# Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,[1]* Svetlana V. Shinkareva,[2] Andrew Carlson,[1] Kai-Min Chang,[3,4] Vicente L. Malave,[5] Robert A. Mason,[3] Marcel Adam Just[3]

The question of how the human brain represents conceptual knowledge has been debated in many scientific fields. Brain imaging studies have shown that different spatial patterns of neural activation are associated with thinking about different semantic categories of pictures and words (for example, tools, buildings, and animals). We present a computational model that predicts the functional magnetic resonance imaging (fMRI) neural activation associated with words for which fMRI data are not yet available. This model is trained with a combination of data from a trillion-word text corpus and observed fMRI data associated with viewing several dozen concrete nouns. Once trained, the model predicts fMRI activation for thousands of other concrete nouns in the text corpus, with highly significant accuracies over the 60 nouns for which we currently have fMRI data.

The question of how the human brain represents and organizes conceptual knowledge has been studied by many scientific communities. Neuroscientists using brain imaging studies (*1*–*9*) have shown that distinct spatial patterns of fMRI activity are associated with viewing pictures of certain semantic categories, including tools, buildings, and animals. Linguists have characterized different semantic roles associated with individual verbs, as well as the types of nouns that can fill those semantic roles [e.g., VerbNet (*10*) and WordNet (*11*, *12*)]. Computational linguists have analyzed the statistics of very large text corpora and have demonstrated that a word's meaning is captured to some extent by the distribution of words and phrases with which it commonly co-occurs (*13*–*17*). Psychologists have studied word meaning through feature-norming studies (*18*) in which participants are asked to list the features they associate with various words, revealing a consistent set of core features across individuals and suggesting a possible grouping of features by sensory-motor modalities. Researchers studying semantic effects of brain damage have found deficits that are specific to given semantic categories (such as animals) (*19*–*21*).

This variety of experimental results has led to competing theories of how the brain encodes meanings of words and knowledge of objects, including theories that meanings are encoded in sensory-motor cortical areas (*22*, *23*) and theories that they are instead organized by semantic categories such as living and nonliving objects (*18*, *24*). Although these competing theories sometimes lead to different predictions (e.g., of which naming disabilities will co-occur in brain-damaged patients), they are primarily descriptive theories that make no attempt to predict the specific brain activation that will be produced when a human subject reads a particular word or views a drawing of a particular object.

We present a computational model that makes directly testable predictions of the fMRI activity associated with thinking about arbitrary concrete nouns, including many nouns for which no fMRI data are currently available. The theory underlying this computational model is that the neural basis of the semantic representation of concrete nouns is related to the distributional properties of those words in a broadly based corpus of the language. We describe experiments training competing computational models based on different assumptions regarding the underlying features that are used in the brain for encoding of meaning of concrete objects. We present experimental evidence showing that the best of these models predicts fMRI neural activity well enough that it can successfully match words it has not yet encountered to their previously unseen fMRI images, with accuracies far above those expected by chance. These results establish a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings.

**Approach.** We use a trainable computational model that predicts the neural activation for any given stimulus word $w$ using a two-step process, illustrated in Fig. 1. Given an arbitrary stimulus word $w$, the first step encodes the meaning of $w$ as a vector of intermediate semantic features computed from the occurrences of stimulus word $w$ within a very large text corpus (*25*) that captures the typical use of words in English text. For example, one intermediate semantic feature might be the frequency with which $w$ co-occurs with the verb "hear." The second step predicts the neural fMRI activation at every voxel location in the brain, as a weighted sum of neural activations contributed by each of the intermediate semantic features. More precisely, the predicted activation $y_v$ at voxel $v$ in the brain for word $w$ is given by

$$y_v = \sum_{i=1}^{n} c_{vi} f_i(w) \qquad (1)$$

where $f_i(w)$ is the value of the $i$th intermediate semantic feature for word $w$, $n$ is the number of semantic features in the model, and $c_{vi}$ is a learned scalar parameter that specifies the degree to which the $i$th intermediate semantic feature activates voxel $v$. This equation can be interpreted as predicting the full fMRI image across all voxels for stimulus word $w$ as a weighted sum of images, one per semantic feature $f_i$. These semantic feature images, defined by the learned $c_{vi}$, constitute a basis set of component images that model the brain activation associated with different semantic components of the input stimulus words.

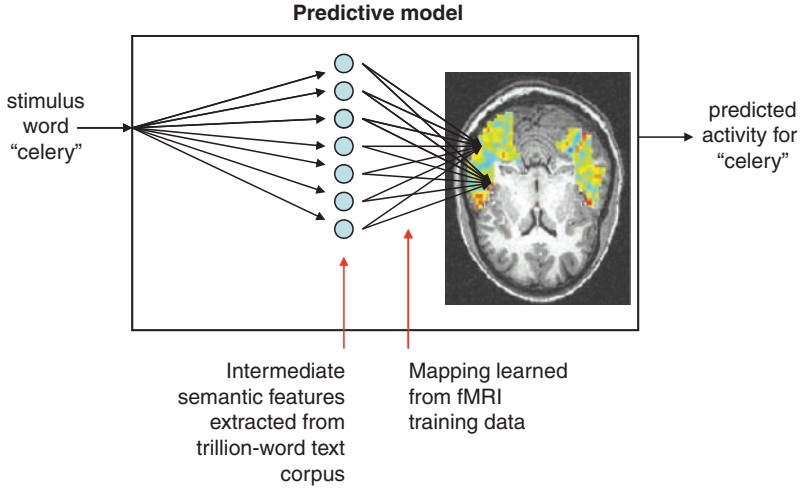



**Fig. 1.** Form of the model for predicting fMRI activation for arbitrary noun stimuli. fMRI activation is predicted in a two-step process. The first step encodes the meaning of the input stimulus word in terms of intermediate semantic features whose values are extracted from a large corpus of text exhibiting typical word use. The second step predicts the fMRI image as a linear combination of the fMRI signatures associated with each of these intermediate semantic features.

[1]Machine Learning Department, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Psychology, University of South Carolina, Columbia, SC 29208, USA. [3]Center for Cognitive Brain Imaging, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [5]Cognitive Science Department, University of California, San Diego, La Jolla, CA 92093, USA.

*To whom correspondence should be addressed. E-mail: Tom.Mitchell@cs.cmu.edu

To fully specify a model within this computational modeling framework, one must first define a set of intermediate semantic features $f_1(w)$ $f_2(w)$…$f_n(w)$ to be extracted from the text corpus. In this paper, each intermediate semantic feature is defined in terms of the co-occurrence statistics of the input stimulus word $w$ with a particular other word (e.g., "taste") or set of words (e.g., "taste," "tastes," or "tasted") within the text corpus. The model is trained by the application of multiple regression to these features $f_i(w)$ and the observed fMRI images, so as to obtain maximum-likelihood estimates for the model parameters $c_{vi}$ (*26*). Once trained, the computational model can be evaluated by giving it words outside the training set and comparing its predicted fMRI images for these words with observed fMRI data.

This computational modeling framework is based on two key theoretical assumptions. First, it assumes the semantic features that distinguish the meanings of arbitrary concrete nouns are reflected in the statistics of their use within a very large text corpus. This assumption is drawn from the field of computational linguistics, where statistical word distributions are frequently used to approximate the meaning of documents and words (*14*–*17*). Second, it assumes that the brain activity observed when thinking about any concrete noun can be derived as a weighted linear sum of contributions from each of its semantic features. Although the correctness of this linearity assumption is debatable, it is consistent with the widespread use of linear models in fMRI analysis (*27*) and with the assumption that fMRI activation often reflects a linear superposition of contributions from different sources. Our theoretical framework does not take a position on whether the neural activation encoding meaning is localized in particular cortical regions. Instead, it considers all cortical voxels and allows the training data to determine which locations are systematically modulated by which aspects of word meanings.

**Results.** We evaluated this computational model using fMRI data from nine healthy, college-age participants who viewed 60 different word-picture pairs presented six times each. Anatomically defined regions of interest were automatically labeled according to the methodology in (*28*). The 60 randomly ordered stimuli included five items from each of 12 semantic categories (animals, body parts, buildings, building parts, clothing, furniture, insects, kitchen items, tools, vegetables, vehicles, and other man-made items). A representative fMRI image for each stimulus was created by computing the mean fMRI response over its six presentations, and the mean of all 60 of these representative images was then subtracted from each [for details, see (*26*)].

To instantiate our modeling framework, we first chose a set of intermediate semantic features. To be effective, the intermediate semantic features must simultaneously encode the wide variety of semantic content of the input stimulus words and factor the observed fMRI activation into more primitive com-

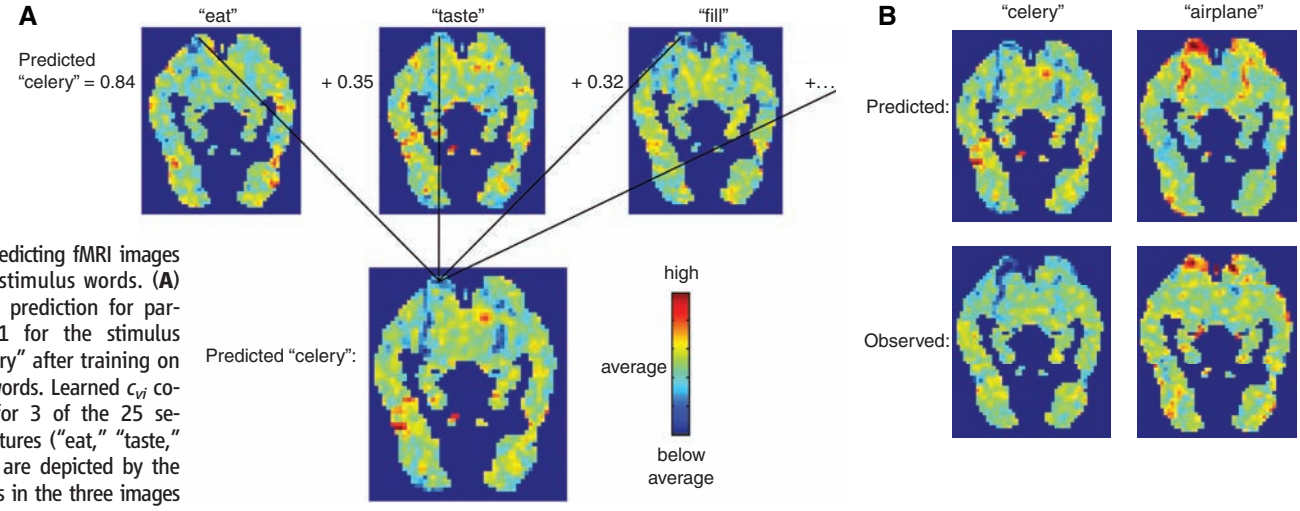
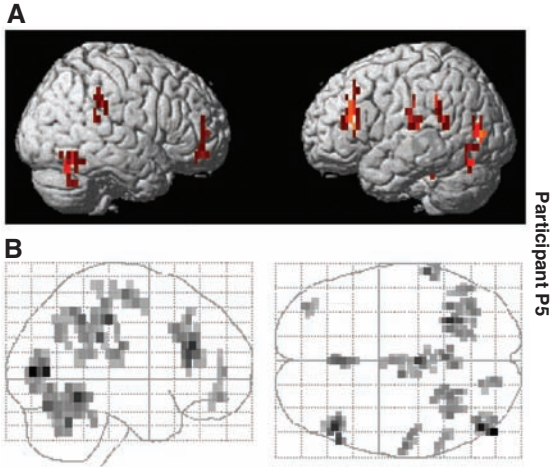



**Fig. 2.** Predicting fMRI images for given stimulus words. (**A**) Forming a prediction for participant P1 for the stimulus word "celery" after training on 58 other words. Learned $c_{vi}$ coefficients for 3 of the 25 semantic features ("eat," "taste," and "fill") are depicted by the voxel colors in the three images at the top of the panel. The co-occurrence value for each of these features for the stimulus word "celery" is shown to the left of their respective images [e.g., the value for "eat (celery)" is 0.84]. The predicted activation for the stimulus word [shown at the bottom of (A)] is a linear combination of the 25 semantic fMRI signatures, weighted by their co-occurrence values. This figure shows just one horizontal slice [$z$ = −12 mm in Montreal Neurological Institute (MNI) space] of the predicted three-dimensional image. (**B**) Predicted and observed fMRI images for "celery" and "airplane" after training that uses 58 other words. The two long red and blue vertical streaks near the top (posterior region) of the predicted and observed images are the left and right fusiform gyri.



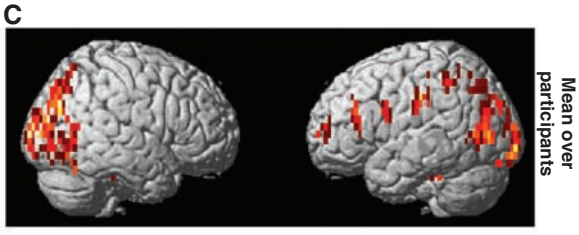




**Fig. 3.** Locations of most accurately predicted voxels. Surface (**A**) and glass brain (**B**) rendering of the correlation between predicted and actual voxel activations for words outside the training set for participant P5. These panels show clusters containing at least 10 contiguous voxels, each of whose predicted-actual correlation is at least 0.28. These voxel clusters are distributed throughout the cortex and located in the left and right occipital and parietal lobes; left and right fusiform, postcentral, and middle frontal gyri; left inferior frontal gyrus; medial frontal gyrus; and anterior cingulate. (**C**) Surface rendering of the predicted-actual correlation averaged over all nine participants. This panel represents clusters containing at least 10 contiguous voxels, each with average correlation of at least 0.14.

ponents that can be linearly recombined to successfully predict the fMRI activation for arbitrary new stimuli. Motivated by existing conjectures regarding the centrality of sensory-motor features in neural representations of objects (*18*, *29*), we designed a set of 25 semantic features defined by 25 verbs: "see," "hear," "listen," "taste," "smell," "eat," "touch," "rub," "lift," "manipulate," "run," "push," "fill," "move," "ride," "say," "fear," "open," "approach," "near," "enter," "drive," "wear," "break," and "clean." These verbs generally correspond to basic sensory and motor activities, actions performed on objects, and actions involving changes to spatial relationships. For each verb, the value of the corresponding intermediate semantic feature for a given input stimulus word $w$ is the normalized co-occurrence count of $w$ with any of three forms of the verb (e.g., "taste," "tastes," or "tasted") over the text corpus. One exception was made for the verb "see." Its past tense was omitted because "saw" is one of our 60 stimulus nouns. Normalization consists of scaling the vector of 25 feature values to unit length.

We trained a separate computational model for each of the nine participants, using this set of 25 semantic features. Each trained model was evaluated by means of a "leave-two-out" cross-validation approach, in which the model was repeatedly trained with only 58 of the 60 available word stimuli and associated fMRI images. Each trained model was tested by requiring that it first predict the fMRI images for the two "held-out" words and then match these correctly to their corresponding held-out fMRI images. The process of predicting the fMRI image for a held-out word is illustrated in Fig. 2A. The match between the two predicted and the two observed fMRI images was determined by which match had a higher cosine similarity, evaluated over the 500 image voxels with the most stable responses across training presentations (*26*). The expected accuracy in matching the left-out words to their left-out fMRI images is 0.50 if the model performs at chance levels. An accuracy of 0.62 or higher for a single model trained for a single participant was determined to be statistically significant ($P < 0.05$) relative to chance, based on the empirical distribution of accuracies for randomly generated null models (*26*). Similarly, observing an accuracy of 0.62 or higher for each of the nine independently trained participant-specific models would be statistically significant at $P < 10^{-11}$.

The cross-validated accuracies in matching two unseen word stimuli to their unseen fMRI images for models trained on participants P1 through P9 were 0.83, 0.76, 0.78, 0.72, 0.78, 0.85, 0.73, 0.68, and 0.82 (mean = 0.77). Thus, all nine participant-specific models exhibited accuracies significantly above chance levels. The models succeeded in distinguishing pairs of previously unseen words in over three-quarters of the 15,930 cross-validated test pairs across these nine participants. Accuracy across participants was strongly correlated ($r$ = –0.66) with estimated head motion (i.e., the less the participant's head motion, the greater the prediction accuracy), suggesting that the variation in accuracies across participants is explained at least in part by noise due to head motion.

Visual inspection of the predicted fMRI images produced by the trained models shows that these predicted images frequently capture substantial aspects of brain activation associated with stimulus words outside the training set. An example is shown in Fig. 2B, where the model was trained on 58 of the 60 stimuli for participant P1, omitting "celery" and "airplane." Although the predicted fMRI images for "celery" and "airplane" are not perfect, they capture substantial components of the activation actually observed for these two stimuli. A plot of similarities between all 60 predicted and observed fMRI images is provided in fig. S3.

The model's predictions are differentially accurate in different brain locations, presumably more accurate in those locations involved in encoding the semantics of the input stimuli. Figure 3 shows the model's "accuracy map," indicating the cortical regions where the model's predicted activations for held-out words best correlate with the observed activations, both for an individual participant (P5) and averaged over all nine participants. These highest-accuracy voxels are meaningfully distributed across the cortex, with the left hemisphere more strongly represented, appearing in left inferior temporal, fusiform, motor cortex, intraparietal sulcus, inferior frontal, orbital frontal, and the occipital cortex. This left hemisphere dominance is consistent with the generally held view that the left hemisphere plays a larger role than the right hemisphere in semantic representation. High-accuracy voxels also appear in both hemispheres in the occipital cortex, intraparietal sulcus, and some of the inferior temporal regions, all of which are also likely to be involved in visual object processing.

