*Social Cognitive and Affective Neuroscience*, 2020, 487–508

doi: 10.1093/scan/nsaa057
Advance Access Publication Date: 25 April 2020
Original article

# *Tools of the Trade*
# Multivoxel pattern analysis in fMRI: a practical introduction for social and affective neuroscientists

Miriam E. Weaverdyck,[1] Matthew D. Lieberman,[1] and Carolyn Parkinson[1,2]

[1]Department of Psychology, University of California, Los Angeles, CA 90095, USA and [2]Brain Research Institute, University of California, Los Angeles, CA 90095, USA

Correspondence should be addressed to Miriam E. Weaverdyck, Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Box 951563, Los Angeles, CA 90095, USA. E-mail: mweaverdyck@ucla.edu.



## Abstract

The family of neuroimaging analytical techniques known as multivoxel pattern analysis (MVPA) has dramatically increased in popularity over the past decade, particularly in social and affective neuroscience research using functional magnetic resonance imaging (fMRI). MVPA examines patterns of neural responses, rather than analyzing single voxel- or region-based values, as is customary in conventional univariate analyses. Here, we provide a practical introduction to MVPA and its most popular variants (namely, representational similarity analysis (RSA) and decoding analyses, such as classification using machine learning) for social and affective neuroscientists of all levels, particularly those new to such methods. We discuss how MVPA differs from traditional mass-univariate analyses, the benefits MVPA offers to social neuroscientists, experimental design and analysis considerations, step-by-step instructions for how to implement specific analyses in one's own dataset and issues that are currently facing research using MVPA methods.

**Key words:** multivoxel pattern analysis; representational similarity analysis; classification; fMRI; social neuroscience

## Introduction

Over the past two decades, the field of social neuroscience has grown rapidly, with an explosion of research linking neuroimaging data to social psychological phenomena. In such a quickly expanding field, and as techniques become more computationally sophisticated, analytical methods can become increasingly inaccessible to scientists not in labs already using them. This article aims to provide an accessible and practical introduction to the family of analyses known as multivoxel pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data for a broad audience of researchers, particularly social and affective neuroscientists new to such methods.

Traditionally, univariate or mass-univariate approaches to analyzing fMRI data have been used to examine the changes in average or peak neural responses across conditions of a study (e.g. the amygdala shows greater activation to fear-inducing than neutral stimuli). This type of approach is referred to as 'univariate' because the corresponding statistical tests only consider one value per condition (e.g. the average signal of a region or voxel) at a time. Recently, more and more researchers are using analyses that consider patterns of responses across multiple voxels, called MVPA, rather than single voxel- or region-based values. Since the decisions made while analyzing and designing MVPA studies dramatically shape the final results, it is essential to understand and think carefully about these issues, rather than relying solely on default software settings or common lab practices. Here, we will focus on the fundamentals of how to use these methods by covering (i) what MVPA is and the opportunities it offers social and affective neuroscience research, (ii) practical explanations and tips for how to implement it,

**Received:** 30 October 2019; **Revised:** 8 April 2020; **Accepted:** 15 April 2020

(iii) types of questions that may be answered with this approach, and (iv) issues currently facing research using MVPA.

## What is MVPA?

fMRI data consist of a single blood-oxygen-level dependent (BOLD) signal value at each point (i.e. voxel) in the brain for every time point (i.e. TR). Instead of looking at each voxel separately, or averaging across voxels' varying signals within a region, as in univariate analyses, MVPA looks for information in the patterns of neural responses across voxels (Figure 1). To illustrate this difference, imagine typing the words 'dog' and 'cat' on a keyboard. If you simply sum the number of keyboard presses, you find no difference between the two words (i.e. no univariate effect), but if you look at which keys were pressed (i.e. the pattern across the whole keyboard), you find distinct patterns that communicate distinct meanings. In the same way, MVPA examines the information carried in the pattern of responses, while univariate analyses only consider the overall magnitude of the response. While any analysis that considers multiple voxel values at a time may fall in the category of MVPA, the two most widely used varieties of MVPA, which are often used in tandem on the same datasets, are decoding analyses and representational similarity analyses (RSA; Kriegeskorte *et al.*, 2008a), as described below.

### Decoding analyses

Decoding analyses, such as classification and regression (Table 1), try to identify what condition elicited a given neural response. In other words, the direction of inference common in traditional univariate analyses—*P*(*brain*|*condition*)—is reversed in decoding analyses, *P*(*condition*|*brain*). An analysis often used in MVPA decoding is classification, which involves attempting to predict (i.e. classify) which categories correspond to which observations—e.g. was a given neural response elicited by an angry or surprised face? However, decoding analyses also encompass methods (e.g. regression analyses) that treat data as continuous—e.g. how angry was the face that elicited a given neural response? Here, we will describe decoding analyses in very general terms; see Practical Implementation section for more details. Generally in MVPA, decoding analyses entail applying supervised machine learning algorithms using out-of-sample prediction, which involves separating the data into training and testing datasets. The training dataset is used to train an algorithm to distinguish between data corresponding to different conditions (classification) or along a continuous scale (regression). The resulting model is then tested on the testing data, which it has never seen. That is, the algorithm tries to detect generalizable systematic differences in the neural response patterns elicited by each condition. There are numerous ways of implementing out-of-sample prediction, including *k*-fold cross-validation and cross-classification (see Practical Implementation section; Table 2). The model's ability to correctly predict which conditions produced the multivoxel patterns in new data reflects the extent to which this information is reliably carried in the neural response patterns.

Such information may not be carried in the overall response magnitude of the region and thus missed by traditional univariate analyses. For example, in their seminal MVPA study, Haxby *et al.* (2001) showed participants images of faces, houses and other objects. They found that the distributed response patterns in ventral temporal cortex distinguished between several categories of visual objects. This included categories for which the overall response magnitude in this brain region did not differ substantially and were thus not separable using univariate analyses (e.g. chairs, shoes, houses). Since the publication of this paper, the use of MVPA to decode the contents of participants' mental states has rapidly expanded, with applications ranging from decoding what people are dreaming about (Horikawa *et al.*, 2013) to the sounds that they are hearing (Giordano *et al.*, 2013), the faces they are seeing (Goesaert and Op de Beeck, 2013) and the people they are imagining (Hassabis *et al.*, 2014).

### Representational similarity analysis

***First- vs second-order isomorphisms.*** Instead of looking directly at the neural response patterns elicited by the different classes of stimuli, RSA examines the relative similarity of the patterns across stimuli. A direct relationship between neural responses and stimuli (e.g. seeing faces elicits more activity in the fusiform face area (FFA) than seeing houses) is called a first-order isomorphism (Figure 2B) and is the basis of most neuroscience research. RSA, on the other hand, considers the relations among neural response patterns, often comparing them to the relations among a stimulus property (or, alternatively, comparing the relations among neural response patterns across individuals). A correspondence between two sets of relations is referred to as a second-order isomorphism (Figure 2D; Shepard and Chipman, 1970). To illustrate this, imagine a picture of a large face on a billboard *vs* a small doodle of a smiley face. The details of the facial features themselves may vary widely across the two images (i.e. no first-order isomorphism); however, there will be relationships between facial features within each image that are consistent across the two representations of faces (e.g. the eyes will be closer to the nose than to the mouth; a second-order isomorphism). In the same way, a brain region that encodes some property (e.g. the presence of a human face) may not show a direct correspondence between the intensity of that property (e.g. how much an image resembles a human face) and its neural response but may show a correspondence between the relations among stimuli in terms of that property (e.g. how similar two faces are in terms of how human they look—two people are very similar to one another while they are both very different from a giraffe; Figure 2C) and the relations among neural responses (e.g. the relative similarity of neural response patterns evoked by those three faces). Similarly, we can compare these sets of relations among neural responses from one subject to those from another subject to test if two people represent the stimuli similarly. In other words, in a brain region that plays the same functional role across individuals, patterns of neural activity evoked by the same stimuli may show little or no correspondence across individuals (Figure 2B), but the relations among those patterns may be consistent (Figure 2D).

***Comparison across modalities.*** To quantify the relations among neural responses, we create a representational dissimilarity matrix (RDM), which shows how similar the neural response patterns elicited by each condition are to each other (Figure 2C). In this way, we can characterize how a brain region distinguishes between a set of stimuli (i.e. its 'information signature'; Kriegeskorte *et al.*, 2008a). In addition to neural RDMs, we can create other RDMs that reflect the differences between stimuli based on specific properties of interest (e.g. objective attributes, behavioral ratings, model predictions, responses measured with other modalities). These may be compared to neural RDMs to test which distinctions the brain 'cares about' (e.g. does a particular

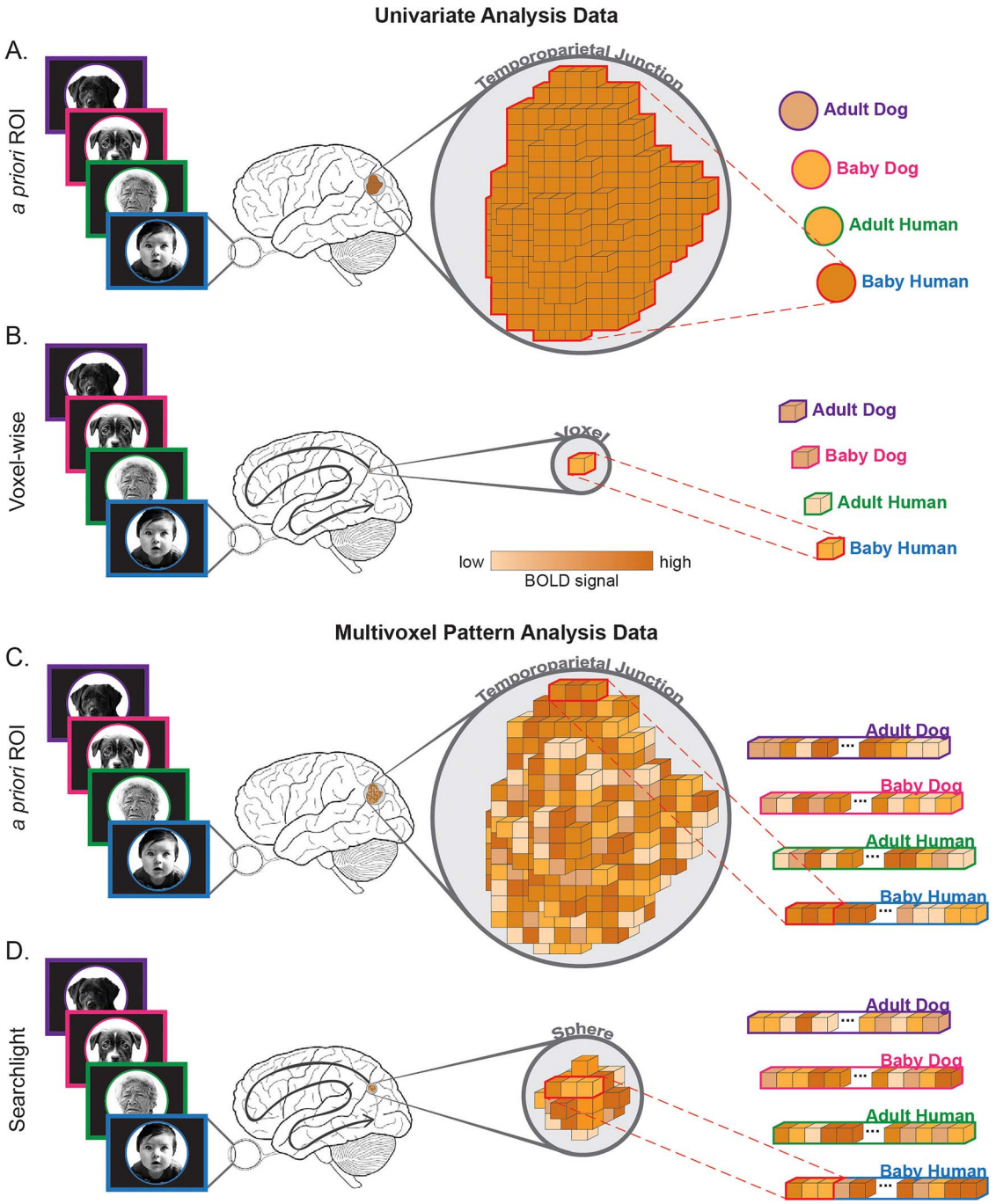



**Fig. 1**. Comparing data in univariate analyses and MVPA. This figure illustrates the differences between how data elicited by four stimuli or experimental conditions (i.e. viewing young and old human and dog faces) is used in univariate analyses (A, B) and MVPA (C, D) as well as how to test a region defined a priori (A, C) *vs* at every point in the brain (B, D). Each method results in data for each condition (right) that is analyzed and compared (see Figures 3 and 4). (A) Univariate analyses in regions defined a priori use a summary statistic (e.g. mean or peak value) to describe the response magnitude across the entire region. (B) Univariate analyses may also be performed on every voxel independently (mass-univariate analyses). (C) MVPA in regions defined a priori use the pattern of neural responses across all voxels strung out into a vector. (D) In a searchlight analysis, a sphere (here with a radius of two voxels) is defined around every voxel, and the pattern of responses in this sphere is strung out in a vector for each condition. The resulting values displayed on the right are then used in the analyses described in other figures.

brain region distinguish faces based on how human they look? Figures 2D and 4).

***Data-driven exploration of representational structure.*** Neural RDMs may also be used to explore how representations are structured in particular brain regions in a data-driven way using tools such as multidimensional scaling (MDS), clustering or dimensionality reduction methods (Figure 4B). For example, when using MDS to visualize how different faces are represented in a given brain region, we may discover that faces cluster by age rather than affect, a phenomenon we may miss when only testing if the affective state of the face could be decoded from

**Table 1.** MVPA methods

| Type of MVPA | | Discrete or continuous condition labels | Supervised or unsupervised learning[a] | Result of analysis |
|---|---|---|---|---|
| Decoding analysis[b] | Classification | Discrete | Supervised | Out-of-sample prediction |
| | Regression[c] | Continuous | Supervised | Out-of-sample prediction |
| RSA | MDS/clustering | Either | Unsupervised | Data-driven description of how representations are organized |
| | RDM comparison | Either | Neither (direct test of relationship) | Relationship between neural and non-neural RDMs |

[a]'Supervised learning' is used to uncover a mapping between a set of observations and a set of labels. This requires training data to learn the relevant distinctions and testing data to assess its ability to accurately predict the labels for previously unseen observations. 'Unsupervised learning' is used to discover the underlying structure/representation of a set of observations and does not require any model or training data.
[b]Decoding analyses are also referred to as 'machine learning' or 'statistical learning'. This approach uses algorithms to learn mappings between data and labels. In the context of this manuscript, these three terms are synonymous.
[c]While some forms of regression analyses are used to predict the probabilities that samples belong to discrete categories (e.g. logistic regression), in this manuscript, we use the term 'regression' to encompass decoding analyses that predict continuous variables.

