Published in final edited form as:
Neuroimage. 2011 May 15; 56(2): 400–410. doi:10.1016/j.neuroimage.2010.07.073.

Encoding and decoding in fMRI

Thomas Naselaris[a], **Kendrick N. Kay**[b], **Shinji Nishimoto**[a], and **Jack L. Gallant**[a,b]

[a]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA

[b]Department of Psychology, University of California, Berkeley, CA 94720, USA

Abstract

Over the past decade fMRI researchers have developed increasingly sensitive techniques for analyzing the information represented in BOLD activity. The most popular of these techniques is linear classification, a simple technique for decoding information about experimental stimuli or tasks from patterns of activity across an array of voxels. A more recent development is the voxel-based encoding model, which describes the information about the stimulus or task that is represented in the activity of single voxels. Encoding and decoding are complementary operations: encoding uses stimuli to predict activity while decoding uses activity to predict information about stimuli. However, in practice these two operations are often confused, and their respective strengths and weaknesses have not been made clear. Here we use the concept of a linearizing feature space to clarify the relationship between encoding and decoding. We show that encoding and decoding operations can both be used to investigate some of the most common questions about how information is represented in the brain. However, focusing on encoding models offers two important advantages over decoding. First, an encoding model can in principle provide a complete functional description of a region of interest, while a decoding model can provide only a partial description. Second, while it is straightforward to derive an optimal decoding model from an encoding model it is much more difficult to derive an encoding model from a decoding model. We propose a systematic modeling approach that begins by estimating an encoding model for every voxel in a scan and ends by using the estimated encoding models to perform decoding.

Keywords

fMRI; encoding; decoding; linear classifier; multivoxel pattern analysis; computational neuroscience

1. Overview

The goal of many fMRI studies is to understand what sensory, cognitive or motor information is represented in some specific region of the brain. Most current understanding has been achieved by analyzing fMRI data from the mirror perspectives of encoding and decoding. When analyzing data from the encoding perspective, one attempts to understand how activity varies when there is concurrent variation in the world. When analyzing data from the decoding perspective, one attempts to determine how much can be learned about

Corresponding author: Jack L. Gallant 3210 Tolman Hall #1650 University of California Berkeley, CA 94720-1650 Tel: (510) 642-2606 FAX: (510) 542-5293 gallant@berkeley.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

the world (which includes sensory stimuli, cognitive state, and movement) by observing activity.

Associated with each perspective are a host of computational techniques. On the encoding side, voxel-based *encoding models* (Dumoulin and Wandell, 2008; Gourtzelidis et al., 2005; Jerde et al., 2008; Kay et al., 2008; Mitchell et al., 2008; Naselaris et al., 2009; Schönwiesner and Zatorre, 2009; Thirion et al., 2006) have recently emerged as a promising computational technique. Voxel-based encoding models predict activity in single voxels that is evoked by different sensory, cognitive or task conditions. Thus, encoding models provide an explicit, quantitative description of how information is represented in the activity of individual voxels. (Throughout the rest of this review we often refer to encoding models rather than voxel-based encoding models, but it should be understood that all of the encoding models discussed here are voxel-based.)

Many conventional data processing pipelines estimate a restricted form of an encoding model. For example, the statistical parametric mapping (SPM) approach developed by Friston et al. (1995) begins by fitting a general linear model (GLM) to each voxel within an ROI. In the SPM approach the parameters of the GLM are directly related to the levels of the independent variables manipulated in the experiment. The GLM parameters are estimated for each voxel. Statistical significance of the GLM is then assessed for each voxel and aggregated across an ROI. The GLMs estimated for individual voxels could theoretically be used to predict the activity in the voxels, so GLMs can be viewed as encoding models. In this paper we show that encoding models can provide more information about the features represented by specific voxels than can be obtained by using conventional approaches.

On the decoding side, the most commonly used computational technique is the linear classifier (see De Martino et al., 2008; Formisano et al., 2008a; Haynes and Rees, 2006; Haynes, 2009; Hansen, 2007; Mitchell et al., 2004; Norman et al., 2006; O'Toole et al., 2007; Pereira et al., 2009 for reviews; see Kippenhan et al., 1992 and Lautrup et al., 1994 for early examples). The linear classifier is an algorithm that uses patterns of activity across an array of voxels to discriminate between different levels of stimuli, experimental or task variables. Because classifiers exploit systematic differences in voxel selectivity within a region of interest (ROI), in principle they can detect information that would be missed by conventional analyses that involve spatial averaging (Kriegeskorte and Bandettini, 2007).

A linear classifier can be viewed as one specific and restricted form of a *decoding model*, a model that uses voxel activity to predict sensory, cognitive, or motor information. Decoding models may also be used to perform identification (Kay et al., 2008) and reconstruction (Miyawaki et al., 2008; Naselaris et al., 2009; Thirion et al., 2006). In identification, patterns of activity are used to identify a specific stimulus or task parameter from a known set. In reconstruction, patterns of activity are used to produce a replica of the stimulus or task. These more general forms of decoding are themselves special cases of multi-voxel pattern analysis, which encompasses many unsupervised methods for analyzing distributed patterns of activity (Kriegeskorte and Bandettini, 2007; Kriegeskorte et al., 2008a; Kriegeskorte et al., 2008b).

Encoding and decoding models are complementary (Kay and Gallant, 2009). However, the relationship between these two types of model has rarely been discussed in the context of fMRI (but see and Dayan and Abbot, 2001 for a more general discussion of this issue). In this article we provide a conceptual overview of the relationship between encoding and decoding models. We clarify the relationship by invoking an abstract linearizing feature space that describes how stimulus, experimental or task variables are nonlinearly mapped into measured activity. We then present a critical comparison of encoding and decoding models that answers several fundamental questions about their relative utility for fMRI. Is there any difference between the sensory or cognitive representations that can be studied with encoding and decoding models? Are there any advantages to using either type of model? Are there any contexts in which it is appropriate to use both types of model?

2. Encoding models and the linearizing feature space

To illustrate the relationship between encoding and decoding we will discuss a concrete example, a recent study from our laboratory that investigated how natural scenes are represented in the early and intermediate visual system (Kay et al. 2008). The stimuli consisted of a long series of briefly flashed gray-scale natural scenes. BOLD activity (hereafter referred to as “voxel activity”) evoked by these scenes was measured in voxels located near the posterior pole, including areas V1, V2, V3, V4, LO and MT+. To interpret the data Kay et al. constructed models of individual voxels that described the information about natural scenes represented by the voxel activity. To model voxels they first mapped stimuli into an over-complete nonlinear basis consisting of many phase-invariant Gabor wavelets that varied in location, orientation, and spatial frequency. These Gabor wavelets reflected neural mechanisms known to exist at early stages of cortical visual processing (Adelson and Bergen 1985; Carandini et al., 2005; Jones and Palmer 1987). They then used linear regression to find a set of weights that mapped these Gabor features into responses of individual voxels. Kay et al. showed that these encoding models predicted responses to novel stimuli with unprecedented accuracy.

Encoding models like the one developed in Kay et al. consist of several distinct components. First is the set of stimuli (or the various task conditions) used in the experiment. In Kay et al. these stimuli were natural scenes drawn at random from a continuous distribution of natural scenes. However, most fMRI studies use stimuli drawn from discrete classes such as faces or houses (Downing et al., 2006), or they probe discrete levels of a cognitive variable such as the allocation of spatial attention to several different locations (Brefczynski and DeYoe, 1999). The second component is a set of features that describes the abstract relationship between stimuli and responses. In Kay et al. the features were phase-invariant Gabor wavelets. However, in most fMRI studies the features consist of labels that reflect different levels of the independent variable (e.g. faces versus houses, different locations of attention, etc.). The third component is one or more regions of interest (ROI) in the brain from which voxels are selected. The final component is the algorithm that is actually used to estimate the model from the data. In the Kay et al. study the model was estimated by linear regression of the Gabor wavelet outputs against the activity of each voxel.

