Decoding mental activity from neuroimaging data—the science behind mind-reading

Kay H. Brodersen[1,2,3]

[1] Oxford Centre for Functional MRI of the Brain (FMRIB), University of Oxford, United Kingdom
[2] Department of Computer Science, ETH Zurich, Switzerland
[3] Institute for Empirical Research in Economics, University of Zurich, Switzerland

Abstract

At the interface of neuroscience and computer science, a new method of analysis has evolved. The idea of reading out mental activity from neuronal measurements has led to increasingly impressive feats of mind-reading. What sounds like science fiction is well-positioned to become a major tool in future brain research.

1. Understanding the brain

The mechanisms by which the brain makes sense of the world rely on computational abilities far beyond any human piece of engineering. It has therefore been a fundamental goal in neuroscience to understand just how our central nervous system analyses sensory inputs, forms an internal cognitive state, and produces behavioural outputs. In recent decades, an overwhelming amount of evidence has accumulated that supports the early hypothesis that populations of nerve cells, or neurons, provide the basic functional processing unit of the brain. Neurons drive one another's activity in highly interconnected groups that assemble and disperse on a millisecond scale, and the dynamics of these ensembles of cells are believed to give rise to the cognitive abilities of the brain. With more and more experimental and theoretical results coming in, many scientists believe that we may eventually solve the neural puzzle—and achieve a detailed understanding of the brain.

But what exactly do we mean by an 'understanding of the brain?' Can we be said to 'understand' the brain once we have come up with a wiring diagram of its 100 billion neurons? Do we 'understand' it once we have written down the differential equations governing the dynamics of its 100 trillion synapses? Such approaches must inevitably fail. A more fruitful way of thinking about the question is: can we demonstrate how the inner workings of the brain relate to cognitive abilities? In other words: can we establish a mapping between structure and function, between brain and mind?

The idea of reading out, or decoding, mental activity from neuronal measurements has been driving the formation of increasingly multidisciplinary research groups [8]. However, as neuroscientists and computer scientists team up, two fundamental challenges have become apparent. First, there are currently no methods available to record the activity of a larger number of individual nerve cells in an awake human being, let alone to obtain high-resolution whole-brain footage of neural activity. Second, the brain displays immense natural variability in structure, connectivity, and dynamics. Your brain is very different from your friends, and it is very different from itself as it was just a few minutes ago. Yet, more recently, increasingly marked advances in decoding have been achieved.

"We show that [our] models make it possible to identify, from a large set of completely novel natural images, which specific image was seen by an observer," Kendrick Kay and colleagues lately reported in the journal Nature [9]. Their ability to tell, by scanning someones brain, which picture they were looking at, is the result of a study carried out at the University of California, Berkeley. It is about decoding information from the visual system—the part of the brain that processes what we are currently looking at. And being able to tell which image was seen out of a fixed set of images is not the end of the story: "Our results suggest that it may soon be possible to reconstruct a picture of a persons visual experience from measurements of brain activity alone." [9]

The idea of engineering a general brain-reading device has long been stimulating researchers' imaginations. Psychologists claim it could be used to investigate perception and consciousness [6, 7, 15]. Neurologists say it could be used to construct brain-computer interfaces for paralyzed patients [18]. Lawyers wonder whether it could be used for lie detection [1]. As basic research increasingly elucidates

the neural mechanisms underlying cognition, we may begin to use this knowledge in reverse: to decipher a cognitive process from its neural correlates.

Decoding relies on two techniques. First, neuroimaging has made it possible to obtain correlates of the summed activity of populations of neurons across the whole human brain [16]. Second, the theory of machine learning has given rise to powerful algorithms that are able to recognize patterns in measured brain activity, and associate them with mental states [13]. The combination of these two fields comes with many challenges, and results require extremely careful interpretation. But it opens up a treasury of exciting applications, and never before have we been so close to their realization.

2. What is the brain thinking about? Measuring neural activity using fMRI

High-level phenomena such as memory or consciousness are difficult to localize: they emerge from the distributed activity of many parts of the brain. By contrast, more basic functional building blocks have been pinpointed to particular cortical areas. Sensory inputs, for example, are known to arrive in dedicated hierarchical structures of the brain including the visual cortex (seeing), the auditory cortex (hearing), and the somatosensory cortex (touching). Similarly, behavioural outputs are passed on to the spinal cord by an area referred to as the motor cortex. In between are association areas that effectively allow any sensory input to trigger any motor output. One technology in particular has fuelled these insights: functional magnetic resonance imaging (fMRI) makes it possible to record neural activity from the brain of a participant who is happily performing some kind of cognitive task.

