Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data

Alice J. O'Toole, Fang Jiang Hervé Abdi [a,*] Nils Pénard, Joseph P. Dunlop & Marc A. Parent

[a] *The University of Texas at Dallas Richardson, TX, 75083–0688, USA.*

Received August 08, 2006; accepted May 02, 2007

Abstract

The goal of pattern-based classification of functional neuroimaging data is to link individual brain activation patterns to the experimental conditions experienced during the scans. These "brain-reading" analyses advance functional neuroimaging on three fronts. From a technical standpoint, pattern-based classifiers overcome fatal flaws in the status quo inferential and exploratory multivariate approaches, by combining pattern-based analyses with a direct link to experimental variables. In theoretical terms, the results that emerge from pattern-based classifiers can offer insight into the nature of neural representations. This shifts the emphasis in functional neuroimaging studies away from localizing brain activity toward understanding how patterns of brain activity encode information. From a practical point of view, pattern-based classifiers are already well established and understood in many areas of cognitive science. These tools are familiar to many researchers and provide a quantitatively sound and qualitatively satisfying answer to most questions addressed in functional neuroimaging studies. Here, we examine the theoretical, statistical, and practical underpinnings of pattern-based classification approaches to functional neuroimaging analyses. Pattern-based classification analyses are well positioned to become the standard approach to analyzing functional neuroimaging data.

Introduction

Pattern-based classification analyses are appearing with increasing frequency in the functional neuroimaging literature and are being applied across a diverse span of topics. These techniques first garnered wide attention in studies of face and object perception, where researchers agree about brain activation sites for faces/objects, but are engaged in a vigorous debate about the nature of the underlying high-level visual representations (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Gauthier, Tarr, Anderson, Skudlarski & Gore, 1999; Hanson, Matsuka, & Haxby, 2004; Haxby, Gobbini, Furey, Ishai, Schouten & Pietrini, 2001; O'Toole, Jiang, Abdi, & Haxby, 2005; Spiridon & Kanwisher, 2002; Kamitani & Tong, 2005). In vision science, pattern-classification analyses have produced striking and important results on the neural processes underlying low-level visual aftereffects (Haynes & Rees, 2005a), predicting conscious visual perceptions (Haynes & Rees, 2005a, Haynes, & Rees, 2005b, Kamitani & Tong, 2005), dissociating brain areas responsive to biological motion (Peelen, Wigget & Downing, 2006), and distinguishing brain states underlying face matching and location matching tasks (Mourão-Miranda, Bokde, Born, Hampel, & Stretter, 2005). In other areas of cognitive

* Corresponding Author. Send correspondence to
Hervé Abdi,
The University of Texas at Dallas,
Program in Cognition and Neuroscience, MS: Gr:4.1,
Richardson, TX, 75083–0688.
Email address: herve@utdallas.edu (Hervé Abdi).
URL: www.utd.edu/∼herve (Hervé Abdi).

[1] Thanks are due to Jim Haxby and Randy McIntosh for insightful conversations about the theory and practice of data analysis for functional neuroimaging studies. We also thank Stephen Strothers for pointers to useful references from earlier uses of classifiers for brain image data and two anonymous reviewers for comments on a previous version of the manuscript. We would also like to gratefully acknowledge Jim Haxby's willingness to make data from his laboratory available to us and to others for independent and open-ended scrutiny and re-analysis—from which we, and others, have learned a great deal. Thanks are also due to Dana Roark for helpful comments on the manuscript.

science, pattern-based classifiers have been able to detect lies from brain activity patterns (Davatzikos, Ruparel, Fan, Acharyya, Loughead, Gur & Langleben, 2005), classify brain states underlying the experience of reading different words (Mitchell, Hutchinson, Niculescu, Pereira, Wang, Just, & Newman, 2004), and predict conscious decisions about emotional faces at accuracy levels comparable to neuronal data (Pessoa & Padmala, 2005).

Pattern-based classification methods have also pushed the spatial resolution of functional neuroimaging data beyond conventional limits, by offering converging "human" evidence for findings from single-unit neurophysiological studies in animals (e.g., Kamitani & Tong, 2005; Haynes & Rees, 2005a). These kinds of approaches can offer interpretative grounding for recently reported data from high-resolution functional neuroimaging studies (e.g., Grill-Spector, Sayres, & Rees, 2006; Schwartzlose, Baker, & Kanwisher, 2005).

The use of pattern-based classification analyses for functional neuroimaging has been reviewed recently with emphasis on the ability of these methods to "decode conscious experience" (Haynes & Rees, 2006) and to address questions about neural representation (Norman, Polyn, Detre, & Haxby, 2006). Haynes and Rees (2006) also offer a thoughtful critique of the ethical issues entailed in brain-decoding approaches. Both of these previous reviews highlight the accomplishments of this new approach across the domain areas to which it has been applied. Here we take a closer look at the theoretical, statistical, and practical underpinnings of pattern-based classification approaches to functional neuroimaging analyses. We argue that pattern-based classification analyses solve some long-standing technical and statistical problems for functional neuroimaging data and provide a more accurate accounting of the data at hand. We also look at the relationship among different classifiers and the issues relevant for implementing classifiers with appropriate pre-processing and cross-validation schemes.

This paper is organized as follows. We begin with a look at what qualifies as a pattern-classification approach to functional neuroimaging analysis. Next we detail the advances made by this approach on technical, theoretical, and practical grounds, focusing on the question of whether these techniques will move the field forward or simply add unnecessary complexity to the diverse methods already in place. Finally, we consider the challenges ahead for using pattern-based classifiers as a standard in functional neuroimaging analysis.

Pattern-based classification for functional neuroimaging analysis: What qualifies?

The goal of pattern-based classification of functional neuroimaging data is to link individual brain activation patterns to the experimental conditions experienced during the scans. These "brain-reading" (Cox & Savoy, 2003) analyses address a fundamentally different experimental question than traditional exploratory or inferential analyses. They ask, "How reliably can *patterns* of brain activation *indicate* or *predict* the task in which the brain is engaged or the stimulus which the experimental participant is processing?" Pattern-based classifiers fit the "brain-reading" label assigned to them by Cox and Savoy (2003), because they allow us to "peer into the brain" and determine the likelihood that a particular perceptual or cognitive state is being experienced.

From a practical point of view, classifiers use individual patterns of brain activity from trials in a functional neuroimaging experiment to predict the experimental conditions in effect when the scans were taken. Most commonly, a standard, "off-the-shelf" classifier algorithm is applied to the task of *learning* the statistical relationship between patterns of brain activity and the occurrence of particular experimental conditions. Individual brain activity patterns are input to the classifier and predictions of the stimulus or experimental condition are generated. The accuracy of these predictions can be measured in standard performance measures (e.g., percent correct, *d'*). The data that emerge from this analysis are, therefore, remarkably similar in form to the data that emerge from behavioral experiments in psychology and cognitive science.

The earliest uses of pattern-classifiers for neuroimaging analysis date back to the early nineties and were implemented to classify PET data (e.g., Azari, Pettigrew, Schapiro, Haxby, Grady, Pietrini, et al., 1993; Clark, Ammann, Martin, Ty & Hayden 1991; Kippenhan, Barker, Pascal, Nagel & Duara, 1992; Moeller & Strother, 1991). The problem considered in these papers was to classify or "diagnose" clinical populations (e.g., Alzheimer, Huntington disease, or AIDS patients) using patterns of brain activation from PET scans. This problem leads naturally to the use of a standard discriminant analysis, which is a linear classifier. The immediate issue faced by these researchers when using standard discriminant analysis for this purpose was the larger number of voxels by comparison to the number of observations [2]. This motivated preprocessing schemes based on predefined regions of interest, hierarchical multiple regression analysis (in order to reduce and optimize the number of predictors), and a principal component analysis (PCA) (or both). The sophisticated techniques used in these early papers are equivalent to the brain decoding methods that have recently attracted wide attention in the literature as a novel approach to functional neuroimaging analysis.

[2] The larger number of voxels than observations would require the inversion of a singular matrix (which is equivalent to a division by zero and is therefore impossible).

A daunting challenge to understanding commonalities among pattern-based approaches to functional neuroimaging analyses is that classifiers go by a wide variety of names including neural network classifiers (NN), connectionist networks, support vector machines (SVM), correlation-based classifiers, backpropagation (BP), and linear discriminant analysis (LDA), among others. The diversity of labels and lack of cross-citations have obscured obvious connections among these analyses, leaving many readers to concentrate more on the specifics of particular analyses than on the general pattern-based classification approach they implement.

Remarkably, the labeling issue has also obscured a rather transparent connection between pattern-based classification approaches and partial least squares (PLS) regression analysis—a statistically driven pattern-based classification algorithm that has been used in functional neuroimaging studies for over a decade now. The now classic paper introducing PLS to the functional neuroimaging community (McIntosh, Bookstein, Haxby, & Grady's, 1996) has been cited 217 times and used in roughly two-thirds of these papers. Few studies using pattern-based classification even cite the McIntosh et al. (1996) paper on PLS. Notwithstanding, pattern-based methods are introduced uniformly in recent papers as a "novel approach" to functional neuroimaging analysis. Concomitantly, few studies using PLS regression cite pattern-based classification studies. A recent review of PLS regression analysis of neuroimaging data (McIntosh & Lobaugh, 2004), for example, does not cite any of the recent pattern-based classification papers or point out the equivalent aspects of the approach. It is not surprising, therefore, that there is limited general recognition of the related nature of the approaches among readers of the neuroimaging literature. From the perspective of researchers with a solid grounding in statistics, however, the labeling of methods in the functional neuroimaging literature is becoming confusing and potentially misleading.

The tendency in science to reinvent the wheel is especially prevalent in literatures that are inherently interdisciplinary and where experimental and statistical methods are borrowed intermittently across domains. One purpose of this paper is to *inform* a general readership in the behavioral and brain sciences about the similar nature of the wide variety of pattern-based classification analyses being used for functional neuroimaging data—and to point out critical differences among the approaches, where they exist. Seen as a body of work, these papers represent a paradigm shift in the way functional neuroimaging data are being analyzed and interpreted. These analyses come from different disciplines and go by different names but they accomplish the same thing. What they accomplish is precisely what is needed—a pattern-based analysis of functional neuroimaging data in terms that relate directly and quantitatively to experimental design variables.