It is interesting to consider whether these trained computational models can extrapolate to make accurate predictions for words in new semantic categories beyond those in the training set. To test this, we retrained the models but this time we excluded from the training set all examples belonging to the same semantic category as either of the two held-out test words (e.g., when testing on "celery" versus "airplane," we removed every food and vehicle stimulus from the training set, training on only 50 words). In this case, the cross-validated prediction accuracies were 0.74, 0.69, 0.67, 0.69, 0.64,

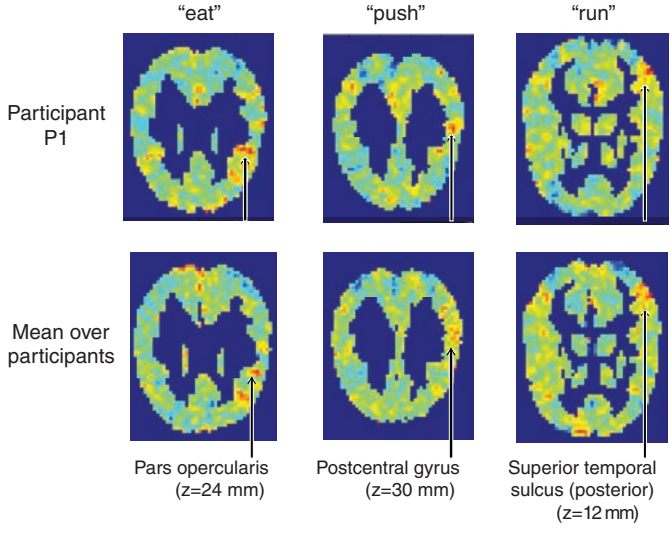



**Fig. 4.** Learned voxel activation signatures for 3 of the 25 semantic features, for participant P1 (top panels) and averaged over all nine participants (bottom panels). Just one horizontal $z$ slice is shown for each. The semantic feature associated with the verb "eat" predicts substantial activity in right pars opercularis, which is believed to be part of the gustatory cortex. The semantic feature associated with "push" activates the right postcentral gyrus, which is believed to be associated with premotor planning. The semantic feature for the verb "run" activates the posterior portion of the right superior temporal sulcus, which is believed to be associated with the perception of biological motion.

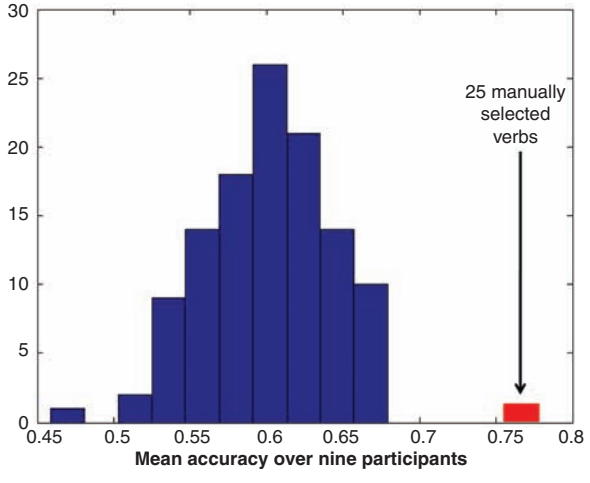



**Fig. 5.** Accuracies of models based on alternative intermediate semantic feature sets. The accuracy of computational models that use 115 different randomly selected sets of intermediate semantic features is shown in the blue histogram. Each feature set is based on 25 words chosen at random from the 5000 most frequent words, excluding the 500 most frequent words and the stimulus words. The accuracy of the feature set based on manually chosen sensory-motor verbs is shown in red. The accuracy of each feature set is the average accuracy obtained when it was used to train models for each of the nine participants.

0.78, 0.68, 0.64, and 0.78 (mean = 0.70). This ability of the model to extrapolate to words semantically distant from those on which it was trained suggests that the semantic features and their learned neural activation signatures of the model may span a diverse semantic space.

Given that the 60 stimuli are composed of five items in each of 12 semantic categories, it is also interesting to determine the degree to which the model can make accurate predictions even when the two held-out test words are from the same category, where the discrimination is likely to be more difficult (e.g., "celery" versus "corn"). These within-category prediction accuracies for the nine individuals were 0.61, 0.58, 0.58, 0.72, 0.58, 0.77, 0.58, 0.52, and 0.68 (mean = 0.62), indicating that although the model's accuracy is lower when it is differentiating between semantically more similar stimuli, on average its predictions nevertheless remain above chance levels.

In order to test the ability of the model to distinguish among an even more diverse range of words, we tested its ability to resolve among 1000 highly frequent words (the 1300 most frequent tokens in the text corpus, omitting the 300 most frequent). Specifically, we conducted a leave-one-out test in which the model was trained using 59 of the 60 available stimulus words. It was then given the fMRI image for the held-out word and a set of 1001 candidate words (the 1000 frequent tokens, plus the held-out word). It ranked these 1001 candidates by first predicting the fMRI image for each candidate and then sorting the 1001 candidates by the similarity between their predicted fMRI image and the fMRI image it was provided. The expected percentile rank of the correct word in this ranked list would be 0.50 if the model were operating at chance. The observed percentile ranks for the nine participants were 0.79, 0.71, 0.74, 0.67, 0.73, 0.77, 0.70, 0.63, and 0.76 (mean = 0.72), indicating that the model is to some degree applicable across a semantically diverse set of words [see (*26*) for details].

A second approach to evaluating our computation model, beyond quantitative measurements of its prediction accuracy, is to examine the learned basis set of fMRI signatures for the 25 verb-based signatures. These 25 signatures represent the model's learned decomposition of neural representations into their component semantic features and provide the basis for all of its predictions. The learned signatures for the semantic features "eat," "push," and "run" are shown in Fig. 4. Notice that each of these signatures predicts activation in multiple cortical regions.

Examining the semantic feature signatures in Fig. 4, one can see that the learned fMRI signature for the semantic feature "eat" predicts strong activation in opercular cortex (as indicated by the arrows in the left panels), which others have suggested is a component of gustatory cortex involved in the sense of taste (*30*). Also, the learned fMRI signature for "push" predicts substantial activation in the right postcentral gyrus, which is widely assumed to be involved in the planning of complex, coordinated movements (*31*). Furthermore, the learned signature for "run" predicts strong activation in the posterior portion of the right superior temporal lobe along the sulcus, which others have suggested is involved in perception of biological motion (*32*, *33*). To summarize, these learned signatures cause the model to predict that the neural activity representing a noun will exhibit activity in gustatory cortex to the degree that this noun co-occurs with the verb "eat," in motor areas to the degree that it co-occurs with "push," and in cortical regions related to body motion to the degree that it co-occurs with "run." Whereas the top row of Fig. 4 illustrates these learned signatures for participant P1, the bottom row shows the mean of the nine signatures learned independently for the nine participants. The similarity of the two rows of signatures demonstrates that these learned intermediate semantic feature signatures exhibit substantial commonalities across participants.

The learned signatures for several other verbs also exhibit interesting correspondences between the function of cortical regions in which they predict activation and that verb's meaning, though in some cases the correspondence holds for only a subset of the nine participants. For example, additional features for participant P1 include the signature for "touch," which predicts strong activation in somatosensory cortex (right postcentral gyrus), and the signature for "listen," which predicts activation in language-processing regions (left posterior superior temporal sulcus and left pars triangularis), though these trends are not common to all nine participants. The learned feature signatures for all 25 semantic features are provided at (*26*).

Given the success of this set of 25 intermediate semantic features motivated by the conjecture that the neural components corresponding to basic semantic properties are related to sensory-motor verbs, it is natural to ask how this set of intermediate semantic features compares with alternatives. To explore this, we trained and tested models based on randomly generated sets of semantic features, each defined by 25 randomly drawn words from the 5000 most frequent words in the text corpus, excluding the 60 stimulus words as well as the 500 most frequent words (which contain many function words and words without much specific semantic content, such as "the" and "have"). A total of 115 random feature sets was generated. For each feature set, models were trained for all nine participants, and the mean prediction accuracy over these nine models was measured. The distribution of resulting accuracies is shown in the blue histogram in Fig. 5. The mean accuracy over these 115 feature sets is 0.60, the SD is 0.041, and the minimum and maximum accuracies are 0.46 and 0.68, respectively. The random feature sets generating the highest and lowest accuracy are shown at (*26*). The fact that the mean accuracy is greater than 0.50 suggests that many feature sets capture some of the semantic content of the 60 stimulus words and some of the regularities in the corresponding brain activation. However, among these 115 feature sets, none came close to the 0.77 mean accuracy of our manually generated feature set (shown by the red bar in the histogram in Fig. 5). This result suggests the set of features defined by our sensory-motor verbs is somewhat distinctive in capturing regularities in the neural activation encoding the semantic content of words in the brain.

**Discussion.** The results reported here establish a direct, predictive relationship between the statistics of word co-occurrence in text and the neural activation associated with thinking about word meanings. Furthermore, the computational models trained to make these predictions provide insight into how the neural activity that represents objects can be decomposed into a basis set of neural activation patterns associated with different semantic components of the objects.

The success of the specific model, which uses 25 sensory-motor verbs (as compared with alternative models based on randomly sampled sets of 25 semantic features), lends credence to the conjecture that neural representations of concrete nouns are in part grounded in sensory-motor features. However, the learned signatures associated with the 25 intermediate semantic features also exhibit significant activation in brain areas not directly associated with sensory-motor function, including frontal regions. Thus, it appears that the basis set of features that underlie neural representations of concrete nouns involves much more than sensory-motor cortical regions.

Other recent work has suggested that the neural encodings that represent concrete objects are at least partly shared across individuals, based on evidence that it is possible to identify which of several items a person is viewing, through only their fMRI image and a classifier model trained from other people (*34*). The results reported here show that the learned basis set of semantic features also shares certain commonalities across individuals and may help determine more directly which factors of neural representations are similar and different across individuals.

Our approach is analogous in some ways to research that focuses on lower-level visual features of picture stimuli to analyze fMRI activation associated with viewing the picture (*9*, *35*, *36*) and to research that compares perceived similarities between object shapes to their similarities based on fMRI activation (*37*). Recent work (*36*) has shown that it is possible to predict aspects of fMRI activation in parts of visual cortex based on visual features of arbitrary scenes and to use this predicted activation to identify which of a set of candidate scenes an individual is viewing. Our work differs from these efforts, in that we focus on encodings of more abstract semantic concepts signified by words and predict brain-wide fMRI activations based on text corpus features that capture semantic aspects of the stimulus word, rather than visual features that capture perceptual aspects. Our work is also related to recent research that uses machine learning algorithms to train classifiers of mental states based on fMRI data (*38*, *39*), though it differs in that our models are capable of extrapolating to predict fMRI images for mental states not present in the training set.

This research represents a shift in the paradigm for studying neural representations in the brain,

moving from work that has cataloged the patterns of fMRI activity associated with specific categories of words and pictures to instead building computational models that predict the fMRI activity for arbitrary words (including thousands of words for which fMRI data are not yet available). This is a natural progression as the field moves from pretheoretical cataloging of data toward development of computational models and the beginnings of a theory of neural representations. Our computational models can be viewed as encoding a restricted form of predictive theory, one that answers such questions as "What is the predicted fMRI neural activity encoding word *w*?" and "What is the basis set of semantic features and corresponding components of neural activation that explain the neural activations encoding meanings of concrete nouns?" Although we remain far from a causal theory explaining how the brain synthesizes these representations from its sensory inputs, answers even to these questions promise to shed light on some of the key regularities underlying neural representations of meaning.

### References and Notes


1. J. V. Haxby *et al.*, *Science* **293**, 2425 (2001).
2. A. Ishai, L. G. Ungerleider, A. Martin, J. L. Schouten, J. V. Haxby, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9379 (1999).
3. N. Kanwisher, J. McDermott, M. M. Chun, *J. Neurosci.* **17**, 4302 (1997).
4. T. A. Carlson, P. Schrater, S. He, *J. Cogn. Neurosci.* **15**, 704 (2003).
5. D. D. Cox, R. L. Savoy, *Neuroimage* **19**, 261 (2003).
6. T. Mitchell *et al.*, *Mach. Learn.* **57**, 145 (2004).
7. S. J. Hanson, T. Matsuka, J. V. Haxby, *Neuroimage* **23**, 156 (2004).
8. S. M. Polyn, V. S. Natu, J. D. Cohen, K. A. Norman, *Science* **310**, 1963 (2005).
9. A. J. O'Toole, F. Jiang, H. Abdi, J. V. Haxby, *J. Cogn. Neurosci.* **17**, 580 (2005).
10. K. Kipper, A. Korhonen, N. Ryant, M. Palmer, *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genoa, Italy, 24 to 26 May 2006.
11. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, *Int. J. Lexicography* **3**, 235 (1990).
12. C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database* (Massachusetts Institute of Technology Press, Cambridge, MA, 1998).
13. K. W. Church, P. Hanks, *Comput. Linguist.* **16**, 22 (1990).
14. T. K. Landauer, S. T. Dumais, *Psychol. Rev.* **104**, 211 (1997).
15. D. Lin, S. Zhao, L. Qin, M. Zhou, *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, Acapulco, Mexico, August 2003 (Morgan Kaufmann, San Francisco, 2003), pp. 1492–1493.
16. D. M. Blei, A. Y. Ng, M. I. Jordan, *J. Mach. Learn. Res.* **3**, 993 (2003).
17. R. Snow, D. Jurafsky, A. Ng, *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, 17 to 21 July 2006.
18. G. S. Cree, K. McRae, *J. Exp. Psychol. Gen.* **132**, 163 (2003).
19. A. Caramazza, J. R. Shelton, *J. Cogn. Neurosci.* **10**, 1 (1998).
20. S. J. Crutch, E. K. Warrington, *Brain* **126**, 1821 (2003).
21. D. Samson, A. Pillon, *Brain Lang.* **91**, 252 (2004).
22. A. Martin, L. L. Chao, *Curr. Opin. Neurobiol.* **11**, 194 (2001).
23. R. F. Goldberg, C. A. Perfetti, W. Schneider, *J. Neurosci.* **26**, 4917 (2006).
24. B. Z. Mahon, A. Caramazza, in *The Encyclopedia of Language and Linguistics*, K. Brown, Ed. (Elsevier Science, Amsterdam, ed. 2, 2005).
25. T. Brants, A. Franz, www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 (Linguistic Data Consortium, Philadelphia, PA, 2006).
26. See Supporting Online Material.
27. K. J. Friston *et al.*, *Hum. Brain Mapp.* **2**, 189 (1995).
28. N. Tzourio-Mazoyer *et al.*, *Neuroimage* **15**, 273 (2002).
29. A. Martin, L. G. Ungerleider, J. V. Haxby, in *The New Cognitive Neurosciences*, M. S. Gazzainga, Ed. (Massachusetts Institute of Technology Press, Cambridge, MA, ed. 2, 2000), pp. 1023–1036.
30. B. Cerf, D. LeBihan, P. F. Van de Moortele, P. Mac Leod, A. Faurion, *Ann. N.Y. Acad. Sci.* **855**, 575 (1998).
31. K. A. Pelphrey, J. P. Morris, C. R. Michelich, T. Allison, G. McCarthy, *Cereb. Cortex* **15**, 1866 (2005).
32. L. M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, J. Belliveau, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 11656 (2001).
33. K. Sakai *et al.*, *Magn. Reson. Med.* **33**, 736 (1995).
34. S. V. Shinkareva *et al.*, *PLoS One* **3**, e1394 (2008).
35. D. R. Hardoon, J. Mourao-Miranda, M. Brammer, J. Shawe-Taylor, *Neuroimage* **37**, 1250 (2007).
36. K. N. Kay, T. Naselaris, R. J. Prenger, J. L. Gallant, *Nature* **452**, 352 (2008).
37. S. Edelman, K. Grill-Spector, T. Kushnir, R. Malach, *Psychobiology* **26**, 309 (1998).
38. J. D. Haynes, G. Rees, *Nat. Rev. Neurosci.* **7**, 523 (2006).
39. K. A. Norman, S. M. Polyn, G. J. Detre, J. V. Haxby, *Trends Cogn. Sci.* **10**, 424 (2006).


40. This research was funded by grants from the W. M. Keck Foundation, NSF, and by a Yahoo! Fellowship to A.C. We acknowledge Google for making available its data from the trillion-token text corpus. We thank W. Cohen for helpful suggestions regarding statistical significance tests.


### Supporting Online Material

www.sciencemag.org/cgi/content/full/320/5880/1191/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
References


12 November 2007; accepted 3 April 2008
10.1126/science.1152876






# The Cassiopeia A Supernova Was of Type IIb


**Oliver Krause,[1]* Stephan M. Birkmann,[1] Tomonori Usuda,[2] Takashi Hattori,[2] Miwa Goto,[1] George H. Rieke,[3] Karl A. Misselt[3]**



**Cassiopeia A is the youngest supernova remnant known in the Milky Way and a unique laboratory for supernova physics. We present an optical spectrum of the Cassiopeia A supernova near maximum brightness, obtained from observations of a scattered light echo more than three centuries after the direct light of the explosion swept past Earth. The spectrum shows that Cassiopeia A was a type IIb supernova and originated from the collapse of the helium core of a red supergiant that had lost most of its hydrogen envelope before exploding. Our finding concludes a long-standing debate on the Cassiopeia A progenitor and provides new insight into supernova physics by linking the properties of the explosion to the wealth of knowledge about its remnant.**


The supernova remnant Cassiopeia A is one of the most-studied objects in the sky, with observations from the longest radio waves to gamma rays. The remnant expansion rate indicates that the core of its progenitor star collapsed around the year 1681 ± 19, as viewed from Earth (*1*). Because of its youth and proximity of $3.4^{+0.3}_{-0.1}$ kpc (*2*), Cas A provides a unique opportunity to probe the death of a massive star and to test theoretical models of core-collapse supernovae. However, such tests are compromised because the Cas A supernova showed at most a faint optical display on Earth at the time of explosion. The lack of a definitive sighting means that there is almost no direct information about the type of the explosion, and the true nature of its progenitor star has been a puzzle since the discovery of the remnant (*3*).