Neural activity is expressed in terms of increased signalling between nerve cells, which, in turn, leads to an increased demand in oxygen. As a result, the level of blood oxygenation rises. The precise details of the underlying cascade of biochemical events are not fully understood, but the effect is of immense use: an MRI scanner is able to pick up subtle changes in blood oxygenation as direct correlates of neural activation [16]. How can we employ this technology to infer something interesting about the brain?

In a typical fMRI experiment, a participant lies in a large magnetic coil and is asked to watch a screen, listen to a sound, press some buttons, navigate in a 3D maze, or perform any other kind of task. In the same way as a digital camera divides up an image into a grid of small pixels, the MRI scanner divides up the brain into voxels, small cubes with a volume of, e.g., $3 \times 3 \times 3$ mm^3. A complete recording then contains a time series of neural activity from each voxel throughout the duration of the experiment.

In a cognitive neuroscience setting, for example, participants might be asked to play a gambling game in which they have to place a bet on either of two cards, and, by trial and error, adopt a successful strategy to maximize their winnings. Given their recorded neural activity, we might now begin by looking for those regions that display systematic differences in activity between periods when participants are at rest and periods when they are making a decision. Technically, we predict what the signal in a voxel should look like if the nerve cells in that voxel were concerned with decision making: low activity during rest, and high activity just before a decision. As a result of our analysis we may find various areas in the brain whose activity appears to follow our prediction: low blood oxygenation during rest, and high oxygenation just before a decision. We might then conclude that these regions are involved in the mental process of making a decision.

There are many caveats associated with this kind of analysis and the interpretation of its results. Nevertheless, careful experimental design and the use of converging evidence have established fMRI as the method of choice for human brain research.

3. Decoding mental activity

When trying to decode mental activity from neural recordings, the conventional analysis described above is modified in two ways. First, rather than predicting the time course of neural activity from a design matrix, we aim to predict parts of the design matrix from the time course of neural activity [8]. Second, rather than considering all voxels independently, we aim to understand how patterns of voxel activities jointly encode information. The first modification is important when the aim is prediction per se, that is, in applications such as lie detection. The second modification is key when it comes to inference on structure-function mappings, that is, in basic research. In either case, we take a snapshot of the activity measured simultaneously at many locations in the brain, and map it onto a particular mental state. These states are often defined in terms of discrete classes, which, in the example above, could be labelled 'rest' and 'decide.'

In this way, decoding can be viewed as classification, a key problem studied by a branch of computer science known as machine learning [13]. How does it work?

The learning methodology. It takes no more than a few years until children can easily recognize digits and letters, or detect a single female face in a series of male ones. For computers, however, tasks of pattern recognition are among the most difficult ones. It is unknown how to teach a machine to flawlessly recognize faces or separate personal e-mails from unwanted spam because no mathematical model of the problem is available, or its implementation is compu-

