Inferential Pitfalls in Decoding Neural Representations

Vencislav Popov (vencislav.popov@gmail.com)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Markus Ostarek (markus.ostarek@mpi.nl)
Max Planck Institute for Psycholinguistics, Nijmegen, Netherlands

Caitlin Tenison (ctenison@andrew.cmu.edu)
Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

A key challenge for cognitive neuroscience is to decipher the representational schemes of the brain. A recent class of decoding algorithms for fMRI data, stimulus-feature-based encoding models, is becoming increasingly popular for inferring the dimensions of neural representational spaces from stimulus-feature spaces. We argue that such inferences are not always valid, because decoding can occur even if the neural representational space and the stimulus-feature space use different representational schemes. This can happen when there is a systematic mapping between them. In a simulation, we successfully decoded the binary representation of numbers from their decimal features. Since binary and decimal number systems use different representations, we cannot conclude that the binary representation encodes decimal features. The same argument applies to the decoding of neural patterns from stimulus-feature spaces and we urge caution in inferring the nature of the neural code from such methods. We discuss ways to overcome these inferential limitations.

Introduction

A key challenge for cognitive neuroscience is to decipher the representational schemes of the brain, to understand the neural code that underlies the encoding and representation of sensory, motor, spatial, emotional, semantic and other types of information. To address these issues researchers often employ neuroimaging techniques like functional magnetic resonance imaging (fMRI), which measures the blood oxygenation level-dependent (BOLD) activation in the brain that is elicited when participants engage with different stimuli. A common assumption has been that the underlying neural representation of each stimulus has measurable but complex effects on the BOLD activation patterns. In order to understand what those patterns of activity can tell us about how the brain processes and represents information, researchers have used various analytical tools such as univariate subtraction methods, multivariate pattern (MVP) classification, representational similarity analysis (RSA) and, recently, explicit stimulus-feature-based encoding and decoding models (for reviews, see Davis & Poldrack, 2013, Haxby, Connolly, & Guntupalli, 2014, or Naselaris, Kay, Nishimoto, & Gallant, 2011). Despite their differences, these methods aim to quantify how changes in task conditions and the properties of the stimuli relate to changes in BOLD activation and vice versa. Where these methods differ is in how they achieve that mapping and in what inferences they allow us to draw.

In this article, we review some of the known inferential limitations of existing fMRI analysis methods and we highlight a previously unrecognized issue in interpreting results from stimulus-feature-based encoding and decoding models. The latter are steadily becoming the de facto gold standard for investigating neural representational spaces (Haxby et al. 2014, Naselaris & Kay, 2015).

Univariate vs. multivariate analysis

Before the advent of the more advanced techniques we review below, the main fMRI analysis tool was based on comparing how activity in a single voxel or averaged activity in a contiguous area of voxels differs between task conditions or stimuli. These univariate subtraction methods have been informative about the relative engagement of certain brain areas in specific tasks. Unfortunately, the coarse nature of this method precludes fine-grained inferences about the underlying representational content and computations that give rise to the observed BOLD signal. By ignoring the possibility that information might be represented in a distributed manner across voxels, the assumptions underlying univariate subtraction methods limit their use in understanding neural representations. In addition, these methods cannot tell us whether changes in activation are due to representational preferences, processing differences, or attentional variation among conditions (Coutanche, 2013).

In contrast, multivoxel pattern analysis (MVPA) techniques have attempted to overcome this limitation by looking at how various categories of stimuli or task conditions lead to differences (i.e. MVP classification) or similarities (i.e. RSA) in distributed patterns of activity over multiple voxels. These methods have become popular because they allow researchers to study neural representational spaces with increasing sensitivity and resolution. For example, a seminal study by Haxby et al. (2001) found that visual object categories can be classified based on the pattern of activation that their exemplars elicited in the ventral temporal cortex. The classification was successful despite the lack of overall activation differences in that region. Similar methods have been used to show that concepts have language-invariant representations in the anterior temporal lobe (Correia et al., 2014), that very similar visual scenes can be discriminated in the hippocampus (Bonnici et al., 2012) and that during their retrieval from memory, the shape, color and identity of visual objects can be differentially decoded across several cortical areas (Coutanche & Thompson-Schill, 2015).

