Pattern-information analysis: from stimulus decoding to computational-model testing

Nikolaus Kriegeskorte
Medical Research Council, Cognition and Brain Sciences Unit,
15 Chaucer Road, Cambridge, CB2 7EF, UK

Citation: Kriegeskorte N (in press) Pattern-information analysis: from stimulus decoding to computational-model testing. *NeuroImage*.

Abstract

Pattern-information analysis has become an important new paradigm in functional imaging. Here I review and compare existing approaches with a focus on the question of what we can learn from them in terms of brain theory. The most popular and widespread method is stimulus decoding by response-pattern classification. This approach addresses the question whether activity patterns in a given region carry information about the stimulus category. Pattern classification uses generic models of the stimulus-response relationship that do not mimic brain information processing and treats the stimulus space as categorical – a simplification that is often helpful, but also limiting in terms of the questions that can be addressed. We can address the question whether representations are consistent across different stimulus sets or tasks by cross-decoding, where the classifier is trained with one set of stimuli (or task) and tested with another. Beyond pattern classification, a major new direction is the integration of computational models of brain information processing into pattern-information analysis. This approach enables us to address the question to what extent competing computational models are consistent with the stimulus representations in a brain region. Two methods that test computational models are voxel receptive-field modeling and representational similarity analysis. These methods sample the stimulus (or mental-state) space more richly, estimate a separate response pattern for each stimulus, and can generalize from the stimulus sample to a stimulus population. Computational models that mimic brain information processing predict responses from stimuli. The reverse transform can be modeled to reconstruct stimuli from responses. Stimulus reconstruction is a challenging feat of engineering, but the implications of the results for brain theory are not always clear. Exploratory pattern analyses complement the confirmatory approaches mentioned so far and can reveal strong, unexpected effects that might be missed when testing only a restricted set of predefined hypotheses.

Introduction

Perceptual, cognitive, and motor representations are thought to reside in neuronal population codes (e.g. Averbeck et al., 2006). This provides a straightforward motivation for pattern-information analysis in functional imaging (e.g. Haxby et al., 2001; Cox & Savoy 2003; Carlson et al., 2003; Mitchell et al., 2004; Kriegeskorte, 2004; Kamitani & Tong, 2005; Haynes & Rees, 2006; Norman et al., 2006; Kriegeskorte et al., 2006; Mur et al., 2009) and in cell recording (Hung et al., 2005; Kiani et al., 2007): to elucidate (more fully than single-unit or single-voxel analyses can) what information is present in a given brain region. In this commentary, I review our current toolbox of pattern-information approaches with a focus on what we can learn from them about brain function. The paper is divided into four sections, each of which discusses a different approach to pattern-information analysis.

The first section discusses methods that test for information about a particular stimulus dimension in brain response patterns (hypothesis-driven goal 1). This approach includes pattern-classifier decoding, the most popular and widespread type of pattern-information analysis. Classifier decoding treats the stimulus space as categorical and "predicts" the stimulus category from the response pattern. More generally, the stimulus space could be treated as continuous, and I argue that the direction, in which the dependency between stimulus and response is modeled (decoding or encoding) is largely inessential to the neuroscientific interpretation (Fig. 1). Most applications have used generic linear models. I argue in favor of linear models on the basis of their stability and interpretability.

The second section discusses methods that test whether a computational model of brain information processing can account for a region's response patterns (hypothesis-driven goal 2). Traditionally, statistical analysis of brain activity uses generic, often linear, models that do not simulate brain information processing (as in goal 1). Brain-experimental results are then related to computational models only at the level of verbal theory. To directly test computational models with brain-activity data, we need to integrate these (typically nonlinear) models into data analysis. Two methods that achieve this are voxel-receptive-field modeling (Dumoulin & Wandell, 2008; Kay et al., 2008; Mitchell et al., 2008) and representational similarity analysis (Kriegeskorte et al., 2008a, 2008b). These methods sample the stimulus (or mental-state) space more richly, estimating a separate response pattern for each stimulus and forgoing any predefined stimulus grouping (Fig. 2). As these approaches are just beginning to gain momentum, there are few examples in the literature. I therefore take a different approach in this section and review three studies in detail (Mitchell et al., 2008; Kay et al., 2008; Kriegeskorte et al., 2008a). Voxel-receptive-field modeling predicts response patterns; representational similarity analysis predicts response-pattern dissimilarities, providing alternative statistical tests of the same conceptual claim, namely that a computational model can account for the representational space of a brain region (Figs. 3, 4).

The shorter third and fourth sections discuss exploratory analysis of population activity patterns and stimulus reconstruction, respectively. These two approaches do not test explicit hypotheses about brain function. Exploratory analysis requires fewer assumptions and can yield unexpected discoveries. It can reveal stimulus-response relationships that explain a lot of variance, but might have been missed in an overly restricted hypothesis-driven approach. Stimulus reconstruction models perceptual processing in reverse, predicting the stimulus from the response pattern (a form of decoding that generalizes to novel stimuli). Reconstruction is a tough engineering challenge and provides an intuitive illustration (the reconstructed stimulus) of the information represented in a region.

However, it is not clear how exactly stimulus reconstruction results constrain neuroscientific theory. Fig. 5 compares the entire range of pattern-information approaches discussed in this paper.

For simplicity, this commentary focuses on the relationship between “stimulus” and “response” in considering pattern-information analysis. However, the arguments apply to other scenarios as well, where the mental states investigated are not directly elicited by stimuli (e.g. mental imagery), or where a brain-behavior relationship is analyzed. The application of pattern-information methods to the relationships between brain and behavior and between different brain regions, individuals, and species have recently been discussed elsewhere (Raizada & Kriegeskorte, in press; Kriegeskorte 2009).

(1) Goal 1: Testing for specific stimulus information in response patterns

A popular pattern-information analysis is pattern classification (e.g. Haxby et al., 2001; Kamitani & Tong, 2005; for a textbook see Duda et al., 2001). In this approach, the stimuli are “predicted” from the activity patterns they elicit. I put “prediction” in quotes here, because it does not usually refer to foretelling a future event or the trajectory of brain dynamics (but see Haynes et al., 2007; Soon et al., 2008). We can interpret the term in the context of an imaginary game of “Give me the response patterns, and I will tell you the stimuli.” This paradigm is also referred to as “decoding” (e.g. Mitchell et al., 2004; Kamitani & Tong, 2005; Haynes & Rees, 2006; Friston et al., 2008). The rationale for this approach is that if decoding works better than chance, then there must be information about the stimuli in the response patterns.

Pattern-classifier decoding differs in three respects from classical activation analyses, which use a univariate general linear model of the response:

(1) The response is treated multivariately: as a pattern (and the stimulus space as categorical).

(2) The model operates in reverse direction: from responses to stimuli.

(3) The data are divided into independent training and test sets, where the training set is used to fit the model parameters, and the test set to estimate decoding accuracy and test for stimulus information in the response pattern.

When the goal is to test for stimulus information, the most important of these three features is (1), the analysis of response-pattern information. The other two features are of a more technical nature: a test for stimulus information in the response pattern could use a model operating in either direction and, regardless of the model direction, inference could be performed using either independent training and test sets or a single data set (and stronger assumptions).

A key statistical advantage of using an independent test set is that the assumptions of the model are implicitly tested when we assess prediction performance. To the extent that the assumptions are violated, prediction will suffer. The test of the presence of pattern information provided by this approach depends on the assumptions of the model for its sensitivity, but not for its specificity: If the assumptions are violated, the test is still valid.

Demonstrating a statistical dependency between stimulus and response pattern

Words like “prediction”, “decoding”, and “brain reading” should not be taken to imply that what is demonstrated goes beyond a statistical dependency between stimulus and response. Whether we are “predicting” the stimulus from the response or the response from the stimulus, all that is demonstrated is a statistical dependency (or, equivalently, mutual information) between the two (Kriegeskorte & Bandettini, 2007).

In a univariate scenario (Fig. 1a), it is easy to see that a correlation between two variables implies predictability in either direction. Note that for a function f: x->y, we can deterministically predict y from x but not necessarily vice versa (as f may not be invertible). In a stochastic setting, by contrast, a statistical dependency implies above-chance predictability in either direction (knowing either variable constrains the possible states of the other variable, even if deterministic prediction fails because multiple states of the other variable remain possible). The ability to “predict” and “decode” could thus equally be claimed on the basis of any classical activation result, such as Kanwisher et al. (1997): If the fusiform face area is specifically activated by face stimuli, then activation there "predicts" that the stimulus was a face.

