classifiers with cross-validation. In summary, although the term "prediction" is often used when classifiers analyse fMRI data, the underlying logic of trying to reason beyond a given data sample to a broader population is no different from that underlying standard GLM analyses. The key difference is that pattern-information analyses consider information distributed across multiple voxels, whereas standard GLM analyses consider each voxel on its own.

It should be noted that there do exist pattern-information studies which use the word "predict" in its valid and literal sense, by making inferences about the subjects' future states, and following up and testing those predictions in a longitudinal study. These have typically been studies attempting to predict the progression of disorders, such as Alzheimer's (Fan et al., 2008a), depression (Costafreda et al., 2009) and psychosis (Koutsouleris et al., 2009).

Buzz words like "prediction", "decoding", and "brain reading" make pattern-information results attractive to a broad audience including the general public and the media. However, these terms should not be taken to imply that what is demonstrated goes beyond a statistical dependency between stimulus and response. The ability to "predict" and "decode" could equally be claimed on the basis of any classical activation analysis, such as Kanwisher et al. (1997). Consider the following potential title claims:

- "Fusiform cortex responds more strongly to faces than to other objects"
- "Fusiform activity predicts the perception of faces"
- "Face percepts can be decoded from fusiform activity"

The fact that face stimuli are correlated with stronger activation in the fusiform gyrus would justify each of these titles. The words "predict" and "decode" in the second and third title do not have any deeper implications about fusiform cortex than the activity difference claimed in the first title.

Whether we are "predicting" the stimulus from the response or the response from the stimulus, all that is demonstrated is a statistical dependency (or, equivalently, mutual information) between the two. In a univariate scenario, it is easy to see that a correlation between two variables implies predictability in both directions. In the multivariate scenario, the same holds. Demonstrating above-chance predictability in either direction implies a statistical dependency and thus above-chance predictability in both directions. The direction in which the model operates has no implications for the neuroscientific interpretation of the result. What is novel about pattern-information analysis is not "prediction" or "decoding", but the joint analysis of multiple responses as a population code. This does have neuroscientific implications distinct from those of traditional univariate analyses.

6 Beyond stimulus representations: relating pattern information to behaviour

Most pattern-information studies have focused on the relationship between stimuli and the neural response patterns that they elicit. However, a larger goal is to understand how the brain gives rise to behaviour. Instead of relating the observed activity patterns to the stimuli that elicited them, we can relate them to subjects' behavioural responses. Furthermore, we can study the extent to which brain representations of stimuli and behavioural responses vary with subject traits (e.g. personality measures or group memberships).

Along these lines, one might hypothesize that visual objects whose inferior temporal representations are similar give rise to similar behavioural responses. In a study of monkeys viewing visual objects, Op de Beeck et al. (2001) compared behavioural and neural population code similarity measures and found them to be somewhat congruent. In human fMRI studies, a number of recent studies have directly compared and found positive correlations between behavioural judgments of similarity and multivoxel fMRI pattern similarity in ventral temporal cortex (Op de Beeck et al., 2008; Weber et al., 2009; Walther et al., 2009), including changes in pattern-similarity following learning by the subjects (Op de Beeck et al., 2006; Li et al., 2008). Learning-related changes can also affect similarity judgments indirectly, by inducing changes in neural and behavioural discrimination criteria (Li et al., 2009).

Similarity judgment is only one form of behavioural response. More generally, we can ask whether distributed fMRI patterns can be related to success and failure in many other types of behaviour, and whether a given set of neural representations is well-structured or poorly structured for performing a particular task.

Figure 5A shows a cartoon example, namely that of using the stimulus dimensions of height and weight to separate sumo wrestlers from basketball players. Two points are worth noting. First, no individual dimension on its own is sufficient to separate one category from the other. It is necessary to take both height and weight into account, as evidenced by the fact that the dividing class boundary is diagonal, rather than vertical or horizontal. In the case of fMRI, the activity of one voxel plays the role of a dimension: standard voxel-by-voxel analysis uses only one voxel's activation at a time, whereas a multi-voxel pattern based analysis jointly uses information from many voxels at once.

The second point of note is that for performing the task of discriminating between sumo wrestlers and basketball players, this height/weight representation is an excellent one. Armed with such a representational structure, a person would be able to carry out this task very successfully.

However, a representational structure can also be less suitable for performing a task. In Fig. 5B, the same dimensions of height and weight are now used in an attempt to separate faculty from students. Clearly this representation will only permit poor task performance: the two categories are strongly overlapping in height/weight space, and are much less separable.

