Generating descriptive text from functional brain images

Francisco Pereira[†] Greg Detre[†] Matthew Botvinick[†]

Summary Recent work has shown that it is possible to take brain images of a subject acquired while they saw a scene and reconstruct an approximation of that scene from the images. Here we show that it is also possible to generate *text* from brain images. We began with images collected as participants read names of objects (e.g., “Apartment”). Without accessing information about the object viewed for an individual image, we were able to generate from it a collection of semantically pertinent words (e.g., ”door,” ”window”). Across images, the sets of words generated overlapped consistently with those contained in articles about the relevant concepts from the online encyclopedia Wikipedia. The technique described, if developed further, could offer an important new tool in building human-computer interfaces for use in clinical settings.

[†]Psychology Department and Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540

Introduction Over the last decade, functional magnetic resonance imaging (fMRI) has become a primary tool for identifying the neural correlates of mental activity. Traditionally, the aim of fMRI experiments has been to identify discrete, coherent neuroanatomic regions engaged during specific forms of information processing. More recently, it has become clear that important information can be extracted from fMRI by attending instead to broadly distributed patterns of activation. The application of machine learning techniques for pattern classification [20] has enabled impressive feats of "brain reading," making it possible to infer the class of object viewed by an experimental participant, to track the process of memory retrieval, to predict decisions or mistakes, or even (controversially) to detect lies [5 8 13 17].

The key step in "brain-reading" applications of fMRI involves classifying brain images into a set of discrete categories. For example, given a brain image collected during single-word reading, the task might be to decide which among a set of candidate words triggered the image [13]. This approach, which continues to be highly fruitful, has recently benefitted from the application of sophisticated models that allow prediction of brain activation patterns induced by stimuli from outside the initial training set [9 14].

In a dramatic departure from the standard approach, a small set of recent studies has demonstrated the feasibility of a *generative* approach to fMRI decoding. Beginning with fMRI data collected as participants viewed complex images, Naselaris [16] and colleagues constructed entire images that strikingly resembled the original stimuli (see also [15 22]). The crucial ingredient in this approach is a generative model, which captures the way in which specific aspects of the stimulus give rise to particular sub-patterns of distributed brain activity. Once established, this generative model can then be *inverted*, in order to synthesize a complex artifact (e.g., a picture) from a single pattern of brain activity, as diagrammed in Figure 1.

To date, the generative approach to fMRI decoding has been applied only in the visual/pictorial domain, as just described. In the present work, we extend it to the generation of written text. The long-range aspiration is to begin with a brain image encoding some mental content, and to generate from it a verbal description of that content. In the present work, we focused on a simplified version of this challenge: We began with brain images collected during viewing of single words naming concrete concepts (e.g., *house* or *dog*, together with a line drawing of the item named), and from these attempted to generate text describing the relevant concept, in the spirit of an encyclopedia entry. The online encyclopedia Wikipedia served as a gold-standard reference, against which our text-generation results could be compared. As a further simplification of the problem, we followed a step common in machine learning work on text representation [12] by ignoring syntax and word order, treating texts as simple collections of words.

Figure 1 here

Approach At the procedural level, our approach followed a set of steps analogous to those employed by Naselaris [16] and colleagues to reconstruct visual stimuli from fMRI data, but tailored to the task of mapping from fMRI to text:

1. Beginning with a corpus of naturalistic images, learn a generative model for them; this represents individual images as weighted combinations of a set of underlying latent factors, which were discovered through unsupervised learning. Analogously, our work begins with a corpus of texts (i.e., Wikipedia articles), using this to parameterize a form of generative model referred to as a *topic* model [3]. This represents individual texts as a weighted combination of underlying factors or "topics".

2. The next step is to learn a mapping from each latent factor in the model from Step 1 to a corresponding brain image, using a training set of brain images

3. Finally, for each image in a new, test set of brain images, the results from Step 2 are used to infer a weighting over latent factors. These are imposed on the generative model from Step 1, and the model is inverted in order to map from this latent-factor representation to the original representational domain. In the work of Naselaris [16] and colleagues, this resulted in a synthetic image. In our work, it results in a probability distribution over words, i.e., a probabilistic representation of a text.