An efficient way to visualize the components of encoding models like the one presented in Kay et al. is to think of the stimuli, features, and ROIs existing in three separate abstract spaces (Figure 1, middle). The experimental stimuli exist in an *input space* whose axes correspond the stimulus dimensions. For the Kay et al. study each axis of the input space corresponds to the luminance of one pixel, and each natural scene is represented by a single point in the input space. The activity of all the voxels within an ROI exists in an *activity space* whose axes correspond to the individual voxels. For the Kay et al. study the ROI includes the visual areas listed earlier, each axis of the activity space corresponds to a single voxel, and the pattern of activity across the ROI is represented by a single point in the activity space. Interposed between the input space and the activity space is an abstract *feature space*. Each axis of the feature space corresponds to a single feature, and each stimulus is represented by one point in the feature space. For the Kay et al. study the axes of the feature space correspond to the phase-invariant Gabor wavelets.

In Kay et al., the input, feature and activity spaces are linked together like a chain, where each link represents a mapping – a mathematical transformation – between spaces (Figure 1, middle). The mapping between the input space and the feature space is nonlinear, while the mapping between the feature space and the activity space is linear. The feature space is called *linearizing*, because the nonlinear mapping into feature space linearizes the relationship between the stimulus and the response (Wu et al. 2006). Encoding models based on linearizing feature spaces are referred to as *linearizing encoding models*.

Linearizing encoding models have a simple interpretation and are relatively easy to estimate. The mapping between the input space and the feature space is assumed to be nonlinear because most of the interesting computations performed by the brain are nonlinear. The mapping between feature space and activity space is assumed to be linear because the features that are represented by an ROI should have the simplest possible relationship to its activity. The nonlinear mapping is the same for each voxel; only the linear mapping has to be estimated from measured voxel activity. Thus, linearizing encoding models require only linear estimation. This can be performed by readily available algorithms for linear regression (Wu et al., 2006). Once estimated, the linear mapping between feature space and activity space describes the particular mix of features that evoke activity in each voxel.

As far as we know all of the encoding models that have been published in the field of fMRI thus far make use of a linearizing feature space. That is, they assume that there is a nonlinear mapping from the stimulus space to the feature space, and a linear mapping between the feature space and the activity space. We have already discussed the study of Kay et al. (2008) in detail. A subsequent study by Naselaris et al. (2009) reanalyzed the data collected as part of the Kay et al. study. However, Naselaris et al. constructed two different models for each voxel: a model based on phase-invariant Gabor wavelets, and a semantic model that was based on a scene category label for each natural scene. Naselaris et al. showed that the Gabor wavelet and semantic models predicted voxel activity equally well, but for different populations of voxels (Figure 2, right). The Gabor wavelet model provided good predictions of activity in early visual areas, while the semantic model predicted activity at higher stages of visual processing. Mitchell et al. (2008) also used a semantic encoding model based on a linearizing feature space. Their stimuli were labeled pictures of everyday objects, and the feature space consisted of co-occurrence measures between the object label and a set of 25 common verbs. Mitchell et al. showed that this semantic feature space accurately predicted voxel activity in several brain areas. These various fMRI studies are a natural outgrowth of a long line of neurophysiological experiments that have used linearizing feature spaces (Aertsen and Johannesma, 1981; Bredfeldt and Ringach, 2002; David et al., 2004; David and Gallant, 2005; Kowalski et al., 1996; Machens et al., 2004; Mazer et al., 2002; Nishimoto et al., 2006; Nykamp and Ringach, 2002; Ringach et al., 1997; Theunissen et al., 2000; Willmore et al. 2010; Wu et al., 2006).

Linearizing encoding models have an interesting interpretation as a means of hypothesis testing. Under this view any linearizing feature space reflects some specific hypothesis about the features that might be represented within an ROI. Testing an hypothesis with an encoding model simply requires estimating the linear mapping between the hypothesized feature space and measured voxel activity (i.e., activity space). For a single voxel, the linear mapping will consist of a weight for each feature. Once these weights are estimated the quality of the model can be examined by testing model predictions against a separate validation data set reserved for this purpose. If the feature space provides an accurate description of the mapping between stimuli and responses, then the linearizing model based on that feature space will accurately predict responses in the validation data set.

3. Decoding models and the linearizing feature space

Linearizing feature spaces are also helpful for thinking about decoding models. In these terms, the key difference between encoding and decoding models is the direction of the linear mapping between feature space and activity space. In an encoding model the linear mapping projects the feature space onto the activity space (Figure 1, middle). In a decoding model the linear mapping projects the activity space onto the feature space (Figure 1, bottom).

Consider as a concrete example the linear classifier study of Cox and Savoy (2003). The stimuli consisted of pictures of objects drawn from several different categories (birds, chairs, garden gnomes, horses, teapots and so on). Voxel activity evoked by these pictures was measured in both retinotopic and object-selective visual cortex. To interpret the data Cox and Savoy constructed several different types of classifiers that discriminated distributed patterns of voxel activity evoked by each category. Cox and Savoy provided an early demonstration that it is feasible to decode stimulus categories by applying classifiers to patterns of voxel activity.

The Cox and Savoy (2003) study has the same components as those found in studies using linearizing encoding models. The features are the levels of the independent variable, which in this case are stimulus category labels. The ROI consists of early and higher-order object selective visual cortex. The statistical algorithm used to fit the data is the linear classifier.

As with the Kay et al study, the components of the Cox and Savoy study can be described in terms of a linearizing feature space. The experimental stimuli are pictures, so in this case each axis of the input space corresponds to the luminance of one pixel and each picture is represented by a single point in the space. The ROI consist of much of visual cortex, so each axis of the activity space corresponds to a single voxel, and the pattern of activity across the visual cortex is represented by a single point in the space. The axes of the feature space correspond to the category labels assigned to specific subsets of the stimulus set (e.g., bird, teapots, etc.). Note that linear classification is a restricted form of decoding in which the decoded features are always discrete. Therefore, in a classification experiment the point corresponding to a decoded feature always lies along one of the axes in the feature space. Finally, the mapping from pictures (the input space) to category labels (the feature space) is highly nonlinear (DiCarlo and Cox, 2007), while the mapping from activity space to feature space is supplied by the linear classifier. Thus, the linear classifier of Cox and Savoy can be considered a *linearizing decoding model*. Although there is enormous distance between Kay et al. and Cox and Savoy in terms of the scientific questions addressed, both studies used linearizing models. As far as linearizing models are concerned, the most salient difference between the studies is the direction of the mapping between feature space and activity space.

Linearizing decoding models also have a simple interpretation and are relatively easy to estimate. The mapping between the input space and the feature space is assumed to be nonlinear because most of the interesting computations performed by the brain are nonlinear. The mapping between feature space and activity space is assumed to be linear because the features that are represented by an ROI should have the simplest possible relationship to its activity. Only the linear mapping has to be estimated using measured voxel activity. This can be performed by readily available algorithms for linear classification (Hastie et al., 2001). Two algorithms are used commonly (Misaki et al. 2010; Mur et al., 2009; Pereira et al., 2009). Linear discriminant analysis assumes that there is a linear relationship between the activity space and the feature space, but that the responses evoked within each class are Gaussian distributed (Hastie et al., 2001; Carlson et al. 2003). The linear support vector machine also assumes that there is a linear relationship between the activity space and the feature space, but it makes no assumptions about the distribution of responses within each class (Cox and Savoy 2003; Hastie et al., 2001). In both cases, the linear classifier aims to find a hyperplane in the response space that discriminates between the patterns of activity evoked under different stimulus, experimental or task conditions.

Some studies have used nonlinear classifiers that assume the relationship between the feature space and the activity space is nonlinear (Cox and Savoy, 2003; Davatzikos et al. 2005; Hanson et al. 2004). Nonlinear classifiers are not linearizing decoding models and their results are difficult to interpret. In theory a sufficiently powerful nonlinear classifier could decode almost any arbitrary feature from the information contained implicitly within an ROI (Kamitani and Tong, 2005). Therefore, a nonlinear classifier can produce significant classification even if the decoded features are not explicitly represented within the ROI. For example, suppose we measured retinal activity evoked by pictures of faces and houses, and we found that a nonlinear classifier could decode these two object categories from the measured activity. Although information about these categories is available implicitly in retinal activity, it would be incorrect to conclude that the retina represents these categories explicitly. It is widely accepted that an explicit representation of these categories arises in the brain only after a series of nonlinear mappings across several stages of visual processing (Felleman and Van Essen 1991). Any data analysis procedure that attributed this series of mappings to a single stage of processing would result in a serious error of interpretation (see Norman et al., 2006 for a similar argument). This kind of error of interpretation can be avoided by using linear classifiers.