Moving form that cartoon example to a real study with data linking fMRI activation to behavioural ability, Raizada et al. (2009a) tested the hypothesis that the more separable the neural patterns elicited by /ra/ and /la/ are in a person's brain, the better that person should be behaviourally at hearing the difference between those phonemes. Thus, in native English speakers the neural patterns should be highly separable, but in Japanese speakers the patterns should be less separable,

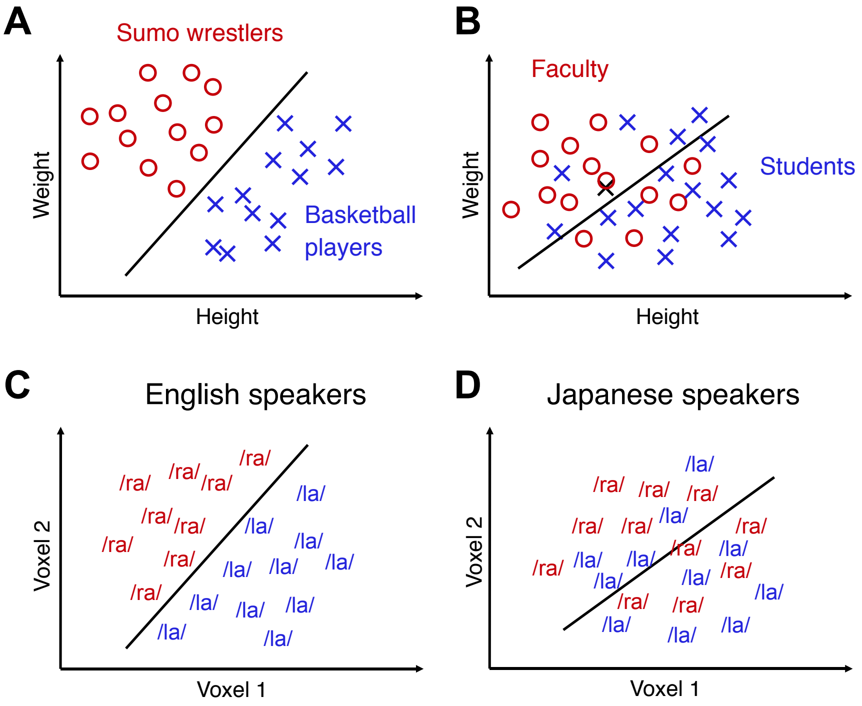

Figure 5: Cartoon examples showing a representational structure which will allow very good performance of one task **(A)**, but only poor performance of a different task **(B)**. Raizada et al. (2009a) found fMRI evidence that analogous representational structures may explain why the task of discriminating /ra/ from /la/ is easy for English speakers, but is difficult for Japanese speakers.

correspondingly the fact that Japanese speakers are less able to perceive that phonetic contrast. This hypothesis is schematically illustrated in Figs. 5C and D, and was found to hold true in the actual data. Moreover, the neural pattern-separability correlated not only with group-differences (English vs. Japanese), but also with individual differences in perceptual discrimination ability. The more separable a persons neural representations for /ra/ and /la/ were, the better they were at hearing the difference between the two sounds.

Note that this type of brain-behaviour relationship is quite different from the type of correlation that has often been found in standard fMRI analysis, in which increasing levels of neural activity in a given brain area correlate with better behavioural performance in the task that the region subserves. This type of result has been found across multiple domains: for example, more intense neural activity can predict better language ability (Demb et al., 1997; Crinion & Price, 2005). In Raizada et al. (2009a), the average *intensity* of fMRI activation was the same for /ra/ as it was for /la/. It was the distinctness of the spatial patterns of activation elicited by those sounds which made the difference.

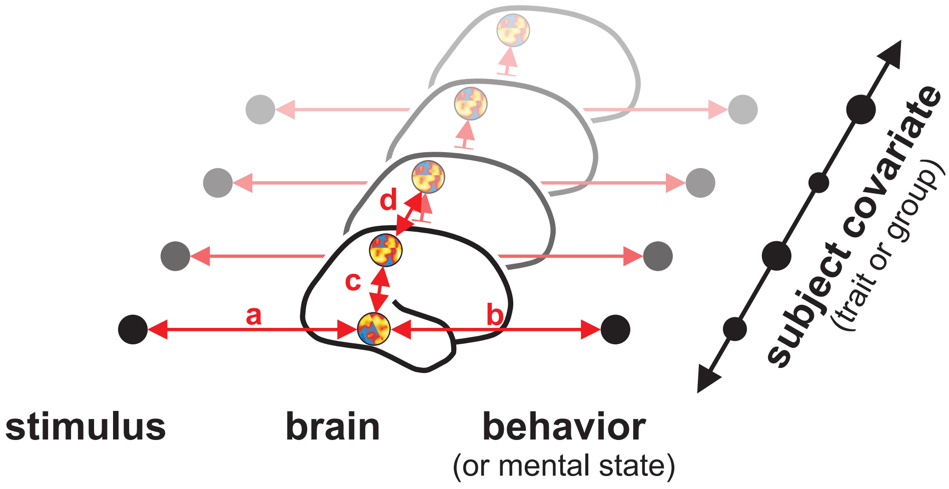

Figure 6: Representational relationships that can be investigated with pattern-information analyses. Most pattern-information studies to date have investigated the representation of experimental stimuli in neuronal population codes (a). Emerging and future applications include (b) relating population codes to behavioural variables (e.g. Op de Beeck et al., 2008; Raizada et al., 2009a), (c) relating population codes between two different brain regions (e.g. "representational connectivity", Kriegeskorte et al., 2008a), and (d) relating the population codes in corresponding brain regions between different subjects (e.g. "intersubject information", Kriegeskorte et al., 2006). Moreover, we can study how these multivariate relationships depend on interindividual differences (e.g. subject traits, group memberships, disease variables).

