Decoding the large-scale structure of brain function by classifying mental states across individuals

Russell A. Poldrack, Yaroslav Halchenko, and Stephen José Hanson

in press, *Psychological Science*

Abstract

Brain imaging research has largely focused on localizing patterns of activity related to specific mental processes, but recent work has shown that mental states can be identified from neuroimaging data using statistical classifiers. We asked whether this approach could be extended to predict the mental state of an individual using a statistical classifier trained on other individuals, and whether the information gained in doing so can provide new insights into how mental processes are organized in the brain. Using a variety of classifier techniques, we achieved cross-validated classification accuracy greater than 80% across individuals (where chance = 13%). Based on classifier sensitivity analysis we recovered a low dimensional representation common to all cognitive/perceptual tasks, and used an ontology of cognitive processes to determine the cognitive concepts most related to each dimension. These results reveal a small ordered set of large-scale networks that map cognitive processes across a highly diverse set of mental tasks, suggesting a novel way to characterize the neural basis of cognition.

1 Introduction

Neuroimaging has long been used to test specific hypotheses about brain-behavior relationships. However, it is increasingly being used to infer the engagement of specific mental processes. This is often done informally, by noting that previous studies have found an area to be engaged for a particular mental process and inferring that this process must be engaged whenever that region is found to be active. Such "reverse inference" has been shown to be problematic, particularly when regions are unselectively active in response to many different cognitive manipulations (Poldrack, 2006). However, recent developments in the application of statistical classifiers to neuroimaging data provide the means to directly test how accurately mental processes can be classified (e.g. O'Toole et al., 2007; Haynes & Rees, 2006). In this paper, we first examine how well classifiers can predict which of a set of eight cognitive tasks a person is engaged in, based on patterns from other individuals, and we show that it is possible with high accuracy. Second, we examine the dimensional representation of brain activity that underlies this classification accuracy, and find that

the differences between these tasks can be described in terms of a small set of underlying dimensions. Finally, we examine how these distributed neural dimensions map onto a the component cognitive processes that are engaged by these diverse tasks, by mapping each task onto an ontology of mental processes. The results demonstrate how neuroimaging can in principle be used to map brain activity onto cognitive processes rather than tasks.

There is increasing interest in using the tools of machine learning to identify signals that can allow "brain-reading", or prediction of mental states or behavior directly from neuroimaging data (O'Toole et al., 2007). These classifiers are first trained on "in-sample" fMRI data and then used to make predictions about "out-of-sample" patterns within the same experimental task. Such methods typically show perfect classification on the in-sample training data, whereas for out-of-sample cases classification ranges between 70-90% correct, which is quite exceptional given the noisiness of the fMRI signal. For example, it is now well-established that fMRI data from the ventral temporal cortex provide sufficient information to accurately predict what class of object (e.g., faces, houses etc) a person is viewing (Hanson & Halchenko, 2008; Hanson et al., 2004; Haxby et al., 2001). In other kinds of tasks one can tell whether the subject is conscious of visual information (Haynes & Rees, 2005) or even "read out" the intention of the subject prior to their behavioral response (Haynes et al., 2007). Thus, it is possible to reliably identify mental states for a given individual within a specific task, using training data from the same individual. There have also been some demonstrations of accurate classification across individuals (Mourao-Miranda et al., 2005; Shinkareva et al., 2008), which have distinguished between a relatively limited set of classes.

2 Classifying tasks across individuals

To investigate classification of tasks across individuals, we combined data across eight fMRI studies performed in the first author's laboratory, including a total of 130 participants (Table 1) performing a wide range of mental tasks. The data were collected on the same 3T MRI scanner with consistent acquisition parameters and analyzed using the same data analysis procedures (see Materials and Methods). This large-scale and methodologically consistent dataset allowed us to ask the following question: Is it possible to tell which mental task a person is engaged in, solely on the basis of fMRI data?