In a multivariate scenario, as well, demonstrating above-chance predictability in either direction implies a statistical dependency and thus above-chance predictability in the other direction (Fig. 1b; Fig. 4, top box). The direction in which a generic model operates therefore often does not matter to the qualitative neuroscientific interpretation. The terms "prediction" and "decoding" denote that the direction of the model is inverted (as compared to the direction of causality: from stimulus to response). However, the key novel feature of pattern-information analysis is the joint analysis of multiple responses as a population code, whether we use decoding or encoding models.

Encoding versus decoding models

Although encoding and decoding models both demonstrate a stimulus-response dependency, they elucidate complementary aspects of the stimulus-response relationship (Fig. 1c). An encoder can reveal how much of a region's total response-pattern entropy the stimulus can explain. Moreover, an encoder might simulate the actual causal process from stimulus to response (Naselaris et al., this issue; Gallant et al., in press; see *Goal 2* below). Conversely, a decoder can reveal how much of the total stimulus information is present in a brain region. Moreover, a decoder might simulate representational readout and can help us relate the region's pattern information to trial-by-trial behavior (e.g. predicting behavioral errors).

Consider the frequent case of a binary stimulus distinction: two categories. If the two categories occur with equal frequency, the stimulus entropy is 1 bit. The response-pattern entropy will generally be much greater, because the response-pattern space is multidimensional and continuous. Assume we can decode the stimulus category from a given brain region with perfect accuracy for single trials (this is rarely the case in reality, but instructive to imagine). This would mean that the single-trial pairwise stimulus information (i.e. the mutual information between the stimulus dichotomy and the single-trial response-pattern measurement; Kriegeskorte et al., 2007) is 1 bit, i.e. we can explain the entire stimulus-category entropy. Note that the proportion of response-pattern entropy explainable by the stimulus category is nevertheless going to be very small: 1 bit of mutual information is a small proportion of the total response-pattern entropy. This indicates that the region's response patterns are dominated by a combination of (a) other information than our stimulus category and (b) noise. This is trivial for the two-category case, but useful in the context of more complex representational models (including those discussed below, under *Goal 2*). It is useful, because it enables us to assess how far away we are from a complete functional account of a brain region's responses (Gallant et al., in press).

Modeling continuous stimulus-response relationships

Pattern classification treats the stimulus space as categorical – a simplification that is often helpful, but also limiting in terms of the questions that can be addressed. The grouping of the stimuli into categories for classification is often artificial and we may have continuous parameters describing the stimuli. For example, instead of grouping object images into conventional categories, we might describe them by multiple properties (e.g. describing color, shape, contrast, size).

This motivates the modeling of continuous stimulus-response relationships. A continuous approach to pattern-information analysis is provided by classical multivariate statistics (for a textbook, see Krzanowski, 1988). For example, multivariate analysis of covariance (MANCOVA) could be used to model multivariate response patterns as a linear combination of stimulus parameters. The classical multivariate methods extend the framework of the general linear model into the multivariate domain and allow efficient tests of complex hypotheses. They are arguably more elegant and versatile, and definitely less cumbersome (requiring no data splitting) and less computationally expensive than typical classifier analyses. Both stimulus description and response pattern can be modeled as multivariate and continuous. Complex hypotheses can easily be tested in this framework. For example, we can test whether adding a set of predictors to the model explains additional multivariate variance in the response patterns (extra-sums-of-squares-and-products test).[1] However, the classical multivariate methods rely on the assumption of multivariate normality, which may not always hold for functional imaging data and fMRI in particular. In order to avoid relying on multivariate normality, we can use a randomization test to perform inference on classical multivariate models (Kriegeskorte et al., 2006). Such tests require no distributional assumptions, but are computationally expensive (Nichols & Holmes, 2002).

Advantages of linear models: stability, interpretability

The pattern-classification and continuous multivariate approaches discussed so far utilize generic models from statistics and machine learning, which do not attempt to mimic brain information processing. The models serve merely to test for a statistical dependency between stimulus and response pattern. Most applications so far use linear models. Misaki et al. (2010) suggest that different linear models (e.g. Fisher linear discriminants and linear support vector machines for pattern classification) often give similar results for fMRI data. More complex nonlinear models don't tend to perform better at prediction and often perform worse than linear models for fMRI data (Cox & Savoy, 2003; LaConte, 2005; Misaki et al., 2010, but see Hanson et al., 2004). Given limited data, a simpler model can outperform a complex model, even if the complex model is correct – because of overfitting. This problem is severe in fMRI, because the number of repeated pattern measurements is not typically much larger than the number of voxels. Even linear models can substantially overfit the data in this scenario. They can therefore benefit from regularization (Misaki et al., 2010) and need to be tested with independent data. Overall, the benefits of assuming linearity to the stability of the estimates (and thus prediction performance) appear to outweigh the cost of not being able to capture nonlinear relationships in fMRI pattern analysis.

The argument against linear models is that they can miss pattern information encoded in a more complex way. However, a restriction to linearly decodable information also facilitates

[1] The test would involve reducing the model space by removing the predictor set to be tested, fitting the full and the reduced model and determining the extra-sums-of-squares-and-products matrix associated with the set of predictors. This matrix is related to the error sums-of-squares-and-products matrix and inference can be performed via Wilk's Λ, Bartlett's statistic, and the χ^2 distribution.

interpretation: What is linearly decodable is “explicit” in the representational pattern (if not in any single neuron or voxel), in the sense that it can be read out in a single biologically plausible step by a neuron at the next stage of processing. The readout filter can easily be visualized as a weight map. Given that different linear classifiers (e.g. Fisher linear discriminant, linear support vector machine) all fit a hyperplane for discrimination (albeit by somewhat different optimization criteria), the particular model used is typically of marginal relevance to the neuroscientific interpretation of the result.

Linear models, thus, are attractive when the goal is to test for pattern information that is available for immediate readout (goal 1). They are appropriate for this function precisely because they do not perform any complex transformation that would require multiple stages of processing in the brain. When the goal is to test a computational theory (goal 2), however, we will need to integrate a computational model that implements the theory into the analysis. Such models are typically nonlinear and their greater complexity poses a challenge if they are to be fitted to the data. However, we can rely on prior empirical findings and neuroscientific theory to constrain their parameters.

Inferring a causal role of pattern information

The presence of stimulus information in a brain region does not imply that this information serves the function of representing the stimulus in the context of the brain’s overall operation. The latter interpretation implies that the information has a causal role. As in the univariate scenario, we can argue for a causal influence based on experimental manipulation of the activity, causal modeling techniques, or prior assumptions about the role of the region in question. The same general techniques used to infer causal influences in the univariate scenario can be applied in the multivariate scenario. However, multivariate causal approaches are not yet well developed in systems neuroscience.

At the experimental level, we face a difficult challenge. Inferring a causal role of brain activity patterns (e.g. “the population code in region X forms the basis of perceptual decision Y”) would require experimental control of the brain activity. Transcranial magnetic stimulation (TMS) enables us to experimentally influence brain activity in humans. However, TMS has low spatial precision and doesn't enable us to impose patterns of activity. Electrical microstimulation (e.g. Afraz et al., 2006) has high precision and its extension to multiple sites is a promising avenue. Optogenetic techniques for controlling activity (Deisseroth et al., 2006) are under development. However, the latter two techniques are highly invasive and not in general suitable for use in humans. We do not presently have methods to impose arbitrary precise patterns of activity in humans.

In the absence of direct experimental evidence for a causal influence, we could rely on reasonable assumptions to constrain the causal relationships to be considered and extend advanced techniques of modeling directed interactions between brain regions (also known as “effective connectivity”) to the multivariate domain. As a simple example, we could test whether the non-stimulus-driven component of the pattern response is related to behavioral responses on a trial-by-trial basis. This idea is generalized in the framework of structural equation modeling. Alternatively, Granger causality (Roebroeck et al., 2005; Ramsey et al., 2009) exploits the temporal lag between cause and effect to infer causality (relying on the assumption that the model does not omit relevant alternative causal pathways). As another example, dynamic causal modeling (Friston et al., 2003) allows us to test and compare prespecified causal models of interactions between brain regions. In neuroimaging, however, these models of directed interactions are typically based on univariate activation time courses (fluctuations of spatially-averaged activation of the analyzed brain regions). The development of pattern-information approaches to modeling directed interactions is an important future direction. A pattern-information equivalent to undirected interactions (i.e. “functional connectivity”: correlated fluctuations of overall activation between two brain regions) is provided by “representational connectivity” (Kriegeskorte et al., 2008a). Because a causal role of activity-pattern information is difficult to infer with present experimental and analysis techniques, our representational interpretations often rest on prior empirical findings and general brain theory.

Cross-decoding: testing whether representations are consistent across different tasks

An important step beyond demonstrating a statistical dependency between the stimulus and the response patterns can be taken by cross-decoding. In cross-decoding, a decoder is trained with one set of stimuli and then tested with another, or the task eliciting the response patterns is changed. In a conventional decoding analysis, the test data set serves only to prevent overfitting from inflating the estimate of decoding performance (for a detailed discussion of this, see Kriegeskorte et al., 2009a), so what is tested is generalization to new measurements using the same stimuli. Cross-decoding, by contrast, tests generalization to novel stimuli or tasks. This allows us to assess whether representations are *consistent* between the two stimulus sets or tasks, for example between perception and memory retrieval (Polyn et al., 2005; Norman et al., 2006) or between perception and imagery (Stokes et al., 2009) of the same visual content. Successful cross-decoding with a linear pattern classifier suggests that the dimension of response-pattern space, along which the representational categories can be discriminated, is at least somewhat consistent between training and test scenario.[2]

Generalization from a sample to a population of stimuli

Conventional inferential analyses of activation and pattern-information do not generalize from the sample of stimuli used in an experiment to a population of stimuli that could have been used (Bedny et al., 2007; Kriegeskorte et al., 2008b; see also Clark, 1973). While cross-decoding tests generalization from one scenario to another (e.g. perception to imagery), we can also use separate samples of stimuli drawn from the same population of stimuli as training and test sets. This enables us to avoid overfitting to the stimulus set.

Let’s say we wanted to assess whether a brain region distinguishes animate and inanimate object images. We present 50 images of each category and train a classifier to distinguish them. Consider the null hypothesis that the region distinguishes individual images, but does not allow linear readout of animacy (as might be expected for the retina or for early visual cortex).[3] This means that there is no linear combination of the voxels that would yield a positive correlation with animacy if the correlation were computed over the entire population of object images. Even if this null hypothesis were true, we would be able to decode animacy on an independent test set of responses to the same stimuli. This is because of overfitting to the stimulus set (although overfitting to the noise is avoided by using independent data). However, linear decoding of animacy would perform at chance

[2] This does not mean that the representations are invariant to the difference between training and test scenario and that the categories are associated with the same activity patterns in both scenarios. It only shows that a linear readout mechanism can provide the cross-decoded information with invariance to the scenario. Training and test patterns could vary along an orthogonal (or approximately orthogonal) dimension. This latter possibility can be investigated by attempting to decode the scenario or by inspecting and visualizing the response-pattern dissimilarities for all pairs of stimuli in either scenario in the framework of representational similarity analysis (Kriegeskorte et al., 2008b).

[3] Linear readout is a key concept here. Because the region distinguishes all stimuli, there necessarily is a complex nonlinear classifier that discriminates animates and inanimates, or any arbitrary division of the stimulus space into two subsets.

level for an independent random sample of stimuli – leading us to correctly accept the null hypothesis.[4]

Beyond pattern classification with separate training and test sets, we can perform generalization to a population of stimuli by using a sufficiently large random sample of stimuli and modeling the variability of responses across individual stimuli. This is analogous to analyses that treat subject as a random factor so as to generalize to the human population. Generalization to novel stimuli is a central theme also in the next section, where generic models of the stimulus-response relationship are replaced by neuroscientifically motivated computational models.

(2) Goal 2: Testing computational models of brain information processing

Several recent pattern-information studies have gone beyond testing for the presence of information (goal 1) and tested computational models that mimic brain information processing (goal 2). The methods described in this section essentially test whether a computational model correctly predicts what information is present and what information is absent, or, in other words, what dimensions of the stimulus space the representation is sensitive to and what dimensions it is invariant (or less sensitive) to.

I focus on three fMRI studies by Mitchell et al. (2008), Kay et al. (2008), and Kriegeskorte et al. (2008a, 2008b). In addition to incorporating neuroscientifically motivated computational models, these studies are similar in that they treat every stimulus as a separate condition (Fig. 2), sample the stimulus-space more richly than previous fMRI studies, and attempt to generalize to the stimulus population that the experimental stimuli can be considered a random sample of.

Mitchell et al.: Predicting brain response patterns for novel stimuli

Mitchell et al. model the brain representation of noun concepts by means of 25 semantic features. Subjects were presented with word-picture pairs to evoke the representations of the noun concepts. Each of the 25 semantic features of the representational model measures the co-occurrence frequency of the input noun with one of 25 manually selected verbs. (Co-occurrences between the nouns and the 25 verbs were counted in a trillion-word text corpus.) The model is based on current theory about semantic brain representations. Each voxel's response is modeled as a linear combination of the 25 features. The model is fitted using a training set of 58 nouns. The fitted model predicts brain response patterns for arbitrary nouns, for which responses have not yet been measured. Prediction, thus, is a meaningful claim here and requires no quotes.

In order to demonstrate that response-pattern prediction works better than chance, the authors use the predicted patterns to identify a novel noun among two novel alternative nouns. To this end, the novel noun's measured pattern (not used for fitting the model) is matched up to the more similar one of the two nouns' predicted patterns. This identification

[4] In fact, early visual cortex does appear to allow above-chance-level linear readout of animacy even for an independent test set of different stimuli (Misaki et al., 2010). This suggests either category differences in low-level image statistics or feedback from higher regions that distinguish the categories. However, representational similarity analysis shows that response-pattern dissimilarities are only slightly larger between than within the two categories in early visual cortex. In the ventral stream, by contrast, animacy is a major variance-explaining factor, and animates and inanimates fall into separate response-pattern clusters (Kriegeskorte et al., 2008a).

among two novel nouns is shown to work better than chance (77% correct on average across subjects). Better-than-chance stimulus identification implies better-than-chance response-pattern prediction. In other words, the response pattern predicted for a novel noun tends to be more similar to the measured pattern for that noun than to measured patterns for other novel nouns – justifying the authors' title claim.

What do we learn about brain function from this study? The study provides some support for the neuroscientific model of semantic representation it is based upon. The model posits that noun-concept representations are similar to the extent that the nouns tend to co-occur with the same verbs. The study's title focuses on prediction of brain activity patterns, and this claim is entirely justified. However, better-than-chance prediction is a low bar and might be obtained with many competing models, each of which may explain a portion of the response-pattern variance.

For example, the competing category-representation model mentioned in the introduction of the paper (but not tested with these stimuli to my knowledge) is also expected to predict activity patterns for novel nouns better than chance based on previous studies: Haxby et al. (2001) showed that ventral-temporal patterns reflect visual object category. Spiridon & Kanwisher (2002) showed that category-average patterns are consistent even when computed for different sets of particular object images. Kriegeskorte et al. (2008a) showed that the representation is inherently categorical with patterns forming natural clusters in response-pattern space that correspond to conventional categories. These findings also suggest generalization to novel stimuli.

It would be useful to directly compare the semantic and category models in the framework of Mitchell et al. (2008). To this end, we could simply predict the category-average pattern of the training examples of the same category (based on living versus nonliving, or on a more fine-grained categorical structure) for each novel noun. Whether this naive model would predict the response patterns for novel nouns better or worse than the semantic model is an open empirical question. Results of Kriegeskorte et al. (2008a) suggest that the ventral-temporal representation of object images combines a categorical and a continuous component. Perhaps the category and semantic models account for the categorical and a continuous component, respectively, and could be combined to form a more complete theoretical account. From a neuroscientific perspective, the key advance of Mitchell et al. (2008) lies in the computational implementation of the semantic representational model and in providing a method that will allow us to adjudicate among alternative computational models in the future.

Kay et al.: Identifying novel stimuli from brain response patterns

Kay et al. presented subjects with a large number of natural images (real-world photos) while measuring early visual cortex with fMRI. They model the brain representation of natural images in V1 as a set of units that combine the outputs of multiple Gabor-filters applied to the image. The Gabor filters span the space of visual-field locations, orientations, and spatial frequencies. The model is based on current theory about the visual representation in V1. Each unit corresponds to an fMRI voxel and its parameters are fitted to predict the voxel response across 1,750 training images. The fitting of the model is constrained by prior neuroscientific knowledge about the nature of the V1 representation. For the general methodological framework underlying this study, see also Naselaris et al. (this issue) and Gallant et al. (in press).

Like Mitchell et al.'s model, the model of Kay et al. predicts brain response patterns for arbitrary stimuli, for which responses have not yet been measured. Like Mitchell et al., Kay et al. use the predicted brain activity patterns to identify stimuli. However, they study how identification accuracy drops off as a function of the size of the set of novel images among which a given novel image is identified. For one of the subjects, for example, identification among 200 images based on just a single perceptual trial is correct in about 50% of the cases (chance level: 1/200).

Note that although the title claim of Mitchell et al. is prediction of responses and the title claim of Kay et al. is identification of stimuli, both studies perform both of these feats for novel stimuli – and by essentially the same method (though using different representational models, appropriate to the respective domains). In analogy to cell-recording studies, Kay et al. refer to their methodology as voxel receptive-field modeling (see also Dumoulin & Wandell, 2008). This reflects the fact that a separate model is fitted to predict the responses of each voxel. This aspect, too, is shared between the studies by Kay et al. and Mitchell et al., and I will therefore refer to both studies' methodology as voxel receptive-field modeling.

What do we learn about brain function from the Kay et al. study? The results are consistent with what is known about V1, namely that the representation is composed of detectors of Gabor-like small visual features varying in location, orientation, and spatial frequency. It further confirms our expectation based on previous studies that these features are reflected in fMRI patterns (e.g. Sereno et al., 1995; Singh et al., 2000; Kamitani & Tong, 2005). These studies clearly imply that it must be possible to use the fMRI information to identify a novel stimulus among alternatives with above-chance performance. Another previous study even reconstructed contrast-defined images based on fMRI patterns using a simpler modeling approach (Thirion et al., 2006, for a more advanced reconstruction techniques, see Miyawaki et al., 2008; Naselaris et al., 2009). However, Kay et al.'s ingenious combination of prior neuroscientific theory and generic statistical techniques constitutes an impressive engineering achievement relevant to the development of brain-computer interfaces.

From a neuroscientific perspective, Kay et al. find evidence for a model consistent with widely accepted theory and show that reduced versions of this model perform worse. The key neuroscientific contribution of this study lies in the methodology of voxel receptive-field modeling (similar to the approaches of Dumoulin & Wandell, 2008; and Mitchell et al., 2008), which promises tests of alternative computational models in the future.

Kriegeskorte et al.: Representational similarity structure matches between a brain region and model

Kriegeskorte et al. (2008a, 2008b) model the human inferior-temporal representation of visual object images by means of a range of conceptual and computational models. The models include category models that posit categorical distinctions without explaining how the representation is computed, naive computational transformations of the bitmap stimuli, and neuroscientifically motivated computational models for the primary visual representation (Gabor-based filters modeling simple and complex cells) and an inferior-temporal-level representation (intermediate-complexity natural image features) (Riesenhuber & Poggio, 2002; Serre et al., 2007). In addition, they include an animal model: the monkey inferior-temporal representation of object images as reflected in single-cell recordings (reanalyzing data from Kiani et al., 2007). The stimuli presented to human subjects and models are 92 photos of real-world objects.

In contrast to the voxel receptive-field modeling approaches of Kay et al. (2008) and Mitchell et al. (2008), Kriegeskorte et al. (2008) do not use the models to predict brain response patterns from stimuli or to identify stimuli from brain response patterns. Instead the models are used to predict the similarity structure of the stimuli in the brain representation. To this end, each of the 92 response patterns in a brain or model representation is compared to each other response pattern, so as to obtain a representational dissimilarity matrix (RDM). The RDM reflects to what extent the brain or model representation distinguishes each pair of stimuli. How well a model predicts a brain region's representational similarity structure is assessed by simply correlating the RDMs of the representations of the brain region and the model. This approach, called representational similarity analysis, is briefly described in Kriegeskorte (2009), and more in detail in Kriegeskorte et al. (2008a, b). The model-based analyses are complemented by data-driven techniques that visualize the representational similarity structure and reveal natural clusters of similar response patterns.

What do we learn about brain function from this study? The human and monkey inferior temporal object representations match closely in terms of their representational similarity structure. Data-driven analysis suggests that a hierarchical categorical structure is inherent to the representation. The top-level categorical distinction (explaining most variance) is animate versus inanimate. Within the animates, faces and bodies form subclusters. Previous studies (e.g. Haxby et al., 2001; Spiridon & Kanwisher, 2002) had built the assumption of a categorical structure into the experimental design and analysis and therefore could not assess inherent categoricality. The inherent categorical structure matches between man and monkey. Within each category cluster, exemplars are distinguished and the within-category pairwise representational dissimilarities are also correlated between man and monkey.

These findings provide substantial constraints for computational theory. A simple categorical model ignoring the finer distinctions and positing only that animate objects are distinguished from inanimate objects explains more dissimilarity variance than any other single model tested, including the inferior-temporal model based on intermediate-complexity natural image features. This reminds us of the limits of our current computational understanding of high-level object representations in inferior temporal cortex and suggests that the representation may utilize features optimized for distinguishing categories that are behaviorally important to primates.

Comparing voxel receptive-field modeling and representational similarity analysis

How does representational similarity analysis as used by Kriegeskorte et al. (2008a, 2008b) relate to voxel receptive-field modeling as used by Kay et al. (2008) and Mitchell et al. (2008)? Both methods test computational models of brain information processing on the basis of brain response patterns estimated for single stimuli.

The key difference is that voxel receptive-field modeling uses computational-model representations to predict the measured response patterns, whereas representational similarity analysis uses the model representations to predict response-pattern dissimilarities. From a technical perspective, this is a substantial difference. Representational similarity analysis avoids the challenge of predicting either the measured response patterns.[5] From a neuroscientific perspective, both approaches serve the same purpose: the testing and comparing of computational models.

[5] The computational models simulating brain information processing do have internal response patterns in representational similarity analysis. However the number of units of the model representation may differ from the number of measured responses (voxels in fMRI), and the model response patterns are not used to predict the measured response patterns.

Evaluating computational models by predicting activity patterns is complicated by the need to define the correspondency mapping between the features of the model and the measured responses (voxels or neurons). Voxel-receptive field mapping therefore requires a linear model (predicting the measured responses from the model representation) to be fitted with one data set and tested with a separate data set (different stimuli).

Testing a computational model by predicting pattern dissimilarities instead of the patterns themselves greatly simplifies the analysis. Representational similarity analysis can test a computational model directly with data from a single stimulus set (no separate training and test sets required), because the pattern dissimilarity matrices of model and brain region are indexed (horizontally and vertically) by the stimuli and can be compared without fitting a linear model. If the stimulus set is a random sample from a population of stimuli, then appropriate statistical inference on the correlation between model pattern dissimilarities and brain pattern dissimilarities can generalize to the stimulus population. Representational similarity analysis requires separate training and test data sets (based on different stimulus samples) only if the computational model has parameters to be fitted to the brain-activity data.

Kay et al. (2008) estimate the maximum possible proportion of pattern variance that any model can explain given the noise in the data. This provides a helpful reference frame for evaluating the quality of a model. In representational similarity analysis, we can similarly estimate the maximum possible proportion of pattern-dissimilarity variance that any model can explain (or equivalently the noise floor; Fig. 8 in Kriegeskorte et al., 2008b).

The fact that the two methods use different criteria to test a computational model (predicting patterns versus predicting pattern dissimilarities) suggests that results might diverge and that the neuroscientific interpretation should perhaps be different. In fact, the two criteria are very closely related. It is easy to see that a match of response patterns between model and brain region implies a match of response-pattern dissimilarities. Conversely, if a computational model accounts for the response-pattern dissimilarities of a brain representation, then we can also use this model to predict the response patterns themselves for novel stimuli (Fig. 3). We can achieve this, for example, by interpolating among the response patterns of the training stimuli that are closest to the novel stimulus in the model representation. In practice, this approach may require many training stimuli, and more sophisticated techniques may achieve better response-pattern prediction. The close relationship between patterns and pattern dissimilarities suggests that, although the two techniques are not mathematically equivalent, the qualitative empirical claims they test, and thus the implications for brain theory, are essentially the same (Fig. 4, bottom box).

(3) Exploratory analysis of population activity patterns

Goals 1 and 2 cover the hypothesis-driven side of pattern-information analysis. Testing for information about a particular stimulus dimension in regional response patterns (goal 1) is driven by a hypothesis about the stimulus dimension represented in a brain region. Testing a computational model (goal 2) is driven by the hypothesis that the model explains the data for a brain region. For both goals, the hypothesis may also specify the brain region to be analyzed.

In cognitive neuroscience, a popular approach is to contrast two competing theories and use the data to decide between them. This is an excellent approach when all other theories can really be excluded a priori. More often, the two theories are merely two points in a much larger space of similarly plausible possibilities that have yet to be tested empirically. In that case, the focus on two theories creates a veneer of conceptual rigour and clarity, but the decisiveness of the experiment is just a fantasy.

Hypothesis-driven analysis is like looking at a scene through a drinking straw:[6] It narrows our perspective on the data. This is useful when we already know what might be going on and when the straw is pointed at the crucial part of the scene. However, we certainly don't get to see the big picture. Exploratory analysis can widen our perspective and keep us more broadly in touch with the data.

As a complement to the hypothesis-driven analyses serving goals 1 and 2, we therefore need more exploratory analyses. Such techniques are data-driven in that their results are determined to a greater degree by the data and to a lesser degree by prior assumptions. We can think of exploration as searching of a space of hypotheses. This perspective shows that there is a continuum between confirmatory and exploratory analysis: We can make our analyses more exploratory by simply applying our hypothesis-driven methods to a greater number of hypotheses. When we fit the parameters of a complex model to the data, we explore a continuous space of candidate models.

While the hypothesis-driven approach of pattern classification (serving goal 1) can very sensitively detect small amounts of information about predefined stimulus categories, it can miss major variance-explaining alternative dimensions of the stimulus space. This motivates us to minimize the assumptions built into the design of the experiment. While block-designs and conventional event-related designs typically assume a stimulus grouping, ungrouped-events designs (Kriegeskorte et al., 2008b; see also Aguirre 2007) avoid this assumption. An ungrouped-events design enables us to discover stimulus dimensions represented in a given region, for example using data-driven multivariate techniques like multidimensional scaling (for applications to fMRI data see Edelman et al., 1998; Tagaris et al., 1998; O'Toole et al., 2005; Kriegeskorte et al., 2008a, 2008b).

Conversely, we can look for regions representing a given stimulus dimension (or conforming to a given computational model), for example with a searchlight approach (Kriegeskorte et al., 2006). This prevents us from overinterpreting a weak effect in a predefined region when stronger effects are present elsewhere. It can also lead us to discover regions in unexpected locations in the brain (Kriegeskorte et al., 2007; Haynes et al., 2007). More generally, complementing a hypothesis-driven approach by exploratory analysis helps keep our theory consistent with the major variance-explaining factors in the data.

(4) Stimulus reconstruction

Pattern-classifier decoding, distinguishing small number of categories (typically two), captures only a tiny subset of the information we expect to be present in a brain representation. Reconstruction is decoding without such limitation (although current reconstruction methods still somewhat restrict the space). Being able to decode arbitrary mental content from a brain representation is arguably the ultimate test of our understanding of the code. Several studies have attempted to reconstruct stimuli from fMRI response patterns (Thirion et al., 2006; Miyawaki et al., 2008; Naselaris et al., 2009).

How does stimulus reconstruction go beyond stimulus identification? The difference is (a) the set of stimuli to choose among is much larger and (b) to justify the term reconstruction,

[6] Regional-average activation analysis additionally mounts a blurring lens to the end of the straw.

we require much more than merely better-than-chance performance. Let’s consider this in a little more detail. If we can predict response patterns from stimuli, then we can in principle enumerate the stimulus set (or sample a continuous space at a finite number of locations), predict the response pattern for each stimulus, and choose the best match to the measured response pattern, thus "reconstructing" by applying identification (as in Kay et al., 2008) to a very large set of candidate stimuli. Better-than-chance response pattern prediction therefore logically implies that better-than-chance stimulus reconstruction is possible.

However, this is like saying that if we can recognize a good poem, then we can write one: by enumerating all possible letter sequences and selecting the first good poem. While true in principle, the method does not work in practice. For one thing, it takes too long (for both poems and general stimulus reconstruction). Moreover, while in neuroscience we are often satisfied with better-than-chance performance (i.e. a significant result: the model explains at least some of the variance), literature and engineering have higher standards: a merely better-than-chance poem is not likely to be a good one and a merely better-than-chance stimulus reconstruction is not likely to deserve being called a reconstruction at all.

The cited studies have shown that reconstructions that deserve to be called such are possible from fMRI data. Thirion et al. (2006) and Miyawaki et al. (2008) restrict the stimulus space to contrast-defined images of low spatial complexity, but allow arbitrary images within this space. Thirion et al. (2006) perform reconstruction by means of a point-by-point inversion of the retinotopic mapping. Miyawaki et al. (2008) use a multiple-module decoder, in which each module decodes one of multiple overlapping image features from the joint response of multiple voxels. The reconstruction, thus, exploits multivariate relationships within both the stimulus and the response domain. Naselaris et al. (2009) reconstruct natural images using complex natural-image priors to constrain the problem.

Brain theory informs engineering in stimulus reconstruction. Generic statistical methods by themselves would not do nearly as well. Reconstruction requires the amalgamation of simplified brain theory and generic statistical methods into a model of just the right level of complexity to be stably fitted with the amount of data available to us. This engineering feat may be a good test of the sum of our current understanding of a given brain representation. Moreover, the technology may elevate brain-computer interfaces to a new level – with promising medical applications. What is less clear is how we can draw specific insights about brain information processing from stimulus reconstruction.

Conclusion

For goal 1 of testing a region for pattern information about a predefined stimulus dimension, we can use generic statistical models. Linear models are attractive because of their stability and interpretability. Pattern classification treats the stimulus space as categorical. This simplification is often helpful, but also limiting in terms of the questions that can be addressed. Classical multivariate techniques are attractive for modeling continuous relationships between stimulus and response patterns. For the more ambitious goal 2 of testing computational models of brain representations, we can use voxel receptive-field modeling or representational similarity analysis. These methods model every stimulus as a separate condition, account for the variability across individual stimuli, and can generalize to populations of stimuli. Hypothesis-driven techniques need to be complemented by exploratory analyses that allow discovery, for example of an unexpected brain region representing a given stimulus dimension or of an unexpected stimulus dimension represented in a given region.

Acknowledgment

I thank Kendrick Kay and Tom Mitchell for helpful comments on a draft of this manuscript.

References

Afraz, S.R., Kiani, R., Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature. 442*(7103):692-5.

Aguirre, G.K. (2007). Continuous carry-over designs for fMRI. *Neuroimage 35*, 1480-1494.

Averbeck, B.B., Latham, P.E., Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience 7*, 358-366.

Bedny, M., Aguirre, G. K., Thompson-Schill, S. L. (2007). Item analysis in functional magnetic resonance imaging. *Neuroimage* 35(3), 1093-1102.Mapp, 25(1), 155-164.

Carlson, T.A., Schrater, P., He, S. (2003). Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci.* 15, 704–717.

Clark, H.H. (1973). The language as fixed effects fallacy: A critique of language statistics in psychological research. *J. Verb. Learn & Verb. Behav.* 12, 335-359.

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

Deisseroth, K., Feng, G., Majewska, A.K., Miesenböck, G., Ting, A., Schnitzer, M.J. (2006). "Next-generation optical technologies for illuminating genetically targeted brain circuits". *J. Neurosci.* 26 (41): 10380–6.

Duda, R.O., Hart, P.E., Stork, D.G. (2001). *Pattern Classification.* New York, NY: John Wiley and Sons.

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

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

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

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

Gallant, J.L., Nishimoto, S., Naselaris, T., Wu, M.C.K. (in press). System identification, encoding models and decoding models, a powerful new approach to fMRI research. In *Understanding visual population codes – toward a common multivariate framework for cell recording and functional imaging*. Kriegeskorte, N., Kreiman, G. (Editors) MIT Press.

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

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

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

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

Hung, C.P., Kreiman, G., Poggio, T., DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science* 310, 863–866.

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

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

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

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

Kriegeskorte, N. (2004). Functional magnetic resonance imaging of the human object-vision system. *PhD Thesis*. Universiteit Maastricht.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences USA, 103*, 3863-3868.

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 USA, 104*, 20600-20605.

Kriegeskorte, N., Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *Neuroimage, 38*, 649-662.

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

Kriegeskorte, N., Mur, M., Bandettini, P.A. (2008b). Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* doi:10.3389/neuro.06.004.2008.