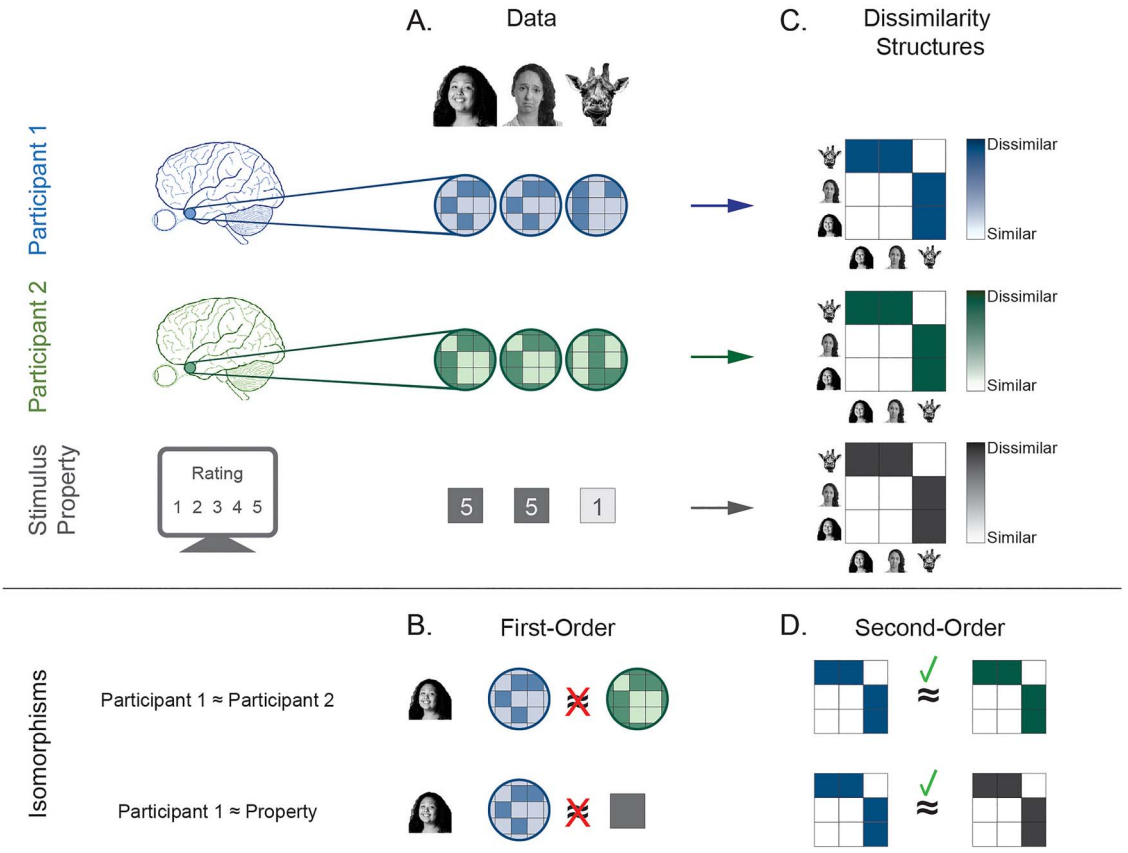



**Fig. 2.** First- and second-order isomorphisms. (A) Neural data from two participants (sets of blue and green squares represent response patterns across voxels) and behavioral ratings (gray) for three stimuli: a happy human face, a sad human face and a giraffe face. (B) Testing for a first-order isomorphism involves directly comparing neural response patterns across people or to behavioral ratings (e.g. testing whether a happy face elicits the same pattern or magnitude of neural responses across participants or whether stimuli that elicit a higher rating also elicit greater neural responses). (C) RDMs capture how, and to what extent, a measure (e.g. responses in a particular brain region, behavioral ratings) distinguishes between stimuli. (D) Testing for a second-order isomorphism involves comparing the relations among stimuli (i.e. comparing RDMs). Here, we see that there are no direct correspondences (i.e. first-order isomorphisms) between neural responses across people or between neural responses and behavioral ratings (B), but there are second-order isomorphisms (D). In other words, even though participants 1 and 2 show different neural response patterns, the highlighted brain region in both participants treats the two human faces as similar to one another, but distinct from a giraffe face. In the same way, even though the amount of activity in this brain region does not directly correspond with behavioral ratings (B), the behavioral data show that the two human faces are identical to each other, and quite distinct from the giraffe face along the rated dimension, mirroring the neural dissimilarity structures. RSA tests for second-order isomorphisms, thus facilitating comparisons across people, modalities of measurement and models, when direct correspondences are impractical or impossible to establish.

responses in that brain region (i.e. using supervised rather than unsupervised learning; Tables 1 and 2). Discovering the structure from the data in this manner is referred to as 'unsupervised learning'. This opportunity to see how the data are naturally structured without imposing a model means RSA can be more data-driven than decoding analyses (Table 1).

**Table 2.** Glossary

| Term | Description |
|---|---|
| Cross-classification | Training a machine learning algorithm on data from one condition and then testing it in another. |
| Cross-validation | The process of iteratively training a machine learning algorithm on training data and testing the resulting model on testing data. Tests for generalizability of the model. |
| Decision boundary; hyperplane | A surface in the representational space that separates samples associated with different labels (used, e.g. in SVM classification). |
| Feature | The units of observation that form a single sample. The input variables when using decoding (supervised learning) to predict labels. Typically voxels. |
| Feature weights; model parameters | The variables that a machine learning algorithm learns in the training data. A set of values assigned to the features, which are used to compute the predicted labels for samples in the testing data. |
| Fold[a] | A division of data into training data and testing data. |
| Hyperparameters | Settings that define how a machine learning algorithm learns the best feature weights (i.e. parameters) in the training data. |
| Labels | The outcome variable that a machine learning algorithm tries to predict. In classification, this is discrete. In regression, this is a continuous measure.<br>Also referred to as 'targets'.<br>Typically condition names or stimuli. |
| Margin | The distance between the support vectors (closest samples with different labels) and the decision boundary. SVM algorithms try to maximize this margin. |
| Representational space | A space defined by features. For *m* features, the representational space will have *m* dimensions. Samples' coordinates in this space are defined by the value of each feature (e.g. the response in each voxel). |
| Sample | A single observation. A set of features associated with a label. Typically a single neural response pattern from a single trial or condition. |
| Supervised learning | A machine learning method of finding a mapping between features and labels based on training data and then testing the trained model on separate testing data.<br>Decoding analyses are types of supervised learning. |
| Support vectors | The samples (vectors in the representational space) that define the hyperplane in SVM learning. These are the correctly classified samples nearest the decision boundary. |
| Testing data | A subset of the full dataset, completely independent from the training data, that is used to test a trained machine learning algorithm. Samples are provided without labels to test how well the algorithm can predict the correct labels. |
| Training data | A subset of the full dataset that is used to train a machine learning algorithm. Samples from the training data are provided with their correct labels so that the algorithm can learn a mapping between features and labels, which is then used to try to predict the labels of the testing data samples. |
| Unsupervised learning | A method of discovering the underlying structure of data without considering labels. The experimenter may then consider the labels after learning is complete (e.g. to name dimensions in MDS based on the data labels).<br>MDS and clustering analyses are types of unsupervised learning. |

[a]Note that 'fold' can have multiple meanings. Some refer to the individual partitions of a dataset used for model training and testing as 'training folds' and 'testing folds', respectively. The act of partitioning the data into these sub-datasets (i.e. for training and testing) is also referred to as 'folding'. Finally, 'fold' may also refer to just a single partitioning of the data into training and testing sub-datasets. To avoid ambiguity, in this paper, we only use 'fold' in the latter sense.

***Comparison across individuals.*** Finally, RSA provides a way to examine and compare neural data from one participant to neural data from other people that is not affected by response idiosyncrasies across participants (e.g. between-participant differences in the neural response patterns themselves). That is, two people may have different response patterns to a set of stimuli, but similar RDMs (Figure 2). This is particularly relevant to many fMRI researchers, given that aligning fMRI data across subjects to a common anatomical template (e.g. Talairach space, Talairach and Tournoux, 1988; MNI space, Mazziotta *et al.*, 1995) may not be sufficiently precise to align fine-scale patterns across people: whereas coarse-scale functional organization is relatively consistent across people, fine-scale spatial patterns may lack substantial person-to-person correspondence (Kriegeskorte and Bandettini, 2007). For example, although the FFA may be in a relatively consistent location across people, fine-scale response patterns within the FFA likely vary substantially across people. By abstracting away from the response patterns themselves (and, thus, from the spatial layout of the data), RSA provides a way to compare representations across models, people and modalities of measurement when direct correspondences are difficult or impossible to establish (Kriegeskorte *et al.*, 2008a).

## Spatially mapping effects

An important aspect of most neuroimaging studies involves locating an effect within the brain. This is done in two primary ways: region-based analyses, in which regions of interest (ROIs) are defined a priori (Figure 1A and C), or point-by-point analyses (i.e. voxel-wise for univariate analyses or searchlight for MVPA; Figure 1B and D). In the searchlight method, the researcher typically defines a region as a sphere of a given radius (often 8–12 mm) surrounding each voxel in the brain and tests if and how the activation pattern within this sphere differs across conditions (Figure 1D). The resulting test statistic (e.g. a *t*-value) is then assigned to the center voxel. This results in a map showing the extent to which the sphere centered on each voxel distinguishes between conditions. It is important to note that because of this mapping process in searchlight analyses (unlike voxel-wise analyses, which can be thought of as a searchlight with a radius of 0 voxels), the significant voxels in the resulting map do not directly correspond to the voxels that differentiate between conditions, but rather are voxels around which spheres (comprised of additional voxels) differentiate between conditions.

## Benefits of MVPA

Analyzing multivoxel patterns, rather than the overall response magnitudes, can provide additional insight into how a brain region processes information. This is achieved by considering reliable (and potentially submaximal) response patterns, examining which physical or conceptual properties that region uses to distinguish between stimuli, and elucidating the functional significance of overlapping responses to distinct stimuli. Each of these benefits is considered in more detail below.

***Sensitivity to information carried in distributed response patterns.*** A variety of information is carried in the spatially distributed pattern of responses in a brain region, not just the overall response magnitude of that voxel or region. That is, even if a single voxel does not significantly change across conditions when considered on its own, the signal variability of this voxel may still contribute to a reliable response pattern that does discriminate between conditions. These voxels are washed out or excluded in univariate approaches, but fully considered in MVPA. Indeed, research using direct measures of neuronal activity demonstrates that the brain encodes many types of information in distributed neuronal population codes (i.e. neuron activity patterns; Pouget *et al.*, 2000), from low-level sensory information (Uchida *et al.*, 2000) and motor plans (Georgopoulos *et al.*, 1988) to high-level category information (Kiani *et al.*, 2007) and subjective decisions (Kiani *et al.*, 2014). Importantly, most social and affective neuroscience research in humans uses indirect and relatively spatially coarse modalities of measurement (e.g. fMRI) rather than directly measuring neuronal firing. Fortunately, work comparing the information content of multi-neuron and multivoxel population codes suggests that applying MVPA to fMRI data can extract much of the same information carried in actual neuronal population codes (Kriegeskorte *et al.*, 2008b; cf. Dubois *et al.*, 2015). In other words, even though each voxel contains tens of thousands of neurons, the same principles that apply to distributed neuronal population codes still apply at the level of distributed voxel patterns.

Recently, researchers have started to combine these methods by training a decoding algorithm on test data to find the feature weights (e.g. a linear transformation of the response pattern) that best distinguishes the conditions of interest. Using the weighted response patterns from the test data, you can then use RSA to compare the transformed neural data to specific models, thereby potentially increasing sensitivity to information carried in those response patterns. This method, known as 'mixed RSA', has been best developed in regions that process lower-level visual information (Khaligh-Razavi *et al.*, 2017), but may offer benefits to social neuroscientists, especially as the field continues to develop better computational models of brain regions that process social and affective information.

***Uncovering the information content of brain regions.*** MVPA provides a rich characterization of how particular brain regions organize information (i.e. the distinctions that brain regions make about various classes of stimuli). For example, Peelen *et al.* (2010) found that the medial prefrontal cortex (mPFC) and left superior temporal sulcus (STS) have similar univariate responses to different emotional cues conveyed using a variety of modalities (e.g. body, face, voice), but encode which emotion is being expressed (anger, disgust, fear, happiness or sadness) in multivoxel response patterns that are consistent irrespective of how a given emotion was conveyed (i.e. in the face, voice or body). These results provide evidence that the left STS and mPFC signal emotion in a modality-independent manner, organizing sensory information in terms of its abstract emotional value. Examining how relations among response patterns change (i.e. what kinds of stimuli are treated as distinct from one another and to what extent they are differentiated) as they progress throughout different stages of processing can provide insight into how information is transformed as it progresses from early sensory cortex (where neural population codes reflect low-level sensory properties) to later stages of processing (where neural population codes reflect higher-level, more abstract categories). More generally, examining the relative discriminability of response patterns associated with a set of stimuli or cognitive states can shed light on the contributions different brain regions make to neural information processing.

***Testing the significance of overlapping activations across tasks, domains and contexts.*** Numerous studies have now shown an overlap in the ventral striatum between social (e.g. praise, happy faces) and non-social (e.g. money, juice) rewards (e.g. Lin *et al.*, 2012; Bhanji and Delgado, 2014). Is that brain region really encoding both types of rewards in the same way? While it is difficult to test this with univariate analyses, MVPA can elucidate differences by examining the extent to which similar patterns of activity across voxels are elicited by the two kinds of reward. Indeed, Wake and Izuma (2017) found that social and monetary rewards elicit similar multivoxel response patterns as well, suggesting that they might rely on a shared neural mechanism for reward representation. In other cases, MVPA has revealed that functional overlap was likely reflective of distinct underlying mechanisms (Peelen *et al.*, 2006; Downing *et al.*, 2007). It is important to note that similar response patterns do not mean two stimuli or conditions are represented in exactly the same way, nor do discrepant response patterns mean that two processes are psychologically unrelated. Rather, these similarities are relative, and by assessing these relative differences across conditions or stimuli, we can gain a more detailed and nuanced picture of a brain region's functional role and response profile.

