Decoding Neural Representational Spaces Using Multivariate Pattern Analysis

James V. Haxby,[1,2] Andrew C. Connolly,[1] and J. Swaroop Guntupalli[1]

[1]Department of Psychological and Brain Sciences, Center for Cognitive Neuroscience, Dartmouth College, Hanover, New Hampshire 03755; email: james.v.haxby@dartmouth.edu, andrew.c.connolly@dartmouth.edu, swaroopgj@gmail.com

[2]Center for Mind/Brain Sciences (CIMeC), University of Trento, Rovereto, Trentino 38068, Italy

Annu. Rev. Neurosci. 2014. 37:435–56

First published online as a Review in Advance on June 25, 2014

The *Annual Review of Neuroscience* is online at neuro.annualreviews.org

This article's doi: 10.1146/annurev-neuro-062012-170325

Keywords

neural decoding, MVPA, RSA, hyperalignment, population response, fMRI

Abstract

A major challenge for systems neuroscience is to break the neural code. Computational algorithms for encoding information into neural activity and extracting information from measured activity afford understanding of how percepts, memories, thought, and knowledge are represented in patterns of brain activity. The past decade and a half has seen significant advances in the development of methods for decoding human neural activity, such as multivariate pattern classification, representational similarity analysis, hyperalignment, and stimulus-model-based encoding and decoding. This article reviews these advances and integrates neural decoding methods into a common framework organized around the concept of high-dimensional representational spaces.

Contents

INTRODUCTION

Information is encoded in patterns of neural activity. This information can come from our experience of the world or can be generated by thinking. One of the great challenges for systems neuroscience is to break this code. Developing algorithms for decoding neural activity involves many modalities of measurement—including single-unit recording, electrocorticography (ECoG), electro- and magnetoencephalography (EEG and MEG), and functional magnetic resonance imaging (fMRI)—in various species. All decoding methods are multivariate analyses of brain activity patterns that are distributed across neurons or cortical regions. These methods are referred to generally as multivariate pattern analysis (MVPA). This review focuses on the progress made in the past decade and a half in the development of methods for decoding human neural activity as measured with fMRI. We make occasional references to decoding analyses of single-unit recording data in monkeys and of ECoG and MEG data in humans to illustrate the general utility of decoding methods and to indicate the potential for multimodal decoding.

Prior to the discovery that within-area patterns of response in fMRI carried information that afforded decoding of stimulus distinctions (Haxby et al. 2001, Cox & Savoy 2003, Haxby 2012), it was generally believed that the spatial resolution of fMRI allowed investigators to ask only which task or stimulus activated a region globally. Thus, fMRI studies focused on associating brain regions with functions. A region's function was identified by determining which task activated it most strongly. The introduction of decoding using MVPA has revolutionized fMRI research by changing the questions that are asked. Instead of asking what a region's function is, in terms of a single brain state associated with global activity, fMRI investigators can now ask what information is represented in a region, in terms of brain states associated with distinct patterns of activity, and how that information is encoded and organized.

Multivariate pattern (MVP) classification distinguishes patterns of neural activity associated with different stimuli or cognitive states. The first demonstrations of MVP classification showed that different high-level visual stimulus categories (faces, animals, and objects) were associated with distinct patterns of brain activity in the ventral object vision pathway (Haxby et al. 2001, Cox & Savoy 2003). Subsequent work has shown that MVP classification can also distinguish many other brain states, for example low-level visual features in the early visual cortex (Haynes & Rees 2005, Kamitani & Tong 2005) and auditory stimuli in the auditory cortex (Formisano et al. 2008,

MEG: magnetoencephalography

fMRI: functional magnetic resonance imaging

Multivariate pattern analysis (MVPA): analysis of brain activity patterns with methods such as pattern classification, RSA, hyperalignment, or stimulus-model-based encoding and decoding

REPRESENTATIONAL SPACE

Representational space is a high-dimensional space in which each neural response or stimulus is expressed as a vector with different values for each dimension. In a neural representational space, each pattern feature is a measure of local activity, such as a voxel or a single neuron. In a stimulus representational space, each feature is a stimulus attribute, such as a physical attribute or semantic label.

Staeren et al. 2009), as well as more abstract brain states such as intentions (Haynes et al. 2007, Soon et al. 2008) and the contents of working memory (Harrison & Tong 2009).

Whereas MVP classification simply demonstrates reliable distinctions among brain states, more recently introduced methods characterize how these brain states are organized. Representational similarity analysis (RSA) (Kriegeskorte et al. 2008a) analyzes the geometry of representations in terms of the similarities among brain states. RSA can show that the representations of the same set of stimuli in two brain regions have a different structure (Kriegeskorte et al. 2008a,b; Connolly et al. 2012a,b), whereas MVP classification may find that the classification accuracy is equivalent in those regions. Stimulus-model-based encoding and decoding algorithms show that brain activity patterns can be related to the constituent features of stimuli or cognitive states. This innovation affords predictions of patterns of brain response to novel stimuli based on their features (Kay et al. 2008, Mitchell et al. 2008). It also affords reconstruction of stimuli from brain activity patterns based on predictions of the stimulus features (Miyawaki et al. 2008, Naselaris et al. 2009, Nishimoto et al. 2011, Horikawa et al. 2013).

Several excellent reviews have focused on MVP classification (Norman et al. 2006, Haynes & Rees 2006, O'Toole et al. 2007, Pereira et al. 2009, Tong & Pratte 2012), RSA (Kriegeskorte & Kievet 2013), or stimulus-model-based encoding and decoding (Naselaris et al. 2011). Here we integrate neural decoding methods into a common framework organized around the concept of high-dimensional representational spaces (see sidebar). In all these methods, brain activity patterns are analyzed as vectors in high-dimensional representational spaces. Neural decoding then analyzes these spaces in terms of (*a*) reliably distinctive locations of pattern response vectors (MVP classification), (*b*) the proximity of these vectors to each other (RSA), or (*c*) mapping of vectors from one representational space to another—from one subject's neural representational space to a model space that is common across subjects (hyperalignment) or from stimulus feature spaces to neural spaces (stimulus-model-based encoding).

Representational similarity analysis (RSA): analysis of the pattern of similarities among response vectors

Response vector: a brain activity pattern expressed as the response strengths for features of that pattern, e.g., voxels, single neurons, or model dimensions

Hyperalignment: transformation of individual representational spaces into a model representational space in which each dimension has a common tuning function

CORE CONCEPT: REPRESENTATIONAL SPACES

The core concept that underlies neural decoding and encoding analyses is that of high-dimensional representational vector spaces. Neural responses—brain activity patterns—are analyzed as vectors in a neural representational space. Brain activity patterns are distributed in space and time. The elements, or features, of these patterns are local measures of activity, and each of these local measures is a dimension in the representational space. Thus, if neural responses measured with fMRI have 1,000 voxels, the representational space is 1,000-dimensional. If a population response has spike rates for 600 cells, the representational space is 600-dimensional. If fMRI responses with 1,000 voxels include six time points, the response vectors are analyzed in a 6,000-dimensional space.

For fMRI, measures of local activity are usually voxels (volume elements in brain images), but there are numerous alternatives, such as nodes on the cortical surface, the average signal for an area, a principal or independent component, or a measure of functional connectivity between a pair

of locations. For single-unit recording, local measures can be single-neuron spike rates, multiple-unit spike rates, or local field potentials, among other possibilities. Similarly, EEG and MEG responses can be analyzed as time-varying patterns of activity distributed over sensors or sources, with numerous possibilities for converting activity into frequencies, principal or independent components, or measures of synchrony between sources.

The computational advantages of representational vector spaces extend beyond neural representational spaces to representational spaces for stimuli or cognitive states. For example, a visual stimulus can be modeled as a set of features based on response properties of neurons in V1, as higher-order visual features, or as a set of semantic labels. Sounds—voices and music—can be modeled as sets of acoustic features, words can be modeled as sets of semantic features, actions as sets of movement and goal features, etc. Once the description of the stimulus is in a representational space, various computational manipulations can be applied for relating stimulus representational spaces to neural representational spaces.

All the major varieties of neural decoding and encoding analyses follow from this conversion of patterns of brain activity or stimuli to single points in high-dimensional representational vector spaces. MVP classification uses machine learning methods to define decision boundaries in a neural representational space that best distinguish a set of response vectors for one brain state from others. RSA analyzes the similarity between response vectors as distances in the representational space. Stimulus-model-based encoding predicts the location of the neural response vector for a new stimulus on the basis of the coordinates of that stimulus in a stimulus feature space. Stimulus-model-based decoding tests whether the encoding-based prediction allows correct classification of neural response vectors to new stimuli. Building a model of a neural representational space that is common across brains requires hyperalignment to rotate the coordinate axes of individual representational spaces to minimize the difference in the locations of response vectors for the same stimuli. Thus, stimuli and other cognitive events are represented as vectors in neural representational spaces as well as in stimulus representational spaces, and the computational task for understanding representation becomes one of characterizing the geometries within spaces and relating the geometries of these spaces to each other.

Numerically, a set of response vectors in a representational space is a matrix in which each column is a local pattern feature (e.g., voxel) and each row is a response vector (**Figure 1**). The values in each column reflect the differential responses of that pattern feature to conditions or stimuli. This profile of differential responses is called the tuning function. All the neural decoding and encoding methods can be understood in terms of analyzing or manipulating the geometry of the response vectors in a high-dimensional space. Computationally, these analyses and manipulations are performed using linear algebra. Here we illustrate the concepts related to high-dimensional representational spaces in two-dimensional figures by showing only two dimensions at a time—the equivalent of a two-voxel brain or a two-neuron population. The linear algebra for these two-dimensional toy examples is the same as the linear algebra for representational spaces with many more dimensions and larger matrices. Most of the algorithms that we discuss here can be implemented using PyMVPA (**http://www.pymvpa.org**; Hanke et al. 2009), a Python-based software platform that includes tutorials and sample data sets.