For each subject, a single statistical parametric (Z) map was obtained for a contrast comparing the task condition to a baseline condition. These Z-statistic data were submitted to classification using a multi-class linear support vector machine (SVM) (Boser et al., 1992; Hsu & Lin, 2002; Schölkopf & Smola, 2002). SVM provides a computationally tractable means to classify extremely high-dimensional data (in this case, over 200,000 features). Accuracy at predicting which task a subject was performing was computed using leave-one-out crossvalidation; the classifier was trained on all subjects except for one, and then tested on that left-out individual, which was repeated for each individual (see Supplementary

Figure 1 for overview of analysis).

Using data from the intersection set of in-mask voxels (which included 214K voxels across the entire brain), 80% classification accuracy was achieved for subjects in the out-of-sample generalization set. Similar levels of accuracy were obtained using neural network classifiers; see Supplementary Table 1 for an exhaustive list of classifier results with this dataset. Table 2 presents the confusion matrix for this analysis, which shows that all tasks were classified with relatively high accuracy, though there was some variability between tasks. Statistical significance of classification accuracy versus chance was assessed using a randomization approach to obtain an empirical null distribution; mean chance accuracy was 13.3%, and according to this analysis accuracy greater than 18.5% is significantly greater than chance at $p < .05$. When the classifier was trained on one run and then generalized to a second run for the same individuals (for the six tasks that had multiple runs), 90% classification accuracy was achieved for the second run (see Supplementary Table 2a for the confusion matrix for this analysis). Thus, the generalizability of task classification across individuals was nearly as high as the accuracy of generalization across runs within individuals. It is difficult to compare these accuracy levels to previous studies of within-subject classification, since those studies have often used much smaller image sets or single images to perform classification, whereas we used summary statistic images in the present analysis.

If this classification ability relies upon general cognitive features of the tasks, then it should be possible to classify individuals performing different versions of the same tasks on which the classifier was trained. This was examined using data from two additional studies, which used similar (but not identical) versions of two of the trained tasks (Tasks 2 and 8). These datasets were collected from subjects who had been included in the original training set, but performing different tasks (Tasks 4 and 7, respectively). When the classifier was trained excluding the data from these subjects from the original set, accurate classification was obtained for the new datasets (84%) (Supplementary Table 2b). This demonstrates that the classifier trained on the original data can accurately generalize to different studies using different versions of the same mental tasks. When the datasets from those same subjects (on different tasks) were included in the training set, accuracy was reduced but still high (66%; Supplementary Table 2c). This decrease in accuracy reflects the fact that the classifier was somewhat sensitive to individual characteristics of the training examples; in particular, when the same subjects performed task 8 in the test data but task 7 in the training data, they were often (7/20) misclassified as performing task 7 in the test. No such misclassification occurred for task 2. These results demonstrate that the classifier is more sensitive to task-relevant information than to idiosyncratic activation patterns of individual subjects, but does retain some sensitivity to task-independent patterns within individuals.

3 Localizing the sources of classification accuracy

In order to determine the anatomical sources of the information that drove classification accuracy, we used three independent sensitivity methods to identify anatomical areas that were potentially diagnostic for the SVM classifier performance. First we looked at the predictive power within each of a set of independent anatomical regions of interest (ROIs) using an SVM classifier within each ROI. Second we applied a localized SVM centered at every voxel with a fixed 4mm radial ROI (Hanson & Gluck, 1990; Kriegeskorte et al., 2006; Poggio & Girosi, 1990). These analyses tended to agree and indicated that many regions throughout the cortex provided information that allowed some degree of accurate prediction (30-50%) of cognitive states (Figure 1; Supplementary Table 3). Substantial predictability was present in sensory cortices; given the fact that the different studies varied substantially in their visual stimulus characteristics and the presence of auditory stimuli, this was not surprising, and it suggested that the classification does not necessarily reflect the higher-order cognitive aspects of the tasks. However, a number of regions in the prefrontal cortex also showed substantial predictability, including the premotor and anterior cingulate cortices. When the local kernel extent was expanded to 8 mm, it was striking that one of the only regions not providing substantial classification accuracy was in the dorsolateral prefrontal cortices (Supplementary Figure 2). This could either reflect the fact that those regions are equally engaged across mental tasks (Duncan & Owen, 2000), that substantial individual variability renders them non-predictive across subjects, or that the radial ROIs are too small to detect relevant inter-regional interactions (since generalization accuracy was so much lower than full brain SVM).