Thus, analyzing fMRI responses within particular brain regions in terms of distributed spatial patterns, rather than the overall magnitude, can be useful in uncovering how those regions process information. Whereas these fine-scale neural response patterns are thought to be relatively idiosyncratic to individuals, MVPA can also be applied at a relatively coarser spatial scale (e.g. on data that have been spatially smoothed and aligned to common anatomical templates) to reveal signatures or 'biomarkers' of particular states of mind that generalize across individuals (e.g. Wager *et al.*, 2013; Chang *et al.*, 2015; Woo *et al.*, 2017). That is, both univariate and decoding analyses may be used to interpret what a brain region is doing (e.g. does a region respond to/distinguish stimuli based on a given feature?), while decoding analyses can also be used for the sole purpose of out-of-sample prediction (e.g. to identify biomarkers; Hebart and Baker, 2018). In addition, as described in more detail earlier in this manuscript, methods such as RSA allow researchers to compare data from different models, neuroimaging techniques, individuals and even species (Kriegeskorte *et al.*, 2008a). Thus, MVPA is a tool that offers diverse opportunities for gaining insights into neural information processing over and above what can be achieved with traditional univariate approaches.

### A note on terminology

It is important to mention that the methods discussed here were not invented for, and are not isolated to, the analysis of fMRI data. What we refer to as 'decoding analyses' are also referred to as (supervised) 'machine learning' and 'statistical learning' (Hastie *et al.*, 2017) and are widely used across academic disciplines and industries. Likewise, the analysis of similarity structures to characterize mental representations and compare data across modalities or people has a long history in psychological research (Shepard, 1963, 1964; Shepard and Chipman, 1970; Shepard and Cooper, 1992). 'MVPA' simply refers to using these data analytic techniques, which have long histories of their own and wide-ranging applications, to analyze multivoxel response patterns. Given that any sort of data could be analyzed using decoding or similarity analyses, it is of course true that the same data analytic techniques used in MVPA could be used to analyze other types of fMRI data (e.g. region-averaged time series, Yeshurun *et al.*, 2017; inter-subject similarities of response time series, Parkinson *et al.*, 2018; patterns of functional connectivity, Shirer *et al.*, 2012) or data from other neuroimaging modalities (e.g. fNIRS, MEG, EEG; Wang *et al.*, 2004; Wardle *et al.*, 2016; Emberson *et al.*, 2017). That said, given that MVPA specifically refers to the analysis of response patterns across fMRI voxels, in the current paper, we focus on the application of decoding and similarity analyses to study multivoxel response patterns. When learning more about these methods, we encourage the readers to seek out the excellent training resources and reference texts that focus on these statistical techniques in a manner that is not specific to fMRI data (e.g. Hastie *et al.*, 2017).

## Practical implementation

Here, we discuss general design and analytical considerations, such as how stimuli are presented within and across fMRI runs, when and how much to smooth, algorithm choices, hyperparameter tuning, and feature selection. Importantly, choices regarding these considerations will be greatly impacted by your research question and paradigm and thus require an in-depth understanding of both your research topic and the data analytic approach you are using.

### Representational spatial scale

The spatial scale at which the phenomenon you are researching is represented will have important consequences for many methodological decisions, including those related to both study design and data analysis. For example, given that person-to-person correspondences in functional brain organization are more limited at finer spatial scales (e.g. voxel-to-voxel), compared with coarser ones (e.g. region-to-region), when analyzing information thought to be carried at a relatively fine spatial scale (e.g. relatively nuanced visual distinctions, such as the encoding of facial identity in regions of the temporal cortex, any information likely carried in neuronal population codes based on related literature; Kriegeskorte *et al.*, 2007, Nestor *et al.*, 2011), it would make sense to conduct all aspects of analyses that involve comparing neural response patterns (e.g. decoding, computing neural RDMs) within each participant, in their native brain space (i.e. without alignment to a standard anatomical template to avoid the distortion and averaging that spatial normalization can introduce). Relatedly, very little or no spatial smoothing of fMRI response pattern data is preferable in cases where the information of interest is carried in fine-grained response patterns, which may be idiosyncratic to individuals (that said, functional alignment methods, such as hyperalignment and the shared response model, can make response patterns more comparable across individuals and, thus, improve between-subject decoding of such patterns; Haxby *et al.*, 2011; Chen *et al.*, 2015). On the other hand, more extensive spatial smoothing of evoked neural response patterns may be beneficial in cases where the information of interest (e.g. the affective state a participant is experiencing, other information that is likely encoded in the relative activity of multiple brain regions) is thought to be signaled in coarser spatial patterns, which may be more generalizable across people. If you are unable to estimate the granularity of the neural representation of interest, even after looking at the past literature using other approaches (e.g. direct neuronal recordings, meta-analyses) or considering your experimental goals (e.g. building a biomarker *vs* determining the distinctions made by a brain region), it may be useful to explore multiple possibilities and report all results.

The spatial scale at which the phenomena you are studying are represented has consequences not just for data analysis but also, relatedly, for experimental design. When studying phenomena signaled by relatively fine-scale neural response patterns, which may be relatively idiosyncratic to each participant, more trials per participant are often needed to obtain robust estimates of response patterns and to have sufficient data for decoding analyses (e.g. since classification analyses would typically need to be performed independently within each participant). On the other hand, effects carried at coarser spatial scales may benefit from larger samples, potentially with fewer trials per participant, since, in such cases, response patterns can be estimated by aggregating across participants. The former approach (i.e. many trials per participant with a smaller sample) could be used to treat each participant as their own experiment and all other participants as replications of that experiment, similar to approaches often used in psychophysics (Smith and Little, 2018). One could also test if summary statistics (e.g. decoding accuracies for corresponding brain regions) reliably exceed a given value (e.g. the level of accuracy expected based on random chance) across participants. This approach, like the latter approach described above (i.e. fewer trials per participant with a larger sample), is more similar to most social psychology research and tests the entire sample against the null, rather than each participant.

As evidenced by the fact that a single factor (the spatial scale of representations of interest) can have widespread implications for both the study design and analysis, it can be difficult to provide general recommendations for best practices, as choices are highly dependent on the specific study. As such, when making design and analysis decisions, it is very important to carefully consider your topic of interest, how related research has approached this topic and the tools and methods you are using. Below, we discuss additional important considerations for the design and analysis of MVPA studies.

## Design considerations

Since MVPA tests if patterns of activation reliably and systematically differ across conditions, it is necessary to ensure that the experimental design is optimized to obtain reliable neural activation patterns for each condition. This requires minimizing noise and even sampling of noise across conditions.

***Minimizing noise.*** *Including a sufficient number of trials per condition.* Just as it is important to have enough participants to minimize the impact of noise associated with small sample studies, an MVPA experiment must include enough samples of each condition (e.g. trials) to reliably calculate the typical activation within each voxel per condition. As alluded to above, presenting each participant with as many examples of each condition as possible is particularly important in cases where pattern analyses are to be carried out within each participant before aggregating results across participants. The exact number of trials necessary will differ based on the signal-to-noise ratio (SNR) of the experimental design, although, generally, more is better, as long as there is sufficient separation of trials to model the evoked neural response. This is particularly important when trials are not averaged to create a single response pattern per condition. The SNR is affected by many factors, including the scanner, acquisition sequence and brain region. The granularity of the phenomena of interest and the relative distinctiveness of experimental conditions being compared are also important considerations. For example, making subtle within-category distinctions, such as between different facial identities, would generally require more examples/trials than making more dramatic between-category distinctions, such as between human and dog faces.

While we discuss the opportunities for optimizing design, it is difficult to give general recommendations for the number of trials needed. Just as you would not recommend that 20 data points are sufficient for a *t*-test without knowing how large an effect is, we cannot recommend a specific number of trials without knowing how robust the difference is between the multivoxel response patterns evoked by a set of conditions. As mentioned above, more samples will generally be needed for algorithms to learn subtler distinctions between conditions; however, the optimal trade-off between the number of trials and the interstimulus interval may depend on your analytic approach (e.g. greater spacing may benefit single-trial-level analyses, while more tightly packed trials may benefit averaged category-level analyses; see Zeithamova *et al.*, 2017 and Kriegeskorte *et al.*, 2008a; for more detailed discussions of design optimization).

Another key determinant of the number of samples needed for decoding is the number of features (typically, voxels). A rule of thumb for cases where the actual separability of categories is unknown is that the number of samples in a training dataset should be at least 5–10 times greater than the number of features (Jain and Chandrasekaran, 1982). In practice, the number of samples that is feasible to acquire in fMRI research is considerably smaller than this (e.g. even in a small, 100-voxel ROI, having 5–10 times more samples than features would translate into 500–1000 trials). Since it is not possible to scan a participant indefinitely, the best rule of thumb we can provide is to maximize the ratio of samples to features. One way to maximize the number of samples is to include as many trials and runs as the study parameters allow. Another strategy is using methods that facilitate the aggregation of data across participants (e.g. using functional alignment to match fine-scale response patterns across participants; using coarse-spatial scale summaries of activity—such as patterns of activity across regions rather than across individual voxels, if appropriate). The ratio of samples to features can also be increased by reducing the number of features (e.g. through feature selection, dimension reduction, etc.). Given that in neuroimaging, studies are often underpowered, we suggest pursuing strategies like those outlined above to maximize the ratio of samples to features. It is worth noting that some algorithms, such as support vector machines (SVMs), work relatively well with large numbers of features and that methods other than decoding, such as RSA, may need fewer samples per condition since they do not entail partitioning the data into training and testing sets.

*Other ways to mitigate noise.* In addition to including more trials, another technique to minimize the impact of stimulus-unrelated noise is to have many short runs, rather than fewer long runs. Since noise is independent across runs, averaging across these patterns can help achieve a robust estimate of the distributed neural response patterns associated with specific experimental conditions, which can improve estimations of neural RDMs and reduce the effect of noise on pattern classifiers (Coutanche and Thompson-Schill, 2012). This also minimizes biases that arise in within-run comparisons (Mumford *et al.*, 2014).

***Even sampling of noise.*** One potential pitfall of fMRI that can be particularly impactful on MVPA occurs when noise systemat-

ically covaries with certain conditions. This could be external noise (e.g. differences in instrument-related noise between runs or between the start and end of the experiment or run) or stem from the participant's behavior (e.g. one condition causes the participant to move more, causing more fluctuation in signal) or cognition (e.g. reaction time differences between conditions; Todd *et al.*, 2013). Such confounds can lead to seemingly significant results or other erroneous conclusions and should be avoided as much as possible. Thus, (i) all conditions should ideally be included in every run to sample variation in the signal across runs as evenly as possible, (ii) all conditions should have the same number of trials in every run, (iii) the order of these trials should be optimized for your psychological question and to minimize order effects (discussed in more detail below), and (iv) special attention should be paid to avoid confounds in experimental designs (e.g. differences in task difficulty/reaction times or in extraneous stimulus properties between conditions).

MVPA can be more sensitive to noise than univariate analyses, since decoding analyses will pick up on any information that discriminates between conditions, including differences in neural patterns elicited by confounds in the stimuli (e.g. visual characteristics unrelated to one's research question that systematically differ between conditions) or instrument-related noise (e.g. uneven distribution of conditions across the scan). As such, traditional methods for sampling noise evenly may not be sufficient or appropriate (see Görgen *et al.*, 2018 for further discussion and methods for detecting design confounds). One common source of instrument-related noise in fMRI experiments is scanner drift (slow changes in signal throughout a run). Often, an effective way to reduce the effect of scanner drift is to minimize autocorrelation in stimulus ordering by randomizing the order of events within a run (Mumford *et al.*, 2014). However, randomizing events within a run is not always sufficient to negate order effects (Cai *et al.*, 2019) or to produce the most efficient design (Buračas and Boynton, 2002). While a detailed discussion of ordering strategies is beyond the scope of this paper (yet is very important and should be carefully considered when designing an MVPA fMRI study), there are other resources that discuss various options in detail (de Bruijn cycles, Aguirre *et al.*, 2011; m-sequences, Buračas and Boynton, 2002). Generally, strategies for mitigating order effects do not substantially differ between studies designed for univariate or pattern analyses, but MVPA results are especially vulnerable to the impact of confounding noise (Todd *et al.*, 2013), so it can be especially beneficial to consider these issues during study design.[1]

## Analytical considerations

***Analytical considerations specific to decoding analyses.*** There are many different machine learning algorithms that may be used in decoding analyses. These algorithms differ in terms of how they systematically assign labels (e.g. condition names in classification; values in regression) based on the training data and thus may significantly influence your results (Douglas *et al.*, 2011). While a detailed review of these methods is beyond the scope of this article, here, we will briefly discuss a few algorithms that are commonly used in fMRI studies.

*Types of algorithms.* Most algorithms learn to distinguish between conditions by placing weights on each feature (typically, voxels are the features used) that best predict the correct labels corresponding to the data. In commonly used linear classification algorithms, feature weights describe the projection of each sample (e.g. each observed multivoxel pattern) onto a decision boundary that best separates the data into the correct conditions. In regression, weights are used to predict the value of a continuous variable (e.g. how old a face is) as a function of the features (e.g. as a weighted combination of responses at each voxel). The best weights for predicting the labels are determined during model training. Next, during model testing, new multivoxel patterns are given to the trained model, which then tries to predict which category or value corresponds to that sample. The model's predictive performance in the test data indicates the extent to which the neural response patterns distinguish between the experimental conditions.

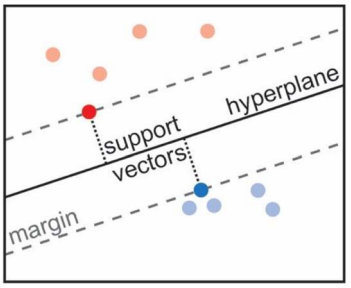
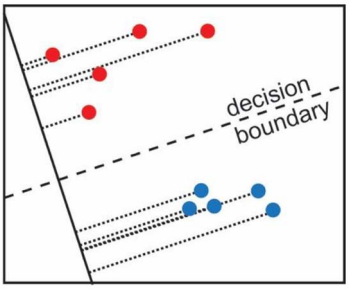
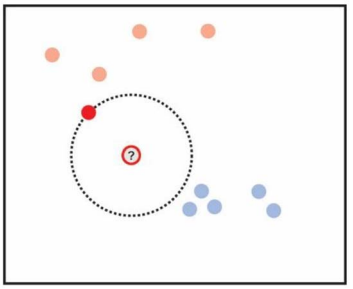
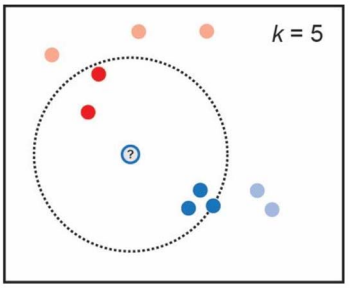
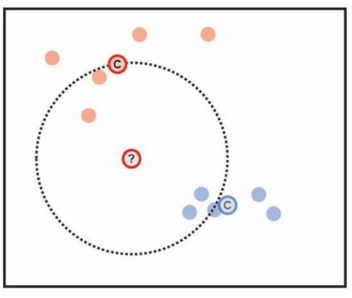
Two commonly used linear classification algorithms are linear SVMs and linear discriminant analysis (LDA; Table 3), which both determine the weights for each feature that linearly project the data onto a decision boundary that maximally separates the data into the corresponding categories. Very generally, LDA (which assumes that all variables are normally distributed and have the same variance) does so by finding the solution that maximizes the between-category variance relative to the within-category variance. Linear SVM learning does so by attempting to find a hyperplane (i.e. boundary) that separates the data points according to their category. In cases where categories are perfectly separable, this is achieved by maximizing the hyperplane's margin (i.e. the distance between the nearest data points of either category and the decision boundary). As such, only samples closest to the decision boundary (also referred to as the 'support vectors', because they support the dividing hyperplane) are what define the decision boundary in linear SVMs. Consequently, samples far from the decision boundary (e.g. multivoxel patterns that clearly belong to either category or outliers) have no effect. In contrast, when using other methods, such as LDA, all samples impact the definition of the decision boundary (Table 3).

Some other commonly used classification algorithms simply assign each sample the label of the closest (e.g. based on correlation or Euclidean distances between multivoxel response patterns) sample from the training data (nearest neighbor classification), or of the majority of the *k*-nearest samples in the training data (*k*-nearest neighbor classification) or of the category whose multivariate mean (i.e. centroid) is closest in the training data (Table 3; Haxby *et al.*, 2001). Because nearest neighbor approaches base distances between patterns on all features, and weight all features equally, these algorithms often work best after having performed feature selection (described in more detail below) so that distances are calculated (and thus, classifications are made) based on the most informative features.

Importantly, one should be very cautious in interpreting the feature weights given by a trained classifier or regression model (e.g. which voxels the algorithm most considered within a given

1 MVPA decoding methods could successfully distinguish between conditions because of differences in the multivariate means of response patterns between conditions or because of differences in the variability of those patterns across trials. Perhaps counterintuitively, the latter could constitute signal rather than noise (e.g. if this information is read out and used by other brain areas). However, differences in response variability between conditions could also easily arise from extraneous factors of the sort described in this section (e.g. differences between conditions in how equally trials are distributed across runs). Thus, it is critically important to minimize factors that could lead to differences in both the mean and variability of response patterns between conditions. See Hebart and Baker (2018) for an in-depth discussion of this issue.

**Table 3.** Overview of a few simple classification algorithms

| Algorithm | Description | Visualization |
|---|---|---|
| Support vector machines (SVMs) | Attempts to draw a decision boundary (hyperplane) that separates categories by maximizing the margin between the hyperplane and the support vectors (i.e., the closest samples that were accurately classified in each category). |   |
| Linear discriminant analysis (LDA) | Separates categories by finding the axis that maximizes the between-category variance relative to the within-category variance (i.e., the distance between category distributions). |   |
| Nearest neighbor classification | Assigns each test sample the label of the closest training sample. |   |
| *K*-nearest neighbor classification | Assigns each test sample the label that is held by the majority of the *k*-nearest training samples. |   |
| Centroid-nearest neighbor | Assigns each test sample the label that is held by the closest centroid (i.e., multivariate mean) of the training samples. |   |

brain region or searchlight sphere). Since voxels are not independent of each other, a voxel may receive a low (or zero) weight from some algorithms because it does not contribute to patterns that distinguish between conditions *or* it may simply carry redundant information. That is, sometimes, voxels are assigned weights of zero if another voxel within the tested region was assigned a larger weight and carries the same information; other times, the weights of correlated voxels decrease as a function of the number of voxels that are correlated with one another (Pereira *et al.*, 2009). The relative voxel weights will also be impacted by whether or not data have been normalized for each voxel (i.e. to have a mean of 0 and s.d. of 1) prior to analysis. Thus, it is often difficult or misleading to interpret the resulting weights.

*Overfitting.* Assessing model performance only in the previously untouched test data is always necessary, and is *particularly* critical in fMRI, because, typically, there are far more features than samples. Indeed, whereas it is common to have hundreds or thousands of voxels in an ROI, it is rare to have such a high number of trials for each participant. Consequently, it will typically be possible to find some combination of feature weights such that the model performs very well within the training data, without generalizing well to new data—i.e. to overfit the model to the data and generate false positives. In addition to having more features (typically, voxels) than samples (typically, trials), another factor that can heighten the risk of overfitting is how flexibly the model can conform to the shape of the data. As such, it is often preferred to use relatively simple (e.g. linear, rather than non-linear) models in decoding analyses of fMRI data.

*Hyperparameter tuning.* Another important analysis consideration pertains to hyperparameter tuning. Whereas the values of model parameters (also known as feature weights) are learned from the training data, additional parameters, known as 'hyperparameters', control how this learning works and, thus, need to be set before a model is trained. The hyperparameters that need to be set, and the reasonable ranges of values to consider, vary by algorithm. For example, in *k*-nearest neighbor classification, the number of neighboring patterns to consider when labeling a test pattern (i.e. *k*) is the relevant hyperparameter to set. Similarly, a linear SVM has a regularization hyperparameter, *C*, which controls the trade-off between maximizing predictive accuracy in the training data and minimizing the norm of the feature weights (i.e. maximizing the margin of the hyperplane). Hyperparameters should be tuned within the training data to ensure that the model performs well on the test data. If you do not explicitly set values for model hyperparameters, most software packages will use default values (e.g. the default value for *C* for SVMs is often 1), but it is not possible to know in advance the best values to use for a particular problem. The process of finding the hyperparameters that result in the best performance for a model in a particular dataset is known as 'hyperparameter tuning'. By tuning the hyperparameters, researchers can determine the ideal hyperparameters in a data-driven manner without basing such decisions on the same data on which inferences will be made, thereby increasing sensitivity while reducing false-positive rates. In practice, this approach can facilitate an exploratory but constrained approach; for example, one could pre-register the set of hyperparameter values to be considered while still using the (training) data to calibrate the eventual choice of those values to the current dataset.

The simplest way to perform hyperparameter tuning is via a 'grid search' in which a model is repeatedly trained and tested using all values (or all possible combinations of values, in cases where a model has multiple hyperparameters) from a user-specified list (or 'grid'). The process of hyperparameter tuning must be performed separately within each fold's training dataset (i.e. nested cross-validation; Figure 5), which can potentially lead to different hyperparameters being selected for different folds. As shown in Figure 5, you can divide training data further into sub-training and validation datasets (see Data Splitting section) and repeat model fitting in the smaller training dataset with many possible hyperparameter values to determine which hyperparameter values perform best across the validation datasets (still within the training data). Following selection of hyperparameters in this manner (i.e. hyperparameter tuning within the training data), you would then use these values when training the algorithm on the entire training set before finally assessing its predictive performance on the test data.

### *Analytical considerations specific to RSA.*

*RDM distance metrics.* When creating a neural RDM, we must calculate the dissimilarity between every pair of response patterns (see RSA step-by-step instructions below). The way that this is typically done is to first calculate the Pearson correlation coefficient, *r*, which corresponds to similarity, and then convert it into the correlation distance, 1-*r*, which corresponds to *d*issimilarity. Importantly, there are other ways to define the distance between two neural response patterns (Nili *et al.*, 2014; Walther *et al.*, 2016). These include the Euclidean distance (calculated by squaring each voxel's difference value, summing these values, then taking the square root), the Mahalanobis distance (similar to the Euclidean distance with normalization) and classification accuracies from decoding analyses (where a classifier being at-chance at discriminating between two conditions would imply zero distance between those conditions). These various metrics are sensitive to different aspects of the data. For example, when comparing multivoxel patterns to generate a neural RDM, the Pearson correlation distance is only sensitive to the difference in the spatial pattern, but not changes in the overall neural response magnitude, while the Euclidean distance is sensitive to both. They can also provide varying degrees of reliability (e.g. continuous measures are more reliable than discretized measures, such as classification accuracies; Walther *et al.*, 2016).

***General analytical considerations: processing and selecting features.*** In this section, we will discuss the considerations related to pre-processing and voxel selection. The following considerations apply to both decoding analyses and RSA, although the term 'features' specifically refers to the predictors that are used by a machine learning algorithm. For most fMRI studies, voxels (or a transformation of the voxels) are the standard features.

*Which voxel-wise summary statistic to use.* Once the data are pre-processed, a general linear model (GLM) using a hemodynamic response function is typically used to create one contrast map (i.e. image) per condition or stimulus. Each voxel in this contrast map signifies the average estimated level of neural activity that was elicited in that voxel across all trials of this condition (or for that trial, in cases where individual trials are modeled). As with any GLM, a beta- and a *t*-statistic are calculated, either of which may be used as feature measures. Betas are the raw values that come out of the GLM (i.e. a quantification of the

relationship between the hemodynamic response function and the condition). *t*-values, on the other hand, scale these betas by dividing each value by its standard error across trials. Thus, if a particular voxel had a lot of variations in neural activity across trials, the *t*-value penalizes this voxel (i.e. the voxel will have a lower *t*-value) compared to a voxel with the same beta value but less fluctuation across trials (i.e. that voxel will have a higher *t*-value). This scaling of the betas into *t*-values can help with pattern detection (Chadwick *et al.*, 2012).

*How and when to smooth.* Smoothing is a form of spatial averaging that recalculates each voxel's signal by summing the weighted values of its neighboring voxels (these weights and how many voxels are included in the smoothing are determined by the Gaussian kernel). Spatial smoothing in univariate analyses is often performed as part of pre-processing to reduce noise and increase signal detection (i.e. power). This reduces the granularity of the signal patterns, however, and can be detrimental when using MVPA. Thus, it is often recommended to use no or minimal smoothing during pre-processing (Misaki *et al.*, 2013; cf. Op de Beeck, 2010; Hendriks *et al.*, 2017) and only smooth after the first-level pattern analyses are complete. Additionally, since there is less anatomical correspondence across participants than within subjects (and smoothing is more beneficial when there is lower spatial correspondence between images), the unsmoothed (or minimally smoothed) images are often used throughout the first-level (i.e. within-subject) analyses, and then smoothed before the second-level (i.e. group-level) analyses, thereby increasing power to detect convergent results across people. The amount of smoothing is dependent on the type of task or how localized the relevant psychological process is (Gardumi *et al.*, 2016).

*Feature selection.* In most cases, the whole-brain contrast images are masked to remove voxels that are uninformative (e.g. voxels in ventricles; all voxels that are not in a specific ROI). Sometimes, you may want to use a functional mask, based on an independent dataset or meta-analysis (e.g. from neurosynth.org, a separate study or a functional localizer). Feature selection (i.e. selecting specific voxels) is useful in reducing the dimensionality of your data (where the number of dimensions of a multivoxel pattern is synonymous with the number of features/voxels it describes) and increasing sensitivity to the question of interest. Reducing the overall number of features also helps decrease the time required to perform analyses and mitigates the risk of overfitting in decoding analyses.

Importantly, feature selection generally must be defined on separate data (i.e. data not included in training and testing datasets) from that being used for pattern analysis (i.e. no double dipping; Kriegeskorte *et al.*, 2009) in order to avoid false positives due to circular analyses in which researcher degrees of freedom are exploited. For example, it is not appropriate to create an ROI of the voxels that respond to faces based on all runs in your dataset and then use MVPA to test if this ROI significantly discriminates between faces and other images within the same data. Instead, independent subsets of the data (e.g. data from distinct runs or participants) generally must be used for voxel selection and model testing. This is true whether or not feature selection is based on which voxels generally respond most to the stimuli (e.g. all conditions *vs* rest), have stable (i.e. low variance) response patterns within a condition across runs/trials (Mitchell *et al.*, 2008), are more variable across conditions (Pereira *et al.*, 2009) or distinguish between conditions most (De Martino *et al.*, 2008).

In cases where researchers wish to use the same data for both feature selection and decoding analyses, feature selection should be performed independently within the training data for each data fold. Different strategies and/or thresholds for feature selection could be assessed independently within each training data fold (by dividing the training data for each fold into training and validation sets, similar to, and potentially in tandem with, hyperparameter tuning, as described above; Figure 5). It is, of course, important not to try out multiple feature selection strategies on the data on which inferences will be made (i.e. the testing data).

*Dimension reduction.* Whereas feature selection reduces the number of features in a model by selecting a subset of features to include in model training, without changing those features in any way, a related family of approaches, called dimension reduction, reduces the number of features in a model by transforming them into fewer dimensions. For example, principal components analysis (PCA) transforms features into a set of orthogonal values (i.e. principal components), allowing much of the variance in correlated variables (e.g. voxels) to be explained by a smaller number of components. Before model training, you can specify how many components you would like to keep as features in your model or what proportion of the variance you would like the retained components to explain. You can base decisions about how to set these thresholds by exploring different possibilities within the training data, using the same nested data folding techniques described in the above discussion of hyperparameter tuning. Dimension reduction techniques, such as PCA, can be beneficial in moving from a situation common in fMRI studies, where you have far more features than samples, to one where you have substantially fewer features in your model, but still retain the majority of the information contained in the entire feature set. Just as with feature selection, this can be useful in preventing overfitting your model to the training data. In addition, transforming features that are correlated with one another into a smaller number of orthogonal components can be beneficial for improving the performance of algorithms that perform best when features are independent of one another (e.g. naive Bayes, some linear regression algorithms).

## Analytical steps

Now we will discuss how to implement MVPA in your own research. There are several software packages now available to help researchers use MVPA methods, including python-based packages [e.g. Nilearn, which facilitates the use of scikit-learn for neuroimaging data (https://nilearn.github.io), PyMVPA (http://www.pymvpa.org), BrainIAK (https://brainiak.org/)], as well as MATLAB toolboxes [e.g. CoSMoMVPA (http://cosmomvpa.org), Toolbox for RSA (http://www.mrc-cbu.cam.ac.uk/methods-and-resources/toolboxes/license/)]. Each software package differs somewhat in terms of its default methods or parameters and how easily certain tests are run. Below are step-by-step instructions for running RSA and linear SVM classification analysis. For clarity, we will describe the steps in terms of a simple experiment in which participants view faces of humans and dogs of varying ages. The first two steps are always necessary, and the following steps (i.e. Step 3 and beyond) differ based on whether one is conducting (A) classification or (B) RSA.

***First steps.*** *Step 1. Define the conditions.* In our example, we will consider response patterns elicited by four different stimulus

conditions: pictures of baby and adult humans and dogs. Suppose stimuli from each of these conditions were presented multiple times in each of 10 runs. We could model responses to each condition (resulting in 40 samples) or individual trials (since each stimulus was presented more than once per run).[2]

*Step 2. Select the region of interest*. Next, we must decide which regions of the brain to test. The analysis is run on each region separately, irrespective of the number of regions being tested (see above for the discussion of selecting features). That is, if analyzing a single brain region (or set of brain regions as a single ROI), then the analysis is only run once, while a whole-brain searchlight analysis consists of completing the analysis as many times as there are voxels in the brain. In the remainder of this section, all steps before significance testing will be described as if being conducted within a single ROI within a single participant.

For each condition, the voxels within the selected region are systematically rearranged into a vector for each condition, such that the first voxel in the resulting vectors corresponds to the same point in the brain for each condition (Figure 1C and D).

***Classification analysis.*** As described earlier in this manuscript, a classification algorithm is iteratively trained on one subset of the data and then tested on an independent subset of the data via cross-validation.

*Step 3. Data splitting*. The simplest method for partitioning a dataset into training and testing data is the holdout method, in which you select one subset of your data for model training and one for model testing (e.g. use one half of trials for model training and the other half for model testing). While this method is simple and fast, the definition of the training and test sets (i.e. which trials happen to end up in either partition of the data) can be very influential on results. As such, it is more common to use *k*-fold cross-validation, in which the data are divided into training and testing sets multiple (*k*) times, and the training and testing procedure is performed in each subsetting of the data (Table 2, Figure 5). Data within each of the *k* subsets are used as test data once and as training data *k*-1 times. It is often recommended to leave 10–20% of data out of the training set for a given fold (Hastie *et al.*, 2017). For example, using 5-fold cross-validation, data from our 10-run fMRI study would be divided into 5 subsets (e.g. runs 1–2, 3–4, 5–6, 7–8, 9–10), and each subset would be used as testing data once and included in the training data 4 times. Leave-one-sample-out cross-validation is a version of *k*-fold cross-validation where *k* is the total number of samples, and, similarly, in leave-one-run-out cross-validation (Figure 3), *k* is the number of runs in the fMRI study. In cases where pattern information can be aggregated across participants, leave-one-participant-out cross-validation is also an option.

To avoid biasing algorithms toward predicting one particular category, it is important to avoid class imbalance in the training data by including the same number of samples of each category (e.g. stimulus, condition) in the training data. A simple strategy is to have an equivalent number of samples of each category in each run and use leave-one-run-out cross-validation (Figure 3). In our study, this would amount to 10-fold cross-validation, with data from each of our four stimuli present 9 times in each training set and once in each testing set.

If performing feature selection or hyperparameter tuning on this data, then the training data within each fold must be split into sub-training and validation sub-folds (i.e. nested cross-validation; Figure 5). Within each of these sub-folds, the algorithm is trained on the sub-training data and tested on the validation data iteratively to find the most predictive features and/or optimal hyperparameters. After this iterative testing is completed on every sub-fold within the training data, the best hyperparameters (and features, if conducting feature selection within the training data) are selected to be used when training the algorithm—i.e. when determining the feature weights for that fold (see Step 4). Note that this process may result in different features, feature weights and hyperparameters being used in each fold.

*Step 4. Train model*. Within each training set, we label the samples with their correct labels and give this information to our algorithm. Essentially, the model considers each multivoxel pattern as a point in a multidimensional representational space, such that each voxel corresponds to one dimension (Figure 3). That is, the coordinate of a sample in this space is defined by each voxel's value (i.e. the *x*-value is the magnitude of voxel 1, the *y*-value is equal to the magnitude of voxel 2, etc.). If we have *m* voxels in our sample, then, we have an *m*-dimensional space. The algorithm tries to select model parameters such that samples are most often assigned the correct labels.

*Step 5. Test model*. Once the model has been trained, we give the algorithm the testing data, which is provided without any labels. The model categorizes each of these new samples based on where they fall in the representational space relative to the boundaries that were estimated from the training data (as in SVM learning) or relative to its neighbors in the training data (as in *k*-nearest neighbor classification; Table 3). We then count the number of errors it made in its categorization and calculate the classification accuracy of that model. Although classification accuracy is the most commonly used measure of decoding success in MVPA of fMRI data, other methods, such as the area under the ROC (receiver operating characteristic) curve, can sometimes be preferable (Ling *et al.*, 2003), particularly in cases where some categories are overrepresented in the data, which could cause high classification accuracies to be misleading.

Next the average classification accuracies across testing sets are compared to what would be expected due to random chance (e.g. 50% if one has two equivalently sampled categories—note that small sample sizes may falsely increase chance accuracy; Combrisson and Jerbi, 2015), as described in the next section. If classification is reliably above chance, this suggests that response patterns in this brain region distinguish between the categories or, in other words, that this brain region 'contains information about' these categories.

***Representational similarity analysis.*** *Step 3. Create RDMs*. An RDM represents the relative differences between the stimuli (or conditions). For *N* stimuli, it is an $N \times N$ matrix with each row and column corresponding to a single stimulus. The cell corresponding to row *i* and column *j* is the difference (i.e. dissimilarity, distance) between stimulus *i* and stimulus *j*.

*Step 3a. Neural RDM*. To create a neural RDM, we compare the pattern of neural responses associated with each stimulus (or condition) with every other stimulus' neural response pattern. Therefore, we first obtain a single response pattern for each stimulus (rather than one response pattern per stimulus per run) by averaging the neural response patterns for each stimulus across runs.[2] Neural RDMs are often constructed by calculating the Pearson correlation distance, 1-*r*, between each pair of neural

2 Note that if the trial order is not fully randomized, response patterns should only be compared across runs for both decoding and similarity analyses; see Mumford *et al.* (2014) for further discussion of this issue.

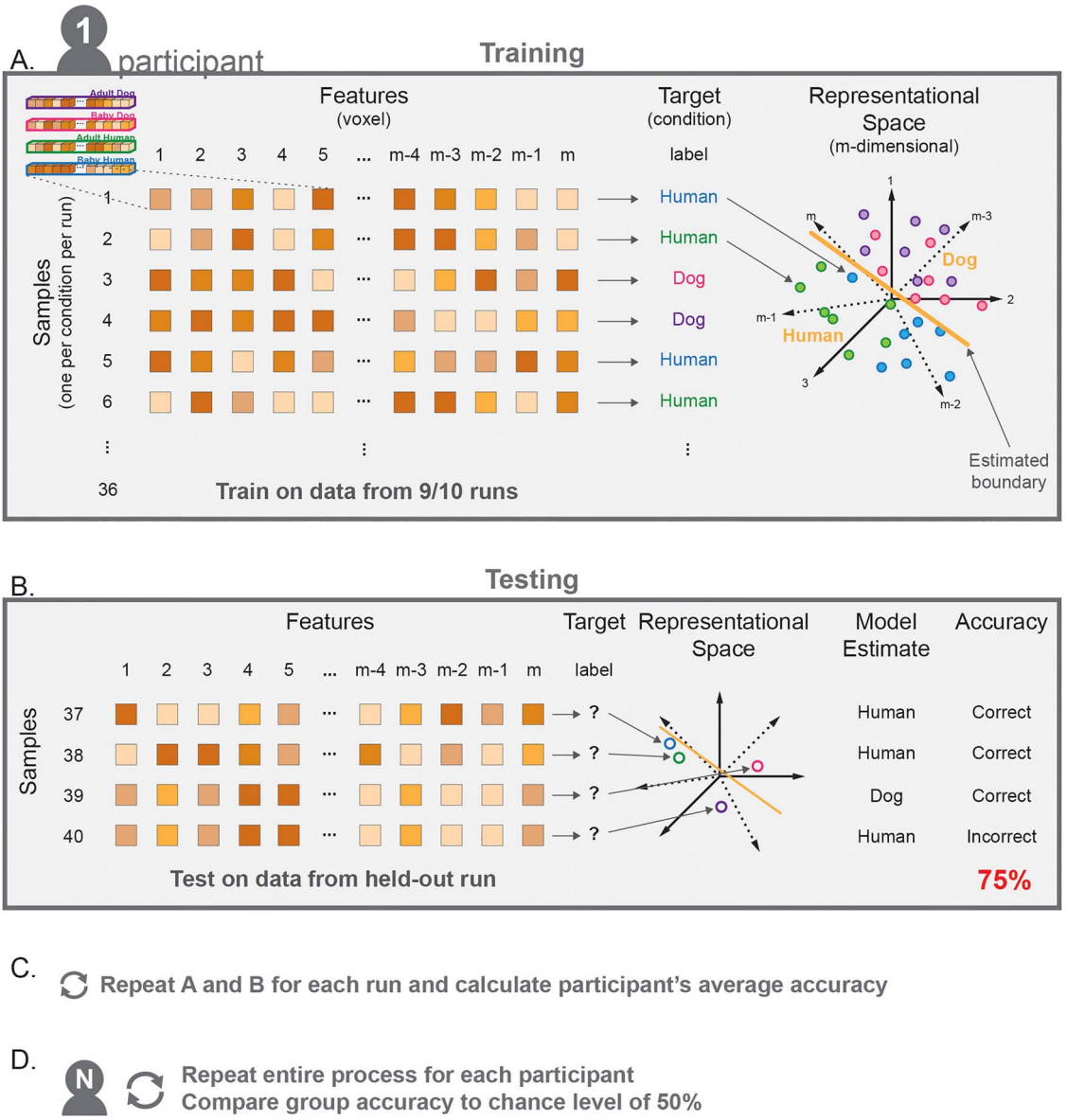



**Fig. 3**. Classification analysis. (A) Within each participant, an algorithm is trained on a subset (here, 9 out of 10 runs) of a participant's data and then tested on a previously unseen subset (here, the heldout run). In the training phase, each sample (here, the multivoxel pattern for each condition in each run) is treated as a point in a representational space. For *m* voxels per sample, there are *m* dimensions in the representational space. Each sample's coordinates are defined by the magnitude of each voxel's response (i.e. voxel 1's response magnitude = coordinate along axis 1, etc.). In many commonly used classification methods, the algorithm then tries to define a boundary (in linear SVM learning, a (*m*-1)-dimensional hyperplane) in the space such that each sample is classified with its correct label (note that the illustration is merely a conceptual example; please see the main text for a more specific discussion of how particular classification algorithms work). (B) After calibrating model parameters on the training data, the algorithm is then fed the testing data, which it has never seen, without the correct labels. Depending on where those samples fall in the representational space, the algorithm classifies them based on the distinctions it has learned from the training data. If a sample was incorrectly classified, it is counted as an error. (C) The average accuracy across all data folds is calculated for each participant. (D) Repeat this process for each participant, and compare the group-level accuracy to what would be expected based on random chance.

response patterns, but there are other distance metrics that may be used (see RDM Distance Metrics section; Figure 4A). Theoretically, the higher the correlation distance, the more the brain region distinguishes between those two concepts.

Once calculated, these distance values are organized into an RDM. Note that this will result in a symmetric matrix across the diagonal because the difference between a baby human and a baby dog is the same as that between a baby dog and a baby human. Note also that the diagonals will all be zero, because each condition is perfectly correlated with itself, and thus has a correlation distance of zero. If comparing two RDMs, this symmetry and diagonal of zeros would falsely increase the correlation between the full RDMs, so only the lower off-diagonal triangles of the RDMs are extracted for further analyses (Figure 4A).

*Step 3b. Non-neural RDMs.* In order to determine what the structure of the neural RDM corresponds to, we can compare it with similarly prepared RDMs from participant data (e.g. perceived age; Figure 4C), objective data (e.g. species; Figure 3E) or data generated from a model (e.g. hypothesized interaction of perceived age and species). In our example, we might want to test if a brain region organizes faces by perceived age. To

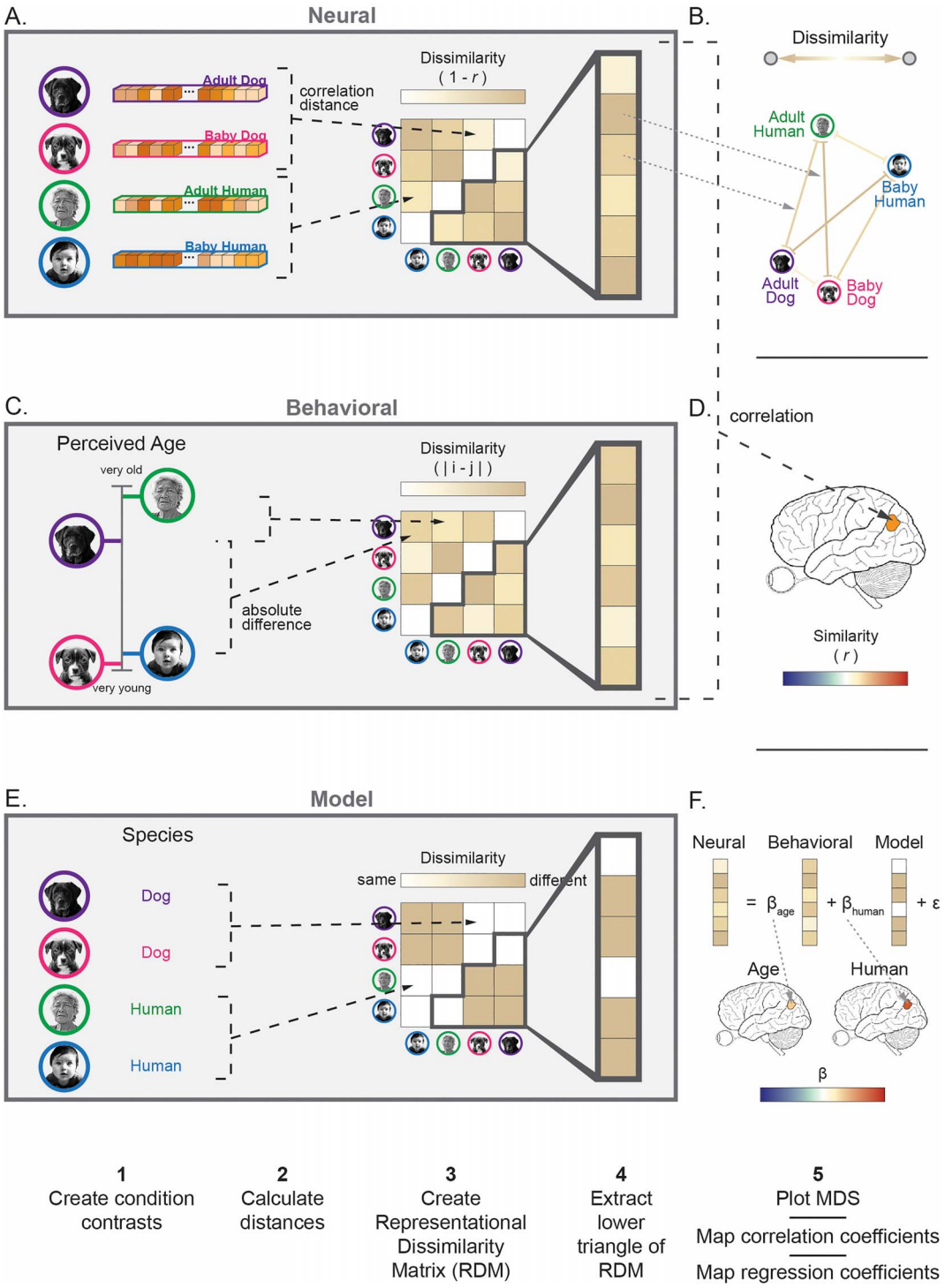



**Fig. 4**. Representational similarity analysis. Representational similarity analysis (RSA) can be used to create (and, often, compare) RDMs summarizing (A) neural, (C) behavioral and (E) model-based data. (A) To create the neural RDM, the patterns of neural responses elicited by each condition within a particular region are compared with each other to estimate their relative distinctiveness (e.g. the correlation distance between them, 1-r). These distances are organized into a neural RDM. Since the RDM is symmetric about a diagonal of zeros, only the lower off-diagonal triangle of this matrix is extracted, which can be (B) visualized in a low-dimensional space using MDS or (D) compared to a behavioral dissimilarity structure. (B) The MDS plot visualizes the dissimilarity structure by plotting conditions that are more similar closer together. Here, we can see that human faces cluster together and are separate (i.e. dissimilar) from dog faces. We can also see that there seems to be an effect of age, such that young faces are similar to each other and separate from older faces. (C) This effect of perceived age can be tested by creating a behavioral dissimilarity structure. This is achieved by finding the absolute difference between the perceived youth of each pair of faces. Again, the lower off-diagonal triangle is extracted. (D) The lower off-diagonal triangles of the neural and behavioral RDMs are compared with one another, often using the Spearman correlation, as it does not assume a linear mapping between RDMs. This correlation coefficient is mapped back into the region, creating a map of how closely the neural data matches the behavioral ratings. (E) A model RDM of species reflects if two pictures are of the same species or not. (F) Multiple RDMs can be included as predictors in a regression, and the resulting betas may be mapped back into the ROI as an indicator of how much that variable predicted the neural data over and above the other predictor(s).

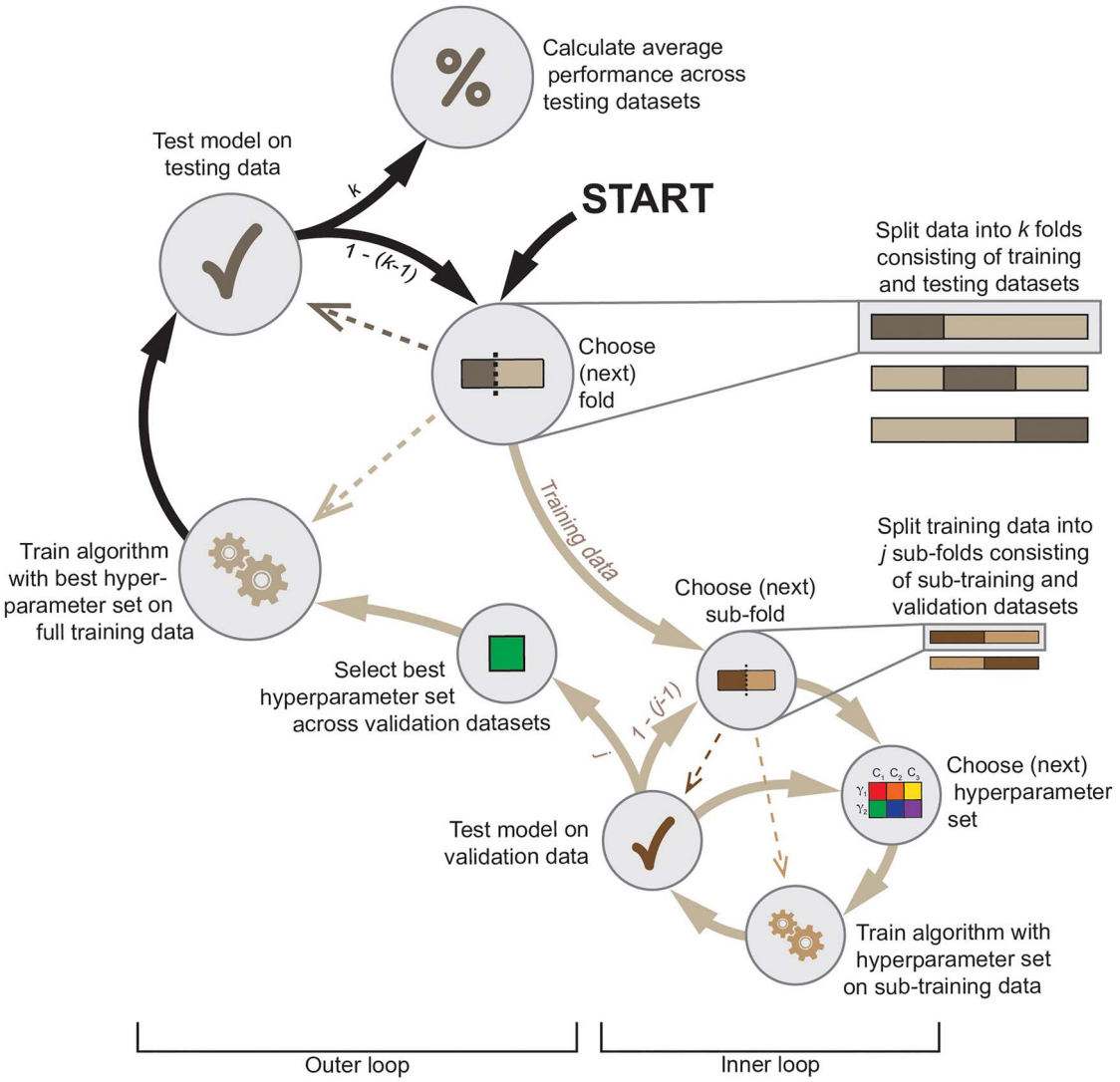



**Fig. 5**. Nested *k*-fold cross-validation with hyperparameter tuning. Cross-validation consists of iteratively splitting data into training and testing datasets, training an algorithm on the training data and then testing the resulting model on the testing data. For each of the *k* divisions of the data (i.e. folds), hyperparameter tuning may be performed within the training data for that fold. To perform hyperparameter tuning, one would further split the training data into a number of 'sub-folds' consisting of sub-training and validation datasets. Within each of these 'sub-folds', the algorithm is trained on the sub-training data and tested on the validation data once per hyperparameter set. Once every unique combination of hyperparameters has been tested in every 'sub-fold', the hyperparameter set with the best performance across the validation datasets (within the training data) is selected. The selected hyperparameter set is then used to train the algorithm on the entire set of training data for that fold. The resulting model is then tested on the testing data in that fold. This process is repeated for each fold (i.e. *k* times). Finally, the average performance of the algorithm across all testing datasets is calculated.

create a behavioral RDM based on the perceived age for each participant, we could calculate the absolute difference between the participant ratings of age for each pair of conditions and then organize this into a matrix and extract the lower off-diagonal triangle (Figure 4C).

*Step 4 (option 1). Compare neural and non-neural RDMs.* Now we can test how well an individual's behavioral RDM predicts that same individual's neural RDM. This is typically achieved by correlating the lower triangles of both RDMs (Figure 4D). Note that since behavioral and neural RDMs likely use different scales, using Spearman rather than Pearson correlations to determine how well they correspond can be beneficial, as this does not assume a linear relationship. If you have multiple predictors (e.g. various stimulus properties) that you would like to test, you can enter them into an RSA regression and test how well each predictor RDM predicts the neural data over and above the other predictors by examining the relevant beta coefficients (Chikazoe *et al.*, 2014; Nastase *et al.*, 2017; Parkinson *et al.*, 2017). For example, we could include both the age RDM (Figure 4C) and another model RDM that reflects whether or not two faces are of the same species (Figure 4E), as predictors in a regression with the neural RDM as the predicted variable (Figure 4F). The beta associated with perceived age would reflect the extent to which age predicts the neural data over and above species. If doing so, it is important to first ensure that the predictor RDMs are independent of one another, as, otherwise, their respective regression coefficients will be difficult to interpret.

*Step 4 (option 2). Visualize RDMs.* RDMs can also be used to visualize the structure of the data. When visualizing RDMs, each cell is often colored based on its value to visually indicate

which conditions are similarly represented (e.g. lighter colors in Figures 2 and 4) and which are represented more distinctly (e.g. darker colors in Figures 2 and 4). Techniques such as MDS can also be used to view the overall structure of the data. MDS plots data points in a low-dimensional space based on how similar they are: two stimuli that elicit similar response patterns are plotted close together, while two stimuli that elicit different response patterns are plotted further apart (Figure 4B). This can help identify how stimuli are organized in the brain (e.g. dog faces may cluster together and may be separate from human faces in a particular brain region).

***Statistical testing.*** After you have completed the above steps, you are ready for significance testing. In many cases, this can be done in the same way as in univariate experiments: results from a searchlight analysis, for example, have a similar data structure to other statistical parametric maps (e.g. one value at each voxel) and can therefore undergo somewhat similar statistical testing. Of course, the exact approach you use will depend on the specifics of your data. For example, it is important to make sure the test you are using is appropriate for the range and distribution of your data's values. Since correlation coefficients and classification accuracy values are bounded by zero and one, it is appropriate to transform them (e.g. using the arcsine transformation) or use non-parametric tests (e.g. permutation testing).

The statistical significance of the results of RSA or decoding analyses can be assessed within each subject or across subjects, and these methods test fundamentally different questions. More specifically, within-subject significance testing of MVPA results assesses if and by how much that participant's data (e.g. correlation of a neural RDM with a model- or behavior-based RDM, decoding accuracy) differ from the null, while between-subject significance testing assesses if and by how much this effect differs from the null across people (similar to univariate analyses on activation-based estimates).

*Within-subject testing.* Generally, within-subject significance testing entails randomly shuffling the labels associated with all samples in a participant's data (e.g. the patterns that are used to create RDMs in RSA and that comprise the training data in decoding analyses) many times. The relevant statistical test (e.g. correlation of the permuted neural RDM with a model RDM, classification analysis) is then performed on each iteration to create a null distribution to which the true test statistic (generated from the non-shuffled data) can be compared. The result is considered significant if it surpasses the critical value in this null distribution (e.g. for $\alpha = 0.05$, one-tailed, it is significant if it is above the 95th percentile).

*Across-subject testing.* Testing the significance of MVPA results across subjects can be accomplished in much the same way that data from corresponding ROIs or statistical parametric maps are tested for significance across participants in univariate studies. It should be noted that if voxel-wise results were obtained for each participant (e.g. searchlight analyses) and decoding analyses and/or RSA were performed on unsmoothed or minimally smoothed data in each participant's native space, each participant's data should be aligned to an anatomical template, and potentially, subjected to additional spatial smoothing prior to group-level significance testing. Note that if using multiple comparison correction methods that require estimating smoothness (e.g. family-wise error correction) of MVPA data (e.g. searchlight maps), it is important to base these estimates on the relevant residuals (e.g. of searchlight results rather than of responses themselves; for an example, see Linden *et al.*, 2012).

## What questions can we ask with MVPA?

In this section, we will discuss the types of research questions that are particularly amenable to MVPA. We will consider how MVPA can be used to answer different kinds of research questions. For clarity, we will continue to consider our example experiment in which participants considered human and dog faces of varying ages while undergoing fMRI.

### Brain-reading

Many researchers are inherently interested in *what* the participant is currently thinking about or attending to (i.e. 'brain-reading') since the ability to determine what cognitive state the participant is in can provide valuable information about where and how information is neurally processed. In our example above, we used classification analysis to determine if people are considering human or dog faces. If we can successfully train a predictive model to decode this information based on response patterns evoked in a given brain region, then there are likely fundamental differences in how human and dog faces (or some covariate) are represented in that region. This can provide valuable insight into how the brain encodes such information.

### Stages of neural processing

We can also examine how information is transformed as it travels through different brain regions. In the Benefits of MVPA section, we discussed a study that provided evidence that the left STS and mPFC represent emotion in terms of its abstract emotional value independent of the modality through which the emotion is expressed (Peelen *et al.*, 2010) and, in general, how we can use MVPA to elucidate how representations change across brain regions as they progress through different stages of processing. For example, this approach can illustrate how information is transformed as it progresses from early sensory cortex (where neural population codes reflect low-level sensory properties, such as modality) to later stages of processing (where neural population codes reflect higher-level, more abstract categories, such as emotional content). This can be tested by comparing each brain region's neural RDM with a model RDM (e.g. whether two stimuli were presented in the same modality) to see which model matches best at each stage. We could also visualize the differences by plotting the neural RDMs with MDS. This could allow us to see how the stimuli are represented at each neural processing stage.

