CellPress

A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives

John-Dylan Haynes[1,2,3,4,5,6,7,*]
[1]Bernstein Center for Computational Neuroscience, Charité – Universitätsmedizin, 10117 Berlin, Germany
[2]Berlin Center for Advanced Neuroimaging, Charité – Universitätsmedizin, 10117 Berlin, Germany
[3]Berlin School of Mind and Brain, Humboldt Universität zu Berlin, 10117 Berlin, Germany
[4]Department of Neurology, Charité – Universitätsmedizin, 10117 Berlin, Germany
[5]Department of Psychology, Humboldt Universität zu Berlin, 10117 Berlin, Germany
[6]Cluster of Excellence NeuroCure, Charité – Universitätsmedizin, 10117 Berlin, Germany
[7]SFB 940, Volition and Cognitive Control, Technische Universität Dresden, 01069 Dresden, Germany
*Correspondence: haynes@bccn-berlin.de

http://dx.doi.org/10.1016/j.neuron.2015.05.025

Human fMRI signals exhibit a spatial patterning that contains detailed information about a person's mental states. Using classifiers it is possible to access this information and study brain processes at the level of individual mental representations. The precise link between fMRI signals and neural population signals still needs to be unraveled. Also, the interpretation of classification studies needs to be handled with care. Nonetheless, pattern-based analyses make it possible to investigate human representational spaces in unprecedented ways, especially when combined with computational modeling.

Introduction

The invention of functional magnetic resonance imaging (fMRI) marked an important milestone in cognitive neuroscience (Ogawa et al., 1990). fMRI made it possible to measure human brain activity with a considerably higher spatial resolution than previous noninvasive neuroimaging techniques, such as electroencephalography (EEG). In its early days, fMRI was used mainly to study brain activity at the level of macro-anatomical regions based on group data that were spatially smoothed and anatomically warped to standard templates. Statistical analyses were performed separately for each brain voxel using a general linear model (GLM; Friston et al., 1995b). In this mass-univariate approach, local spatial dependencies are nonetheless present due to the smoothing and the spatial spread of the hemodynamic response. The main focus was not on single-voxel activity, but on smooth, regional differences in brain activity between experimental conditions.

However, smoothed group results were not suitable for addressing a key question in cognitive neuroscience, that is, how specific and individual cognitive representations (or mental contents) are encoded and transformed in the human brain. For example, activity in lateral prefrontal regions has been shown to be increased under higher versus lower working memory load (Braver et al., 1997). Prefrontal activity might reflect the storage of working memory contents across a delay (Lee et al., 2013). Alternatively, the signal might reflect unspecific control or updating signals, while the contents might be stored elsewhere in the brain. To distinguish between these two possibilities, it is important to test whether the signal patterns in prefrontal cortex allow one to distinguish between different working memory contents (Lee et al., 2013). The key problem that impeded content-based fMRI studies was that neural encoding of specific contents occurs in cortical columns at a submillimeter scale (Mountcastle, 1957), which is below the spatial resolution of standard fMRI, especially when data are smoothed and averaged across subjects. A few work-arounds were developed to attempt to study content processing with neuroimaging techniques, such as fMRI adaptation or tagging by frequencies or categories (Grill-Spector et al., 1999; O'Craven et al., 1999; Tononi et al., 1998). However, the applicability of category-tagging is highly limited to a few categories (Downing et al., 2006), and fMRI adaptation leaves the underlying physiological mechanisms unclear (De Baene and Vogels, 2010).

Pattern-based fMRI analyses make it possible to address content-based processing in the human brain without relying on selective adaptation or frequency tagging. The idea is to directly study the link between mental representations and corresponding multivoxel fMRI activity patterns. This content-selective spatial patterning is conceptually related to theories of neural representation involving population codes, where each content involves the distributed activation of more than one representational unit (Pouget et al., 2000). The aim of this primer is to provide a concise introduction to the key concepts of pattern-based analysis. Another goal is to present an overview of challenges and limitations in the interpretation of decoding results, especially with respect to underlying neural population signals. Where appropriate, the reader is pointed to more in-depth reviews on specialized topics, such as cognitive neuroscience applications (Haynes and Rees, 2006; Norman et al., 2006; Tong and Pratte, 2012), neural coding (Haxby et al., 2014; Kriegeskorte, 2009; Kriegeskorte and Kievit, 2013; Naselaris et al., 2011; Serences and Saproo, 2012), and methods and algorithms (Pereira et al., 2009; Mur et al., 2009).

Spatial Patterning of fMRI Signals

The spatial resolution of fMRI is highly limited compared to invasive measurements of neural activity. A single voxel with a size of a few millimeters can sample up to several million neurons (Logothetis, 2008). The spatial resolution of standard fMRI measurements is also lower than would be required to sample

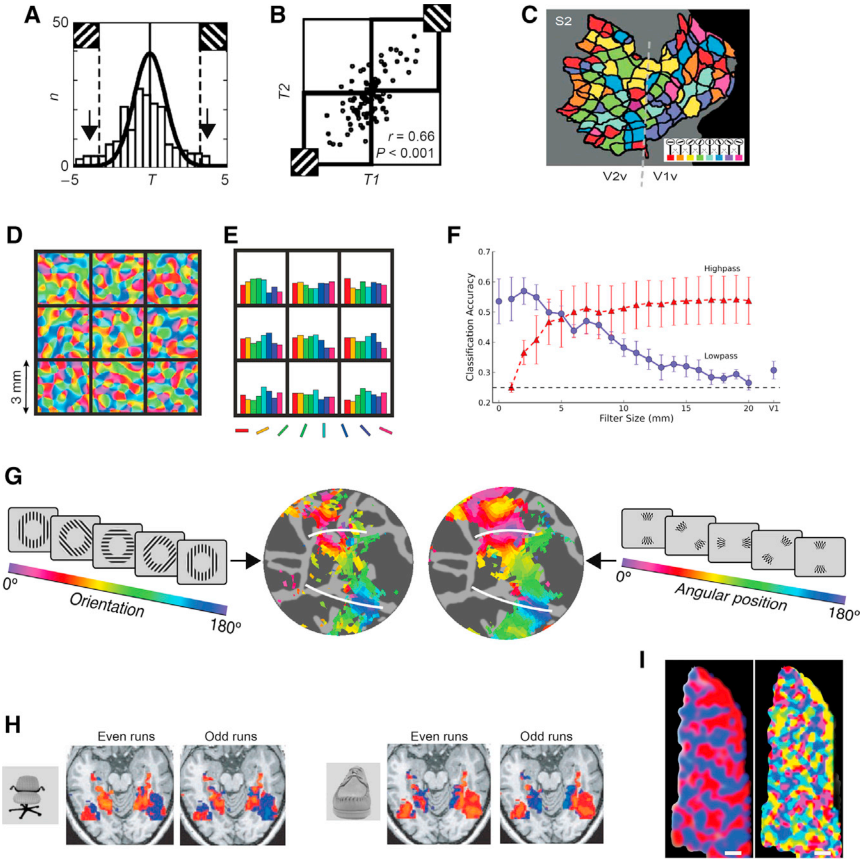

Figure 1. Spatial Patterning of fMRI Signals

(A) Histogram of orientation biases T of 100 voxels in V1 for two orthogonal orientations. The orientation bias T reflects the degree to which a voxel responds more strongly to either a left-tilted or a right-tilted orientation stimulus (Haynes and Rees, 2005).

(B) A scatterplot of biases of the 100 voxels for two independent measurement periods. This reveals a high correlation and thus reproducibility.

(C) This map shows a flattened representation of V1 where the orientation eliciting the strongest response is color coded (see inset) for each voxel (Kamitani and Tong, 2005).

(D) A simulated map of cortical orientation columns where the dominant orientation at each location is color coded. Note the coarse sampling of the columns by the comparatively large fMRI voxels (black grid; Boynton, 2005).

(E) Due to slight irregularities in the orientation columnar map, each voxel samples a slightly different number of cells of each orientation preference (shown here as a histogram). This is known as the biased sampling model (Haynes and Rees, 2005; Kamitani and Tong, 2005).

(F) High- and low-pass filtering of fMRI signals at different filter cutoffs and the resulting orientation classification accuracy. Low-pass filtering at increasingly lower spatial scales decreases the orientation information in human V1 (blue curve). High-pass filtering removes the lower spatial frequencies but still maintains orientation information (red curve; Swisher et al., 2010; filter size shown as millimeters; error bars represent SEM).

(G) Macroscopic biases contribute to voxel selectivity (Freeman et al., 2011). The experiment on the left presented grating stimuli with different orientations. The map shows which orientation yielded the strongest responses in different regions of early visual cortex. For comparison, the right shows the results of a standard retinotopic mapping stimulus with color-coded visual angle. Regions coding the horizontal meridian (blue/purple in the right map) respond strongest to vertical gratings, as would be expected if the map on the right were predominantly due to radial biases (Freeman et al., 2011).

(H) Patterning of fMRI signals in temporal cortex during observation of two objects (left, chairs; right, shoes). Hot/cold colors indicate stronger/weaker responses compared to the mean across all conditions. The brain response patterns averaged separately for even and odd runs are similar for the same objects, but different between objects (Haxby et al., 2001).

(I) Ocular dominance (left) and orientation maps (right) in human primary visual cortex obtained with ultra-high-field fMRI at 7 T (Yacoub et al., 2008; Copyright 2008 National Academy of Sciences, USA). The color on the left indicates which eye a given region responds to most strongly (red, right eye; blue, left eye). The color on the right indicates the orientation that elicits the strongest response in a given voxel (as in C–E). Both maps are obtained using a contrastive analysis by assessing whether voxels respond stronger to one versus all other conditions. It is currently impossible with fMRI to obtain voxels that are solely dependent on left or right eye stimulation.

individual human cortical columns that have a sub-millimeter spatial scale (Adams et al., 2007). Due to this comparatively low resolution, it might seem surprising that the responses of single voxels can be modulated by stimulus features that are encoded at a columnar level (see Figures 1A–1C; Haynes and Rees, 2005; Kamitani and Tong, 2005). For example, voxels in V1 can respond stronger to a visually presented grating of one orientation than to others (Haynes and Rees, 2005; Kamitani and Tong, 2005). Figure 1A shows an example of orientation biases T for 100 voxels in a single person's primary visual cortex. Participants viewed grating stimuli that were either left tilted or right tilted. The T value expresses the degree to which each voxel responds stronger to either left- or right-tilted gratings. Several voxels express orientation biases that are stronger than would be expected by chance (arrows in Figure 1A). The orientation biases are reproducible across different measurements (Figure 1B). Figure 1C shows a spatial map of a single person's V1/V2 region where each voxel is color coded according to the orientation that elicits the strongest response (Kamitani and Tong, 2005).