Despite early enthusiasm that MVPA methods could be used to understand the structure of the neural code and the nature of the underlying representations (Norman, Polyn, Detre, & Haxby, 2006), conventional MVP classification and RSA techniques share one of the same fundamental inferential limitations of univariate methods. Successful classification or careful inspection of confusions/similarity matrices can indicate that some relevant information about the stimulus class is present in the population of analyzed voxels, but it cannot identify exactly what that information is, or how it is represented and organized (Naselaris & Kay, 2015; Poldrack, 2011; Tong & Pratte, 2012). Because neural data is correlational, many different properties of the stimuli might lead to successful classification of the stimulus category, the task condition, or the brain state in question. For example, successfully categorizing whether a word represents an animate or an inanimate object does not necessarily mean that the region of interest encodes that category distinction. There are many differences between animate and inanimate objects, such as differences in their sensory and functional features (Farah & McClelland, 1991) that could be responsible for the successful classification.

Another limitation of conventional MVP classifiers is that they cannot generalize and predict behavioral responses to novel *types* of stimuli or task conditions. To understand why, we can conceptualize classifiers in terms of types and tokens. An MVP classifier is usually trained on stimuli that are tokens from several types. For example, the stimuli tokens might be different category exemplars, and the classifier is trained to predict the type of category to which they belong. Alternatively, the tokens might be multiple presentations of the same word in different modalities or languages and the types are the unique words themselves. In the first case, the classifier can only be used to predict category membership of words that belong to one of the categories on which it was trained. In the second case even though the classifier could be used to predict exemplars in novel languages or modalities, it is again restricted only to exemplars of the words on which it was trained in the first place. In general, while the tokens being tested might be novel, they will be potentially decoded only if they are exemplars of a type that has already been trained on.

For example, if one trains a classifier to predict the color of objects and trains it on yellow and orange objects (Coutanche & Thompson-Schill, 2015), one will not be able to predict the color of novel objects that are green. This methodological limitation is important - just as understanding how the decimal system represents numbers allows people to understand and manipulate numbers they have never seen before, a complete understanding of any neural representational system should allow researchers to use the neural pattern associated with novel stimuli to predict their identity, even if those stimuli are not exemplars of the types on which a particular model was trained on.

Stimulus-feature-based encoding models

To overcome this limitation many researchers are turning to a novel analysis method that is known by a few different names – voxelwise modelling (Naselaris & Kay, 2015), stimulus-model based encoding and decoding (Haxby et al., 2014), voxel-based encoding and decoding models (Naselaris et al., 2011), and forward models (Brouwer & Heeger, 2009; Fernandino, Humphries, Conant, Seidenberg, & Binder, 2016). This approach can decode the identity of novel *types* of stimuli from neural activity by predicting activity not for the stimuli themselves, but for a set of simpler features into which they can be decomposed. In a seminal study, Mitchell et al. (2008) predicted the neural activity associated with individual novel words based only on the activation of other words. To achieve that, they decomposed each word into a vector of weights on 25 sensory-motor semantic features (verbs such as “eat”, “taste”, “run”, “fear”, etc.). The weights were estimated from co-occurrence statistics of the word with each verb feature in a large corpus. They trained a classifier to predict the neural activity associated with *each constituent feature* of a training set of words, which resulted in separate neural activation maps for each feature. Neural activity for novel test words was then predicted highly accurately as a linear combination of the semantic feature activation maps weighted by the association of the word with each feature. Based on these results, Mitchell et al. (2008) concluded that the neural representation of concrete nouns might be based on sensory-motor features.