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

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

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* 3, 3: 363–373. doi: 10.3389/neuro.01.035.2009

Krzanowski, W. J. (1988). *Principles of Multivariate Analysis: A User's Perspective.* Clarendon Press: Oxford.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X. (2005). Support vector machines for temporal classification of block design fMRI data. *Neuroimage, 26*, 317-329.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *Neuroimage* 53(1):103-18. Epub 2010 May 23.

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

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

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

Mur, M., Bandettini, P., Kriegeskorte, N. (2009). Revealing Representational Content with Pattern-Information fMRI – an Introductory Guide. *Social Cognitive and Affective Neuroscience 4*(1): 101-9.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. *Neuron*. 63(6):902-15.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant J.L. (this issue). Computational approaches to fMRI: a comparison of encoding and decoding. *Neuroimage*.

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

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

O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci* 17: 580–590.

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

Raizada, R.D.S., Kriegeskorte, N. (in press). Pattern-information fMRI: new questions which it opens up, and challenges which face it. International *Journal of Imaging Systems and Technology.*

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

Riesenhuber, M., Poggio, T. (2002). Neural mechanisms of object recognition. *Curr. Opin. Neurobiol.* 12, 162–168.

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

Sereno, M. I. et al. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268, 889–893.

Serre, T., Oliva, A., Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. USA* 104, 6424–6429.

Singh, K. D., Smith, A. T. Greenlee, M. W. (2000). Spatiotemporal frequency and direction sensitivities of human visual areas measured using fMRI. *Neuroimage* 12, 550–564.

Soon CS, Brass M, Heinze HJ, Haynes JD. Unconscious determinants of free decisions in the human brain. *Nat Neurosci* 11(5): 543-5.

Spiridon, M., Kanwisher, N. (2002). How distributed is visual category information in human occipito-temporal cortex? An fMRI study. *Neuron* 35(6): 1157-65.

Stokes, M., Thompson, R., Cusack, R., Duncan, J. (2009). Top-down activation of shape-specific population codes in visual cortex during mental imagery. *J Neurosci* 29:1565–1572.

Tagaris, G. A., Richter, W., Kim, S. G., Pellizzer, G., Andersen, P., Ugurbil, K., Georgopoulos, A. P. (1998). Functional magnetic resonance imaging of mental rotation and memory scanning: a multidimensional scaling analysis of brain activation patterns. *Brain Res. Rev.* 26, 106–112.

Thirion, B. et al. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage* 33, 1104–1116.

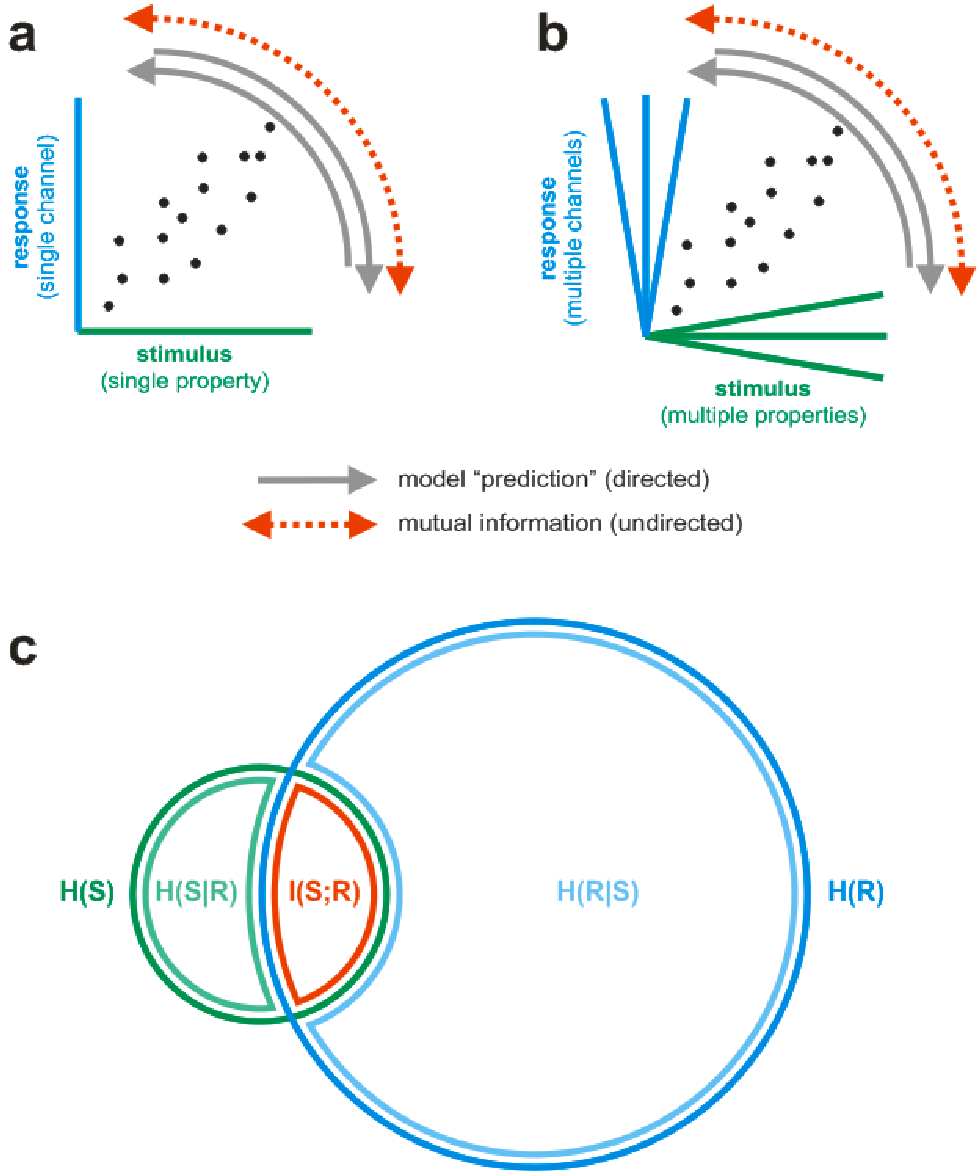

Figure 1: Decoding and encoding models both demonstrate a statistical dependency between stimulus and response pattern, but elucidate complementary aspects of the stimulus-response relationship. (**a**) In a univariate scenario, it is easy to see that a correlation between two variables implies predictability in both directions. More generally, we can say that there is *mutual information* (or, equivalently, a *statistical dependency*) between stimulus and response. (**b**) In a multivariate scenario, the same holds. In either case (a,b), demonstrating above-chance predictability in either direction implies mutual information and, thus, above-chance predictability in the opposite direction. (**c**) The mutual information can be construed as the entropy overlap (red) between the stimulus entropy H(S) (green) and the response-pattern entropy H(R) (blue). This perspective reminds us of two important facts: (1) Stimulus and response entropies are not equal in general. In this illustration the stimulus entropy is somewhat smaller. In decoding studies, the stimulus variable often specifies, which of two equally probable categories the stimulus belongs to. In that case, H(S) is 1 bit. H(R) is typically much greater. (2) Encoding and decoding models elucidate complementary aspects of the stimulus-response relationship. An encoding model can demonstrate, to what extent the stimulus description suffices to explain the response patterns. Conversely, a decoding model can demonstrate, to what extent the response pattern suffices to determine the stimulus that elicited it. In either case the estimate will be a lower bound (since we do not know that the model is optimal). We can estimate the portion of the response variability explained by the stimulus (or vice versa) in terms of either explained entropy (i.e. mutual information), or, more simply, in terms of explained variance.

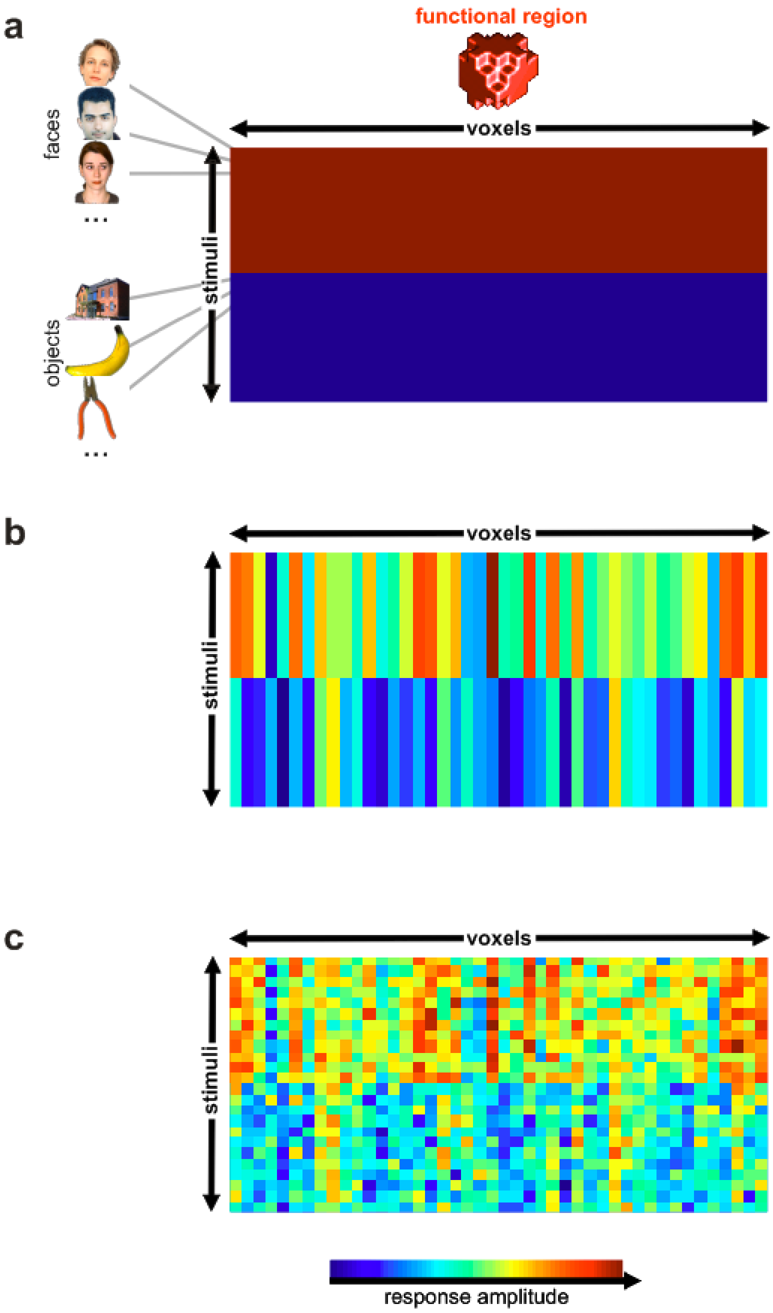

Figure 2: Advanced pattern-information methods reveal the forest while honoring the trees. The three panels show stimulus-response matrices with warm colors indicating high activity and cold colors indicating low activity. Different approaches to analysis (panels a-c) group and average the data to different degrees. (**a**) Activation-based neuroimaging relies on averaging of activity across the voxels of a given brain region and typically also across different stimuli within a given experimental condition. The activation-based paradigm has been successful in revealing the big picture of task-related regional activation; it has shown us the forest – at the expense of honoring the trees. (**b**) Pattern-information-based neuroimaging analyzes patterns of activity across voxels. It honors the differences between individual voxels (trees), while combining the evidence across voxels in order to summarize the information (forest) and in order to gain statistical power. However, the popular approach of pattern classification still requires stimuli to be grouped into a small number of predefined categories. (**c**) Advanced pattern-information methods, including voxel-receptive-field modeling and representational similarity analysis, additionally honor the distinctions between individual stimuli (even for large numbers of stimuli). Although they honor distinctions between individual voxels (trees) and between individual stimuli (a different sort of trees), they combine the evidence across both voxels and stimuli (to summarize and to gain statistical power) when testing computational models (forest).

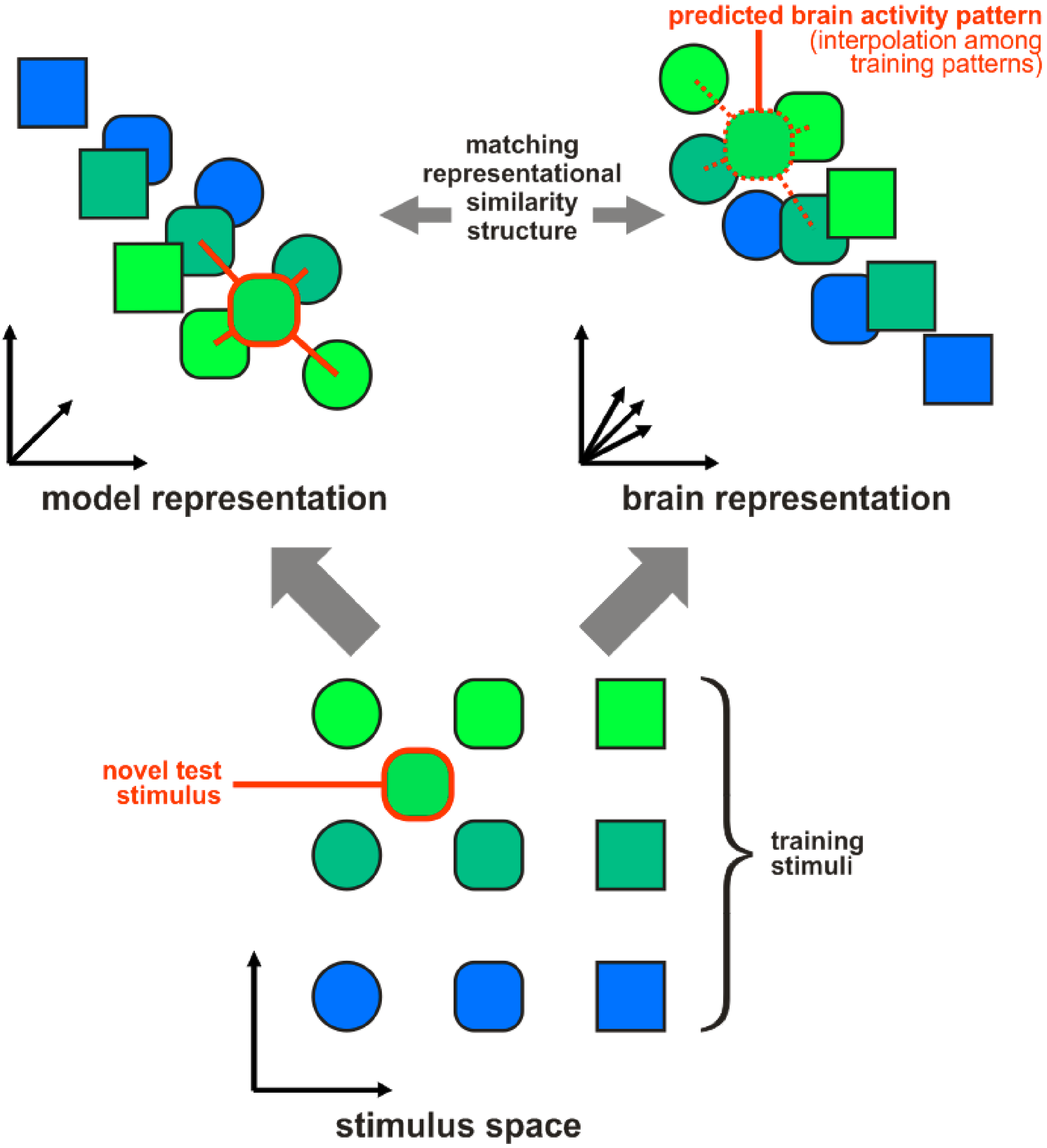
predicted brain activity pattern
(interpolation among training patterns)
matching representational similarity structure
model representation
brain representation
novel test stimulus
training stimuli
stimulus space

Figure 3: If a model accounts for the representational similarity structure of a brain representation, then it can in principle be used to predict brain response patterns for novel stimuli. The stimulus space (bottom) spans two dimensions (color and shape) and is sampled here by training stimuli placed at regular intervals. The brain representation (top right) emphasizes a diagonal dimension of the stimulus space (the axis from top left to bottom right in the stimulus space) and deemphasizes the orthogonal dimension. A model representation (top left) accurately mimics the similarity structure of the brain representation, i.e. the representational distances are similar between model and brain representation. We can test for matching representational similarity structures by simply correlating the representational distance matrices (not shown). If the representational similarity structure matches, then we can also predict response patterns for novel test stimuli. This can be achieved by interpolating among the response patterns of the training stimuli closest to the novel stimulus in the model representation, or by fitting a generic statistical model. Note that the dimensions of the model representation need not correspond to the dimensions of the brain representation. For example, the model could contain a much smaller number of abstract units, which nevertheless capture the representational similarity structure.