Since the Cox and Savoy (2003) study, classification studies have become ubiquitous in fMRI, and we review several of these studies below. Decoding is not limited to classification, however. Other more general forms of decoding such as stimulus reconstruction can also be performed using linearizing decoding models. For example, the study of Miyawaki et al. (2008) used a linearizing decoding model to reconstruct flashing black-and-white geometrical patterns from activity measured in visual cortex. (See Thirion et al., 2006 for a similar study.) Their feature space reflected stimulus energy measured at different spatial locations and scales. They constructed a decoding model by using linear regression to estimate the mapping from voxel activity into the feature space. They then used the decoding model to reconstruct various geometric patterns. Ganesh et al. (2008) provides another example of a linearizing decoding model that performs reconstruction. They recorded surface electromyography (EMG) from muscles in the wrist while simultaneously recording voxel activity in motor cortex. The feature space consisted of the EMG traces recorded during an isometric tension task. They constructed a decoding model by using linear regression to estimate the mapping from voxel activity into feature space, and then used the decoding model to reconstruct EMG traces.

Decoding models can also be interpreted as a means of hypothesis testing. Once again the linearizing feature space reflects a specific hypothesis about the features that might be represented within an ROI, and the hypothesis testing strategy simply requires estimating the linear mapping between the hypothesized feature space and measured voxel activity. The linear relationship between the activity measurements and the features is estimated by linear regression or by using a linear classifier. In this case linear regression maps from the activity space to the feature space, so there will be one weight estimated for each voxel. Once the weights are estimated the quality of the model can be examined by classifying, identifying or reconstructing the features. If the feature space provides an accurate description of the nonlinear mapping between stimuli and voxel activity, then the linearizing model based on that feature space should accurately decode the features.

4. Comparing encoding and decoding models

The similarities between linearizing encoding and decoding models suggests that they might play similar roles in scientific investigations of the brain. To explore this issue we consider five questions that are commonly addressed in studies using encoding or decoding models. (1) Does an ROI contain information about some specific set of features? (2) Is the information represented within some ROI important for behavior? (3) Are there specific ROIs that contain relatively more information about a specific set of features? (4) What specific features are preferentially represented by a single ROI? (5) What set of features provides a complete functional description of an ROI?

4.1 Does an ROI contain information about some specific set of features?

A fundamental goal of any modeling effort is to establish whether an ROI represents any information at all about some specific set of features. To establish this, it is necessary to construct an encoding or decoding model whose prediction accuracy is significantly greater than chance. If an encoding model based on the set of features in question generates significantly accurate predictions for the voxels in an ROI, then it must be possible to decode some information about the features from voxel activity. If a decoding model generates significantly accurate predictions then it follows that the constituent voxels within the ROI must represent some information about the decoded features. Thus, when used as a tool to establish significance, there is in principle very little difference between the effectiveness of encoding and decoding models (Friston, 2009).

4.2 Is the information represented within some ROI important for behavior?

Significant prediction alone does not prove that behavioral performance related to a specific set of features depends critically on the ROI. It may therefore be important to test for direct relationships between patterns of activity and behavioral performance. It is difficult to use encoding models to do this. Encoding models produce predictions of activity, but it is difficult to interpret what accurate prediction of activity implies for behavior unless one also has a valid model that links activity to behavior. However, decoding models generate predictions about features or task outcomes, and these predictions can be directly compared to a subject's behavior (Raizada et al., 2009; Walther et al., 2009; Williams et al., 2007). Thus, an important advantage of decoding models is that they can be used to assess if the activity in an ROI is related to behavioral performance.

4.3 Are there specific ROIs that contain relatively more information about a specific set of features?

One way to show that an ROI contains more information about a specific set of features than can be found in some other ROI is to compare the predictions of encoding or decoding models across ROIs. To make this comparison with an encoding model, a model based on the specific features in question is estimated for each available voxel and prediction accuracy is calculated for each voxel individually (see below for a more detailed discussion of prediction accuracy). ROIs that contain relatively more information about the features in question should yield more accurate predictions. These ROIs can be identified by plotting prediction accuracy on a cortical map (Figure 2, left).

Comparisons between ROIs may also be performed using decoding models and linear classifiers. To compare ROIs with linear classifiers, the activity measured in each ROI is used to assess decoding accuracy with respect to some specific feature of interest. If one ROI yields significantly higher classification performance than the others, then it is legitimate to conclude that the ROI contains relatively more information about those specific features. (Schemes for systematically comparing multiple ROIs are discussed in

Kriegeskorte et al. 2006 and Pereira et al. 2009). Since the procedure for making comparisons across ROIs is the same whether encoding or decoding models are used, encoding and decoding models are likely to be equally useful tools for comparing ROIs.

4.4 Are there specific features that are preferentially represented by a single ROI?

Just as it is interesting to compare prediction accuracy for a single model across different ROIs, it is also interesting to compare the prediction accuracy of different models within a single ROI. Consider constructing two separate encoding models, one based on discrete features (e.g., semantic categories), and one based on continuous features (e.g., Gabor wavelets). Each model is estimated for all of the voxels in the ROI. The prediction accuracy of each model is then compared to determine which specific features—if any—are preferentially represented within the ROI. Because prediction accuracy is measured in terms of activity, it is directly comparable across these two models even though they are based on very different feature spaces (see Figure 2, right).

It is much more difficult to use decoding models to address this issue. Consider a linear classifier that decodes discrete features (e.g., semantic categories), and another decoding model that decodes continuous features (e.g., Gabor wavelets). For the linear classifier, the measure of prediction accuracy is percent correct classification. For the other decoding model, the measure of prediction accuracy is some metric appropriate for continuous quantities (e.g., Pearson's correlation). Because the discrete and continuous features require different measures of prediction accuracy, it is difficult to compare the two models to determine which set of features is preferentially represented. Thus, encoding models are better than decoding models for determining which set of features is preferentially represented within a specific ROI.

4.5 What set of features provides a complete functional description of an ROI?

A complete functional description of an ROI would consist of a list of all the features that it represents. Only encoding models can be used to obtain a complete functional description of an ROI, and this is one of the main differences between encoding and decoding models. To see why this is true, imagine trying to obtain a complete functional description of a specific ROI by constructing a series of encoding models, each based on a different feature space. Eventually, an optimal model that reflects just those features that are represented in the ROI would be constructed. This model would provide predictions that account for all of the explainable (*i.e.*, non-noise) variance in activity. Since there would be no more variance in the activity left to explain, it would not be necessary to test any more feature spaces. The features used to construct the optimal encoding model would constitute a complete list of all features represented in the ROI.

Now consider constructing a series of decoding models. Each decoding model is used to decode features in a different feature space. Eventually, a model that decodes perfectly the features in some feature space would be constructed. This would certainly be an important and interesting result. However, it would not indicate that a complete functional description had been achieved, as the features in some other feature space, yet untested, might also be perfectly decoded. In fact, there is no upper limit on the number of feature spaces whose features might potentially be decoded from an ROI. Thus, even after achieving perfect decoding for one feature space, it is still necessary to continue testing other feature spaces. Because there are an unlimited number of feature spaces that can be tested it is impossible to obtain a complete functional description by decoding alone.

5. Experimental designs that exploit the major advantage of encoding models

As our comparison of encoding and decoding shows, the major advantage of encoding models is that they can be easily compared to one another. By comparing multiple encoding models, it is possible to discover what features are preferentially represented by an ROI and it is even possible to discover the features that provide a complete functional description of an ROI. However, the conventional approach to experimental design in fMRI research does not lend itself to multiple model comparisons. The conventional approach is to select two or more discrete sets of stimuli or task conditions that correspond to the levels of an independent variable (Friston et al., 1995). The main hypothesis is that each level of the independent variable will evoke different levels of average activity within an ROI or different patterns of activity across an array of voxels. If activity evoked by the different levels of the independent variable are significantly different then the experimental results are judged to be consistent with the main hypothesis.