Similar approaches have been used to predict the neural response to novel natural images using Gabor filter features (Kay, Naselaris, Prenger, & Gallant, 2008), to novel colors based on color tuning curve features (Brouwer & Heeger, 2009), to novel music clips based on acoustic timbre features (Casey, Thompson, Kang, Raizada, & Wheatley, 2012), to natural sounds based on frequency, spectral and temporal modulations (Santoro et al., 2014), to novel faces based on a PCA decomposition of face features (Lee & Kuhl, 2016), to novel words based on subjective sensory-motor ratings (Fernandino et al., 2016). The motivating question behind the majority of these studies has been about the nature of the representations used by the brain in encoding the experimental stimuli, and the results have been used to argue that the neural representation is based on the constituent features of the stimuli used in the model.

To summarize, stimulus-feature encoding models generally use the following analysis procedure: 1) Specify a set of features and dimensions that hypothetically underlie the representation of a stimulus set in brain. 2) Decompose a set of stimuli into vectors of weights for each feature. 3) Select a region of interest (ROI) in the brain from which to analyze neural activation. 4) Train a model to predict activity in each voxel for a training set of stimuli, using the weights of their features as predictors. 5) Derive activation pattern maps (e.g. regression coefficients) associated with each feature. 6) Predict neural activity in the ROI for novel stimuli, based on their feature weights and the activation pattern maps for each feature. 7) Compare predicted neural activity for each novel stimulus with their observed neural activity and derive a measure of fit and accuracy. In essence, stimulus-feature-based encoding models attempt to map a stimulus feature representational space, where each feature is a separate dimension, and each stimulus is a point in that space, to a neural activation space, where each voxel is a separate dimension, and the activation pattern elicited by each stimulus is a point in that space.

What can we infer about neural representations?

What can a successful mapping between a stimulus feature space and a neural activation space tell us about the nature of the representation used by the brain? A common inference in these studies has been that if you can predict the identity of novel stimuli based on that mapping, then the neural representation is likely based on the feature set used by the model. Put formally, the inferential claim goes as follows:

1) We can represent certain stimuli as a combination of lower-level features
2) We can show that it is possible to predict the neural pattern caused by a novel stimulus in brain area A from an encoding model based on these features
3) *Therefore, brain area A encodes those features and uses a representational scheme based on them.*

This claim has been made to different degrees both in theoretical and methodological papers reviewing the approach (e.g., Haxby et al., 2014; Naselaris & Kay, 2015; Naselaris et al., 2011; Norman et al., 2006; Tong & Pratte, 2012), as well as in empirical studies that use it to address representational questions (Fernandino et al., 2016; Kay et al., 2008; Mitchell et al., 2008; Santoro et al., 2014; although some are more cautionary, e.g. Lee & Kuhl, 2016). If this inference is valid, then encoding models could be an extremely powerful tool for understanding the nature of neural representations.

A useful illustrative example of this inference in practice comes from a recent study by Fernandino et al. (2016). The authors wanted to understand how conceptual information is represented in a set of higher-order non-modality-specific brain regions in General Semantic Network (Binder, Desai, Graves, & Conant, 2009). An encoding model based on subjective ratings for 5 sensory-motor features of training words (“color”, “motion”, “sound”, “shape”, “action”) was used to predict activation patterns related to novel individual words. The model successfully predicted above chance the brain activity patterns for concrete words in the semantic network regions (61% mean accuracy), but not in a set of control regions associated with visual word form processing. Based on this finding, Fernandino et al. (2016) suggested that “the brain represents concepts as multimodal combinations of sensory and motor representations” and that “heteromodal areas involved in semantic processing encode information about the relative importance of different sensory-motor attributes of concepts, possibly by storing particular combinations of sensory and motor features”.