7 Challenges facing the field of pattern-information fMRI

7.1 Inferring the causal role of pattern information

If the stimuli in an experiment are under our control, then by making changes to the stimuli (e.g. turning them on and off) and observing consistent changes in fMRI responses, we can infer that the stimuli were causally involved in triggering the responses. In order to be able to infer a causal role of the brain activity patterns (e.g. "the population code in region X forms the basis of perceptual decision Y"), we would similarly require experimental control of the brain activity. Transcranial magnetic stimulation (TMS) allows us to experimentally influence brain activity in humans. However, TMS has low spatial precision and its effects are akin to a temporary brain lesion. It does not presently allow us to impose arbitrary precise patterns of activity. Alternatively, we can rely on assumptions to constrain the causal relationships to be considered, and then use techniques for modeling directed interactions between brain regions (also known as "effective connectivity"). For example, Granger causality (Roebroeck et al., 2005; Ramsey et al., 2009) exploits the temporal lag between cause and effect to infer causality (relying on the assumption that the model does not omit relevant alternative causal pathways). As another example, dynamic causal modeling (Friston et al., 2003) allows us to test and compare prespecified causal models of interactions between brain regions. In neuroimaging, however, these models of directed interactions are typically based on univariate activation time courses (fluctuations of spatially-averaged overall activation of the analyzed brain regions). The development of a pattern-information approach to modeling directed interactions is an important future direction. A pattern-information equivalent to undirected interactions (i.e. "functional connectivity": correlated fluctuations of overall activation between two brain regions) is provided by "representational connectivity" (Kriegeskorte et al., 2008a). A causal role of activity-pattern information, thus, is difficult to infer with present empirical and analysis techniques. However, we will see in the following parts of the paper that relating activity patterns to behavioural variables in a (multivariate) correlational framework already provides important constraints for brain theory.

7.2 Towards fMRI having diagnostic power at the level of individuals

Although fMRI has been able to reveal many important links between brain and behaviour, almost all of these results hold at the level of groups, not of individuals. Even when brain-behaviour correlations are found to hold across individuals, such as in Raizada et al. (2009a), there is almost always a prior step in which the individual-differences analysis is preceded by, and depends upon, an analysis at the group level. In the group-level analysis, potential regions of interest (ROIs) are identified on the basis of whether they show experimental effects. The individual difference results are then based upon these group-level derived ROIs, but this two-stage analysis process can potentially lead to problems of selection-bias (Vul et al., 2009; Kriegeskorte et al., 2009). Even in studies which omit the preceding group-level step by using pre-defined ROIs, those ROIs will typically have been chosen on the basis of previously published group-level studies.

The most direct way to avoid needing to select an ROI is to analyse the whole brain at once, as a single brain-wide multivoxel pattern. This approach has been extensively pursued in studies of anatomical differences between patient groups and controls. For example, Davatzikos and colleagues have investigated whole-brain multivoxel pattern differences in schizophrenia (Davatzikos et al., 2005; Khurd et al., 2007; Fan et al., 2008c) and in Alzheimer's (Davatzikos et al., 2008; Fan et al., 2008a,b; Misra et al., 2009). Other groups have also investigated multivoxel anatomical differences in several disorders, including depression (Costafreda et al., 2009) fragile X syndrome (Hoeft et al., 2008), and psychosis (Sun et al., 2009). A review of the application of such approaches to Alzheimer's can be found in Klöppel (2009).

Brain-wide functional MRI patterns, in addition to the structural MRI studies described above, can also be used to distinguish between patient and control groups, for example in schizophrenia Demirci et al. (2008a,b). An experiment of this sort in normal subjects, but with potentially direct clinical applicability, is the study of pain perception by Marquand et al. (2009).

Perhaps the most direct application to individual differences of studying brain-wide multivoxel patterns is to relate these patterns to people's different levels of ability to perform a given task. Although the study by Raizada et al. (2009a) found a relation across individuals between fMRI pattern-separability and levels of task performance, it used an ROI that needed first to be derived at the group level. However, in follow-up work, Raizada et al. (2009b) found that the same brain-behaviour correlation held true even when no ROI was used, with pattern-separability assessed using just one brain-wide multivoxel classification in each subject (note that in all such whole-brain studies, the use of cross-validation makes the number of actual computations performed per subject be more than one. The key point is that only a single brain-wide set of voxels is used for each subject). Increased brain-wide pattern separability was found to correlate with improved behavioural performance in two different tasks, from two different datasets: the /ra/-/la/ task examined in Raizada et al. (2009a), and a numerical distance-effect task from Holloway et al. (2010).