In the language of the linearizing feature space, the conventional approach is to select stimuli or task conditions that test whether the ROI contains information about a single feature space. The main hypothesis determines the feature space, and each axis of the feature space corresponds to one level of the independent variable. In this case, the encoding model will be the familiar GLM of the SPM approach. (Note that with the conventional approach, it makes little difference if the hypothesis is tested using an encoding model or a linear classifier; the experimental design and the interpretation of a significant result will be the same (Friston, 2009)). Because the stimuli or task conditions have been selected to evaluate a single feature space, encoding models based on different feature spaces are difficult or impossible to construct. Thus, the conventional approach cannot fully exploit the major advantage of encoding models.

To exploit the major advantage of encoding models, the stimuli or task conditions should admit multiple feature spaces. For vision studies, natural scenes are an appropriate stimulus set (Kay et al., 2008; Naselaris et al., 2009). A random selection of natural scenes admits low-level structural features (e.g., Gabor wavelets), high-level semantic features (e.g., scene categories), and any other intermediate features that might be of interest to the experimenter. The only drawback to using natural stimuli or tasks is that the resulting data can be difficult to analyze without sophisticated mathematical techniques (Wu et al., 2006).

6. The upper limit of encoding model prediction accuracy

The prospect of obtaining a complete functional description of a specific ROI is compelling; however, a complete functional description can only be provided by an encoding model that has achieved the upper limit of prediction accuracy. In practice, the upper limit of prediction accuracy will not be the same as *perfect* prediction accuracy. This is because fMRI data are noisy. For the purposes of developing encoding models, noise is simply activity that is not reliably associated with the stimuli or task. Noise in fMRI is caused by physical factors related to MR imaging such as thermal noise, physiological factors such as respiration and to neural sources (Buxton, 2002). Finally, various cognitive factors such as arousal, attention and memory may reduce the reliability of activity if they are not completely controlled in the experiment.

The relationship between features, activity, and noise, can be summarized succinctly by an *encoding distribution*, $p(r \mid f(s))$. Here r denotes the pattern of activity across an array of voxels (i.e., a point in activity space), s denotes the external stimulus or task variable (i.e., a point in input space), and $f(s)$ denotes the features (i.e., a point in feature space). The encoding distribution gives the likelihood of a pattern of activity, *r*, given the features, *f*(*s*). Note that here the pattern of activity *r* refers to the activities of each of the voxels in an ROI. (For generality and convenience the rest of this discussion focuses on patterns of activity across an array of voxels instead of activity in single voxels.) Encoding models that predict patterns of activity across an array of voxels are called *multi-voxel encoding models* (see Naselaris et al, 2009).

To gain an intuition for the encoding distribution, consider the pattern of voxel activity that is evoked by a single image (i.e., a single point in input space). Because experimental noise varies across trials, different presentations of the image will evoke a slightly different pattern of activity on each trial (Figure 3, left). If we represent the activity measured on a single trial as a point in activity space, then many repetitions of the stimulus will produce a cloud of points in this space. The most likely pattern of activity will be given by the point at the center of this cloud. The encoding distribution describes the size and shape of these clouds and the locations of the most likely patterns of activity. Although a description of the size and shape of these clouds is important for training the encoding model (see Wu et al., 2006 for an extensive discussion), the best that an encoding model can do is to predict the *most likely* pattern of activity. Thus, an encoding model that reaches the upper limit of prediction accuracy is one that can perfectly predict the most likely pattern of activity. A linearizing encoding model that reaches this upper limit can be written explicitly in terms of the encoding distribution:

$$H^T f(s) = \underset{r}{\operatorname{argmax}}\, p(r|f(s)).$$

Here, *H* is the set of weights that defines the linear mapping from feature space to activity space. *H* is a matrix, and each column of *H* contains the weights for a single voxel. Because there is one weight per feature the number of rows of *H* is equal to the number of features (or axes in the feature space).

In practice the weights, *H*, of a linearizing multi-voxel encoding model must be estimated from experimental data. The optimal method for inferring the weights of a linear model is determined by the specific form of the encoding distribution. (Wu et al. 2006 discusses this issue in detail for single-neuron encoding models but their conclusions are also applicable to multi-voxel encoding models.) For voxels recorded using fMRI it is generally safe to assume that the specific form of the encoding distribution is Gaussian:

$$p(r|f(s)) \sim \exp\left(-\frac{1}{2}\left(r - H^T f(s)\right)^T \Sigma^{-1} \left(r - H^T f(s)\right)\right).$$

Here, Σ is a noise covariance matrix that describes the size and shape of the cloud in Figure 3 (left). In this case the optimal method for inferring the weights reduces to a least-squared-error minimization procedure (Wu et al., 2006).

Once an optimal set of weights is obtained the multi-voxel encoding model can be used to produce predictions of the most likely pattern of activity evoked by the stimuli in the validation data set. Testing the accuracy of these predictions requires an empirical estimate of the most likely pattern of activity for each stimulus in the validation data set. If noise is Gaussian, an estimate of the most likely pattern of activity for a specific stimulus can be obtained by averaging patterns of activity over repeated presentations.

By testing the prediction accuracies of many models that use different feature spaces, it should be possible to eventually uncover just those features that are represented in the ROI. As long as there were enough training data to obtain an accurate estimate of the weights, a model that incorporated these features would achieve the upper limit of prediction accuracy. This optimal model would constitute a complete functional description of the ROI.

7. Converting an encoding model to a decoding model

Although we have pointed out the advantages of encoding models, decoding models are appealing for several reasons. As discussed earlier, decoding models can be used to directly compare a subject's behavioral performance to the decoding accuracy of an ROI, but encoding models cannot be used for this purpose. Decoding models are also interesting because they can potentially reveal everything that can be learned about a specific feature space by observing brain activity (Rieke et al. 1999). Finally, decoding models provide the foundation for potential brain-reading and neuroprosthetic technologies (deCharms, 2008; Haynes and Rees, 2006).

Fortunately, there is no need to make a choice between encoding and decoding models. Given an encoding distribution, $p(r \mid f(s))$, it is possible to derive a complementary *decoding distribution* that can be used to perform decoding. The key to this derivation is Bayes' theorem:

$$p(f(s)|r) \sim p(r|f(s))\, p(f(s))$$

The distribution on the left hand side is the decoding distribution. In the language of probability theory, it expresses the *posterior probability* that the features, $f(s)$, evoked the measured activity, r. On the right hand side are the encoding distribution (discussed earlier), and a second distribution, $p(f(s))$, called a prior. Bayes' Theorem shows that the decoding distribution is proportional to the product of the encoding distribution and the prior. This fact will be familiar to those with expertise in classification, where the use of Bayes' theorem to derive a classifier from an underlying encoding model is referred to as “generative” classification (Bishop, 2006; Friston et al, 2008).

To gain an intuition for the decoding distribution, consider a hypothetical thought experiment intended to measure all of the various features that evoke one specific pattern of activity from an array of voxels (Figure 3, right). If we represent these data as a cloud of points in feature space then the decoding distribution characterizes the size and shape of the cloud. (Note that the structure of the decoding distribution may bear little resemblance to the structure of the encoding distribution.) The densest region of the cloud corresponds to the features that most often evoke the specific pattern of activity. These are the most probable features, given the specific pattern of activity. Decoding by extracting the most probable features from the decoding distribution is known as *maximum a posteriori* (MAP) decoding. MAP decoding is a powerful and theoretically well-developed technique (Ma et al., 2006) that has also been used in neurophysiological (Zhang et al., 1998) and in voltage-sensitive dye imaging studies (Chen et al., 2006).

The prior reflects the probability that each feature will occur. This distribution is only related to the input space and the feature space; it is completely independent of brain activity. For example, if the input space consists of natural scenes and the feature space consists of oriented edges then the prior will indicate which edges tend to occur most frequently in natural scenes. If the input space consists of natural scenes and the feature space consists of the names of scene categories then the prior will indicate which scene categories tend to occur most frequently. If all features have an equal chance of occurring then the prior distribution is flat, and it has no influence on decoding. However, in many experiments that use complex stimuli some features will tend to occur—or co-occur—much more often than others. In these cases the prior will have a large influence on the quality of the decoded result (Mesgarani et al., 2009; Naselaris et al., 2009). Even so, relatively few decoding studies have incorporated an explicit prior.

In principle Bayes' theorem could also be used to convert a decoding model to an encoding model:

$$p(r|f(s)) \sim p(f(s)|r)\,p(r)$$

However, converting a decoding model to an encoding model would be difficult to do in practice, as it is impractical to determine the form of the decoding distribution empirically. The decoding distribution describes variance in features that evoke the same specific pattern of activity (Figure 3, right). Estimating a decoding distribution to describe this variance would require identifying all the features that evoke one specific pattern of activity in an array of voxels, but noise in voxel activity will make this quite difficult. Thus, another advantage of encoding over decoding is that it is easier to derive a decoding distribution from an encoding distribution than the other way around.

8. The combined encoding / decoding approach

Given our discussion of encoding and decoding, we propose a procedure (Figure 4) for analyzing fMRI data that consists of four steps. (1) Collect data and divide it into training and validation sets. These data sets will be used to estimate and evaluate both encoding and decoding models. In both cases, the data used to train the models should be kept separate from the data used to validate their predictions. (2) Use the training data to estimate one or more encoding models for each voxel. We recommend estimating encoding models first because it is much easier to derive a decoding model from an encoding model than the other way around. (3) Apply the estimated encoding models to the validation data and evaluate prediction accuracy. Prediction accuracy for any single model can be compared across ROIs. Prediction accuracy of multiple models can be compared within a single ROI, and can be used to determine the set of features that provides the most complete functional description of the ROI. (4) Use the encoding models to derive decoding models and apply them to the validation data to decode features. Decoding permits direct comparison to behavior, and may also be used to corroborate any conclusions drawn from the encoding models. Decoding also capitalizes on the increased sensitivity obtained by pooling the activity of many voxels without eliminating information by averaging (Kriegeskorte and Bandettini, 2007). This combined encoding / decoding approach exploits the relative strengths of both encoding and decoding, and requires almost no effort beyond constructing either type of model alone.

9. fMRI studies that use a combined encoding / decoding approach

We are certainly not the first to suggest that encoding and decoding approaches should be combined during analysis. This general idea goes back at least to the neurophysiological studies of Georgopoulos and colleagues on population vector decoding (Georgopoulos et al., 1986). As far as we are aware, Georgopoulos and colleagues were also the first to use combined encoding and decoding to analyze fMRI data. Gourtzelidis et al. (2005) modeled voxel activity in the superior parietal lobule (SPL) evoked by mentally traversing a path through a maze. Their encoding model was based on features that reflected the direction of the traversed path (this was in fact a linearizing encoding model, in the sense that the direction of a traversed path can only be extracted from an image of a maze via a nonlinear mapping). Gourtzelidis et al. used this encoding model to identify a spatially organized distribution of voxels in the SPL that were tuned for path direction (see also Jerde et al., 2008). They then used population vector decoding to reconstruct path direction from voxel activity. Their results provide evidence for an orderly functional organization of the SPL with respect to mental tracing.

Several vision studies have also used encoding models to decode brain activity. Thirion et al. (2006) modeled voxel activity in early visual areas evoked by flashing black-and-white geometric patterns. Their encoding model was based on features that reflected stimulus energy at a variety of spatial locations. Thirion et al. used this encoding model to reveal the location and extent of the spatial receptive field for each voxel. They then used a Bayesian decoding approach described earlier to reconstruct both observed and imagined geometric patterns. Their results provide evidence that mental imagery evokes retinotopically organized activation in early visual areas.

Kay et al. (2008) modeled voxel activity in early visual areas evoked by complex natural scenes. Their encoding model was based on the Gabor wavelet features described earlier. They used their encoding model to perform identification of natural scenes. Their encoding model enabled highly accurate identification performance, even when the natural scene was drawn from a potential set of hundreds of millions. Their results provide evidence that fine-grained visual information is represented in the activity of single voxels.

Naselaris et al. (2009) also modeled voxel activity in visual cortex evoked by natural images. Their two encoding models were based on the Gabor wavelet features and the semantic features described earlier. They developed a generalization of the Bayesian decoding approach that combined the Gabor wavelet and semantic models with a natural image prior to accurately reconstruct natural images. Their results provide evidence that combining activity from functionally distinct areas can produce reconstructions of natural scenes that are both structurally and semantically accurate.

Mitchell et al. (2008) modeled voxel activity across the whole brain evoked by line drawings of everyday objects. Their encoding model was based on word co-occurrence features described earlier. They used their encoding model to perform identification of arbitrary nouns (using an identification approach similar to that in Kay et al., 2008). Their results provide evidence for a relationship between the statistics of word co-occurrence in written language and the representation of the meaning of nouns.

Brouwer and Heeger (2009) modeled voxel activity in retinotopic visual areas evoked by wide-field color patterns. Their encoding model was based on a nonlinear perceptual color space (specifically, the L*a*b color space; Commission Internationale de l'Eclairage, 1986). They used their encoding model to reconstruct novel colors that were not present in the training data. They found that activity in visual area V4 enabled the most accurate reconstruction of novel colors. Their results provide evidence that V4 represents a distinct transition from the color representation in V1 into a perceptual color space.

10. fMRI studies that use linear classifiers

We have emphasized the major advantage of encoding models over decoding models. Nonetheless, linear classifiers (perhaps the simplest kind of decoding model) are one of the most commonly used data analysis techniques in fMRI research. Linear classifiers have been used in virtually every area of research in systems and cognitive neuroscience. A incomplete tally of some current work includes studies of vision (Brouwer and Heeger, 2009; Carlson et al., 2003; Cox and Savoy, 2003; Eger et al., 2008; Haxby et al., 2001; MacEvoy and Epstein, 2009; Peelen et al, 2009; Walther et al., 2009), somatosensation (Beauchamp et al., 2009: Rolls et al., 2009), olfaction (Howard et al., 2009), audition (Ethofer et al., 2009; Formisano et al., 2008b; Raizada et al., 2009), movement (Dehaene et al., 1998; Dinstein et al., 2008), attention (Kamitani and Tong 2005; Kamitani and Tong 2006), consciousness (Haynes and Rees, 2005b; Schurger et al., 2010), memory (Harrison and Tong, 2009; Johnson et al, 2009), intention (Haynes et al., 2007), cognitive control (Esterman et al., 2009), decision making (Hampton and O'Doherty, 2007; Soon et al., 2008) and imagery (Reddy et al., 2010; Stokes, 2009). Here, we discuss several specific studies where the use of linear classifiers has led to important advances.

Haxby et al. (2001) used a linear classifier to decode the categories of various commonplace objects. They found that object category could be decoded from voxel activity in ventral temporal cortex, even when activity from object-specific modules (e.g., the fusiform face area; Kanwisher et al., 1997) were excluded from the analysis. Their results provide evidence that object representation is distributed across cortex rather than entirely localized within object-specific modules.

Kamitani and Tong (2005) used a linear classifier to decode which of two simultaneously presented gratings was attended on any trial. Haynes and Rees (2005a) used a linear classifier to decode the orientation of a grating rendered subjectively invisible by a mask. Subsequent studies by Haynes and colleagues used linear classifiers to decode the intention to add or subtract two numbers (Haynes et al, 2007) and to decode the outcomes of decisions made several seconds before subjects became aware of them (Soon et al., 2008). The success of these remarkable studies demonstrates the usefulness of linear classifiers for investigating covert mental processing.

Stokes et al. (2009) measured voxel activity in higher-level visual areas, and showed that a linear classifier trained to decode perceived images of Xs and Os could successfully decode mental images of Xs and Os. Reddy et al. (2010) also measured activity in higher-level visual areas, and showed that a linear classifier trained to decode the category of perceived images of commonplace objects could successfully decode the category of imagined objects. Along with previous results (O'Craven and Kanwisher, 2000), these studies provide evidence that perception and imagery evoke similar patterns of activity in higher-level visual areas.

11. Conclusions

We introduced this paper by posing several general questions about the differences between encoding and decoding models. Is there any difference in the kinds of sensory or cognitive representations that can be studied with either model? If the goal is to establish that the activity in an ROI represents some amount of information about a specific sensory or cognitive state then the answer is “no”. The only difference between linearizing encoding and decoding models is the direction of the linear mapping between activity and feature space. Thus, any sensory or cognitive representation can be studied using either encoding or decoding models. Are there any advantages to using one type of model instead of the other? Encoding models have one great advantage over decoding models. A perfect encoding model provides a complete functional description of a specific ROI but a perfect decoding model does not. Are there any contexts in which both types of model should be used? We believe that it should be common practice to construct both encoding and decoding models. The encoding model describes how information is represented in the activity of each voxel. Once an encoding model is constructed Bayes’ theorem can be used to derive a decoding model with little effort. The decoding model can then be used to investigate the information represented in patterns of activity across an array of voxels. Application of the decoding model validates the encoding model and provides a sanity check on the conclusions drawn from it. The decoding model also provides a way to link activity directly to behavior, and provides easily accessible intuitions about the role of a specific brain area in the context of our overall experience.

REFERENCES

Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A. 1985; 2(2):284–299. [PubMed: 3973762]

Aertsen A, Johannesma P. The spectro-temporal receptive field. A functional characteristic of auditory neurons. Biol. Cybern. 1981; 42(2):133–143. [PubMed: 7326288]

Beauchamp MS, LaConte S, Yasar N. Distributed representation of single touches in somatosensory and visual cortex. Hum. Brain Mapp. 2009; 30(10):3163–3171. [PubMed: 19224618]

Bishop, CM. Pattern recognition and machine learning. Springer; New York: 2006.

Bredfeldt CE, Ringach DL. Dynamics of spatial frequency tuning in macaque V1. J. Neurosci. 2002; 22:1976–1984. [PubMed: 11880528]

Brouwer GJ, Heeger DJ. Decoding and reconstructing color from responses in human visual cortex. J. Neurosci. 2009; 29(44):13992–14003. [PubMed: 19890009]

Buxton, RB. Introduction to functional magnetic resonance imaging: principles and techniques. Cambridge University Press; Cambridge: 2002.

Brefczynski JA, DeYoe EA. A physiological correlate of the ‘spotlight’ of visual attention. Nat. Neurosci. 1999; 2(4):370–374. [PubMed: 10204545]

Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, Gallant JL, Rust N. Do we know what the early visual system Does? J. Neurosci. 2005; 25(46):10577–10597. [PubMed: 16291931]

Carlson TA, Schrater P, He S. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 2003; 15(5):704–717. [PubMed: 12965044]

Chen Y, Geisler WS, Seidemann E. Optimal decoding of correlated neural population responses in the primate visual cortex. Nat. Neurosci. 2006; 9(11):1412–1420. [PubMed: 17057706]

Commission Internationale de l'Eclairage. Colorimetry. 2. Commission Internationale de l'Eclairage; Vienna: 1986. CIE No. 152

Cox DD, Savoy RL. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage. 2003; 19(2): 261–270. [PubMed: 12814577]

Davatzikos C, Ruparel K, Fan Y, Shen DG, Acharyya M, Loughead JW, Gur RC, Langleben DD. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. NeuroImage. 2005; 28(3):663–668. [PubMed: 16169252]

David SV, Vinje WE, Gallant JL. Natural stimulus statistics alter the receptive field structure of V1 neurons. J. Neurosci. 2004; 24(31):6991–7006. [PubMed: 15295035]

David SV, Gallant JL. Predicting neuronal responses during natural vision. Network. 2005; 16(3):239–260. [PubMed: 16411498]

Dayan, P.; Abbott, L. Theoretical neuroscience: Computational and mathematical modeling of neural systems. MIT Press; Cambridge: 2001.

deCharms R. Applications of real-time fMRI. Nat. Rev. Neurosci. 2008; 9(9):720–729. [PubMed: 18714327]

Dehaene S, Le Clec'H G, Cohen L, Poline J, van de Moortele P, Le Bihan D. Inferring behavior from functional brain images. Nat. Neurosci. 1998; 1(7):549–550. [PubMed: 10196560]

De Martino F, Valente G, Staeren N, Ashburner J, Goebel R, Formisano E. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage. 2008; 43(1):44–58. [PubMed: 18672070]

DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn. Sci. 2007; 11(8):333–341. [PubMed: 17631409]

Dinstein I, Gardner JL, Jazayeri M, Heeger DJ. Executed and observed movements have different distributed representations in human aIPS. J. Neurosci. 2008; 28(44):11231–11239. [PubMed: 18971465]

Downing PE, Chan AW, Peelen MV, Dodds CM, Kanwisher N. Domain specificity in visual cortex. Cereb. Cortex. 2006; 16(10):1453–1461. [PubMed: 16339084]

Dumoulin SO, Wandell BA. Population receptive field estimates in human visual cortex. NeuroImage. 2008; 39(2):647–660. [PubMed: 17977024]

Eger E, Ashburner J, Haynes J, Dolan RJ, Rees G. fMRI Activity patterns in human LOC carry information about object exemplars within category. J. Cogn. Neurosci. 2008; 20(2):356–370. [PubMed: 18275340]

Esterman M, Chiu Y, Tamber-Rosenau BJ, Yantis S. Decoding cognitive control in human parietal cortex. Proc. Natl. Acad. Sci. USA. 2009; 106(42):17974–17979. [PubMed: 19805050]

Ethofer T, Van De Ville D, Scherer K, Vuilleumier P. Decoding of emotional information in voice-sensitive cortices. Curr. Biol. 2009; 19(12):1028–1033. [PubMed: 19446457]

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cereb. Cortex. 1991; 1(1):1–47. [PubMed: 1822724]

Formisano E, De Martino F, Valente G. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Mag. Res. Imag. 2008a; 26(7):921–934.

Formisano E, De Martino F, Bonte M, Goebel R. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science. 2008b; 322(5903):970–973. [PubMed: 18988858]

Friston KJ, Holmes AP, Poline J, Grasby PJ, Williams SCR, Frackowiak RSJ, Turner R. Analysis of fMRI time-series revisited. NeuroImage. 1995; 2(1):45–53. [PubMed: 9343589]

Friston KJ, Chu C, Mourão-Miranda J, Hulme O, Rees G, Penny W, Ashburner J. Bayesian decoding of brain images. NeuroImage. 2008; 39(1):181–205. [PubMed: 17919928]

Friston KJ. Modalities, modes, and models in functional neuroimaging. Science. 2009; 326(5951): 399–403. [PubMed: 19833961]

Ganesh G, Burdet E, Haruno M, Kawato M. Sparse linear regression for reconstructing muscle activity from human cortical fMRI. NeuroImage. 2008; 42(4):1463–1472. [PubMed: 18634889]

Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science. 1986; 233(4771):1416–1419. [PubMed: 3749885]

Gourtzelidis P, Tzagarakis C, Lewis SM, Crowe DA, Auerbach E, Jerde TA, Uğurbil K, Georgopoulos AP. Mental maze solving: directional fMRI tuning and population coding in the superior parietal lobule. Expr. Brain Res. 2005; 165(3):273–282.

Hampton AN, O'Doherty JP. Decoding the neural substrates of reward-related decision making with functional MRI. Proc. Natl. Acad. Sci. USA. 2007; 104(4):1377–1382. [PubMed: 17227855]

Hansen LK. Multivariate strategies in functional magnetic resonance imaging. Brain Lang. 2007; 102(2):186–191. [PubMed: 17223190]

Hanson SJ, Matsuka T, Haxby JV. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? NeuroImage. 2004; 23(1):156–166. [PubMed: 15325362]

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458(7238):632–635. [PubMed: 19225460]

Hastie, T.; Tibshirani, R.; Friedman, JH. The elements of statistical learning: data mining, inference, and prediction. Springer; New York: 2001.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 2001; 293(5539):2425–2430. [PubMed: 11577229]

Haynes J, Rees G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 2005a; 8(5):686–691. [PubMed: 15852013]

Haynes J, Rees G. Predicting the stream of consciousness from activity in human visual cortex. Curr. Biol. 2005b; 15(14):1301–1307. [PubMed: 16051174]

Haynes J, Rees G. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 2006; 7(7):523–534. [PubMed: 16791142]

Haynes J, Sakai K, Rees G, Gilbert S, Frith C, Passingham RE. Reading hidden intentions in the human brain. Curr. Biol. 2007; 17(4):323–328. [PubMed: 17291759]

Haynes J. Decoding visual consciousness from human brain signals. Trends Cogn. Sci. 2009; 13(5): 194–202. [PubMed: 19375378]

Howard JD, Plailly J, Grueschow M, Haynes J, Gottfried JA. Odor quality coding and categorization in human posterior piriform cortex. Nat. Neurosci. 2009; 12(7):932–938. [PubMed: 19483688]

Jerde T, Lewis S, Goerke U, Gourtzelidis P, Tzagarakis C, Lynch J, Moeller S, Van de Moortele P, Adriany G, Trangle J, Uğurbil K, Georgopoulos AP. Ultra-high field parallel imaging of the superior parietal lobule during mental maze solving. Exp. Brain Res. 2008; 187(4):551–561. [PubMed: 18305932]

Johnson JD, McDuff SG, Rugg MD, Norman KA. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron. 2009; 63(5):697–708. [PubMed: 19755111]

Jones JP, Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophys. 1987; 58(6):1233–1258.

Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 2005; 8(5):679–685. [PubMed: 15852014]

Kamitani Y, Tong F. Decoding seen and attended motion directions from activity in the human visual cortex. Curr. Biol. 2006; 16(11):1096–1102. [PubMed: 16753563]

Kowalski N, Depireux DA, Shamma SA. Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J. Neurophysiol. 1996; 76(5):3503–3523. [PubMed: 8930289]

Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 1997; 17(11):4302–4311. [PubMed: 9151747]

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452(7185):352–5. [PubMed: 18322462]

Kay KN, Gallant JL. I can see what you see. Nat. Neurosci. 2009; 12(3):245. [PubMed: 19238184]

Kippenhan JS, Barker WW, Pascal S, Nagel J, Duara R. Evaluation of a neural-network classifier for PET scans of normal and Alzheimer's disease subjects. J. Nucl. Med. 1992; 33(8):1459–1467. [PubMed: 1634935]

Kriegeskorte N, Goebel R, Bandettini P. Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA. 2006; 103(10):3863–3868. [PubMed: 16537458]

Kriegeskorte N, Bandettini P. Analyzing for information, not activation, to exploit high-resolution fMRI. NeuroImage. 2007; 38(4):649–662. [PubMed: 17804260]

Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2008a; 2(4):1–28.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini P. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 2008b; 60(6):1126–1141. [PubMed: 19109916]

Lautrup B, Hansen LK, Law I, Morch N, Svarer C, Strother SC. Massive weight sharing: a cure for extremely ill-posed problems. Workshop on supercomputing in brain research: From tomography to neural networks. 1994:137–144.

Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nat. Neurosci. 2006; 9(11):1432–1438. [PubMed: 17057707]

MacEvoy SP, Epstein RA. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. Curr. Biol. 2009; 19(11):943–947. [PubMed: 19446454]

Machens CK, Wehr MS, Zador AM. Linearity of cortical receptive fields measured with natural sounds. J. Neurosci. 2004; 24(5):1089–1100. [PubMed: 14762127]

Mazer JA, Vinje WE, McDermott J, Schiller PH, Gallant JL. Spatial frequency and orientation tuning dynamics in area V1. Proc. Natl. Acad. Sci. USA. 2002; 99(3):1645–1650. [PubMed: 11818532]

Mesgarani N, David SV, Fritz JB, Shamma SA. Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. J. Neurophys. 2009; 102(6):91128–2008.

Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage. 2010 doi:10.1016/j.neuroimage.2010.05.051.

Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S. Learning to decode cognitive states from brain images. Machine Learning. 2004; 57(1):145–175.

Mitchell TM, Shinkareva SV, Carlson A, Chang K, Malave VL, Mason RA, Just MA. Predicting human brain activity associated with the meanings of nouns. Science. 2008; 320(5880):1191–1195. [PubMed: 18511683]

Miyawaki Y, Uchida H, Yamashita O, Sato M, Morito Y, Tanabe HC, Sadato N, Kamitani Y. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron. 2008; 60(5):915–929. [PubMed: 19081384]

Mur M, Bandettini PA, Kriegeskorte N. Revealing representational content with pattern-information fMRI–an introductory guide. Soc. Cogn. Affect. Neurosci. 2009; 4(1):101–109. [PubMed: 19151374]

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009; 63(6):902–915. [PubMed: 19778517]

Nishimoto S, Ishida T, Ohzawa I. Receptive field properties of neurons in the early visual cortex revealed by local spectral reverse correlation. J. Neurosci. 2006; 26(12):3269–3280. [PubMed: 16554477]

Norman KA, Polyn SM, Detre GJ, Haxby JV. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 2006; 10(9):424–430. [PubMed: 16899397]

Nykamp DQ, Ringach DL. Full identification of a linear-nonlinear system via crosscorrelation analysis. J. Vis. 2002; 2(1):1–11. [PubMed: 12678593]

O'Craven KM, Kanwisher N. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. J. Cogn. Neurosci. 2000; 12(6):1013–1023. [PubMed: 11177421]

O'Toole AJ, Jiang F, Abdi H, Pénard N, Dunlop JP, Parent MA. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 2007; 19(11):1735–1752. [PubMed: 17958478]

Peelen MV, Fei-Fei L, Kastner S. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature. 2009; 460(7251):94–97. [PubMed: 19506558]

Pereira F, Mitchell T, Botvinick M. Machine learning classifiers and fMRI: A tutorial overview. NeuroImage. 2009; 45(1):199–209.

Raizada RDS, Tsao F, Liu H, Kuhl PK. Quantifying the adequacy of neural representations for a cross-language phonetic discrimination task: prediction of individual differences. Cereb. Cortex. 2009; 20(1):1–12. [PubMed: 19386636]

Reddy L, Tsuchiya N, Serre T. Reading the mind's eye: Decoding category information during mental imagery. NeuroImage. 2010; 50(2):818–825. [PubMed: 20004247]

Rieke, F.; Warland, D.; Bialek, W. Spikes: exploring the neural code. The MIT Press; Cambridge: 1999.

Ringach DL, Hawken MJ, Shapley R. Dynamics of orientation tuning in macaque primary visual cortex. Nature. 1997; 387(6330):281–84. [PubMed: 9153392]

Rolls ET, Grabenhorst F, Franco L. Prediction of subjective affective state from brain activations. J. Neurophysiol. 2009; 101(3):1294–1308. [PubMed: 19109452]

Schönwiesner M, Zatorre RJ. Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. Proc. Natl. Acad. Sci. USA. 2009; 106(34):14611–14616. [PubMed: 19667199]

Schurger A, Pereira F, Treisman A, Cohen JD. Reproducibility distinguishes conscious from nonconscious neural representations. Science. 2010; 327(5961):97–99. [PubMed: 19965385]

Soon CS, Brass M, Heinze H, Haynes J. Unconscious determinants of free decisions in the human brain. Nat. Neurosci. 2008; 11(5):543–545. [PubMed: 18408715]

Stokes M, Thompson R, Cusack R, Duncan J. Top-down activation of shape-specific population codes in visual cortex during mental imagery. J. Neurosci. 2009; 29(5):1565–1572. [PubMed: 19193903]

Theunissen FE, Sen K, Doupe AJ. Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. J. Neurosci. 2000; 20(6):2315–2331. [PubMed: 10704507]

Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, Lebihan D, Dehaene S. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. Neuroimage. 2006; 33(4): 1104–1116. [PubMed: 17029988]

Walther DB, Caddigan E, Fei-Fei L, Beck DM. Natural scene categories revealed in distributed patterns of activity in the human brain. J. Neurosci. 2009; 29(34):10573–10581. [PubMed: 19710310]

Williams MA, Dang S, Kanwisher NG. Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. 2007; 10(6):685–686. [PubMed: 17486103]

Willmore BDB, Prenger RJ, Gallant JL. Neural representation of natural images in visual area V2. J. Neurosci. 2010; 30(6):2102–2114. [PubMed: 20147538]

Wu MC, David SV, Gallant JL. Complete functional characterization of sensory neurons by system identification. Ann. Rev. Neurosci. 2006; 29:477–505. [PubMed: 16776594]

Zhang K, Ginzburg I, McNaughton BL, Sejnowski TJ. Interpreting neuronal population activity by reconstruction: unified framework with application to hippocampal place cells. J. Neurophysiol. 1998; 79(2):1017–1044. [PubMed: 9463459]

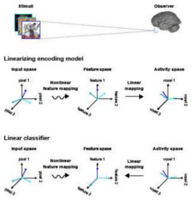

Figure 1. Linearizing encoding and decoding models

[Top] The brain can be viewed as a system that nonlinearly maps stimuli into brain activity. According to this perspective a central task of systems and cognitive neuroscience is to discover the nonlinear mapping between input and activity. [Middle] Linearizing encoding model. The relationship between encoding and decoding can be described in terms of a series of abstract spaces. In experiments using visual stimuli the axes of the input space are the luminance of pixels and each point in the space (here different colors in the input space) represents a different image. Brain activity measured in each voxel is represented by an activity space. The axes of the activity space correspond to the activity of different voxels and each point in the space represents a unique pattern of activity across voxels (different colors in the activity space). In between the input and activity spaces is a feature space. The mapping between the input space and the feature space is nonlinear and the mapping between the feature space and activity space is linear. [Bottom] Linear classifier. The linear classifier is a simple decoding model that can also be described in terms of input, feature and activity spaces. However, the direction of the mapping between activity and feature space is reversed relative to the encoding model. Because the features are discrete all points in the feature space lie along the axes

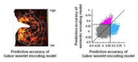

Figure 2. Comparative analyses using encoding models

[Left] Prediction accuracy for a Gabor wavelet encoding model. The stimuli used in the experiment were grayscale natural scenes. A Gabor wavelet encoding model was estimated for each voxel. Here the prediction accuracy for each voxel has been projected onto a digitally flattened map of visual cortex. Known visual areas are outlined in white. Prediction accuracy for the Gabor wavelet model is highest in early visual areas such as primary visual cortex, and declines in higher areas. Maps such as this one can be used to compare the representation of a specific set of features (such as Gabor wavelets) across many regions of interest. [Right] Comparison of prediction accuracy for two different encoding models. The horizontal axis gives prediction accuracy for the Gabor wavelet encoding model shown at left. The vertical axis gives prediction accuracy for a semantic encoding model. Each dot indicates an individual voxel. Cyan dots indicate voxels that are better modeled by the Gabor wavelet encoding model while magenta dots indicate voxels that are better modeled by the semantic encoding model. Gray dots indicate voxels that are not modeled well by either of these two encoding models. The two models provide good predictions for different populations of voxels. As this figure shows, it is easy to compare the accuracy of predictions for distinct encoding models even when the models are based upon very different features. In contrast, it is difficult to compare predictions of decoding models that decode different features.

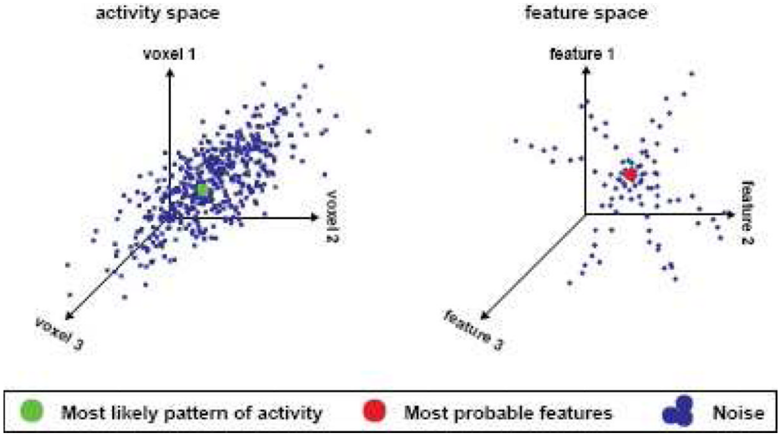

Figure 3. Encoding and decoding distributions

[Left] The encoding distribution describes variance in the patterns of activity (blue dots) that are evoked by repeated presentations of the same stimulus. It also describes the location of the most likely pattern of activity (green dot) given the stimulus. A perfect encoding model would be able to predict the most likely pattern of activity evoked by any arbitrary stimulus. [Right] The decoding distribution describes variance in the features (blue dots) that evoke the same pattern of activity. It also describes the most probable features (red dot) given the pattern of activity. Maximum *a posteriori* decoding attempts to predict the most probable feature, given any arbitrary pattern of activity.

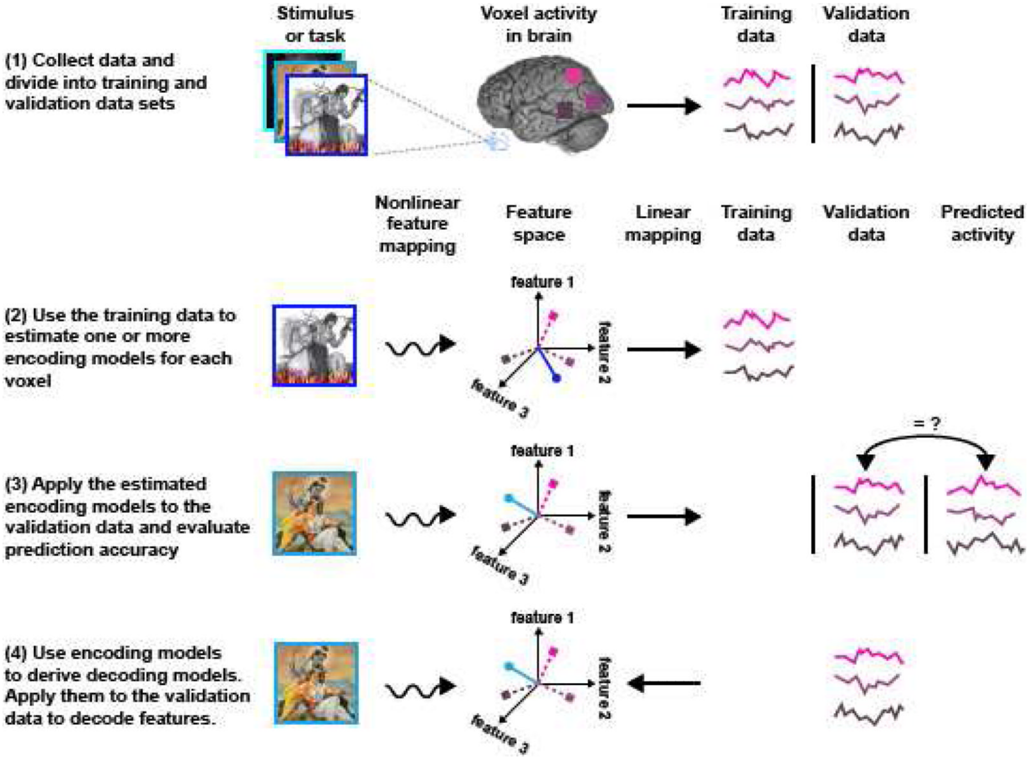
Stimulus or task
Voxel activity in brain
Training data
Validation data
(1) Collect data and divide into training and validation data sets
Nonlinear feature mapping
Feature space
Linear mapping
Training data
Validation data
Predicted activity
(2) Use the training data to estimate one or more encoding models for each voxel
feature 1
feature 2
feature 3
= ?
(3) Apply the estimated encoding models to the validation data and evaluate prediction accuracy
(4) Use encoding models to derive decoding models. Apply them to the validation data to decode features.

Figure 4. The combined encoding / decoding approach

The relationship between encoding and decoding models suggests an ideal procedure for analyzing fMRI data that consists of four steps (one step for each row in the figure). [Row 1] Voxel activity (jagged lines) evoked by experimental stimuli (scenes at left) is divided into a training data set and a validation data set. [Row 2] Encoding models are specified by a nonlinear mapping (curvy arrow) of the stimuli into an abstract feature space (labeled axes represent hypothetical feature space; stimuli depicted by line with circular end). Model weights (dashed lines with square ends) estimated from training data specify a linear mapping (straight arrows) from feature space to voxel activity. [Row 3] Prediction accuracy is measured by comparing the activity in the validation data set to the predicted activity (far right). [Row 4] Decoding models are derived by using Bayes' theorem to reverse the direction of the linear mapping (straight arrow).