Here lies the problem – this inference is not formally valid. We need to consider what the data would have looked like if the underlying neural representation was actually different (Mahon, 2015). In this example, the successful decoding of conceptual identity in the GSN based on an encoding model of sensory-motor features does not necessitate the representational format in the GSN to be sensory-motor in nature. The results might be obtained even if the GSN uses amodal representations, as long as there is a non-arbitrary mapping between representations in the GSN and sensory-motor features. To illustrate, let us hypothetically assume that the GSN literally encodes word co-occurrence statistics. As co-occurrence statistics correlate with sensory-motor feature ratings, it would be possible to predict GSN activity patterns based on these features, even if they are not driving the activity patterns. In contrast, successful decoding would be impossible if the mapping between the GSN representations and sensory-motor features was arbitrary. Thus, Fernandino et al.'s (2016) results constitute evidence against the possibility that conceptual representations in heteromodal areas bear an arbitrary relation to sensory-motor features, as has been argued by some proponents of symbolic systems (Fodor & Pylyshyn, 1988), but should not be taken as conclusive evidence that the GSN encodes multimodal sensory-motor information.

This issue is not limited to the specific study discussed above. To put the claim more generally, we argue that information in one representational system might be decoded based on features from another, even if they use different representational schemes, *as long as there is at least a partially systematic mapping between them*. Specifically, while such encoding models should be able to predict the neural activation from the features of a stimulus if the brain uses a representational scheme based on those features, the reverse is not guaranteed[1]. A successful prediction can also occur when the stimulus feature space is *systematically related* to the features that underlie the neural representational scheme. However, that relationship need not be one of equivalence. There are at least three ways in which mappings between representational systems can be made and successful prediction can occur in two of those cases.

Types of mappings

Arbitrary mappings between representations. First, items from two representational systems might be related in an entirely arbitrary way. For example, the meaning of words is mostly unrelated to their orthographic features[2], and the geographic locations of countries are not predictive of their names, etc. More generally, consider two unordered sets of items, $A = \{A_1, A_2, \dots, A_n\}$ and $B = \{B_1, B_2, \dots, B_n\}$. An arbitrary mapping between these two sets exists when the mapping from a specific item in set A to a corresponding item in set B is unrelated to the mappings between the remaining items in the sets. In the context of encoding models and the brain, decoding of novel items from one set would be impossible based on a feature model from the other set, if these two sets are arbitrarily related.

[1] this problem is similar, but not identical, to the problem of reverse inference (Poldrack, 2006)

[2] Whereas a minor degree of systematicity does seem to exist in this domain (e.g., Monaghan et al., 2014), word meanings cannot be systematically predicted based on their orthography and vice versa.

Table 1 Examples of studies that use feature encoding models

Source	Item	Features	Response vector
Mitchell et al., (2008)	Concrete words *(dog)*	Co-occurrence statistics with 25 sensory-motor verbs	Pattern of activation in all cortical voxels
Fernandino et al., (2016)	Concrete words *(dog)*	5 sensory-motor relevance ratings	Pattern of activation in the GSN (Binder et al., 2009)
Current simulation	Numbers *(3497)*	5 decimal digits [0 3 4 9 7]	17 binary digits [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 1 0 1 0 0 1]

Sets that use the same representational format. In contrast, a successful prediction can occur if the two sets use the same representational format. Consider the set of multi-digit numbers in the decimal system, $A = \{10, 11, \dots, 427, \dots\}$, and the set of 10 digits in the decimal system, $B = \{0,1,2,3,4,5,6,7,8,9,10\}$. These sets use the same representational format to represent quantities (the decimal system), and there is a systematic linear mapping from the features (the digits), to the multi-digit numbers, such that:

$$\overline{d_n d_{n-1} \dots d_1 d_0} = \sum_{i=0}^{n} (d_i \times 10^i)$$

$$3491 = 3 \times 1000 + 4 \times 100 + 9 \times 10 + 1 \times 1$$

When we have such systematic mappings between systems that use the same representational format, knowing the mapping function allows us to decompose any item from set A as a combination of features from set B. An example of such a mapping would be Fernandino et al.'s (2016) suggestion that the general semantic network encodes multimodal combinations of sensory-motor features by integrating information from modality-specific sensory-motor areas. If this were true, then you could predict the neural pattern of novel items from their featural representations, which is what that study found as well.