Probably the biggest potential advantage of a brain-wide approach is in the domain of diagnosis. When scanning an individual subject, the goal is to be able to make inferences immediately, without needing first to derive a group-level ROI which will tell us where in the brain to look. The numerous studies cited above showing whole-brain multivoxel pattern differences between patients and controls shows that this approach can be successful at the group level; the study showing correlations with people's levels of behavioural performance suggests that it may also hold promise at the level of individuals.

7.3 How do we know that pattern information is functionally significant?

Finding that the activity pattern in a brain region reflects some stimulus variable means that the brain region contains information about that stimulus variable. However, it does not strictly imply that this information serves the function of representing the stimulus variable in the context of the brain's overall operation.

Although we have seen that a causal role of pattern information is hard to pin down, additional evidence can come from correlations with other brain regions and with behaviour. If fMRI pattern-differences are found to correlate with behavioural differences, then that supports the claim that the pattern differences genuinely reflect processes that are functionally significant for the brain. However, behaviour is not the only such source of potential corroboration.

One method via which pattern-information analyses try to make sure that they discover real aspects of brain processing, as opposed to the idiosyncrasies of a particular data set, is cross-validation: fitting a model to a training set, and evaluating it on a testing set. However, cross-validation need not be the only approach. Brain-computer interface (BCI) systems aim to transform neural signals into the output movements of an actuator such as a robot arm in the real physical world. In that sense, they use reality as their testing set: if the robot arm moves to the wrong place, then the classifier has failed.

Because of its bulky and immobile nature, fMRI has not been as widely used for BCI as the much more lightweight and portable arrangement of EEG, although there has been some work (Ganesh et al., 2008; Sitaram et al., 2009; Lee et al., 2009). Perhaps the most promising line of fMRI investigation in this area does not involve the more usual BCI goal of controlling a motor actuator, but instead "closes the loop" by giving subjects visual feedback about their own levels brain activation, which they can then try to control in real-time. This approach may be useful for helping people to self-regulate pain (deCharms, 2008), emotion (Johnston et al., 2010), and even cognitive processes such as language (Rota et al., 2009). The fMRI signals being regulated through neurofeedback need not only be univariate activations which become more or less intense, but can also be multivoxel spatial patterns which become more or less dissimilar (LaConte et al., 2007). This promises to be an exciting but challenging area for future work, and could potentially have fruitful cross-fertilisation with principles of multi-neuronal coding derived from neurophysiological BCI studies in animals (Nicolelis & Lebedev, 2009).

7.4 Finding structure underlying high-dimensional neural representations

Pattern-information analysis deals with very high-dimensional spaces, as a single fMRI volume of the brain typically contains tens of thousands of voxels. This poses particular a challenge to analysis known as the “curse of dimensionality”: as the number of dimensions increases, the size of the space grows exponentially. When trying to find statistical structure in a high-dimensional space, one often runs into what is called “the small sample size problem,” in which there are more dimensions than there are data points (Raudys & Jain, 1991). For example, in fMRI we may have only a few hundred data points (the number of TRs), but tens of thousands of dimensions (the number of voxels). In such a space, non-degenerate classes will always be linearly separable (Cover, 1965). A classifier can then trivially obtain 100% correct on any training set, without it necessarily having captured any aspect of the underlying structure of the data. Training-set performance therefore tells us very little about how well the classifier will perform on a test set. Other fields in which this situation also commonly arises are genetics, in which we may have thousands of genes from only a few dozen patients (Li et al., 2004), face-recognition, in which a few hundred face images each contain tens of thousands of pixels (Howland et al., 2006), and chemometrics, in which spectroscopy can yield thousands of measurements per sample (Frank & Friedman, 1993). This problem is usually tackled by some combination of dimension-reduction and regularisation to constrain the space of solutions (for an overview, see the Chapter 18 of Hastie et al., 2009), and, crucially, by cross-validation for assessing decoding performance (and, thus, the presence of pattern information).

In many cases, it is reasonable to suspect that there may be a lower dimensional structure lying hidden inside the much higher dimensional space. Methods have been devised for trying to find low-dimensional manifold embeddings of this sort (Tenenbaum et al., 2000; Belkin & Niyogi, 2003; Saul et al., 2006). However, such methods can be difficult to use when the data is sparse and noisy (Balasubramanian & Schwartz, 2002), as is the case for fMRI. The best indication that a real structure has been found underlying high-dimensional fMRI data is when that structure can be used to interpolate and predict fMRI activation for novel stimuli which were not in the original training set. The studies which have achieved this to date have incorporated domain-specific knowledge into their models (Kay et al., 2008; Mitchell et al., 2008; Brouwer & Heeger, 2009), rather than deriving their structure automatically using a manifold-embedding approach.

