A critique of multi-voxel pattern analysis

Michael L. Anderson (michael.anderson@fandm.edu)
Department of Psychology, Franklin & Marshall College
Lancaster, PA 17604 USA

Tim Oates (oates@cs.umbc.edu)
Department of Computer Science, University of Maryland, Baltimore County
Baltimore, MD 21250 USA

Abstract

Multi-voxel pattern analysis (MVPA) is a popular analytical technique in neuroscience that involves identifying patterns in fMRI BOLD signal data that are predictive of task conditions. But the technique is also frequently used to make inferences about the regions of the brain that are most important to the tasks in question, and our analysis shows that this is a mistake. MVPA does not provide a reliable guide to what information is being used by the brain during cognitive tasks, nor where that information is. This is due in part to inherent run to run variability in the decision space generated by the classifier, but there are also several other issues, discussed here, that make inference from the characteristics of the learned models to relevant brain activity deeply problematic. These issues have significant implications both for many papers already published, and for how the field uses this technique in the future.

Keywords: neuroscience, machine learning, inference, philosophical issues.

Introduction

Multi-voxel pattern analysis (MVPA) is an increasingly popular analytical technique in neuroscience. MVPA involves searching through the Blood Oxygenation Level Dependent (BOLD) signal data produced in fMRI experiments to identify patterns that are highly predictive of task conditions. To illustrate, consider a simple experiment in which participants are asked to view pictures representing various object categories (e.g. faces, houses, chairs, shoes, etc.). One early MVPA study showed it was possible to determine, by looking only at BOLD data, which class of object an experimental participant was viewing when that data was collected (Haxby et al., 2001). The technique has since been used to predict the orientation of lines being viewed by a participant (Haynes & Rees, 2005), to differentiate between lying and truth-telling (Davatzikos et al., 2005), and to predict which action a participant was about to take (Haynes et al., 2007), among many other things (see Pereira, Mitchell & Botvinick, 2009; Norman et al., 2006; Haynes & Rees, 2006 for reviews of the technique and its applications).

This is indeed impressive, and we expect that MVPA will have many important experimental and diagnostic applications (Lao et al., 2004). It has become commonplace to make certain inferences about the way differences in BOLD signal patterns correspond to differences in mental states. For instance, by finding the set of voxels that are most predictive of a certain task outcome, studies have claimed to discover the "cognitive states associated with perception of tools and dwellings" (Shinkareva et al., 2008), "localizable task-specific representations of freely chosen intentions" (Haynes at al., 2007), and the regions of the brain that "contain information" (Preston et al., 2008) relevant to the cognitive or perceptual task under investigation.

To put it bluntly, however, such inferences are at best misleading and at worst entirely unwarranted. The issues dovetail with, but are distinct from, the more general concerns about the unreliability of "reverse inference" from neuroimaging data (Poldrack, 2006), and have significant implications both for how we ought to interpret some of the many papers already published, and for how the field applies this technique in the future.

Of course, not every MVPA study is governed by the logic that we will criticize here. For instance, Mitchell et al. (2008) take something like the opposite approach, and see if they can predict the pattern of brain activity that will be caused by listening to novel words. Here the point of the study is not to discover which brain regions are responsible for understanding; rather, they are testing the hypothesis that meanings of words are based on sets of "semantic features" that can be inferred from word co-occurrence in language corpora. McDuff, Frankel & Norman (2009) are likewise focused on hypothesis testing, in their case about the characteristics of targeted memory retrieval. We think that MVPA has a very promising future both as a diagnostic tool, and as a useful dependent variable—in part because the technique is sensitive to contingencies beyond classical single-voxel effects—but that for the reasons outlined in this paper it is a very poor tool for reliably localizing information or identifying cognitive states.

Information and the brain

There are three general ways in which information could inhere in the BOLD signal. First, the information could be non-local, that is, carried by irreducibly relational features of the signal like regional co-variance. We might expect this to occur when large-scale neural synchrony is the relevant aspect of brain activity (Varela, et al., 2001; Gross et al., 2004). Second, it could be local and distributed, that is, the information could be carried by the activity of individual voxels, and the information-carrying voxels could be spread throughout the brain. We might expect this for cognitive processes that require the cooperation of many different brain regions. Third, the information could be local and concentrated, that is, carried by individual voxels that are grouped together in one or a few clumps. This might happen when the work done by local neural circuits is most important to the cognitive task(s) in question. In this essay, we will consider the performance of MVPA in all three

situations, and discuss what can, and cannot, be inferred from features of the learned model in each case.

Local, distributed information

Consider the problem of differentiating between the following two patterns of hypothetical voxel-level activation data, each presented with two *versions* of the same pattern *type* (see Figure 1). In this simple example each of the 25 "voxels" can be in one of two states (active or inactive, if you like). Suppose "brain scans" like these had been observed during an experiment in which participants were asked to classify pictures as "living" or "non-living". If these judgments reliably corresponded to the two patterns, respectively, could we use MVPA to read the mind of the participants?

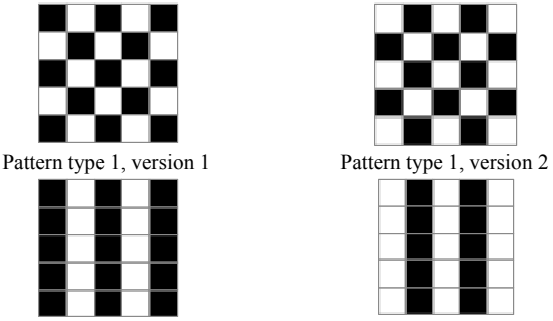

Figure 1: Simple patterns for use in MVPA test

Now, when the ratio of pattern versions within each pattern type is 1:1, every voxel is in both of its possible states in every task condition. That is: no voxel is by itself predictive of any cognitive state, and thus in this condition all information is non-local. In this condition *linear* MVPA cannot distinguish between these two patterns; it is blind to non-local information (Kamitani & Tong, 2005; Norman et al. 2006). For linear classifiers, since the evidence provided by each voxel is integrated separately, linear MVPA is successful only when there are individual voxels that are sensitive to the difference between classes. In general, a (binary) linear classifier over an input space of dimension n looks like this:

$$\text{prediction} = \text{sign}\left(b + \sum_{i=1}^{n} w_i * x_i\right)$$

where the i^{th} weight is w_i and the i^{th} component of the input vector (the list of numbers that describe the patterns to be classified) is x_i and the bias value is b. If the sum above is positive, the instance is classified one way; if it is negative, the instance is classified the other way.

However, manipulating the version ratio changes the situation from one in which no voxel is more informative than any other—a situation in which linear classifiers fail—to one in which there is indeed a set of voxels, scattered through the patterns, that are informative about class membership. That is to say, although there is still non-local information in the patterns—and it is arguable that the non-local co-variance structure is the crucial, relevant distinction between these patterns—the initial test situation is one in which there nevertheless is also relevant local information, distributed across many voxels.

For our analysis of the performance of MVPA with local, distributed information, we generated 20 sets of 80 "scans"—that is, 20 datasets, each containing 40 instances of each pattern type. Patterns were corrupted with 5% noise—a 5% chance for each voxel that it will be in a state inconsistent with the pattern. For each dataset, we used 40 of the 80 scans for training and 40 for test, and classified them using a Support Vector Machine. Because classification accuracy roughly tracks the relative proportion of pattern versions, our scans contained a 4:1 ratio of pattern versions within each type, and classification accuracy averaged 80%.

Thus, our hypothetical experiment would have produced a solid predictive success; we would be able to tell, 80% of the time, which task condition the participant was in just by looking at the fMRI data. But what, if anything, would we be permitted to conclude about the local neural conditions—representations, information content, activity, etc.—contributing to the differences in cognitive tasks (thinking about or judging the difference between living vs. non-living things)?

Although any of the input components could contribute to the prediction breaking one way or the other in an given case (and it needn't be the same components for each instance), in practice there can be a small number of voxels that contribute most to the classifier performance because they (literally) carry the most weight—that is, they have the highest values of w_i. In linear MVPA, this set of highly weighted voxels is considered the "most informative".