A second purpose of this paper is to understand why analyses that have been available for decades are only now beginning to take hold in the functional neuroimaging literature. We argue that one reason behind the recent popularity of pattern-based classifiers in this field is that the time is ripe for moving functional neuroimaging research beyond the era of cortical localization, to a new level where questions about neural representation dominate questions about neural locus. Pattern-based classification analyses have the potential to support this next step and to become *the* standard approach in functional neuroimaging analysis.

A third purpose is to evaluate how the shift will likely affect progress in the field. On technical, theoretical, and practical grounds, we argue that the use of pattern-based classifiers will make important strides toward setting a minimum quantitative standard to which functional neuroimaging analysis must adhere and toward making full use of the potential of neuroimaging technologies.

Before proceeding, it is worth noting that the largest concentration of papers using pattern-based classification analyses for functional neuroimaging have appeared in domain of face and object processing. This is likely due to the recent emphasis in that literature on understanding the *pattern structure* of brain responses for deciding among alternative theoretical hypotheses about the neural representation of faces/objects. In what follows, we make liberal use of the studies from this domain to illustrate our points.

Why pattern-based classification analyses are attracting attention

The widespread and accelerating popularity of pattern-based classification analyses in functional neuroimaging can be attributed to three factors, which provide an organizational structure for this paper. First, pattern-based classifiers overcome fatal flaws in the status quo inferential and exploratory multivariate approaches. Second, the results that emerge from pattern-based classifiers can provide insight into the nature of neural representations. Third, pattern-based classifiers are already well established and understood in many areas of cognitive science. These tools are familiar to many researchers and provide a qualitatively and quantitatively satisfying answer to most questions addressed in functional neuroimaging studies.

We present these three factors first as tenets and then in more detail. We concentrate on how these factors support the adoption of pattern-based classifiers as the standard approach to analyzing functional neuroimaging data.

Tenet 1: Fatal flaws in the status quo

Voxel-based inferential statistics (e.g., analysis of variance, ANOVA) and multivariate exploratory methods (e.g., principal/independent components analysis, PCA/ICA) constitute the status quo in functional neuroimaging data analysis. Voxel-based inferential analyses are flawed because they treat brains data from neuroimaging studies as independent voxels. Exploratory multivariate analyses are flawed because they fail to provide quantifiable links to experimental design variables. Pattern-based classifiers address these shortcomings by treating brain images as patterns *and* by providing a quantifiable link to experimental conditions. *This is a technical advance in the quality of analyses available for functional neuroimaging data.*

Tenet 2: Understanding neural representation

Status quo analyses are focused on identifying brain regions that are active during perceptual and cognitive tasks. Pattern-based classification approaches to functional neuroimaging are focused more on understanding *how*, rather than *where*, the brain encodes information. If successful, pattern-based classifier approaches can offer insight into the nature of neural representations. This is a watershed issue for high-level neural codes that can go beyond trivial extensions of single unit neural response profiles to the nature of the neural codes that underlie perceptual and cognitive brain states. *This is a theoretical advance in the kinds of questions that functional neuroimaging studies can address.* It is a by-product of the technical advance, but one that relies on appropriately framing the experimental questions, rather than simply implementing a classifier.

Tenet 3: Familiarity, comfort level, and "attractiveness" of the approach

To ignore the influence of the sociology of science in understanding why certain methodological approaches "catch on" and others (even rigorous, well-respected ones) are left to "specialists" is to miss a solid slice of causality in the progress of science. The practitioners of functional neuroimaging have come from traditions in psychology, cognitive science, and neuroscience, bringing analysis methods from these domains with them. From a technical point of view, these methods are ill suited to the analysis of functional neuroimaging data. Statisticians have long bemoaned the inadequacies of status quo approaches to functional neuroimaging data analysis and have succeeded in publishing, but not popularizing, sensible alternatives. Pattern-based classifiers took center stage in cognitive science under the name of neural networks in the eighties and early nineties. As techniques, they are familiar to broad and diverse groups of researchers in cognitive science and are statistically appropriate for the analysis of functional neuroimaging data. They are also functionally equivalent to statistically based techniques, such as PLS, which achieved the technical advance of Tenet 1 over a decade ago, but have not become the standard in functional neuroimaging analysis. The technical advance of Tenet 1 is "catching on" and will increase in popularity, under the name of pattern-based classifiers, "brain-reading" and "brain decoding" approaches. *This is a practical advance as research expertise from a broader range of behavioral and brain sciences can be brought to bear on issues in the analysis of functional neuroimaging data*.

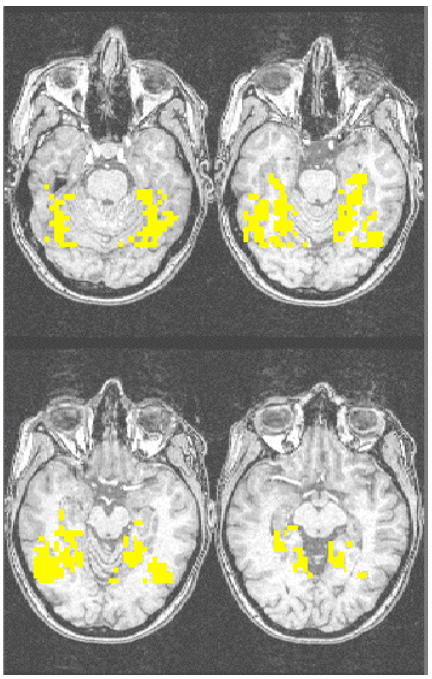
A sub-tenet of the "catch-on" issue is that the brain-reading metaphor suggested by pattern-based classification algorithms is an attractive metaphor that might be able to supplant less positive metaphors for functional neuroimaging, such as the "new phrenology" (Uttal, 2001). Whether a change in metaphor represents an advance (of any kind) in science depends on whether it alters the way researchers and the general scientific public view the value of the results that emerge. This may indeed be the most tangible advance to emerge from a paradigm shift.

Technical Advance: What's wrong with the status quo? Tenet 1

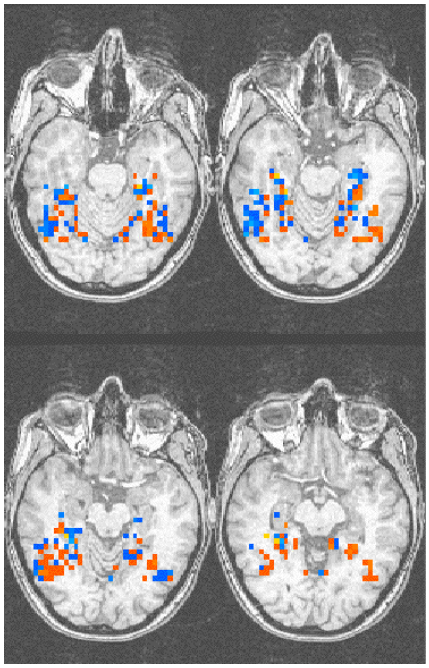
Pattern-based classification methods represent a major advance in the analysis of functional neuroimaging data, because they *combine* pattern-based data analysis of brain responses with quantifiable links to the experimental conditions. The two most common approaches to functional neuroimaging analysis are flawed because they provide *either* a quantitative link between the data and the experimental conditions *or* a pattern-based analysis, but not both. Specifically, inferential or "hypothesis-led" methods (Friston, 1998; Petersson, Nichols, Poline & Holms, 1999b) provide quantification of functional neuroimaging data in terms of the experimental variables, but do not operate on patterns of neural activation. Inferential analyses operate on single voxels, which are treated incorrectly in these analyses as independent measures.

Exploratory or "data-led" analyses (Friston, 1998; Petersson, Nichols, Poline, & Holms, 1999a) are applied directly to the neural activity patterns, but are limited in their ability to quantify patterns in terms of the original condition-based, experimental variables. These analyses quantify free-floating variance in neural activation patterns without regard to their source, experimental or otherwise. In what follows, we present the assumptions, advantages, and disadvantages of the status quo approaches.

a.)



b.)



c.)

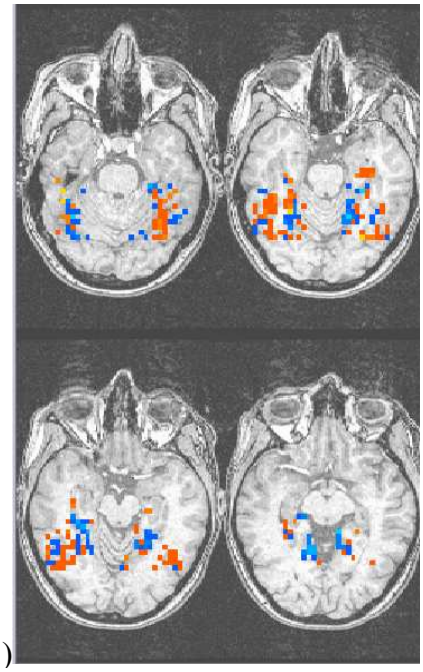


Fig. 1. *Figure 1.* Pattern based visualization. a.) Voxels from the Haxby et al. (2001) that differed significantly as a function of the object category being viewed by the participant when the scan was taken; b.) a principal component computed from the scans taken while participants viewed faces and houses that explains almost half of the variance in the set of scans from this participant, but is unrelated to the experimental variable. Intensity indicates the weighting of each voxel on this component (positive values in orange, negative values in blue); c.) a principal component that proved useful for classifying scans by condition (face versus house), but explained only 3 percent of the variance in the PCA.