If the aim of cognitive neuroscience is to elucidate the underlying principles of neural activation, not just to “decode” that activation by assigning labels to brain states, then finding the deeper structures hidden in high-dimensional multivoxel spaces is likely to be an important challenge in the years ahead.

7.5 Relating one person’s neural patterns to another’s

One type of structure which may lie hidden in high-dimensional multivoxel spaces is a possible mapping between one person’s fine-grained neural activation patterns and another’s. Just as all healthy human brains share common large-scale anatomical features, such as the central sulcus and the sylvian fissure, they also have a shared coarse-grained functional topography, with functionally-defined areas such as the Fusiform Face Area being present in all subjects in a roughly similar location. However, the fine-grained functional patterns (just like the finer shapes and folds of gyri and sulci) are likely to be subject-unique. A classifier trained on one person’s fine-grained spatial patterns will therefore not in general perform well on the fine-grained patterns in someone else’s brain. This raises the challenge of whether it might be possible to find a more subtle indirect mapping between one person’s fine-grained fMRI patterns and another’s.

A number of studies have found across-subject commonalities in distributed spatial patterns at a large or coarse-grained scale, using "leave-one-subject-out cross-validation," in which a classifier is trained on all but one of the subjects, and tested on the remaining one. Mourão-Miranda et al. (2005), using two tasks, and Poldrack et al. (2009), using multiple tasks drawn from several studies, were able to decode which task a subject was performing after having trained a classifier using the other subjects. Shinkareva et al. (2008) went further, and, again using leave-one-subject-out cross-validation, were able to decode not only which general category of object a person was looking at (tool vs. dwelling), but also, for eight of the twelve subjects, which specific object within that category they were looking at. However, this decoding was limited. As they wrote: "The category and exemplar classification accuracies when training across participants were on average lower than when training within participants, indicating that a critical diagnostic portion of the neural representation of the categories and exemplars is still idiosyncratic to individual participants" (Shinkareva et al., 2008, p.8).

Hasson et al. (2004) studied coarse-scale consistencies in subjects' brain-activity patterns during movie viewing. The authors correlated activity in corresponding brain regions across subjects so as to find regions driven with similar time-courses by the movie. This approach is interesting because it allows an inferential statistical brain mapping without a design matrix describing the stimulus variables. For a complex natural stimulus stream such as a movie, a design matrix is difficult or impossible to define, because there are so many potentially relevant variables describing the variation of the stimulus over time. Hasson et al.'s (2004) "intersubject correlation mapping" is a massively univariate, activation-based approach. However, it has a multivariate pattern-information equivalent, "intersubject information mapping" (Kriegeskorte & Bandettini, 2006b), where corresponding brain regions of multiple subjects watching the same movie are analyzed for shared pattern information using canonical variates analysis. This approach requires coarse-scale intersubject correspondence of functional regions, but not a fine-grained spatial consistency between the patterns encoding the shared information. Along similar lines, more recent work has sought to find between-subject mappings between seemingly idiosyncratic fine-grained patterns (Guntupalli & Haxby, 2009), although it remains to be seen whether mappings of this sort can be found which are simple and robust. If the patterns themselves are idiosyncratic to each subject, then a direct point-to-point comparison may fail despite functional correspondence of the regions considered. It may then be more appropriate to compare representations between subjects at the more abstract level of information or pattern similarity structure (Kriegeskorte et al., 2008b).

8 Conclusion

By dealing head-on with the intrinsically multivariate nature of the brain's representations, pattern-information analysis is helping us to study population codes in greater depth and breadth. The approach allows us to answer new questions, and helps the field of fMRI to shake off the old accusation of being an overly simplistic "new phrenology." Moreover, the pattern-information approach promises to help us build bridges across the traditional divides between human fMRI, animal neurophysiology, and computational modeling. However, the "bleeding edge" nature of the work brings with it several challenges, which we need to take very seriously if the approach is to fulfill its promise.

References

Balasubramanian, M. & Schwartz, E. L. (2002). The isomap algorithm and topological stability. *Science*, *295*(5552), 7.

Belkin, M. & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, *15*(6), 1373–1396.

Bray, S., Chang, C., & Hoeft, F. (2009). Applications of multivariate pattern classification analyses in developmental neuroimaging of healthy and clinical populations. *Frontiers in human neuroscience*, *3*, 32.

Brouwer, G. J. & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *J Neurosci*, *29*(44), 13992–4003.

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *J Cogn Neurosci*, *15*, 704–17.

Carroll, J. & Chang, J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, *35*(3), 283–319.