The geometries in a neural representational space, as defined here, are distinctly different from the geometries of cortical anatomy and of cortical topographies. A cortical topography can be thought of as a two-dimensional manifold. Neural encoding can be thought of as the problem of projecting high-dimensional representations into this low-dimensional topography. Decoding is the problem of projecting a neural response in a two-dimensional topography into a high-dimensional representational space. The methods that we discuss here are computational algorithms that attempt to model these transformations.

Machine learning: a branch of artificial intelligence that builds and evaluates induction algorithms that learn data patterns and associate them with labels

Pattern feature: a single element in a distributed pattern, such as a voxel, a single neuron, or a model space dimension

Tuning function: the profile of differential responses to stimuli or brain states for a single pattern feature

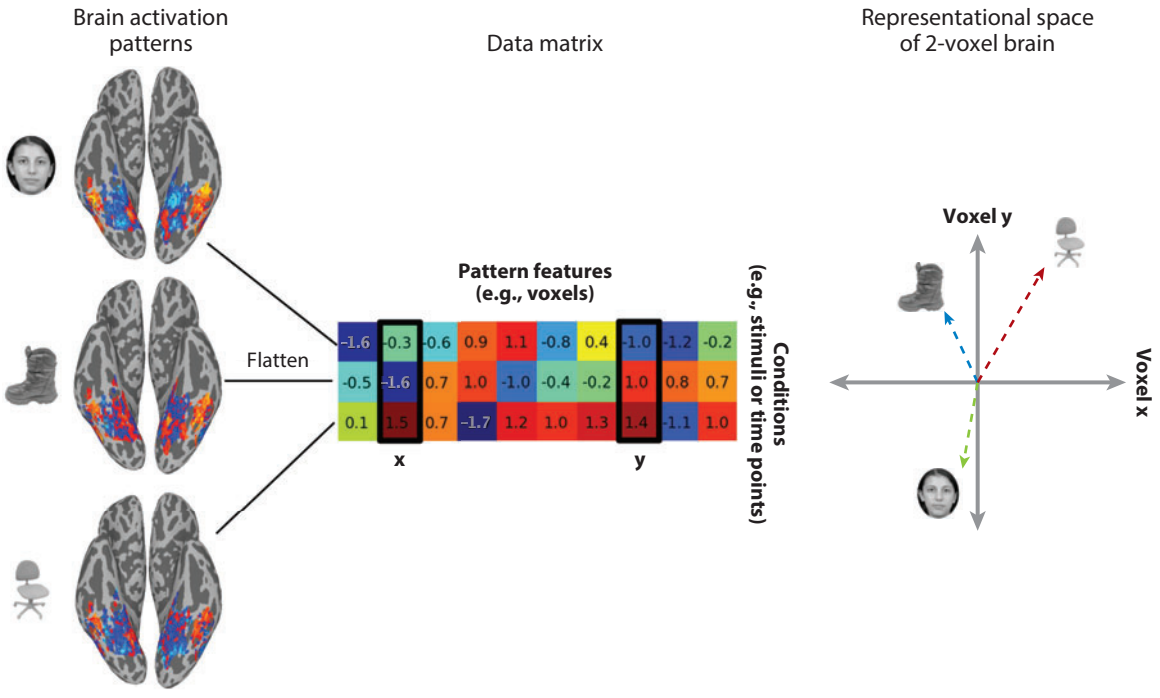


Figure 1

Multivariate pattern analysis (MVPA) is a family of methods that treats the measured fMRI signal as a set of pattern vectors stored in an N × M matrix with *N* observations (e.g., stimulus conditions, time points) and *M* features (e.g., voxels, cortical surface nodes) define an *M*-dimensional vector space. The goal of MVPA analyses is to analyze the structure of these high-dimensional representational spaces.

In a neural representational space, brain responses are vectorized, thus discarding the spatial relationships among cortical locations and the temporal relationships among time points. Thus, the approaches that we present here do not attempt to model the spatial structure of cortical topographies or how high-dimensional functional representations are packed into these topographies. An approach to modeling cortical topographies based on the principle of spatial continuity of function can be found in work by Aflalo & Graziano (2006, 2011) and Graziano & Aflalo (2007a,b). Anatomy and, in particular, cortical topography are important aspects of neural representation, of course. Although decoding methods discard anatomical and topographic information when brain responses are analyzed in high-dimensional representational spaces, the anatomical location of a representational space can be investigated using searchlight analyses (Kriegeskorte et al. 2006, Chen et al. 2011, Oosterhof et al. 2011), and the topographic organization of that representation can be recovered by projecting response vectors and linear discriminants from a common model representational space into individual subjects' topographies (Haxby et al. 2011).

MULTIVARIATE PATTERN CLASSIFICATION

MVP classification uses machine learning algorithms to classify response patterns, associating each neural response with an experimental condition. Pattern classification involves defining sectors in the neural representational space in which all response vectors represent the same class of information, such as a stimulus category (e.g., Haxby et al. 2001, Cox & Savoy 2003), an attended stimulus (e.g., Kamitani & Tong 2005), or a cognitive state (e.g., Haynes et al. 2007).

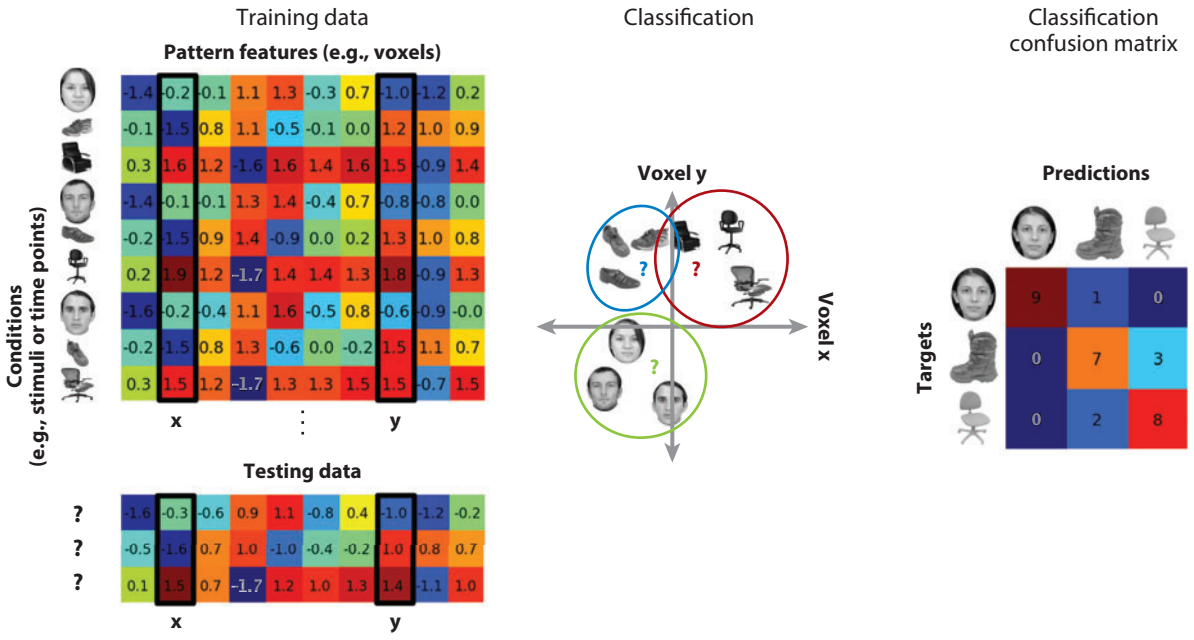


Figure 2

MVP classification analyses involve partitioning data matrices into different sets for training and testing a pattern classifier. A classifier is trained to designate sectors of the vector space to the labels provided for the samples in the training set. Test samples are then classified as belonging to the labeled class associated with the sector in which they reside. Classification accuracy is measured as the proportion of predicted labels that match the actual label (target) for each test item. A confusion matrix provides information about the patterns of correct classifications (on the diagonal) and misclassifications (off diagonal).

An MVP classification analysis begins with dividing the data into independent training and test data sets (**Figure 2**). The decision rules that determine the confines of each class of neural response vectors are developed on training data. The border between sectors for different conditions is called a decision surface. The validity of the classifier is then tested on the independent test data. For valid generalization testing, the test data must play no role in the development of the classifier, including data preprocessing (Kriegeskorte et al. 2009). Each test data response vector is then classified as another exemplar of the condition associated with the sector in which it is located.

Classifier accuracy is the percentage of test vectors that are correctly classified. A more revealing assessment of classifier performance is afforded by examining the confusion matrix. A confusion matrix presents the frequencies for all classifications of each experimental condition, including the details about misclassifications. Examination of misclassifications adds information about which conditions are most distinct and which are more similar. This information is analyzed using additional methods in RSA (see next section). Examination of classification accuracy for each condition separately can alert the investigator to whether average accuracy is really dependent on a small number of conditions, rather than an accurate reflection of performance across all or most conditions. Thus, average classification accuracy is a useful metric but discards information that can be discovered by examining the classification confusion matrix. Confusion matrices are shown in **Figure 3** for two category perception experiments (Haxby et al. 2011; Connolly et al. 2012a,b). From the first experiment, on the perception of faces and objects, the confusion matrix reveals that if misclassified, faces are classified as other faces and objects as other objects. Moreover, the classifier cannot distinguish female from male faces. From the second experiment, on the

Test data: the data set used to test the validity of a decision rule that was derived on training data

Training data: the portion of a data set that is used to derive the decision rule for pattern classification

Decision surface: a surface that defines the boundary between sectors in a representational space and is used to classify vectors

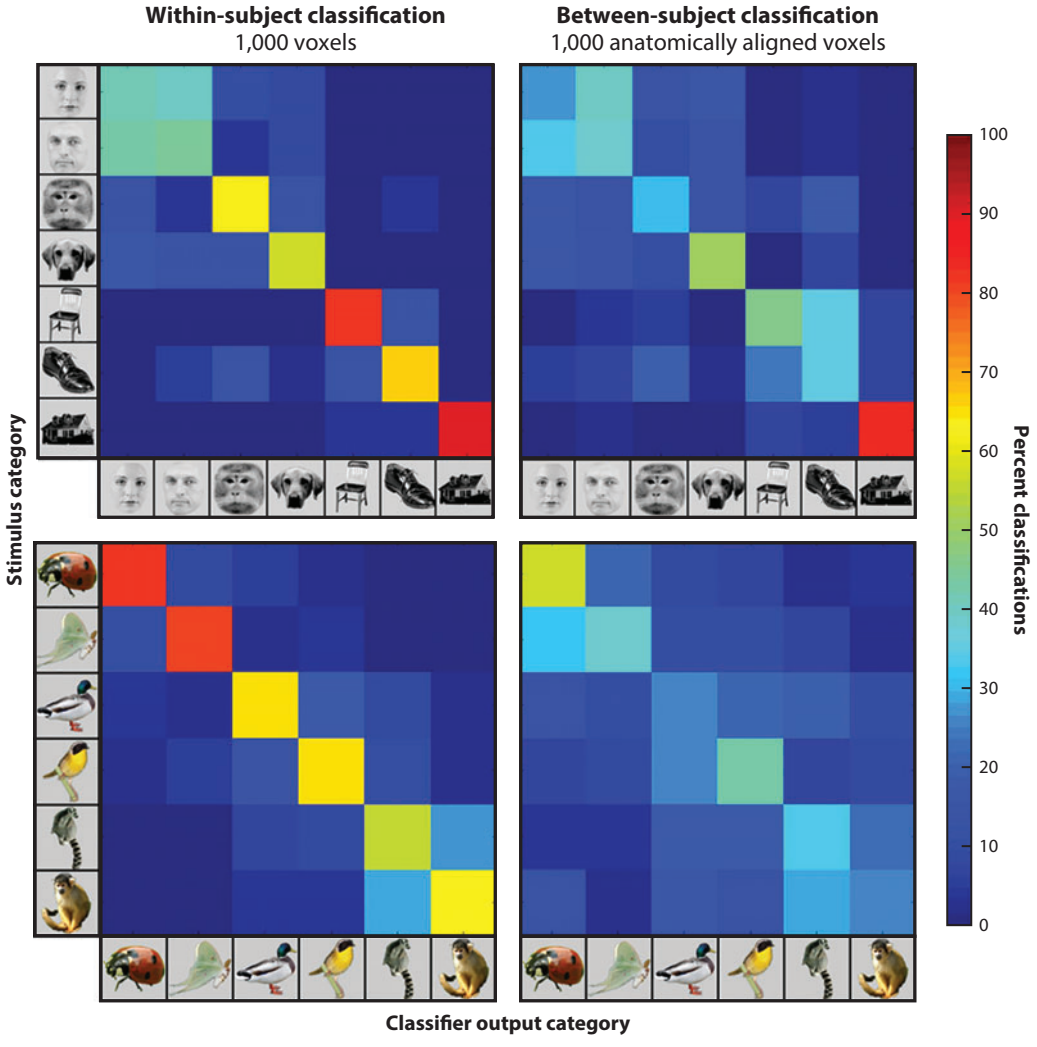


Figure 3

Confusion matrices for two experiments that measured responses to visual objects in human ventral temporal (VT) cortex. The patterns of misclassifications show that when items are misclassified they are more likely to be confused with items from the same superordinate category: faces and small objects (*top*); and primates, birds, and bugs (*bottom*). Classification performed on data matrices from the same subject (i.e., the same set of features) produces higher overall accuracies (*a*) than does between-subject classification (BSC) (*b*) where the features (voxels) have been aligned on the basis of a standard anatomical template.

perception of animals, the confusion matrix reveals that misclassifications are usually within animal class (primates, birds, and insects) and rarely across classes.

MVP classification uses machine learning classifiers to develop the decision rules. In general, different classifiers produce similar results, and some classifiers tend to perform better than others. A seminal MVP classification study (Haxby et al. 2001) used a one-nearest-neighbor classifier that classified a test response vector as the category for the training data vector that was closest in the neural representational space. Distance between vectors was measured using correlation, which is the cosine of the angle between mean-centered vectors. Because a single vector was used for each

class in the training data—the mean pattern for each class in half the data—the decision surfaces were linear hyperplanes separating each pair of unique classes. Nearest-neighbor methods are fast and conceptually clear, but most have found that other classifiers provide slightly higher accuracies. Cox & Savoy (2003) were the first to use support vector machine (SVM) (Cortez & Vapnik 1995) classifiers for fMRI. SVM classifiers fine-tune the position of the decision surface on the basis of the vectors that are closest to the surface, i.e., the support vectors, by maximizing the distances from the surface to these borderline cases. Other regression-based methods, such as linear discriminant analysis (LDA) (e.g., Carlson et al. 2003, O'Toole et al. 2005), are also effective and can include regularization methods for selecting features, such as sparse multinomial logistic regression (SMLR) (Yamashita et al. 2008). Most MVP classification analyses have used linear classifiers—meaning that the decision surface is planar—some for theoretically driven reasons (Kamitani & Tong 2005) but mostly for simplicity and to avoid overfitting the noise in the training data, which leads to larger performance decrements in generalization testing.

Between-subject classification (BSC): classification of a subject's response vectors based on a classifier built on other subjects' data

Within-subject classification (WSC): classification of a subject's response vectors based on a classifier built on that subject's own data

Dissimilarity matrix (DSM): the set of all pairwise dissimilarities between response vectors

Until recently, almost all MVP classification analyses had built a new classifier for each individual brain. Cox & Savoy (2003) showed that classifier performance dropped drastically if based on other subjects' data. The performance decrement for between-subject classification (BSC) relative to within-subject classification (WSC) shows that the structure of activity patterns differs across subjects. This variance could be due to the inadequacy of methods for aligning cortical topographies based on anatomical features. Some of the more successful BSC analyses are of large-scale patterns that involve many, widely distributed areas (Shinkareva et al. 2008, 2011), suggesting that larger-scale topographies may be aligned adequately based on anatomy and that poor BSC performance occurs when distinctions are found in finer-scale topographies. Low BSC accuracies could also be due to idiosyncratic neural codes. **Figure 3** shows the confusion matrices for both WSC and BSC of two category-perception experiments, illustrating the severity of the problem (Haxby et al. 2011). Accuracies dropped from 63% to 45% for 7 categories of faces and objects and from 69% to 37% for 6 animal species. A recently developed method for aligning the neural representational spaces across brains—hyperalignment—affords accuracies for BSC that are equal to, and sometimes higher than, the accuracies for WSC (Haxby et al. 2011), suggesting that the neural codes for different individuals are common rather than idiosyncratic. Use of hyperalignment to build a model of a common neural representational space is reviewed in a later section.

REPRESENTATIONAL SIMILARITY ANALYSIS

RSA examines the structure of representations within a representational space in terms of distances between response vectors (**Figure 4**). The complete set of distances among all pairs of response vectors is known as the dissimilarity matrix (DSM) (**Figure 5*b*,*c***). Whereas MVP classification analyzes whether the vectors for different conditions are clearly distinct, RSA analyzes how they are related to each other. This approach confers several advantages. First, RSA can reveal that representations in different brain areas differ even if MVP classification is equivalent in those areas (Kriegeskorte et al. 2008a,b; Connolly et al. 2012a,b). Second, by converting the locations of response vectors from a set of feature coordinates to a set of distances between vectors, the geometry of the representational space is now in a format that is not dependent on the feature coordinate axes. This conversion allows comparison to DSMs for the same conditions in other spaces that have different feature coordinate axes, such as the voxels of another subject's brain or of another brain region (Kriegeskorte et al. 2008a,b; Connolly et al. 2012a,b). It even affords comparison of representational spaces based on stimulus feature models or on other types of brain activity measurement, such as single-unit recordings or MEG.