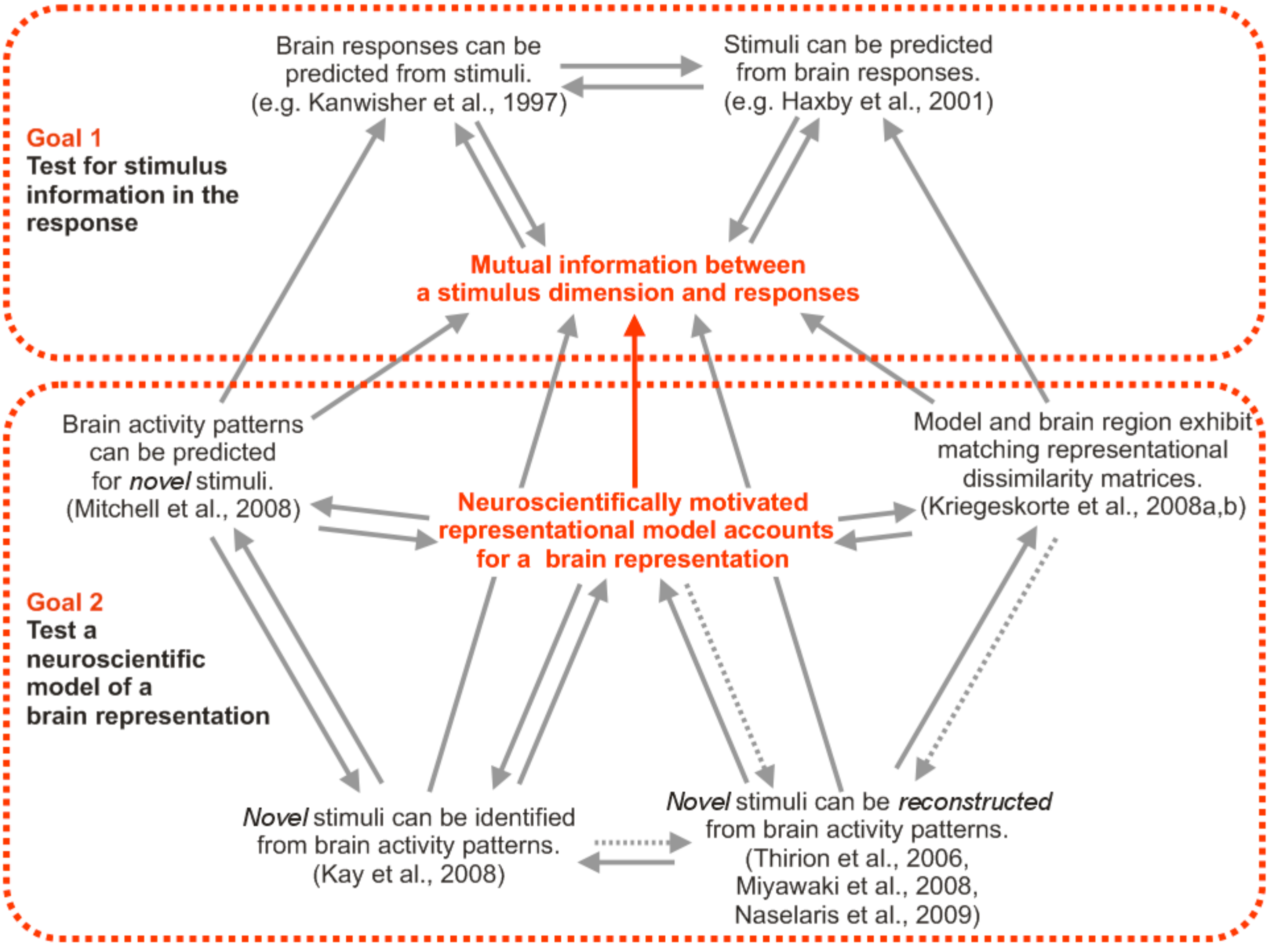
Brain responses can be predicted from stimuli. (e.g. Kanwisher et al., 1997)
Stimuli can be predicted from brain responses. (e.g. Haxby et al., 2001)
Goal 1
Test for stimulus information in the response
Mutual information between a stimulus dimension and responses
Brain activity patterns can be predicted for *novel* stimuli. (Mitchell et al., 2008)
Neuroscientifically motivated representational model accounts for a brain representation
Model and brain region exhibit matching representational dissimilarity matrices. (Kriegeskorte et al., 2008a,b)
Goal 2
Test a neuroscientific model of a brain representation
Novel stimuli can be identified from brain activity patterns. (Kay et al., 2008)
Novel stimuli can be *reconstructed* from brain activity patterns. (Thirion et al., 2006, Miyawaki et al., 2008, Naselaris et al., 2009)

Figure 4: Neuroscientific implications of different classes of pattern-information result. This figure summarizes the logical implications (arrows) of different types of pattern-information result. The claims form two clusters of logical equivalence, each of which corresponds to a type of neuroscientific insight (bold red font) and to one of the two goals of hypothesis-driven pattern-information analysis (as stated on the left). Each class of empirical finding is characterized by a generic claim (black font), followed by one or several example studies. The implications hold for the generic claims, but not for the particular studies, because the studies differ in stimuli, designs, and specific questions (e.g. Kanwisher et al. (1997) does not imply Haxby et al. (2001) or vice versa). Each empirical claim is assumed to be justified if it holds to a greater degree than expected by chance (i.e. if a significance test rejects an alternative null hypothesis). The implication arrows pointing toward the claim "*Novel* stimuli can be *reconstructed* from brain activity patterns" (bottom right) are dashed to indicate that although the other statements imply that stimulus reconstruction will work better than chance, our intuitive criterion for successful stimulus reconstruction requires reconstructions of much higher quality.

<table>
<tr><th></th><th>stimulus decoding with response-pattern classifier</th><th>cross-decoding with response-pattern classifier</th><th>voxel receptive-field (RF) modeling</th><th>stimulus reconstruction</th><th>representational similarity analysis</th></tr>
<tr><td><b>example studies</b></td><td>Haxby et al., 2001; Kamitani & Tong, 2005</td><td>Polyn et al., 2005; Stokes et al., 2009</td><td>Kay et al., 2008; Mitchell et al., 2008</td><td>Miyawaki et al., 2008; Naselaris et al., 2009</td><td>Kriegeskorte et al., 2008a, 2008b</td></tr>
<tr><td><b>model direction</b></td><td colspan="2">decoding</td><td>encoding</td><td>decoding</td><td>symmetrical</td></tr>
<tr><td><b>stimulus sample</b></td><td colspan="2">few stimuli (or many stimuli grouped into few categories for classification)</td><td colspan="3">many stimuli (every one of which is treated as a distinct entity)</td></tr>
<tr><td><b>stimulus space generalized to</b></td><td>no generalization to new stimuli (generalization only to new responses elicited by the same particular stimuli)</td><td>generalization from one set of stimuli (or conditions) to an alternative set (e.g. perception to imagery)</td><td colspan="3">generalization to the theoretical population of stimuli
(of which the experimental stimuli are a random sample)</td></tr>
<tr><td><b>general stimulus reconstruction?</b></td><td colspan="3">no</td><td>yes</td><td>no</td></tr>
<tr><td><b>test of computational models of brain information processing?</b></td><td>no</td><td>no</td><td>yes</td><td>no</td><td>yes</td></tr>
<tr><td><b>strengths</b></td><td>very sensitive detection of small amounts of information about predefined stimulus categories</td><td>testing whether representations are consistent across different stimulus sets or tasks that invoke them</td><td>accurate prediction of activity patterns for novel stimuli utilizing prior neuroscientific knowledge;
tests of different computational RF models</td><td>high-quality, generalized decoding utilizing prior neuroscientific knowledge and multivariate response relationships</td><td>discovery of major variance-explaining stimulus dimensions; tests of conceptual and computational models; relating brain regions, individuals and species, and brain to behavior</td></tr>
<tr><td rowspan="2"><b>weaknesses</b></td><td colspan="2">predefined category grouping may be artificial and may miss major variance-explaining factors</td><td rowspan="2">difficult to apply to higher regions, where computational models are lacking or have prohibitive parameter complexity for RF-model fitting</td><td rowspan="2">engineering,
not neuroscience focus:
unclear how to test theories or draw specific neuroscientific conclusions</td><td rowspan="2">neuroscience,
not engineering focus:
no prediction of activity patterns or decoding</td></tr>
<tr><td>no generalization to novel stimuli</td><td>limited generalization to novel stimuli</td></tr>
</table>

Figure 5: Comparison of a range of pattern-information methods. This figure compares a range of basic and advanced methods of pattern-information analysis.