The discovery of light echoes due both to scattering and to absorption and re-emission of the outgoing supernova flash (*4*, *5*) by the interstellar dust near the remnant raised the possibility of conducting a postmortem study of the last historic Galactic supernova by observing its scattered light. Similarly, the determination of a supernova spectral type long after its explosion using light echoes was recently demonstrated for an extragalactic supernova (*6*).

We have monitored infrared echoes around Cas A at a wavelength of 24 μm with use of the multiband imaging photometer (MIPS) instrument aboard the Spitzer Space Telescope (*4*). The results confirm that they arise from the flash emitted in the initial explosion of Cas A (*5*). An image taken on 20 August 2007 revealed a bright (flux density $F_{24\mu m} = 0.36 \pm 0.04$ Jy, 1 Jy $\equiv 10^{-26}$ W m$^{-2}$ Hz$^{-1}$) and mainly unresolved echo feature located 80 arc min northwest of Cas A (position angle 311° east of north). It had not been detected ($F_{24\mu m} < 2$ mJy; 5-σ) on two previous images of this region obtained on 2 October 2006 and 23 January 2007 (Fig. 1).

An image obtained on 7 January 2008 shows that the peak of the echo has dropped in surface brightness by a factor of 18 and shifted toward the west. Transient optical emission associated with the infrared echo was detected in an *R*-band image obtained at a wavelength of 6500 Å at the Calar Alto 2.2-m telescope on 6 October 2007


[1]Max-Planck-Institut für Astronomie, Königstuhl 17, 69117 Heidelberg, Germany. [2]National Astronomical Observatory of Japan, 650 North A'ohoku Place, Hilo, HI 96720, USA. [3]Steward Observatory, 933 North Cherry Avenue, Tucson, AZ 85721, USA.

*To whom correspondence should be addressed. E-mail: krause@mpia.de

# Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, Marcel Adam Just

*Science,* vol. 320

May 30, 2008.

## Supporting online material

## Table of Contents

# 1. Materials and Methods

## 1.1 fMRI Data collection and processing

Nine right-handed adults (5 female, age between 18 and 32) from the Carnegie Mellon University community participated in the fMRI study, and gave informed consent approved by the University of Pittsburgh and Carnegie Mellon Institutional Review Boards. Data from two additional participants exhibiting head motion of 2.2 mm and 3.0 mm were excluded.

The stimuli were line drawings and noun labels of 60 concrete objects from 12 semantic categories with 5 exemplars per category, as shown in Figure S1. Most of the line drawings were taken or adapted from the Snodgrass and Vanderwart set (*S1*) and others were added using a similar drawing style. The entire set of 60 stimulus items was presented six times, randomly permuting the sequence of the 60 items on each presentation. Each stimulus item was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate on an X displayed in the center of the screen. There were twelve additional presentations of a fixation X, 31s each, distributed across the session to provide a baseline measure.

When an exemplar was presented, the participants' task was to think about the properties of the object. To promote their consideration of a consistent set of properties across the 6 presentations, they were asked to generate a set of properties for each item prior to the scanning session (for example, for the item castle, the properties might be cold, knights, and stone). Each participant was free to choose any properties they wished, and there was no attempt to obtain consistency across participants in the choice of properties.

Functional images were acquired on a Siemens (Erlangen, Germany) Allegra 3.0T scanner at the Brain Imaging Research Center of Carnegie Mellon University and the University of Pittsburgh using a gradient echo EPI pulse sequence with TR = 1000 ms, TE = 30 ms and a 60° flip angle. Seventeen 5-mm thick oblique-axial slices were imaged with a gap of 1 mm between slices. The acquisition matrix was 64 x 64 with 3.125-mm x 3.125-mm x 5-mm voxels.

Initial data processing was performed using Statistical Parametric Mapping software (SPM2, Wellcome Department of Cognitive Neurology, London, UK). The data were corrected for slice timing, motion, and linear trend, and were temporally filtered using a 190s cutoff. The data were spatially normalized into MNI space and resampled to 3x3x6 $mm^3$ voxels. The percent signal change (PSC) relative to the fixation condition was computed at each voxel for each stimulus presentation. A single fMRI mean image was created for each of the 360 item presentations by taking the mean of the images collected 4s, 5s, 6s, and 7s after stimulus onset (to account for the delay in the hemodynamic response).

### 1.2 Text corpus data

The text corpus data was provided by Google Inc., and is available online at http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13. It consists of a set of n-grams (sequences of words and other text tokens) ranging from unigrams (single tokens) up to five-grams (sequences of five tokens), along with counts giving the number of times each n-gram appeared in a large corpus containing over a trillion total tokens. The corpus consisted of publicly available English text web pages. N-grams occurring fewer than 40 times were not provided. We used this data to calculate co-occurrence counts for words occurring within five tokens of one another. These are the co-occurrence counts used in all experiments reported in this paper.

### 1.3 Training the model

Once the semantic features $f_i(w)$ are specified, the parameters $c_{vi}$ that define the neural signature contributed by the $i^{th}$ semantic feature to the $v^{th}$ voxel are estimated. This is accomplished by training the model using a set of observed fMRI images associated with known stimulus words. Each training stimulus $w_t$ is first re-expressed in terms of its feature vector $< f_1(w_t) \ldots f_n(w_t) >$, and multiple regression is then used to obtain maximum likelihood estimates of the $c_{vi}$ values; that is, the set of $c_{vi}$ values that minimize the sum of squared errors in reconstructing the training fMRI images. If the number of semantic features is less than the number of training examples, then this multiple regression problem is well posed and a unique solution is obtained. If the number of semantic features is greater than the number of training examples, a solution can be obtained by introducing a regularization term such as a penalty equal to the sum of squares of the learned regression weights.

Once trained, the resulting computational model can be used to predict the full fMRI activation image for any other word found in the trillion ($10^{12}$) token text corpus, as shown in Figure 2A of the main text. Given an arbitrary new word $w_{new}$ the model first extracts the intermediate semantic feature values $< f_1(w_{new}) \ldots f_n(w_{new}) >$ from the corpus statistics database, then applies the above formula using the previously learned values for the parameters $c_{vi}$. The computational model and corresponding theory can be directly evaluated by comparing their predictions for words outside the training set to observed fMRI images associated with those words. Different predefined sets of intermediate semantic features can be directly compared by training competing models and evaluating their prediction accuracies.

The detailed list of intermediate semantic features vectors for each of the 60 stimulus words can be found at www.cs.cmu.edu/~tom/science2008.

## 1.4 Training and Evaluating Computational Models

Alternative computational models were trained based on different sets of intermediate semantic features. Each model was trained and evaluated using a cross validation approach, in which the model was repeatedly trained using only 58 of the 60 available stimulus items, then tested using the two stimulus items that had been left out. On each iteration, the trained model was tested by giving it the two stimulus words it had not yet seen ($w_1$ and $w_2$), plus their observed fMRI images ($i_1$ and $i_2$), then requiring it to predict which of the two novel images was associated with which of the two novel words, using a matching procedure described in the following section. This leave-two-out train-test procedure was iterated 1770 times, leaving out each of the possible word pairs. The expected accuracy in matching the two left-out words to their left-out fMRI images is 0.50 if the matching is performed at chance levels.

## 1.5 Matching predicted to actual images

Given a trained computational model, two new words (w1 and w2) and two new images (i1 and i2), the trained model was first used to create predicted image p1 for word w1 and predicted image p2 for word w2. It then decided which was a better match: (p1=i1 and p2=i2) or (p1=i2 and p2=i1), by choosing the image pairing with the best similarity score. Because we do not expect every voxel in the brain to be involved in representing the meaning of the stimulus, only a subset of voxels was used for assessing the similarity between images. This subset of voxels was selected automatically during training, using only the data for the 58 training words, and excluding the data from the two test words. The voxel selection method is described below. Let sel(i) be the vector of values of the selected subset of voxels for image i. The similarity score between a predicted image, p, and observed image, i, was calculated as the cosine similarity between the vectors sel(p) and sel(i). Cosine similarity between two vectors is defined as the cosine of the angle between the vectors, and was computed as the dot product of these vectors normalized to unit length. Finally, the similarity match score for a candidate pairing of predicted to actual images, (e.g., p1=i2 and p2=i1), was computed as the sum of the two cosine similarities:

match(p1=i2 and p2=i1) = cosineSimilarity(sel(p1),sel(i2))+cosineSimilarity(sel(p2),sel(i1)).

Cosine similarity was the first similarity measure we considered, but we subsequently also considered the Pearson correlation between two images and found that the two yielded similar results. All results reported in the current paper use cosine similarity.

## 1.6 Voxel selection

As described above, similarity between two images was calculated using only a subset of the image voxels. Voxels were selected automatically during training, using only the 58 training words on each of the leave-two-out cross validation folds. To select voxels, all voxels were first

assigned a "stability score" using the data from the 6 presentations of each of the 58 training stimuli. Given these 6 x 58 = 348 presentations represented as 348 fMRI images, each voxel was assigned a 6x58 matrix, where the entry at row i, column j, is the value of this voxel during the ith presentation of the jth word. The stability score for this voxel was then computed as the average pairwise correlation over all pairs of rows in this matrix. In essence, this assigns highest scores to voxels that exhibit a consistent (across different presentations) variation in activity across the 58 training stimuli. For example, if a voxel were to exhibit the same 58 responses during each presentation, it would have an average pairwise correlation of 1.0. Of course the noise inherent in fMRI activations prevents this from happening in practice, and high pairwise correlations tend to be found only when there is a strong and repeatable voxel response pattern of signals that outweighs this noise. Note that high pairwise correlations can occur even among voxels that activate similarly for some of the 58 stimuli, so long as they activate differently (and consistently so) for at least some other subset of the 58 stimuli. The 500 voxels ranked highest by this stability score were used in the cosine similarity test described above. Although individual selected voxels might distinguish among only some subset of the stimuli, the entire set of voxels selected in this fashion tends to distinguish fairly well in practice among all stimuli, as is evident from the reported results.

### 1.7 Empirical distribution to determine statistical significance and p values

The expected chance accuracy of an uninformed model correctly matching two stimuli outside the training set to their two fMRI images is 0.5. The observed accuracies of our trained models, based on 1770 iterations of a leave-two-out cross validation train/test regime, are higher than 0.5. Here we consider the question of how to determine p values based on observed accuracies, to reject the null hypothesis that the trained model has true accuracy of 0.5. Given our leave-two-out train/test regime, no closed-form formula is available to assign such a p value. Therefore, we computed p values based on an empirical distribution of observed accuracies obtained from 768 independently trained single-participant models that we expect will have true accuracy very close to 0.5. The empirical distribution of accuracies for these null models was 0.501, with standard deviation 0.070, indicating that observed accuracies above 0.62 for a single participant model is statistically significant at $p<0.05$. Below we describe our approach in more detail.

We created this empirical distribution of accuracies by training multiple models using the observed fMRI images for the 60 stimulus words, but using different word labels and different intermediate semantic features. This approach is similar to a form of permutation test, except that instead of permuting the 60 stimulus labels, we chose 60 new words from the vocabulary of tokens in our text corpus. In particular, each model was trained by first choosing one of our nine participant data sets uniformly at random, then selecting 60 words uniformly at random from the 500 through 5000 most frequent words in the text corpus, then selecting 25 intermediate semantic feature words uniformly at random from the 500 through 5000 most frequent words in the corpus. The model was then trained and tested, substituting the 60 randomly drawn words

for the 60 correct word labels, and using the 25 randomly drawn intermediate semantic feature words. Models were trained and tested using the leave-two-out test regime, exactly as elsewhere in this paper, with one minor exception: in these models the 500 most stable voxels were selected using data from all 60 words, whereas elsewhere this selection of stable voxels was based only on the 58 training words. This exception was made because it dramatically improves the tractability of training hundreds of such random models, leading to a 1000-fold speedup. Note the net effect is that the expected observed accuracy of the random models evaluated in this way will be slightly positively biased, and the p values calculated from the resulting distribution will therefore be slightly conservative. In fact, we found this bias to be very small, as the empirical mean accuracy of models trained and tested in this way was 0.501, very close to the expected chance accuracy of 0.500.

We trained and tested 768 such randomly generated models. The mean accuracy over these 768 models was 0.501, with standard deviation 0.070. The distribution of observed accuracies is plotted in Figure S2. Examining the cumulative distribution, we found that 95% of these models had accuracies below 0.621, and therefore assign a p value of $p<0.05$ to single subject models with observed accuracies above 0.621. As a consistency check, we also modeled the empirical distribution of accuracies as a Gaussian with $\mu=0.501$ and $\sigma=0.070$, and, based on the cumulative distribution for a Gaussian found that $p<0.05$ corresponds to accuracies greater than $\mu+1.645\sigma = 0.617$, which is very close to the 0.621 obtained from the empirical cumulative distribution. Under this same Gaussian model, an accuracy of 0.719 for a single-participant model would be significant at $p < .001$. Notice the above analysis applies to the accuracy of a single model trained for a single participant. The p value associated with observing that all nine independently trained participant models exhibit accuracies greater than 0.62 is $p< 10^{-11}$.

### 1.8 Computing the accuracy map of Figure 3 in the main paper

The accuracy map in Figure 3 of the main text shows voxel clusters with the highest correlation between predicted and actual voxel values. We first calculated sixty predicted images for the sixty words, training a model on the other 59 words, then using this to predict the remaining word. For each voxel, this produced a set of 60 predicted values. The accuracy score of each voxel was calculated as the Pearson correlation between this vector of its predicted values and the corresponding vector of its observed values. An image map containing these voxel scores was created, and the clusters shown in Figure 3 were then produced using standard SPM tools, to identify clusters containing at least 10 contiguous voxels whose score was greater than a threshold value (0.28 for Figure 3A and 3B, and 0.14 for Figure 3C).

# 2. Additional Results and Observations

## 2.1 Experiment with randomly generated intermediate semantic features

Features in this experiment, summarized in Figure 5 of the main text, were defined by 25 randomly selected words. These 25 words were chosen uniformly at random from the 5000 most frequently occurring tokens in the text corpus, and omitting the 500 most frequent tokens (which include many function words such as "the" and "of") as well as the 60 stimulus nouns. Models were trained and tested exactly as described for our 25 manually selected verbs, with one exception which introduced a slight optimistic bias in the measured accuracy of models trained with these randomly generated features: Instead of performing voxel selection using just the 58 training words on each cross-validation fold, the voxels were instead selected just once for each participant and feature set, using all 60 words. This change was introduced in order to reduce the computational cost of training and testing models, enabling us to explore a larger variety of randomly generated feature sets. As discussed in the above section on "Empirical distribution to determine statistical significance and p values," we estimate that the positive bias in observed accuracies due to performing voxel selection is this way is negligible.

Compared to our manually generated set of 25 semantic features, the randomly generated feature sets differed in two ways worth noting. First, whereas our manually generated features were all verbs, the randomly generated features contained tokens of all kinds, including many adjectives, verbs, adverbs, nouns, proper names, slang words, and some tokens frequently found on the web which may not be commonly thought of as English words (e.g., "html"). Second, whereas we defined the features for our verbs using three forms of the verb (e.g., the feature for the verb "eat" used the sum of co-occurrences with the three forms of the verb "eat," "eats," and "ate"), we did not attempt to expand randomly selected tokens into such sets of related tokens. In general there is no obvious way to automatically expand arbitrary word tokens in an analogous fashion, and for many words (e.g., "partly," "news") it is unclear how to do this even manually.

For each randomly generated feature set, models were trained for each of the nine participants. Among the 115 randomly generated feature sets, the greatest mean accuracy achieved across the nine participants was 0.68, compared to 0.77 for the 25 manually selected verbs. The set of 25 randomly selected feature tokens that achieved this 0.68 accuracy is: *seems, productions, lots, various, counts, seek, lab, arizona, body, pieces, drop, disabled, lol, venture, finally, arts, eating, infrastructure, xml, nikon, ericsson, partly, governments, ladies*, and *ft*. The feature set with the lowest nine-participant mean accuracy achieved an accuracy of 0.46. This feature set used the tokens: *outcome, sessions, schedule, failure, characteristics, statistics, med, beauty, mt, alternative, richard, responsible, god, parties, candidates, towards, governments, fred, father, seeking, kim, hunt, xxx, keeps,* and *summary*. In scanning the feature sets with higher versus lower accuracies, we found no obvious regularities.

## 2.2 Learned Feature Signatures

Figure 4 in the main text shows some of the feature signatures for participant P1, and averaged over nine participants. Voxels that were absent in any participant were excluded from the image displaying the mean over participants. The full set of 25 feature signatures for participant P1 and averaged over nine participants is available online at www.cs.cmu.edu/~tom/science2008

## 2.3 Plot of similarities between predicted and actual images

To provide more insight into the power of the trained computational model, Figure S3 depicts for participant P1 the cosine similarity score between each of the 60 predicted images and each of the 60 observed images, using the 500 most stable voxels as described above. Here the entry at row i and column j gives the cosine similarity between the predicted image for stimulus word i, and the observed image for word j, using a model trained without either word (training on the other 58 words). Thus, this figure contains only similarity scores between pairs of words outside the training set. Note high positive values along the diagonal indicate correct predictions at the word level. High values in blocks around the diagonal reflect similarities between images from the same semantic category. Note also the dark blue regions generally indicate category pairs where the predicted images for words from category A are very different from (have negative cosine similarity with) category B. Whereas Figure S3 shows the similarities for participant P1, Figure S4 shows the similarities averaged over all nine participants.