### Underlying neurocognitive mechanisms

Earlier, in the step-by-step instructions, we discussed how RSA may be used in our example to discover that a brain region clusters stimuli by age as well as by species and how to test this using explicit models. That is, RSA allows us to test what type of information a given brain region uses to organize state or stimulus representations. Decoding analyses can be used in a complementary fashion to RSA. Using cross-classification, we can ask if we represent age in the same way across species. Cross-classification involves training a model within one condition (e.g. to distinguish human faces based on age) and then testing the model on another condition (e.g. to distinguish dog faces based on age). If the model can reliably decode the age of dog

faces after being trained on human faces, then there are likely consistent underlying patterns that encoded age across species in this region. This would suggest that people represent the age of human and dog faces similarly at this level of processing.

### Individual differences

Does everyone see and process the world in the same way? Just like univariate analyses, individual differences may be integrated with any type of MVPA to better understand how individuals process information. That is, results from MVPA can be used to predict individual differences. For example, Ersner-Hershfield *et al.* (2009) found that individuals with greater continuity between their present and future selves save more for retirement. One could, for example, follow up on this with an fMRI study by examining the similarity of response patterns associated with participants' present and future selves. The similarity of these response patterns may reflect future self-continuity and thus predict retirement savings. MVPA is a useful tool when studying individual differences because these differences may manifest in the distinctiveness of neural patterns, not just the overall response magnitude of a region.

## Issues in MVPA

Although we have focused largely on the benefits of MVPA, like any data analytic technique, there are important issues and potential pitfalls to consider. For instance, in decoding analyses, it can be difficult to interpret anything about the model itself, beyond the yes or no question of whether or not a region distinguishes between the stimuli (Carlson and Wardle, 2015). MVPA also introduces many more researcher degrees of freedom.

It is important to remember that MVPA is not simply a replacement of univariate analyses. For example, many common methods of implementing MVPA are not very sensitive to the shift of an ROI's mean magnitude across conditions, which univariate analyses capture easily (Davis *et al.*, 2014; Naselaris and Kay, 2015). In addition, it is important to consider that two conditions might evoke similar univariate responses but different multivoxel response patterns. This does not necessarily imply that these conditions have nothing in common psychologically or that they do not entail shared processing demands. As such, both techniques may be used in a complementary manner. In this section, we describe issues that may be helpful to consider when planning and interpreting MVPA.

### What are we measuring, content or process?

In both MVPA and univariate analyses, it is often difficult to ascertain when the apparent neural encoding of stimulus characteristics reflects the computation or representation of those characteristics themselves and when it reflects systematic (and perhaps subtle) effects on processes that typically follow the computation of those characteristics. For example, much previous research suggests that the activity in parietal and premotor regions associated with planning and executing actions is associated with viewing tools; whether this activity reflects encoding part of the tool concept itself (Mahon and Caramazza, 2008), or downstream processes that typically follow tool identification (e.g. prediction of future actions; Martin, 2016), is the subject of ongoing debate. Thus, the same results may be interpreted by some authors as information encoding and by other authors as processes being affected by that information. In the same way, if a decoding analysis can distinguish between human and dog faces, it is unclear if that brain region encodes the content of those stimuli differently (e.g. physical features) or if this result reflects differences in other related processes (e.g. behavioral predictions, knowledge retrieval).

### Beyond static multivoxel response patterns

In some cases, researchers may be interested in elucidating the neural basis of psychological processes that unfold over time, rather than 'snapshots' of perceptions (e.g. someone's apparent emotional state) or retrieved knowledge (e.g. the social group to which someone belongs). In such instances, it can be appropriate to use characterizations of fMRI data that capture how responses change over time, such as how multivoxel patterns ebb and flow over time (e.g. Chang *et al.*, 2018; R. Hyon *et al.*, 2020) or how patterns of functional connectivity vary across tasks or conditions (Richiardi *et al.*, 2011; Shirer *et al.*, 2012). The same methods used in MVPA can be used to analyze patterns of functional connectivity. A particular benefit of characterizing task-evoked fMRI responses using patterns of functional connectivity, where each feature is a correlation between two brain regions' response time series (e.g. the functional connectivity between the amygdala and prefrontal cortex), is that, unlike voxels, these features are abstracted away from the spatial layout of the data and, thus, readily generalize across participants. As such, when analyzing patterns of functional connectivity to decode psychological processes or states, data can be easily aggregated across participants, which can substantially increase the amount of training data available for decoding analyses and, in turn, the ability of machine learning algorithms to learn generalizable distinctions between conditions.

### Within- *vs* between-subject decoding

As alluded to above, decoding analyses generally benefit from having more training data in which to learn distinctions between conditions, and one way to achieve this is by analyzing response patterns that can be well-aligned across participants. This includes cases where patterns of functional connectivity, rather than multivoxel response patterns, are used for decoding (Richiardi *et al.*, 2011; Shirer *et al.*, 2012), and also cases where functional alignment methods are employed (Haxby *et al.*, 2011; Chen *et al.*, 2015), or where RSA is used to generate features for decoding in 'similarity space' rather than 'voxel space' (Raizada and Connolly, 2012). In addition to the tendency for techniques that facilitate between-subject decoding to produce relatively high classification accuracies (due to the increase in the amount of training data available), between-subject decoding approaches could also have practical benefits in cases where researchers wish to predict things about new individuals. That said, within-subject analyses may be preferable in cases where response patterns are thought to be idiosyncratic to particular participants, either because of between-subject heterogeneity in fine-scale functional brain organization (see Comparison Across Individuals section and Representational Spatial Scale section) or because the stimuli in question connote meaning that is inherently specific to each participant (e.g. the personal meaning of objects, Charest *et al.*, 2014; real-world social relationships, Parkinson *et al.*, 2017).

### Imaging resolution

While MVPA can detect information carried at a finer-grained spatial scale than most univariate fMRI analyses, it is still relatively coarse when compared with methods that analyze individual neurons. Patterns of neuronal activity have been shown to carry a diversity of information (Georgopoulos *et al.*, 1988; Pouget *et al.*, 2000; Kiani *et al.*, 2007), yet each fMRI voxel contains hundreds of thousands of neurons. Thus, while we are gaining some nuanced signal in MVPA compared with univariate tests, we are missing information carried at a much finer spatial scale.

***Examining multivoxel, rather than multi-neuron, patterns can systematically produce both false positives and false negatives.*** The relatively low spatial resolution of fMRI data can engender misleading results in the context of MVPA, potentially leading to false positives in some cases and false negatives in others. For example, neurophysiological studies in monkeys show that nearby but largely non-overlapping sets of neurons in the orbitofrontal cortex encode the value of social and non-social rewards (Watson and Platt, 2012). Given that many thousands of neurons comprise each voxel in a multivoxel pattern, using MVPA (or univariate analyses) on fMRI data to study such phenomena may lead researchers to erroneously conclude the presence of a common encoding scheme. This could be an issue in any cases where distinct, but nearby or interdigitated, populations of neurons encode different kinds of information.

On the other hand, analyzing multivoxel patterns, rather than multi-neuron patterns, can also systematically produce false negatives. Dubois *et al.* (2015) compared the analyses of populations of single units and of multivoxel patterns of fMRI data while monkeys viewed faces. Both the identity and viewpoint of faces could be decoded from multi-neuron response patterns, but that only facial viewpoint, and not identity, could be reliably decoded using MVPA. This appeared to be due to the fact that whereas similarly tuned cells signaling facial viewpoint were tightly clustered in space, similarly tuned cells signaling facial identity were weakly spatially clustered. Thus, MVPA of fMRI data may not be sensitive to information present in the actual underlying neural population codes when neurons are not strongly clustered based on their selectivity to a given stimulus dimension.

### Uncertainty about the timing of social and affective processes

When studying the neural basis of social and affective processes, researchers will sometimes be uncertain when exactly a psychological process occurred. For example, if a participant is given an 8 s window to reappraise a stressful event, it can be difficult or impossible to ascertain when during that 8 s window the start, end, and duration of that reappraisal process actually took place (Lieberman and Cunningham, 2009). In such cases, how should researchers approach MVPA methods? One option is to simply estimate the multivoxel response pattern from an entire block or event, as one might do for a univariate analysis, and submit those event-wise patterns to MVPA. Another option would be to estimate the multivoxel response patterns at each time point within a block or event and then perform decoding or RSA separately at each time point to test when information that distinguishes between conditions is reliably present across participants (Soon *et al.*, 2008; Cichy *et al.*, 2014). Note that this approach assumes consistency across participants and events in the timing of the psychological process at hand.

If the timing of a psychological process is thought to differ across people and events, and the researcher does not wish to estimate a single multivoxel response pattern for the entire event (i.e. the first option summarized above), complications can arise. For instance, there would be a very large number of possible combinations of durations and onsets to choose from for each event when generating multivoxel patterns to analyze. Trying out different possibilities of onsets and durations for each event would be both computationally intensive and necessitate correcting for an untenably large number of statistical tests. One strategy would be to concatenate multivoxel response patterns from different time points within each event into a longer feature vector (i.e. a single sample) and then use supervised learning to identify which spatiotemporal features distinguish between conditions. Although such a method would assume consistency across events in the temporal profile of how a psychological process unfolds, it would allow for heterogeneity in the timing of psychological processes across participants if carried out within each participant separately. We encourage researchers concerned with such questions to explore emerging approaches being developed to analyze neuroimaging modalities with greater temporal resolution than fMRI (e.g. Su *et al.*, 2012; King *et al.*, 2014; Grootswagers *et al.*, 2017), as they may be adapted for fMRI and foster a richer consideration of how the spatiotemporal dynamics of neural response patterns relate to psychological processes.

## Conclusion

In this article, we aimed to provide a practical and accessible introduction to the popular family of analyses known as multivoxel pattern analysis, or MVPA, for social and affective neuroscientists of all levels, particularly those new to such methods. We explained what MVPA is by comparing it to commonly used univariate analyses, explored different types of questions that can be answered with MVPA and covered practical steps and considerations required to implement MVPA in one's own data. Many others have discussed the finer points of specific analyses that were referred to more generally here and should be studied for a deeper understanding of these tools (e.g. RSA, Kriegeskorte *et al.*, 2008a, Diedrichsen and Kriegeskorte, 2017; decoding, Pereira *et al.*, 2009; hyperalignment, Haxby *et al.*, 2011) and new ways in which to implement them (e.g. pattern-based biomarkers, van der Miesen *et al.*, 2019; representational connectivity between brain regions, Anzelloti *et al.*, 2017; real-time decoded neurofeedback, Watanabe *et al.*, 2017; Taschereau-Dumouchel *et al.*, 2018). In recent years, MVPA has grown rapidly in popularity. As these techniques are expanded and applied in new ways, we will be able to use neuroimaging to explore many new types of questions in the field of social and affective neuroscience.

## Funding

M.E.W. is supported by a Graduate Research Mentorship award and a Dean's Scholar award from the University of California, Los Angeles (UCLA). C.P. is supported by NSF grant SBE-1835239 and a Sloan Foundation Research Fellowship.

## Conflict of interest

The authors declare that they have no conflicts of interest.

## References


Aguirre, G.K., Mattar, M.G., Magis-Weinberg, L. (2011). de Bruijn cycles for neural decoding. *NeuroImage*, **56**(3), 1293–300. doi: 10.1016/j.neuroimage.2011.02.005.

Anzellotti, S., Caramazza, A., Saxe, R. (2017). Multivariate pattern dependence. *PLoS Computational Biology*, **13**(11), 1–20. doi: 10.1371/journal.pcbi.1005799.

Bhanji, J.P., Delgado, M.R. (2014). The social brain and reward: social information processing in the human striatum. *Wiley Interdisciplinary Reviews: Cognitive Science*, **5**(1), 61–73. doi: 10.1002/wcs.1266.

Buračas, G.T., Boynton, G.M. (2002). Efficient design of event-related fMRI experiments using m-sequences. *NeuroImage*, **16**, 801–13. doi: 10.1006/nimg.2002.1116.

Cai, M.B., Schuck, N.W., Pillow, J.W., Niv, Y. (2019). Representational structure or task structure? Bias in neural representational similarity analysis and a Bayesian method for reducing bias. *PLoS Computational Biology*, **15**(5), e1006299. doi: 10.1371/journal.pcbi.1006299.

Carlson, T.A., Wardle, S.G. (2015). Sensible decoding. *NeuroImage*, **110**, 217–8. doi: 10.1016/j.neuroimage.2015.02.009.

Chadwick, M.J., Bonnici, H.M., Maguire, E.A. (2012). Decoding information in the human hippocampus: a user's guide. *Neuropsychologia*, **50**(13), 3107–21. doi: 10.1016/j.neuropsychologia.2012.07.007.

Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., Wager, T.D. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biology*, **13**(6), 1–28. doi: 10.1371/journal.pbio.1002180.

Chang, L.J., Jolly, E., Cheong, J.H., Rapuano, K., Greenstein, N., Chen, P.-H.A., Manning, J.R. (2018). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *BioRxiv*. doi: 10.1101/487892.

Charest, I., Kievit, R.A., Schmitz, T.W., Deca, D., Kriegeskorte, N., Ungerleider, L.G. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(40), 14565–70. doi: 10.1073/pnas.1402594111.

Chen, P.H., Chen, J., Yeshurun, Y., Hasson, U., Haxby, J.V., Ramadge, P.J. (2015). A reduced-dimension fMRI shared response model. In: *Advances in Neural Information Processing Systems*, pp. 460–8.

Chikazoe, J., Lee, D.H., Kriegeskorte, N., Anderson, A.K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, **17**(8), 1114–22. doi: 10.1038/nn.3749.

Cichy, R.M., Pantazis, D., Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, **17**(3), 455–62. doi: 10.1038/nn.3635.

Combrisson, E., Jerbi, K. (2015). Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, **250**, 126–36. doi: 10.1016/j.jneumeth.2015.01.010.

Coutanche, M.N., Thompson-Schill, S.L. (2012). The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. *NeuroImage*, **61**(4), 1113–9. doi: 10.1016/j.neuroimage.2012.03.076.

Davis, T., LaRocque, K.F., Mumford, J.A., Norman, K.A., Wagner, A.D., Poldrack, R.A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, **97**, 271–83. doi: 10.1016/j.neuroimage.2014.04.037.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, **43**(1), 44–58. doi: 10.1016/j.neuroimage.2008.06.037.