Sets that use different but systematically related representational formats. However, there is an alternative, which would also allow you to make a successful prediction from encoding models. Two sets can use different representational schemes, while maintaining a systematic mapping between themselves that allows us to predict the mapping of any one pair of items from knowledge of the mapping function. Within the context of conceptual representations in the brain, higher-level heteromodal areas might use a representational code that is different from the one used by sensory-motor cortices, but there might be a systematic mapping between representations in each system[3].

For a simplified example, consider the relation between the decimal and the binary systems for representing numeric values. A binary represented value can be transformed into a decimal number by applying the following formula:

$$\left(\overline{d_n d_{n-1} \dots d_0}\right)_2 \rightarrow \left(\sum_{i=0}^{n} (d_i \times 2^i)\right)_{10}$$

$$10011_2 \rightarrow 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0$$
$$= 16_{10} + 2_{10} + 1_{10} = 19_{10}$$

Clearly, there is a systematic but non-linear mapping between the decimal and the binary system, and yet, these two systems use different codes to represent numbers. If our argument is correct then it should be possible to predict the binary representation of a number based on a decimal feature encoding model. Below we present a simulation that achieves this by applying the encoding model approach often used in neuroimaging studies. Within the simulation, binary vectors are analogous to voxel activation patterns, and the encoding model is based on decimal representations (Table 1).

Simulation: Decoding binary representations with a decimal feature encoding model

As detailed previously, encoding models predict stimulus identity from brain activation by modelling the relationship between the constituent features of the training stimuli and their corresponding BOLD activation in a group of voxels. Then they use that relationship to estimate the expected neural activation patterns for novel test items based on their feature representations. The predicted activation pattern for each stimulus is compared to the observed patterns for all test stimuli. For the following simulation, let us consider the numbers from 0 to 99 999 as our stimulus set. They can be decomposed into 5-dimensional feature vectors where each feature is a decimal digit (e.g., 3497 can be decomposed as [0 3 4 9 7]. These features can be considered analogous to the 5 sensory-motor relevance ratings of words used by Fernandino et al. (2016) or to the co-occurrence statistics with sensory-motor verbs used by Mitchell et al. (2008). Further, let us consider the binary representation numbers as 17-dimensional vectors (e.g. [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 1 0 1 0 0 1], to be analogous to the BOLD activation pattern in a set of 17 voxels in an ROI under investigation. The correspondence between these patterns and actual neuroimaging studies using this approach is demonstrated in Table 1.

We trained an encoding model to predict the binary activation pattern for a given number, based on its 5-dimensional decimal

[3] What makes representational codes different is a surprisingly difficult question to answer. Due to space limitations we will briefly cover this issue in the general discussion, but a more in-depth treatment is needed

feature representation. The modelling followed 4 steps: 1) splitting the stimuli into a training (90%) and a test (10%) set, 2) fitting multiple linear regression models on the training set with the 17 binary features as response variables, and the 5 decimal features as predictors, 3) calculating predicted activation pattern (predicted maps, PMs) for each test item from its decimal features and the multivariate regression model, 4) comparing the PMs with the actual binary patterns for all test items (observed maps, OMs). In the comparison stage, we computed the Euclidean distance between each PM and the OMs for all test items, and we calculated the percentile rank of the similarity between the PM and the OM of each item. For example, if the PM for the number 29782 were most similar to OM for that number, then the percentile rank for it would be 10 000/10 000 = 1. However, if it were more similar to the OMs of 1 000 other items, then its percentile rank would be 9 000/10 000 = 0.9.

The encoding model was successful in decoding the binary representation of untrained items based only on their decimal features. The prediction accuracy of the linear regression model was 0.7 (SD = 0.24) and a wilcoxon signed rank test showed that it was above chance ($p < .0001$). Since by definition binary and decimal number systems use different representational formats, we cannot conclude that the representation of binary numbers encodes decimal features. By analogy, the successful decoding of patterns of neural activation based on a stimulus feature space, cannot be used to infer that the brain encodes information about these features or that its neural representational space is organized along the dimensions of that feature space.

Discussion

Stimulus-feature based encoding models (Haxby et al., 2014, Naselaris et al., 2011) are a powerful new tool for studying how the constituent features of stimuli relate to the neural activation patterns elicited by these stimuli. They represent a significant methodological advance over more traditional MVPA methods because they allow us to predict neural activation for novel items and because they can be used to decode the identity of such items from neural data alone. While this is an impressive feat and an incredibly useful tool, we have to be cautious in interpreting what such successes mean for our understanding of the representational system of the brain. Both theorists (e.g., Haxby et al., 2014; Naselaris & Kay, 2015; Naselaris et al., 2011; Norman et al., 2006; Tong & Pratte, 2012) and practitioners (e.g. Fernandino et al., 2016; Kay et al., 2008; Mitchell et al., 2008; Santoro et al., 2014) have suggested that we can infer that the brain uses a certain set of features to encode information, if we can successfully decode the activity of novel items from such features. However, as we have argued here, this inference is not formally valid. Successful decoding might be the result of a systematic relationship between the representational system of the brain and the stimulus feature set, even if those utilize different representational schemes.

How do we know whether two representational systems are truly different? It could be argued that in our example, both binary and decimal number systems share many properties, and that they are merely different implementations of the same fundamental representation. For example, both systems use the position of a digit to encode its magnitude, and as a result, all arithmetic procedures that can be performed with decimal numbers can be applied to binary numbers as well. We propose that the key issue in determining whether two representations are the same is whether you can establish a one-to-one mapping relation between features at different levels of representation in each system. For example, if you substitute each decimal digit with a unique letter, the resulting system would appear different from the decimal system only on the surface, but the relation between multi-digit numbers and their features would be the same in both cases[4] In contrast, decimal and binary features have a qualitatively different relation to the numbers they represent. Despite this, binary representations can be decoded based on decimal features, illustrating the inferential problem of encoding models we address here.

It is important to clarify that the "one-to-one" mapping is an abstract requirement. We are not claiming that to establish representational equivalence between the brain and a certain set of features that it is necessary to find a one-to-one mapping between the basic feature components of stimuli and activation in individual voxels or groups of voxels. The brain does not compute and represent information at the voxel level – voxel activations are the result of averaged activity over hundreds of thousands of neurons. The general lack of access to large-scale neural level activity in the living human brain makes it even more important to not only discover analytical tools that helps us relate voxel activation to possible representations, but also to understand the limitations of those tools and what they can and cannot tell us.

An important question that naturally arises from the caveats we discussed is how one can maximize confidence in the outcome of a forward encoding model approach, or conversely, guard oneself against unjustified inferences. We propose that it is crucial to compare the performance of several possible encoding models. To that aim, it is not sufficient to use a "baseline model" that is unrelated to the domain of interest (i.e., compare a semantic feature model to a low-level visual word form model). Instead, one or several alternative representational models should be tested that are derived from competing theories (i.e., semantic model A vs. semantic model B). To illustrate, an elegant comparison of a sensory-based vs. non-sensory-based semantic model was achieved by Anderson et al. (2015). These authors contrasted a visual model with a word co-occurrence model to investigate which brain regions represent modality-specific visual features, and which do not (using differential correlation in RSA rather than an encoding model). The relative superiority of a particular model at predicting activation patterns in a brain region makes it more likely that the brain is using the representational scheme of the better performing model rather than the alternative. However, it is important to keep in mind that such comparisons only provide

[4] in fact, because of that linear one-to-one relationship, replicating our simulation with these two examples leads to perfect decoding accuracy; compare that to the 0.7 decoding accuracy for the decimal-to-binary model

evidence for the relative likelihood of each model, but, due to the limitations discussed above, still do not allow us to infer that the winning model is the "true" model.

For that reason, besides the assessment of relative model performance based on model comparison, a second crucial step is to evaluate absolute prediction performance. In particular, the observed decoding accuracy can be compared to the "noise ceiling", or to the "upper limit of prediction accuracy" (Naselaris et al., 2011), reflecting the maximal performance that can be feasibly achieved given the noise present in the signal. The gap between the two can be thought of as the variance that is not explained by the current model, which should motivate and guide the search for an improved or alternative version of the model. Until such maximal performance is obtained, we should be careful in making strong representational inferences about the brain from the currently available analytic methods.