Examining the entries in Figure S4, one can determine how well the similarity scores resolve on average the correct word out of the 60 candidates. In particular, each row shows the similarity scores of the predicted word's image to each of the 60 observed images (each calculated by a model that omitted the two words being compared). Sorting these similarity scores for each row from most to least similar, the score of the correct word appears at the 79th percentile on average, indicating an imperfect but strong ability of the model to predict images whose features resolve among the 60 words. The percentile rank of the correct image for each of the 60 words is shown below. Words here are numbered according to their position in Figures S3, S4 and S5.

1. 0.283 bear
2. 0.767 cat
3. 0.517 cow
4. 0.950 dog
5. 0.950 horse
6. 0.750 arm
7. 0.583 eye
8. 0.933 foot
9. 0.883 hand
10. 0.833 leg
11. 0.917 apartment
12. 0.950 barn
13. 0.950 church
14. 0.950 house
15. 0.400 igloo
16. 0.900 arch
17. 0.933 chimney
18. 0.983 closet
19. 0.967 door
20. 0.983 window
21. 0.850 coat
22. 0.967 dress
23. 0.933 pants
24. 0.850 shirt
25. 0.867 skirt
26. 0.717 bed
27. 0.783 chair
28. 0.833 desk
29. 0.833 dresser
30. 0.550 table
31. 0.867 ant
32. 0.900 bee
33. 0.917 beetle

34. 0.317 butterfly
35. 0.783 fly
36. 0.983 bottle
37. 0.817 cup
38. 0.983 glass
39. 0.900 knife
40. 0.967 spoon
41. 0.383 bell
42. 0.267 key
43. 0.867 refrigerator
44. 0.283 telephone
45. 0.867 watch
46. 0.883 chisel
47. 0.833 hammer
48. 0.933 pliers
49. 0.067 saw
50. 0.967 screwdriver
51. 0.783 carrot
52. 0.767 celery
53. 0.950 corn
54. 0.567 lettuce
55. 0.150 tomato
56. 0.867 airplane
57. 0.983 bicycle
58. 0.883 car
59. 0.983 train
60. 0.983 truck

Note the word producing the worst prediction above is "saw" (word 49). This is primarily due to the fact that although we presented "saw" to our subjects as a tool, the token co-occurrence counts for "saw" used by the model are dominated by its more frequent use as a verb (past tense of "see"). This suggests that future refinements to our model might achieve even greater accuracy by using an enriched set of corpus features that distinguish different meanings of word tokens.

For comparison to Figure S4, Figure S5 shows the similarities between the sixty observed images (and is therefore a summary of the data, rather than the trained models). More specifically, the entry at row i and column j shows the mean, over the nine participants, of the similarity between the observed images for words i and j for that participant. Comparing Figure S5 to Figure S4, it is possible to see that some of the confusions in the predicted versus actual images (off-diagonal red and yellow entries) are the result of similarities in the actual observed images for the two stimuli, whereas other confusions reflect errors in the model in failing to predict differences that do exist in the actual images. For example, it appears that the similarities visible in Figure S4 between the predicted and observed images for furniture items and building parts may be due to actual similarities between the neural encodings of these objects as seen in Figure 4. In comparing Figures S3, S4 and S5, note the color scale is customized to each figure, setting the brightest red to the maximum in the matrix, and the darkest blue to the minimum.

## 2.4 Resolving among 1000 candidate words

As described in the main paper, we also performed a leave-one-out test in which the model was repeatedly trained using 59 of the 60 available stimuli, and was then asked to rank a set of 1001 candidate words according to which candidate was most likely to have produced the held out fMRI image. The ranking was based on the cosine similarity between the held out fMRI image and the predicted images for each of the candidate words (as usual, using only the 500 most stable voxels over the training data). For this experiment we used the 1300 most frequent tokens in the text corpus, omitting the 300 most frequent (which contain many function words such as "for" and "the"). As noted in the main paper, the mean percentile rank of the correct word in the

model's ranked list was 0.72 on average, across all nine participants. The median rank accuracy of the correct word across all participants was 0.79, reflecting the fact that most words were ranked fairly highly, and a smaller number were ranked very poorly. Below is the list of all 60 words, sorted by their average percentile rank across all nine participants (the number next to each word is the mean percentile rank of this word in the sorted list of candidates, when it was the correct candidate word). As can be seen, some words, such as "glass" are very accurately predicted on average across all participants, with only 26 of the 1000 candidates on average ranked more likely to have generated the test fMRI image. Other words, such as "saw" and "bear" are ranked very poorly on average. Notice that the accurately and inaccurately ranked words below correlate highly with the words ranked accurately and inaccurately in the list above, associated with Figure S4. The difference between these two lists is that the list below involves ranking 1001 predicted images by their similarity to the single observed fMRI image for the held-out word. In contrast, the list above involves ranking the 60 observed fMRI images by their similarity to the single predicted image for the held-out word. .

1. 0.974 glass
2. 0.955 chimney
3. 0.914 church
4. 0.905 train
5. 0.898 bicycle
6. 0.890 dress
7. 0.889 closet
8. 0.889 screwdriver
9. 0.886 foot
10. 0.884 bottle
11. 0.878 arch
12. 0.868 house
13. 0.856 airplane
14. 0.852 horse
15. 0.851 door
16. 0.849 spoon
17. 0.846 barn
18. 0.837 window
19. 0.825 hammer
20. 0.824 knife
21. 0.822 chisel
22. 0.821 car
23. 0.819 dresser
24. 0.814 skirt
25. 0.810 truck
26. 0.802 leg
27. 0.799 hand
28. 0.796 refrigerator
29. 0.796 bee
30. 0.792 dog
31. 0.791 cup
32. 0.775 watch
33. 0.771 apartment
34. 0.769 pants
35. 0.765 pliers
36. 0.751 desk
37. 0.743 bed
38. 0.743 coat
39. 0.738 corn
40. 0.732 shirt
41. 0.718 carrot
42. 0.703 chair
43. 0.702 ant
44. 0.673 fly
45. 0.668 celery
46. 0.628 arm
47. 0.585 cat
48. 0.585 beetle
49. 0.570 table
50. 0.533 eye
51. 0.512 bell
52. 0.512 key
53. 0.476 cow
54. 0.453 lettuce
55. 0.434 igloo
56. 0.345 tomato
57. 0.307 butterfly
58. 0.295 telephone
59. 0.242 bear
60. 0.171 saw

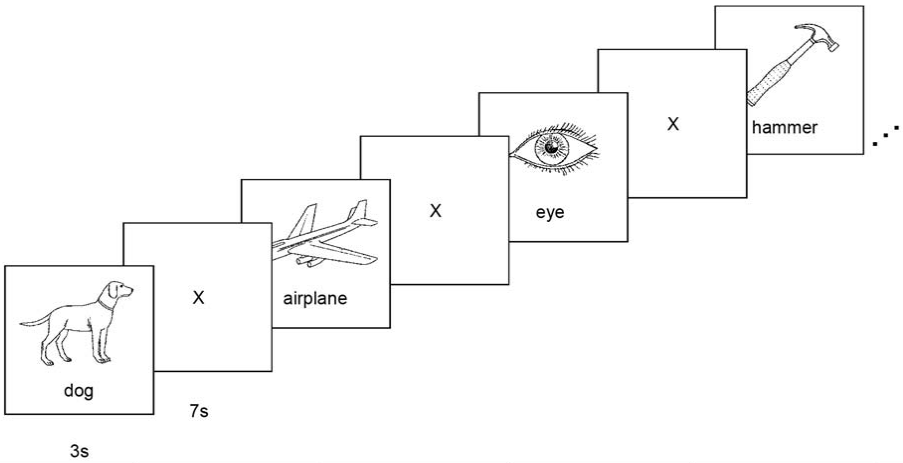
## 2.5 Note on use of co-occurrence counts to define semantic features

Although using co-occurrence counts to approximate the semantic content of a word or document is a common technique in computational linguistics, this remains a crude approach with several shortcomings. One is due to the fact that simple co-occurrence within a specified window fails to resolve the syntactic relation between the two words. For example, the relation between "mouse" and "ate" is very different in the sentence "The mouse ate the cheese" versus "The cat ate the mouse." Our co-occurrence counts fail to resolve, for example, cases where the noun is the subject, versus the direct object of the verb with which it co-occurs. Second, many words have multiple meanings, and our approach fails to resolve these. For example, although the token "saw" can refer to a noun (a tool), it more commonly refers to a verb (past tense of "see"), resulting in a semantic feature vector that is unrepresentative of its intended meaning as a tool, and to resulting poor prediction for this word. Despite these shortcomings, the co-occurrence data collected from the very large corpus appears to suffice in capturing enough of the meaning of our stimulus words to support a reasonable model. We believe stronger models can be developed in the future by considering more sophisticated linguistic features (e.g., by parsing the sentences to determine the relationship between verb and noun, and by automatically resolving among different word senses).

## 2.6 Availability of additional online materials

Additional information is available at www.cs.cmu.edu/~tom/science2008. At the time of publication of this paper, additional information available at this site included the detailed list of intermediate semantic feature vectors for each of the 60 stimulus words, displays of the 25 semantic feature signatures (similar to those shown in Figure 3 of the main paper) for participant P1, and displays of the 25 semantic feature signatures averaged over all nine participants.

# 3. Additional Figures and legends




| Category | Exemplar 1 | Exemplar 2 | Exemplar 3 | Exemplar 4 | Exemplar 5 |
|---|---|---|---|---|---|
| animals | bear | cat | cow | dog | horse |
| body parts | arm | eye | foot | hand | leg |
| buildings | apartment | barn | church | house | igloo |
| building parts | arch | chimney | closet | door | window |
| clothing | coat | dress | pants | shirt | skirt |
| furniture | bed | chair | desk | dresser | table |
| insects | ant | bee | beetle | butterfly | fly |
| kitchen utensils | bottle | cup | glass | knife | spoon |
| man made objects | bell | key | refrigerator | telephone | watch |
| tools | chisel | hammer | pliers | saw | screwdriver |
| vegetables | carrot | celery | corn | lettuce | tomato |
| vehicles | airplane | bicycle | car | train | truck |

***Figure S1. Presentation and set of exemplars used in the experiment.*** Participants were presented 60 distinct word-picture pairs describing common concrete nouns. These consisted of 5 exemplars from each of 12 categories, as shown above. A slow event-related paradigm was employed, in which the stimulus was presented for 3s, followed by a 7s fixation period during which an X was presented in the center of the screen. Images were presented as white lines and characters on a dark background, but are inverted here to improve readability. The entire set of 60 exemplars was presented six times, randomly permuting the sequence on each presentation.

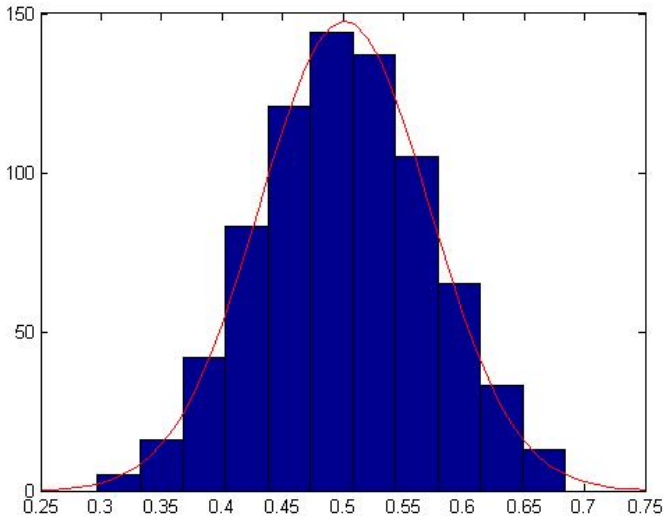



***Figure S2. Empirical distribution of accuracies for null models, and Gaussian approximation.*** The blue histogram shows the observed accuracies for the 768 randomly generated single-participant null models (mean = 0.501, standard deviation = 0.070). The red line shows a Gaussian distribution with this mean and standard deviation. This empirical distribution was used to determine p values for the observed model accuracies.

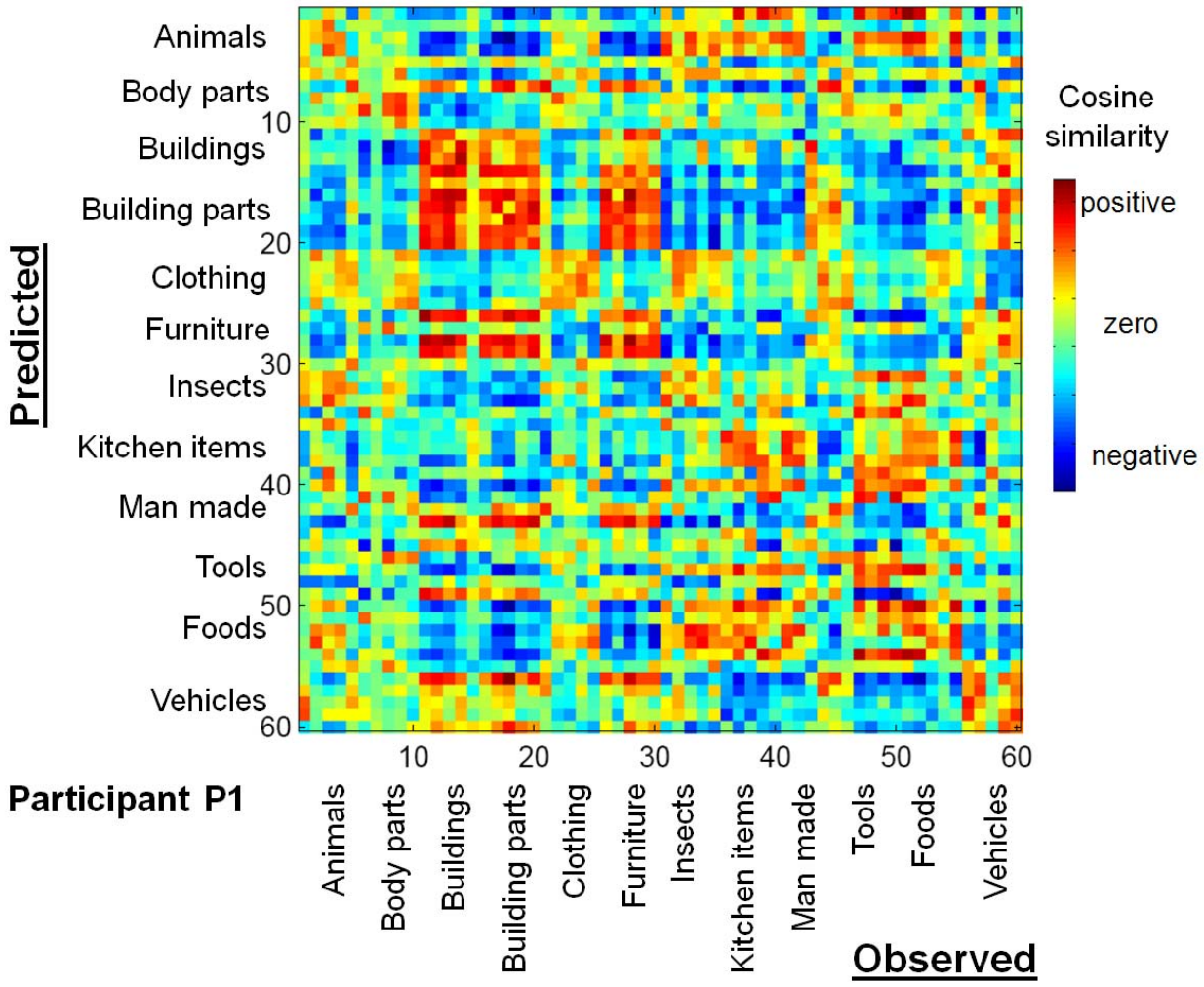



***Figure S3. Cosine similarities between predicted and actual images for participant P1.*** The point at row i, column j, shows the cosine similarity between the image predicted for word i, and the image observed for word j, when using a model trained on the other 58 words and excluding words i and j. Numbering of exemplars of each category follows the chart shown in Figure S1, and similarity was calculated over the 500 most accurate voxels measured over the 58 word training set. High positive values along the diagonal indicate that predicted images for a given word are similar to the observed image for that word.