Hypothesis-led approaches: A closer look

Hypothesis-led approaches to the analysis of functional neuroimaging data draw on inferential statistical methods similar to those used in the behavioral sciences. A standard hypothesis-testing perspective underlies this approach. A researcher posits that different brain states will result as a function of different experimental conditions. An inferential statistical test, such as an analysis of variance (ANOVA), is applied to answer the question, "Do the brain states that occur in response to experimental Condition A differ significantly from the brain states that occur in response to experimental Condition B?" The operational definition of this question reduces to assessing differences in the activation level of any voxel or subset of voxels or "region of interest" as evidence for rejecting the null hypothesis of "no difference" in brain states. Hypothesis-led methods follow a "label and analyze" approach. Voxel activations are *labeled* by condition and then *analyzed* in a way that detects statistically reliable differences between the activation levels as a function of experimental condition.

Figure 1a shows the results of an inferential analysis applied to data from Haxby et al. (2001) on face and object processing. Participants in that study viewed eight categories of objects (faces, houses, cats, chairs, scissors, bottles, shoes, and scrambled objects). An inferential analysis of these data asks the question, "Does brain activation vary as a function of the object category being viewed?" Inferential methods answer this question, one voxel at a time. The figure displays the result of applying one-factor (object category) ANOVAs separately to each voxel in a pre-selected target area of the brain, with voxel activation as the dependent variable in each analysis. Specifically, we see the anatomy of a participant, with an overlay of the voxels that were statistically significant in their respective ANOVAs. Figure 1a, therefore, shows the parts of the brain that are affected by changes in the experimental condition, (i.e., the object being viewed).

The advantage of inferential approaches is that they provide a statistical method for making statements about the reliability of brain response as a function of experimental conditions. The disadvantages of this approach are well known, but rarely taken seriously enough to limit the use of these techniques. First, these inferential analyses are voxel-based and therefore rely on the assumption that voxels are independent, which they are not—a fact that is universally accepted by functional neuroimaging researchers.

The consequence of accepting the independence assumption for co-dependent voxels is that it eliminates any possibility of measuring or assessing neural codes that reside in the *interactions among voxels*. The communication among neurons is, arguably, the preeminent strength of neural/brain computation. The power of functional neuroimaging techniques is that they simultaneously measure *patterns* of brain activity across large regions of the brain. The use of voxel-based inferential statistics, therefore, systematically eliminates *most* of the data a researcher gathers in a functional neuroimaging study, reducing the power of neuroimaging techniques to the level of single unit recordings.

The independence assumption leads to the second more technical, but equally vexing problem of multiple comparisons across statistical tests (e.g., Petersson et al., 1999b). The application of multiple inferential tests to the many voxels measured in functional neuroimaging experiments inflates the alpha level for rejecting the null hypothesis. The more comparisons made, the more likely it is to reject the null when it is correct (a Type I error). This leads to a liberal test for significance that overestimates effects. There are a number of ways of correcting for inflated values, but these swing the pendulum in the opposite direction toward Type II errors. This leads to excessively conservative tests that tend to underestimate effects (e.g., Petersson et al., 1999b).

Data-led approaches: A closer look

Data-led approaches to the analysis of functional neuroimaging data employ pattern-based, multivariate exploratory methods to "characterize" the nature of the signal present in the data, including "unsuspected effects" (Petersson et al., 1999a). Typical exploratory analyses used with functional neuroimaging data include principal component analysis (PCA) and independent component analysis (ICA; e.g., McKeown, Makeig, Brown, Jung, Kindermann, Bell, & Sejnowski, 1998). These kinds of multivariate analyses have a long and distinguished history and been applied ubiquitously to problems in engineering and cognitive science for decades. They are appropriate for functional neuroimaging analyses because of the multiple measures of brain activity that are taken across time and brain space in a typical study.

At the outset, it is worth stressing that brain activity data from a standard functional neuroimaging experiment do indeed vary across *both* space and time. As such, brain activity patterns can be defined relative to the temporal, spatial, or spatiotemporal dimensions of the brain data. We consider the importance of this spatiotemporal issue in more detail shortly. For present purposes, however, data-led methods are applied most commonly to the entire series of brain maps (spatially arrayed voxels) available from an experiment, (e.g., a matrix of n brain maps in the columns with m voxels for each in the rows). The signal characterization that results usually takes the form of n brain activity patterns [3] (e.g., principal or independent components) that explain different amounts of variance in the set of brain scans.

[3] Assuming $n < m$.

These components can be ordered according to the proportion of variance they explain, but must be interpreted "by eye" in a post-hoc fashion.

The process of interpretation is usually undertaken visually by projecting individual principal or independent components onto the participant's brain anatomy and looking at the projections. The patterns of activity captured in the exploratory analysis are considered meaningful positive indications of the experimental effects when they explain substantial proportions of variance and "resemble" or "echo" the hypothesized regions of brain activation and/or the temporal structure expected from the experimental conditions. Additional components may be deemed meaningful when they indicate experimental artifacts (e.g., from head movement). These pattern-based artifacts can be removed from the data when they are interpretable, allowing for a cleaner interpretation of the experimental data. Un-interpretable components are ignored.

A simple example appears in Figure 1b, using the data from the Haxby et al. (2001) experiment and the participant presented in Figure 1a. A PCA was applied to the brain scans taken while the participant viewed faces and houses. Because PCA was applied to brain scans, it is possible to interpret and visualize the resultant principal components as brain scans by projecting them back onto the anatomy of the participant. The first principal component explains roughly half of the variance in the scans and is displayed in Figure 1b. It highlights areas near the fusiform and parahippocampal place areas—these are areas known from previous studies to respond well to faces and houses, respectively. As we will see shortly, despite the general resemblance of the principal component to preconceived ideas about the locations of the relevant brain areas, a pattern-based classification approach shows that this principal component is unrelated to the experimental manipulation.

Exploratory analyses such as PCA and ICA are "blind" to the structure of the experimental design, and so the goal is to uncover the experimental design in the context of other important components of the data that may or may not be expected. In contrast to the inferential approach, exploratory methods can be considered examples of an "analyze and label" approach. The pattern data are analyzed first and are then labeled post-hoc with an interpretation that fits them into aspects of the experimental design (e.g., which category of objects a participant was viewing).

The fact that exploratory analyses are blind to the structure and parameters of the experiment is both an advantage and a weakness. The advantage is that components explaining variance highlight relationships among voxels at both a local and global scale. These relationships can be indicative of the co-activation of multiple brain areas to the stimulus being presented and also can indicate unsuspected or non-hypothesized effects, which might escape notice in a voxel-based analysis.

The primary disadvantage of exploratory approaches is that the interpretation of the derived components, in terms of the experimental conditions and artifacts, is left completely to the experimenter. Components that are compelling in terms of the experimental context are retained. Components that are not easy to interpret are discarded or ignored.

A second more practical disadvantage of multivariate analyses such as PCA and ICA is the voluminous and sometimes unwieldy nature of the resultant output. In fact, as much data comes out of the analysis as goes in. Although the proportion of explained variance serves as a guide to the "importance" of the components, this "importance" is defined by explained variance in the set of brain images analyzed. There is no systematic or prescribed method for determining whether explained variance in the set of images has anything to do with the manipulation of the experimental variables. As is well known, the functional effects detected in *f*MRI alter voxel activation level by less than a few percentage points. Consequently, the experimental effects of interest may explain a relatively small amount of variance in a linear analysis by comparison to a variety of other sources of variance in the scans (e.g., head movements). Thus, the proportion of explained variance may not be the best guide to finding components that relate to the experimental conditions.

Finally, the need to interpret components by eye, from an inherently three-dimensional functional neuroimaging brain map, is a daunting and potentially error-prone task. Given the volume of data and the number of components to choose from, it is generally possible to find something that looks "interesting" and meets a researcher's expectations about the location(s) of brain activity changes that should accompany particular experimental manipulations. Though patterns may "look" interesting, and may meet the expectation of a hypothesis, these methods include no formal way to establish a link between the patterns and experimental variables.

Synopsis of the status quo and technical advantages of pattern-based classifiers

The problem with status quo analyses to functional neuroimaging data is that there are fatal flaws in what they accomplish. The inordinate loss of information in voxel-based analyses is an unacceptable waste of functional neuroimaging data, which are expensive and challenging to collect. The need to interpret multivariate exploratory analyses without built-in data-driven constraints makes for an unacceptably weak tie between brain activation maps and experimental variables. Both voxel-based inferential and multivariate exploratory approaches are ill suited for functional neuroimaging analysis.

Two critical differences in pattern-based classification approaches over the status quo address the technical shortcomings of previous approaches. First, pattern-based classifiers ask a different question than either inferential or exploratory methods. Inferential methods ask, "Does the activation level of a voxel vary significantly as a function of experimental condition?" Exploratory methods ask, "What patterns of activity explain variations across the set of brain maps?" Pattern-based classification methods ask, "How reliably can *patterns* of brain activation *indicate* or *predict* the task in which the brain is engaged or the stimulus which the experimental participant is processing?" The answer provided by pattern-based classification methods is a quantitative measure of the discriminability of brain maps in terms of experimental variables. These data can be expressed as a percent correct classification or, when appropriate, in signal detection terms as a discrimination index, such as d'. The extracted measures of pattern similarity can be submitted to inferential statistics and compared to behavioral data from human participants.

Second, pattern-based classification algorithms operate on brain activity patterns, allowing for the relationship or interaction among voxels to contribute to the classification success of the algorithm. The ability of pattern-based classifiers to use this enormous quantity of data, which voxel-based analyses discard, makes them less likely than traditional voxel-based analyses to underestimate the discriminability of brain states as a function of experimental variables. We qualify this claim shortly to make clear that it applies only to certain types of pattern-based classifiers, which are a subset of those used in recent analyses of functional neuroimaging data.

Spatial, Temporal, or Spatiotemporal? Again, we stress that the brain activity data measured in a standard functional neuroimaging experiment vary both in time and space. Consequently, there can be significant pattern structure in time between voxels and in space between time points. Temporal variation typically marks modulations in the stimulus/experimental conditions in a dataset, whereas spatial variations mark the structure of voxel activation across different brain locations (assuming a constant stimulus/experimental condition). Most pattern-based approaches have been limited to the analysis of either the spatial or temporal structure of the design (though see Lobaugh, West, and McIntosh, 2001, for spatiotemporal analyses that encompass the entire spatiotemporal design of a study). In most cases, looking at the spatial or temporal structure alone is a reasonable first order analysis. For example, there can be valuable information gained by looking at the stability of a spatial activation structure in a stimulus/experimental condition and contrasting it to other conditions. Concomitantly, there can be valuable information in the time course activation of voxels as experimental conditions vary.