Several explanations for this patterning of fMRI signals have been offered. In the biased sampling account (also misleadingly referred to as "hyperacuity" or "aliasing"), the differential responses reflect the fact that each voxel samples cells arranged in cortical columns. Due to slight fluctuations in the columnar maps, each voxel will sample a slightly different number of each cell type (Figures 1D and 1E; Boynton, 2005; Haynes and Rees, 2005; Kamitani and Tong, 2005). For example, one voxel might sample more cells coding for horizontal orientations, whereas another might sample more cells coding for vertical orientations. As a consequence, voxels are expected to respond slightly more to one orientation than to others, which is the case (Haynes and Rees, 2005; Kamitani and Tong, 2005). The degree of this orientation bias depends on the spatial distribution of cells with different tuning properties within individual voxels. If a voxel samples a homogenous population of neurons with

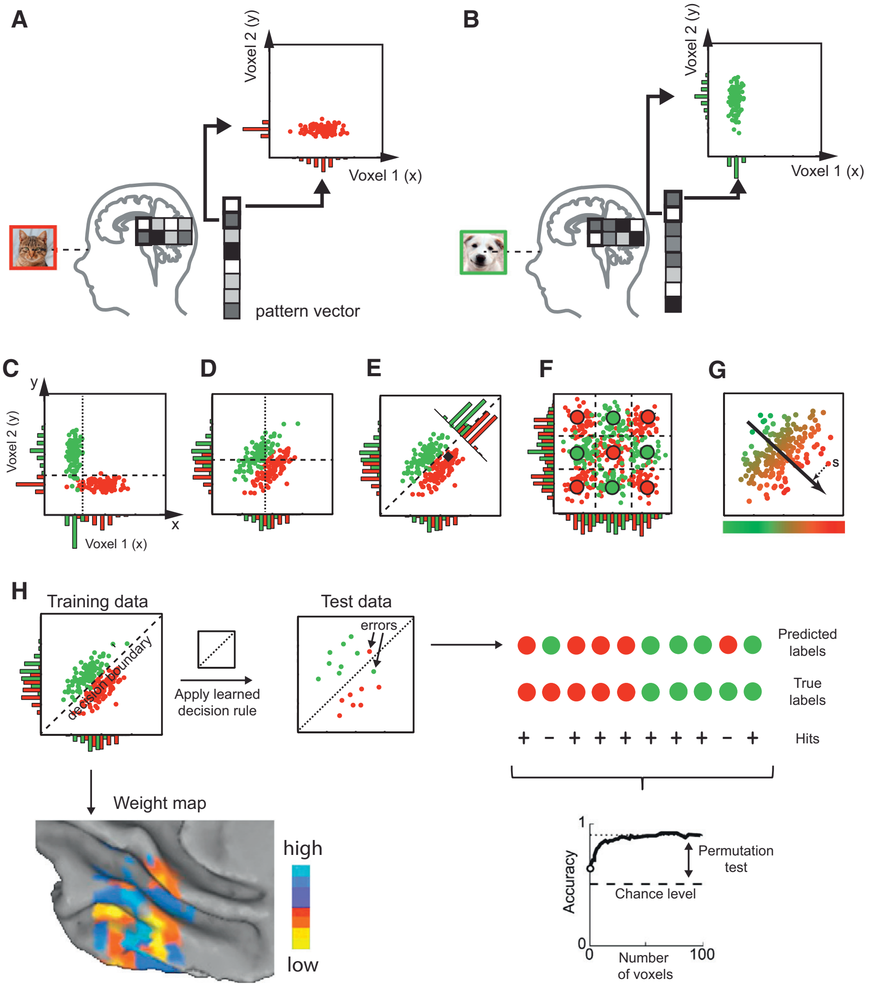

Figure 2. Multivariate Pattern Classification

(A and B) Hypothetical brain activity measured in eight voxels for two different conditions while a person is viewing images of cats or dogs (labels). The response amplitudes in the first two voxels are plotted in a two-dimensional coordinate system, separately for both conditions (red/green).

(C and D) A distribution of measurements can be separable from the brain activity in single voxels (vertical dotted and horizontal dashed lines) if the marginal distributions are not overlapping (C, red/green distributions on the axes), but not if they are overlapping (D).

(E) In this case, a linear decision boundary (dashed line) can be used to separate the response distributions by taking into account the activity in both voxels simultaneously. Here a linear decision boundary was estimated using linear discriminant analysis (LDA; Fisher, 1936) that maximizes the between-class to within-class variance. Support vector machines (SVMs, see Duda et al., 2000) are also commonly used as linear classifiers. While LDA and SVM differ in the algorithms for parameter estimation, the classification itself is identical in both cases and involves a linear projection of the data onto a decision axis.

(F) In certain cases as here, a linear decision boundary is not suitable for classification and nonlinear classifiers can be employed (see Duda et al., 2000).

(G) Multivariate regression can be used in cases where one is interested in continuous rather than discrete labels. The plot shows the continuous label as a graded color code. The regression is computed by projecting a sample (**s**) to the decision axis that explains maximal variance in the continuous labels.

(H) Cross-validation. The classifier is trained on part of the data (training data) and then applied to a statistically independent test dataset. This yields a predicted label for each sample that can be compared to the true label (right). The proportion of correct classifications then yields a classification accuracy that can be tested against chance performance. The bottom right graph shows a plot of accuracy obtained for an increasing number of voxels. The more voxels enter the classification the higher the accuracy. At some point the classification saturates. The weights of a linear classifier can be used to plot a weight map on the cortical surface (bottom left, taken from Kahnt et al., 2010). The different colors indicate the weights of the classifier. In this study, linear classification weights were either positively or negatively predictive of high reward (Kahnt et al., 2010).

similar response profiles, then the voxel tuning will be strong. If it samples neurons with a very divergent set of properties, the voxel-level tuning will be very weak (Chaimow et al., 2011).

This biased sampling account has been debated for several reasons. For example, if the voxel tuning effects were to reflect a sampling of cortical columns, one would intuitively expect that smoothing substantially decreases the effect. Generally speaking, the rationale is that the random biases between neighboring voxels should average out with smoothing (but see Kamitani and Sawahata, 2010). Experimental measurements of the effects of smoothing have yielded quite divergent results. In one report (Op de Beeck, 2010), smoothing actually improved orientation information that could be decoded from fMRI signals V1. In contrast, a different study compared high-pass and low-pass filtering and found a dominant spatial scale of information around 2–10 mm (Figure 1F), which is compatible with biased sampling (Swisher et al., 2010).

Several other studies have found evidence that voxel tuning might reflect macroscopic biases instead of local sampling biases (Freeman et al., 2011; Freeman et al., 2013; Sasaki et al., 2006). The idea is that, for example, orientation preferences of voxels are influenced by global biases in orientation processing across different regions of the visual field, such as the oblique effect (Furmanski and Engel, 2000) or radial bias effect (Sasaki et al., 2006; Freeman et al., 2011; see Figure 1G). This in turn might give a false impression of columnar sampling. The question of the origin and spatial scale of the patterning of fMRI signals is still under debate (Alink et al., 2013; Wang et al., 2014). Patterning also has been observed for high-level visual features, such as object stimuli in occipito-temporal cortex (Figure 1H; Haxby et al., 2001) or for reward representations in orbitofrontal cortex (Figure 2H, bottom left; Kahnt et al., 2010).

One interesting future direction is to attempt to go beyond biased sampling and directly image columnar signals in human fMRI using higher-resolution imaging sequences (Figure 1I). Columnar imaging has been reported for eye-of-origin and orientation (Yacoub et al., 2008) and motion direction (Zimmermann et al., 2011), but improving day-to-day reliability using optimized

MRI imaging sequences remains an important challenge (Zimmermann et al., 2011). To further clarify the spatiotemporal dynamics with which fMRI samples cortical tissue, a combination of invasive recordings and computational modeling is needed (Chaimow et al., 2011; Gardner, 2010; Kriegeskorte et al., 2010; Nevado et al., 2004; Shmuel et al., 2010). One important question is how the vascular geometry samples the spatial topography with which individual neurons are distributed on the cortex (Gardner, 2010).

Analyzing Patterned fMRI Signals

To analyze the full information contained in spatially distributed fMRI signal patterns, a multivariate (as opposed to mass-univariate) framework is required (Allefeld and Haynes, 2015; Cox and Savoy, 2003; Friston et al., 1995a; Haynes and Rees, 2006; Haxby et al., 2001; Haxby et al., 2014; Kamitani and Tong, 2005; Kriegeskorte et al., 2006; Mitchell et al., 2003; Norman et al., 2006; Tong and Pratte, 2012). In multivariate pattern classification (Figure 2), brain activity is analyzed at the level of patterns consisting of a number of voxels. The multivariate samples of brain activity are then assigned labels that indicate the condition under which they were acquired. For example, a sample acquired while someone was thinking about a cat or a dog would be labeled "cat" or "dog," respectively. Classification algorithms where such a class structure is imposed by the experimenter are referred to as "supervised learning," as opposed to unsupervised learning algorithms (e.g., Kohonen, 1989).

Figures 2A and 2B show hypothetical fMRI signals in eight voxels in visual cortex while a participant is viewing a picture of a cat (A, red) or dog (B, green). To understand how the classifier operates, it can help to consider the samples as defining points in a coordinate system. Each sample of brain activity can be thought of as a pattern vector, which is a one-dimensional array of numbers. This allows one to mathematically treat the list of activation values as values on different dimensions of a coordinate system. The first two entries in the pattern vectors (x,y), i.e., the measurements in the first two voxels, are shown in Figure 2 as scatterplots. The repeated measurements of the response yield two clouds of points, one for each category. The task of the classifier is to find a way to separate the two point distributions while at the same time avoiding overfitting (Bishop, 1995; Duda et al., 2000; see below).

Figure 2C shows a case where the two response distributions are separable based on each voxel (x,y) alone using a suitable decision boundary that is parallel either to the x or y axis. This single-voxel separation is possible because the marginal distributions (red and green) are not overlapping. However, in many cases, the marginal distributions for both conditions are highly overlapping (Figure 2D), and thus the classification cannot be performed by only considering one voxel. A multivariate solution is required that takes into account the activation values in both voxels simultaneously. One approach is to estimate a linear decision boundary (Figure 2E, dashed line) that partitions the response plane into regions with different labels. Two common ways to estimate linear decision boundaries are linear discriminant analysis (LDA; Fisher, 1936) and linear support vector machines (SVMs; for details, see Allefeld and Haynes, 2015; Duda et al., 2000; Pereira et al., 2009; and Mur et al., 2009). For classification, both LDA and SVM use a weight at each voxel to linearly project the data points to a single decision axis (top right of Figure 2E). However, the algorithms for estimating the weights from the training data are different. LDA identifies projection weights that maximize the between-class to within-class variance. SVM identifies weights that define a so-called maximum margin hyperplane (see Duda et al., 2000 for details). For visualization, the weights of the classifier for each voxel can be plotted as a weight map (Figure 2H, bottom left). In certain cases, response distributions cannot be sufficiently partitioned using single linear decision boundaries (Figure 2F). In these cases, nonlinear approaches such as nearest-neighbor classifiers or nonlinear SVMs can be used (for an overview, see Duda et al., 2000).

The logic of classification requires that the training data can be grouped into discrete categories, each with a unique label. However, often one might be interested in predicting a continuous (rather than discrete, categorical) variable from a multivariate signal (Figure 2G). This can be achieved using multivariate regression approaches. Here the decision boundary is replaced by a continuous variable to which each sample is projected (Smola and Schölkopf, 2004; Chu et al., 2011; Marquand et al., 2014).

To train the classifier, only a part of the samples is used, the training data. The remaining samples, constituting test data, are left out and used to assess whether the classifier can correctly assign the labels (Figure 2H). The reason for the separation into training and test data is to see whether the classifier can generalize to new test samples. The obtained decision boundary is applied to the individual samples in the left-out and statistically independent test dataset (Figure 2H, test data). The proportion of test samples that is correctly predicted is known as the classification accuracy. Typically, the procedure is repeated again using a different partitioning of data into training and test. This is known as cross-validation. It is absolutely vital that the training and test data are independent and stationary in order to avoid overfitting and circular inference.

To test whether the classifier can indeed extract information from the data, the classification accuracy is then compared to a chance accuracy (Figure 2H, bottom right), which is the proportion of samples that would have been labeled correctly based on guessing alone. In the case of two alternative categories, the chance accuracy is 0.5; in the case of n categories, it is 1/n. Even a classifier operating at chance level will have some variability in the accuracy for a fixed sample size. Thus, it is important to statistically test whether the accuracy is significantly above chance. For testing against chance, permutation tests have proven to be more valid than binomial tests or t tests (Nichols and Holmes, 2002; Noirhomme et al., 2014; Pereira and Botvinick, 2011; Schreiber and Krekelberg, 2013; Stelzer et al., 2013). The idea of these tests is to permute the assignment between labels and samples and thus obtain a distribution of chance accuracies against which a specific accuracy can be tested. Permutation testing has additional advantages because it requires only very weak distributional assumptions. Furthermore, it can help to reveal biases in the processing pipeline. For example, if the independence between training and test data were violated, this could lead to above-chance baseline-level

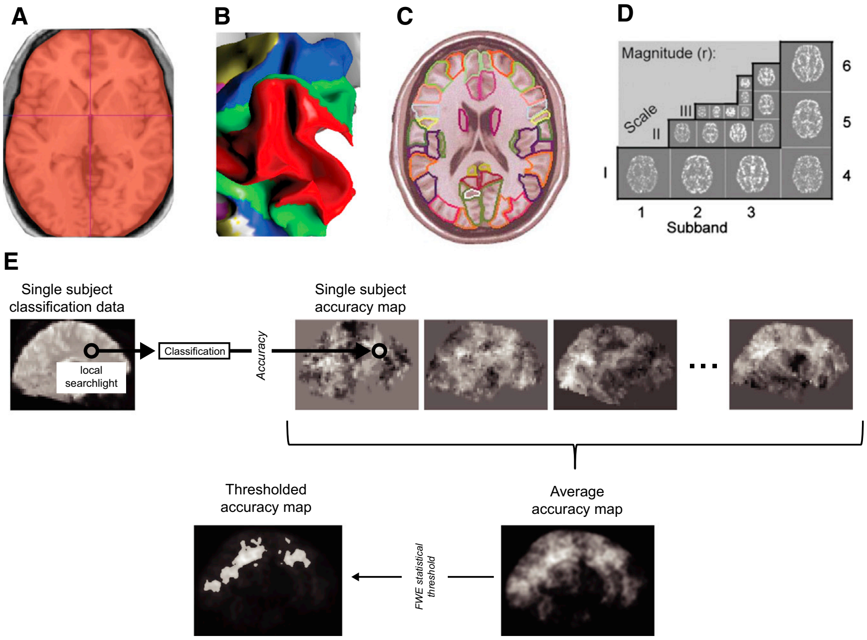

Figure 3. Spatial Selection for Pattern Classification

(A) In whole-brain classification, all voxels enter the pattern analysis simultaneously (shown here in red). Some form of dimensionality reduction is typically required for whole-brain classifiers (Mourão-Miranda et al., 2005).

(B) A region of interest (ROI) can be chosen based on functional localizers (the red region V1 is obtained by retinotopic mapping).

(C) ROIs also can be obtained based on anatomical criteria. The color-coded regions here are obtained using automated anatomical labeling (Tzourio-Mazoyer et al., 2002).

(D) Using wavelet pyramids (Hackmack et al., 2012) or other forms of spatial filtering (Swisher et al., 2010), it is possible to perform classification at multiple spatial scales and assess which scales contain the most information. The figure shows a brain image at different spatial scales used for a wavelet analysis (see Hackmack et al., 2012 for details).

(E) Searchlight analyses. Classification is performed separately for each local spherical cluster of voxels (indicated here by the circle). The classification accuracy is then entered into a single-subject searchlight accuracy map (middle), and the procedure is repeated across all brain locations. The maps for different subjects can then be averaged and subjected to a second-level statistical test (bottom). The result is a map (bottom left) that shows where in the brain local clusters of voxels contain significant information about the chosen conditions (see also Challenges and Pitfalls). Searchlight classification can be performed in 3D voxel space (Kriegeskorte et al., 2006) or on the cortical surface (Chen et al., 2011; Oosterhof et al., 2010), which can improve localization accuracy.

accuracies. In such a case, the whole brain might mistakenly appear to have decodable information about a cognitive variable.

What Goes into the Classifier?

An important choice in pattern classification is the format in which the data are entered into the classifier. The first choice is which voxels to use, i.e., the spatial selection (Figure 3). One might intuitively believe that it is safe to enter all available brain voxels into the classifier. Such whole-brain classification (Figure 3A), however, suffers from the curse of dimensionality (e.g., Scott, 1992) because of the high number of voxels in fMRI experiments (typically >100,000). Due to the low number of samples (i.e., time points) and the high number of dimensions (voxels), the activity in many voxels spuriously correlates with the labels. There are several ways to reduce the dimensionality of the whole-brain classification problem. One established way is to preprocess the data using a principal component analysis (PCA) and to enter a smaller set of components into the classifier (e.g., Mourão-Miranda et al., 2005). A different approach is to use an algorithm that automatically selects only a subset of voxels for the classification (De Martino et al., 2008). This is known as feature selection.

After dimensionality reduction or feature selection, classifiers generally still reflect information distributed across large-scale brain networks. There are several approaches that allow for more localized assessment of information coding. One way is to use regions of interest (ROIs), defined either functionally, say with retinotopic mapping (Sereno et al., 1995; Figure 3B), or based on anatomical criteria (Tzourio-Mazoyer et al., 2002; Figure 3C). A different approach is to spatially filter the data, for example, using three-dimensional (3D) wavelet pyramids (e.g., Hackmack et al., 2012; Figure 3D). This allows one to compare the information encoded at different spatial scales.

Another way to assess the information encoded in small regional networks is to use searchlight decoding (Figure 3E; Haynes et al., 2007; Kriegeskorte et al., 2006). In this approach, a classifier is applied to a small local cluster of voxels, typically a small spherical voxel cluster with a radius of a few millimeters centered on one brain location. The resulting classification accuracy at this brain location is entered into the corresponding position in a 3D brain map. The procedure is then repeated for different searchlight centers, thus yielding a whole-brain searchlight map that depicts the information contained at each local cluster of voxels. These searchlight analyses make the simplifying assumption that information is contained in local clusters of voxels, possibly reflecting local population codes (Pouget et al., 2000). However, they cannot access information encoded in a more distributed fashion across multiple brain regions (see also Challenges and Pitfalls).

A second important choice is the level of temporal aggregation (Figure 4). The individual samples entering a classifier analysis can stem from single fMRI volumes, single trials, single blocks, entire scanning runs, or even from single subjects (see Figure 4 for details). It is even possible to use entire spatiotemporal patterns for classification (Mourão-Miranda et al., 2007). The level of temporal aggregation is important when it comes to ensuring the statistical independence between training and test datasets in each of the cross-validation folds (Figure 4; see also Mumford et al., 2014). An extreme violation of independence would be to mix data from closely spaced trials in training and test data (Figure 4A). In this case, the temporal independence would be violated for several reasons. The signals in different trials would

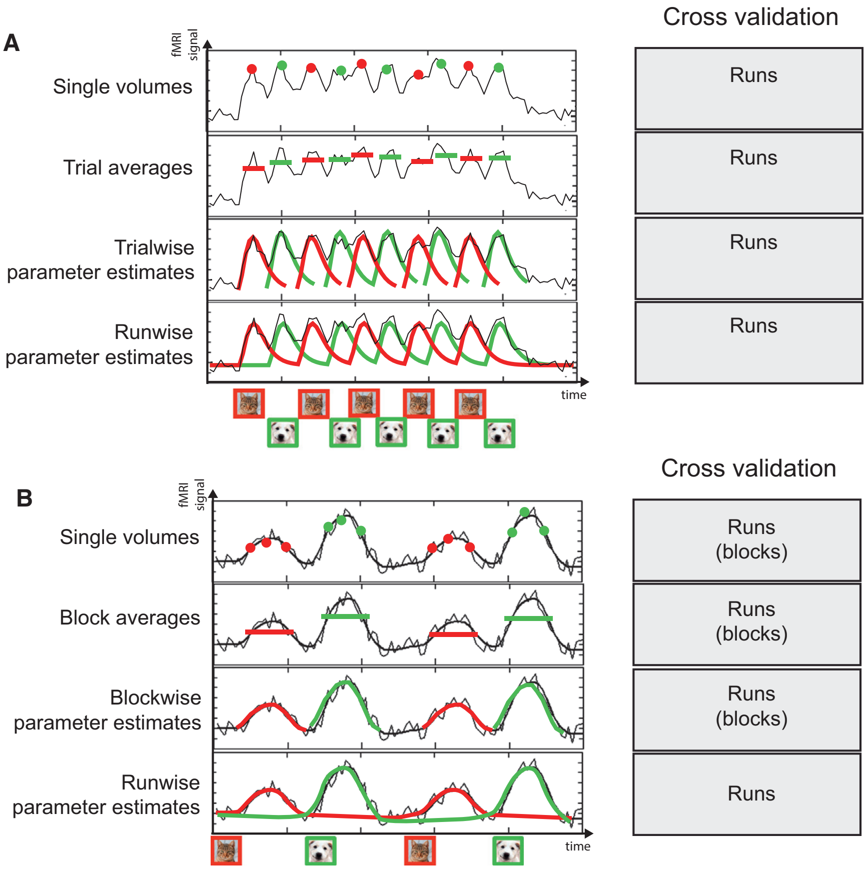

Figure 4. Temporal Selection of Classification Samples Based on Single Events or on Extended Blocks of Stimulation

(A) Schematic fMRI responses to two different stimulus conditions (red/green) in an event-related design, where individual trials occur more rapidly than the temporal extent of the hemodynamic response function. The brain responses to different trials are thus overlapping. Typical approaches to defining classification samples include the voxel-wise responses in single fMRI volumes, averages within a selected time window after onset of the trial, parameter estimates of canonical hemodynamic response functions for each trial, and parameter estimates of a general linear model for the entire scanning run. Due to the temporal overlap of trials and the autocorrelation of the fMRI signal, great care has to be taken to avoid spillover of information between training and test datasets. For this reason, cross-validation is best performed across different scanning periods (runs) between which the scanning is briefly stopped.

(B) In block designs, similar conditions are presented in extended temporal sequences. Here similar choices for samples can be made as for single events, with either single-volume activity, temporal averages across blocks, parameter estimates across individual blocks, or parameter estimates across entire runs. In contrast to event-related designs, for block designs a block-wise cross-validation might be an option if the danger of temporal overlap of information can be excluded. However, the safer option is to use a splitting based on runs also here in order to avoid spillover of information between training and test data.

be correlated due to the low-frequency autocorrelation of the fMRI signal, the temporal extent of the hemodynamic response, plus potential cognitive factors such as slow fluctuations in arousal or attention. These dependencies are lower for block designs (Figure 4B) but can still be observed (e.g., Goldfine et al., 2013). The safest solution is to perform a cross-validation across independent fMRI measurement periods (runs) for both trial-based and block-based experiments (Mumford et al., 2014).

Challenges and Pitfalls

Despite the significant advances made possible by fMRI decoding, it is important to also highlight a number of challenges and pitfalls in the design and interpretation of classification experiments. The question of what it means if a cognitive factor can be decoded from an fMRI signal pattern needs to be approached with care. Neither the information contained in single voxels nor in ensembles of voxels can be directly related to the information encoded in single neurons. The sampling of neural activity by fMRI voxels is highly indirect and involves the magnetization level of blood as a marker of neural activity, a pooling of many thousand neurons per voxel, and a sluggish and nonlinear hemodynamic response (Logothetis and Wandell, 2004). There are many complicating factors, as outlined below.

Interpreting Accuracies: Underestimating Information

An absence of information at the level of fMRI does not mean that the local neural populations do not contain information. For example, if neurons with different tuning properties were mixed randomly in a salt-and-pepper fashion, then no macroscopic information would be expected at the voxel level (Chaimow et al., 2011). Also, in an extreme case, a single neuron might contain substantial information that is drowned out by other neurons only contributing noise (e.g., Etzel et al., 2013). The tuning of a single voxel thus depends on the sampling of neurons in a complex way that can only be unraveled by direct invasive measurement of population signals in combination with computational modeling (Chaimow et al., 2011; Kriegeskorte, 2011; Nevado et al., 2004; Ramírez et al., 2014).

Interpreting Accuracies: Overestimating Information

There are several ways in which an observed accuracy with fMRI might overestimate the information that is computationally available at the neural level. For example, a voxel might sample a large blood vessel that drains a large population of neurons (Gardner, 2010) that share no direct anatomical connections. This could yield an aggregation of information that is not computationally used at the neural level. Also, the low sampling rate of fMRI signals and the sluggishness of the hemodynamic response might temporally integrate information beyond the relevant timescales of neural signal processing. Classifiers based on whole-brain activity could potentially be integrating information from widely disparate brain regions that are not anatomically connected, thus reflecting information that is only

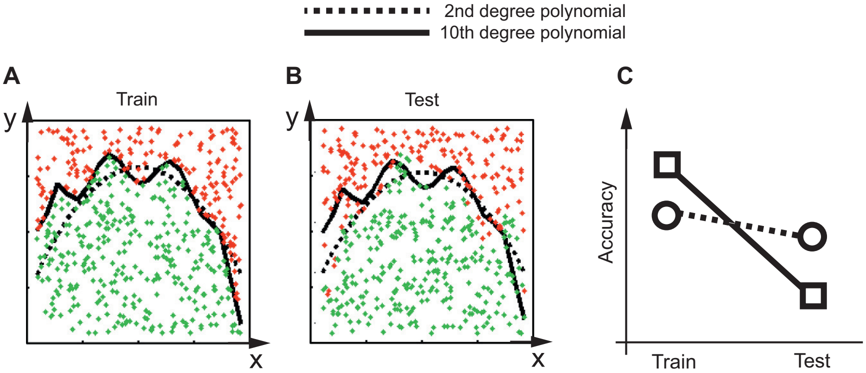

Figure 5. Overfitting of Training Data

(A) Two different nonlinear polynomial classifiers are fit to a two-class dataset (red and green) for activity measured at two voxels, x and y. One classifier is a second-degree and one a tenth-degree polynomial. The tenth-degree polynomial has more parameters and fits the training data better with fewer misclassifications.

(B) When applied to an independent test dataset, the second-degree polynomial exhibits better generalization performance and achieves higher accuracies in the test data than the tenth-degree classifier. This is presumably because the tenth-degree classifier is fitting spurious noise in the training data.

(C) Schematic plot of the crossover effect of accuracies. The tenth-degree polynomial (squares) performs better in the training data, but worse in the test data than the second-degree classifier (circles).

accessible to the scientist as an external observer, but not computationally usable within the brain itself.

Interpreting Accuracies: Comparing Different Brain Regions

There are several factors that limit the comparison of accuracies between different brain regions. For example, the size of regions generally is different, so the number of voxels entering into the classifier and thus the dimensionality of the classification problem differs. Also, the sensitivity of fMRI to neural activity in different brain regions might be quite variable, which is known as the local hemodynamic response efficiency (Logothetis and Wandell, 2004). The signal-to-noise levels also generally differ between regions, thus further limiting the interpretation of accuracies. For each classification task and each region, there might thus be a different noise ceiling that limits the maximum possible accuracy given the noise in the data (Kay et al., 2008; Nili et al., 2014).

Interpreting Accuracies: Other Limitations

There are more reasons why the overall level of accuracy of a classifier is difficult to interpret (Allefeld and Haynes, 2014). For example, the obtained accuracy depends on the partitioning of data into training and test. Less training data generally yield lower accuracies because the classifier has less ability to learn an optimal decision boundary. Factors such as experimental design efficiency (Josephs and Henson, 1999; Liu et al., 2001), the level of temporal aggregation (see above), or smoothing (Op de Beeck, 2010; Swisher et al., 2010) also have an impact on the accuracies obtained. Accuracy has an absolute ceiling at 100%, whereas it could be interesting to assess distances between pattern vectors obtained for two classes even beyond perfect classification. For this reason, multivariate distance measures between brain activity patterns might be more suitable to assess the information contained in voxel patterns (Allefeld and Haynes, 2014; Kriegeskorte et al., 2006).

Circularity and Overfitting

The importance of ensuring independence between training and test data was already discussed above. Any dependencies are likely to cause false-positive classification of the test data even in the absence of information (e.g., Mumford et al., 2014). Due to leakage of information between training and test data, the classifier would be training and testing on the same data. This is referred to as "double dipping" and it constitutes a circular inference (Kriegeskorte et al., 2009).

Classification analyses also might suffer from a related phenomenon known as overfitting. As outlined above (Figure 2), the aim of a classifier is to separate the neural response distributions belonging to several classes. Overfitting can occur if a too-complex classifier is fit to the training dataset that works well in the training data, but then fails to generalize to the test data. Figure 5 shows examples of two classifiers, a second-order and a tenth-order polynomial applied to a training dataset. The tenth-order polynomial has more parameters and fits the training data better than the second-order polynomial. However, when this classifier is applied to an independent test dataset, it becomes apparent that the second-order polynomial generalizes better. The reason is that the tenth-order polynomial is more flexible and also adjusts to the spurious noise in the training data. Testing the generalizability of a classifier on independent test data thus protects against overfitting.

However, cross-validation does not protect against overfitting if different classifiers are tried out on the same data. For example, a researcher might try out different options for spatial or temporal selection of samples, or might try out different classifiers (e.g., linear versus nonlinear) and different partitioning schemes of training and test data. Once several classifiers have been tried out, overfitting only can be revealed by testing the accuracy on a further independent test dataset. While overfitting is not unique to fMRI classification (Kriegeskorte et al., 2009) or even to neuroscience (Ioannidis et al., 2001), it is exacerbated by the high number of free parameters in fMRI classification.

One way to maintain the flexibility of trying out different classifiers while at the same time avoiding such overfitting is to use a nested cross-validation (e.g., Pereira et al., 2009). The idea is to divide the data into training and test, and then to further subdivide the training data into a second-level training and test set in order to try out different classification approaches (say linear versus nonlinear classification). Then, the best classifier from the second level can be used for classification of the test data at the first level. This allows for optimization of classification while at the same time avoiding overfitting and false-positive classification. A different solution to such overfitting is to use an approach that substantially decreases the number of free parameters (e.g., Allefeld and Haynes, 2014).

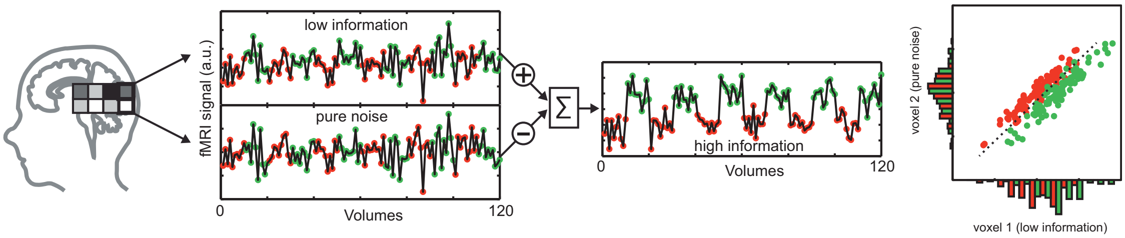

Figure 6. The Role of Noise Filtering in the Interpretation of Weight Maps
The activity in two voxels plotted as arbitrary fMRI signal units as a function of volumes. One of the voxels contains weak label-related information and one contains no information. Both are contaminated by the same noise source (which could reflect scanner noise or background activity). By subtracting the signal of the pure-noise voxel from the low-information voxel (middle plot), it is possible to recover substantially more label-related information. The right plot shows the data in both voxels plotted as x and y coordinates. The x voxel has weak information and slightly separated marginal distributions, and the y voxel has no information and fully overlapping marginal distributions. A linear classifier can nonetheless achieve high classification accuracy, which is only possible if the signal from the pure-noise voxel is included. For further details see Haufe et al. (2014).

Interpreting Classification Maps

Another challenge lies in the interpretation of the maps obtained through classification analyses. In a linear classifier such as LDA or SVM, the weight at each voxel directly reflects the contribution of that voxel to the classification result (provided the data are normalized separately for each voxel). The weights at each voxel are often plotted as a weight map (Figure 2H, bottom left). However, when the output of a classifier is statistically tested against chance level, this pertains to the classifier as a whole, and does not permit a conclusion as to whether an individual voxel contributed *significantly* to the result. To test whether a single voxel contributes significantly to the performance, it is necessary to test whether it makes a significant difference if the voxel is included in the classifier (for a related approach, see Pereira et al., 2009).

A further complication in interpreting weight maps lies in the fact that a voxel might have a significant weight despite not having label-related information. Voxels that are not informative on their own can contribute to a classification by de-noising or by removing the effects of global variations in unspecific internal states (Haufe et al., 2014; Yamashita et al., 2008). At first sight, this might appear counter-intuitive, but it can be demonstrated easily (see Figure 6). Consider one voxel that contains information about a label, but the signal in this voxel is contaminated by noise that is not related to the label (Figure 6, low information). This noise could have many sources, ranging from thermal noise in MRI recording to ongoing physiological background activity that is unrelated to the task (e.g., Raichle, 2010). Although this noise impairs the classification, it might sometimes be possible to regain high classification rates by subtracting the signal from a second voxel that purely reflects the same noise source but does not contain label-related information (Figure 6, pure noise). In this case, a linear classifier would assign a positive weight to the informative voxel and a negative weight to the noise voxel. But the weight at this second voxel would only reflect a de-noising process and not the presence of information. The plot on the far right in Figure 6 shows this from the viewpoint of a classifier. The voxel on the y axis carries no information about the two categories, and the voxel on the x axis has low information (note that the centroids for the classes only differ along the x axis, but not the y axis). Nonetheless, the activity in the noise voxel contributes to the decision boundary (Haufe et al., 2014; Yamashita et al., 2008).

Searchlight analyses (Figure 3E; Kriegeskorte et al., 2006) also yield maps that can cause confusion and need to be interpreted with care. Each point in a searchlight map depicts the accuracy with which mental contents can be decoded from local clusters of voxels surrounding that point. They depict the centers of informative voxel clusters, but not the informative voxels themselves. Within these clusters, information might be encoded in quite different ways, and this might or might not include the voxel at the searchlight center itself (Etzel et al., 2013). For example, an informative searchlight center could even reflect a single informative voxel somewhere in the searchlight, a phenomenon referred to as the "needle-in-the-haystack effect" (Viswanathan et al., 2012). The resulting maps also depend on processing options. For example, larger searchlights yield more extended, smoother maps (e.g., Chen et al., 2011).

Controlling for Nuisance Variables

Due to the increased sensitivity of multivariate analyses, a more detailed control for confounding factors is also necessary. In contrast to GLM analyses, classifiers can extract information even if the sign of an effect randomly varies across subjects (Todd et al., 2013). For example, in task-set decoding, a classifier might be able to exploit subject-by-subject differences in difficulty between tasks that average out in the mean (say one subject might find task 1 easier, whereas another subject might find task 2 easier; Todd et al., 2013). Thus, more elaborate controls are needed to avoid that decoding results merely reflect nuisance variables, such as difficulty or attention. Two solutions are to either regress out the nuisance variable (Todd et al., 2013; Woolgar et al., 2014), or to directly compare decoding for nuisance variables and for the cognitive factor of interest (Wisniewski et al., 2014).

Emerging Directions: Encoding, Reconstruction, and Computational Modeling

Multivariate pattern classification of fMRI signals has been applied in diverse fields of cognitive neuroscience (see Haxby et al., 2014; and Tong and Pratte, 2012 for recent reviews). The logic of classification has proven useful in studying generalization between conditions (Cichy et al., 2012), compositionality (Reddy et al., 2009; Reverberi et al., 2012), temporal buildup of information (Polyn et al., 2005; Soon et al., 2008), and information

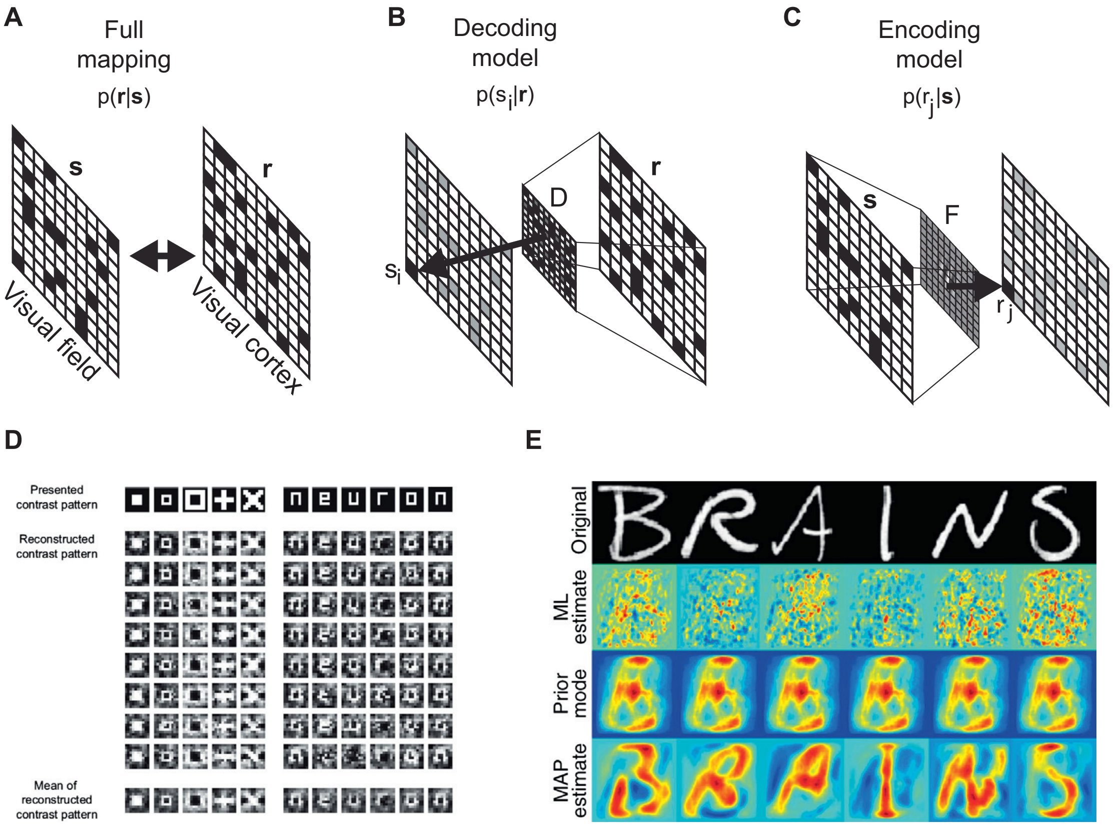

Figure 7. Encoding and Decoding Models in Reconstruction of Visual Images

(A) An ideal reconstruction would require obtaining the conditional probability p(**s**|**r**) of a visual image **s** given a brain response **r**. However, obtaining p(**s**|**r**) for a set of ten-by-ten black-and-white random images is not tractable because it would require measuring the brain activity to each of 2^{100} possible visual images.
(B) Decoding models simplify the mapping problem by decoding one image point s_i at a time, say by using a linear decoder D applied to the full set of voxels **r**.
(C) Encoding models simplify the full mapping by predicting the response in a single voxel r_j at a time based on a spatial filter F optimized for that voxel.
(D) Reconstruction of arbitrary symbols and letters on a ten-by-ten grid using an ensemble of decoding models for each position in the visual field (Miyawaki et al., 2008). The top row shows the presented pattern and beneath are eight reconstructions. The bottom row shows the mean across the reconstructions.
(E) Reconstruction of letters with encoding models and stimulus priors (taken from Schoenmakers et al., 2013). The maximum likelihood estimate (ML) is equivalent to the stimulus **s** with maximal probability p(**r**|**s**) (thus neglecting priors). The prior is the stimulus **s** with highest prior probability p(**s**). The ML and the prior can be combined to yield the maximum a posteriori estimate (MAP), p(**s**|**r**). The inclusion of the prior clearly improves the estimation.

flow between brain regions (Heinzle et al., 2011) and even between brains (Anders et al., 2011).