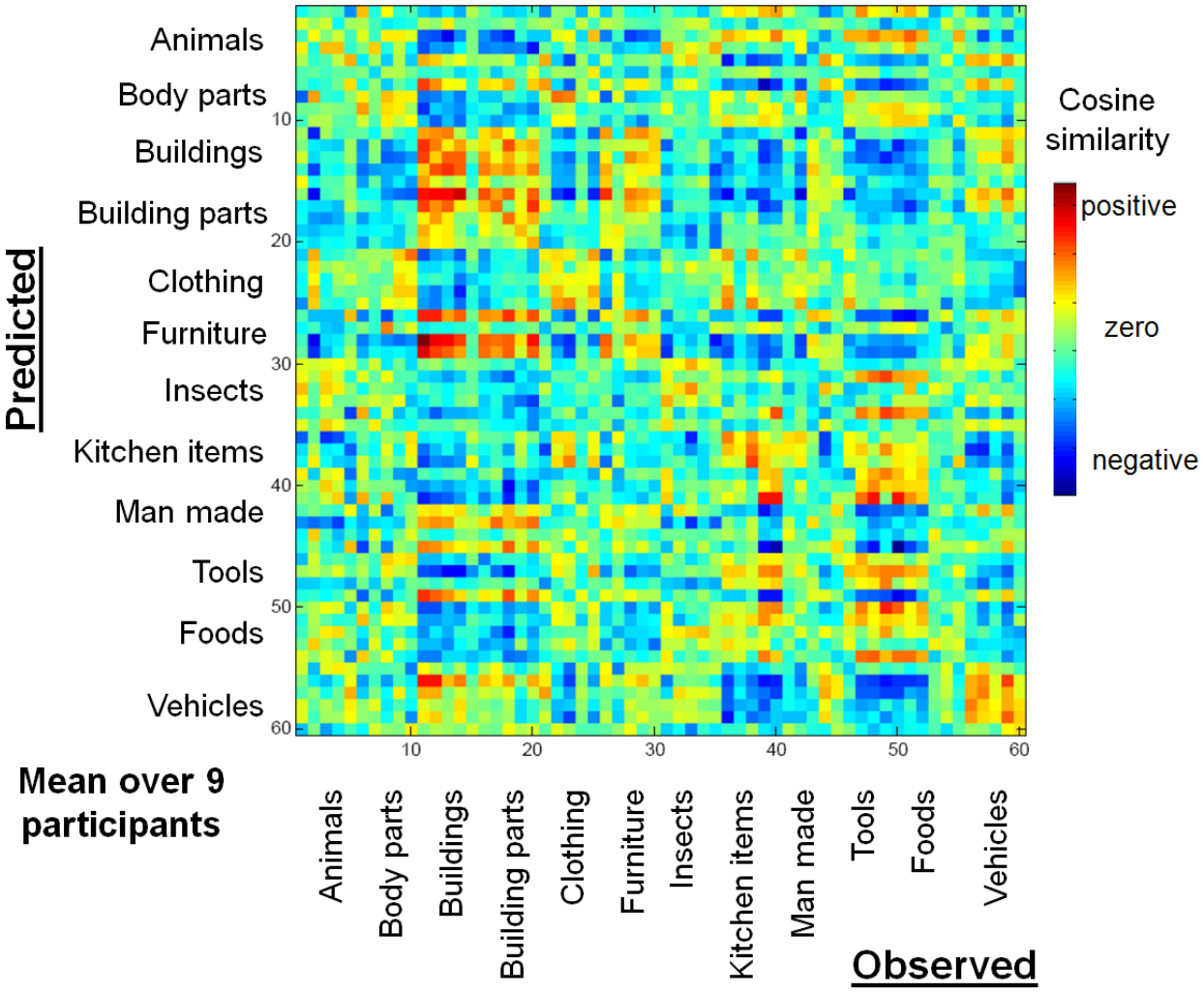



***Figure S4. Cosine similarities between predicted and actual images, averaged over all participants.*** This figure follows the same conventions as Figure S3, except that it reflects the average similarities between predicted and observed images, averaged over the nine participants. The mean of the diagonal values is 0.179, whereas the mean over the entire matrix is -0.016, indicating that on average the predicted image is more similar to the actual image than to others. The maximum (most red) value in the matrix is 0.65, and the minimum (most blue) is -0.60.

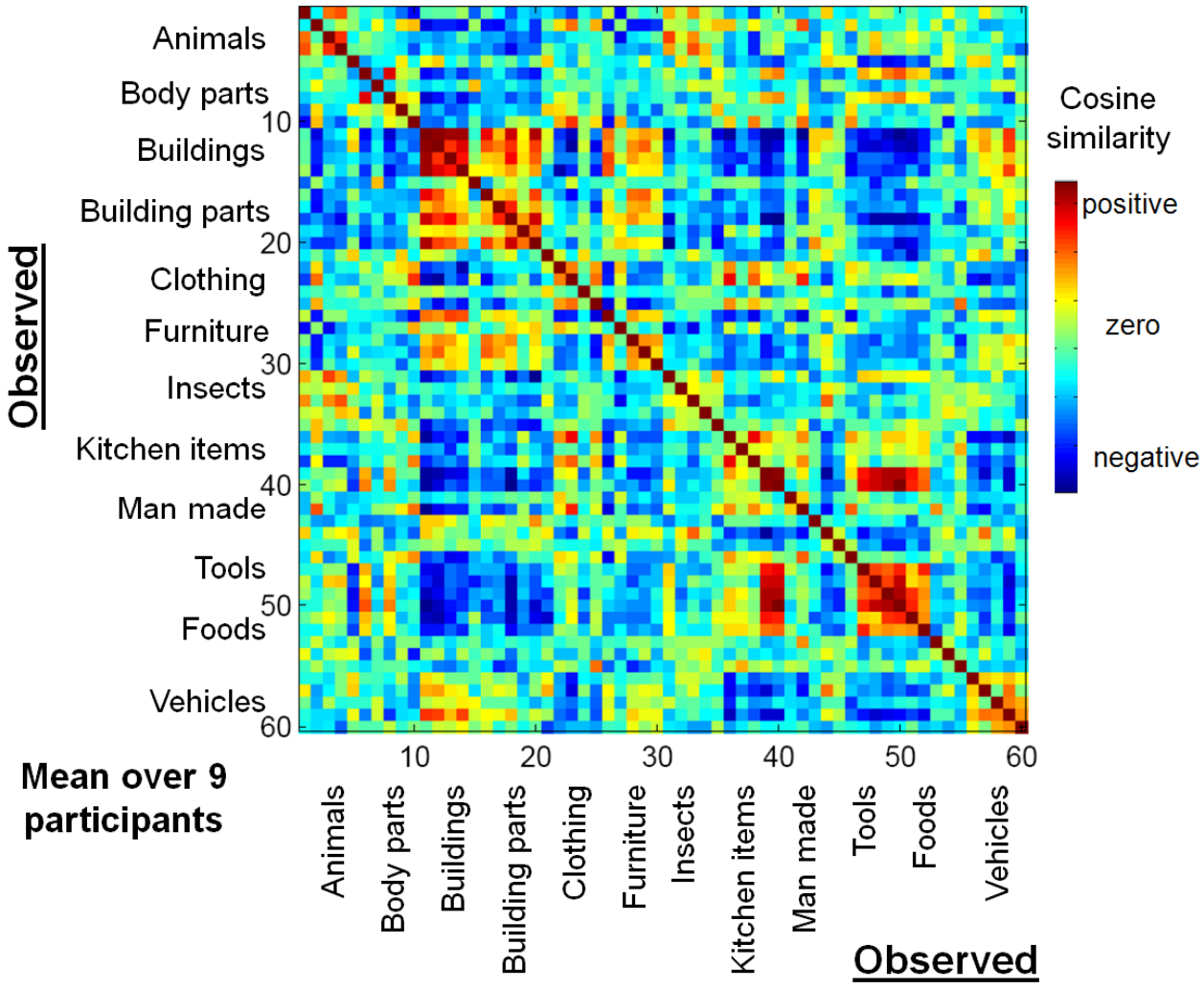



***Figure S5. Cosine similarities between actual images, averaged over all participants.*** This figure follows the same conventions as Figures S3 and S4, except that it reflects the average similarities between pairs of observed images, averaged over the nine participants. High values in blocks along the diagonal reflect similarities between images from the same semantic category. Ignoring the diagonal entries, whose similarity values are 1.0, the maximum off-diagonal value is 0.52, and the minimum is -0.41.

## 4. Additional References

S1. J.G. Snodgrass & M. Vanderwart, *J. Exp. Psy.: Human Learning & Memory*, **6,** 174. (1980).

**Predicting Human Brain Activity Associated with the Meanings of Nouns**
Tom M. Mitchell, *et al.*
*Science* **320**, 1191 (2008);
DOI: 10.1126/science.1152876

# Supporting Online Material for

## Predicting Human Brain Activity Associated with the Meanings of Nouns

Tom M. Mitchell,* Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, Marcel Adam Just

*To whom correspondence should be addressed. E-mail: Tom.Mitchell@cs.cmu.edu

Published 30 May 2008, *Science* **320**, 1191 (2008)

DOI: 10.1126/science.1152876

**This PDF file includes:**

title: Predicting Human Brain Activity Associated with the Meanings of Nouns author: Tom M. Mitchell and Svetlana V. Shinkareva and Andrew Carlson and Kai-Min Chang and Vicente L. Malave and Robert A. Mason and Marcel Adam Just date: 30 MAY 2008 journal: SCIENCE VOL 320

# 名詞の意味に関連した人間の脳活動の予測

## 要旨

人間の脳が概念的な知識をどのように表現しているかという問題は， 多くの科学分野で議論されてきた。 これまでの研究では， 絵や言葉の意味カテゴリ (道具，建物， 動物など) を考えるときに， 異なる空間パターンの神経活性化が生じることが示されている。 本研究では fMRI のデータがまだ得られていない単語に関連する機能的磁気共鳴画像 (fMRI) の神経活性化を予測する計算モデルを提案する。 このモデルは、1 兆語規模のテキストコーパスのデータと，数十個の具体的な名詞を見たときに観測された fMRI データを組み合わせて学習される。 このモデルは， テキストコーパスに含まれる数千の具体的な名詞の fMRI 活性化を予測し， 現在 fMRI データが得られている 60 の名詞よりも高い精度で予測することができる。

人間の脳が概念的な知識をどのように表現し， 整理しているのかという問題は， 多くの科学者によって研究されてきた。 脳のイメージング研究を行っている神経科学者（1-9）は，道具， 建物， 動物などの特定の意味カテゴリの写真を見たときに， fMRI 活動の明確な空間パターンが関連することを示している。 言語学者は，個々の動詞に関連する様々な意味上の役割と，それらの意味上の役割を果たすことができる名詞の種類を特徴付けてきた (例えば VerbNet (10) や WordNet (11, 12)。 計算言語学者は， 非常に大規模なテキストコーパスの統計を分析し， ある単語の意味は， その単語がよく共起する単語やフレーズの分布によってある程度把握されることを実証している (13-17)。 心理学者は，参加者にさまざまな単語から連想される特徴を挙げてもらう特徴規範研究 (18) を通じて， 単語の意味を研究してきた。 その結果， 個人間で一貫した中核的な特徴があることが明らかになり， 感覚運動のモダリティによって特徴がグループ化される可能性が示唆された。 脳損傷による意味論的影響を研究している研究者たちは， 特定の意味論的カテゴリー (動物など) に特異的な障害を発見している [@CaramazzaShelton1998;@2003CrutchWarrington_fruits_vegetables;@2004Samson_Pillon]。

このような様々な実験結果から， 脳が言葉の意味や物の知識をどのように符号化しているかについて， 感覚運動皮質領域で意味が符号化されているという説 (22,23) や， 生物や非生物などの 意味カテゴリで整理されているという説 (18, 24) など， 相反する理論が生まれた。 これらの競合する理論は， 時に異なる予測 (例えば 脳障害を受けた患者にどのような呼称障害が併発するか) を導き出すこともある。 だが， これらは主に記述的な理論であり， 被験者が特定の単語を読んだり， 特定の物体の絵を見たりしたときに生じる具体的な脳の活性化を予測しようとするものではない。

本研究では， 現在 fMRI データがない多くの名詞を含む， 任意の具体的な名詞について考えることに関連する fMRI 活動を直接検証可能な形で予測する計算モデルを発表する。 この計算モデルの基礎となる理論は， 具体的な名詞の意味表現の神経基盤が， その言語の広範なコーパスにおけるそれらの単語の分布特性に関連しているというものである。 具体的な対象物の意味を符号化するために脳内で使用される基本的な特徴に関する異なる仮定に基づいて， 競合する計算モデルを訓練する実験について説明する。 その結果， これらのモデルが fMRI の神経活動を十分に予測し， まだ未知の単語とその fMRI 画像を， 偶然をはるかに上回る精度で照合できることを実験的に示した。 これらの結果は， テキスト中の単語の共起の統計量と， 単語の意味を考えることに関連する神経活動との間に，直接的な予測関係を確立している。

# 1. アプローチ

本研究では，訓練可能な計算モデルを用いて，任意の刺激語 $w$ に対する神経活性化を 2 段階のプロセスで予測する。 任意の刺激語 $w$ が与えられた場合、最初のステップでは、英語のテキストにおける単語の典型的な使用法を捉えた大規模テキストコーパス(25)内の刺激語 $w$の出現頻度から計算された中間的な意味的特徴のベクトルとして $w$ の意味をエンコードする。 例えば $w$ が動詞 "hear" と共起する頻度が中間的な意味的特徴として挙げられる。 第 2 段階 では 脳内の各ボクセル位置における fMRI 活性値を 中間的意味特徴のそれぞれが寄与する神経活性化の加重和として予測する。 より正確には、単語 $w$ に対する脳内ボクセル $v$ での予測される活性化 $y_v$ は以で与えられる:

$$y_v = \sum_{i=1}^{n} c_{vi} f_i(w), \quad (1)$$

ここで $f_i(w)$ は単語 $w$ に対するi番目の中間的な意味特徴の値 $n$ はモデル内の意味特徴の数 $c_{vi}$ は i番目の中間的な意味特徴がボクセル $v$ を活性化する程度 を指定する 学習済みのスカラーパラメータである。 この式は 単語 $w$ に対する全ボクセルの fMRI 画像を 意味特徴量 $f_i$ ごとに 1つ ずつの画像の加重和として予測していると解釈できる。 学習した $c_{vi}$ で定義されるこれらの意味的特徴画像は、入力刺激語の異なる意味的成分に関連する脳の活性化をモデル化する成分画像の基礎集合を構成する。

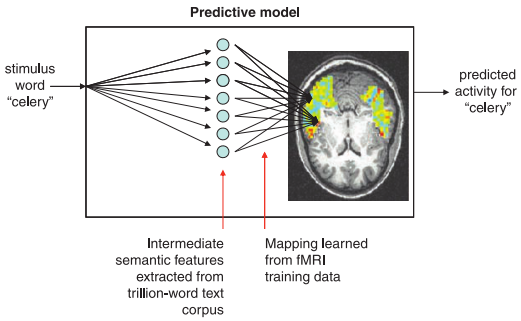


図 1. 任意の名詞刺激に対するfMRI活性化を予測するモデルの形式。 fMRI の活性化は、2段階 プロセスで予測される。 第 1 段階では，入力刺激語の意味を，典型的な単語使用を示す大規模なテキストコーパスから値を抽出した中間的な意味的特徴の観点から符号化する。 第 2 段階では，これらの中間的な意味的特徴のそれぞれに関連する fMRIシグネチャ の線形結合として，fMRI 画像を予測する。

この計算モデルの枠組みでモデルを完全に規定するためには、まず、テキストコーパスから抽出する中間意味特徴 $f_1(w)f_2(w)\cdots f_n(w)$ のセットを定義しなければならない。 本論文では、各中間意味特徴は、入力刺激語 $w$ と、 テキストコーパス内の特定の他の単語 (例えば "taste") または単語のセット (例えば "taste", "taste", "tasted") との共起統計の観点から定義される。 モデルの学習は，これらの特徴量 $f_i(w)$ と観測された fMRI 画像に重回帰を適用して，モデルのパラメータ $c_{vi}$ の最尤推定値を得ることで行われる。 学習後、学習セット以外の単語を与え、その単語に対する予測fMRI画像と観測された fMRIデータ を比較することで、計算モデルを評価することができる。

この計算モデルの枠組みは 2つ の重要な理論的仮定に基づいている。 まず，任意の具体的な名詞の意味を区別する意味的特徴が，大規模テキストコーパス内での使用の統計に反映されていると仮定している。 この仮定は，文書や単語の意味を近似するために統計的な単語分布が頻繁に用いられる計算言語学の分野から導き出されたものである (14-17)。 次に、具体的な名詞について考えているときに観察される脳活動は、その名詞の各意味的特徴からの寄与の加重線形和として得られると仮定する。 この線形性の仮定が正しいかどうかは議論の余地がある。しかし fMRI 解析で線形モデルが広く使われていること (27) や、 fMRI の活性化が異なるソースからの寄与の線形的な重ね合わせを反映していることが多いという仮定と一致している。 我々の理論的枠組みは、意味を符号化する神経活性化が特定の皮質領域に局在するかどうかについては見解を示さない。 その代わりに、すべての皮質ボクセルを考慮し、どの場所が単語の意味のどの側面によって系統的に変調されるかを訓練データによって決定する。

## 2. 結果

この計算モデルを，60種類 の単語と絵のペアを 6回 ずつ提示した 9人 の大学生の健康な被験者の fMRIデータ を用いて評価した。 解剖学的に定義された関心領域は，(28) の方法にしたがって自動的にラベル付けされた． ランダムに並べられた 60個 の刺激には、12 の意味カテゴリ (動物, 体の一部, 建物, 建物の部品, 衣服, 家具, 昆虫, 台所用品, 道具, 野菜, 乗り物, その他の人工物) からそれぞれ 5つ の項目が含まれていた。 各刺激の代表的な fMRI画像 は その 6回 の呈示における fMRI 応答の平均値を計算することで作成した。 この代表的な画 像60個 すべての平均値をそれぞれから差し引いた［詳細は（26）参照］。

我々のモデル化の枠組みを具体化するために、我々はまず、中間的な意味的特徴を選んだ。 この中間的な意味特徴は， 入力刺激語の多種多様な意味内容を同時に符号化し， 観測されたfMRI活性化をより原始的な要素に分解し， 線形的に再結合することで， 任意の新しい刺激に対する fMRI 活性化 を予測するのに有効である。 物体の神経表現における感覚と運動の特徴の重要性に関する既存の仮説 (18,29) に基づいて， 我々は 25 の動詞によって定義される 25 の意味的特徴のセットを設計した。 (見る, 聞く, 聴く, 味わう, 嗅ぐ, 食べる, 触る, こする, 持ち上げる, 操作する, 走る, 押す, 満たす, 動かす, 乗る, 言う, 恐れる, 開く, 近づく, ,近い, 入る, 運転する, 着る, 壊す, 掃除する)。 (see, hear, listen, taste, smell, eat, touch, rub, ,lift, manipulate, run, push, fill, move, ride, say, fear, open, approach, near, enter, drive, wear, break, clean.) これらの動詞は一般的に， 基本的な感覚や運動， 物に対して行う動作， 空間的な関係の変化を伴う動作に対応している。 各動詞について，入力刺激語 $w$ に対応する中間的な意味特徴の値は，テキストコーパス上で動詞の 3つ の形式（例えば "taste", "tests", "tasted") のいずれかと $w$ の正規化された共起回数である。 ただし "see" という動詞は例外であった。 これは "saw" が 60 個 の刺激名詞の 1つ であるためである。 正規化とは，25 個の特徴量のベクトルを単位長さにスケーリングすることである。

この 25 の意味的特徴を用いて，9人 の被験者それぞれに個別の計算モデルを訓練した。 学習したモデルの評価は，leave-two-out 交差検証法 で行い，60個 の単語刺激とそれに関連する fMRI画像 のうち 58個 だけを使ってモデルを繰り返し学習した。 学習されたモデルは まず 「保留された」2つ の単語の fMRI画像 を予測し、次に、それらの単語と保留された fMRI画像 を正しく照合することを要求してテストされた。 図2A は、ホールドアウトされた単語の fMRI画像 を予測するプロセスを示している。 予測された2つの fMRI画像 と観察された 2つの fMRI画像 の一致は，どちらの一致がより高いコサイン類似度 を持つかによって決定され，訓練提示の間に最も安定した反応を示した 500 の画像ボクセル で評価された (26)。 左の単語と左の fMRI画像 のマッチングの予想精度は モデルの性能が偶然レベルであれば 0.50 である。 ランダムに生成されたヌルモデルの精度の経験的な分布に基づいて，1人 の参加者に対して訓練された 1つのモデルの精度が 0.62以上 であれば，偶然に比べて統計的に有意（$P < 0.05$）であると判断した (26)。 同様に，独立して訓練された 9つ の参加者固有のモデルのそれぞれについて 0.62以上 の精度を観測した場合，$P < 10^{-11}$ で統計的に有意となる.