Ideally, pattern-based analyses should consider the full range of spatiotemporal patterns. However, to limit the complexity of this paper and to follow the thread inherent in most of the literature on this topic, we discuss pattern-based classification analyses in this paper only for spatial brain patterns. To date, most pattern-based classifiers in functional neuroimaging concentrate on the spatial layout of brain activity by condition. It is worth bearing in mind, however, that *both* space and time are variables in all functional neuroimaging studies—though they may have differential importance in the context of different experimental designs. In principle, all of the analyses discussed here apply to temporal and spatial variations in brain activity data, with the caveat that interpreting the analyses will differ substantially as a function of design type (e.g., block versus event).

Neural representation from functional neuroimaging data: Tenet 2

Much has been made recently of the fact that pattern-based analyses of functional neuroimaging data provide a kind of neural "read-out" of conscious experience (e.g., Haynes & Rees, 2006). This is a compelling metaphor for a pattern-based classifier approach to the analysis of functional neuroimaging data, but one that threatens to direct the attention of researchers away from the potential of these analyses for making progress on questions of neural code beyond the limits of established methods. Single unit neurophysiology, neuropsychology, and status quo treatments of functional neuroimaging data can confirm the importance of brain locations for a perceptual or cognitive process, but are limited in their potential for elucidating the contribution of the location to the neural code. In rare instances, neurophysiological single unit recordings can show a link between the response of a neuron and conscious perception (e.g., Newsome, Britten, & Movshon, 1989). The motivation of these studies is analogous to brain-decoding, but on the scale of individual neurons. Although these cases establish strong ties between conscious experience and a neural response, they constitute instances of grandmother cell codes (Barlow, 1972), and may represent only a small fraction of neural codes.

A verifiable link between a real-valued neural activity pattern and an experimental variable (or conscious perception) (e.g., this brain activity pattern predicts that a person is experiencing this ambiguous stimulus as *x*) can provide information about neural codes that is not available with previous approaches. This information comes in the form of interactions among voxels, the degree to which neural codes are shared at the level of stimuli and/or psychological tasks, and the ability to visualize pattern-based data that relate directly to experimental variables. The challenge is

to use the information to advance understanding of how the brain works and to expand the range of questions that can be addressed in functional neuroimaging.

The literature on face and object processing offers an entry into understanding how pattern-based classifiers can provide insights into neural representation that are not possible with other methods. It also serves as a useful base for a discussion of the relative merits of individual classifiers for analyzing functional neuroimaging data. We take up these issues in the next three sections. In this section, we offer a brief sketch of the neural representation issue at question in the domain of face and object processing. In so far as possible, we separate issues about neural representation from issues about the quality and adequacy of classifiers. We evaluate the various classifiers used in this literature in the section on "Familiarity, comfort level, and practical aspects of the approach." For clarity, however, we provide a brief description of the classifier(s) for each study as they are discussed, leaving the methodological issues and evaluation of different classifiers to the third section.

Face and Object Processing: Modular versus distributed?

The recent popularity of pattern-based classification approaches for analyzing functional neuroimaging data can be traced directly to Haxby et al.'s (2001) influential [4] study of face and object processing in ventral temporal (VT) cortex. The Haxby et al. (2001) paper weighed into an existing debate about whether the neural representation of faces and objects is *modular* or *distributed*. The modular account of representation posits that specific areas of VT cortex are specialized for representing certain categories of objects (Spiridon & Kanwisher, 2002). Prior to the use of pattern-based classification for this problem, support for the modular approach was based on findings that certain regions of cortex respond preferentially to particular categories of objects (e.g., faces and houses). By viewing statistically significant voxels projected onto brain anatomy, it is possible to locate cortical "hot spots" or maximal activation sites for faces, houses, and other objects. Faces, for example, elicit maximal response from an area in VT cortex now known as the fusiform face area (FFA, Kanwisher, McDermott, & Chun, 1997). Houses and scenes, for example, maximally activate an area in VT known as the parahippocampal place area (PPA, Epstein & Kanwisher, 1998).

The distributed or *object form topography model* posits that the representations of objects are distributed widely across VT cortex (Haxby et al., 2001). Prior to the use of pattern-based classifiers, support for the distributed account was based on findings that the brain regions preferring certain categories of objects also respond to other "non-preferred" categories of objects (e.g., Ishai, Ungerleider, Martin, Schouten, & Haxby, 1999).

Remarkably, there is little disagreement in the literature about the existence of brain areas that respond maximally to certain objects and little disagreement about where these areas are located. At issue is whether different categories of objects are coded with dedicated (modular) or shared (distributed) neural resources.

Pattern-based classifier approaches to understanding neural representation

Dissect and classify

Haxby et al.'s (2001) approach to the neural representation issue was simple and direct. They used a simple classifier to categorize scans from an *f*MRI experiment in which participants viewed eight categories of objects (faces, houses, cats, bottles, scissors, shoes, chairs, and scrambled control stimuli). This classifier was based on comparisons of individual brain map correlations, both within and across object categories (Haxby et al., 2001). The idea behind this approach is that accurate pattern-based discrimination of the scans is possible if there is a higher correlation among scans taken while participants viewed stimuli from "within a category of objects" than while they viewed stimuli "across different categories of objects." Haxby et al. (2001) found that the patterns of brain responses to object categories were highly discriminable.

To look at the distributed versus modular representation of faces and objects, the classifier was applied to different subsets of voxels. The modular hypothesis predicts that the information for classifying objects by category should be contained primarily in voxels that respond maximally to the category in question. Haxby et al. (2001) found that the voxels maximally activated in response to particular categories could be deleted from the classifier input with only minor cost to classification accuracy. This finding supports a distributed account of neural representation because it suggests that the regions of brain that respond maximally to particular categories of objects are not required to categorize objects.

Following the study of Haxby et al. (2001), Spiridon and Kanwisher (2002) also used the same type of classifier with a related, but not identical, operational definition of "distributed" versus "modular". Haxby et al. asked, "To what extent do voxels that respond non-maximally to a particular category of objects affect performance in discriminating preferred and non-preferred objects?" Spiridon and Kanwisher (2002) asked, "How useful are voxels from specialized areas for discriminating among objects from non-preferred categories?" Spiridon and Kanwisher (2002) found that regions that give their maximal response to a particular object (e.g., faces) were only minimally able to

[4] Cited 221 times as of this writing!

classify other objects (e.g., chairs and shoes). They concluded in favor of a modular organization of ventral temporal cortex.

These two studies illustrate a *classify-and-dissect* approach to understanding neural representation. The first step is to show that the classifier can discriminate brain activation patterns with reasonable accuracy. This provides a quantitative link between patterns and experimental conditions. The accuracy of prediction gives a measure of the strength of the link. The second step is to selectively ablate the voxels input to the classifier in a theoretically motivated way. The effect of this ablation on classification accuracy is taken as an indication of how patterns of brain activity contribute functionally to the neural code.

Differences in the conclusions of Haxby et al. (2001) and Spiridon and Kanwisher (2002) are likely due to differences in their operational definitions of distributed and modular. Although classifiers are capable of providing the kind of data needed to test hypotheses about neural representation, as for other analysis methods in science, the definitiveness of the test lies not with the method itself, but with the framing of the question. A positive outcome of any new method in science is that it opens a dialog on how to frame questions in the context of a novel tool. To provide more definitive answers to questions about neural representations that are pattern-based, more precise formulations of these theoretical concepts are required. In the case of the face and object literature, these formulations came quickly in the form of a flurry of follow-up studies to Haxby et al (2001) and Spiridon and Kanwisher (2002). These studies introduced a host of intriguing refinements to the general notion of how (or whether) pattern-based neural codes distribute information.

Hanson et al. (2004) moved the representation question to a higher level of precision by proposing an explicit taxonomy of neural code types. Working with the data from Haxby et al. (2001), Hanson et al. defined four neural instantiations of distributed and modular codes. The first comprises a spatially local/compact code that indicates the presence or absence of an object type. The second adds a potential likelihood estimate to this spatially local/compact representation. The third code is distributed (i.e., spread out), but non-overlapping. Voxels in this code are dedicated to object categories, although voxels dedicated to coding specific object categories are not (necessarily) spatially contiguous. The fourth type of code is distributed and partially or completely overlapping. Hanson et al. (2004) refer to the completely overlapped version of this code as "combinatorial" and stress that this code type is "intensity variable." Object categories are indicated, therefore, not only by the patterns of active voxels, but by the relative intensities of the voxels as well. At its extreme, all voxels might be active for all object categories, with distinctions among categories coded *only* by the relationship among voxel intensities across cortex.

Hanson et al. (2004) implemented a variety of feedforward, neural network architectures with both linear and nonlinear decision rules. The architecture for which they report the most interesting results starts with input from the voxels, follows with a bottleneck of hidden units, and ends with projections onto output units indicating the experimental conditions. The sensitivity of individual voxels was assessed by adding Gaussian noise to the voxels and recomputing classification accuracy. This sensitivity analysis showed a high degree of overlap among voxels recruited for all object categories and exemplars. Hanson et al. (2004) conclude, therefore, in favor of a combinatorial code.

Knock-out and classify

Carlson et al. (2003) took the representation enterprise a step further using an elegant "knock-out" procedure that is capable of deleting entire systems of voxel activation patterns involved in certain kinds of classifications. They modeled data from a face and object processing study in which participants viewed three categories of objects (faces, houses, and chairs) in either a passive viewing procedure or a delayed matching task (Ishai et al., 1999).

The knock-out procedure was implemented in two steps. First, Carlson et al. (2003) created a multidimensional representation space of the scans using PCA. Next, a LDA classifier learned to predict experimental conditions from the projections of the individual scans on the principal components. The important difference in this approach is that voxel values are not directly input to the classifier. Rather, classifier input is based on an extracted exploratory multivariate analysis of the scans (see Section "Data-led approaches"). In this context, the interpretation difficulties we discussed previously for multivariate exploratory analyses are reduced, because the classifier quantitatively links the principal components to the experimental variables. A researcher does not have to guess whether or not a pattern of activation (e.g., principal component) is linked to the experimental manipulation. Moreover, the good points of exploratory multivariate analyses, like their ability to remove experimental artifacts, return. Factors that explain variance, (e.g., experimental artifact and experimental condition manipulations), will be detected in the multivariate exploratory analysis. The classifier then weights the information useful for predicting experimental condition strongly and pushes the weights on the artifactual information toward zero. Because classifier input is in the form of projections of scans on the multivariate axes, the weighting applies, not to individual voxels, but to entire (overlapping) patterns of voxel activations specified by the multivariate axes.

The knock-out procedure operates by systematically eliminating or "lesioning" the best discriminant *axes* until

the classification performance is close to chance [5] . Each "knock-out" deletes an entire complex pattern of brain responses, rather than deleting individual voxels. This allowed Carlson et al. (2003) to look at the interdependence among neural codes for the various kinds of classifier tasks. With this procedure, it becomes possible to delete all of the information needed for a particular discrimination task and then to assess the effects of this deletion on any other discrimination task. Carlson et al. (2003) defined *category-specific classifiers,* which discriminated a specific category of items (e.g., faces) from the other categories of items (e.g., houses and chairs). *Pairwise classifiers* discriminated a specific category of items (e.g., faces) from another category of items (e.g., houses). *Object-control classifiers* discriminated a specific category of items (e.g., faces) from scrambled items from the category (e.g., scrambled faces). Carlson et al. (2003) found that category-specific knock-outs reduced classification performance for all three object-specific control tasks. This illustrates that there are aspects of the neural representations that are shared across all object categories.

The knock-out technique adds a powerful tool to the repertoire of methods available for investigating the nature of the patterns involved in object representation. Of note, it sets up a framework for investigating the degree to which neural representations sub-serve multiple tasks, thereby providing a window into the structure of information processing in the brain. Questions of task independence, which are key in many areas of psychology, can be addressed with this method.

Representations: Merging Brain Space with Stimulus Space

An important advantage of pattern-based classification analyses of functional neuroimaging data is that many of these analyses create a "brain space". When compared appropriately to perceptual spaces or computationally derived stimulus spaces, brain spaces can ground hypotheses about neural representation (cf., Edelman, Grill-Spector, Kushnir & Malach, 1998; Young & Yamane, 1992). By "brain space," we mean simply that in geometric terms, pattern-based classifiers classify patterns based on *where* they are in a multidimensional space that represents brain states. The success of the classifier depends upon how neatly the scans/brain states from particular experimental conditions cluster. Failures of classification indicate that scans from different conditions intermix in the brain space. The confusability of brain scans can be used to leverage information about representations, when the hypothesized representations or known perceptual data suggest that certain stimulus conditions *should be* more or less confusable.

To illustrate a brain space using the data of Haxby et al. (2001), O'Toole et al. (2005) used the "distances" between the brain states that resulted when viewing different object categories. These distances were derived with a PCA followed by a pattern-based classifier, similar to that used by Carlson et al. (2003). The *d'*s (distances) for discriminating object categories were based on the functional neuroimaging scans collected by Haxby et al. (2001). These distances can be combined across the individual participants to give an idea of the consistency of brain representations across different observers, avoiding the notoriously error-prone process of physically aligning participant brains.

Figure 2a shows a profile of the similarity of *brain responses* to the object categories, combined across participants. It is clear that the pattern-based representations of the face and house categories are highly distinct from each other and from the cluster of other objects [6] . As indicated by the participant dispersion lines, it is also clear that the neural response to these categories was relatively consistent across participants in the study. Viewed in computational object recognition terms, questions about the nature of face and object representations in VT cortex can be considered in the context of stimulus predictions. How confusable *should we expect* the neural representations of object categories to be, assuming a distributed or modular hypothesis (O'Toole et al., 2005)? The object form topography model of Haxby et al. (2001) posits that the representations of faces and other objects are widely distributed and overlapping, because VT cortex contains a topographically organized representation of the attributes that underlie face and object recognition. It follows that voxels should share information about object categories as a function of the degree to which the object categories share features or attributes. "Similar" object categories should share more voxels than "dissimilar" object categories. An unexpected consequence of this logic is that a distributed coding of objects in VT cortex actually predicts modular brain activations, *when the object categories share few common attributes*.

To define "physically similar", O'Toole et al. (2005) implemented a computational model to categorize the stimuli used by Haxby et al. (2001). The resultant stimulus space was remarkably similar to the brain space (Figure 2b). This is consistent with a distributed code in which the brain response patterns for different object categories share voxels in roughly equal measure to the extent to which they share attributes. It also accounts for brain activation patterns that look modular, but which are actually based on distributed coding principles.

[5] It is worth noting that Hanson et al. (2004) constrain the number of axes available for coding by varying the number of hidden units, a procedure related to that used by Carlson et al. (2003). The Hanson et al. results have interesting implications for finding the minimum dimensionality of representations.

[6] See O'Toole et al. (2005) for a discussion of the cat and scrambled categories, which are not well-fit by the computational model.

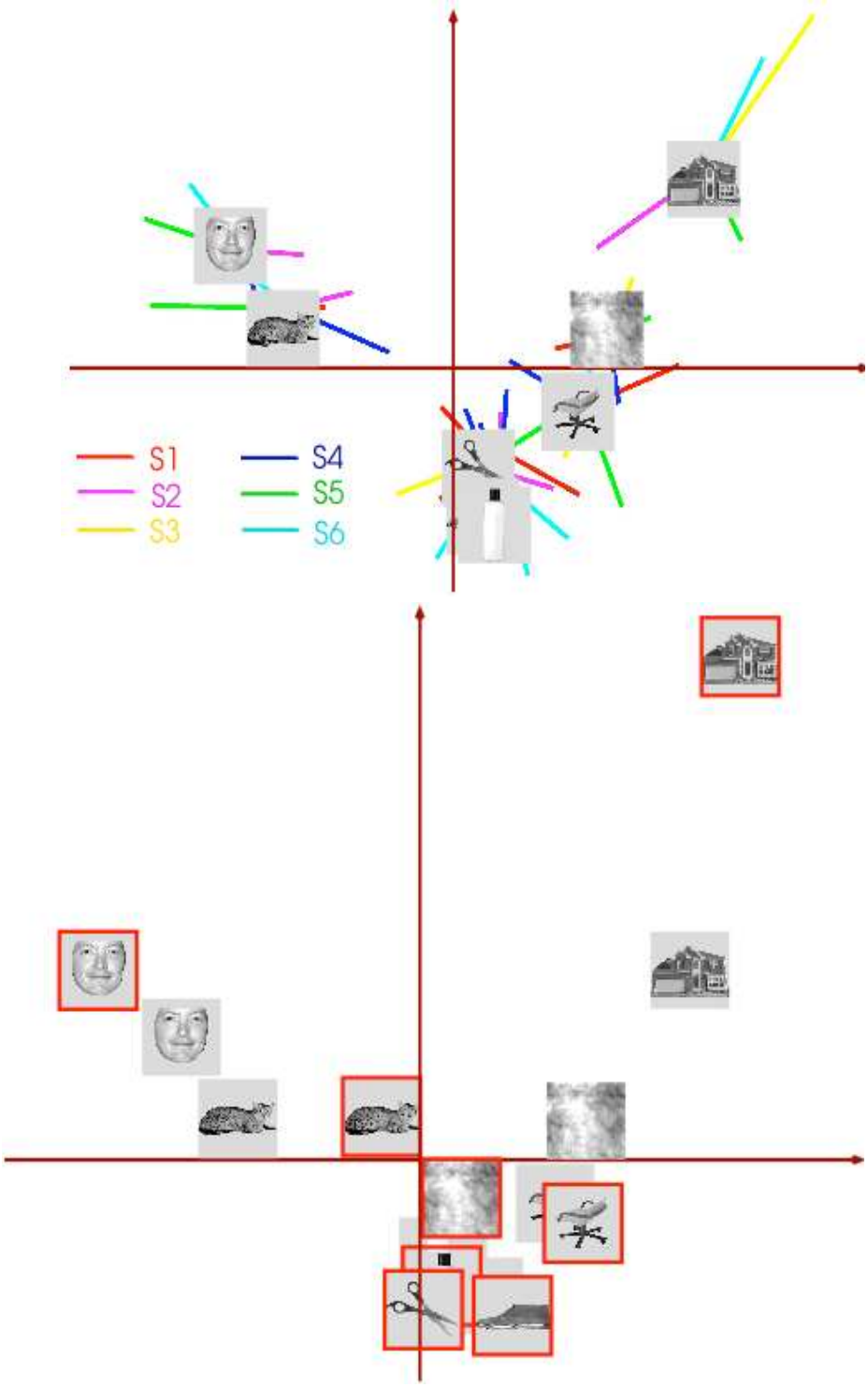


Fig. 2. A schematic view of the neural separability of representations for faces and objects derived from O'Toole et al. (2005) based on data from Haxby et al. (2001). a.) shows a DISTATIS plot derived from the neuroimaging-based confusability profiles of the scans from the eight categories (Abdi, Valentin, O'Toole & Edelman, 2005; Abdi, 2003, 2007; Kherif, Poline, Mériaux, Benali, Flandin, & Brett 2003, Shinkareva, Ombao, Sutton, Mohanty, & Miller, 2006). Dispersion lines indicate the consistently of individual participants with respect to the center points of the categories; b.) a combined neuroimaging and stimulus space that shows the compatibility of the neural activation maps from Figure 2a and the stimulus model (outlined pictures). The distance matrix from the stimulus model projected onto the neural activation distance map (Abdi & Valentin, 2006) reveals a close fit between brain and stimulus spaces.

The use of patterns as the primary unit of analyses offers advantages in visualizing pattern-based brain activity with a stronger tie to interpretation than is possible with multivariate exploratory analyses. We can now return to the information in patterns for discriminating experimental conditions. From the PCA-LDA classifier used by O'Toole et al. (2005), the principal component displayed in Figure 1c was quantitatively linked to the experimental variable—it achieved a high degree of separation (*d'* = 3.3) between face and house brain maps in the brain of a participant in Haxby et al. (2001). This component, however, explained only 3 percent of the variance in the PCA. This stands in stark contrast to the principal component displayed in Figure 1b, which explained over 50 percent of variance but was unrelated to the experimental manipulation.

In summary, a verifiable link between a real-valued neural activity pattern and an experimental variable (or conscious perception) can provide information about neural codes that is not available with previous approaches. The information provided by the interaction among voxels, the degree to which neural codes are shared at the level of stimuli and/or psychological tasks (e.g., detection of an object, discriminating pairs of objects, open-ended object classification), and visualization of supportive pattern-based data expand the range of questions that can be addressed directly in functional neuroimaging studies.

Familiarity, comfort level, and practical aspects of the approach: Tenet 3

In the introduction, we noted that pattern-based classifiers are familiar to many researchers via the neural network boom of the eighties and early nineties. Shared knowledge of methods in a field fosters progress through the common ground it provides for dialog. The quick spread of pattern-based classifiers across a diverse range of domains in functional neuroimaging is probably due in part to the comfort level aspect of these analyses. If it were simply a question of making a technical advance, PLS regression would now be the standard. Pattern-based classification analyses satisfy long-standing valid complaints about the inadequacy and inappropriateness of status quo approaches and are reasonably well understood in the behavioral sciences. They can therefore support a relatively painless transition from the status quo to a better way of analyzing functional neuroimaging data. It is important, nonetheless, to see these approaches in a correct historical context, where functionally related analyses are part of the overall change in perspective for functional neuroimaging analysis.

In this section, we look at some of the practical issues involved in choosing an appropriate classifier and in implementing it in ways that enable conclusions that answer the questions at hand. We begin with the topic of choosing input for a classifier from the enormous number of available brain voxels. These pre-processing decisions are critical for determining the scope of conclusion possible and are a necessary prerequisite for training well-behaved classifiers. Next, we consider cross-validation methods, which must be implemented to evaluate the robustness of classifier results across samples. Finally, we consider some critical features of the classifiers and the representations on which they operate.

Preprocessing or Input Conditioning

Because the number of voxel activations measured in a standard fMRI study can be on the order of several hundred-thousand voxels, the input used by classifiers almost never includes the entire set of voxels. Rather, it is common for researchers to pare down the activity pattern to a smaller subset of the available voxels (e.g., Haxby et al., 2001), a low-dimensional or compressed representation of the voxels (achieved usually by PCA) (e.g., Strother, Anderson, Schaper, Sidtis, Rottenberg, 1995), or both (O'Toole, et al., 2005). There are important technical and statistical motivations for the selection or compression of the voxel input to a classifier. One, in particular, is that the number of voxels is very large by comparison to number of scans. This creates a problem of over-fitting, whereby perfect classification can be obtained for the learning set, because there are too many free parameters (i.e. voxels) relative to the number of observations (i.e., scans). Because of this problem, the solution obtained with a large number of parameters can fit a sample dataset, regardless of its relevance or generality across other sample datasets extracted from the same population.

A second, related but more vexing, problem occurs when the number of predictors is larger than the number of observations. This "multicolinearity" problem makes it impossible to invert the between-voxel covariance matrix, due to the fact that it is not full rank. In practice, techniques such as discriminant analysis or multiple regression analysis fail in these cases. Both multicolinearity and over-fitting problems can be handled by input conditioning (i.e., preprocessing), cross-validation (see section below), or by a combination of both.

Preprocessing or voxel selection is a common first step in many pattern-based classification models of functional neuroimaging data. Methods for selecting voxels for input to a classifier include the use of inferential statistics as a screening device for deciding which voxels merit further pattern-based examination. For example, Haxby et al., (2001) applied an inferentially-based regressor to the raw *f*MRI voxel activations to locate voxels that differed significantly across the eight object categories he used. Cox and Savoy (2003) used inferential statistics to target two subsets of voxels: a.) all voxels that varied significantly over at least one cate-

gory of object stimuli; and b.) functionally localized voxels that discriminated significantly between object categories and their scrambled controls. In both cases, this reduced input dimensionality from several hundred thousand to a few hundred, with the caveat that the classifier performance is conditional on the set of voxels that vary significantly with respect to some experimental condition variable(s).

Reduction of the dimensionality of the classifier input using PCA-based compression is also a common preprocessing method. A short digression is useful here to recall that PCA produces a set of principal components derived from the statistical structure of the *set* of brain scans. These principal components form an orthogonal basis set of brain response patterns, which can be ordered by the degree of variance they explain in the set of scans. Because individual scans in the training/example set can be reconstructed (exactly) as a weighted combination of the principal components, it is possible to use these weights in lieu of the vector of voxel activations as input to a classifier algorithm.

The use of these weights can vastly reduce the "size/dimensionality" [7] of the input to the classifier in two distinct ways. First, from a practical point of view, full vectors of hundreds or even thousands of voxels can be condensed to consist of vectors containing only the weights of individual scans on the principal components (i.e., discarding the PCs from further analysis). Second, PCA is a dimensionality reduction technique in a more generic sense, because the original scans can be approximated using only a subset of principal components—those explaining relatively large amounts of variance in the dataset. This further reduces the "size" of the input vector by restricting the weight-based representation to contain a smaller number of elements than the number of available principal components [8] . The work of Hanson et al. (2004) evaluating a non-linear hidden-unit neural network suggests that the amount of data reduction possible for fMRI studies is substantial.

As an example, Strother, Anderson, Hansen, Kjems, Kustra, Sidtis, et al., (2002) used PCA compression to reduce input dimensionality and noted that it is particularly helpful in reducing noise in the input, thereby boosting the signal to noise ratio (cf., also Moeller & Strother, 1991). A further example of this approach can be seen in the work of Carlson et al. (2003) who selected a low dimensional representation of the information contained in the voxel activations by using the first 40 principal component axes. This representation allows for their knock-out procedure to target individual PC axes of voxel activations rather than single voxels.

Finally, an example of combining compression schemes can be found in O'Toole et al. (2005), who used Haxby et al.'s scheme to select significant voxels and then applied a PCA compression in which they retained the PCs with the most discriminative power.

In summary, pre-processing reduces the dimensionality of the input and can attenuate noise in the signal. It can also provide a convenient representation of the input (i.e., "eigenbrains") with strong value for visualizing patterns of brain activity that have discriminative power (cf., Figure 1c). On the less positive side, pre-processing can create statistical problems related to the exclusive use of a particular sample of variables in deriving predictors (i.e., voxels or PCs) of the experimental conditions.

Cross-validation

Cross-validation is the most commonly used method for evaluating the accuracy and generalizability of results from pattern-based classifiers. Cross-validation involves a separation of the available data into training and test set(s). The quality of a trained mapping between brain activation maps and experimental variables can be assessed, therefore, by evaluating the accuracy of the classifier on the "left-out" test-set data. Versions of cross-validation procedures include jackknifing, bootstrapping (Efron & Tibshirani, 1993), and reproducibility resampling (e.g., NPAIRS, Strother et al., 2002).

Cross-validation techniques are needed to evaluate the generality of the results from classifiers, because these data are not easily amenable to standard statistical inferential approaches. Specifically, as in any standard experiment, it is important to know how well the system will perform on a different sample. Experimental psychologists typically answer this question by performing an inferential statistical test (e.g., ANOVA). A "significant" result is interpreted as a potentially "replicable" result. At present, there is no standard statistical practice for assessing the inferential validity of pattern classifiers (e.g., Duda, Hart & Stork 2001). Cross-validation has been used, therefore, to address questions about the robustness and replicability of classifiers across samples.

Although the generalizability problem has been considered extensively in the neural networks and machine learning literatures (cf., e.g., Duda et al., 2001, Ripley, 1996; Hastie, Tibshirani, & Friedman, 2001; Bishop, 2006), brain imaging data pose unique challenges. The most critical problem to consider in implementing cross-validation on spatially arrayed brain map data is the issue of temporal structure of the data set. Specifically, the temporal struc-

[7] The word size is used here because "dimensionality" could refer either to the dimensionality or number of elements in an input vector (of either voxels or weights) or to the true dimensionality of the problem, in more mathematical terms (the rank of the covariance matrix).

[8] This is equal to the rank of the matrix, which is less than or equal to the lesser of n and m, where n is the number of voxels per scan and m is the number of scans included in the PCA.

ture of experimental conditions combined with the hemodynamic correlation among temporally contiguous scans (e.g., two scans performed closely in time will be similar for reasons unrelated to the experimental condition) makes the choice of training and test sets difficult and potentially subject to confounds. These sampling issues play out in different ways for event-related versus block designs. Notwithstanding, the temporal correlation problem must be taken into account in implementing cross-validation procedures. Strother et al. (2002) and LaConte, Anderson, Muley, Ashe, Frutiger, Rehm, et al. (2003) discuss this issue in detail and present a comprehensive approach to the problem for brain imaging data.

Critical features of the classifiers

In what follows, we overview some important features of classifiers applied to the analysis of functional neuroimaging data. The section is meant to provide an accessible and general entry into the kinds of classifiers used in this literature, with the caveat that more detailed statistical references are available in other work (cf., Bishop, 2006; Duda, Hart & Stork, 2001; Hastie et al., 2001; Ripley, 1996).

Assessing Brain Map Similarity

Nearly all classifiers operate on measures of the similarity (or inversely, distances) among the stimuli to be classified. An important difference among pattern-based classifiers is the way these distances are computed. In the functional neuroimaging literature, computations vary from using simple correlations to compute brain map similarity (e.g., Haxby et al. 2001; Spiridon & Kanwisher, 2002) to comparisons based on multivariate representations of the brain scans (e.g., Cox & Savoy, 2003; Strother et. al., 1998).

Similarity based on correlation

Classifiers that categorize inputs based on correlation treat brain maps as patterns, but use only a fraction of the data available in the patterns for classification. Correlation provides a measure of the similarity of two patterns. The contribution of any given voxel to the computation of similarity is based on the deviation of its response from the average response across voxels—a computation that can be inordinately affected by the presence of outliers.