One limitation in many pattern-based fMRI studies is that only a small set of different cognitive states is considered, for example, which of several images a person is viewing (Haxby et al., 2001) or which of two intentions a person is holding (Haynes et al., 2007). The reason is that it is only possible to obtain brain responses for a limited number of cognitive states within the limited scanning time of a typical fMRI experiment. An important question is thus whether it might be possible to go beyond only a few alternatives and study the encoding of large numbers of cognitive states. Due to the limited scanning time, this would require generalizing from training data obtained with few classes to new cognitive states. One solution to this problem is to define a formal space in which the cognitive states occur. This could be a space defining a large set of possible images, or any other domain that can be suitably formalized in a model, such as, for example, sounds (Schönwiesner and Zatorre, 2009) or semantics (Mitchell et al., 2008).

One field that has pioneered this approach is visual image reconstruction (Thirion et al., 2006; Miyawaki et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011; Schoenmakers et al., 2013). Image reconstruction refers to the attempt at decoding arbitrary images (rather than a few known samples) from brain signals. To understand the scope of the challenge this poses, it can help to first dramatically simplify the problem and consider random images of ten-by-ten black-and-white squares (Figures 7A–7C, visual field). The task of reconstructing an arbitrary pattern of ten-by-ten black-and-white squares that a person is viewing might seem easy, but it constitutes a formidable challenge (Attneave, 1954). The number of such potential black-and-white images is 2^{100} (or $\approx 10^{30}$). Even if the brain response to each image could be measured in one second, it would take 10^{30} seconds (or 10^{13} times the age of the universe) to measure the brain response (Figures 7A–7C, visual cortex) to each stimulus once. Visual reconstruction thus faces the problem that it is impossible to measure the brain response associated with each possible image.

One way to solve this problem is to use decoding models (Miyawaki et al., 2008). In this case, a decoder is used to predict the brightness at a single local region of the visual field (and not the entire image space) from the ensemble of fMRI voxels in visual cortex (Figure 7B). The classifier is based on the full ensemble of fMRI voxels in visual cortex (Figure 7B). In the case of a linear classifier, each voxel would be assigned a weight in order to optimally predict a single region of the visual field. Here, the problem of knowing the brain response associated with *each* full ten-by-ten image is simplified by learning the brain response associated with each local image region at a time. This simplification makes the assumption that responses to each region of the visual field are independent, which is not the case at the level of single cells (Carandini and Heeger, 2011; Albright and Stoner, 2002). Such decoding models have been successfully applied to the reconstruction of ten-by-ten-pixel patterns of letters from signals in early visual cortex (Figure 7D; Miyawaki et al., 2008).

Another solution to reconstruction is to use *encoding* models, which use the inverse direction of inference. The idea is to simplify the problem of obtaining the full mapping of many voxels to many image pixels in the other direction, now by using models that predict the activity in a *single voxel* based on the image intensities at all locations in the visual field (Figure 7C). An encoding model is a filter that is applied to the entire visual image and tuned to optimally predict the fMRI response in a single voxel (e.g., Dumoulin and Wandell, 2008; Kay et al., 2008; Nevado et al., 2004; Thirion et al., 2006). Such a filter is also referred to as the population-receptive field (pRF) of a voxel (Dumoulin and Wandell, 2008), where the term "population" indicates that the tuning pertains to the summed effect of the population of neurons sampled by the voxel, not to individual neurons. The receptive fields of voxels in early visual cortex have been characterized as simple two-dimensional Gaussian filters (Dumoulin and Wandell, 2008; Thirion et al., 2006), difference of Gaussians (Zuiderbaan et al., 2012), or as standard multi-parameter Gabor filter banks (Kay et al., 2008). Voxel-wise encoding models have been extended to effects of color (Brouwer and Heeger, 2009), facial identity (Gratton et al., 2013), attention (Sprague and Serences, 2013), working memory (Sprague et al., 2014), numerosity (Harvey et al., 2013), semantics (Huth et al., 2012), and even to other modalities (Schönwiesner and Zatorre, 2009; Thomas et al., 2015). Encoding models also can incorporate nonlinear processing stages (Nishimoto et al., 2011).

As outlined above, encoding models predict the responses of *single* voxels. However, the task for a reconstruction algorithm is to solve the opposite inference: which image most likely led to an activation pattern observed across many voxels in visual cortex? For a reconstruction based on encoding models, a second step is needed in order to combine the many single-voxel models into a single inference. Thus, the encoding model needs to be inverted. Depending on the complexity of the encoding model, this can be done using simple matrix inversion (Brouwer and Heeger, 2009; Sprague and Serences, 2013) or Bayesian estimation (for detailed examples, see Thirion et al., 2006; Naselaris et al., 2009; and Schoenmakers et al., 2013). Bayesian estimation also makes it possible to extend encoding models and incorporate prior knowledge of stimulus probabilities, which can substantially improve reconstruction (Thirion et al., 2006; Naselaris et al., 2009; Schoenmakers et al., 2013; Figure 7E).

The importance of priors can be explained using Bayes' rule. Let p(**r**|**s**) be the conditional probability of observing a brain response **r** to stimulus **s**. This is also referred to as the likelihood. Ultimately, p(**r**|**s**) is the basic output of most neuroimaging studies, but with restricted sets of stimulation conditions. As mentioned above, estimating the full p(**r**|**s**) is not tractable because it would require measuring brain responses to all possible stimuli (but see Yarkoni et al., 2011). Let p(**s**) be the prior probability of a stimulus, which reflects how well one could guess the stimulus without any brain responses just based on the frequency of their occurrence. Let p(**r**) be the probability of a certain brain response, independent of the stimulus condition. What one wants to infer is p(**s**|**r**), the conditional probability of a visual image **s** given a brain response **r**. For limited data, this inverse inference is problematic (Poldrack, 2006); however, if the full p(**r**|**s**) is known, then p(**s**|**r**) can be obtained according to Bayes' rule as follows:

$$p(\mathbf{s}|\mathbf{r}) = p(\mathbf{r}|\mathbf{s})p(\mathbf{s})/p(\mathbf{r}). \qquad \text{(Equation 1)}$$

This can be transformed to

$$p(\mathbf{s}|\mathbf{r}) \propto p(\mathbf{r}|\mathbf{s})p(\mathbf{s}), \qquad \text{(Equation 2)}$$

because for a given brain response **r**, p(**r**) is a constant and can be ignored. This reveals that the probability of a stimulus **s** given a brain response **r** is proportional to the product of the likelihood p(**r**|**s**) and the prior p(**s**). One conventional approach to deciding which stimulus **s** might have caused the brain response **r** is to choose the stimulus that has the highest associated likelihood p(**r**|**s**). This is known as the maximum likelihood estimate (ML). However, as can be seen from Equation 2 above, also the prior plays an important role in determining the most probable cause. So a better approach is to choose the stimulus that maximizes p(**s**|**r**), which is referred to as the maximum a posteriori estimate (MAP). This is illustrated in Figure 7E that provides a reconstruction of handwritten letters by combining a brain-based ML with a stimulus prior to yield a MAP (Schoenmakers et al., 2013). The reconstruction based on brain activity alone (ML) is poor; but, after combination with the prior, the MAP reconstruction is very good (Figure 7E).

The encoding and decoding models mentioned above show that computational approaches can substantially improve decoding of cognitive states from brain activity. The encoding approach does so by directly formulating computational models for single-voxel responses. A related approach to studying the link between computational models and brain activity is representational similarity analysis (RSA; Figure 8; Kriegeskorte et al., 2008a; Kriegeskorte and Kievit, 2013). RSA is not primarily aimed at reconstructing cognitive states from brain activity. It provides a different approach for testing how well specific computational models fit with the distributed activity pattern in a given brain area. RSA achieves this by testing whether the similarity between the brain responses to different stimuli matches the similarity between these stimuli according to a person's perceptual judgements, or according to a specific computational model of representation.

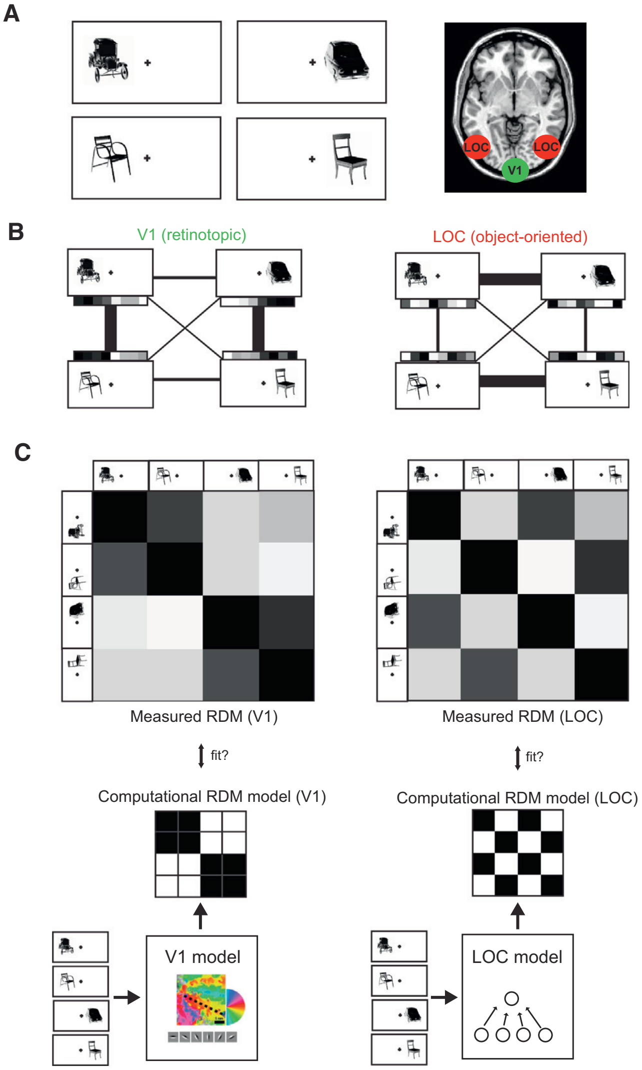

Figure 8. Representational Similarity Analysis

(A) This example shows four stimuli (adapted from Cichy et al., 2011), two with cars and two with chairs, one of each presented either to the left or the right of a central fixation cross.

(B) (Left) The brain response patterns in primary visual cortex (green) should largely reflect the retinotopic location of the stimuli. Thus, the correlation of the response patterns among the four different images (symbolically indicated by the thickness of the bars between each pair) should be highest for the stimuli that share the same retinotopic location (left/right of fixation) independent of their identity (car/chair). (Right) In contrast, in LOC (red), responses should largely reflect the identity of stimuli. In this case, the correlation should be higher between response patterns for images of the same objects, whereas the correlation between brain responses to different objects at the same location should be low.