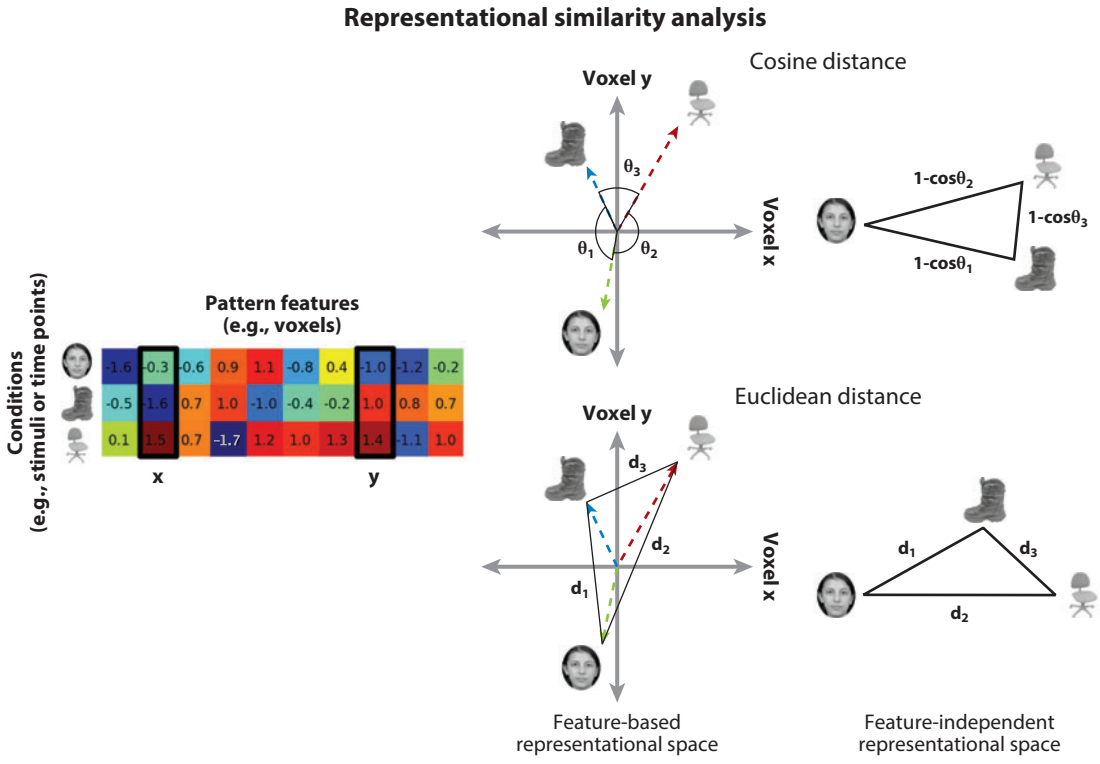


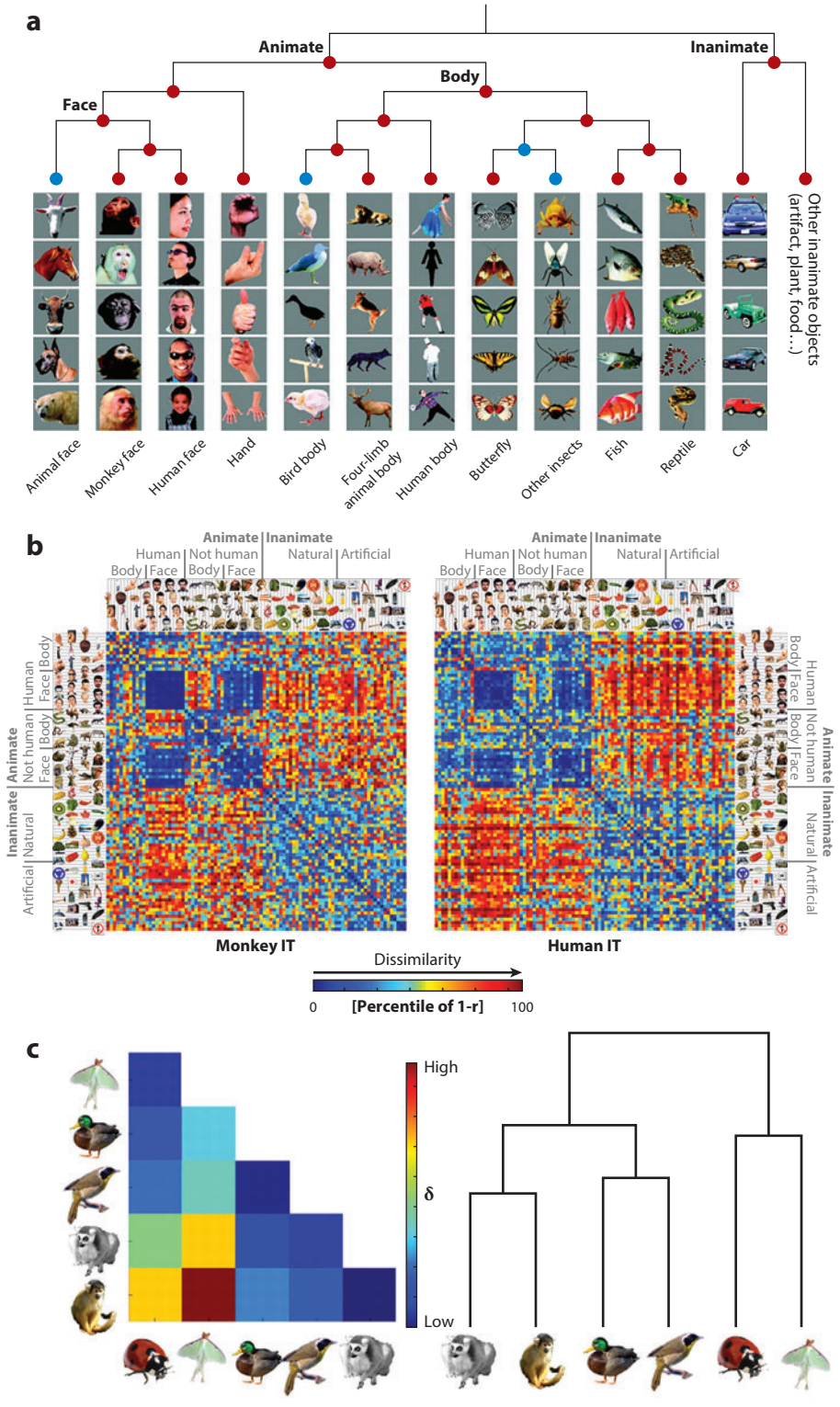
Figure 4

Representational similarity analysis examines the patterns of distances between vectors in the high-dimensional vector space. Measures of angular similarity such as cosine and Pearson product-moment correlation are standard measures that are most sensitive to the relative contributions of feature dimensions. These similarity measures are transformed into dissimilarities by subtracting them from 1. Another standard measure of the distance between vectors is Euclidean distance, which is more sensitive to overall differences in vector length or magnitude.

Investigation of the similarity of neural representations from fMRI data dates back to an early paper by Edelman et al. (1998), which was the first to use multidimensional scaling to visualize the representational space for visual objects. Two groups reanalyzed data from an early MVP classification study on the distributed representation of faces and objects (Haxby et al. 2001; reanalyzed by Hanson et al. 2004; O'Toole et al. 2005, 2007) and found similar similarity structures using different distance measures, one based on intermediate-layer weights in a neural network classifier and the other based on misclassifications. Both found that the strongest dissimilarity was between the animate (human faces and cats) and inanimate (houses, chairs, shoes, bottles, and scissors) categories and, within the inanimate domain, a strong dissimilarity between houses and all the smaller objects. These basic elements of the structure of face, animal, and object representations were corroborated and greatly amplified by subsequent studies in monkeys (Kiani et al. 2007) and humans (Kriegeskorte et al. 2008b). Kiani et al. (2007) measured the responses of single neurons in the monkey inferior temporal (IT) cortex to a large variety of faces, animals, and objects and calculated correlations among response vectors as indices of the similarity of the population response vectors. The results revealed the major distinctions between animate and inanimate stimuli, with a clear distinction between faces and bodies, but went deeper to show a similarity structure for the

Figure 5

Three examples of representational similarity analysis (RSA). (*a*) Dendrogram derived from multiple single-unit recordings in macaque inferior temporal (IT) cortex (from Kiani et al. 2007) shows hierarchical category structure with remarkable detail to the level of different classes of animal body type. (*b*) An example of cross-modal and cross-species RSA analysis for a common set of stimuli (from Kriegeskorte et al. 2008b) shows a high degree of common structure in the representational spaces between humans and monkeys, however, with less definition of subordinate categories within humans. (*c*) A targeted study of subordinate class structure for animal categories (from Connolly et al. 2012b) shows detailed structure for animal classes in human VT cortex.



representations of animals that appears to reflect knowledge about similarities among species (see **Figure 5*a***). Kriegeskorte et al. (2008a,b) used a subset of the stimuli from the Kiani et al. study in an fMRI study in humans and showed that the basic similarity structure of representations in human ventral temporal cortex is quite similar to that for representations in the monkey IT cortex (**Figure 5*b***). Although Kriegeskorte et al. did not show the detailed structure among animal species that was evident in the monkey data, subsequent targeted studies by Connolly et al. (2012b) show that this structure is also evident in the human ventral temporal (VT) cortex (**Figure 5*c***).

RSA can be applied in many different ways to discover the structure of neural representational geometries. These approaches include data-driven analyses and model-driven analyses. Data-driven RSA discovers and describes the similarity structures that exist in different cortical fields. Model-driven RSA searches for cortical fields whose similarity structures are predicted using stimulus or cognitive models, including behavioral ratings of perceived similarity.