By comparison to voxel-based inferential approaches, classifiers that categorize based on correlation increase the quantity of information considered from *n* independent voxels to 2*n*. Specifically, as implemented in the classifiers we have mentioned, correlation measures the accord between the neural activity in two experimental conditions *at the same locations* in the brain. The decision of whether a brain activity map was generated during a particular experimental condition depends, therefore, on the combined estimate of how similar the brain responses are across the *n* pairs of location-matched voxels.

Similarity based on multivariate representations

Other pattern-based classifiers *learn* a mapping function from a multivariate representation of the brain activity directly onto a variable indicating the experimental condition from which the scans originate. The learning makes use of "example" mappings from scans to experimental conditions. Test data are computed on mappings that were not part of the example data. The primary technical advantage of multivariate representation is that a classifier can learn, not only the discriminative power of individual voxels, but also the dicriminative power of *voxel combinations*. Classifiers can seek, therefore, an "optimal" separation of categories in a *multivariate* space by finding a set of weights for combining voxels to accurately predict experimental conditions.

The substantive advantage of these classifiers is that they make use of the relationship among responses at *different* locations in the brain (cf., Cox & Savoy, 2003). Considering the covariance of all possible pairs of voxels in brain activities, rather than attending selectively only to the pairs located at corresponding brain locations (as in correlation), effectively increases the information available for each measure from $2n$ in correlation, to n^2 in a multivariate analysis.

Most of the classifiers used for functional neuroimaging data analysis operate on a multivariate representation of brain activity. These include SVM's (e.g., Cox & Savoy, 2003), LDA's (Carlson et al., 2003; O'Toole et al., 2005), feed-forward neural networks (Hanson et al., 2004), and PLS (McIntosh et al., 1996). To summarize, classifiers that categorize based on correlation between brain maps (Haxby et al., 2001; Spiridon & Kanwisher, 2002) operate on patterns, but not in a multivariate space. Classifiers like LDA, SVM, and PLS are pattern-based and have the advantage of making use of the multivariate structure of the space in categorization.

In choosing the parameters involved in implementing a particular classifier, a reappearing principle for guiding the reader through the next few sections can be summarized as follows. *If information is available and useful for classification, but not used by a particular classifier, that classifier is deficient by comparison to a better one that can exploit the available and useful information*. This says both nothing and everything about which classifier to use. It says nothing, because the true answer to the question of what information is available and useful for classifying a particular dataset is entirely empirical. To find out what information is available and useful, one has to implement classifiers that exploit various kinds of information and see which one works best.

In the absence of the data one could gain by taking the impractical route of implementing all possible classifiers, an educated guess about an adequate classifier is needed. This brings us to the second axiom. *Implement the simplest classifier able to effectively classify the data at hand, with the caveat that the classifier be consistent with theoretical notions about the nature of the information in neural codes.* Although beyond the scope of the present paper to justify, we claim that there is more than ample evidence to suggest that an important part of the neural code lies in the interaction of brain response across different spatial locations in the cortex. We recommend, therefore, that multivariate classifiers be the standard for functional neuroimaging analysis.

Linear or Nonlinear Classifiers?

Classifiers differ also in whether they are linear or nonlinear. This is a feature of the classifier algorithm itself. Classifiers seek an "optimal" separation of categories in a space by finding a set of weights for combining voxels. Suffice to say that in geometric terms, linear classifiers work in a representational space to find the best (hyper)-plane for separating the scans by condition. Non-linear classifiers can bend this (hyper)-plane in various (classifier-limited) ways to seek a better separation of scans by condition.

We re-iterate that it is impossible to know, a priori, what kind of surface is needed to classify brain maps in any given dataset. So, for linear versus non-linear, we recommend the following approach. Start with a linear classifier. If it fails, try a nonlinear classifier. Failure means that accuracy is either not above chance or is seriously below the accuracy humans achieve on the best-fit analogous task. Human accuracy is sometimes at ceiling in functional neuroimaging tasks, (e.g., we assume that humans never misclassify common objects). In many other cases, however, human accuracy varies by condition, so perfect accuracy in all conditions may not be a sensible goal for a classifier.

On the question of whether classifiers should be make use of the multivariate features of the data, we argued that there are ample data to support the view that at least some part of the neural code resides in interactions among neural activity at different locations in the brain. For the question of whether or not important principles of neural coding will be missed without considering *nonlinear* dividers between brain states, the literature is less clear. Two things are certain. First, if a linear classifier works to levels of accuracy that are theoretically acceptable for a given problem, without a theoretically motivated reason for seeking a nonlinear separator, there is little to be gained by using a nonlinear classifier.

Second, if linear separation does not work up to expectations, nonlinear classifiers are an available and sensible alternative. We cannot think of neural data that preclude brain state separation by nonlinear dividers. In this case, the best practice is to compare classifiers. The approach of Cox and Savoy (2003) comparing several classifier algorithms for the task of separating object categories from a functional neuroimaging experiment serves as a good example. They implemented three classifiers (LDA, SVM, and a cubic polynomial SVM classifier). LDA and SVM are linear classifiers, whereas the cubic polynomial SVM is a nonlinear classifier. The output of the classifier was a prediction of the object category viewed while the scan was taken. Cox and Savoy did not find a performance advantage for the nonlinear cubic polynomial SVM over its linear counterparts. There are several interpretations of for this lack of difference, including the possibility that the neural signal itself is linearly separable by object category or that the cubic polynomial function failed to capture the true nature of the decision boundaries.

Representational Spaces or Not?

Analyses that use representational spaces, (e.g., based on PCA, Carlson et al., 2003; O'Toole et al., 2005, ICA, or multi-layer feed-forward networks (Hanson et al., 2004) offer some advantages over direct voxel-based classifiers (e.g., SVM or LDA alone). Perhaps the primary one is that representational spaces allow for a richer analysis of the pattern-structure of functional neuroimaging brain maps than direct voxel-based classifiers. They also allow for visualization of brain states that are proven to be relevant for classifying by scans by experimental condition. This puts the interpretation of brain activity patterns on firmer ground.

Going from good to better to best, the following sums up the popcorn trail a classifier leaves behind when it is successful for classifying brain states. Classifiers based on correlation leave no popcorn trail for interpretation, beyond that available from the average brain map in each condition. In their simplest implementation, direct voxel based classifiers like LDA and SVM leave behind a pattern of weights that specifies how to combine the voxels to predict condition. Weights with larger absolute values are more important for the classification than weights closer to zero.

Classifiers that operate on a representational space leave behind a set of patterns and a set of corresponding weights indicating the importance of each of the patterns in predicting experimental condition. As noted, some of these patterns will be useful for predicting experimental condition and some (artifactual or simply unrelated to experimental manipulation) will not be useful. Representational spaces can prove helpful in understanding the dimensionality and complexity of the brain task being undertaken (see Hanson et al., 2004 for an excellent discussion of dimensionality reduction). Further, as illustrated by Carlson et al. (2003), there are important advantages in being able to selectively manipulate whole patterns linked to performance in a task, with the goal of assessing the side effects of deletion on

other tasks.

On the downside, representational spaces add complexity to the analysis that may or may not be helpful in answering the experimental questions at hand. This returns us to the principle "*implement the simplest classifier able to effectively classify the data at hand with the caveat that the classifier be consistent with theoretical notions about the nature of the information in neural codes*".

Challenges for pattern-based classifiers in functional neuroimaging analysis

Keeping inferential analyses in the picture

Researchers in the behavioral sciences are trained to embrace inferential statistics with great fervor—with good reason (Cohen, 1994). A major challenge ahead is to develop and standardize inferential methods for the pattern-based classifier methods in ways that solidify and complement current cross-validation approaches. Inferential methods are ultimately the best approach, when they can be applied appropriately. However, the application of a quantitatively sound analysis to the correct data is preferential to an application of inferential analyses to the incorrect type of data. Individual voxels are the wrong unit of analysis for functional neuroimaging.

Pattern-based classification results have been tested inferentially with a number of methods. At least two routes have been proposed and implemented. The first route measures the stability of the brain patterns contributing to the prediction (e.g., McIntosh et al., 1996) and the second measures the stability of classifier performance across participant brains.

For the first, PLS regression, as introduced to the brain imaging community by McIntosh et al. (1996), was described with an accompanying bootstrapping procedure for testing the inferential stability of the latent vectors (similar to principal components). Dedicated users of PLS typically employ this type of inferential analyses, as a matter of course.

Another way to test pattern-based classifiers inferentially is to apply inferential analyses across participant brains, to a measure of classifier success in different experimental conditions. As noted, classifier success is measured by experimental condition in much the same way as it is for human participant success—with a percent or number correct per condition. Applied to the brains of individual participants, classifiers can yield data that are structurally analogous to data from behavioral experiments (cf., O'Toole et. al., 2005, for an example), allowing for a direct comparison between behavioral and neural data. The primary weakness of this approach is that the extracted measures do not necessarily refer to brain activations that are at the same anatomical locations in the brains of different participants.

To retain staying power, an acceptable standard of inferential analyses for the results of pattern-based classifiers will ultimately need to emerge. In the interim, we stress that inferential analysis of voxels, though statistically rigorous, is based on such a small proportion of the available data from functional neuroimaging experiments and is not guaranteed to provide a meaningful assessment of the data.

So many pattern-based classifiers!

The science of pattern recognition has flourished for decades in engineering, statistics, physics, and cognitive science. Relatively minor classifier refinements that improve performance in an engineering or computer application from excellent to perfect performance are noteworthy in these literatures (e.g., face and fingerprint analysis), but are potentially an enormous distraction for the emergent use of these methods in functional neuroimaging analysis. A challenge for this field will be to *stay on task* and to use classifiers that get the job done without forcing the researcher into too many unwarranted assumptions. The job is to link patterns to experimental variables in ways that capture as much of the richness and complexity of the neural code as possible. Used in this least-common-denominator way, pattern-based classifiers can be used to address questions about neural codes, avoiding the potential pitfall of using neural codes to address questions about the quality of pattern-based classifiers. In no way do we wish to suggest that all classifiers are equivalent—they are not. We suggest only that achieving the highest possible level of classification accuracy may be less important in many functional neuroimaging applications than arriving at understandable and transparently interpretable solutions to the problem of separating brain states.

Concluding remarks

A large part of the appeal of pattern-based classifiers for the analysis of functional neuroimaging data is the claim that they are acting like "brain-readers." This is a real change in experimental question from previous approaches. We have argued in the context of the recent literature on face and object processing that pattern-based prediction has advanced our understanding of the neural coding principles underlying high level visual representations of faces and objects. This is true even if there is still controversy about the answers to these questions. The most important outcome of the use of classifiers in the face and object processing area is that it is making researchers think in new ways about the representation questions. This extends the appeal of the brain-reading metaphor beyond prediction and will be of long-term value to cognitive science.

In summary, pattern-based classifiers make a fundamental technical advance in the state of the art by linking patterns of brain activity to experimental design variables. In the context of appropriately framed questions, these analyses open a door toward advancing functional neuroimaging studies beyond cortical localization toward questions that offer insight into neural codes. The time is ripe for allowing these analyses to replace the status quo.

References

[1] Abdi, H. (2003) Multivariate analysis. In M. Lewis-Beck, A. Bryman, T. Futing (Eds): *Encyclopedia for research methods for the social sciences*. Thousand Oaks (CA): Sage.

[2] Abdi, H. (2007). Metric Multidimensional Scaling. In N. Salkind (Ed) : *Encyclopedia of measurement and statistics.* Thousand Oaks (CA): Sage

[3] Abdi, H. & Valentin, D. (2006) *Mathématiques pour les sciences cognitives*. Grenoble: Presses Universitaires de Grenoble (PUG). (357p.) Collection: Sciences et techniques de la connaissance.

[4] Abdi, H., Valentin, D., O'Toole, A.J., & Edelman, B. (2005) DISTATIS: The analysis of multiple distance matrices. *Proceedings of the IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition*. (San Diego, CA, USA).

[5] Azari, N.P., Pettigrew, K.D., Schapiro, M.B., Haxby, J.V., Grady, C.L., Pietrini, P., Salerno, J.A., Heston, L.L., Rapoport, S.I., & Horwitz, B. (1993). Early detection of Alzheimer disease: A statistical approach using Positron Emission Tomographic data. *Journal of Cerebral Blood Flow and Metabolism, 13, 438-447.*

[6] Barlow, H. B. (1972) Single units and sensation: a neuron doctrine for perceptual psychology? *Perception, 1*, 371-394. Bishop, C. M. (2006). *Pattern recognition and machine learning.* New York: Springer Verlag.

[7] Carlson, T.A., Schrater, P., & He, S. (2003) Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*, 704-717.

[8] Clark, C.M., Ammann, W., Martin W.R.W., Ty, P., & Hayden, M.R. (1991). The FDG/PET methodology for early detection of disease onset: A statistical model: *Journal of Cerebral Blood Flow and Metabolism, 11,* A96-A102.

[9] Cohen, J. (1994). The earth is round (p<.05). *American Psychologist. 49*, 997-1003.

[10] Cox, D. & Savoy, R. (2003) Functional magnetic resonance imaging (fMRI) "Brain reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex. *Neuroimage, 19*, 261-270.

[11] Davatzikos, C., Ruparel, K., Fan, Y. Shen, D. G., Acharyya, M. Loughead, J. W., Gur, R. C., & Langleben, D.D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *Neuroimage*, .

[12] Duda, R. O., Hart, P. E. & Stork, D.G. (2001) *Pattern classfication*. New York: Wiley.

[13] Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Towards direct visualization of the internal shape representation space by fMRI. *Psychobiology, 26*, 309-321.

[14] Efron, B. & Tibshirani, R. (1993). *An Introduction to the bootstrap*. London: Chapman and Hall.

[15] Epstein, R. & Kanwisher, N. (1998) A Cortical Representation of the Local Visual Environment. *Nature, 392,* 598-601.

[16] Friston, K. J. (1998) Modes or models: A critique of independent component analysis for fMRI. *Trends in Cognitive Sciences, 2(10),* 373-375.

[17] Gauthier, I., Tarr, M.J., Anderson, A.W., Skudlarski, P. & Gore, J.C. (1999) Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nature Neuroscience, 2(6),* 568-573.

[18] Grill-Spector, K., Sayres, R., & Rees, D. (2006). High resolution imaging reveals highly selective nonface clusters in the fusiform face area. *Nature Neuroscience, 9,* 1177-1185.

[19] Hanson, S., 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.

[20] Hastie, T., Tibshirani, R., & Friedman, J. (2001) *The elements of statistical learning: Data-mining, inference, and prediction*. New York: Springer-Verlag.

[21] Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L. & Pietrini, P. (2001) Distributed and Overlapping Representation of Faces and Objects in Ventral Temporal Cortex. *Science, 293*, 2425-2430.

[22] Haynes, J. & Rees, G. (2005a) Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience, 8(5),* 686-691.

[23] Haynes, J. & Rees, G. (2005b) Predicting the stream of consciousness from activity in human visual cortex. *Current Biology, 15*, 1301-1307.

[24] Haynes, J.D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Review Neuroscience*, *7(7)*, 523-534.

[25] Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L. & Haxby, J.V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Science*, *USA, 96,* 9379-9384.

[26] Kamitani, Y. & Tong, F. (2005) Decoding the visual and subjective contents of the human brain. *Nature Neuroscience, 8(5),* 679-685.

[27] Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A Module in human extrastriate cortex specialized for the perception of faces. *Journal of Neuroscience, 17,* 4302-4311.

[28] Kherif, F., Poline, J.P., Mériaux, S., Benali, H., Flandin, G., & Brett, M. (2003) Group analysis in functional neuroimaging: Selecting subjects using similarity measures. *NeuroImage, 20,* 2197-2208.

[29] Kippenhan, J. S., Barker, W. W., Pascal, S., Nagel, J., & Duara, R. (1992). Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects. *Journal of Nuclear Medicine, 33(8)*, 1459-1467.

[30] LaConte, S.L., Anderson, J., Muley, S., Ashe, J., Frutiger, S., Rehm, K., Hansen, L.K., Yacoub, E., Hu, X., Rottenberg, D., & Strother, S. (2003). The evaluation of preprocessing choices in single subject BOLD fMRI using NPAIRS performance metrics. *NeuroImage, 18,* 10-27

[31] Lobaugh, N.J., West, R., & McIntosh, A.R. (2001). Spatiotemporal analysis of experimental differences in event-related potential data with partial least squares. *Psychophysiology, 38,* 517-530.

[32] McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J., and Sejnowski, T. J. (1998) Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping, 6*, 160-188.

[33] McIntosh, A. R. & Lobaugh, N. J. (2004) Partial least squares analysis of neuroimaging data: applications and advances. *Neuroimage, 23 Suppl 1*, S250-63.

[34] McIntosh, A. R., Bookstein, F. L., Haxby, J. V., and Grady, C.L. (1996) Spatial pattern analysis of functional brain images using partial least squares. *Neuroimage, 3,* 143-157.

[35] 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.

[36] Moeller, J.R., Strother, S.C., (1991). A regional covariance approach to the analysis of functional patterns in positron emission tomographic data. *Journal of Cerebral Blood Flow and Metabolism, 11,* A121-A135.

[37] 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 machines on functional MRI data. *Neuroimage, 28,* 980-995.

[38] Newsome, W.T., Britten, K.H., & Movshon, J.A. (1989). Neuronal correlates of a perceptual decision. *Nature, 341,* 52-54.

[39] 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.

[40] O'Toole, A. J., Jiang, F., & Abdi, H. (2005) Partially distributed representations of objects and faces in ventral temporal cortex. *Journal of Cognitive Neuroscience, 17,* 580-590.

[41] Peelen, M. V., Wiggett, A. J., & Downing, P. E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas. *Neuron, 49*, 815-822.

[42] Pessoa, L. & Padmala, S. (2005). Quantitative prediction of perceptual decisions furing near threshold fear detection. *Proceedings of the National Academy of Sciences, 102*, 5612-5617.

[43] Petersson, K. M., Nichols, T. E., Poline, J., & Holms, A. P. (1999a). Non-inferential methods and statistical models I. Signal detection and statistical inference. *Philosophical Transactions of the Royal Society of London B 354***,** 1261-1281.

[44] Petersson, K. M., Nichols, T. E., Poline, J., & Holms, A. P. (1999b). Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philosophical Transactions of the Royal Society of London B 354,* 1261-1281.

[45] Ripley, B. D. (1996) *Pattern recognition and neural networks*. Cambridge, U.K: Cambridge University Press.

[46] Schwartzlose, R. F., Baker, C. & Kanwisher, N. (2005). Separate face and body selectivity on the fusiform gyrus. *The Journal of Neuroscience, 25,* 11055-11059.

[47] Shinkareva, S.V., Ombao, H.C., Sutton, B.P., Mohanty, A., & Miller, G.A. (2006). Classification of functional brain images with a spatio-temporal dissimilarity map. *NeuroImage, 33,* 63-71.

[48] Spiridon, M. & Kanwisher, N. (2002). How Distributed is Visual Category Information in Human Occipito-Temporal Cortex? An fMRI Study. *Neuron, 35*, 1157-1165.

[49] Strother, S., Anderson, J., Hansen, L.K., Kjems U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., & Rottenberg D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. *NeuroImage, 15,* 747-771.

[50] Strother S.C., Anderson, J.R., Schaper, K.A., Sidtis, J.S., Rottenberg, D.A., (1995). Linear models of orthogonal subspaces and networks from functional activation PET studies of the human brain. In Bizals Y. (Ed.) *Information Processing in Mdeical Imaging.* New York: Kluwer. Pp 299-310.

[51] Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain.* MIT Press, Cambridge: MA.

[52] Young, M. P. & Yamane, S. (1992) Sparse population coding of faces in the inferotemporal cortex. *Science, 256,* 1327-1331.