Our use of topic models had a dual motivation. First, as generative statistical models, topic models support Bayesian inversion, a critical operation in generative fMRI analysis. Second, it has been suggested that the latent representations discovered by topic models may bear important similarities with human semantic representations [7]. This encourages the idea that the latent factors discovered by the topic models in our study would bear a meaningful relationship to patterns of neural activation carrying conceptual information.

We learned our models on a corpus derived from a set of 3500 Wikipedia pages, each dealing with a concrete, imageable concept. As further described in the Methods section, the training texts were stripped of closed-class or function words, and were lemmatized by converting each word to a root form (e.g.,*painted* becomes *paint*). The result of this training was a representation for each article, in the form of a probability distribution over topics, each of which itself defined a probability distribution over individual words. An illustration is presented in Figure 2A. Each column in the figure corresponds to a topic, each row to an article (a small subset of the articles used), with articles grouped into general categories, as labeled on the left. Below, the figure shows the ten most highly weighted words for three topics. The pattern of topic weightings makes clear that the model has captured the category structure implicit in the corpus; through unsupervised learning, several topics have aligned with specific semantic categories (e.g., topic 1 with the *vegetable* category). Topic probabilities for all concepts and topic word distributions can be examined in detail through a model browser available online (`http://www.princeton.edu/~matthewb/wikipedia`).

Armed with the topic model, we used ridge regression to establish a mapping between each topic and a corresponding pattern of brain activation. Our fMRI dataset was derived from an experiment [14] in which participants viewed word-picture pairs, each indicating a specific concrete object (see Figure 1). The stimulus set included a total of 60 objects, corresponding to the Wikipedia articles included in Figure 2A. A representative fMRI image for each stimulus was constructed by averaging across all images collected during trials where the stimulus was presented (a subset of voxels was selected for analysis using a reproducibility criterion detailed in the Supplementary Information). We used the resulting set of 60 images (reserving two images for the test set, as further explained below) as the prediction targets and the set of topic probabilities describing the corresponding Wikipedia articles as regression inputs.

`Figure 2 here`

The resulting regression weights effectively represent each topic in terms of a *basis image*, or representative pattern of brain activation. This makes it possible to decompose the fMRI image for any stimulus object into a set of topic-specific basis images, with combination weights quantifying the contribution of the relevant topic, as illustrated in Figure 2B. Critically, because each topic defines a weighting over specific words, the topic weights inferred from an image can be further translated into an overall probability distribution over words. The process can be reversed and topic probabilities estimated from test brain images, from which we then produce such a probability distribution over words. This procedure is described more formally in the Supplementary Information.

Results Text outputs were generated for each of the 60 brain images in the dataset (when it was in the test set), and an illustrative example is presented as part of Figure 3. The data shown are based on brain images collected during presentation of the stimuli *apartment* and *hammer* for one of the participants. The tag clouds shown in the figure indicate the words most heavily weighted in their respective output distribution. As in this case, text outputs for many stimuli appeared strikingly well aligned with the presumptive semantic associations of the stimulus item. Full results for all 60 concepts are available for inspection online (`http://www.princeton.edu/~matthewb/wikipedia`).

`Figure 3 here`

To more objectively evaluate the quality of the text generated for each stimulus, we used a classification task where the word distributions derived from the brain images for each pair of concepts (test set) were used to match them with the two corresponding Wikipedia pages. The classification was done by considering the total probability of all the words in each Wikipedia article under each probability distribution, and selecting the article deemed most probable.

The idea is illustrated in Figure 3, for the stimuli *apartment* and *hammer*. The text for each of the corresponding Wikipedia articles is presented in colors that indicate the likelihood ratio for each word, given the fMRI-derived text for each stimulus. In this case, each output text matched most closely with the appropriate Wikipedia article. This, indeed, was the case for the majority of stimulus pairs. Plots comparable to Figure 3 for all concept pairs are available online (`http://www.princeton.edu/~matthewb/wikipedia`).