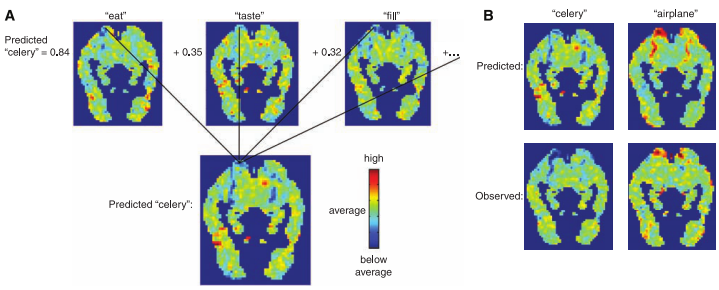



図 2. 与えられた刺激語に対する fMRI 画像の予測。 (A) 参加者 P1 が 「セロリ」刺激語に対して、他の 58 の単語で学習した後に予測を行う。 25 個の意味的特徴のうち 3 つの特徴量のベクトルを単位長にスケーリングすることである。 (食べる, 味わう, 満たす) について学習した $c_{vi}$ 係数は， パネル上部の 3 つの画像のボクセルの色で示されている。 刺激語「セロリ」に対する各特徴量の共起値は， それぞれの画像の左側に表示されている (例えば 「食べる（セロリ）」の 共起値は 0.84)。 刺激語の活性化予測値 ((A) の下部に表示) は 25個 の意味的 fMRI シグネチャを線形結合し， その共起値で重み付けしたものである。 この図は 予測された三次元画像の1つの水平方向のスライス [z=-12 mm in Montreal Neurological Institute (MNI) space] を示している。 (B) 「セロリ」と「飛行機」について， 他の 58 個の単語を使った訓練後に予測された fMRI 画像と観察された fMRI 画像。 予測画像と観測画像の上部（後方領域）付近にある赤と青の 2本 の長い縦筋は、左右の楔状回である。 }

参加者 P1〜P9 で学習したモデルの 2 つ の未見の 単語刺激と未見の fMRI 画像 を照合する交差検証の精度は 0.83, 0.76, 0.78, 0.72, 0.78, 0.85, 0.73, 0.68, 0.82 (平均=0.77) であった。 このように，参加者ごとに設定した 9 つ のモデルすべてが，偶然のレベルを大幅に上回る精度を示した。 これらのモデルは，9 人 の参加者の間で 交差検証 された 15,930 組 のテスト対のうち，4 分の 3 以上 において，以前に見たことのない単語のペアを識別することに成功した。 参加者間の精度は，推定された頭部の動きと強い相関 (r=-0.66)

を示した (すなわち，参加者の頭部の動きが少ないほど，予測精度は高くなる)。 これは，参加者間の精度のばらつきが，頭部の動きによるノイズによって少なくとも部分的には説明されることを示唆している。

訓練されたモデルが作成した fMRI の予測画像を目視で確認すると， これらの予測画像は， 訓練セット以外の刺激語に関連する脳の活性化のかなりの部分を捉えていることが分かる。 図2B は， 参加者 P1 の 60 個の刺激のうち， "セロリ" と "飛行機" を除いた 58 個の刺激でモデルを学習させた場合の例である。 "セロリ" と "飛行機" の fMRI 画像の予測値は完璧ではないが，この 2 つ の刺激で実際に観測された活性化のかなりの部分を捉えている。 予測された 60 個の fMRI 画像と観察された fMRI 画像の類似性をプロットしたものを 図S3 に示す。

このモデルの予測は，脳のさまざまな場所で異なる精度を示し，入力刺激の意味を符号化する場所ではより精度が高いと考えられる。 図3 は， モデルの「精度マップ」を示したもので， 個々の参加者 (P5) と 9 人 の参加者全体の平均値の両方で， 保留された単語に対するモデルの予測活動が観察された活動と最もよく相関する大脳皮質領域を示している。 これらの高精度ボクセルは， 大脳皮質全体に分布しており， 左半球がより強く表れており， 左下側頭， 豆状体， 運動皮質，頭頂内溝，下前頭，眼窩前頭， 後頭皮質 に現れていた。 このような左半球優位性は， 意味表現において左半球が右半球よりも大きな役割を果たしているという一般的な見解と一致する。 また， 高精度ボクセルは， 両半球の後頭皮質， 頭頂内溝， 下側頭領域の一部にも出現しており，これらの領域も視覚的物体処理に関与していると考えられる。

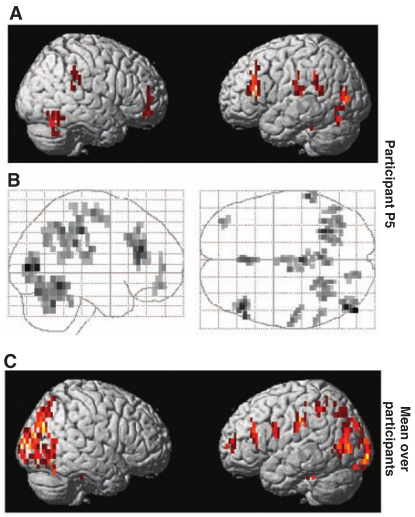



図 3. 最も正確に予測されたボクセルの位置。 参加者 P5 の訓練セット以外の単語について、予測されたボクセルの活性化と実際のボクセルの活性化の相関を表面（A）とグラスブレイン（B）で表したもの。 これらのパネルは、少なくとも 10個 の連続したボクセルを含むクラスタを示しており、それぞれのボクセルの予測-実際の相関は少なくとも 0.28 である。 これらのボクセル・クラスターは、大脳皮質全体に分布しており、左右の後頭葉と頭頂葉、左右の豆状部、中央後葉、中央前葉に位置しています。 左右の後頭葉、頭頂葉、中前頭葉、左下前頭回、内側前頭回、前帯状回に分布している。 (C) 9人の参加者全員で平均化した予測-実測相関の表面表現。 このパネルは、平均相関が 0.14 以上の連続した10 個以上のボクセルを含むクラスターを示している。

このようにして訓練された計算モデルが、訓練セットに含まれていない新たな意味カテゴリの単語に対しても正確な予測を行うことができるかどうかを考えるのは興味深い。 そこで，モデルの再学習を行った。 今回は，2つ のテスト単語のいずれかと同じ意味カテゴリに属する例をすべて学習セットから除外した。 例えば，「セロリ」と「飛行機」のテストを行う場合は，食べ物と乗り物の刺激をすべて学習セットから除外し，50 単語 のみで学習を行った。 この場合、交差妥当性による予測精度は 0.74, 0.69, 0.67, 0.69, 0.64, 0.78, 0.68, 0.64, 0.78 (平均＝0.70) となった。 これは、モデルが学習した単語から意味的に離れた単語まで予測できることから、モデルの意味的特徴とその学習した神経活性化シグネチャが、多様な意味空間をカバーしている可能性を示唆している。

60 個 の刺激が 12 の 意味カテゴリの各 5項目 で構成されていることを考えると、モデルがどの程度まで正確な予測を行うことができるかを判断することも興味深い。 識別が難しいと思われる同じカテゴリーの単語であっても、モデルがどの程度正確な予測を行うことができるかについても興味深いものがある。 例えば セロリ と とうもろこし など。 9 人のカテゴリー内予測精度は, 0.61, 0.58, 0.58, 0.72, 0.58, 0.77, 0.58, 0.52, 0.68 (平均値＝0.62) となり、意味的に類似した刺激を区別する際にモデルの精度は低下するものの、平均的にはチャンスレベル以上の予測が可能であることがわかった。

さらに多様な単語を識別するモデルの能力を検証するために，1000 個の高頻度単語テキストコーパスの中で最も頻度の高い 300 個のトークンを省いた 1300 個のトークン) の間で解決する能力をテストした。 具体的には、60 個の刺激語のうち，59 個の刺激語を用いてモデルを学習させる leave-out-one を行った。 このテストでは，60 個の刺激語のうち 59 個の刺激語を用いてモデルを学習させた後， 残った単語の fMRI 画像 と 1001 個の候補語 (1000個 の頻出トークンと残りの単語) を与えた。 この 1001個 の候補を、まず各候補の fMRI画像 を予測し、次に 1001個 の候補を、予測した fMRI画像 と与えられたfMRI画像の類似度で並べることで、順位付けを行った。 この順位付けされたリストにおける正しい単語の期待されるパーセンタイルランクは、モデルが偶然に動作していた場合、0.50 となる。 9 人の被験者について観測されたパーセンタイルランクは, 0.79, 0.71, 0.74, 0.67, 0.73, 0.77, 0.70, 0.63, 0.76 (平均＝0.72) であり， このモデルが意味的に多様な単語群にある程度適用できることが示された (詳細は（26）参照)。

予測精度の定量的な測定に加えて， 我々の計算モデルを評価する 2 つ目のアプローチは，25 の動詞ベース のシグネチャに対する fMRI シグネチャ の学習ベースセットを調べることである。 この 25 個のシグネチャは， モデルが学習した神経表現を意味的特徴に分解したものであり， モデルのすべての予測の基礎となるものである。 図 4 は eat, push, run という意味上の特徴に対するシグネチャを示したものである。 これらのシグネチャは， それぞれ複数の皮質領域の活性化を予測していることに注目してほしい。

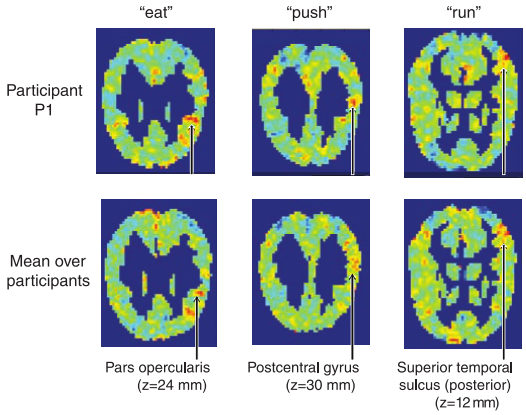



図 4. 被験者 P1 の 25 の意味的特徴 のうち 3 つの特徴について学習したボクセルの活性化信号 (上段) と 9 人 の参加者全体で平均 (下段)。 それぞれ, 水平方向の $z$ スライスを 1 枚だけ示している。 「食べる」という動詞に関連した意味的特徴は, 味覚野の一部と考えられる右のオペクラリス (Pars opercularis) の活動を予測する。 また 「押す」という動詞の意味的特徴は, 運動前計画に関連するとされる右の中心後野を活性化する。 run という動詞の意味的特徴は生物学的運動の知覚に関連するとされる右上側頭溝の後部を活性化する。

図 4 の意味特徴のシグネチャーを見ると, 「食べる」という意味特徴の学習済み fMRI シグネチャーは, 味覚に関わる味覚野の構成要素であると他の研究者が示唆している, 左パネルの矢印で示すように, オペクラ皮質の強い活性化を予測していることがわかる (30)。 また push の学習された fMRI シグネチャーは, 複雑で協調的な動作の計画に関与していると広く想定されている右の中心後野の活性化を予測している (31)。 さらに run の学習されたシグネチャーは, 右側の上側頭葉の後部にある溝に沿った部分の強い活性化を予測する。 これらの学習されたシグネチャにより, 名詞を表す神経活動は, その名詞が動詞 eat と共起するほど味覚野で, push と共起するほど運動野で, run と共起するほど体の動きに関連する皮質領域で活動するとモデルは予測している。 図 4 上段は, 被験者 P1 の学習したシグネチャを示しているが, 下段は 9 人 の参加者が独自に学習した 9 つのシグネチャの平均値を示している。 この 2 列のシグネチャの類似性は, これらの学習された中間的な意味特徴のシグネチャが, 参加者間で実質的な共通性を示すことを示している。

他のいくつかの動詞の特徴も、活性化を予測する皮質領域の機能と動詞の意味との間に興味深い対応関係を示している。 その対応関係が 9人 の被験者の一部でしか成立していない場合もある。 例えば P1 の場合、体性感覚野の活性化を予測する touch のシグネチャ (右後中心回) や、言語処理領域の活性化を予測する listen の シグネチャ (左後上側頭溝, 左三角筋) が追加されている。 これらの傾向は 9人 全員に共通するものではない。 25 個 の意味的特徴すべてに対する学習済み特徴シグネチャを (26) に示す。

基本的な意味特性に対応する神経構成要素が感覚運動動詞に関連しているという推測に基づいて, この 25 個 の中間的な意味特性のセットが成功したことを考えると, この中間的な意味特性のセットが他のものと比較してどうなのかを問うのは自然なことである。 そこで, テキストコーパスに含まれる 5000 語の最頻出単語のうち, 60 の刺激語と 500 の最頻出単語 (多くの機能語や the や have のような特定の意味を持たない単語を含む) を除いた 25 単語 からランダムに生成された意味的特徴のセットを用いて, モデルの学習とテストを行った。 計 115 個のランダムな特徴量セットを作成した。 それぞれの特徴セットについて, 9 人の参加者全員のモデルを学習し, これら 9 つのモデルの平均予測精度を測定した。 予測精度の分布は, 図 5 の青色のヒストグラムに示されている。 この 115 個の特徴量セットの平均精度は 0.60, SD は 0.041, 最小精度と最大精度はそれぞれ 0.46 と 0.68 であった。 最も高い精度と最も低い精度を生み出したランダムな特徴セットを (26) に示す。 平均精度が 0.50 を超えているということは, 多くの特徴量セットが, 60 個の刺激語の意味内容の一部と, それに対応する脳の活性化の規則性の一部を捉えていることを示唆している。 しかし, これらの 115 個の特徴量セットのうち, 手動で生成した特徴量セットの平均精度 0.77 に近いものはなかった。 図 5 のヒストグラムの赤いバーで示されている。 この結果は, 感覚・運動動詞で定義された特徴セットが, 言葉の意味内容をコード化している脳内の神経活動の規則性を捉える上で, やや特徴的であることを示唆している。

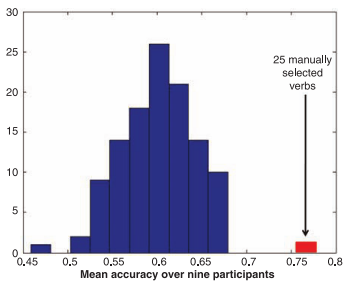



図 5. 別の中間的な意味的特徴のセットに基づくモデルの精度。 115 種類の中間的な意味的特徴のセットをランダムに選んで使用した計算モデルの精度を青いヒストグラムで示す。 それぞれの特徴セットは 5000 個 の最頻出単語から、500個 の最頻出単語と刺激語を除いた 25 個 の単語をランダムに選んだ。 手動で選んだ感覚・運動動詞に基づく特徴セットの精度を赤で示している。 各特徴量セットの精度は, その特徴量セットを 9 人の被験者それぞれのモデルの学習に用いたときに得られた平均精度である。

## 3. 議論

今回報告された結果は、テキスト中の単語の共起の統計量と、単語の意味を考えることに関連する神経活性化との間に、直接的な予測関係を確立したものである。 さらに、これらの予測を行うために訓練された計算モデルは、物体を表す神経活動が、物体の異なる意味的構成要素に関連する神経活性化パターンの基礎セットにどのように分解できるかについての洞察を与えてくれる。

具体的には、25個 の感覚運動動詞を用いたモデルが成功し (25個 の意味的特徴をランダムに抽出したモデルと比較して), 具体的な名詞の 神経表現が感覚運動的特徴に基づいている部分があるという推測に信憑性を与えている。 しかし、25個 の中間的な意味的特徴に

関連した学習シグネチャは、感覚運動機能に直接関連しない脳領域 (前頭領域を含む) でも有意な活性化を示した。 このように、具体的な名詞の神経表現を支える特徴の基礎セットには、感覚運動皮質領域以外にも多くの領域が関与していると考えられる。

また，最近の研究では，具体的な物体を表現する神経エンコーディングは，少なくとも部分的には個人間で共有されていることが示唆されている。 これは，ある人が見ている複数のアイテムのうち，どのアイテムを見ているかを，その人のfMRI画像と他の人から学習した分類モデルだけで識別できるという証拠に基づいている(34)。 今回の結果は、学習された意味的特徴の基礎セットにも個人間の共通点があることを示しており、神経表現のどの要素が個人間で似ていて、どの要素が違うのかをより直接的に判断するのに役立つと考えられる。