The foregoing analyses demonstrated which regions provide information that might be useful for task classification, but do not tell us which regions are diagnostic for particular tasks. In order to determine this, we performed an analysis that measured the diagnosticity of each voxel; that is, how predictive it is of a specific task. This was achieved by determining from the whole-brain dataset which voxels have the greatest effect on the classifier error, which is equivalent to the effect of removing them individually to see which ones have the greatest effect on the classification (see Supplementary Methods for details). The results of this analysis (Supplementary Figure 3) showed that the set of voxels identified as diagnostic for each task is heavily overlapping but much smaller than the set of voxels identified as active in a standard GLM analysis. These analyses have different goals: for the GLM the goal is voxel detection while for classifiers it the identification of voxels that are diagnostic for tasks, which has the potential for higher specificity (Hanson & Halchenko, 2008).

4 Relating neural and psychological similarity spaces

The ability to accurately classify mental tasks based on brain imaging data requires that brain patterns from the same task are more similar in the high-dimensional voxel space than

patterns from different tasks. We next set out to examine how this neural similarity space is related to the psychological similarity space of the specific tasks used in the dataset, by visualizing the location of each individual subject's brain data in a brain activation space with greatly reduced dimensionality.

So far we have shown that SVM provides strong evidence for a valid classification function based on whole-brain data (200,000 features). Although the identified support vectors are diagnostic of the boundaries of the decision surface, they, by design, cannot at the same time provide probablistic information about the underlying feature space or the class-conditional probability distributions. On the other hand, based on the impressive performance of the SVM classifier, it is likely that a conservatively chosen feature selection/extraction set could be used to approximate the classification function identified and at the same time allow visualization of the feature space and information on the class-conditional probability distribution that SVM does not provide. One candidate for this classification approximation is a related learning method which also has the ability to both select exemplar patterns like SVM and find prototypes based on interesting projections in the feature space: Neural networks (which are additive sigmoidal kernel function approximators). Unsupervised dimensionality reduction methods, such as principal or independent components analysis, can also identify lower-dimension projections of fMRI data, but are not constrained at the same time by the particular classification problem, as is a neural network. However, one limitation of neural networks is that depending on the complexity of the decision surface, they will be unable to process more than about 10,000 features due to memory and computational constraints, and thus in the present case, will require feature selection/extraction.

In our study, feature selection was performed by computing the relative entropy over all brains and tasks in each voxel; in comparison to feature selection using variance within features, this provided a more sensitive measure of voxel sensitivity to brain and task variation. This measure resulted in 2173 voxels at a $p < .01$ threshold; the selected voxels were sparsely distributed throughout the brain (Supplementary Figure 4). These voxels were used to train a sigmoidal neural network (with a varied number of hidden units), which was able to produce similar classification accuracy to the SVM analyses using whole brain data (71% at 6 hidden units with little improvement to higher values; see Supplementary Table 1). To further confirm the validity of entropy-based feature selection, we used these voxels with an SVM classifier, which achieved reduced but similar accuracy (72%) to the original analysis using 200K voxels, thus producing a compression factor of 100 to 1.

The NN classifier was trained on all exemplars and was able to achieve high classification accuracy and simultaneously project the data into a lower-dimensional subspace (6 hidden units). To further characterize this space we first performed an agglomerative hierarchical cluster analysis (see Supplementary Methods for more details) in the 6-dimensional space derived from the hidden units of the network (Figure 2). It is clear from this cluster space that the neural activity patterns not only preserve task differences, but also reflect the similarity structure of the mental tasks. For example, the three tasks that require linguistic

processing (READ, RHYME, SEM) are adjacent, as are the two tasks that require attention to auditory stimuli (WM and INH). In order to characterize the derived dimensionality of the task space, we constructed a visualization of the dimensions per task using star plots, which code the contribution of each dimension to each task. These plots (in Figure 3) reveal two important results: (1) brain function across these diverse tasks are ordered on a small set of unknown functional features (3-6) which suggests a recruiting of similar brain networks over all tasks, and (2) the dimensions are ordered from sensory/perceptual (auditory, visual) to more complex function (decision making, categorization, language supporting functions, etc).