Costafreda, S. G., Chu, C., Ashburner, J., & Fu, C. H. Y. (2009). Prognostic and diagnostic potential of the structural neuroanatomy of depression. *PLoS ONE*, *4*(7), e6353.

Cover, T. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, *14*(3), 326–334.

Cox, D. D. & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, *19*(2 Pt 1), 261–70.

Crinion, J. & Price, C. J. (2005). Right anterior superior temporal activation predicts auditory sentence comprehension following aphasic stroke. *Brain*, *128*(12), 2858–2871.

Davatzikos, C., Resnick, S. M., Wu, X., Parmpi, P., & Clark, C. M. (2008). Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *Neuroimage*, *41*(4), 1220–7.

Davatzikos, C., Shen, D., Gur, R. C., Wu, X., Liu, D., Fan, Y., Hughett, P., Turetsky, B. I., & Gur, R. E. (2005). Whole-brain morphometric study of schizophrenia revealing a spatially complex set of focal abnormalities. *Arch Gen Psychiatry*, *62*(11), 1218–27.

deCharms, R. C. (2008). Applications of real-time fMRI. *Nat Rev Neurosci*, *9*(9), 720–9.

Demb, J. B., Boynton, G. M., & Heeger, D. J. (1997). Brain activity in visual cortex predicts individual differences in reading performance. *Proc Natl Acad Sci U S A*, *94*(24), 13363–13366.

Demirci, O., Clark, V., Magnotta, V., Andreasen, N., Lauriello, J., Kiehl, K., Pearlson, G., & Calhoun, V. D. (2008b). A review of challenges in the use of fMRI for disease classification / characterization and a projection pursuit application from multi-site fMRI schizophrenia study. *Brain imaging and behavior*, *2*(3), 147–226.

Demirci, O., Clark, V. P., & Calhoun, V. D. (2008a). A projection pursuit algorithm to classify individuals using fMRI data: Application to schizophrenia. *Neuroimage*, *39*(4), 1774–82.

Downing, P. E., Wiggett, A. J., & Peelen, M. V. (2007). Functional magnetic resonance imaging investigation of overlapping lateral occipitotemporal activations using multi-voxel pattern analysis. *Journal of Neuroscience*, *27*(1), 226–233.

Edelman, S. (1998). Representation is representation of similarities. *The Behavioral and brain sciences*, *21*(4), 449–67; discussion 467–98.

Fan, Y., Batmanghelich, N., Clark, C. M., Davatzikos, C., & Initiative, A. D. N. (2008b). Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, *39*(4), 1731–43.

Fan, Y., Gur, R. E., Gur, R. C., Wu, X., Shen, D., Calkins, M. E., & Davatzikos, C. (2008c). Unaffected family members and schizophrenia patients share brain structure patterns: a high-dimensional pattern classification study. *Biological Psychiatry*, *63*(1), 118–24.

Fan, Y., Resnick, S. M., Wu, X., & Davatzikos, C. (2008a). Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *Neuroimage*, *41*(2), 277–85.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "who" is saying "what"? brain-based decoding of human voice and speech. *Science*, *322*(5903), 970–3.

Frank, I. & Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*(2), 109–135.

Friston, K., Chu, C., Mourão-Miranda, J., Hulme, O., Rees, G., Penny, W., & Ashburner, J. (2008). Bayesian decoding of brain images. *Neuroimage*, *39*(1), 181–205.

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273–302.

Ganesh, G., Burdet, E., Haruno, M., & Kawato, M. (2008). Sparse linear regression for reconstructing muscle activity from human cortical fMRI. *Neuroimage*, *42*(4), 1463–72.

Gardner, J. L. (2009). Is cortical vasculature functionally organized? *Neuroimage*.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science*, *233*(4771), 1416–9.

Guntupalli, J. S. & Haxby, J. V. (2009). Inter-subject hyperalignment of neural representational space for objects. *Society for Neuroscience Abstracts*, *262.20*.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, *23*(1), 156–166.

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., & Malach, R. (2004). Intersubject synchronization of cortical activity during natural vision. *Science*, *303*(5664), 1634–40.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning, second edition: data mining, inference, and prediction* (2nd ed ed.). New York: Springer.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J. & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat Neurosci*, *8*(5), 686–91.

Haynes, J.-D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, *7*(7), 523–534.

Hoeft, F., Lightbody, A. A., Hazlett, H. C., Patnaik, S., Piven, J., & Reiss, A. L. (2008). Morphometric spatial patterns differentiating boys with fragile x syndrome, typically developing boys, and developmentally delayed boys aged 1 to 3 years. *Arch Gen Psychiatry*, *65*(9), 1087–97.

Holloway, I. D., Price, G. R., & Ansari, D. (2010). Common and segregated neural pathways for the processing of symbolic and nonsymbolic numerical magnitude: an fMRI study. *Neuroimage*, *49*(1), 1006–17.