`Figure 4 here`

Overall classification accuracies for each subject are shown in Figure 4, averaged across model parameterizations (number of topics) to avoid bias. Results were statistically significant for all subjects, with p-values calculated using a conservative Monte Carlo procedure being less than 0.01 (see the Supplementary Information for more details about the procedure). As the figure shows, classification performance was best when the comparison was between items belonging to different semantic categories. This indicates that the text outputs for semantically related stimulus items tended to be quite similar.

`Figure 5 here`

The pattern of similarity across items is visualized in Figure 5, which shows the correlation between the topic distributions predicted from each pair of stimulus-specific brain images. The adjacent matrix shows the same correlations for the topic distributions derived from the corresponding pair of Wikipedia articles. The close resemblance between the two matrices indicates that the fMRI-derived text reflected the semantic similarity structure inherent in the stimulus set. The high correlations apparent in the Wikipedia-based matrix also indicate a possible explanation for the relatively weak within-category classification performance we obtained, since our text-generation procedure can only pick up on distinctions if they are made by the underlying topic model. The marginal within-category classification performance may thus reflect the limited granularity of our topic models, rather than a fixed limitation of the overall technique.

Discussion The results we have reported show how a generative, multivariate approach to fMRI image analysis, recently used to generate visual images, can also be applied to the problem of generating text from fMRI data. If this approach can be further developed, it may offer a significant advance over previous efforts to decode patterns of neural activation into language outputs, either letter-by-letter [2] or word-by-word [10], with potential clinical implications for conditions such as locked-in syndrome [18].

The present work serves as a proof of concept, subject to considerable limitations. In order to simplify the problem, we focused only on neural representations of concrete objects. It is therefore an open question how the present technique would perform on a wider range of semantic content. This includes more abstract concepts and relational representations. However, one can also optimistically imagine developing techniques for fMRI-based text generation that might take such factors as emotion or even attitude into account. A second important simplification was to ignore word order and grammatical structure. Although this is a conventional step in text-analysis research, a practical method for text generation would clearly require grammatical structure to be taken into account. In this regard, it is interesting to note that there have been proposals [7,24] of approaches to enriching topic model representations by considering word dependency and order. Integrating such a modeling approach into the present generative approach to fMRI analysis might support more transparently meaningful text outputs.

Methods Summary The stimuli in the fMRI study [14] that originated our dataset were line drawings and noun labels of 60 concrete objects from 12 semantic categories, with 5 exemplars per category, adapted from an existing collection [21]. The 60 stimulus items were presented six times, randomly permuted in each presentation. Each item was presented for 3s, followed by a 7s rest period, during which participants fixated. When an item was presented, the participant's task was to think about the properties the item. Nine subjects participated in the fMRI study. 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). Each image was normalized by subtracting its mean and dividing by its standard deviation, both across all voxels.

The classification procedure uses two types of optimization problem. The first is learning a set of basis images, given example images for 58 concepts and their respective topic probabilities. This can be decomposed into a set of independent ridge regression problems, one per voxel, where one predicts the values of the voxel across examples from the respective topic probabilities; the regression coefficients are the values of the basis images at that voxel. The second problem is predicting the topic probabilities present in an example image, given a set of basis images. This is a linear regression problem where the values of all the voxels in an example are predicted by combining the basis images, using the topic probabilities for that example as the regression coefficients, constrained to be positive and sum to 1. The Supplementary Information contains more details about the study, corpus construction, topic models and classification procedure.

Methods The stimuli in the fMRI study [14] that originated our dataset were line drawings and noun labels of 60 concrete objects from 12 semantic categories with 5 exemplars per category, adapted from an existing collection [21]. The entire set of 60 stimulus items was presented six times, randomly permuting the sequence of the 60 items on each presentation. Each item was presented for 3s, followed by a 7s rest period, during which the participants were instructed to fixate. When an exemplar was presented, the participant's task was to think about the properties of the object. Nine subjects participated in the fMRI study. Functional images were acquired on a Siemens 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 degree 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 mm3 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). Each of these images was normalized by subtracting its mean and dividing by its standard deviation, both across all voxels.