5 Mapping neural and mental spaces using ontologies

To characterize the neural dimensions obtained from the NN analysis in terms of basic mental processes, we coded each task according to the presence or absence of a number of such processes (as depicted in Supplementary Figure 5). These were then projected onto all 6 functional dimensions, in order to characterize which cognitive processes were most strongly related to each neural processing dimension (Figure 4). Dimension 1 loads most heavily on the cognitive concept "audition", and the neural pattern associated with this dimension is primarily centered on the superior temporal gyrus (i.e., auditory cortex) and precentral gyus. Dimension 2 orders tasks related to language (SEM, READ, CAT), and factors a bilateral network including Broca's and Wernicke's areas and their right hemisphere homologs as well as parahippocampal gyrus, medial parietal, and medial prefrontal regions. Dimensions 3 and 4 select tasks related to learning and memory and decision making, and are associated with highly overlapping neural structures including thalamus, striatum, amygdala, medial prefrontal cortex, and parietal cortex. Dimension 5 is mostly strongly associated with memory and vision, and is tightly focused on the dorsomedial thalamus and dorsal striatum. Finally, dimension 6 shows a pattern of loading that is very similar to activation observed in studies of response inhibition (right IFG, basal ganglia, and medial prefrontal cortex; e.g.,(Aron & Poldrack, 2006)), and the mental concept most associated with this pattern is indeed "response inhibition." In each case is it also clear that the pattern is not specific to those concepts, as seen in the relatively strong loading of other concepts as well. These data suggest that the function of these networks is only partially captured by these specific terms; however, the relatively small number of tasks certainly biases the particular associations that were observed.

6 Discussion

The results presented here show that fMRI data contain sufficient information to accurately determine an individual's mental state (as imposed by a mental task), using classifiers trained on data from other individuals. This generalizes previous results, which

have demonstrated significant classification within individuals (Haxby et al., 2001; Hanson et al., 2004; Haynes & Rees, 2005) and between individuals (Mourao-Miranda et al., 2005; Shinkareva et al., 2008), and provides a proof of concept that fMRI could be used to detect a relatively broad range of cognitive states in previously untested individuals. The results also demonstrate how large-scale neuroimaging datasets could be used to test theories about the organization of cognition. Whereas previous imaging studies have nearly always focused on determining the neural basis of a particular cognitive process using specific task comparisons to isolate that process, the approach outlined here shows how data from multiple tasks can be used to examine the neural basis of cognitive processes that span across tasks. To the degree that cognitive theories make predictions regarding the similarity structure of different tasks, these theories could be tested using neuroimaging data.

6.1 Relation to standard neuroimaging analyses

The standard mass-univariate approach to fMRI analysis asks the question: What regions are significantly active when a specific mental process is manipulated? Examination of the statistical maps associated with each of the eight tasks in the present study shows substantial overlap between different sets of tasks, as well as some distinctive features. The classifier analysis used in the present study asks a very different question: What task is the subject engaged in, given the observed pattern of brain activity? It is common in the neuroimaging literature to use univariate maps to infer the engagement of specific mental processes from univariate analyses (i.e. reverse inference), but without using a classification technique it is impossible to determine the accuracy of such inferences. More directly, our diagnosticity analysis shows that the set of voxels that are activated by a task is much larger than the set of voxels whose activity is diagnostic for engagement of a particular task. This suggests that informal reverse inference is almost certain to be highly inaccurate in task domains like those examined here. These results suggest that approaches like the one used here are necessary in order to make strong inferences about cognitive processes from neuroimaging data.

6.2 Ontologies for cognitive neuroscience

The use of formal ontologies (Bard & Rhee, 2004) (such as the Gene Ontology; (Ashburner et al., 2000)) has become prevalent in many areas of bioscience as a means to formalize the relation between structure and function. The results presented here, in which a simple ontology of mental processes was mapped onto dimensions of neural activity, provides a proof of concept for the utility of cognitive ontologies as a means to better understand how mental processes map to neural processes (cf. (Price & Friston, 2005; Bilder et al., 2009)). It is not possible at present to determine how well the present methods could scale to a complete ontology of mental states. Such analyses would require large databases of statis-

tical results from individual subjects, which are not currently available; the present results suggest that such databases could be of significant utility to the cognitive neuroscience community. In addition, it is likely that differences in acquisition parameters will have significant effects on the ability to classify and cluster neuroimaging data across studies. The development of neuroimaging consortia using consistent data acquisition parameters across centers could help reduce this problem.