Diedrichsen, J., Kriegeskorte, N. (2017). Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, **13**(4), e1005508. doi: 10.1371/journal.pcbi.1005508.

Douglas, P.K., Harris, S., Yuille, A., Cohen, M.S. (2011). Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. *NeuroImage*, **56**(2), 544–53. doi: 10.1016/j.neuroimage.2010.11.002.

Downing, P.E., Wiggett, A.J., Peelen, M.V. (2007). Functional magnetic resonance imaging investigation of overlapping lateral occipitotemporal activations using multi-voxel pattern analysis. *Journal of Neuroscience*, **27**(1), 226–33. doi: 10.1523/JNEUROSCI.3619-06.2007.

Dubois, J., de Berker, A.O., Tsao, D.Y. (2015). Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. *Journal of Neuroscience*, **35**(6), 2791–802. doi: 10.1523/JNEUROSCI.4037-14.2015.

Emberson, L.L., Zinszer, B.D., Raizada, R.D.S., Aslin, R.N. (2017). Decoding the infant mind: multivariate pattern analysis (MVPA) using fNIRS. *PLoS One*, **12**(4), 1–23. doi: 10.1371/journal.pone.0172500.

Ersner-Hershfield, H., Garton, M.T., Ballard, K., Samanez-Larkin, G.R., Knutson, B. (2009). Don't stop thinking about tomorrow: individual differences in future self-continuity account for saving. *Judgment and Decision making*, **4**(4), 280–6.

Gardumi, A., Ivanov, D., Hausfeld, L., Valente, G., Formisano, E., Uludağ, K. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, **132**, 32–42. doi: 10.1016/j.neuroimage.2016.02.033.

Georgopoulos, A.P., Kettner, R.E., Schwartz, A.B. (1988). Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. *Journal of Neuroscience*, **8**(8), 2928–37 Available:. http://www.ncbi.nlm.nih.gov/pubmed/3411362.

Giordano, B.L., McAdams, S., Zatorre, R.J., Kriegeskorte, N., Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, **23**(9), 2025–37. doi: 10.1093/cercor/bhs162.

Goesaert, E., Op de Beeck, H.P. (2013). Representations of facial identity information in the ventral visual stream investigated with multivoxel pattern analyses. *Journal of Neuroscience*, **33**(19), 8549–58. doi: 10.1523/jneurosci.1829-12.2013.

Görgen, K., Hebart, M.N., Allefeld, C., Haynes, J.-D. (2018). The same analysis approach: practical protection against the pitfalls of novel neuroimaging analysis methods. *NeuroImage*, **180**, 19–30. doi: 10.1016/j.neuroimage.2017.12.083.

Grootswagers, T., Wardle, S.G., Carlson, T.A. (2017). Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *Journal of Cognitive Neuroscience*, **29**(4), 677–97. doi: 10.1162/jocn_a_01068.

Hassabis, D., Spreng, R.N., Rusu, A.A., Robbins, C.A., Mar, R.A., Schacter, D.L. (2014). Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cerebral Cortex*, **24**(8), 1979–87. doi: 10.1093/cercor/bht042.

Hastie, T., Tibshirani, R., Friedman, J. (2017). *The Elements of Statistical Learning*, 2nd edn, New York, NY: Springer, 10.1007/b94608.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, **293**(5539), 2425–30. doi: 10.1126/science.1063736.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., *et al.* (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, **72**, 404–16.

Hebart, M.N., Baker, C.I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, **180**, 4–18. doi: 10.1016/j.neuroimage.2017.08.005.

Hendriks, M.H.A., Daniels, N., Pegado, F., de Beeck, H.P.O. (2017). The effect of spatial smoothing on representational similarity in a simple motor paradigm. *Frontiers in Neurology*, **8**(222), 1–11. doi: 10.3389/fneur.2017.00222.

Horikawa, T., Tamaki, M., Miyawaki, Y., Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. *Science*, **340**(6132), 639–42. doi: 10.1126/science.1234330.

Hyon, R., Kleinbaum, A.M., Parkinson, C. (2020) Social network proximity predicts similar trajectories of psychological states: evidence from multi-voxel spatiotemporal dynamics. *NeuroImage*, **216**, 116492. https://doi.org/10.1016/j.neuroimage.2019.116492.

Jain, A.K., Chandrasekaran, B. (1982). Dimensionality and sample size considerations. *Pattern Recognition in Practice*, **2**, 835–55.

Khaligh-Razavi, S.M., Henriksson, L., Kay, K., Kriegeskorte, N. (2017). Fixed versus mixed RSA: explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. *Journal of Mathematical Psychology*, **76**, 184–97.

Kiani, R., Esteky, H., Mirpour, K., Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, **97**(6), 4296–309. doi: 10.1152/jn.00024.2007.

Kiani, R., Cueva, C.J., Reppas, J.B., Newsome, W.T. (2014). Dynamics of neural population responses in prefrontal cortex indicate changes of mind on single trials. *Current Biology*, **24**(13), 1542–7. doi: 10.1016/j.cub.2014.05.049.

King, J.R., Gramfort, A., Schurger, A., Naccache, L., Dehaene, S. (2014). Two distinct dynamic modes subtend the detection of unexpected sounds. *PLoS One*, **9**(1), e85791. doi: 10.1371/journal.pone.0085791.

Kriegeskorte, N., Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, **38**(4), 649–62. doi: 10.1016/j.neuroimage.2007.02.022.

Kriegeskorte, N., Formisano, E., Sorger, B., Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences*, **104**(51), 20600–5. doi: 10.1073/pnas.0705654104.

Kriegeskorte, N., Mur, M., Bandettini, P. (2008a). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, **2**(4), 1–28. doi: 10.3389/neuro.06.004.2008.

Kriegeskorte, N., Mur, M., Ruff, D.A., *et al.* (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, **60**(6), 1126–41. doi: 10.1016/j.neuron.2008.10.043.

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S., Baker, C.I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience*, **12**(5), 535–40. doi: 10.1038/nn.2303.

Lieberman, M.D., Cunningham, W.A. (2009). Type I and type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience*, **4**(4), 423–8. doi: 10.1093/scan/nsp052.

Lin, A., Adolphs, R., Rangel, A. (2012). Social and monetary reward learning engage overlapping neural substrates. *Social Cognitive and Affective Neuroscience*, **7**(3), 274–81. doi: 10.1093/scan/nsr006.

Linden, D.E.J., Oosterhof, N.N., Klein, C., Downing, P.E. (2012). Mapping brain activation and information during category-specific visual working memory. *Journal of Neurophysiology*, **107**(2), 628–39. doi: 10.1152/jn.00105.2011.

Ling, C.X., Huang, J., Zhang, H. (2003). AUC: a better measure than accuracy in comparing learning algorithms. In: *Conference of the Canadian Society for Computational Studies of Intelligence* (Vol. **2671**, pp. 329–41). 10.1007/3-540-44886-1_25

Mahon, B.Z., Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology Paris*, **102**(1–3), 59–70. doi: 10.1016/j.jphysparis.2008.03.004.

Martin, A. (2016). GRAPES—grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, **23**(4), 979–90. doi: 10.3758/s13423-015-0842-3.

Mazziotta, J.C., Toga, A.W., Evans, A., Fox, P., Lancaster, J. (1995). A probabilistic atlas of the human brain: theory and rationale for its development. *NeuroImage*, **2**(2), 89–101. doi: 10.1006/nimg.1995.1012.

Misaki, M., Luh, W.M., Bandettini, P.A. (2013). The effect of spatial smoothing on fMRI decoding of columnar-level organization with linear support vector machine. *Journal of Neuroscience Methods*, **212**(2), 355–61. doi: 10.1016/j.jneumeth.2012.11.004.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, **320**(5880), 1191–5. doi: 10.1126/science.1152876.

Mumford, J.A., Davis, T., Poldrack, R.A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. *NeuroImage*, **103**, 130–8. doi: 10.1016/j.neuroimage.2014.09.026.

Naselaris, T., Kay, K.N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences*, **19**(10), 551–4. doi: 10.1016/j.tics.2015.07.005.

Nastase, S.A., Connolly, A.C., Oosterhof, N.N., *et al.* (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, **27**(8), 4277–91. doi: 10.1093/cercor/bhx138.

Nestor, A., Plaut, D.C., Behrmann, M. (2011). Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **108**(24), 9998–10003. doi: 10.1073/pnas.1102433108.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. *PLoS Computational Biology*, **10**(4), 1–11. doi: 10.1371/journal.pcbi.1003553.

Op de Beeck, H.P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? *NeuroImage*, **49**, 1943–8. doi: 10.1016/j.neuroimage.2009.02.047.

Parkinson, C., Kleinbaum, A.M., Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, **1**(5), 72. doi: 10.1038/s41562-017-0072.

Parkinson, C., Kleinbaum, A.M., Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, **9**(1), 332. doi: 10.1038/s41467-017-02722-7.

Peelen, M.V., Wiggett, A.J., Downing, P.E. (2006). Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. *Neuron*, **49**, 815–22. doi: 10.1016/j.neuron.2006.02.004.

Peelen, M.V., Atkinson, A.P., Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, **30**(30), 10127–34. doi: 10.1523/JNEUROSCI.2161-10.2010.

Pereira, F., Mitchell, T., Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, **45**, S199–209. doi: 10.1016/j.neuroimage.2008.11.007.

Pouget, A., Dayan, P., Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, **1**, 125–32. doi: 10.1038/35039062.

Raizada, R.D.S., Connolly, A.C. (2012). What makes different people's representations alike: neural similarity space solves the problem of across-subject fMRI decoding. *Journal of Cognitive Neuroscience*, **24**(4), 868–77. doi: 10.1162/jocn_a_00189.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D. (2011). Decoding brain states from fMRI connectivity graphs. *NeuroImage*, **56**(2), 616–26. doi: 10.1016/j.neuroimage.2010.05.081.

Shepard, R.N. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human Factors: The Journal of Human Factors and Ergonomics Society*, **5**(1), 33–48. doi: 10.1177/001872086300500104.

Shepard, R.N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, **1**(1), 54–87. doi: 10.1016/0022-2496(64)90017-3.

Shepard, R.N., Chipman, S. (1970). Second-order isomorphism of internal representations: shapes of states. *Cognitive Psychology*, **1**(1), 1–17. doi: 10.1016/0010-0285(70)90002-2.

Shepard, R.N., Cooper, L.A. (1992). Representation of colors in the blind, color-blind, and normally sighted. *Psychological Science*, **3**(2), 97–104. doi: 10.1111/j.1467-9280.1992.tb00006.x.

Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D. (2012). Decoding subject-driven cognitive states with whole-brain connectivity patterns. *Cerebral Cortex*, **22**(1), 158–65. doi: 10.1093/cercor/bhr099.

Smith, P.L., Little, D.R. (2018). Small is beautiful: in defense of the small-N design. *Psychonomic Bulletin and Review*, **25**, 2083–101. doi: 10.3758/s13423-018-1451-8.

Soon, C.S., Brass, M., Heinze, H.J., Haynes, J.D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, **11**(5), 543–5. doi: 10.1038/nn.2112.

Su, L., Fonteneau, E., Marslen-Wilson, W., Kriegeskorte, N. (2012). Spatiotemporal searchlight representational similarity analysis in EMEG source space. In: *Proceedings—2012 2nd International Workshop on Pattern Recognition in NeuroImaging*, 97–100. 10.1109/PRNI.2012.26

Talairach, J., Tournoux, P. (1988). *Co-planar Stereotaxic Atlas of the Human Brain*, New York: Thieme.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J.D., Kawato, M., Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences of the United States of America*, **115**(13), 3470–5. doi: 10.1073/pnas.1721572115.

Todd, M.T., Nystrom, L.E., Cohen, J.D. (2013). Confounds in multivariate pattern analysis: theory and rule representation case study. *NeuroImage*, **77**, 157–65. doi: 10.1016/j.neuroimage.2013.03.039.

Uchida, N., Takahashi, Y.K., Tanifuji, M., Mori, K. (2000). Odor maps in the mammalian olfactory bulb: domain organization and odorant structural features. *Nature Neuroscience*, **3**(10), 1035–43. doi: 10.1038/79857.

van der Miesen, M.M., Lindquist, M.A., Wager, T.D. (2019). Neuroimaging-based biomarkers for pain: state of the field and current directions. *PAIN Reports*, **4**(4), e751. doi: 10.1097/PR9.0000000000000751.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.-W., Kross, E. (2013). An fMRI-based neurologic signature of physical pain. *New England Journal of Medicine*, **368**(15), 1388–97. doi: 10.1056/NEJMoa1204471.

Wake, S.J., Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, **12**(10), 1558–64. doi: 10.1093/scan/nsx092.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, **137**, 188–200. doi: 10.1016/j.neuroimage.2015.12.012.

Wang, T., Deng, J., He, B. (2004). Classifying EEG-based motor imagery tasks by means of time-frequency synthesized spatial patterns. *Clinical Neurophysiology*, **115**(12), 2744–53. doi: 10.1016/j.clinph.2004.06.022.

Wardle, S.G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.M., Carlson, T.A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *NeuroImage*, **132**, 59–70. doi: 10.1016/j.neuroimage.2016.02.019.

Watanabe, T., Sasaki, Y., Shibata, K., Kawato, M. (2017). Advances in fMRI real-time neurofeedback. *Trends in Cognitive Sciences*, **21**(12), 997–1010. doi: 10.1016/j.tics.2017.09.010.

Watson, K.K., Platt, M.L. (2012). Social signals in primate orbitofrontal cortex. *Current Biology*, **22**(23), 2268–2273. doi: 10.1016/j.cub.2012.10.016.

Woo, C.-W., Chang, L.J., Lindquist, M.A., Wager, T.D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, **20**(3), 365–77. doi: 10.1038/nn.4478.

Yeshurun, Y., Swanson, S., Simony, E., Chen, J., Lazaridi, C., Honey, C.J., Hasson, U. (2017). Same story, different story: the neural representation of interpretive frameworks. *Psychological Science*, **28**(3), 307–19. doi: 10.1177/0956797616682029.

Zeithamova, D., de Araujo Sanchez, M.-A., Adke, A. (2017). Trial timing and pattern-information analyses of fMRI data. *NeuroImage*, **153**(November 2016), 221–31. doi: 10.1016/j.neuroimage.2017.04.025.