我々のアプローチは，絵刺激の低レベルの視覚的特徴に着目して，絵を見たときの fMRI活性化を解析する研究 (9, 35, 36) や，fMRI 活性化に基づいて， 物体の形状間の知覚された類似性を比較する研究 (37) に類似している。 最近の研究 (36) では、任意のシーンの視覚的特徴に基づいて、視覚野の一部におけるfMRI活性化の側面を予測し、この予測された活性化を用いて、個人が見ているシーンの候補のどれかを特定することが可能であることが示されている。 本研究は、これらの研究とは異なり、より抽象的な意味概念の符号化に焦点を当て、知覚的な側面を示す視覚的特徴ではなく、刺激となる単語の意味的な側面を示すテキストコーパスの特徴に基づいて、脳全体の fMRI活性化 を予測するものである。 我々の研究は、機械学習アルゴリズムを用いて、fMRIデータ に基づいて精神状態の分類法を学習する最近の研究 (38, 39) とも関連している。 我々のモデルは、学習セットに存在しない精神状態の fMRI画像 を予測するために外挿することができるという点で異なっている。

本研究は，脳内の神経表現を研究するパラダイムの転換を意味している。 特定のカテゴリーの単語や絵に関連する fMRI 活動のパターンをカタログ化する研究から， 任意の単語 (fMRIデータがまだ得られていない何千もの単語を含む) の fMRI 活動 を予測する計算モデルを構築する研究へと移行している。 これは，この分野がデータの理論的なカタログ化から計算モデルの開発、そして神経表現の理論の始まりへと移行していく中で、自然な流れである。 我々の計算モデルは， 「言葉の意味を表す fMRI 神経活動の予測値は何か」， 「具体的な名詞の意味を表す神経活動を説明する意味的特徴の基本セットとそれに対応する神経活動の成分は何か」 といった疑問に答える， 限定された形の予測理論を符号化したものと見ることができる。 脳が感覚入力からこれらの表現をどのように合成しているかを説明する因果関係のある理論にはまだ遠いが， これらの質問に対する答えでさえ， 意味の神経表現の基礎となる重要な規則性のいくつかを明らかにすることが期待される。

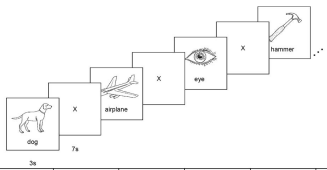
# 付録

## A 1. 材料と方法

### A 1.1 fMRI データの収集と処理

カーネギーメロン大学生右利き成人 9 名 (女性 5 名， 年齢 18 歳から 32 歳) は， ピッツバーグ大学およびカーネギーメロン大学の Institutional Review Board で承認されたインフォームドコンセントを得て， fMRI 研究に参加した。 なお， 頭部の動きが 2.2 mm と 3.0 mmだった2 名の被験者のデータは除外した。

刺激は，図 S1 に示すように，12 の意味カテゴリの 60 個の具体的な物品の線画と名詞ラベルで， カテゴリごとに 5 個の図版があった。 線画のほとんどは Snodgrass and Vanderwart のセット (図 S1) から引用または適応し， その他は同様の描画スタイルで追加した。 60 個の刺激項目セット全体を 6 回提示し， 各提示において 60 個のアイテムの順序をランダムに入れ替えた。 各刺激項目は，3 秒間提示された後，7 秒間の休息時間があり， その間， 被験者は画面中央に表示された X に固視するよう指示された。 基準となる測定値を得るために，31 秒ずつ 12 回の X の注視が行われた。




| Category | Exemplar 1 | Exemplar 2 | Exemplar 3 | Exemplar 4 | Exemplar 5 |
|---|---|---|---|---|---|
| animals | bear | cat | crow | dog | horse |
| body parts | arm | eye | foot | hand | leg |
| buildings | apartment | barn | church | house | igloo |
| building parts | arch | chimney | closet | door | window |
| clothing | coat | dress | pants | shirt | skirt |
| furniture | bed | chair | desk | dresser | table |
| insects | ant | bee | beetle | butterfly | fly |
| kitchen utensils | bottle | cup | glass | knife | spoon |
| man made objects | bell | key | refrigerator | telephone | watch |
| tools | chisel | hammer | pliers | saw | screwdriver |
| vegetables | carrot | celery | corn | lettuce | tomato |
| vehicles | airplane | bicycle | car | train | truck |

図 S1. 実験に使用した例題の提示とセット。 実験参加者には，一般的な具体的な名詞を表す 60 の異なる単語と画像対が提示された。 これらは， 上述のように 12 のカテゴリーからそれぞれ 5 つの模範例で構成されていた。 遅緩事象関連パラダイムを採用し， 刺激を 3 秒間提示した後，7 秒間の固視時間を置き， その間にスクリーンの中央に X を提示した。 画像は， 暗い背景に白い線と文字で提示されたが， ここでは読みやすさを向上させるために反転させている。 60 個の模範例を 6 回提示し，各回の提示で順序をランダムに入れ替えた。

模範例が提示されると， 被験者はその物体の特性を考えることが課題となった。 6 回の提示で一貫した特性を考えるようにするために，スキャンセッションの前に，各アイテムの特性セットを作成するように求められた (例えば， アイテム「城」の場合，特性は「寒さ」「騎士」「石」となる)。 各参加者は好きな特性を自由に選ぶことができ， 特性の選択において参加者間の一貫性を得ようとはしなかった。

カーネギーメロン大学とピッツバーグ大学の脳画像研究センターに設置されたシーメンス社 (ドイツ， エアランゲン) の Allegra 3.0T スキャナーで，TR=1000 ms, TE=30 ms, フリップアングル 60° のグラディエントエコーEPI パルスシーケンスを用いて，機能画像を取得した。 17 枚の 5 mm 厚の斜め軸スライスを， スライス間のギャップ 1 mm で撮影した。 撮影マトリクスは 64x64 で，3.125 mm x 3.125 mm x 5 mmのボクセルを用いた。

初期のデータ処理は，Statistical Parametric Mapping ソフトウェア (SPM2，Wellcome Department of Cognitive Neurology, London, UK) を用いて行った。 データはスライスのタイミング，動き，線形傾向を補正し，190 秒のカットオフを用いて時間的にフィルタリングした。 データは，MNI 空間に空間的に正規化され，3x3x6 $mm^3$ のボクセルにリサンプリングされた。 刺激の提示ごとに，各ボクセルで固定条件に対する信号変化率（PSC）を算出した。 刺激開始から 4 秒後，5 秒後，6 秒後，7 秒後に採取した画像の平均値を取ることで，360 個のアイテム提示ごとに 1 つの fMRI 平均画像を作成した (血行力学的反応の遅延を考慮した)。

### A 1.2 テキストコーパスデータ

テキストコーパスデータは，Google Inc. から提供されたもので，オンラインで http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13 で公開されている。 1 兆個以上の総トークンを含む大規模なコーパスの中に 1 gram (1 つのトークン) から 5 gram (5 つのトークンの系列) までの n-gram (単語やその他のテキストトークンの系列）のセットと，各 n-gram の出現回数を示すカウントで構成されている。 このコーパスは，一般に公開されている英語テキストのウェブページで構成されています。 40 回未満しか出現しない N-gramは 提供されませんでした。 このデータを用いて，5 トークン以内に出現した単語の共起回数を計算した。 このデータは，本論文で報告されているすべての実験で使用されている共有カウントである。

### A 1.3 モデルの訓練

意味的特徴 $f_i(w)$ を指定した後，i 番目の意味的特徴が v 番目のボクセルに寄与する神経シグネチャを定義するパラメータ $c_{vi}$ を推定する。 これは，既知の刺激語に関連して観測された fMRI 画像のセットを使ってモデルを訓練することで達成される。 各訓練刺激 $w_t$ は，まず，その特徴ベクトル $< f_1(w_t) \ldots f_n(w_t) >$ *で再表現され，次に，重回帰を用いて，*{vi}cvi 値の最尤推定値，すなわち，訓練 fMRI 画像を再構成する際の二乗誤差の合計を最小化する $c_{vi}$ 値のセットを得る。 意味的特徴の数が訓練例の数よりも少ない場合，この重回帰問題はよく提起され，一意の解が得られる。 意味的特徴の数が訓練例の数よりも多い場合には，学習した回帰重みの二乗和に等しいペナルティなどの正則化項を導入することで解が得られる。

一度学習された計算モデルは，本文の図 2A に示すように，1 兆 ($10^{12}$) トークンのテキストコーパスに含まれる他の任意の単語の fMRI 活性化画像を予測するために使用することができる。 任意の新しい単語 $w_{new}$ が与えられた場合，モデルはまずコーパス統計データベースから中間的な意味的特徴値 $< f_1(w_{new}) \ldots f_n(w_{new}) >$ を抽出し，次にパラメータ $c_{vi}$ について以前に学習した値を用いて上の式を適用する。 計算モデルとそれに対応する理論は，訓練セット外の単語に対する予測値と，その単語に関連して観測された fMRI 画像を比較することで，直接評価することができる。 事前に定義された異なる中間的な意味的特徴のセットは，競合するモデルを学習し，その予測精度を評価することで直接比較することができる。

60 の刺激語のそれぞれに対する中間的な意味的特徴ベクトルの詳細なリストは，www.cs.cmu.edu/~tom/science2008 に掲載されている。

### A 1.4 計算モデルの訓練と評価

異なった中間的な意味的特徴セットに基づいて，別の計算モデルが学習された。 この方法では，60 個の刺激項目のうち 58 個の刺激項目だけを使ってモデルを繰り返し学習し，次に，残った 2 個 の刺激項目を使ってテストを行う。 各反復において，学習されたモデルは，まだ見たことのない 2 つ の刺激語 ($w_1$ と$w_2$) と，それらが観測された fMRI 画像 ($i_1$と$i_2$) を与えられてテストされ，次節で説明するマッチング手順を用いて，2 つの新規画像のどちらが 2 つの新規語のどちらと関連しているかを予測することを要求された。 この放置型トレーニングテストを 1770 回 繰り返し，可能性のある単語対をそれぞれ放置した。 偶然レベルでマッチングが行われた場合，取り残された 2 つ の単語と取り残された fMRI 画像 のマッチングの期待される精度は 0.50 である。

### A 1.5 予測値と実際の画像のマッチング

訓練された計算モデル，2 つの新しい単語 ($w_1$ と $w_2$)，2 つの新しい画像 ($i_1$ と $i_2$) が与えられると，まず訓練されたモデルを用いて，単語 $w_1$ に対する予測画像 $p_1$ と単語 $w_2$ に対する予測画像 $p_2$ が作成された。 そして、どちらがよりマッチしているかを判断した。 ($p_1 = i_1$ and $p_2 = i_2$) と($p_1 = i_2$ and $p_2 = i_1$) のどちらがよりマッチするかを，最も類似度の高い画像ペアを選択することで決定した。 脳内のすべてのボクセルが刺激の意味を表現することは期待できないので，画像間の類似性を評価するために，ボクセルのサブセットのみを使用した。 このボクセルのサブセットは，トレーニング時に 58 訓練単語のデータのみを用いて自動的に選択され，2 つのテスト単語のデータは除外された。 ボクセルの選択方法は以下の通りである。 sel(i) を画像 $i$ の選択されたボクセルのサブセットの値のベクトルとする。 予測された画像 $p$ と観測された画像 $i$ の間の類似性スコアは，ベクトル sel(p) と sel(i) のコサイン類似度として計算された。 2 つのベクトル間のコサイン類似度は，ベクトル間の角度のコサインとして定義され，単位長さに正規化されたこれらのベクトルのドット積として計算された。 最後に，予測画像と実際の画像のペア候補 (例: $p_1 = i_2$ と $p_2 = i_1$）の類似度マッチスコアを，2 つのコサイン類似度の合計として計算した。

$$\text{match}\left(p_1 = i_2 \text{ and } p_2 = i_1\right) = \text{cosine Similarity}\left(\text{sel}(p_1), \text{sel}(i_2)\right) + \text{cosine Similarity}\left(\text{sel}(p_2), \text{sel}(i_1)\right).$$

最初に検討した類似性指標はコサイン類似性だったが，その後，2 つの画像間のピアソン相関も検討した。 2 つの画像間のピアソン相関も検討したが，同様の結果が得られた。 本論文では、コサイン類似度を用いている。

### A 1.6 ボクセルの選択

上述のように，2 つの画像間の類似性は，画像のボクセルのサブセットのみを用いて計算された。 ボクセル選択は，学習時に，58 個の学習語を用いて自動的に行われ (leave-two-out cross validation folds)。. ボクセルを選択するために，まず，58 個の訓練刺激をそれぞれ 6 回提示したときのデータを用いて，すべてのボクセルに「安定性得点」を割り当てた。 この 6×58＝348 回の呈示が 348 枚の fMR I画像として表現されている場合，各ボクセルには 6×58 の行列が割り当てられ，$i$ 行j列目のエントリは，$j$ 番目の単語の $i$ 回目の呈示時のこのボクセルの値となる。

次に，このボクセルの安定性得点を，この行列のすべての行の対の平均相関として計算した。 これにより，58 個の訓練刺激に対して一貫した (異なる呈示の間で) 活動の変化を示すボクセルに最高の得点が割り当てられる。 例えば，あるボクセルが各提示で同じ 58 の反応を示した場合，平均的なペアワイズ相関は 1.0 となる。 もちろん，fMRI の活動に固有のノイズがあるため，実際にはこのようなことは起こらず，高いペアワイズ相関が見られるのは，このノイズを凌駕するような，強くて再現性のあるボクセル応答パターンの信号がある場合に限られる傾向がある。 なお，58 個の刺激のうち，いくつかの刺激で同じように活性化しているボクセルでも，58 個の刺激のうち，少なくとも他のサブセットで異なる活性化をしていれば，高いペアワイズ相関が発生する可能性がある。 この安定性スコアで上位にランクされた 500 個のボクセルは，上述のコサイン類似性テストに使用された。 選択された個々のボクセ

ルは， 刺激の一部のサブセットの間でのみ区別されるかもしれないが， この方法で選択されたボクセルのセット全体は， 報告された結果から明らかなように， 実際にはすべての刺激の間でかなりよく区別される傾向がある。

### A 1.7 統計的有意性と p 値を決定する経験分布

情報を持たないモデルが，訓練セット外の 2 つの刺激とその 2 つの fMRI 画像を正しくマッチングする期待される偶然の精度は 0.5 である。 しかし，1770 回の繰り返しを行った結果，訓練モデルの精度は 0.5 よりも高くなった。 ここでは， 訓練されたモデルの真の精度が 0.5 であるという帰無仮説を棄却するために， 観測された精度に基づいてどのように p 値を決定するかという問題を考える。 しかし， 訓練とテストを 2 回ずつ行う方法では， このような p 値を求める閉形式は存在しない。 そこで，真の精度が 0.5 に非常に近いと予想される 768 の独立して訓練された単一被験者モデルから得られた観測された精度の経験的な分布に基づいて，p 値を計算した。 これらの帰無仮説の精度の経験分布は 0.501, 標準偏差 0.070 であり, 単一被験者モデルの観測精度が 0.62 以上であれば $p < 0.05$ で統計的に有意であることを示している。 以下では，我々のアプローチをより詳細に説明する。

この経験的な精度分布は，60 個の刺激語について観測された fMRI 画像を用いて， 異なる単語ラベルと異なる中間的な意味的特徴を用いて複数のモデルを学習することで作成した。 この方法は， 並べ替えテストに似ているが， 60 個の刺激ラベルを並べ替えるのではなく， テキストコーパスのトークンの語彙から 60 個の新しい単語を選んだ。 各モデルの学習は， まず，9 つの参加者データセットの 1 つを一様にランダムに選び， 次に， テキストコーパスの 500 から 5000 の最頻値の単語から一様にランダムに 60 個の単語を選び， 次に， コーパスの 500 から 5000 の最頻値の単語から一様にランダムに 25 個の中間的な意味特徴の単語を選んだ。 そして，ランダムに抽出された 60 個の単語を 60 個の正しい単語ラベルに置き換え，ランダムに抽出された 25 個の中間的な意味特徴語を用いて，モデルの学習とテストを行った。 モデルの学習とテストは， 他の論文と全く同じように leave-two-out テスト方式で行われたが， 1 つだけ例外があった。 これらのモデルでは， 最も安定した 500 個のボクセルが 60 個の単語すべてのデータを使って選択されたが， 他の論文では， この安定したボクセルの選択は 58 個の訓練単語のみに基づいていた。 この例外は， このようなランダムなモデルを何百個も学習する際の扱いやすさを劇的に向上させ， 1000 倍のスピードアップにつながるために行われた。 このようにして評価されたランダムモデルの期待される観測精度はわずかに正のバイアスがかかり， 結果として得られる分布から計算される p 値はわずかに保守的になるという正味の効果がある。 実際， この方法で学習・テストされたモデルの経験的な平均精度は 0.501 で， 期待される偶然の精度である 0.500 に非常に近いため， このバイアスは非常に小さいことがわかった。

このようにランダムに生成された 768 個のモデルを学習・テストした。 この 768 個のモデルの平均精度は 0.501 で， 標準偏差は 0.070 であった。 観測された精度の分布を Figure S2 に示す。 この累積分布を見ると，95 %のモデルの精度が 0.621 以下であることがわかり，0.621 以上の精度を持つ被験者モデルには p<0.05 の p 値が割り当てられていることがわかる。 一貫性の確認として， 経験的な精度分布を $\mu = 0.501$, $\sigma = 0.070$ のガウスモデルとし， ガウスの累積分布に基づいて，$p < 0.05$ は $\mu + 1.645\sigma = 0.617$ 以上の精度に対応することがわかった。 これは経験的な累積分布から得られた 0.621 に非常に近い値である。 この同じガウスモデルの下では， 一人参加型モデルの精度が 0.719 であれば， $p < 0.001$ で有意となる。 上記の分析は，1 人の参加者に対して訓練された 1 つのモデルの精度に適用されることに注意。 9 人の独立して訓練された参加者モデルがすべて 0.62 以上の精度を示すことを観察する際の p 値は $p < 10^{-11}$である。