Ultimately, many of these inferential caveats exist because fMRI data is correlational. Comparing alternative models and evaluating absolute prediction performance might eventually converge on the true underlying feature model, but this is not guaranteed. We propose that an even better way to test representational hypotheses might be to introduce experimental manipulations that affect the hypothesized representational dimensions. For example, one could prime participants to weight some features of the stimuli more than others. If that leads to changes in the performance of a classifier based on the primed features, this would constitute much stronger evidence that these features underlie the neural representational scheme in question. This proposal is logical but it has not been experimentally tested yet, and we look forward to seeing how it will fare in practice.

References

Anderson, A. J., Bruni, E., Lopopolo, A., Poesio, M., & Baroni, M. (2015). Reading visually embodied meaning from the brain: visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage*, 120, 309-322.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Bonnici, H. M., Kumaran, D., Chadwick, M. J., Weiskopf, N., Hassabis, D., & Maguire, E. A. (2012). Decoding representations of scenes in the medial temporal lobes. *Hippocampus*, *22*(5), 1143–1153.

Brouwer, G. J., & Heeger, D. J. (2009). Decoding and reconstructing color from responses in human visual cortex. *The Journal of Neuroscience*, *29*(44), 13992–14003.

Casey, M., Thompson, J., Kang, O., Raizada, R., & Wheatley, T. (2012). Population Codes Representing Musical Timbre for High-Level fMRI Categorization of Music Genres. In *Machine Learning and Interpretation in Neuroimaging* (pp. 34–41). Springer Berlin Heidelberg.

Correia, J., Formisano, E., Valente, G., Hausfeld, L., Jansma, B., & Bonte, M. (2014). Brain-Based Translation: fMRI Decoding of Spoken Words in Bilinguals Reveals Language-Independent Semantic Representations in Anterior Temporal Lobe. *The Journal of Neuroscience*, *34*(1), 332–338.

Coutanche, M. N. (2013). Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? *Cognitive, Affective & Behavioral Neuroscience*, *13*(3), 667–673.

Coutanche, M. N., & Thompson-Schill, S. L. (2015). Creating Concepts from Converging Features in Human Cortex. *Cerebral Cortex*, *25*(9), 2584–2593.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. *Annals of the New York Academy of Sciences,* 1296(1), 108–134.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339.

Fernandino, L., Humphries, C. J., Conant, L. L., Seidenberg, M. S., & Binder, J. R. (2016). Heteromodal Cortical Areas Encode Sensory-Motor Features of Word Meaning. *Journal of Neuroscience*, *36*(38), 9763–9769.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*(1), 435–456.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Kay, K. N., Naselaris, T., Prenger, R. J., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, *452*(7185), 352–355.

Lee, H., & Kuhl, B. A. (2016). Reconstructing Perceived and Retrieved Faces from Activity Patterns in Lateral Parietal Cortex. *Journal of Neuroscience*, *36*(22), 6069–6082.

Mahon, B. Z. (2015). The Burden of Embodied Cognition. *Canadian Journal of Experimental Psychology*, *69*(2), 172–178.

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.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Phil. Trans. R. Soc. B*, *369*(1651), 20130299.

Naselaris, T., & Kay, K. N. (2015). Resolving ambiguities of MVPA using explicit models of representation. *Trends in Cognitive Sciences*, *19*(10), 551–554.

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410.

Norman, K., Polyn, S., Detre, G., & Haxby, J. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

Poldrack, R. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends in Cognitive Sciences*, *10*(2), 59–63.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, *72*(5), 692–697.

Santoro, R., Moerel, M., Martino, F. D., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLOS Computational Biology*, *10*(1), e1003412.

Tong, F., & Pratte, M. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*(1), 483–509.