Cortical fields that have representational spaces with similarity structures of interest can be identified in numerous ways. Cortical fields that show significant MVP classification can be further analyzed with RSA to describe the similarity structure of the conditions that can be distinguished (**Figure 6*a***). Cortical fields can also be identified by virtue of having similarity structures that

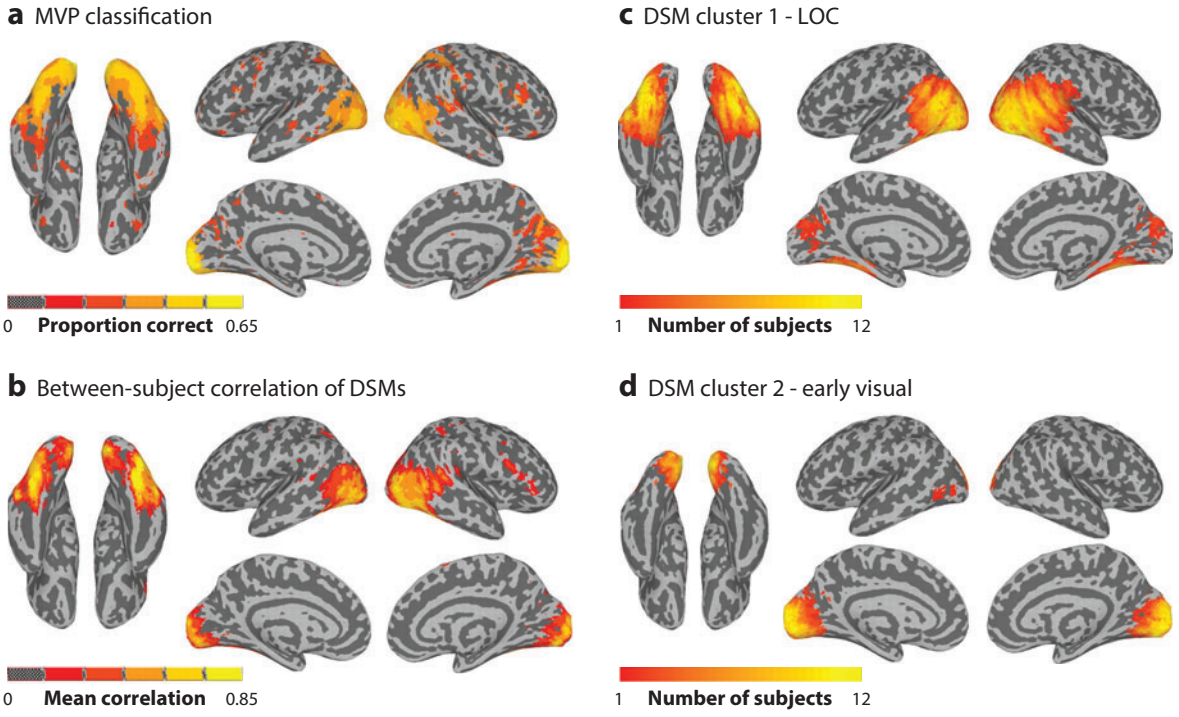


Figure 6

MVPA searchlight analyses (Kriegeskorte et al. 2006) for identifying cortical fields of interest (from Connolly et al. 2012a). MVP classification accuracies (*a*) and consistency in local similarity structures across subjects (*b*) identify similarly large swaths of the visually responsive cortex. Clustering of voxels based on similarities between locally defined searchlight dissimilarity matrix (DSMs) provides a means to identify cortical fields with unique shared structure such as the lateral occipital complex (LOC) (*c*) and the early visual cortex (*d*).

are consistent across subjects (**Figure 6*b***). The similarity structures in different cortical fields identified in these ways, however, may differ. Connolly et al. (2012b) developed a clustering method for finding different similarity structures in different cortical locations that were shared across subjects (**Figure 6*c*,*d***).

Understanding the similarity structure in a cortical field requires examining that structure. Often, the DSM itself is too complicated and high-dimensional to see the structure clearly. The full DSM, therefore, is often distilled into a lower-dimensional illustration to facilitate examination. The most common methods for reducing the dimensionality of a DSM are hierarchical clustering to produce a dendrogram (e.g., **Figure 6*a*,*c***), multidimensional scaling (MDS), and related methods such as DISTATIS (Abdi et al. 2009, 2012a). The dendrogram in **Figure 5*c*** (Connolly et al. 2012b), for example, reveals that animal species from the same class are most similar to each other and that vertebrate classes (primates and birds) are more similar to each other than they are to the invertebrate class (insects). MDS often reveals that a low-dimensional subspace can account for a large portion of a similarity structure. For example, in Connolly et al. (2012b), the similarities among animal classes were captured by a single dimension—the animacy continuum—that ranged from primates to birds to insects and was associated with a distinctive coarse scale topography in human VT cortex that had previously been attributed to the distinction between animate and inanimate stimuli. Finer within-class distinctions (e.g., moths versus ladybugs), however, were based on other dimensions and finer-scale topographies.

The meaning of the similarity structure in a representational space can also be investigated by comparing it to a DSM generated by a model of the experimental conditions based on stimulus features or behavioral ratings (Kriegeskorte et al. 2008a). For example, Connolly et al. (2012b) showed that the DSM in the first cluster (**Figure 6*b***) correlated highly with a DSM based on behavioral ratings of similarities among animals, whereas the second cluster correlated highly with a DSM based on visual features from a model of V1 neuron responses. Carlin et al. (2011) used RSA to investigate the representation of the direction of another's eye gaze that is independent of head angle. They constructed a model similarity structure of gaze direction with invariance across head angle. In addition, they constructed models of similarity structure due to confounding factors, such as image similarity, and searched for areas with a similarity structure that correlated with their DSM of interest after partialling out any shared variance due to confounding factors—illustrating how RSA may be used to test a well-controlled model.

One of the great advantages of RSA is that it strips a cluster of response vectors out of a feature-based representational space into a representational space based on relative distances among vectors. This format allows comparison of representational geometries across subjects, across brain regions, across measurement modalities, and even across species. The second-order isomorphism across these spaces is afforded by the feature-independent format of DSMs. For example, between-subject similarity of DSMs has been exploited to afford between-subject MVP classification (Abdi et al. 2012b, Raizada & Connolly 2012).

The feature-independent second-order isomorphism, however, does have some cost. Stripping representational spaces of features makes it impossible to compare population codes in terms of the constituent tuning functions of those features. Thus, one cannot investigate whether the spaces in different subjects share the same feature tuning functions or how these tuning function codes differ for different brain regions. One cannot predict the response to a new stimulus in a subject on the basis of the responses to that stimulus in other subjects. One cannot predict the tuning function for individual neural features in terms of stimulus features, precluding investigators from predicting the response pattern vector for a new stimulus on the basis of its features. The next two sections review methods that do afford these predictions using hyperalignment and stimulus-model-based encoding and decoding.

BUILDING A COMMON MODEL OF A NEURAL REPRESENTATIONAL SPACE

MVP classification usually builds a new classifier for each individual brain [within-subject classification (WSC)]. Except for distinctions that are carried by large, coarse-scale topographies, BSC based on other subjects' anatomically aligned response vectors yields accuracies that are much lower than those for WSC (see **Figure 3**). Basing decoding on classifiers that are tailored to individual representational spaces leaves open the question of whether the population responses in different brains use the same or idiosyncratic codes, in terms of the tuning functions for individual features within those responses.

Procrustes transformation: aligns two patterns of vectors by finding an orthogonal transformation that minimizes distances between paired vectors in two matrices

Building a common model of a representational space requires an algorithm for aligning the representational spaces of individual subjects' brains into that common space. Anatomical alignment, using affine transformations of the brain volume and rubber-sheet warping of the cortical manifold, does not afford BSC accuracies that approach WSC accuracies (Cox & Savoy 2003, Haxby et al. 2011, Conroy et al. 2013). Algorithms for function-based, rubber-sheet alignment of the cortical manifold, based on either the tuning functions or the functional connectivity of cortical nodes, improve BSC but still do not afford BSC accuracies that are equivalent to WSC accuracies (Sabuncu et al. 2010, Conroy et al. 2013).

A recently developed algorithm for aligning individual neural representational spaces into a common model space, hyperalignment, does afford BSC accuracies that are equivalent to, and sometimes exceed, WSC accuracies (Haxby et al. 2011). The algorithm revolves around a transformation matrix that is calculated for each individual subject that rotates that subject's representational space into the common model space (**Figure 7**). Valid parameters with broad general validity for high-level visual representations can be calculated on the basis of brain responses measured while subjects watch a complex, dynamic stimulus. This method was demonstrated using data collected while subjects watched the full-length action movie *Raiders of the Lost Ark*, reasoning that the subjects' visual cortices represent the same visual information while they watch the movie. The response vectors in different subjects' brains, however, are not aligned because voxels in the same anatomical locations do not have the same tuning functions. Hyperalignment uses the Procrustes transformation (Schönemann 1966) to rotate the coordinate axes of an individual's representational space to bring that subject's response vectors into optimal alignment with another subject's vectors. Iterative alignments of individual representational spaces to each other produced a single common representational space for the VT cortex. Each individual representational space could then be rotated into that common model space. The dimensionality of the common model space was reduced using principal components analysis. Optimal BSC accuracy for validation testing across experiments required more than 30 dimensions. Thus, the common model of the representational space in the VT cortex has 35 dimensions, meaning that the transformation matrix is an orthogonal matrix with 35 columns for the 35 common model dimensions and the same number of rows as the number of voxels in an individual subject's VT cortex (**Figure 7**).