Howard, J. D., Plailly, J., Grueschow, M., Haynes, J.-D., & Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nat Neurosci*, *12*(7), 932–8.

Howland, P., Wang, J., & Park, H. (2006). Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, *39*(2), 277–287.

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–6.

Johnston, S. J., Boehm, S. G., Healy, D., Goebel, R., & Linden, D. E. J. (2010). Neurofeedback: A promising tool for the self-regulation of emotion networks. *Neuroimage*, *49*(1), 1066–72.

Kamitani, Y. & Sawahata, Y. (2009). Spatial smoothing hurts localization but not information: Pitfalls for brain mappers. *Neuroimage*.

Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nat Neurosci*, *8*(5), 679–685.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, *17*(11), 4302–11.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Khurd, P., Verma, R., & Davatzikos, C. (2007). Kernel-based manifold learning for statistical analysis of diffusion tensor images. *Information processing in medical imaging : proceedings of the conference*, *20*, 581–93.

Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J Neurophysiol*, *97*(6), 4296–309.

Klöppel, S. (2009). Brain morphometry and functional imaging techniques in dementia: methods, findings and relevance in forensic neurology. *Curr Opin Neurol*.

Knops, A., Thirion, B., Hubbard, E., Michel, V., & Dehaene, S. (2009). Recruitment of an area involved in eye movements during mental arithmetic. *Science*, 1171599v1.

Koutsouleris, N., Meisenzahl, E. M., Davatzikos, C., Bottlender, R., Frodl, T., Scheuerecker, J., Schmitt, G., Zetzsche, T., Decker, P., Reiser, M., Möller, H.-J., & Gaser, C. (2009). Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Arch Gen Psychiatry*, *66*(7), 700–12.

Kriegeskorte, N. & Bandettini, P. (2006b). Intersubject-information-based brain mapping reveals cortical representations canonically driven by a movie presentation. *Society for Neuroscience Abstracts*, *309.10*.

Kriegeskorte, N. & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI.

Neuroimage, *38*(4), 649–662.

Kriegeskorte, N., Cusack, R., & Bandettini, P. (2009). How does an fMRI voxel sample the neuronal activity pattern: Compact-kernel or complex spatiotemporal filter? *Neuroimage*.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proc Natl Acad Sci U S A*, *104*(51), 20600–20605.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, *103*(10), 3863–3868.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008b). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008a). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–41.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, *12*(5), 535–40.

Kruskal, J. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

LaConte, S. M., Peltier, S. J., & Hu, X. P. (2007). Real-time fMRI using brain-state classification. *Hum Brain Mapp*, *28*(10), 1033–44.

Lee, J.-H., Ryu, J., Jolesz, F. A., Cho, Z.-H., & Yoo, S.-S. (2009). Brain-machine interface via real-time fMRI: preliminary study on thought-controlled robotic arm. *Neurosci Lett*, *450*(1), 1–6.

Li, S., Mayhew, S. D., & Kourtzi, Z. (2009). Learning shapes the representation of behavioral choice in the human brain. *Neuron*, *62*(3), 441–452.

Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, *20*(15), 2429–37.

Li, W., Howard, J. D., Parrish, T. B., & Gottfried, J. A. (2008). Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science*, *319*(5871), 1842–5.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, *62*(2), 281–90.

Lotte, F., Congedo, M., Lecuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *J Neural Eng*, *4*(2), R1–R13.

Luo, W.-L. & Nichols, T. E. (2003). Diagnosis and exploration of massively univariate neuroimaging models. *Neuroimage*, *19*(3), 1014–32.

Marquand, A. F., Howard, M., Brammer, M. J., Chu, C., Coen, S., & Mourao-Miranda, J. (2009). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using gaussian processes. *Neuroimage*.

Medin, D. L. (1989). Concepts and conceptual structure. *The American psychologist*, *44*(12), 1469–81.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J Acoust Soc Am*, *123*(2), 899–909.

Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *J Neurophysiol*, *100*(3), 1407–19.

Misra, C., Fan, Y., & Davatzikos, C. (2009). Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to ad: results from adni. *Neuroimage*, *44*(4), 1415–22.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images,. *Machine Learning*, *57*, 145–175.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195.

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *Neuroimage*, *28*(4), 980–95.

Muller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: from brain-computer interfacing to mental state monitoring. *J Neurosci Methods*, *167*(1), 82–90.

Mur, M., Bandettini, P., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI – an introductory guide. *Social Cognitive and Affective Neuroscience*.

Nicolelis, M. A. L. & Lebedev, M. A. (2009). Principles of neural ensemble physiology underlying the operation of brain-machine interfaces. *Nat Rev Neurosci*, *10*(7), 530–40.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*, *10*(9), 424–430.

Nosofsky, R. (1988). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 54–65.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat Neurosci*, *4*(12), 1244–52.

Op de Beeck, H. P. (2009). Against hyperacuity in brain reading: Spatial smoothing does not hurt multivariate fMRI analyses? *Neuroimage*.