(C) (Top) This can be expressed as a representational dissimilarity matrix (RDM; for details, see Kriegeskorte et al., 2008a). This matrix contains the dissimilarity in brain responses (1 – correlation) for each pairwise combination of stimuli. Here brighter regions indicate higher dissimilarity. The left shows a *measured* RDM in retinotopic visual cortex, the right a *measured* RDM for LOC. (Bottom) Computational models yield predictions for such RDMs. Here the four individual images are subjected either to a V1 model based on a retinotopically specific Gabor filter bank (left) or a computational model of location-invariant object representations in LOC (right; e.g., Riesenhuber and Poggio, 1999). For each model, this yields a *computational* RDM that can then be compared to the measured RDMs in each area (Kriegeskorte et al., 2008a).

For further details see Kriegeskorte et al. (2008a) and Kriegeskorte and Kievit (2013).

As a simple example, Figure 8 shows a basic similarity analysis for four stimuli, separately for brain responses in retinotopic visual cortex and in object-recognition region lateral occipital complex (LOC). The stimuli consist of one of four possible objects, two cars and two chairs, presented either to the left or right of fixation (Figure 8A). In early visual cortex, the responses are dominated by the retinotopic location of the stimuli (Figure 8B, left). For this reason, the brain responses in V1 are more similar and exhibit a higher correlation for the stimuli at the same location than for stimuli with the same identity but different location. In LOC (Figure 8B, right), the responses should largely reflect the identity of the stimuli. So the response patterns to images with the same objects should be more similar than those for different objects, and the retinotopic location should play only a minor role. This set of correlations can be shown more conveniently in a matrix form (Figure 8C, top row). By convention, typically the dissimilarity (1 – correlation) is plotted rather than the similarity. This is referred to as a representational dissimilarity matrix (RDM). These measured RDMs can then be compared to RDMs predicted by computational models (Kriegeskorte et al., 2008a). For example, the set of car and chair images could be subjected to a computational model, such as a Gabor filter bank model (Figure 8C, bottom left) or a location-invariant model of object recognition (Figure 8C, bottom right). The similarities between modeled brain responses for different images yield a computational RDM that can then be compared to the measured RDMs in different brain regions.

RSA provides another interesting constraint on testing the link between mental representations and brain activity patterns. Intuitively, one might assume that a brain region that allows one to best decode a perceptual feature (such as color) is the most likely area to provide a neural explanation for perception of this feature. However, Brouwer and Heeger (2009) showed that this need not be the case. They studied cortical responses to different color stimuli and found that the accuracy for decoding the color was highest in V1. However, an analysis of the neural representational space showed that V1 was incompatible with the perceptual space, despite yielding high accuracies, because the similarity between brain responses to different colors in V1 did not match their perceived similarity. In contrast, region V4 provided a closer match to perception, despite exhibiting smaller overall accuracies (cf. Figure 6 in Brouwer and Heeger, 2009). RSA also provides an important future direction for comparing coding spaces across different brain regions, or even between different modalities (Cichy et al., 2014) and different species (Kriegeskorte et al., 2008b).

Conclusions

Taken together, when handled with care, pattern-based analysis of fMRI patterns can help reveal how cognitive representations are encoded in human brain signals. This extends more conventional approaches to fMRI that typically have focused on mean activation levels in smoothed fMRI images (Friston et al., 1995b). The approach has to be applied carefully in order to avoid overfitting of the large parameter spaces involved. Caution also is required when interpreting the results of classification studies in terms of the information encoded in neural populations or in the tuning of single neurons. For the future, multivariate pattern analysis provides a generic framework for testing computational models with fMRI data, either in the form of encoding models or in combination with RSAs. Given the limitations of fMRI, a next important step needs to be the validation of classification results by direct comparison with recorded population measures, and the comparison of coding spaces across methods and species. Important additional contributions could come from optical imaging and from future developments in high-field MRI that might directly reveal the fine-grained representational topography of the human brain.

ACKNOWLEDGMENTS

The author thanks Carsten Allefeld, Carsten Bogler, Thomas Christophel, Radoslaw Cichy, Kai Görgen, Kerstin Ritter, Philipp Sterzer, and Martin Weygandt for valuable comments on the manuscript. This work was supported by the Bernstein Computational Neuroscience Program of the German Federal Ministry of Education and Research (grant reference 01GQ1001C); the Deutsche Forschungsgemeinschaft (DFG) grants SFB 940, Neurocure (Exc 257), KFO247, and German-Israeli Project Cooperation DIP (JA 945/3-1); and the European Regional Development Funds (grant references 10153458 and 10153460).

REFERENCES

Adams, D.L., Sincich, L.C., and Horton, J.C. (2007). Complete pattern of ocular dominance columns in human primary visual cortex. J. Neurosci. *27*, 10391–10403.

Albright, T.D., and Stoner, G.R. (2002). Contextual influences on visual processing. Annu. Rev. Neurosci. *25*, 339–379.

Alink, A., Krugliak, A., Walther, A., and Kriegeskorte, N. (2013). fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. Front. Psychol. *4*, 493.

Allefeld, C., and Haynes, J.D. (2014). Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. Neuroimage *89*, 345–357.

Allefeld, C., and Haynes, J.D. (2015). Multi-Voxel Pattern Analysis. In Brain Mapping: An Encyclopedic Reference, A.W. Toga, ed. (Waltham: Academic Press), pp. 641–646.

Anders, S., Heinzle, J., Weiskopf, N., Ethofer, T., and Haynes, J.D. (2011). Flow of affective information between communicating brains. Neuroimage *54*, 439–446.

Attneave, F. (1954). Some informational aspects of visual perception. Psychol. Rev. *61*, 183–193.

Bishop, C.M. (1995). Neural Networks for Pattern Recognition (Oxford: OUP).

Boynton, G.M. (2005). Imaging orientation selectivity: decoding conscious perception in V1. Nat. Neurosci. *8*, 541–542.

Braver, T.S., Cohen, J.D., Nystrom, L.E., Jonides, J., Smith, E.E., and Noll, D.C. (1997). A parametric study of prefrontal cortex involvement in human working memory. Neuroimage *5*, 49–62.

Brouwer, G.J., and Heeger, D.J. (2009). Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. *29*, 13992–14003.

Carandini, M., and Heeger, D.J. (2011). Normalization as a canonical neural computation. Nat. Rev. Neurosci. *13*, 51–62.

Chaimow, D., Yacoub, E., Ugurbil, K., and Shmuel, A. (2011). Modeling and analysis of mechanisms underlying fMRI-based decoding of information conveyed in cortical columns. Neuroimage *56*, 627–642.

Chen, Y., Namburi, P., Elliott, L.T., Heinzle, J., Soon, C.S., Chee, M.W., and Haynes, J.D. (2011). Cortical surface-based searchlight decoding. Neuroimage *56*, 582–592.

Chu, C., Ni, Y., Tan, G., Saunders, C.J., and Ashburner, J. (2011). Kernel regression for fMRI pattern prediction. Neuroimage *56*, 662–673.

Cichy, R.M., Chen, Y., and Haynes, J.D. (2011). Encoding the identity and location of objects in human LOC. Neuroimage *54*, 2297–2307.

Cichy, R.M., Heinzle, J., and Haynes, J.D. (2012). Imagery and perception share cortical representations of content and location. Cereb. Cortex *22*, 372–380.

Cichy, R.M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. Nat. Neurosci. *17*, 455–462.

Cox, D.D., and Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) ‘‘brain reading’’: detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage *19*, 261–270.

De Baene, W., and Vogels, R. (2010). Effects of adaptation on the stimulus selectivity of macaque inferior temporal spiking activity and local field potentials. Cereb. Cortex *20*, 2145–2165.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., and Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. Neuroimage *43*, 44–58.

Downing, P.E., Chan, A.W., Peelen, M.V., Dodds, C.M., and Kanwisher, N. (2006). Domain specificity in visual cortex. Cereb. Cortex *16*, 1453–1461.

Duda, R.O., Hart, P.E., and Stork, D.G. (2000). Pattern Classification, Second Edition (New York: Wiley).

Dumoulin, S.O., and Wandell, B.A. (2008). Population receptive field estimates in human visual cortex. Neuroimage *39*, 647–660.

Etzel, J.A., Zacks, J.M., and Braver, T.S. (2013). Searchlight analysis: promise, pitfalls, and potential. Neuroimage *78*, 261–269.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. *7*, 179–188.

Freeman, J., Brouwer, G.J., Heeger, D.J., and Merriam, E.P. (2011). Orientation decoding depends on maps, not columns. J. Neurosci. *31*, 4792–4804.

Freeman, J., Heeger, D.J., and Merriam, E.P. (2013). Coarse-scale biases for spirals and orientation in human visual cortex. J. Neurosci. *33*, 19695–19703.

Friston, K.J., Frith, C.D., Frackowiak, R.S., and Turner, R. (1995a). Characterizing dynamic brain responses with fMRI: a multivariate approach. Neuroimage *2*, 166–172.

Friston, K.J., Holmes, A.P., Poline, J.B., Grasby, P.J., Williams, S.C., Frackowiak, R.S., and Turner, R. (1995b). Analysis of fMRI time-series revisited. Neuroimage *2*, 45–53.

Furmanski, C.S., and Engel, S.A. (2000). An oblique effect in human primary visual cortex. Nat. Neurosci. *3*, 535–536.

Gardner, J.L. (2010). Is cortical vasculature functionally organized? Neuroimage *49*, 1953–1956.

Goldfine, A.M., Bardin, J.C., Noirhomme, Q., Fins, J.J., Schiff, N.D., and Victor, J.D. (2013). Reanalysis of ‘‘Bedside detection of awareness in the vegetative state: a cohort study’’. Lancet *381*, 289–291.

Gratton, C., Sreenivasan, K.K., Silver, M.A., and D’Esposito, M. (2013). Attention selectively modifies the representation of individual faces in the human brain. J. Neurosci. *33*, 6979–6989.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., and Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron *24*, 187–203.

Hackmack, K., Paul, F., Weygandt, M., Allefeld, C., and Haynes, J.D.; Alzheimer’s Disease Neuroimaging Initiative (2012). Multi-scale classification of disease using structural MRI and wavelet transform. Neuroimage *62*, 48–58.

Harvey, B.M., Klein, B.P., Petridou, N., and Dumoulin, S.O. (2013). Topographic representation of numerosity in the human parietal cortex. Science *341*, 1123–1126.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage *87*, 96–110.

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

Haxby, J.V., Connolly, A.C., and Guntupalli, J.S. (2014). Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. *37*, 435–456.

Haynes, J.D., and Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. *8*, 686–691.

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

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R.E. (2007). Reading hidden intentions in the human brain. Curr. Biol. *17*, 323–328.

Heinzle, J., Kahnt, T., and Haynes, J.D. (2011). Topographically specific functional connectivity between visual field maps in the human brain. Neuroimage *56*, 1426–1436.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.

Ioannidis, J.P., Ntzani, E.E., Trikalinos, T.A., and Contopoulos-Ioannidis, D.G. (2001). Replication validity of genetic association studies. Nat. Genet. *29*, 306–309.

Josephs, O., and Henson, R.N. (1999). Event-related functional magnetic resonance imaging: modelling, inference and optimization. Philos. Trans. R. Soc. Lond. B Biol. Sci. *354*, 1215–1228.

Kahnt, T., Heinzle, J., Park, S.Q., and Haynes, J.D. (2010). The neural code of reward anticipation in human orbitofrontal cortex. Proc. Natl. Acad. Sci. USA *107*, 6010–6015.

Kamitani, Y., and Sawahata, Y. (2010). Spatial smoothing hurts localization but not information: pitfalls for brain mappers. Neuroimage *49*, 1949–1952.

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

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

Kohonen, T. (1989). Self-Organization and Asssociative Memory (Berlin: Springer).

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. Front. Neurosci. *3*, 363–373.

Kriegeskorte, N. (2011). Pattern-information analysis: from stimulus decoding to computational-model testing. Neuroimage *56*, 411–421.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron *60*, 1126–1141.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., and Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. *12*, 535–540.

Kriegeskorte, N., Cusack, R., and Bandettini, P. (2010). How does an fMRI voxel sample the neuronal activity pattern: compact-kernel or complex spatiotemporal filter? Neuroimage *49*, 1965–1976.

Lee, S.H., Kravitz, D.J., and Baker, C.I. (2013). Goal-dependent dissociation of visual and prefrontal cortices during working memory. Nat. Neurosci. *16*, 997–999.

Liu, T.T., Frank, L.R., Wong, E.C., and Buxton, R.B. (2001). Detection power, estimation efficiency, and predictability in event-related fMRI. Neuroimage *13*, 759–773.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. Nature *453*, 869–878.

Logothetis, N.K., and Wandell, B.A. (2004). Interpreting the BOLD signal. Annu. Rev. Physiol. *66*, 735–769.

Marquand, A.F., Brammer, M., Williams, S.C., and Doyle, O.M. (2014). Bayesian multi-task learning for decoding multi-subject neuroimaging data. Neuroimage *92*, 298–311.

Mitchell, T.M., Hutchinson, R., Just, M.A., Niculescu, R.S., Pereira, F., and Wang, X. (2003). Classifying instantaneous cognitive states from FMRI data. AMIA Annu. Symp. Proc. *2003*, 465–469.

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

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.A., Morito, Y., Tanabe, H.C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron *60*, 915–929.

Mountcastle, V.B. (1957). Modality and topographic properties of single neurons of cat's somatic sensory cortex. J. Neurophysiol. *20*, 408–434.

Mourão-Miranda, J., Bokde, A.L., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage *28*, 980–995.

Mourão-Miranda, J., Friston, K.J., and Brammer, M. (2007). Dynamic discrimination analysis: a spatial-temporal SVM. Neuroimage *36*, 88–99.

Mumford, J.A., Davis, T., and Poldrack, R.A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. Neuroimage *103*, 130–138.

Mur, M., Bandettini, P.A., and Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. Soc. Cogn. Affect. Neurosci. *4*, 101–109.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron *63*, 902–915.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. Neuroimage *56*, 400–410.

Nevado, A., Young, M.P., and Panzeri, S. (2004). Functional imaging and neural information coding. Neuroimage *21*, 1083–1095.

Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. *15*, 1–25.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS Comput. Biol. *10*, e1003553.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. *21*, 1641–1646.

Noirhomme, Q., Lesenfants, D., Gomez, F., Soddu, A., Schrouff, J., Garraux, G., Luxen, A., Phillips, C., and Laureys, S. (2014). Biased binomial assessment of cross-validated estimation of classification accuracies illustrated in diagnosis predictions. Neuroimage Clin. *4*, 687–694.

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

O'Craven, K.M., Downing, P.E., and Kanwisher, N. (1999). fMRI evidence for objects as the units of attentional selection. Nature *401*, 584–587.

Ogawa, S., Lee, T.M., Kay, A.R., and Tank, D.W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proc. Natl. Acad. Sci. USA *87*, 9868–9872.

Oosterhof, N.N., Wiggett, A.J., Diedrichsen, J., Tipper, S.P., and Downing, P.E. (2010). Surface-based information mapping reveals crossmodal vision-action representations in human parietal and occipitotemporal cortex. J. Neurophysiol. *104*, 1077–1089.

Op de Beeck, H.P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? Neuroimage *49*, 1943–1948.

Pereira, F., and Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. Neuroimage *56*, 476–496.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage *45* (1, Suppl), S199–S209.

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

Polyn, S.M., Natu, V.S., Cohen, J.D., and Norman, K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. Science *310*, 1963–1966.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. Nat. Rev. Neurosci. *1*, 125–132.

Raichle, M.E. (2010). Two views of brain function. Trends Cogn. Sci. *14*, 180–190.

Ramírez, F.M., Cichy, R.M., Allefeld, C., and Haynes, J.D. (2014). The neural code for face orientation in the human fusiform face area. J. Neurosci. *34*, 12155–12167.

Reddy, L., Kanwisher, N.G., and VanRullen, R. (2009). Attention and biased competition in multi-voxel object representations. Proc. Natl. Acad. Sci. USA *106*, 21447–21452.

Reverberi, C., Görgen, K., and Haynes, J.D. (2012). Compositionality of rule representations in human prefrontal cortex. Cereb. Cortex *22*, 1237–1246.

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

Sasaki, Y., Rajimehr, R., Kim, B.W., Ekstrom, L.B., Vanduffel, W., and Tootell, R.B. (2006). The radial bias: a different slant on visual orientation sensitivity in human and nonhuman primates. Neuron *51*, 661–670.

Schoenmakers, S., Barth, M., Heskes, T., and van Gerven, M. (2013). Linear reconstruction of perceived images from human brain activity. Neuroimage *83*, 951–961.

Schönwiesner, M., and Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA *106*, 14611–14616.

Schreiber, K., and Krekelberg, B. (2013). The statistical analysis of multi-voxel patterns in functional imaging. PLoS ONE *8*, e69328.

Scott, D.W. (1992). Multivariate density estimation: Theory, practice, visualization (New York: Wiley).

Serences, J.T., and Saproo, S. (2012). Computational advances towards linking BOLD and behavior. Neuropsychologia *50*, 435–446.

Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science *268*, 889–893.

Shmuel, A., Chaimow, D., Raddatz, G., Ugurbil, K., and Yacoub, E. (2010). Mechanisms underlying decoding at 7 T: ocular dominance columns, broad structures, and macroscopic blood vessels in V1 convey information on the stimulated eye. Neuroimage *49*, 1957–1964.

Smola, A.J., and Schölkopf, B. (2004). A tutorial on support vector regression. Stat. Comput. *14*, 199–222.

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

Sprague, T.C., and Serences, J.T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. Nat. Neurosci. *16*, 1879–1887.

Sprague, T.C., Ester, E.F., and Serences, J.T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. Curr. Biol. *24*, 2174–2180.

Stelzer, J., Chen, Y., and Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. Neuroimage *65*, 69–82.

Swisher, J.D., Gatenby, J.C., Gore, J.C., Wolfe, B.A., Moon, C.H., Kim, S.G., and Tong, F. (2010). Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. J. Neurosci. *30*, 325–330.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage *33*, 1104–1116.

Thomas, J.M., Huber, E., Stecker, G.C., Boynton, G.M., Saenz, M., and Fine, I. (2015). Population receptive field estimates of human auditory cortex. Neuroimage *105*, 428–439.

Todd, M.T., Nystrom, L.E., and Cohen, J.D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. Neuroimage *77*, 157–165.

Tong, F., and Pratte, M.S. (2012). Decoding patterns of human brain activity. Annu. Rev. Psychol. *63*, 483–509.

Tononi, G., Srinivasan, R., Russell, D.P., and Edelman, G.M. (1998). Investigating neural correlates of conscious perception by frequency-tagged neuromagnetic responses. Proc. Natl. Acad. Sci. USA *95*, 3198–3203.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage *15*, 273–289.

Viswanathan, S., Cieslak, M., and Grafton, S.T. (2012). On the geometric structure of fMRI searchlight-based information maps. arXiv:1210.6317.

Wang, H.X., Merriam, E.P., Freeman, J., and Heeger, D.J. (2014). Motion direction biases and decoding in human visual cortex. J. Neurosci. *34*, 12601–12615.

Wisniewski, D., Reverberi, C., Tusche, A., and Haynes, J.-D. (2014). The neural representation of voluntary task-set selection in dynamic environments. Cereb. Cortex. http://dx.doi.org/10.1093/cercor/bhu155.

Woolgar, A., Golland, P., and Bode, S. (2014). Coping with confounds in multivoxel pattern analysis: what should we do about reaction time differences? A comment on Todd, Nystrom & Cohen 2013. Neuroimage *98*, 506–512.

Yacoub, E., Harel, N., and Ugurbil, K. (2008). High-field fMRI unveils orientation columns in humans. Proc. Natl. Acad. Sci. USA *105*, 10607–10612.

Yamashita, O., Sato, M.A., Yoshioka, T., Tong, F., and Kamitani, Y. (2008). Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. Neuroimage *42*, 1414–1429.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., Van Essen, D.C., and Wager, T.D. (2011). Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods *8*, 665–670.

Zimmermann, J., Goebel, R., De Martino, F., van de Moortele, P.F., Feinberg, D., Adriany, G., Chaimow, D., Shmuel, A., Uğurbil, K., and Yacoub, E. (2011). Mapping the organization of axis of motion selective features in human area MT using high-field fMRI. PLoS ONE *6*, e28716.

Zuiderbaan, W., Harvey, B.M., and Dumoulin, S.O. (2012). Modeling center-surround configurations in population receptive fields using fMRI. J. Vis. *12*, 10.