The hyperalignment transformation matrix derived from responses to the movie provides the keys that unlock an individual's neural code. The parameters derived from the movie have general validity across a wide range of visual stimuli and can be applied to data from any experiment, making it possible to decode a subject's response vectors for a wide variety of stimuli based on other subjects' brain responses. BSC of data from two category perception experiments, after transformation into the common model dimensions using parameters derived from movie viewing, was equivalent to WSC. In subsequent work, we have found that the algorithm produces valid common models of representational spaces in early visual cortex, in lateral occipital and lateral temporal visual cortices, in auditory cortices, and in motor cortices.

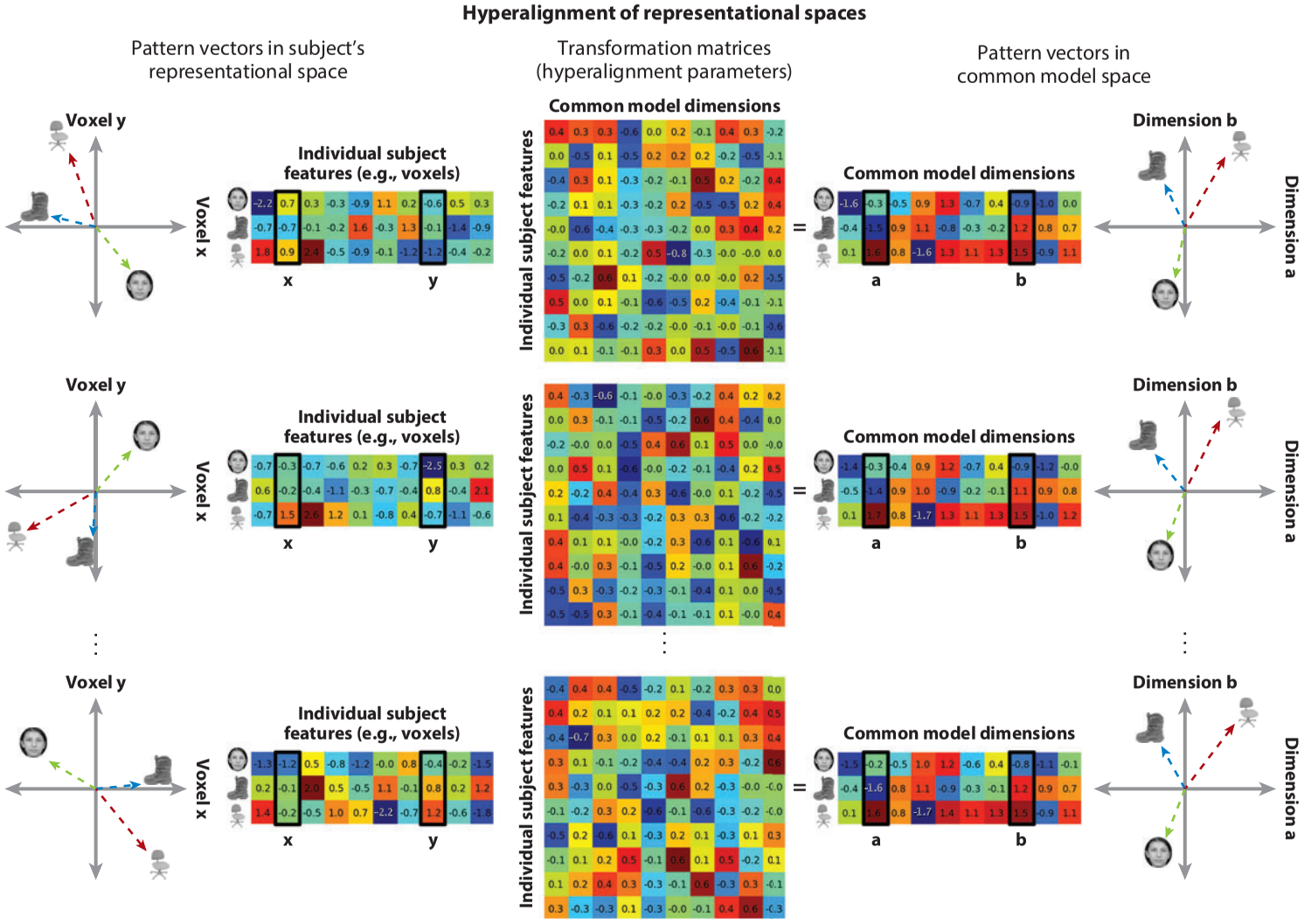


Figure 7

Intersubject hyperalignment of neural representational spaces. Hyperalignment aligns individual subjects' representational spaces into a common model representational space using high-dimensional rotation characterized by an orthogonal transformation matrix for each subject. A dimension in this common model space is a weighted sum of voxels in each individual subject (a column of the transformation matrix), which is functionally equivalent across subjects as reflected by the alignment of pattern vectors in the common model space.

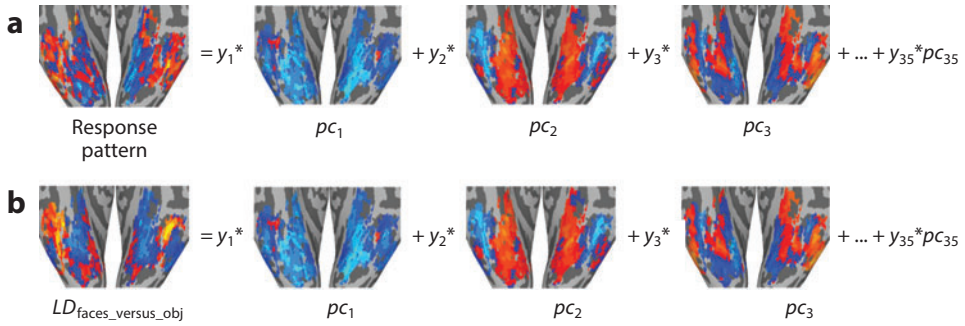


Figure 8

Modeling response patterns as weighted sums of common model dimensions. (*a*) Any response pattern in a subject can be modeled as a weighted sum of patterns representing common model dimensions. (*b*) Category-selective regions defined by contrasts, such as faces-versus-objects for the fusiform face area, can be modeled in the same way.

The transformation matrix provides a set of basis functions that can model any pattern of brain response. Each column in the matrix, corresponding to one common model dimension, is a set of weights distributed across voxels. A pattern of brain response in those voxels is a weighted sum of these patterns of weights (**Figure 8*a***). Models of VT cortex based on single dimensions, such as contrasts that define category-selective regions, are modeled well in the 35-dimensional model space (**Figure 8*b***), but these single dimensions account for only a small portion of the variance in responses to a dynamic and varied natural stimulus such as the movie. For example, the contrast between responses to faces and responses to objects, which defines the fusiform face area (FFA) (Kanwisher et al. 1997), accounts for only 12% of the variance that is accounted for by the 35-dimensional model. This result indicates that models based on simple, univariate contrasts are insufficient as models of neural representational spaces.

The use of a complex, dynamic stimulus is essential for deriving transformation matrix parameters that afford general validity across a wide range of stimuli. Transformation matrices can also be calculated on the basis of responses to more controlled experiments, such as the category perception experiments. These transformation matrices are valid for modeling the response vectors for stimuli in that experiment but, when applied to data from other experiments, do not afford BSC of new stimuli (Haxby et al. 2011). This result indicates that data from a limited sampling of brain states, such as those sampled in a standard category perception experiment, do not provide a sufficient basis for building a common model of a neural representational space.

STIMULUS-MODEL-BASED ENCODING AND DECODING

For MVP classification, RSA, and hyperalignment, a response vector to be decoded is compared with response vectors for that same stimulus measured in the same subject or in other subjects. These methods cannot predict the response pattern for a novel stimulus or experimental condition. Stimulus-model-based methods extend neural decoding to novel stimuli by predicting the response to stimulus features rather than to whole stimuli.

The stimuli used to produce training data for stimulus-model-based decoding are analyzed into constituent features. Feature sets used for this type of analysis include models of V1 neuron response profiles, namely oriented Gabor filters (Kay et al. 2008, Naselaris et al. 2009), visual motion energy filters (Nishimoto et al. 2011), semantic features (Mitchell et al. 2008, Naselaris

et al. 2009), and acoustic features of music (Casey et al. 2012). These features can be continuous or binary. Thus, each stimulus is characterized as a vector in the high-dimensional stimulus feature space (**Figure 9**). The response for each feature in a neural representational space (e.g., voxel) is then modeled as a weighted sum of the stimulus features. For example, a V1 voxel may have the strongest weights for Gabor filters of a certain orientation and spatial frequency in a particular location, reflecting the orientation selectivity and retinotopic receptive field for that voxel. The prediction equation for each voxel is calculated on the basis of regularized regression analysis of responses to the stimuli in the training data. The result is a linear transformation matrix in which each column is a neural pattern feature and each row is a stimulus feature (**Figure 9**). The values in each column are the regression weights for predicting the response of that neural feature given a set of stimulus feature values.

ECoG: electrocorticography

Applying the encoding parameter transformation matrix to the stimulus feature values for a new stimulus thus predicts the response in each voxel (**Figure 9**). This new stimulus can be any new stimulus in the same domain that can be described with the same stimulus features. For example, using Gabor filters, investigators can predict the response to any natural still image (Kay et al. 2008, Naselaris et al. 2009). Using motion energy filters, the response to any video can be predicted (Nishimoto et al. 2011). Using acoustic features, the response to any clip of music can be predicted (Casey et al. 2012). The validity of the transformation can then be tested using either MVP classification or Bayesian reconstruction of the stimulus. Each type of validation testing involves analysis of response vectors for new stimuli. For MVP classification, neural response vectors are predicted for stimuli in the validation testing set using the linear transformation matrix estimated from an independent set of training stimuli. The classifier then tests whether the measured response vector is more similar to the predicted response vector for that stimulus than to predicted response vectors for other stimuli in the testing set. For Bayesian reconstruction, the algorithm identifies predicted response vectors generated from a large set of stimuli (priors) that are most similar to the measured response vector. The reconstructed stimulus is then produced from those matching stimuli. For example, Nishimoto et al. (2011) found the 30 videos that generated predicted response vectors that most closely matched a measured response vector. They then produced a video by averaging those 30 video priors. The reconstructed videos bear an unmistakable resemblance to the viewed video, albeit lacking detail, providing a convincing proof of concept.

Related methods have been used to reconstruct 10 $\times$ 10 contrast images (Miyawaki et al. 2008) and decode the contents of dream imagery (Horikawa et al. 2013). A study of brain activity measured over the auditory language cortex using electrocorticography (ECoG) in surgery patients used related methods to reconstruct the spectrogram for spoken words, producing auditory stimuli that closely resembled the original words (Pasley et al. 2012).

In general, stimulus-model-based decoding is limited to Bayesian reconstruction based on similarities to a set of priors. Simply generating a stimulus based on predicted features is infeasible because the dimensionality of the neural representational space, given current brain measurement techniques, is much lower than that of stimulus feature spaces that are complete enough to construct a stimulus. Some researchers have speculated, however, that perception also involves a process of Bayesian reconstruction based on similarity to prior perceptual experiences, a process that could operate effectively with incomplete specification of stimulus features (Friston & Kiebel 2009).

MULTIPLEXED TOPOGRAPHIES FOR POPULATION RESPONSES

The spatial resolution of fMRI used in most neural decoding studies is 2–3 mm. Consequently, each fMRI voxel contains more than 100,000 neurons and roughly 10–100 cortical columns.

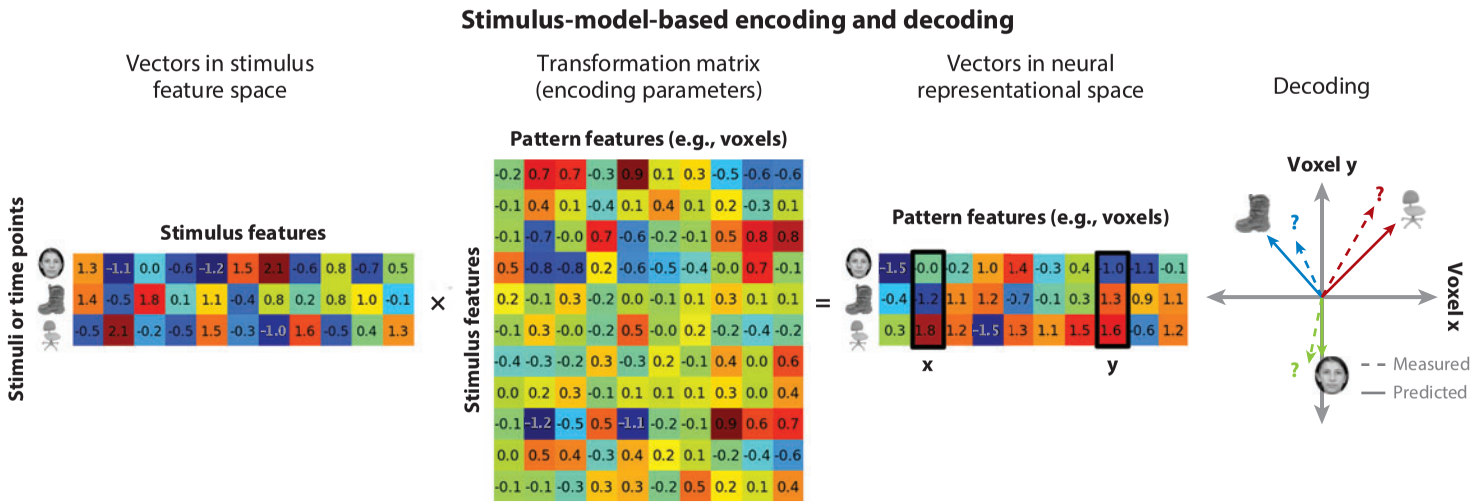


Figure 9

Stimulus-model-based encoding and decoding involve deriving a transformation matrix that affords prediction of responses of neural features from stimulus-model features. Each column in the encoding transformation matrix corresponds to a neural feature and provides the weights for stimulus features to estimate that neural feature's response. Decoding responses to novel stimuli involves comparing measured response patterns to predicted patterns for stimulus priors.

The fact that neural activity resampled into this coarse spatial grid can be effectively decoded suggests that the topographies for distinct neural representations of the decoded information include lower spatial frequency patterns. In fact, decoding of different distinctions appears to be based on topographies of different spatial scales. Coarse distinctions, such as the difference between animate and inanimate stimuli (Martin 2007, Mahon et al. 2009) or animal species at different levels on the animacy continuum (Connolly et al. 2012b), are found in a coarse topography from lateral to medial VT cortex. Finer distinctions, such as those between old and young human faces (Op de Beeck et al. 2010, Brants et al. 2011) or between two types of birds (Connolly et al. 2012b), are carried by finer-scale topographies. The finer-scale topographies for subordinate distinctions, such as old versus young faces, are not restricted to the areas of maximal activity for the superordinate category (Op de Beeck et al. 2010, Brants et al. 2011, Haxby et al. 2011). Thus, the topographies for different distinctions appear to be overlapping and exist at multiple scales.

The topographic organization of the cortex reflects the problem of projecting a high-dimensional representational space, composed of features with complex, interrelated tuning functions, into a two-dimensional manifold. Kohonen (1982, 2001) proposed that cortical maps self-organize to locate neurons with related tuning functions close to each other. Aflalo & Graziano (2006, 2011; Graziano & Aflalo 2007a,b) have used this principle of spatial continuity of function to account for the topographic organization of the motor cortex (Aflalo & Graziano 2006) and the coarse-scale topographic organization of the extrastriate visual cortex (Aflalo & Graziano 2011). Others have used this principle to account for multiplexed functional topographies in the primary visual (Durbin & Mitchison 1990) and auditory (Schreiner 1995) cortices. Accounting for seemingly disordered multiplexed functional topographies requires an adequate model of the high-dimensional representational space. Haxby et al. (2011) showed that the topographies in VT cortex that support a wide range of stimulus distinctions, including distinctions among responses to complex video segments, can be modeled with 35 basis functions. These pattern bases are of different spatial scales and define gradients in different locations. Low-dimensional models of neural representation, such as those exemplified by category-selective regions (Kanwisher 2010, Weiner & Grill-Spector 2011), however, are not sufficient to model complex, multiplexed functional topographies that support these distinctions.

FUTURE DIRECTIONS

Recent advances in computational methods for neural decoding have revealed that the information provided from measurement of human brain activity is far more detailed and specific than was previously thought possible. This review of the current state of the art shows that these methods can be integrated in a framework organized around the concept of high-dimensional representational spaces. The development of algorithms for neural encoding and decoding, however, has only begun. Although the power of neural decoding is limited by brain activity measurement methods—and further technological breakthroughs will bring greater power and sensitivity to neural decoding projects—new computational methods can direct investigators to additional important topics. Three areas for future investigation are addressed below.

Individual and Group Differences

Most neural decoding work to date has focused on the commonality of neural representational spaces across subjects. The methods for aligning representational spaces across subjects, namely RSA and hyperalignment, however, can also be adapted to investigate how an individual's representational space differs from others' or how groups differ. Developing methods for examining

individual and group differences would facilitate studies of how factors such as development, education, genetics, and clinical disorders influence neural representation.

Between-Area Transformations in a Processing Pathway

RSA affords one way to draw distinctions among how representations are structured in different parts of a processing pathway (Kriegeskorte et al. 2008b, Connolly et al. 2012b). Modeling the transformation of representations from one cortical field to another, however, would help elucidate how information is processed within a pathway, leading to the construction of representations laden with meaning from representations of low-level physical stimulus properties. For example, the manifold of response vectors that correspond to different views of the same face in early visual cortex is complex and does not afford easy separation of the responses to one individual face from the responses to another, whereas the manifold in the anterior temporal cortex may untangle these manifolds, affording viewpoint-invariant identity recognition (DiCarlo & Cox 2007, Freiwald & Tsao 2010). Determining the structure of these transformations and the role of input from multiple regions is a major challenge for future work.

Multimodality Decoding

fMRI has very coarse temporal resolution. Neural decoding studies with other measurement modalities, such as single-unit recording (e.g., Hung et al. 2005, Freiwald & Tsao 2010), ECoG (Pasley et al. 2012), and MEG (Carlson et al. 2011, Sudre et al. 2012), have shown how population codes for different types of information emerge over time as measured in tens of milliseconds. Similar tracking of population codes has been demonstrated with fMRI but is severely limited by the temporal characteristics of the hemodynamic response (Kohler et al. 2013). Using multiple modalities, representational spaces could be modeled in which some dimensions reflect different time points for the same spatial feature and other dimensions reflect spatiotemporal gradients or wavelets. The potential of multimodal neural decoding is largely unexplored.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abdi H, Dunlop JP, Williams LJ. 2009. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). *NeuroImage* 45:89–95

Abdi H, Williams LJ, Conolly AC, Gobbini MI, Dunlop JP, Haxby JV. 2012a. Multiple Subject Barycentric Discriminant Analysis (MUSUBADA): how to assign scans to categories without using spatial normalization. *Comp. Math. Methods Med.* 2012:634165

Abdi H, Williams LJ, Valentin D, Bennani-Dosse M. 2012b. STATIS and DISTATIS: optimum multitable principal component analysis and three way metric multidimensional scaling. *Wiley Interdiscip. Rev. Comput. Stat.* 4:124–67

Aflalo TN, Graziano MSA. 2006. Possible origins of the complex topographic organization of motor cortex: reduction of a multidimensional space onto a two-dimensional array. *J. Neurosci.* 26:6288–97

Aflalo TN, Graziano MSA. 2011. Organization of the macaque extrastriate cortex re-examined using the principle of spatial continuity of function. *J. Neurophysiol.* 105:305–20

Brants M, Baeck A, Wagemans J, Op de Beeck H. 2011. Multiple scales of organization for object selectivity in ventral visual cortex. *NeuroImage* 56:1372–81

Carlin JD, Calder AJ, Kriegeskorte N, Nili H, Rowe JB. 2011. A head view-invariant representation of gaze direction in anterior superior temporal cortex. *Curr. Biol.* 21:1817–21

Carlson TA, Hogendoorn H, Kanai R, Mesik J, Turret J. 2011. High temporal resolution decoding of object position and category. *J. Vis.* 11:1–17

Carlson TA, Schrater P, He S. 2003. Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci.* 15:704–17

Casey M, Thompson J, Kang O, Raizada R, Wheatley T. 2012. Population codes representing musical timbre for high-level fMRI categorization of music genres. In *Machine Learning and Interpretation in Neuroimaging*, Ser. Vol. 7263, ed. G Langs, I Rish, M Grosse-Wentrup, B Murphy, pp. 36–41. Berlin/Heidelberg: Springer-Verlag

Chen Y, Namburi P, Elliott LT, Heinzle J, Soon CS, et al. 2011. Cortical surface-based searchlight decoding. *NeuroImage* 56:582–92

Connolly AC, Gobbini MI, Haxby JV. 2012a. Three virtues of similarity-based multi-voxel pattern analysis: an example from the human object vision pathway. In *Understanding Visual Population Codes* (*UVPC*): *Toward A Common Multivariate Framework for Cell Recording and Functional Imaging*, ed. N Kriegeskorte, G Kreiman, pp. 335–55. Cambridge, MA: MIT Press

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, et al. 2012b. The representation of biological classes in the human brain. *J. Neurosci.* 32:2608–18

Conroy BR, Singer BD, Guntupalli JS, Ramadge PJ, Haxby JV. 2013. Inter-subject alignment of human cortical anatomy using functional connectivity. *NeuroImage* 81:400–11

Cortes C, Vapnik V. 1995. Support-vector networks. *Mach. Learn.* 20:273–97

Cox DD, Savoy RL. 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19:261–70

DiCarlo JJ, Cox DD. 2007. Untangling invariant object recognition. *Trends Cogn. Sci.* 11:333–41

Durbin R, Mitchison G. 1990. A dimension reduction framework for understanding cortical maps. *Nature* 343:644–47

Edelman S, Grill-Spector K, Kushnir T, Malach R. 1998. Toward direct visualization of the internal shape space by fMRI. *Psychobiology* 26:309–21

Formisano I, De Martino F, Bonte M, Goebel R. 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. *Science* 322:970–73

Freiwald WA, Tsao DY. 2010. Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science* 330:845–51

Friston K, Kiebel S. 2009. Predictive coding under the free-energy principle. *Phil. Trans. R. Soc. B.* 364:1211–21

Graziano MSA, Aflalo TN. 2007a. Mapping behavioral repertoire onto the cortex. *Neuron* 56:239–51

Graziano MSA, Aflalo TN. 2007b. Rethinking cortical organization: moving away from discrete areas arranged in hierarchies. *Neuroscientist* 13:138–47

An integrated software system based on Python for performing neural decoding.

Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollman S. 2009. PyMVPA: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics* 7:37–53

Initial paper on multivariate pattern classification of fMRI.

Hanson SJ, Toshihiko M, Haxby JV. 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *NeuroImage* 23:156–67

Harrison SA, Tong F. 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 458:632–35

Haxby JV. 2012. Multivariate pattern analysis of fMRI: the early beginnings. *NeuroImage* 62:852–55

Introduces hyperalignment for building a common high-dimensional model of a neural representational space.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–30

Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, et al. 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72:404–16

Haynes JD, Rees G. 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* 8:686–91

Haynes JD, Rees G. 2006. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7:523–34
Haynes JD, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. 2007. Reading hidden intentions in the human brain. *Curr. Biol.* 17:323–28
Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y. 2013. Neural decoding of visual imagery during sleep. *Science* 340:639–42
Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310:863–66
Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8:679–85
Kanwisher N. 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. USA* 107:11163–70
Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17:4302–11
Kay KN, Naselaris T, Prenger RJ, Gallant JL. 2008. Identifying natural images from human brain activity. *Nature* 452:352–55
Kiani R, Esteky H, Mirpour K, Tanaka K. 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97:4296–309
Kohler PJ, Fogelson SV, Reavis EA, Meng M, Guntupalli JS, et al. 2013. Pattern classification precedes region-average hemodynamic response in early visual cortex. *NeuroImage* 78:249–60
Kohonen T. 1982. Self-organizing formation of topologically correct feature maps. *Biol. Cybern.* 43:59–69
Kohonen T. 2001. *Self-Organizing Maps*. Berlin: Springer
Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. *Proc. Natl. Acad. Sci. USA* 103:3863–68
Kriegeskorte N, Kievet RA. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17:401–12
Kriegeskorte N, Mur M, Bandettini P. 2008a. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, et al. 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60:1126–41
Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. 2009. Circular analysis is systems neuroscience: the dangers of double dipping. *Nat. Neurosci.* 12:535–40
Mahon BZ, Anzellotti S, Schwarzbach J, Zampini M, Caramazza A. 2009. Category-specific organization in the human brain does not require visual experience. *Neuron* 63:397–405
Martin A. 2007. The representation of object concepts in the brain. *Annu. Rev. Psychol* 58:25–45
Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–95
Miyawaki Y, Uchida H, Yamashita O, Sato M, Morito Y, et al. 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron* 60:915–29
Naselaris T, Kay KN, Nishimoto S, Gallant JL. 2011. Encoding and decoding in fMRI. *NeuroImage* 56:400–10
Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. 2009. Bayesian reconstruction of natural images from human brain activity. *Neuron* 63:902–15
Nishimoto S, Vu AT, Naselaris T, Bejamini Y, Yu B, Gallant JL. 2011. Reconstructing visual experience from brain activity evoked by natural movies. *Curr. Biol.* 21:1641–46
Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10:424–30
Oosterhof NN, Wiestler T, Downing PE, Diedrichsen J. 2011. A comparison of volume-based and surface-based multi-voxel pattern analysis. *NeuroImage* 56:593–600
Op de Beeck H, Brants M, Baeck A, Wagemans J. 2010. Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. *NeuroImage* 49:3414–25
O'Toole AJ, Jiang F, Abdi H, Haxby JV. 2005. Partially distributed representations of objects and faces in ventral temporal cortex. *J. Cogn. Neurosci.* 17:580–90

First paper to show that MVPA can decode an early visual feature, namely edge orientation.

Introduced stimulus-model-based encoding and decoding of natural images.

Introduces RSA as a common format for the geometry of representational spaces.

First demonstration that RSA affords comparison of representational spaces across species.

This paper introduced decoding of words and concepts based on semantic feature models.

Reviews methods for stimulus-model-based encoding and decoding.

O'Toole AJ, Jiang F, Abdi H, Pénard N, Dunlop JP, Parent MA. 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J. Cogn. Neurosci.* 191:735–52

Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, et al. 2012. Reconstructing speech from human auditory cortex. *PLoS Biol.* 10:e1001251

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

Raizada RDS, Connolly AC. 2012. What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding. *J. Cogn. Neurosci.* 24:868–77

Sabuncu M, Singer BD, Conroy B, Bryan RE, Ramadge PJ, Haxby JV. 2010. Function-based intersubject alignment of human cortical anatomy. *Cereb. Cortex* 20:130–40

Schönemann PH. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika* 31:1–10

Schreiner CE. 1995. Order and disorder in auditory cortical maps. *Curr. Opin. Neurobiol.* 5:489–96

Shinkareva SV, Malave VL, Mason RA, Mitchell TM, Just MA. 2011. Commonality of neural representations of words and pictures. *NeuroImage* 54:2418–25

Shinkareva SV, Mason RA, Malave VL, Wang W, Mitchell TM, Just MA. 2008. Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE* 1:e1394

Soon CS, Brass M, Heinze HJ, Haynes JD. 2008. Unconscious determinants of free decisions in the human brain. *Nat. Neurosci.* 5:543–45

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. 2009. Sound categories are represented as distributed patterns in the human auditory cortex. *Curr. Biol.* 19:498–502

Sudre G, Pomerleau D, Palatucci M, Wehbe L, Fyshe A, et al. 2012. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage* 62:451–63

Tong F, Pratte MS. 2012. Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63:483–509

Weiner K, Grill-Spector K. 2011. The improbable simplicity of the fusiform face area. *Trends Cogn. Sci.* 16:251–54

Yamashita O, Sato M-A, Yoshioka T, Tong F, Kamitani Y. 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *NeuroImage* 42:1414–29

Annual Review of Neuroscience

Volume 37, 2014

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Neuroscience* articles may be found at http://www.annualreviews.org/errata/neuro