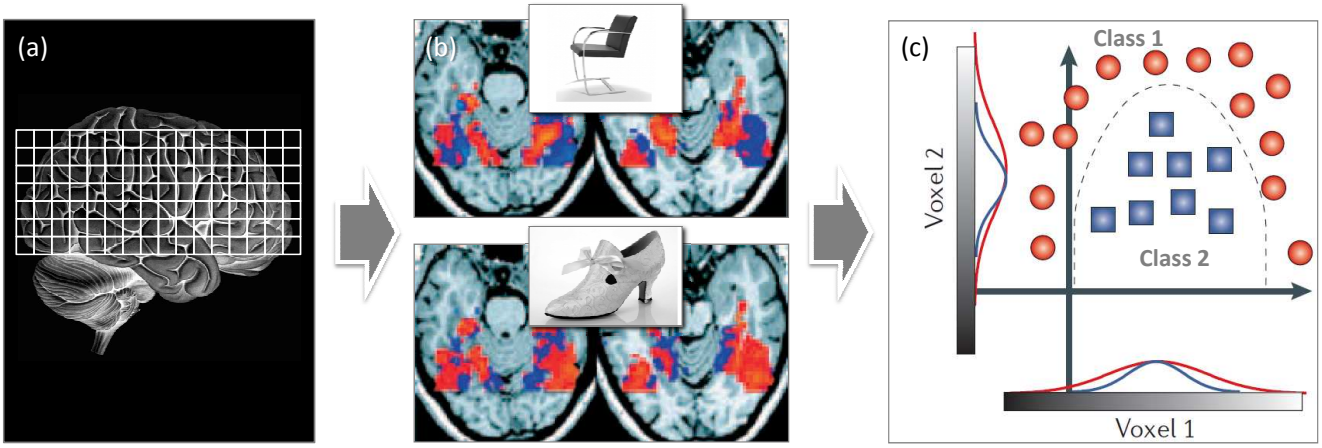

Figure 1. **The classification pipeline.** (a) An MRI scanner splits up the brain into thousands of small three-dimensional voxels. Here they are illustrated by a (two-dimensional) grid. The scanner records neural activity within each voxel while participants are being engaged in different cognitive tasks. For example, they might be presented, one after another, with many different images of chairs and shoes. (b) The learning algorithm is trained on the neural activity that was recorded while participants were looking at chairs or shoes, respectively. Impossible to pick up by eyesight, the algorithm can find subtle, yet systematic, differences in brain activity between the two classes. (c) As a result of the learning phase, a classifier can tell apart the two classes of mental perception. The example illustrates a case where input data are two-dimensional, containing activity from just two voxels. Each voxel on its own is not sufficiently informative to allow for accurate classification, but together they allow for reliable separation. In a real setting, feature spaces of hundreds of dimensions might be considered. (Fig. 1b, c adapted from [8].)

tationally too expensive [10, 20]. In most instances, the human brain outperforms all algorithms devised so far. What is its secret?

When the brain learns to tell apart the 26 letters in the alphabet, it is not given a formal description of the morphology of each and every letter. Instead, it learns by example. This idea has been taken on by statistical learning theory, where it is referred to as the learning methodology. It is the idea that a computer should learn by inductive inference, that it should learn from examples, rather than being programmed to solve a particular problem explicitly. This is precisely what we can make use of when aiming to decode a mental state from neural activity.

Classification and cross-validation. From a statistical learning point of view, the goal of mind reading is to learn about input-output pairings, where the input domain is a high-dimensional feature space of neural activity, and the output range is a set of discrete classes of mental activity. In principle, these classes could be about any type of measurement that can be tied to an individual example of data: which type of stimulus a participant saw; what action they chose; what the outcome of the trial was. To do this, we need training data: a set of examples of neuronal recordings for which we know what the respective cognitive states were. A learning algorithm then attempts to find a systematic relationship between the training examples and their respective classes. But crucially, in order to assess whether the algorithm has in fact picked up a robust relationship, we need to run it on entirely separate test data: previously unseen examples for which, again, we know the respective classes but do not allow the algorithm to see them. Think of the purpose of an exam: the classifier has been taught many individual facts during the training phase; now we should assess whether it is not only able to repeat what it has been taught but has actually understood the underlying principles—by asking it to apply its knowledge to new cases. The percentage of correctly classified test examples is what we can then report as the accuracy of our mind-reading device (see Figure 1).

In the gambling game described above, for example, we might try and decode the choice participants were about to make on a given trial: were they going to pick the blue card or the green card? To the classifier, an input example is a vector containing the neural activity from each voxel in the brain, recorded just before the decision of a particular trial. The two output classes are 'blue' and 'green.' If a participant has played the game 100 times, we could, for instance, train the classifier on 80 trials and test it on the remaining 20 ones. If the classifier correctly predicts whether the participant chose the blue or the green card in 15 cases, we can report an accuracy of $15/20 = 75\%$, corresponding to a significance level of $p < .006$. In other words, it is very unlikely that the classifier was merely guessing—in which case we would expect about 10 correct predictions.

The fact that the dataset must be split up into separate

training and test sets raises an interesting question: how much of the data should be used for training the classifier, and how much should be kept separate to test it? On the one hand, the larger the training set, the more can the classifier learn about what neural activity tends to look like in each of the classes. On the other hand, the larger the test set, the better our estimate of the actual accuracy. In particular, it may easily happen that the classifier overfits by learning an extremely complex relationship between neural activations and cognitive class which generalizes poorly to unseen test examples. Fortunately, there is a clever way of dealing with this problem: a method termed cross-validation that allows us to present the classifier with as much training data as possible, yet obtain a reliable estimate of its generalization ability.

To begin with, we train our classifier on all trials but one—that is, the classifier gets to see 99 examples of neural activity (trials 1 through 99) and their corresponding class: 'blue' or 'green.' We then test the classifier on the single remaining trial that was held out (trial 100), giving us either a correct or an incorrect prediction. And here comes the trick: we can re-run the whole procedure of training and testing while, this time, we are testing on the penultimate trial (trial 99) and training on the 99 other ones (trials 1 through 98 and trial 100). Repeating this 100 times, each time working with a different split of training vs. test data, yields 100 predictions which we can average to get a good estimate of the true accuracy of the classifier. For example, if the classifier has predicted correctly in 64 out of 100 cross-validation folds, we can report an accuracy of 64%. This is slightly worse than the estimated 75% from above. However, it is more accurate and, in fact, corresponds to a better significance level of $p < .002$. In other words, we can be even more confident that the classifier has truly found some relationship between neural activity and chosen card, and we can reject the null hypothesis of it just operating at chance.

Feature selection. One issue we have only implicitly dealt with so far is the problem of feature selection. It turns out to be crucial in many machine-learning applications [5], and it holds particular challenges in a neuroimaging setting [2, 13, 18, 19]. A typical dataset may contain a time series of acquired volumes of, say, $64 \times 64 \times 45$ voxels required to cover the entire cerebral cortex. If we take a snapshot of the neural activity within each voxel as input to the classifier, we end up with a vector of 184,320 values, or features. Each example could be represented as a point in a 184,320-dimensional cube—which usually makes it impossible for a classifier to learn which particular combinations of features belong to which class. The high dimensionality of the input space therefore needs to be reduced to a smaller number of features presented to the learning algorithm. Ideally, this allows the classifier to focus on a few informative features and ignore the other ones [5].

If each feature represents the neural activity measured in a particular voxel, then one way of selecting features is to confine ourselves to certain regions of interest. For example, we might hypothesize that participants who are about to choose the blue card tend to actually look at the blue card just before making their decision, and hence choose voxels from the visual cortex as features. However, this requires some prior knowledge and does not always allow for a strong reduction in dimensionality. It may also defeat the point of using the classifier to find out, rather than presuppose, which parts of the brain contain information about the participant's decision [19]. One solution, among the many skillful methods that have been proposed, is to choose those voxels that allow for above-chance classification just by themselves. Here is how it works: we consider each individual voxel in turn and run a full cross-validation analysis using only the neural activity of this particular voxel. The resulting accuracy is viewed as this voxel's score. Once we have tested all voxels individually, their scores allow us to select the best ones for the main analysis. In doing so, it is important that feature selection only operates on the training set and never sees the test set. This finally sets the stage for the key theme: how can we use a classifier?

4. Tackling new questions

Conventional fMRI analysis predicts neural activity in individual voxels from a model that captures the expected observations. Decoding, by contrast, predicts parts of the model from entire patterns of neural activity measured simultaneously in many voxels. While conventional analysis has proven very successful over the past 15 years, decoding allows us to explore the brain from a new angle. While conventional analysis typically treats each voxel in the brain independently, decoding is a multivariate approach: it allows for the identification of patterns emerging from whole ensembles of voxels, it does not impose any constraints on spatial contiguity, and it is less severely affected by the natural variability in people's brains. Decoding is by no means the only multivariate approach that has been applied to neuroimaging datasets [11, 12]—but it is the one that has attracted most interest in recent years. It allows us to tackle four key questions from a new perspective [18].

Whether, where, when, and how. First, the question of pattern discrimination: does the recorded neural activity carry information about a variable of interest? From the point of view of pattern discrimination, a classifier is an information-extraction device. It can be used for inference on structure-function mappings in the brain, but may also provide new methods for clinical diagnosis, brain-computer interfaces, or lie detection (see Figure 2a).

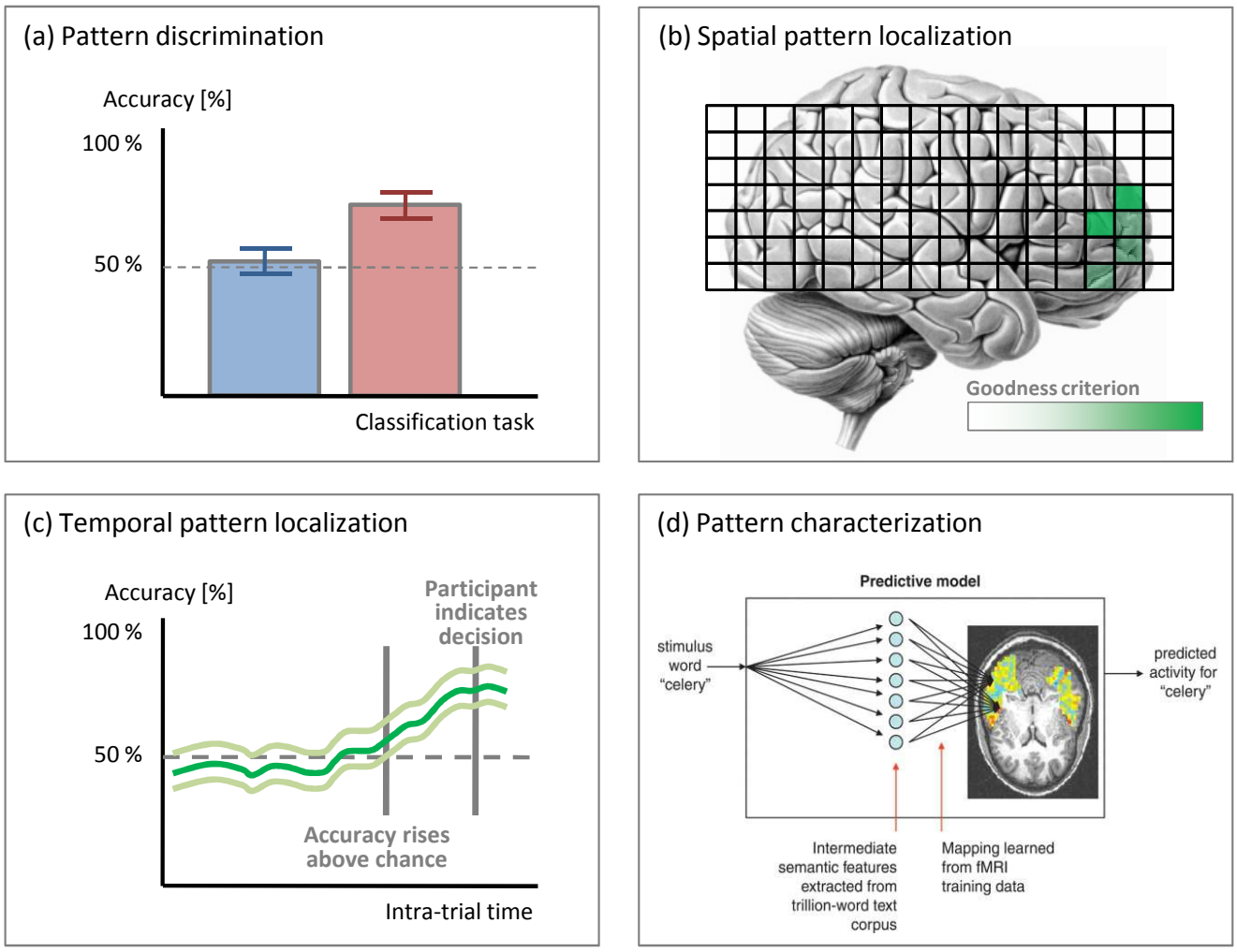

Figure 2. **Applications of decoding analyses.** (a) When asking whether we can decode a variable of interest from neural activity, we need to compare our classifier's performance with chance level. If the algorithm fails to extract information, its accuracy is expected not to differ significantly from chance (left column); by contrast, a classifier performing above chance provides evidence for the presence of information about the variable of interest (right column). (b) Having demonstrated the extraction of information, we can ask where in the brain this information is encoded: by finding those voxels that drive the classifier's success. (c) Within a given area, we can examine when patterns become sufficiently distinct to allow for classification. For example, we might find that a brain area contains information about a participant's choice before they indicate this choice by pressing a button. (d) Characterizing how an abstract stimulus translates into neural activity is the most sophisticated type of analysis. The example shown here illustrates how a word can be decoded [13]. For a given noun (e.g., 'celery'), it is first determined how much this noun generally co-occurs with a number of basic activities (e.g., 'hear' or 'taste' or 'run'). A predictive model then describes how every voxel in the brain appears to be tuned to these activities. When a new word from a dictionary is presented to the participant, the recorded activity can be compared to the predicted activity for all words in the dictionary. The classifier finally chooses the word that matches the recording best. (Fig. 2d adapted from [13].)

Second, the question of spatial pattern localization: where in the brain is class information encoded? Once we have established that neural activity carries information about a variable of interest, we can investigate which voxels contribute most strongly to the classifier's success (see Figure 2b).

Third, the question of temporal pattern localization: when does information take shape in, or become available to, a certain brain area? In the decision-making experiment described earlier, for example, participants are typically asked to indicate their choice by pressing a button. We may now ask how early we can predict their choice. To do this, we would re-run the classifier analysis many times, each time granting the classifier access to another second of

input data, up to the point in time when the decision was finally made. We could then report how early the classifier began to perform significantly above chance level (see Figure 2c).

Fourth, the question of pattern characterization: how is information encoded in the brain? In order to answer this most intricate of all decoding questions, researchers have begun to take the design of classification analyses to yet another level [4]. Rather than training an algorithm to distinguish between neural patterns of, say, two different classes, we can try and map neural activity onto the correct class out of dozens if not hundreds of classes, where the classifier has only seen training examples from a small number of classes. To achieve this, the classifier, during the training phase, constructs a model for each voxel known to contribute to overall prediction accuracy. The model describes what abstract features of a stimulus each particular voxel is tuned to. For a potentially large set of test stimuli, the classifier can then compute the expected activation in each voxel if the brain was exposed to that particular stimulus. Given neural activity recorded while one of these stimuli was actually presented to a participant, the classifier chooses the stimulus whose predicted pattern of neural activity most closely matches the observed pattern (see Figure 2d).

A study by Tom Mitchell and colleagues of Carnegie Mellon University, published in Science in May 2008, illustrates this idea: "We present a computational model that predicts the [...] neural activation associated with words for which fMRI data are not yet available." In other words, given an arbitrary word presented to a participant in an MRI scanner, the algorithm, out of a large corpus of nouns, finds the word that was most probably being presented [14]. This brings us back to the findings by Kendrick Kay's group, mentioned at the beginning, who used precisely the same idea to engineer a decoding algorithm for the visual system: given an image shown to a participant, their algorithm would find, out of a large set of images, the one image that was most likely being shown [9]. Both studies illustrate the idea of prediction for the purpose of inference: classification is not used for its own sake, but allows for new insights about the link between brain and mind [3].

What the future holds. Having evolved at the interface of neuroscience and computer science, decoding mental states from neural activity is well-positioned to become a key tool in brain research [8, 17]. The spatial and temporal resolution of imaging techniques such as fMRI is limited, and will remain so for the foreseeable future. Learning algorithms currently lack the robustness they need for wide applicability, and no principled guidelines for solving the problem of feature selection have been agreed on yet. Nevertheless, skilful analyses have already led to a whole range of impressive findings about information processing in the brain. The field is not only likely to continue doing so in the future; it also demonstrates how multidisciplinary research can give rise to insights each discipline on its own would not have achieved.

References

[1] C. Davatzikos, K. Ruparel, Y. Fan, D. Shen, M. Acharyya, J. Loughead, R. Gur, and D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3):663–668, 2005.

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

[3] K. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and J. Ashburner. Bayesian decoding of brain images. *NeuroImage*, 39(1):181–205, 2008.

[4] A. B. A. Graf, O. Bousquet, G. Rtsch, and B. Schlkopf. Prototype classification: Insights from machine learning. *Neural Computation*, 2008. PMID: 18624661.

[5] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

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

[7] J. Haynes and G. Rees. Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, 15(14):1301–1307, 2005.

[8] J. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7(7):523–534, 2006.

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

[10] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi. Recent advances in visual and infrared face recognitiona review. *Computer Vision and Image Understanding*, 97(1):103–135, 2005.

[11] A. R. McIntosh, F. L. Bookstein, J. V. Haxby, and C. L. Grady. Spatial pattern analysis of functional brain images using partial least squares. *NeuroImage*, 3(3):143–157, 1996.

[12] M. J. McKeown, S. Makeig, G. G. Brown, T. P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski. Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping*, 6(3):160–88, 1998. PMID: 9673671.

[13] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1):145–175, 2004.

[14] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human

brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[15] K. A. Norman, S. M. Polyn, G. J. Detre, and J. V. Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, 10(9):424–30, 2006. PMID: 16899397.

[16] S. Ogawa, T. M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proceedings of the National Academy of Sciences of the United States of America*, 87(24):9868–9872, 1990.

[17] A. J. O'Toole, F. Jiang, H. Abdi, N. Penard, J. P. Dunlop, and M. A. Parent. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, 19(11):1735–1752, 2007.

[18] P. G. Patil and D. A. Turner. The development of brain-machine interface neuroprosthetic devices. *Neurotherapeutics: The Journal of the American Society for Experimental NeuroTherapeutics*, 5(1):137–46, 2008. PMID: 18164493.

[19] F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199–S209, 2009.

[20] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Comput. Surv.*, 35(4):399–458, 2003.