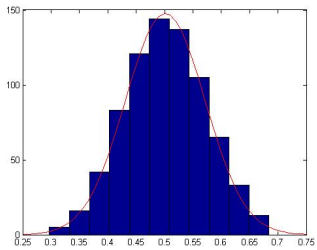



図 S2. 図S2. ヌルモデル精度の経験分布とガウス近似。 青色のヒストグラムは，768 個のランダムに生成された一人参加型のヌルモデルの観測された精度を示している(平均=0.501, 標準偏差=0.070)。 赤い線は， この平均値と標準偏差を持つガウス分布を示している。 この経験的な分布を用いて， 観測されたモデルの精度に対する p 値を決定した。

### A. 1.8 本論の図 3 の精度マップの計算

本文中の図 3 の精度マップは， 予測値と実際のボクセル値の相関が最も高いボクセルクラスターを示している。 まず，60 個の単語に対する 60 個の予測画像を計算し， 残りの 59 個の単語に対してモデルを訓練し， これを使って残りの単語を予測した。 これにより， 各ボクセルに対して 60 個の予測値が得られた。 各ボクセルの精度得点は， この予測値のベクトルと， 対応する観測値のベクトルとのピアソン相関として計算された。 これらのボクセル得点を含むイメージマップを作成し， 標準的な SPM ツールを用いて，スコアがしきい値 (図 3A および 3B では 0.28， 図 3C では 0.14) よりも大きい連続した 10 個以上のボクセルを含むクラスターを特定し， 図 3 に示すクラスターを作成した。

## A 2. 追加結果と観測

### A 2.1 ランダム生成意味特徴を用いた実験

今回の実験では，本文の 図 5 にまとめられているように， 無作為に選ばれた 25 単語で特徴を定義した。 この 25 単語はテキストコーパスの中で最も頻繁に出現する 5000 トークン から一様にランダムに選ばれ， 最頻 500 語のトークン (the や of などの多くの機能語を含む) と 60 個の刺激名詞は除外された。 モデルの学習とテストは， 手動で選択した 25 個の動詞 について説明したのとまったく同じように行われた。 だが，1 つの例外があり， このランダムに生成された特徴で学習したモデルの測定精度にわずかに楽観的なバイアスがかかった。 各交差妥当性検証において，58 個の訓練単語を用いてボクセル選択を行う代わりに，60 個の単語すべてを用いて， 各参加者と特徴セットに対して 1 回だけボクセルを選択した。 この変更は，モデルの訓練と テストの計算コストを削減し，より多くの種類のランダムに生成された特徴セットを探索できるようにするために導入したものである。 上述の「統計的有意性と p 値を決定する経験分布」の項で述べたように，この方法でボクセル選択を行ったことによる観測された精度の正のバイアスは無視できると推定される。

手動で生成した 25個 の意味的特徴と比較して，ランダムに生成された特徴セットには 2つ の違いがある。 まず，手動で生成した特徴量がすべて動詞であったのに対し，ランダムに生成した特徴量には，形容詞，動詞，副詞，名詞，固有名詞，俗語，ウェブ上でよく

見られる英単語とは思えないトークン (html など) など， あらゆる種類のトークンが含まれている。 次に， 動詞の特徴は，動詞の 3 つの形を用いて定義したが (例えば 動詞 eat の特徴は， 動詞 eat, eats, ate の 3つ の形との共起の合計を用いた)， ランダムに選択されたトークンをそのような関連するトークンのセットに展開しようとはしなかった。 一般的に， 任意の単語のトークンを自動的に類似した方法で展開する明白な方法はなく， 多くの単語 (例: partly, news) については、手動でもその方法は不明である。

ランダムに生成されたそれぞれの特徴セットについて，9 人の参加者それぞれについてモデルを学習した。 ランダムに生成された 115 個の素性セットのうち，9 人の参加者の間で達成された最大の平均精度は 0.68 であり，手動で選択された 25 個 の動詞の 0.77 と比較しても遜色ないものであった。 この 0.68 の精度を達成したランダムに選ばれた 25 個の特徴語は， seems, productions, lots, various, counts, seek, lab, arizona, body, pieces, drop, disabled, lol, venture, finally, arts, eating, infrastructure, xml, nikon, ericsson, partly, governments, ladies, and ft である。 参加者 9人 の平均精度が最も低かった特徴セットの精度は 0.46 であった。 この特徴セットは: outcome, sessions, schedule, failure, characteristics, statistics, med, beauty, mt, alternative, richard, responsible, god, parties, candidates, towards, governments, fred, father, seeking, kim, hunt, xxx, keeps, and summary. というトークンを使用した。 精度の高い特徴セットと低い特徴セットを比較したところ， 明らかな規則性は見られなかった。

### A 2.2 学習した特徴シグネチャ

本文中の図 4 は， 参加者 P1 の特徴標識の一部を示しており，9 人の参加者の平均値を示している。 どの参加者にも存在しないボクセルは， 参加者全体の平均値を示す画像から除外した。 参加者 P1 および 9 人の参加者で平均化した 25 個の特徴シグネチャの全セットは， www.cs.cmu.edu/~tom/science2008 に掲載されている。

### A 2.3 予測画像と実際の画像の類似性のプロット

訓練された計算モデルの能力をより理解するため， 図S3 は， 参加者 P1 について， 上述のように最も安定した 500 個のボクセルを用いて， 60 個の予測画像のそれぞれと 60 個の観察画像のそれぞれの間のコサイン類似度得点を描いている。 ここで，i 行 j 列目のエントリは， どちらの単語も使用せずに訓練したモデル (他の 58 単語で訓練) を使用し，刺激語 i の予測画像と単語 j の観察画像の間のコサイン類似度を示している。 したがって，この図には， 訓練セット以外の単語対間の類似性得点のみが含まれている。 対角線上の高い正の値は， 単語レベルでの予測が正しいことを示している。 対角線上のブロックの高い値は，同じ意味的カテゴリーの画像間の類似性を反映している。 また， 紺色の領域は， カテゴリー A の単語に対する予測画像がカテゴリー B と大きく異なる (負のコサイン類似度を持つ) カテゴリー対を示している。 図 S3 は， 参加者 P1 の類似度を示しているが， 図 S4 は，9 人の参加者全体の平均的な類似度を示している。

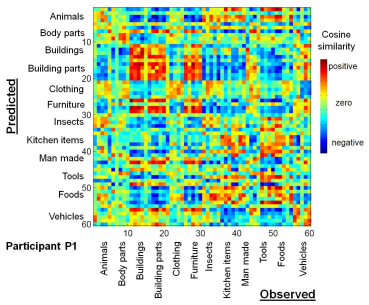



図 S3. 参加者 P1 の予測画像と実際の画像のコサイン類似度。 i 行j列目の点は， 他の 58 個の単語で学習したモデルを用いて， i と j の単語を除いた場合に， i の単語に対して予測した画像と j の単語に対して観察した画像のコサイン類似度を示している。 対角線上にある高い正の値は， ある単語の予測画像がその単語の観察画像と似ていることを示している。

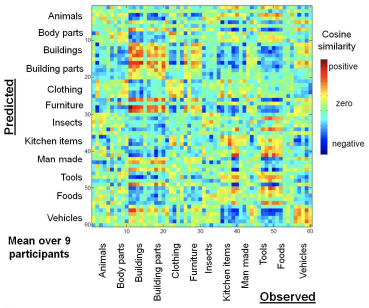



図 S4. 予測画像と実際の画像のコサイン類似度 (全参加者の平均)。 この図は， 図 S3 と同じ規則に従うが， 9 人の参加者で平均した予測画像と観察画像の平均類似度を反映している。 対角線上の値の平均は 0.179 であるのに対し， 行列全体の平均は -0.016 であり， 平均的に予測画像が他の画像よりも実際の画像に類似していることを示している。 また， 行列の最大値 (最も赤い値) は 0.65，最小値 (最も青い値) は -0.60 である。

図 S4 のエントリを見ると， 類似度得点が 60 の候補のうち正しい単語を平均的にどの程度解決しているかを判断することができる。 具体的には， 各行には， 予測された単語の画像と， 60 個の観測された画像 (それぞれ，比較される 2 つの単語を省略したモデルによって計算されたもの) との類似性得点が示されている。 各行の類似度スコアを類似度の高いものから低いものに並べると， 正解の単語のスコアは平均で 79 パーセンタイル目に表示され，60 個の単語の中で特徴が解決する画像を予測するモデルの能力が不完全ながらも高いことがわかる。 60 個の単語のそれぞれについて， 正しい画像のパーセンタイルランクを以下に示す。 ここでの単語は， 図 S3, S4, S5 での位置に応じて番号が付けられている。

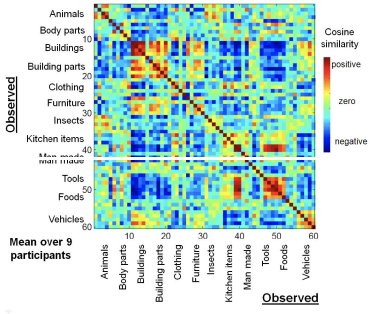



図 S5. 実際の画像間のコサイン類似度 (全参加者の平均値)。 この図は， 図 S3 および S4 と同じ規則に従うが， 観察された画像対間の平均類似度を反映していることを除いて，9 人の参加者で平均化している。 対角線上のブロックが高い値を示しているのは， 同じ意味上のカテゴリーの画像間の類似性を示している。 類似度が 1.0 である対角線上のエントリを無視すると， 対角線外の最大値は 0.52, 最小値は -0.41 となる。

1. 0.283 bear
2. 0.767 cat
3. 0.517 cow
4. 0.950 dog
5. 0.950 horse
6. 0.750 arm
7. 0.583 eye
8. 0.933 foot
9. 0.883 hand
10. 0.833 leg
11. 0.917 apartment
12. 0.950 barn
13. 0.950 church
14. 0.950 house
15. 0.400 igloo
16. 0.900 arch
17. 0.933 chimney
18. 0.983 closet
19. 0.967 door
20. 0.983 window
21. 0.850 coat
22. 0.967 dress
23. 0.933 pants
24. 0.850 shirt
25. 0.867 skirt
26. 0.717 bed
27. 0.783 chair
28. 0.833 desk
29. 0.833 dresser
30. 0.550 table
31. 0.867 ant
32. 0.900 bee
33. 0.917 beetle
34. 0.317 butterfly
35. 0.783 fly
36. 0.983 bottle
37. 0.817 cup
38. 0.983 glass
39. 0.900 knife
40. 0.967 spoon

41. 0.383 bell
42. 0.267 key
43. 0.867 refrigerator
44. 0.283 telephone
45. 0.867 watch
46. 0.883 chisel
47. 0.833 hammer
48. 0.933 pliers
49. 0.067 saw
50. 0.967 screwdriver
51. 0.783 carrot
52. 0.767 celery
53. 0.950 corn
54. 0.567 lettuce
55. 0.150 tomato
56. 0.867 airplane
57. 0.983 bicycle
58. 0.883 car
59. 0.983 train
60. 0.983 truck

予測結果が最も悪かった単語は saw (単語49) である。 これは saw をツールとして被験者に提示したにもかかわらず, モデルが使用した saw のトークン共起数は, より頻繁に使用される動詞 (see の過去形) としての使用に支配されていることが主な原因である。 このことから, 将来的には, 単語の意味を区別するコーパスの特徴を強化することで, さらに精度の高いモデルを実現できる可能性がある。

図 S5 は, 図 S4 と比較するために, 60 個の観察画像間の類似性を示したものである (したがって, 学習したモデルではなく, データの要約である)。 具体的には，i 行j列目のエントリは，9 人の参加者における, その参加者の単語 i と j の観察画像間の類似性の平均値を示している。 図 S5 と図 S4 を比較すると, 予測画像と実際の画像が混同している部分 (非対角線上の赤と黄色のエントリ) は，2 つの刺激について実際に観察された画像が類似していることによるものであり, 他の混同は, 実際の画像に存在する差異を予測できないというモデルのエラーを反映していることがわかる。 例えば，図 S4 の「家具」と「建物の部品」の予測画像と観察画像の類似性は，図 4に見られるように, これらの物体の神経符号化間の実際の類似性によるものであると考えられている。 図 S3, S4, S5 を比較する際には，色の尺度を各図に合わせてカスタマイズし，最も明るい赤を行列の最大値に, 最も暗い青を最小値に設定していることに注意。

### A 2.4 1000語の候補の中から解決する

本論文で述べたように, 我々は leave-one-out テストも行った。 このテストでは, 60 個の利用可能な刺激のうち 59 個の刺激を用いてモデルを繰り返し学習させ, 1001 個の候補語の中から, どの候補がホールドアウトされた fMR I画像を生成した可能性が最も高いかによって順位付けをさせた。 この順位付けは, ホールドアウトされた fMRI 画像と各候補語の予測画像とのコサイン類似度に基づいて行われた (通常通り，訓練データの中で最も安定した 500 個のボクセルのみを使用)。 今回の実験では, テキストコーパスの中で最も頻度の高い 1300 のトークンを使用し, 最も頻度の高い 300 トークン (for や the などの機能語を多く含む) を省いた。 本論で述べたように, モデルのランク付けされたリストにおける正しい単語の平均パーセンタイルランクは, 9 人の参加者全員の平均で 0.72 であった。 これは, ほとんどの単語がかなり高くランク付けされており, 少数の単語が非常に低くランク付けされていることを反映している。 下のリストは, 9 人の参加者全員の平均パーセンタイルランクでソートされた 60 個の単語のリストある (各単語の隣にある数字はソートされた候補リストの中で, の単語が正解候補の単語であった場合の平均パーセンタイルランク)。 見てわかるように glass のようないくつかの単語は , 全参加者の平均で非常に正確に予測されており, 1000 個の候補のうち 26 個だけがテスト fMRI 画像を生成した可能性が高いと平均的にランク付けされた。 一方, saw や bear などの単語は, 平均して非常に低い順位となった。 以下の正確な順位と不正確な順位の単語は, 図 S4 に関連する上のリストの正確な順位と不正確な順位の単語と高い相関があることに注目。 この 2 つのリストの違いは，下のリストでは, 1001 個の予測画像を, 保持された単語について観測された 1 つの fMRI 画像との類似性でランク付けしていることである。 一方, 上のリストでは, 60 枚の観察された fMRI 画像を, ホールドアウトされた単語に対する 1 枚の予測画像との類似性で順位付けしている。

1. 0.974 glass
2. 0.955 chimney
3. 0.914 church
4. 0.905 train
5. 0.898 bicycle
6. 0.890 dress
7. 0.889 closet
8. 0.889 screwdriver
9. 0.886 foot
10. 0.884 bottle

11. 0.878 arch
12. 0.868 house
13. 0.856 airplane
14. 0.852 horse
15. 0.851 door
16. 0.849 spoon
17. 0.846 barn
18. 0.837 window
19. 0.825 hammer
20. 0.824 knife
21. 0.822 chisel
22. 0.821 car
23. 0.819 dresser
24. 0.814 skirt
25. 0.810 truck
26. 0.802 leg
27. 0.799 hand
28. 0.796 refrigerator
29. 0.796 bee
30. 0.792 dog
31. 0.791 cup
32. 0.775 watch
33. 0.771 apartment
34. 0.769 pants
35. 0.765 pliers
36. 0.751 desk
37. 0.743 bed
38. 0.743 coat
39. 0.738 corn
40. 0.732 shirt
41. 0.718 carrot
42. 0.703 chair
43. 0.702 ant
44. 0.673 fly
45. 0.668 celery
46. 0.628 arm
47. 0.585 cat
48. 0.585 beetle
49. 0.570 table
50. 0.533 eye
51. 0.512 bell
52. 0.512 key
53. 0.476 cow
54. 0.453 lettuce
55. 0.434 igloo
56. 0.345 tomato
57. 0.307 butterfly
58. 0.295 telephone
59. 0.242 bear
60. 0.171 saw

### A 2.5 意味的特徴を定義するための共起回数の使用に関する注記

共起回数を使って単語や文書の意味内容を近似することは，計算言語学では一般的な手法だが，これはいくつかの欠点を持つ粗いアプローチである。1 つは，指定されたウィンドウ内での単純な共起が，2 つの単語の間の構文的な関係を解決できないという事実に起因する。例えば，"The mouse ate the cheese" と "The cat ate the mouse" では，"mouse " と "ate" の関係が大きく異なる。例えば，名詞が主語である場合と，名詞が共起する動詞の直接目的語である場合などは，共起回数では解決できない。また，多くの単語には複数の意味があるが，このアプローチではこれらを解決できない。例えば saw という単語は，名詞 (道具) を指すこともあるが，一般的には動詞 (see の過去形) を指すことが多く，その結果，意味的特徴ベクトルは，道具としての意図された意味を代表しておらず，結果的にこの単語の予測がうまくいあなかった。これらの欠点にもかかわらず，大規模コーパスから収集された共起データは，妥当なモデルをサポートするのに十分なほど，刺激語の意味を捉えているように見える。将来的には，より洗練された言語的特徴を考慮することで，より強力なモデルを開発できると考えている (例えば，動詞と名詞の関係を決定するために文を解析したり，異なる語義を自動的に解決したりする)。

### A 2.6 追加のオンライン資料の利用について

追加情報は www.cs.cmu.edu/~tom/science2008 で公開されている。本論文発表時点で，このサイトで入手可能な追加情報は，60 個の刺激語それぞれの中間意味特徴ベクトルの詳細リスト，参加者 P1 の 25 個の意味特徴シグネチャ (本論文の図 3 に示されているものと同様) の表示，および 9 人の参加者全体で平均化された 25 個の意味特徴シグネチャの表示である。