Op de Beeck, H. P., Baker, C. I., DiCarlo, J. J., & Kanwisher, N. G. (2006). Discrimination training alters object representations in human extrastriate cortex. *J Neurosci*, *26*(50), 13025–36.

Op de Beeck, H. P., Torfs, K., & Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. *J Neurosci*, *28*(40), 10111–23.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cogn Neurosci*, *19*(11), 1735–52.

Parker, A. J. & Newsome, W. T. (1998). Sense and the single neuron: probing the physiology of perception. *Annu Rev Neurosci*, *21*, 227–77.

Peelen, M. V. & Downing, P. E. (2007). Using multi-voxel pattern analysis of fMRI data to interpret overlapping functional activations. *Trends Cogn Sci*, *11*(1), 4–5.

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1 Suppl), S199–209.

Poldrack, R. A., Halchenko, Y., & Hanson, S. J. (2009). Classifying and visualizing large-scale brain states from neuroimaging data. *Psychological Science*, *In press*.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, *310*, 1963–6.

Quiroga, R. Q., Reddy, L., Koch, C., & Fried, I. (2007). Decoding visual inputs from multiple neurons in the human temporal lobe. *J Neurophysiol*, *98*(4), 1997–2007.

Racine, J. & Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics*, *119*(1), 99–130.

Raizada, R. D. S., Tsao, F. M., Liu, H. M., Holloway, I. D., Ansari, D., & Kuhl, P. K. (2009b). Linking brain-wide multivoxel activation patterns to behaviour: examples from language and math. *NeuroImage*, *under review*.

Raizada, R. D. S., Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2009a). Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. *Cerebral Cortex*, *Advance online publication*, DOI: 10.1093/cercor/bhp076.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2009). Six problems for causal inference from fMRI. *Neuroimage*.

Raudys, S. & Jain, A. (1991). Small sample-size effects in statistical pattern-recognition - recommendations for practitioners. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, *13*(3), 252–264.

Rencher, A. C. (2002). *Methods of multivariate analysis* (2nd ed ed.). New York: J. Wiley.

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, *25*(1), 230–42.

Rota, G., Sitaram, R., Veit, R., Erb, M., Weiskopf, N., Dogil, G., & Birbaumer, N. (2009). Self-regulation of regional cortical activity using real-time fMRI: the right inferior frontal gyrus and linguistic processing. *Hum Brain Mapp*, *30*(5), 1605–14.

Saul, L. K., Weinberger, K. Q., Ham, J. H., Sha, F., & Lee, D. D. (2006). Spectral methods for dimensionality reduction. In O. Chapelle, B. Schoelkopf, & A. Zien (Eds.), *Semisupervised Learning*. Cambridge, MA: MIT Press.

Schurger, A., Pereira, F., Treisman, A., & Cohen, J. (2009). Reproducibility distinguishes conscious from nonconscious neural representations. *Science*, science.1180029v1.

Serences, J. T. & Boynton, G. M. (2007). Feature-based attentional modulations in the absence of direct visual stimulation. *Neuron*, *55*(2), 301–312.

Shepard, R. N. (1962). The analysis of proximities - multidimensional-scaling with an unknown distance function. 1. *Psychometrika*, *27*, 125–140.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–23.

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.

Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K., & Yacoub, E. (2009). Mechanisms underlying decoding at 7 t: Ocular dominance columns, broad structures, and macroscopic blood vessels in v1 convey information on the stimulated eye. *Neuroimage*.

Sitaram, R., Caria, A., & Birbaumer, N. (2009). Hemodynamic brain–computer interfaces for communication and rehabilitation. *Neural Networks*.

Soon, C. S., Brass, M., Heinze, H.-J., & Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nat Neurosci*, *11*(5), 543–545.

Sun, D., van Erp, T. G. M., Thompson, P. M., Bearden, C., Daley, M., Kushan, L., Hardt, M., Nuechterlein, K., Toga, A. W., & Cannon, T. D. (2009). Elucidating a magnetic resonance imaging-based neuroanatomic biomarker for psychosis: Classification analysis using probabilistic brain atlas and machine learning algorithms. *Biological Psychiatry*.

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327–352.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition (formerly titled ”voodoo correlations in social neuroscience”). *Perspectives on Psychological Science*, *4*(3), 274 – 290.

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *J Neurosci*, *29*(34), 10573–81.

Weber, M., Thompson-Schill, S. L., Osherson, D., Haxby, J., & Parsons, L. (2009). Predicting judged similarity of natural categories from their neural representations. *Neuropsychologia*, *47*(3), 859–68.

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nat Neurosci*, *10*(6), 685–6.

Worsley, K., Poline, J., Friston, K., & Evans, A. (1997). Characterizing the response of PET and fMRI data using multivariate linear models. *NeuroImage*, *6*(4), 305–319.