To derive a corpus from Wikipedia we started with classical lists of words [1 19], as well as modern revisions/extensions thereof [4 23], and compiled words corresponding to concepts that were deemed concrete or imageable, be it because of their score in one of the lists or through editorial decision. We then identified the corresponding Wikipedia article titles (e.g. “airplane” is “Fixed-wing aircraft”) and also compiled related articles which were linked to from these (e.g. “Aircraft cabin”). If there were words in the original lists with multiple meanings we included the articles for at least a few of those meanings. We used Wikipedia Extractor (`http://medialab.di.unipi.it/wiki/Wikipedia_extractor`) to remove HTML, wiki formatting and annotations and processed the resulting text through the morphological analysis tool Morpha [11] (`http://www.informatics.susx.ac.uk/research/groups/nlp/carroll/morph.html`) to lemmatize all the words to their basic stems (e.g. “taste”,”tasted”,”taster” and “tastes” all become the same word).

The resulting text corpus was processed with topic modelling [3] software (`http://www.cs.princeton.edu/~blei/topicmodeling.html`) to produce several models, excluding words that appeared in a single article or were in a stopword list. We ran the software varying the number of topics allowed from 10 to 100, in increments of 10, setting the α parameter to $\frac{25}{\#topics}$ (as suggested in other work modelling a large text corpus for semantic purposes [7], though a range of multiples of the inverse of the number of topics yielded comparable experiment results).

Classification accuracy was measured on the task of matching two example images with the two corresponding wikipedia articles, by considering the probability assigned to the words in each article by the distributions derived from the example images, as illustrated in Supplementary Figure 1 and described in the following steps:

1. leave out one pair of concepts (e.g. “apartment” and “hammer”) as test set

2. use the example images for the remaining 58 concepts, together with their respective topic probabilities under the model, as the training set to obtain a set of basis images (over 1000 stable voxels, selected in this training set)

3. for each of the test concepts (for instance, "apartment"):

   - predict the probability of each topic being present from the "apartment" example image
   - obtain an "apartment"-brain probability distribution for that combination of topic probabilities
   - compute the probability of "apartment" article and "hammer" article under that distribution, respectively $p_{apartment}(\text{``}\texttt{apartment}'')$ and $p_{apartment}(\text{``}\texttt{hammer}'')$

4. assign the article with highest probability to the corresponding test concept, and the other article to the other concept (this will be correct or incorrect)

The steps are repeated for every possible pair of concepts, and the accuracy is the fraction of the pairs where the assignment of articles to example images was correct. For voxel selection we used a reproducibility criterion, which identifies voxels whose activation levels across the training set examples of each concept bear the same relationship to each other over epochs (mathematically, the vector of activation levels across the sorted concepts is highly correlated between epochs). More details are provided in the Supplementary Information.

The classification procedure has two steps that require solving optimization problems. The first is learning a set of basis images, given example images for 58 concepts and their respective topic probabilities. This can be decomposed into a set of independent ridge regression problems, one per voxel, where one predicts the values of the voxel across examples from the respective topic probabilities; the regression coefficients are the values of the basis images at that voxel. The second problem is predicting the topic probabilities present in an example image, given a set of basis images. This is a linear regression problem where the values of all the voxels in an example are predicted by combining the basis images, using the topic probabilities for that example as the regression coefficients (under the constraint that the values need to be greater than or equal to 0 and add up to 1, as they are probabilities). More details are provided in the Supplementary Information.

References

[1] W. F. Battig and W. E. Montague. Category Norms for Verbal Items in 56 Categories. *Journal of Experimental Psychology*, 80(3), 1969.

[2] N. Birbaumer. Breaking the silence: brain-computer interfaces (BCI) for communication and motor control. *Psychophysiology*, 43(6):517–32, November 2006.

[3] D. Blei, M. Jordan, and A. Y. Ng. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] J. M. Clark and A. Paivio. Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior research methods, instruments, & computers : a journal of the Psychonomic Society, Inc*, 36(3):371–83, August 2004.

[5] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *NeuroImage*, 28(3):663–8, 2005.

[6] M. Grant and S. Boyd. CVX users guide, 2009.

[7] T. L. Griffiths, M. Steyvers, and J. B. Tenenbaum. Topics in semantic representation. *Psychological review*, 114(2):211–44, April 2007.