Acknowledgments

This research was supported by the US Office of Naval Research (RP), James S. McDonnell Foundation and National Science Foundation (SJH), and National Institutes of Health (PL1MH083271; R. Bilder, PI). Collection of the datasets included in the analyses here was supported by grants to RP from the National Science Foundation, Whitehall Foundation, and James S. McDonnell Foundation. The authors would like to thank the researchers whose data were included in the analyses: Adam Aron, Fabienne Cazalis, Karin Foerde, Elena Stover, Sabrina Tom, and Gui Xue. Thanks to Alan Castel, Marta Garrido, Dara Ghahremani, Keith Holyoak, Michael Todd, and Ed Vul for helpful comments on an earlier draft.

References

Aron, A. R., & Poldrack, R. A. (2006). Cortical and subcortical contributions to stop signal response inhibition: role of the subthalamic nucleus. *J Neurosci*, *26*(9), 2424–2433.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, *25*(1), 25–29.

Bard, J. B. L., & Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nat Rev Genet*, *5*(3), 213–222.

Bilder, R. M., Sabb, F. W., Parker, D. S., Kalar, D., Chu, W. W., Fox, J., & Poldrack, R. (2009). Cognitive ontologies for neuropsychiatric phenomics research. *Cognitive Neuropsychiatry*, *in press*.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifier. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, (pp. 144–152). ACM Press.

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci*, *23*(10), 475–483.

Hanson, S. J., & Gluck, M. A. (1990). Spherical units as dynamic consequential regions. In R. Lippmann, J. Moody, & D. Touretzky (Eds.) *Advances in Neural Information Processing Systems*, vol. 3, (pp. 656–664).

Hanson, S. J., & Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. *Neural Comput*, *20*(2), 486–503.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, *23*(1), 156–166.

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.

Haynes, J.-D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Curr Biol*, *15*(14), 1301–1307.

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, *7*(7), 523–534.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading hidden intentions in the human brain. *Curr Biol*, *17*(4), 323–328.

Hsu, C.-W., & Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Trans Neural Netw*, *13*(2), 415–425.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proc Natl Acad Sci U S A*, *103*(10), 3863–3868.

Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional mri data. *Neuroimage*, *28*(4), 980–995.

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

Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*(4945), 978–982.

Poldrack, R. A. (2006). Can cognitive processes be inferred from neuroimaging data? *Trends Cogn Sci*, *10*(2), 59–63.

Price, C., & Friston, K. (2005). Functional ontologies for cognition: The systematic definition of structure and function. *Cognitive Neuropsychology*, *22*(3-4), 262–275. URL `http://dx.doi.org/10.1080/02643290442000095`

Schölkopf, B., & Smola, A. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

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.

Van Essen, D. C. (2005). A population-average, landmark- and surface-based (pals) atlas of human cerebral cortex. *Neuroimage*, *28*(3), 635–662.

Table 1: Datasets included in analysis (* - only one scanning run was available for this task)

Task	Task name	Task code	# of subjects	Design type
1	Risky decision making (Balloon analog risk task)	RISK	16	Event-related
2	Probabilistic classification (no feedback)	CLS	20	Event-related
3	Rhyme judgments on pseudowords	RHYME	13	Blocked*
4	Working memory (tone counting)	WM	17	Blocked*
5	50/50 gain-loss gamble decisions	DEC	16	Event-related
6	Living/nonliving decision on mirror-reversed words	SEM	14	Event-related
7	Reading pseudowords aloud	READ	19	Event-related
8	Response inhibition (stop-signal task)	INH	15	Event-related

Table 2: Confusion matrix for SVM analyses: Generalization across subjects using leave-one-out crossvalidation

True	RISK	CLS	RHYME	WM	DEC	SEM	READ	INH	Accuracy
RISK	14	0	0	0	1	0	1	0	87.50%
CLS	0	18	0	0	0	0	1	1	90.00%
RHYME	1	2	8	0	1	1	0	0	61.54%
WM	0	0	0	14	0	0	0	3	82.35%
DEC	0	3	0	0	11	2	0	0	68.75%
SEM	0	2	0	0	1	11	0	0	78.57%
READ	0	1	0	0	0	0	17	1	89.47%
INH	0	0	0	1	0	0	3	11	73.33%

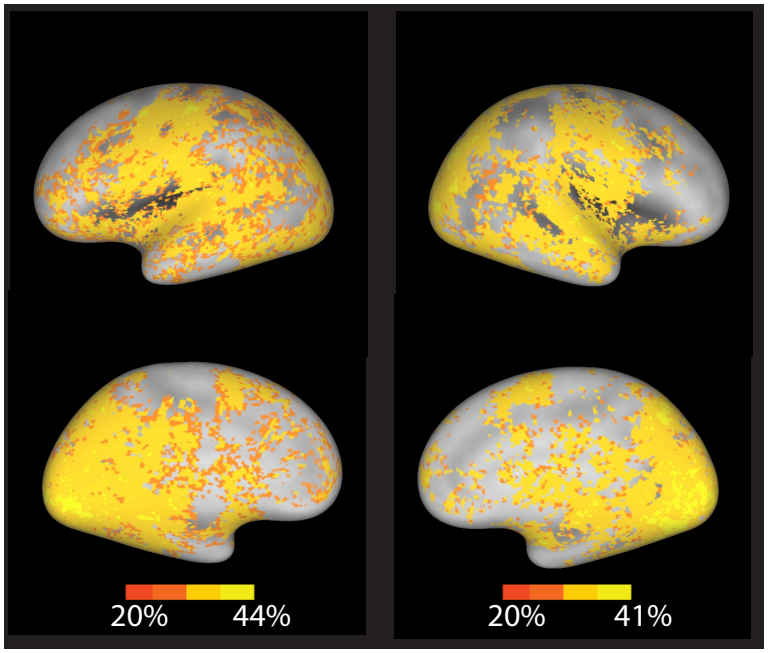

Figure 1: Localized accuracy of "reverse inference" across the eight cognitive tasks, identified using a searchlight technique (localized SVM performed across a 4mm radial ROI centered at each voxel). Results are overlaid on a population-average surface using CARET software (Van Essen, 2005).

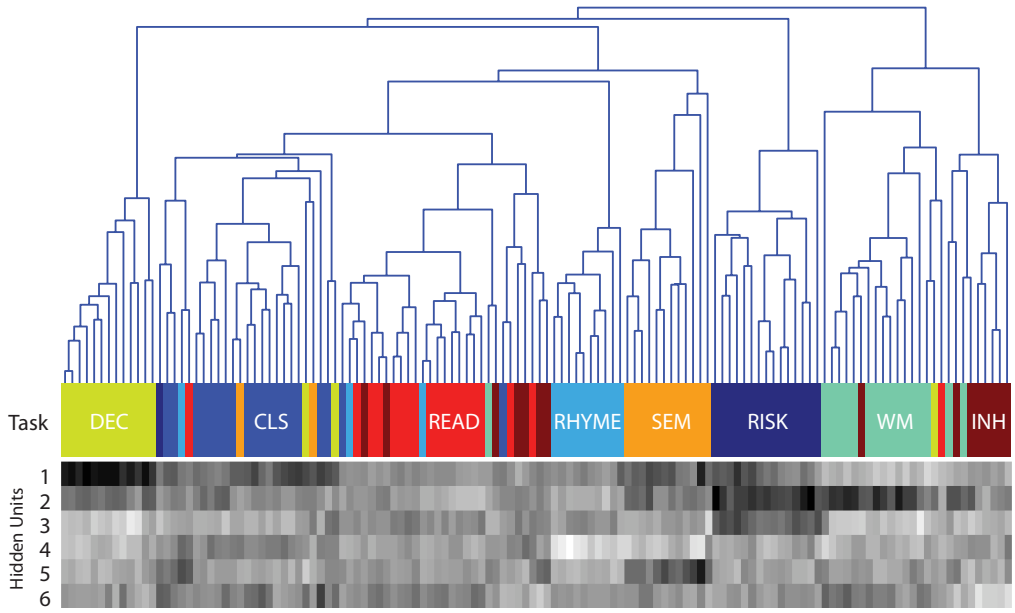

Figure 2: Visualization of the reduced dimension dataset. The cluster tree is based on a hierarchical clustering solution using the 6-dimension data obtained from the hidden unit activity in a neural network when presented with each individual's data. The data on each component for each subject are presented in grayscale form in the lower panel (brighter values represent higher values on each component). Each final branch in the tree and column in the heatmap represents a single individual. Task labels are presented in Table 1.

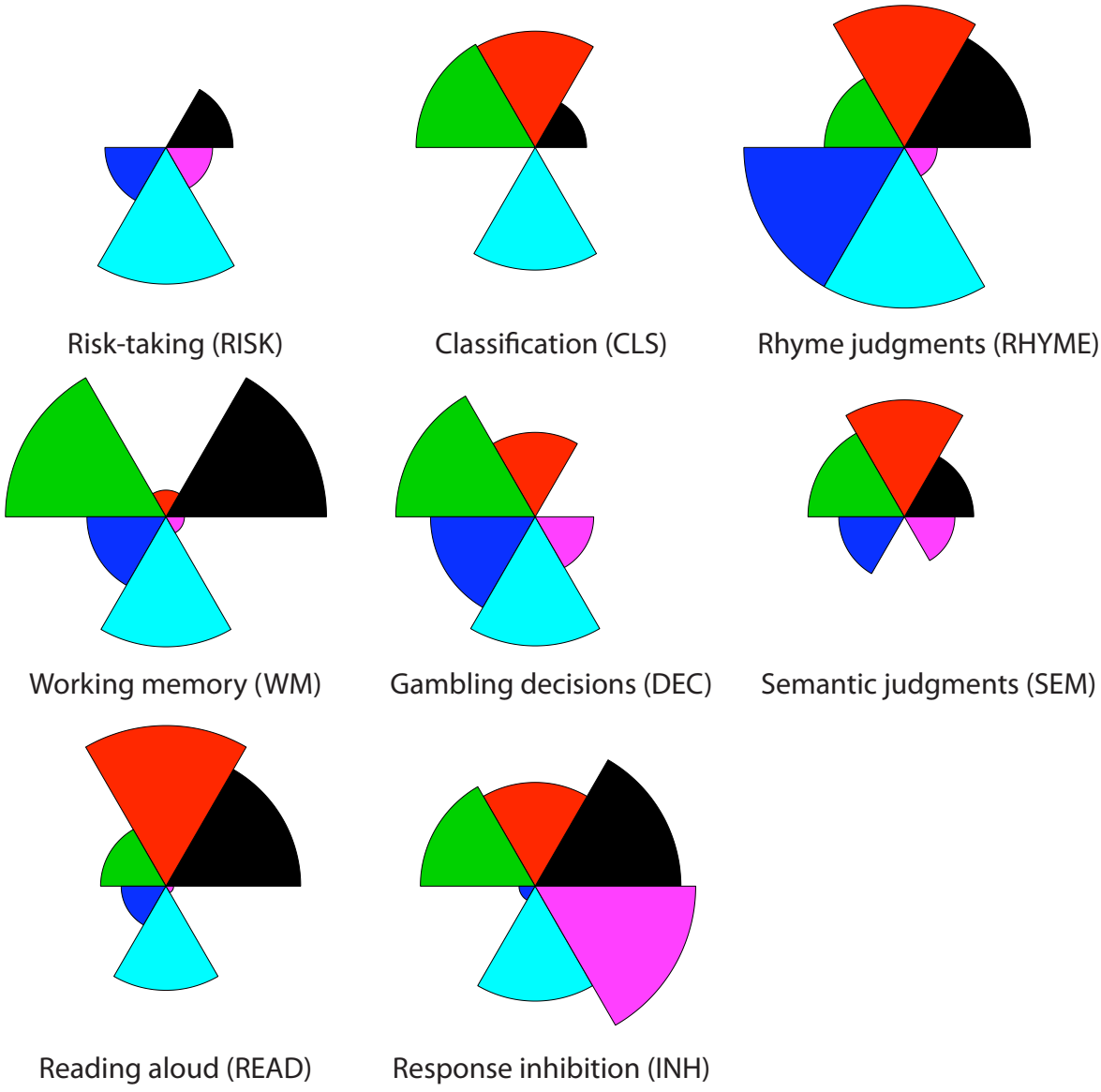

Figure 3: Dimensional loadings per task shown in a star plot display where the coefficient loading for each specific dimension in each task is coded by the relative size of its wedge in the plot.

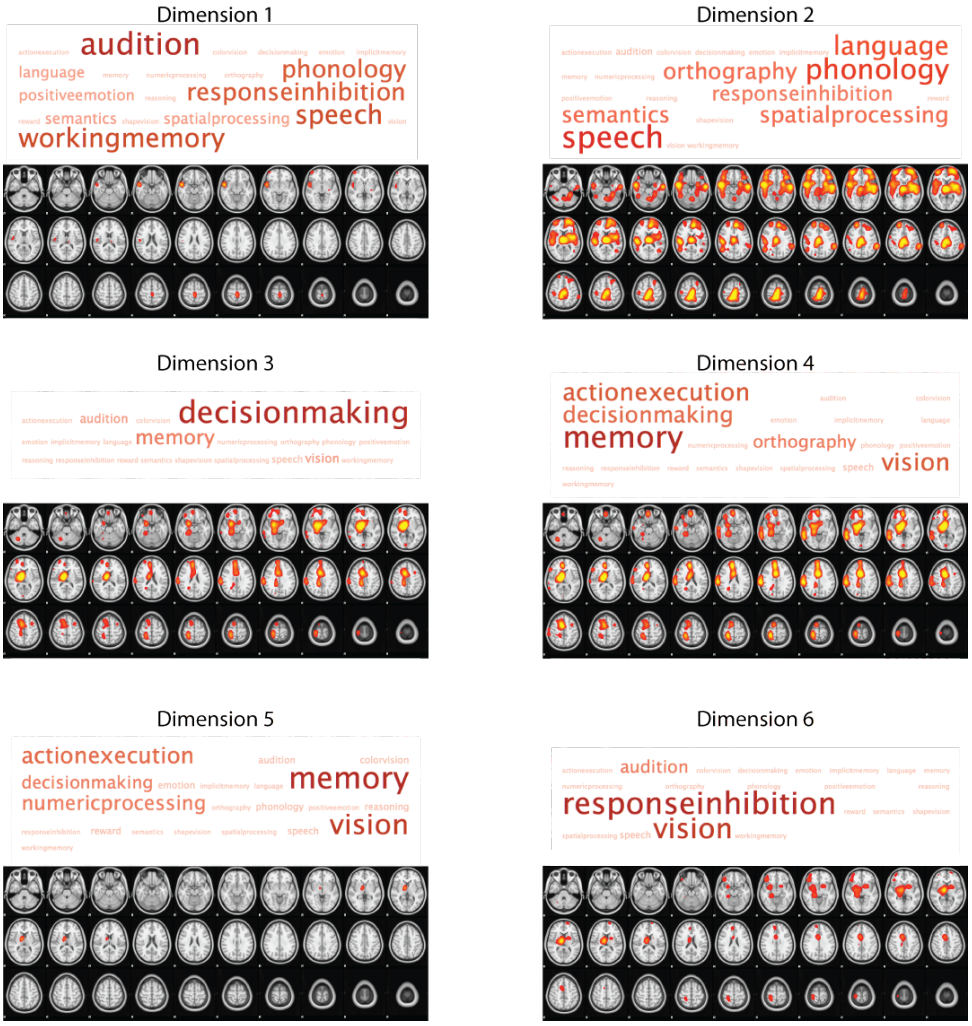

Figure 4: Visualization of the loading of mental concepts onto brain systems. The slice images show regions that exhibited positive (red-yellow) loading on the particular dimension; the original voxel loading maps were smoothed in order to create these images. The tag clouds represent the strength of association between the cognitive concepts and dimensions via the size of the text; larger words are more strongly associated with the dimension direction of the same color.