Figure 2 shows a map of the voxels that were most informative for distinguishing between pattern types 1 and 2 in dataset 1.

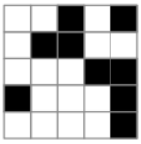

Figure 2: Most informative voxels for an MVPA classification

What is the proper interpretation of these results in the context of MVPA? These are the voxels that, had they been in a different state, would have been most likely to cause the classifier to place the pattern in the other class. But consider the following inference, an inference of similar structure to those being made in the MVPA literature: if the state of these voxels had been different in the right way—and note this picture provides no information about what the right way is—the brain would have been in the relevantly different state (or the participant would have been in the different cognitive state). This inference does not follow, because if covariance is the crucial cognitively relevant property of the activity here, then all the other voxels would also be different when the brain/participant is in the other state: they will be covarying with a different set of partners. And, even if covariance is not the crucial property—if the relevant information is the local information—it seems pretty clear that it isn't all or only the voxels in the "most

informative" set that would need to be in a different state to turn one pattern into the other.

Likewise, consider a similar inference (versions of which can also easily be found in the literature): the information contained in these voxels is the information crucial to the difference between the cognitive states under investigation (judging living vs. non-living things). This inference is also unwarranted, for similar reasons. For one strong possibility is that the relevant information is carried by the covariance structure of the patterns, and this non-local information is not contained in the set of "most informative" voxels. And even if the local information is what is relevant here, we can see from the results above that the set of most informative voxels does not consist of all or only the voxels carrying the relevant information.

The uncertainty of inferences about brain or cognitive states based on which voxels are most highly weighted is driven home even more strongly when one looks at the stability of the set of highly weighted voxels over multiple trials of the same task. Figure 3 shows the most highly weighted voxels from the first three datasets.

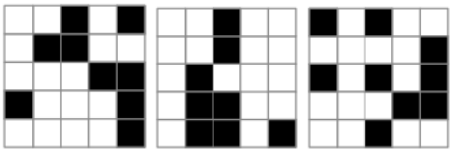

Figure 3: Most informative voxels for three different classification runs

Obviously, the highly weighted voxels vary from run to run. To get a better quantitative handle on the stability of the highly weighted voxel set, we counted the number of times each voxel was among the top 10 most highly weighted. Overall, every voxel was in this set at least twice, and none more than 12 times. 24 voxels were in the set between 6 and 12 times, and 22 between 6 and 10 times. The characteristics of the classification model can vary considerably, driven in part by noise in the training instances, but also by the fact that the algorithm needs only find *some* of the features that discriminate between *some* instances of the patterns *some* of the time. It is not guaranteed to find all of the relevant differentiating features, nor the best. The conclusion seems obvious, but is worth stating clearly: when any voxel can make it into the "most informative" set, and many voxels are more or less equally likely to end up there, this should make us a bit uneasy about their actual informativeness. If there is something stable to the cognitive states differentiating the task conditions, the set of most informative voxels is certainly not tracking it, nor can it therefore be a reliable indicator of the location of the cognitively relevant information.

In should be noted that cross-validation does not alleviate this issue. Cross-validation consists of a family of methods designed to prevent over-fitting of the model to what could be an unusually biased sample. Typically, it involves building multiple models based on multiple partitions of the sample, and averaging results over the range of different partitions (Pereira et al., 2009). For instance, K-fold cross-validation involves splitting the training data into K parts, and training K times on a rotating K-1 of the partitions. We performed 10-fold cross validation on our 20 training sets, and found similar variability in the set of most informative voxels in each fold. The mean number of inclusions among the top 10 most highly weighted voxels was 4 (SD 2.83). 23 of the voxels were among the 10 most highly weighted voxels in at least one fold.

Non-local information

So, that seems to be the situation when information is local and distributed. What about when the only information is non-local, that is, when the ratio of pattern versions is 1:1? It turns out it is possible to classify these patterns with 100% accuracy, applying MVPA using a support-vector machine with a polynomial kernel of degree two. Can we conclude anything in this case about the neural conditions—representations, information, activity, etc.—contributing to the differences in cognitive tasks?

One is of course tempted to simply dismiss the possibility. In our examples every voxel is in both of its possible states in every task condition. That is: no voxel is by itself predictive of any cognitive state, and thus no inference to the special status of activity in any voxel could possibly be supported by the predictive success of MVPA. Yet researchers do extract "most informative" voxel sets even when using polynomial kernels (e.g. Davatzikos et al., 2005), so it is wise to consider the matter more carefully.

In linear classifiers identifying most important voxels for the classifier is easy—the features in the decision space that have the highest weights are the most important, and these features have a 1:1 correspondence with components of the input vector, that is, with the voxel values fed into the classifier. But non-linear SVMs use "kernel functions", whereby a vector input is projected into a kernel-specific high-dimensional space, and the importance of each feature is determined in that space. The original space for a given vector x has one dimension for each component of the input vector, and the value of that feature—its position on dimension i—is just x_i. In contrast, a polynomial kernel of degree 2 (K_2) projects the input vector into a space having a dimension for each unique (unordered) pair of features in the vector, and the value of each of those features—its position on dimensions (i,j)—is $x_i * x_j$.

Thus, the K_2 classifier over an input space of dimension n looks more like this:

$$\text{prediction} = \text{sign}\left(b + \sum_{i,j=1}^{n} w_{i,j} * x_i * x_j\right)$$

In fact, we can get exactly this situation by manually projecting our input vectors into the polynomial space—turning them in this case from n-dimensional vectors into $n+((n^2-n)/2)$ dimensional vectors—and using linear SVMs to classify them. This procedure will produce the same decision surface as using K_2 in the original space, but will allow us to directly inspect the resulting weights to determine which features were most important.

However, given the nature of non-linear SVMs, relating features to individual components of the input vectors is inherently problematic. For note that what gets weighted in the decision function is the product of each pair of

components. So, if a given product turns out to be important to the classifier, shall we attribute this importance to just one of the components, or to both? Either decision seems likely to give misleading results. Nevertheless, for the sake of the discussion, let's adopt the simple rule that when a given feature is highly weighted, both components (voxels) will be counted as "informative". Given this, we can examine the frequency with which voxels are informative, and track the voxels that are frequently informative.

To test this procedure when using K_2, we generated 40, 10x10 versions of the standard patterns from Figure 1, 20 of each pattern type, with a 1:1 ratio of versions, and a noise level of 5%. We projected each of these patterns into the 5,050-dimensional feature space of K_2, and trained a linear SVM on the set. Then we found the top 500 highest weighted features, and projected these back onto the 10x10 pattern following the rule above. Now, it is perfectly legitimate to make the following inference from this procedure: the highly weighted voxels are the ones that, had they been in a different state, would have been most likely to cause the classifier to place the pattern in the other class. The trouble is, the weighting is often taken to tell us something about the relative importance of each voxel to the intrinsic difference between the patterns (and to the underlying cognitive states), and no such inference is warranted in this case.

First, there is a basic problem of interpretation given that the important features are in fact products of two voxels—so, every time a voxel is deemed informative, it has a partner with which it was important, and the set itself gives no information about the distribution of these partners. Second, it is clear in this case (because there is no local information) that the relevant information differentiating between the patterns is non-local, carried in the covariance structure of the pattern, and this information is not contained in the set of frequently informative voxels. Third, the most highly-weighted features are not those that contain the most information. As in the linear case, they are the features that contained sufficient information to drive the classifier on a given set of training examples. Fourth and finally, as should not be surprising, the set of informative features and informative voxels is highly unstable in this case, as well.

To explore the stability of the set of important features when using K_2, we generated 10x10 versions of the standard patterns above, creating 100 sets of 40 (20 of each pattern) with a noise level of 5%. We projected each of these patterns into the 5,050-dimensional feature space of K_2, and trained a linear SVM on each of the 100 sets. From each of these 100 sets, we extracted the top 500 most important features. Doing a pair-wise comparison of the most important features from each set revealed that, on average, only 101.08 (SD 16.94) of these features (20.21%) were common between each pair. Moreover, the common features varied from pair to pair. Doing a 5-wise comparison of the most important features sets reveals an average of just 0.81 (SD 1.09) of the features (0.16%) are shared across all five sets. Note that despite the instability of the "most informative" feature sets, classification accuracy in all cases was 100%.

Given the high degree of variability in the features considered most important, it seems certain that the set of frequently informative components (voxels) is likewise unstable. To confirm this, we generated 500 training sets of the 10x10 patterns, and, following the procedure above, found the top 500 most important features for each set. Then, we counted the number of times each individual component of the input vector was included in a pair that was in this important feature set. On average, each component was included in the set 10.00 times (SD 0.39). No component averaged fewer than 9 inclusions, or more than 11.00. Once again, if there is some stable difference between the cognitive states in the two task conditions, the set of most informative voxels is certainly not tracking it, nor can it therefore be a reliable indicator of the location of the cognitively relevant information.

Admittedly, this example was based on a very simple rule for mapping features in the multi-dimensional space to components of the original vector, and it is true that more sophisticated procedures for uncovering the most informative components have been developed (Davatzikos et al., 2005; Lao et al., 2004). But insofar as these techniques still depend in one way or another on identifying the most highly weighted features in a multi-dimensional space, and insofar as this set is not determinate for a given classification task, then the results of such analyses need to be interpreted with extreme caution.

Before moving on with the remainder of the analysis, it is worth pausing to summarize the findings. In the case where there is local information relevant to distinguishing patterns, linear MVPA does not reliably find it; and in the case where there is relevant non-local information, carried for instance by covariance patterns, linear MVPA cannot find it, and non-linear MVPA models can make it look as if they were using local information. More importantly, having discovered some features whose state matters most to the classification decision is not the same as having discovered the brain regions whose activity matters most (or even relatively more) to the participant (or her brain). Indeed, these two sorts of information need have no regular correspondence to one another; one need not track, be a reliable indicator of, or be otherwise instructive about the nature, scope or location of the other.

Local, concentrated information

How is this disconnect possible? Consider first an example from the MVPA literature meant to showcase the power of the technique. Haynes and Rees (2005) were able to use MVPA to correctly identify the orientation of visually-presented lines, even when the stimuli were presented briefly and masked so that the participant did not consciously perceive them. That is an intriguing result, and may tell us something interesting about the operation of V1 (the ROI they used to make the predictions). But note the broader implication for the method: since the participants cannot judge the orientation of the lines, they cannot be in whatever cognitive state gives the ability to judge the orientation of the lines. Thus, MVPA can be used to infer features of the task environment from characteristics of the

BOLD signal, without being a reliable indicator of the cognitive state of the participant.

Now consider extending the experiment in the following straightforward way: while the visual stimulus is being shown (and masked), experimenters play an auditory tone from which the participant could reliably infer the orientation of the line. If, as seems likely in this particular case, the most informative voxels for the pattern classifier would remain in V1, this outcome would provide a clear instance in which the information used by the participant and the information used by the classifier would not have the expected relation.

But is such an outcome really possible? In fact, this hypothetical example points in the direction of a well-known fact about the way classification algorithms perform. Numerous theoretical results and a tremendous amount of empirical evidence in machine learning demonstrate that there is no universally best learning algorithm (Wolpert, 1996). Every algorithm has a bias that is appropriate for some problems and inappropriate for others. This is true for the brain, and the same is true of kernels. There is no universally best kernel, and changing from one kernel to another can lead to large changes in the learned decision surface and thus to changes in what features in the data set seem to be important.

The relevance of this problem for MVPA is that a particular set of stimuli may elicit different patterns of activity, call them pattern A and pattern B, in different parts of the brain, and one kernel may be able to detect pattern A but not pattern B, whereas another kernel may be able to detect pattern B but not pattern A. Thus, when relating "most informative features" to "most important activity", the area of the brain implicated in the experiment will change depending on which kernel is used.

To make this concrete, consider two patterns with 20 binary features (f_1 - f_{20}) in which for every instance of the first (positive) pattern the following two conditions hold:

(a) Either $f_{19} = 1$ and $f_{20} = -1$, or $f_{19} = -1$ and $f_{20} = 1$
(b) The sum of the first 5 bits is less than or equal to zero

For every instance of the second (negative) pattern, the following two conditions hold:

(a) Either $f_{19} = 1$ and $f_{20} = 1$, or $f_{19} = -1$ and $f_{20} = -1$
(b) The sum of the first 5 bits is greater than zero

The values of the other bits are chosen uniformly at random from $\{-1, 1\}$. Condition (a) is the logical exclusive or (*XOR*) function on bits 19 and 20 and is easily learned by the polynomial kernel of degree two (the class label is $-\text{sign}(f_{19} * f_{20})$) but is impossible to learn with a linear kernel. Condition (b) is easily learned with a linear kernel (the class label is 1 if $f_1+f_2+f_3+f_4+f_5 \leq 0$ and is -1 otherwise), but is extremely difficult for the polynomial kernel of degree two because it has access to individual feature f_i only as $f_i * f_i$ which is 1 regardless of the value of f_i.

We created 100 datasets based on the above rules and trained an SVM with a linear kernel on both the original feature space and the feature space constructed for the polynomial kernel of degree two. In the latter space, the feature corresponding to $f_{19} * f_{20}$ had an average weight of 3.64. The remaining 209 features had average weights in the range (0.05, 0.10). In the former case, the average weights for features f_1 through f_5 were 1.92, 1.94, 1.94, 1.93, and 1.94. The remaining 15 features had average weights in the range (0.03, 0.10). Clearly, the choice of kernel can have a dramatic impact on which features are deemed important and, in the case of MVPA, which voxels are implicated in various cognitive tasks.

Thus, although much of this paper was spent detailing the worrying instability and potential deceptiveness of the most informative voxel set when information is non-local or distributed, the fact is that even if MVPA were perfectly reliable at the task of finding the most informative features in a data set, the inference from this to the brain activity most important determining the outcome in given task would remain fairly weak. This is because inference from most informative features to most important activity apparently relies on the unwarranted additional assumption that the pattern classification algorithm and the brain are classifying on a relevantly similar basis. While of course no one claims that the success of MVPA shows that the brain is implementing an identical classifier, the issue is that the hypothesis space is different for different classifiers, and so different information will be relevant to each. What is relevant in the brain, and what is relevant to classifying an image of the brain, need not bear much relation.

Conclusion

There are very many challenges to the task of reliably relating the features (of the BOLD signal) most important to classification success to the features (of brain activity) most important to cognitive states/outcomes. By way of summation, consider this general list of possible ways in which these features might fail to relate as expected.

(1) The highly informative elements of the pattern as discerned by MVPA are distributed in the brain in such a way that the brain is anatomically or functionally incapable of integrating the information. If people are nevertheless capable of making the relevant discrimination, it must have been on the basis of different information.

(2) There may well be classes of stimuli that differ in ways undetectable to subjects (under any presentation condition, conscious or otherwise), but which nevertheless create patterns in the BOLD signal allowing for successful classification by MVPA. Consider in this regard an experiment run by Hung et al. (2005). Macaques passively viewed picture stimuli in eight different categories while undergoing direct recording of neural activity using microelectrode arrays. Hung et al. were able to successfully classify the stimuli with a linear SVM taking the multi-unit activity as input. But here the macaques did not—indeed, in all likelihood could not—classify the stimuli, because they had not been trained to do so. In this case, the SVM might have been making distinctions that the (untrained) macaques were not.

(3) Stimuli may differ along more than one dimension, both of which lead to differences in the BOLD signal. MVPA classification could rely on patterns relating to one

dimension, while participants use information relating to the other. That is, even when there is information in the BOLD signal that is theoretically accessible by (or that is tracking information accessible by) the participant, this may not be the information that is being used by the participant.

(4) The MVPA classifier may be using a kernel that is significantly different from what is implemented in the brain. As we saw, classifiers with different kernels trained on the very same data will extract different features, and thus come to different decisions about which features (and which elements of the input vectors) are most important.

(5) Since there will always be a set of highly informative voxels produced by the MVPA classifier, the existence of such a set won't tell us whether the relevant information in the brain is local and concentrated, local and distributed, non-local, or some combination of these.

The discussion also raises a much more general issue. As we noted at the outset, MVPA offers an exciting new way to investigate the operation of the brain, by looking at the predictive value of (typically widely) distributed patterns of activity. The problematic inferences generally come in the attempt to reduce such patterns to local features of brain activity. But if the best predictor of cognitive states is not the location of an activated region, but rather the patterns of cooperation and coactivation between them—as the success of MVPA might be said to indicate, and as has been argued for independent reasons (Anderson, 2008; Sporns, et al., 2004; Uttal, 2001)—then perhaps it is time to pay more heed to the patterns than to the partners. We are just beginning to develop the tools to make such an investigation fruitful and rigorous—including not only MVPA but other forms of statistical pattern analysis, machine learning, graph theory, etc.—and it seems a shame instead to use these tools in the service of localization projects for which they are ultimately ill-suited. New tools often come with the opportunity to re-consider the strengths of theoretical perspectives and paradigms, and these are offering a chance to look beyond localization, to what other perspectives on brain organization might have to offer.

References

Anderson, M. (2008). Circuit sharing and the implementation of intelligent systems. *Connection Science*, 20(4): 239-51.

Davatzikos, C., Ruparel, K., Fan, Y., Shen, D., Acharyya, M., Loughead, J., Gur, R. & Langleben, D.D. (2005). Classifying spatial patterns of brain activity with machine learning methods: Application to lie detection. *NeuroImage*, 28(3): 663-68.

Gross, J., Schmitz, F., Schnitzler, I., Kessler, K., Shapiro, K., Hommel, B & , Schnitzler, A. (2004). Modulation of long-range neural synchrony reflects temporal limitations of visual attention in humans. *Proceedings of the National Academy of Sciences-USA*, 101(35): 13050-13055

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: 2425-30.

Haynes, J-D. & Rees, G. (2006). Deconding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523-34.

Haynes, J.-D. & Rees, G. (2005). Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nature Neuroscience*, 8(5): 686-91.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C. & Passingham, R.E. (2007). Reading hidden intentions in the human brain. *Current Biology*, 17: 232-28.

Hung, C.P., Kreiman, G., Poggio, T. & DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, 310: 863-6.

Kamitani, Y. & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5): 679-85.

Kay, K.N., Naselaris, T., Prenger, R.J. & Gallant, J.L. (2008). Identifying natural images from human brain activity. *Nature*, 452: 352-6.

Lao, Z., Shen, D., Xue, Z., Karacali, B., Resnick, S. & Davatzikos, C. (2004). Morphological classification of brains via high-dimensional shape transformation and machine learning methods. *NeuroImage*, 21 (1): 46-57.

McDuff, S.G.R., Frankel, H.C. & Norman, K.A. (2009). Multivoxel pattern analysis reveals increased memory targeting and reduced use of retrieved details during single-agenda source monitoring. Journal of Neuroscience, 29(2):508-516.

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: 1191-5.

Norman, K.A., Polyn, S.M., Detre, G.J. & Haxby, J.V. (2006). Beyond mind-reading: Milti-voxel pattern analysis of fMRI data. *TRENDS in Cognitive Sciences*, 10(9): 424-30.

Pereira, F., Mitchell, T., Botvinick, M. M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45: S199-S209.

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

Preston, T.J., Li, S., Kourti, Z. & Welchman, A.E. (2008). Multivoxel pattern selectivity for perceptually relevant binocular disparities in the human brain. *The Journal of Neuroscience*, 28(44): 11315-27.

Shinkareva, S.V., Mason, R.A., Malave, V.L., Wang, W., Mitchell, T.M. & Just, M.A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLOS One*, 3(1): e1394. doi:10.137/journal.pone.0001394

Sporns, O., Chialvo, D., Kaiser, M. & Hilgetag, C.C. (2004). Organization, development and function of complex brain networks. *Trends in Cognitive Sciences*, 8: 418-425.

Uttal, W. (2001). *The New Phrenology*. Cambridge: MIT Press.

Varela, F., Lachaux J.P., Rodriguez E. & Martinerie J. (2001). The brainweb: phase synchronization and large-scale integration. *Nature Rev. Neuroscience*, 2(4): 229-39.

Wolpert, D. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7): 1341-1390.