[8] J. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature reviews. Neuroscience*, 7(7):523–34, 2006.

[9] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–5, 2008.

[10] S. Kellis, K. Miller, K. Thomson, R. Brown, P. House, and B. Greger. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering*, 7(5):056007, September 2010.

[11] G. Minnen, J. Carroll, and D. Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(03):207–223, 2001.

[12] T. M. Mitchell. *Machine Learning*. 1997.

[13] T. M. Mitchell, R. Hutchinson, R. S. Niculescu, F. Pereira, X. Wang, M. Just, and S. Newman. Learning to Decode Cognitive States from Brain Images. *Machine Learning*, 57(1/2):145–175, October 2004.

[14] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *Science (New York, N.Y.)*, 320(5880):1191–5, 2008.

[15] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H. Tanabe, N. Sadato, and Y. Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60(5):915–29, 2008.

[16] T. Naselaris, R. J. Prenger, K. N. Kay, M. Oliver, and J. L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63(6):902–15, 2009.

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

[18] A. M. Owen, M. R. Coleman, M. Boly, M. H. Davis, S. Laureys, and J. Pickard. Detecting awareness in the vegetative state. *Science*, 313(5792):1402, September 2006.

[19] A. Paivio, J. C. Yuille, and S. A. Madigan. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology*, 76(1), 1968.

[20] F. Pereira, T. M. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, 45(1 Suppl):S199–209, March 2009.

[21] J. G. Snodgrass and M. Vanderwart. A Standardized Set of 260 Pictures : Norms for Name Agreement, Image Agreement, Familiarity, and Visual Complexity. *Human Learning*, 6(2):174–215, 1980.

[22] B. Thirion, E. Duchesnay, E. Hubbard, J. Dubois, J. B. Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *NeuroImage*, 33:1104–1116, 2006.

[23] J. Van Overschelde. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50(3):289–335, 2004.

[24] H.M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, number 1, pages 977–984. ACM, 2006.

Supplementary Information Supplementary Information is linked to the online version of the paper at `http://www.nature.com/nature`.

Acknowledgments We would like to thank Tom Mitchell and Marcel Just for generously agreeing to share their dataset, Dave Blei for help with building and interpreting topic models and Ken Norman and Sam Gershman for several discussions regarding presentation of the material.

Author Contributions F.P and G.D. designed the approach for predicting topics from brain images, F.P selected the corpus, produced topic models and ran all the prediction experiments, F.P and M.B. designed the prediction experiments and wrote the manuscript.

Author Information The dataset we used can be downloaded at `http://www.cs.cmu.edu/~tom/science2008`. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to `francisco.pereira@gmail.com`.

Figure Legends

Figure 1

The approach we follow to generate text (bottom) parallels that used by Naselaris [16] (top, adapted from that paper), by having three stages: creating a model of how stimuli will be represented in the brain, learning how to predict fMRI data in response to the stimuli, given the model, and inverting the process to make a prediction for fMRI data not used to fit the model.

Figure 2

A: Topic probabilities for the wikipedia articles about the 60 concepts for which we have fMRI data. Each concept belongs to one of 12 semantic categories, and concepts are grouped by category (five animals, five insects, etc). Note that the category structure visible is due to how we sorted the columns for display; the model is trained in an unsupervised manner and knows nothing about category structure. Note also that there are topics that are not probable for any of the concepts, which happens because they are used for other concepts in the 3500 concept corpus. Below this are the top 10 most probable words in the probability distributions associated with three of the topics.

B: The decomposition of the brain image for "House" into a weighted combination of topic basis images. The weights allow us to combine the corresponding topic word distributions into an overall word distribution (top 10 words shown).

Figure 3

The inset under each article shows the top words from the corresponding brain-derived distribution (10 which are present in the article (black) and 10 which are not (gray)). Each word of the two articles is colored to reflect the ratio $\frac{P_{apartment}(\texttt{word})}{P_{hammer}(\texttt{word})}$ between the probabilities assigned to it by the brain-derived distributions for concepts "apartment" and "hammer" (red means higher probability under "apartment", blue under "hammer", gray means the word is not considered by the text model).

Figure 4

Average classification accuracy across models using 10 to 100 topics, for each of 9 subjects (chance level is 0.5); the accuracy is broken down into classification of concept pairs where concepts are in different categories ("Between") and pairs where the category is the same ("Within"). Error bars are across numbers of topics.

Figure 5

Similarity between the topic probability representations of each concept learned solely from text (left) and also the representations predicted from the brain images for each pair of concepts, when they were being used as the test set (right). The latter was obtained from subject 1 and a 40 topic model, but the general pattern is similar for the other subjects. Note that the representations for concepts in the same category similar when obtained from brain images but this is also the case when those representations are derived from text.

Figures

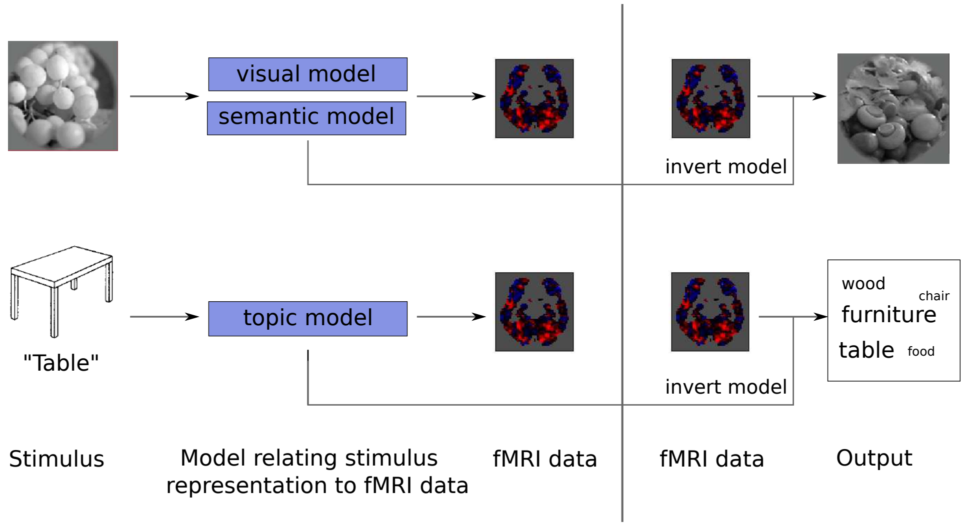

Figure 1:

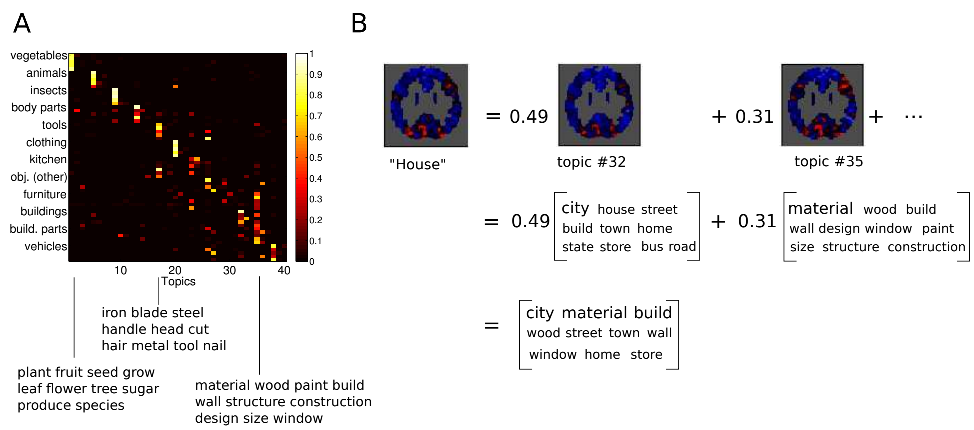

Figure 2:

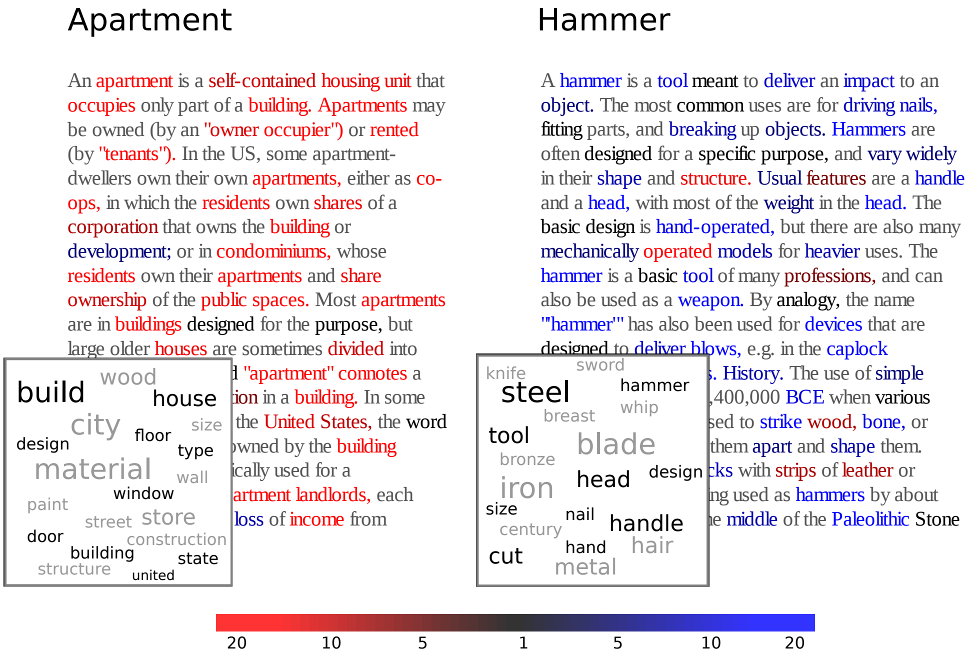

Figure 3:

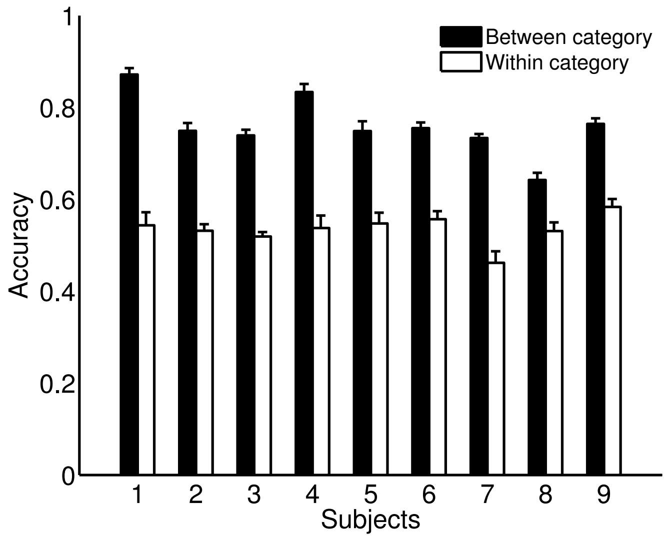

Figure 4:

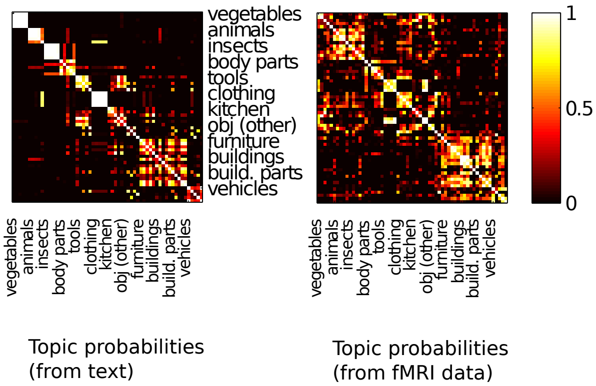
vegetables
animals
insects
body parts
tools
clothing
kitchen
obj (other)
furniture
buildings
build. parts
vehicles
1
0.5
0
Topic probabilities
(from text)
Topic probabilities
(from fMRI data)

Figure 5: