Neuron
Article

Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex

Dustin E. Stansbury,[1] Thomas Naselaris,[2,4] and Jack L. Gallant[1,2,3,*]
[1]Vision Science Group
[2]Helen Wills Neuroscience Institute
[3]Department of Psychology
University of California, Berkeley, CA 94720, USA
[4]Present address: Department of Neurosciences, Medical University of South Carolina, Charleston, SC 29425, USA
*Correspondence: gallant@berkeley.edu

http://dx.doi.org/10.1016/j.neuron.2013.06.034

SUMMARY

During natural vision, humans categorize the scenes they encounter: an office, the beach, and so on. These categories are informed by knowledge of the way that objects co-occur in natural scenes. How does the human brain aggregate information about objects to represent scene categories? To explore this issue, we used statistical learning methods to learn categories that objectively capture the co-occurrence statistics of objects in a large collection of natural scenes. Using the learned categories, we modeled fMRI brain signals evoked in human subjects when viewing images of scenes. We find that evoked activity across much of anterior visual cortex is explained by the learned categories. Furthermore, a decoder based on these scene categories accurately predicts the categories and objects comprising novel scenes from brain activity evoked by those scenes. These results suggest that the human brain represents scene categories that capture the co-occurrence statistics of objects in the world.

INTRODUCTION

During natural vision, humans categorize the scenes that they encounter. A scene category can often be inferred from the objects present in the scene. For example, a person can infer that she is at the beach by seeing water, sand, and sunbathers. Inferences can also be made in the opposite direction: the category "beach" is sufficient to elicit the recall of these objects plus many others such as towels, umbrellas, sandcastles, and so on. These objects are very different from those that would be recalled for another scene category such as an office. These observations suggest that humans use knowledge about how objects co-occur in the natural world to categorize natural scenes.

There is substantial behavioral evidence to show that humans exploit the co-occurrence statistics of objects during natural vision. For example, object recognition is faster when objects in a scene are contextually consistent (Biederman, 1972; Biederman et al., 1973; Palmer, 1975). When a scene contains objects that are contextually inconsistent, then scene categorization is more difficult (Potter, 1975; Davenport and Potter, 2004; Joubert et al., 2007). Despite the likely importance of object co-occurrence statistics for visual scene perception, few fMRI studies have investigated this issue systematically. Most previous fMRI studies have investigated isolated and decontextualized objects (Kanwisher et al., 1997; Downing et al., 2001) or a few, very broad scene categories (Epstein and Kanwisher, 1998; Peelen et al., 2009). However, two recent fMRI studies (Walther et al., 2009; MacEvoy and Epstein, 2011) provide some evidence that the human visual system represents information about individual objects during scene perception.

Here we test the hypothesis that the human visual system represents scene categories that capture the statistical relationships between objects in the natural world. To investigate this issue, we used a statistical learning algorithm originally developed to model large text corpora to learn scene categories that capture the co-occurrence statistics of objects found in a large collection of natural scenes. We then used fMRI to record blood oxygenation level-dependent (BOLD) activity evoked in the human brain when viewing natural scenes. Finally, we used the learned scene categories to model the tuning of individual voxels and we compared predictions of these models to alternative models based on object co-occurrence statistics that lack the statistical structure inherent in natural scenes.

We report three main results that are consistent with our hypothesis. First, much of anterior visual cortex represents scene categories that reflect the co-occurrence statistics of objects in natural scenes. Second, voxels located within and beyond the boundaries of many well-established functional ROIs in anterior visual cortex are tuned to mixtures of these scene categories. Third, scene categories and the specific objects that occur in novel scenes can be accurately decoded from evoked brain activity alone. Taken together, these results suggest that scene categories represented in the human brain

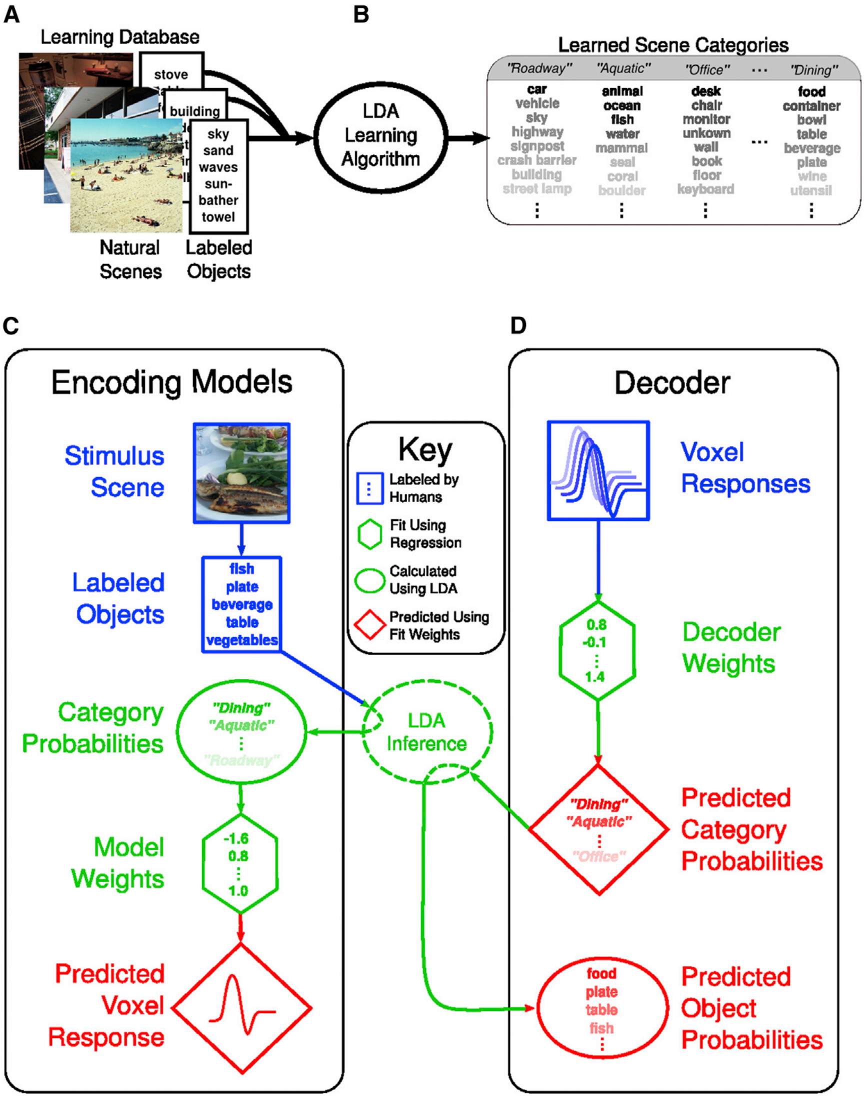


Figure 1. Overview of Analyses

(A) Learning database. We compiled a large database of labeled natural scenes. All objects in each of the scenes were labeled by naive participants. See also Figure S2.

(B) Scene categories learned by LDA. LDA was used to learn scene categories that best capture the co-occurrence statistics of objects in the learning database. LDA defines each scene category as a list of probabilities, where each probability is the likelihood that any particular object within a fixed vocabulary will occur in a scene. Lists of probable objects for four example scene categories learned by LDA are shown on the right. Each list of object labels corresponds to a distinct scene category; within each list, saturation indicates an object's probability of occurrence. The experimenters, not the LDA algorithm, assigned intuitive category names in quotes. Once a set of categories is learned, LDA can also be used to infer the probability that a new scene belongs to each of the learned categories, conditioned on the objects in the new scene. See also Figure S2.

(C) Voxelwise encoding model analysis. Voxelwise encoding models were constructed to predict BOLD responses to stimulus scenes presented during an fMRI experiment. Blue represents inputs to the encoding model, green represents intermediate model steps, and red represents model predictions. To generate predictions, we passed the labels associated with each stimulus scene (blue box) to the LDA algorithm (dashed green oval). LDA is used to infer from these labels the probability that the stimulus scene belongs to each of the learned categories (solid green oval). In this example, the stimulus scene depicts a plate of fish, so the scene categories "Dining" and "Aquatic" are highly probable (indicated by label saturation), while the category "Roadway" is much less probable. These probabilities are then transformed into a predicted BOLD response (red diamond) by a set of linear model weights (green hexagon). Model weights were fit independently for each voxel using a regularized linear regression procedure applied to the responses evoked by a set of training stimuli.

(D) Decoding model analysis. A decoder was constructed for each subject that uses BOLD signals evoked by a viewed stimulus scene to predict the probability that the scene belongs to each of a set of learned scene categories. Blue represents inputs to the decoder, green represents intermediate model steps, and red represents decoder predictions. To generate a set of category probability predictions for a scene (red diamond), we mapped evoked population voxel responses (blue box) onto the category probabilities by a set of multinomial model weights (green hexagon). Predicted scene category probabilities were then used in conjunction with the LDA algorithm to infer the probabilities that specific objects occurred in the viewed scene (red oval). The decoder weights were fit using regularized multinomial regression applied to the scene category probabilities inferred for a set of training stimuli using LDA and the responses to those stimuli.

capture the statistical relationships between objects in the natural world.

RESULTS

Learning Natural Scene Categories

To test whether the brain represents scene categories that reflect the co-occurrence statistics of objects in natural scenes, we first had to obtain such a set of categories. We used statistical learning methods to solve this problem (Figures 1A and 1B). First, we created a learning database by labeling the individual objects in a large collection of natural scenes (Figure 1A). The frequency counts of the objects that appeared in each scene in the learning database were then used as input to the Latent Dirichlet Allocation (LDA) learning algorithm (Blei et al., 2003). LDA was originally developed to learn underlying topics in a collection of documents based on the co-occurrence statistics of the words in the documents. When applied to the frequency counts of the objects in the learning database, the LDA algorithm learns an underlying set of scene categories that capture the co-occurrence statistics of the objects in the database.

LDA defines each scene category as a list of probabilities that are assigned to each of the object labels within an available vocabulary. Each probability reflects the likelihood that a specific

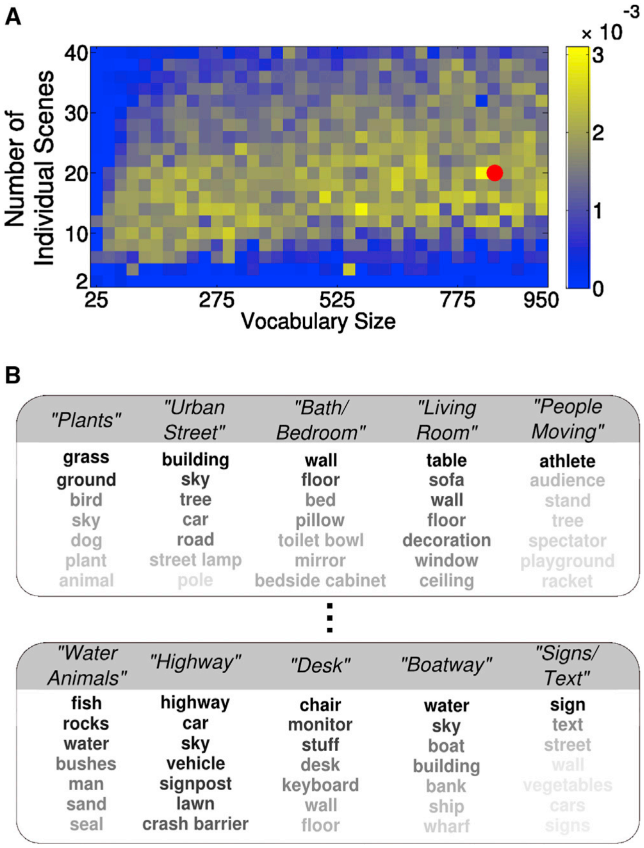


Figure 2. Identifying the Best Scene Categories for Modeling Data across Subjects

(A) Encoding model performance across a range of settings for the specified number of distinct categories learned using LDA (y axis) and vocabulary size (x axis). Each pixel corresponds to one of the candidate scene categories learned by LDA when applied to the learning database. The color of each pixel represents the relative amount of cortical territory across subjects that is accurately predicted by encoding models based on a specific setting for the number of individual categories and vocabulary size. The number of individual categories was incremented from 2 to 40. The object vocabulary was varied from the 25 most frequent to the 950 most frequent objects in the learning database. The red dot identifies the number of individual categories and vocabulary size that produce accurate predictions for the largest amount of cortical territory across subjects. For individual results, see Figure S3.

(B) Ten examples taken from the 20 best scene categories identified across subjects (corresponding to the red dot in A). The seven most probable objects for each category are shown. Format is the same as in Figure 1B. See Figures S4 and S5 for interpretation of all 20 categories.

object occurs in a scene that belongs to that category (Figure 1B). LDA learns the probabilities that define each scene category without supervision. However, the number of distinct categories the algorithm learns and the object label vocabulary must be specified by the experimenter. The vocabulary used for our study consisted of the most frequent objects in the learning database.

Figure 1B shows examples of scene categories learned by LDA from the learning database. Each of the learned categories can be named intuitively by inspecting the objects that they are most likely to contain. For example, the first category in Figure 1B (left column) is aptly named “Roadway” because it is most likely to contain the objects “car,” “vehicle,” “highway,” “crash barrier,” and “street lamp.” The other examples shown in Figure 1B can also be assigned intuitive names that describe typical natural scenes. Once a set of scene categories has been learned, the LDA algorithm also offers a probabilistic inference procedure that can be used to estimate the probability that a new scene belongs to each of the learned categories, conditioned on the objects in the new scene.

Voxelwise Encoding Models Based on Learned Scene Categories

To determine whether the brain represents the scene categories learned by LDA, we recorded BOLD brain activity evoked when human subjects viewed 1,260 individual natural scene images. We used the LDA probabilistic inference procedure to estimate the probability that each of the presented stimulus scenes belonged to each of a learned set of categories. For instance, if a scene contained the objects “plate,” “table,” “fish,” and “beverage,” LDA would assign the scene a high probability of belonging to the “Dining” category in Figure 1B, a lower probability to the “Aquatic” category, and near zero probability to the remaining categories (Figure 1C, green oval).

The category probabilities inferred for each stimulus scene were used to construct voxelwise encoding models. The encoding model for each voxel consisted of a set of weights that best mapped the inferred category probabilities of the stimulus scenes onto the BOLD responses evoked by the scenes (Figure 1C, green hexagon). Model weights were estimated using regularized linear regression applied independently for each subject and voxel. The prediction accuracy for each voxelwise encoding model was defined to be the correlation coefficient (Pearson’s *r* score) between the responses evoked by a novel set of stimulus scenes and the responses to those scenes predicted by the model.

Introspection suggests that humans can conceive of a vast number of distinct objects and scene categories. However, because the spatial and temporal resolution of fMRI data are fairly coarse (Buxton, 2002), it is unlikely that all these objects or scene categories can be recovered from BOLD signals. BOLD signal-to-noise ratios (SNRs) also vary dramatically across individuals, so the amount of information that can be recovered from individual fMRI data also varies. Therefore, before proceeding with further analysis of the voxelwise models, we first identified the single set of scene categories that provided the best predictions of brain activity recorded from all subjects. To do so, we examined how the amount of accurately predicted cortical territory across subjects varied with specific settings of the number of individual scene categories and object vocabulary size assumed by the LDA algorithm during category learning. Specifically, we incremented the number of individual categories learned from 2 to 40 while also varying the size of the object label vocabulary from the 25 most frequent to 950 most frequent objects in the learning database (see Experimental Procedures for further details). Figure 2A shows the relative amount of accurately predicted cortical territory across subjects based on each setting. Accurate predictions are stable across a wide range of settings.

Across subjects, the encoding models perform best when based on 20 individual categories and composed of a vocabulary of 850 objects (Figure 2A, indicated by red dot; for individual subject results, see Figure S3 available online). Examples of these categories are displayed in Figure 2B (for an interpretation of all 20 categories, see Figures S4 and S5). To the best of our knowledge, previous fMRI studies have only used two to eight distinct categories and 2–200 individual objects (see Walther et al., 2009; MacEvoy and Epstein, 2011). Thus, our results show there is more information in BOLD signals related to encoding scene categories than has been previously appreciated.

We next tested whether natural scene categories were necessary to accurately model the measured fMRI data. We derived a set of null scene categories by training LDA on artificial scenes. The artificial scenes were created by scrambling the objects in the learning database across scenes, thus removing the natural statistical structure of object co-occurrences inherent in the original learning database. If the brain incorporates information about the co-occurrence statistics of objects in natural scenes, then the prediction accuracy of encoding models based upon these null scene categories should be much poorer than encoding models based on scene categories learned from natural scenes.

Indeed, we find that encoding models based on the categories learned from natural scenes provide significantly better predictions of brain activity than do encoding models based on the null categories and for all subjects ($p < 1 \times 10^{-10}$ for all subjects, Wilcox rank-sum test for differences in median prediction accuracy across all cortical voxels and candidate scene category settings; subject S1: $W(15{,}025{,}164) = 9.96 \times 10^{13}$; subject S2: $W(24{,}440{,}399) = 3.04 \times 10^{14}$; subject S3: $W(15{,}778{,}360) = 9.93 \times 10^{13}$; subject S4: $W(14{,}705{,}625) = 1.09 \times 10^{14}$). In a set of supplemental analyses, we also compared the LDA-based models to several other plausible models of scene category representation. We find that the LDA-based models provide superior prediction accuracy to all these alternative models (see Figures S12–S15). These results support our central hypothesis that the human brain encodes categories that reflect the co-occurrence statistics of objects in natural scenes.

Categories Learned From Natural Scenes Explain Selectivity in Many Anterior Visual ROIs

Previous fMRI studies have identified functional regions of interest (ROIs) tuned to very broad scene categories, such as places (Epstein and Kanwisher, 1998), as well as to narrow object categories such as faces (Kanwisher et al., 1997) or body parts (Downing et al., 2001). Can selectivity in these regions be explained in terms of the categories learned from natural scene object statistics?

We evaluated scene category tuning for voxels located within the boundaries of several conventional functional ROIs: the fusiform face area (FFA; Kanwisher et al., 1997), the occipital face area (OFA; Gauthier et al., 2000), the extrastriate body area (EBA; Downing et al., 2001), the parahippocampal place area (PPA; Epstein and Kanwisher, 1998), the transverse occipital sulcus (TOS; Nakamura et al., 2000; Grill-Spector, 2003; Hasson et al., 2003), the retrosplenial cortex (RSC; Maguire, 2001), and lateral occipital cortex (LO; Malach et al., 1995).

Figure 3A shows the boundaries of these ROIs, identified using separate functional localizer experiments, and projected on the cortical flat map of one representative subject. The color of each location on the cortical map indicates the prediction accuracy of the corresponding encoding model. All encoding models were based on the 20 best scene categories identified across subjects. These data show that the encoding models accurately predict responses of voxels located in many ROIs within anterior visual cortex. To quantify this effect, we calculated the proportion of response variance explained by the encoding models, averaged across all voxels within each ROI. We find that the average proportion of variance explained to be significantly greater than chance for every anterior visual cortex ROI and for all subjects ($p < 0.01$; see Experimental Procedures for details). Thus, selectivity in many previously identified ROIs can be explained in terms of tuning to scene categories learned from natural scene statistics.

To determine whether scene category tuning is consistent with tuning reported in earlier localizer studies, we visualized the weights of encoding models fit to voxels within each ROI. Figure 3C shows encoding model weights averaged across all voxels located within each function ROI. Scene category selectivity is broadly consistent with the results of previous functional localizer experiments. For example, previous studies have suggested that PPA is selective for presence of buildings (Epstein and Kanwisher, 1998). The LDA algorithm suggests that images containing buildings are most likely to belong to the ‘‘Urban/Street’’ category (see Figure 2B), and we find that voxels within PPA have large weights for the ‘‘Urban/Street’’ category (see Figures S4 and S5). To take another example, previous studies have suggested that OFA is selective for the presence of human faces (Gauthier et al., 2000). Under the trained LDA model, images containing faces are most likely to belong to the ‘‘Portrait’’ category (see Figures S4 and S5), and we find that voxels within OFA have large weights for the ‘‘Portrait’’ category.

Although category tuning within functional ROIs is generally consistent with previous reports, Figure 3C demonstrates that tuning is clearly more complicated than assumed previously. In particular, many functional ROIs are tuned for more than one scene category. For example, both FFA and OFA are thought to be selective for human faces, but voxels in both these areas also have large weights for the ‘‘Plants’’ category. Additionally, area TOS, an ROI generally associated with encoding information important for navigation, has relatively large weights for the ‘‘Portrait’’ and ‘‘People Moving’’ categories. Thus, our results suggest that tuning in conventional ROIs may be more diverse than generally believed (for additional evidence, see Huth et al., 2012 and Naselaris et al., 2012).

Decoding Natural Scene Categories from Evoked Brain Activity

The results presented thus far suggest that information about natural scene categories is encoded in the activity of many voxels located in anterior visual cortex. It should therefore be possible to decode these scene categories from brain activity evoked by viewing a scene. To investigate this possibility, we constructed a decoder for each subject that uses voxel activity evoked in anterior visual cortex to predict the probability that a

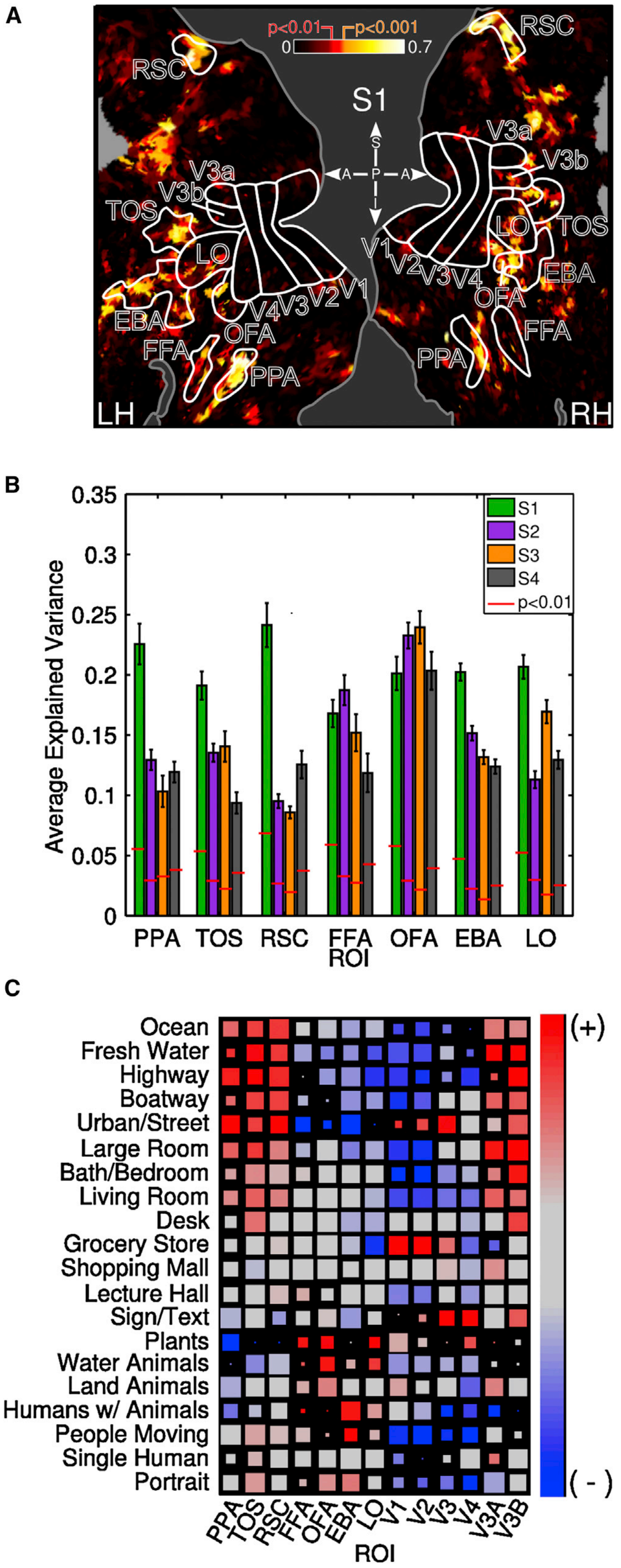


Figure 3. Scene Categories Learned from Natural Scenes Are Encoded in Many Anterior Visual ROIs

(A) Encoding model prediction accuracies are mapped onto the left (LH) and right (RH) cortical surfaces of one representative subject (S1). Gray indicates areas outside of the scan boundary. Bright locations indicate voxels that are accurately predicted by the corresponding encoding model (prediction accuracy at two levels of statistical significance—$p < 0.01$ [$r = 0.21$] and $p < 0.001$ [$r = 0.28$]—are highlighted on the color bar). ROIs identified in separate retinotopy and functional localizer experiments are outlined in white. The bright regions overlap with a number of the ROIs in anterior visual cortex. These ROIs are associated with representing various high-level visual features. However, the activity of voxels in retinotopic visual areas (V1, V2, V3, V4, V3a, V3b) are not predicted accurately by the encoding models. Prediction accuracy was calculated on responses from a separate validation set of stimuli not used to estimate the model. ROI Abbreviations: V1–V4, retinotopic visual areas 1–4; PPA, parahippocampal place area; FFA, fusiform face area; EBA, extrastriate body area; OFA, occipital face area; RSC, retrosplenial cortex; TOS, transverse occipital sulcus. Center key: A, anterior; P, posterior; S, superior; I, inferior. For remaining subjects' data, see Figure S6.

(B) Each bar indicates the average proportion of voxel response variance in an ROI that is explained by voxelwise encoding models estimated for a single subject. Bar colors distinguish individual subjects. Error bars represent SEM. For all anterior visual ROIs and for all subjects, encoding models based on scene categories learned from natural scenes explain a significant proportion of voxel response variance ($p < 0.01$, indicated by red lines).

(C) The average encoding model weights for voxels within distinct functional ROIs. Averages are calculated across all voxels located within the boundaries of an ROI and across subjects. Each row displays the average weights for the scene category listed on the left margin. Each column distinguishes average weights for individual ROIs. The color of each pixel represents the positive (red) or negative (blue) average ROI weight for the corresponding category. The size of each pixel is inversely proportional to the magnitude of the SEM estimate; larger pixels indicate selectivity estimates with greater confidence. SE scaling is according to the data within an ROI (column). ROI tuning is generally consistent with previous findings. However, tuning also appears to be more complex than indicated by conventional ROI-based analyses. For individual subjects' data, see Figure S7; see also Figures S8–S15.

viewed scene belongs to each of 20 best scene categories identified across subjects. To maximize performance, the decoder used only those voxels for which the encoding models produced accurate predictions on a held-out portion of the model estimation data (for details, see Experimental Procedures).

We used the decoder to predict the 20 category probabilities for 126 novel scenes that had not been used to construct the decoder. Figure 4A shows several examples of the category probabilities predicted by the decoder. The scene in the upper right of Figure 4A depicts a harbor in front of a city skyline. The predicted category probabilities indicate that the scene is most likely a mixture of the categories "Urban" and "Boatway," which is an accurate description of the scene. Inspection of the other examples in the figure suggests that the predicted scene category probabilities accurately describe many different types of natural scenes.

To quantify the accuracy of each decoder, we calculated the correlation (Pearson's r) between the scene category probabilities predicted by the decoder and the probabilities inferred using the LDA algorithm (conditioned on the labeled objects in each scene). Figure 4B shows the distribution of decoding accuracies across all decoded scenes, for each subject. The median accuracies and 95% confidence interval (CI) on median estimates are indicated by the black cross-hairs. Most of the novel scenes are decoded significantly for all subjects. Prediction accuracy

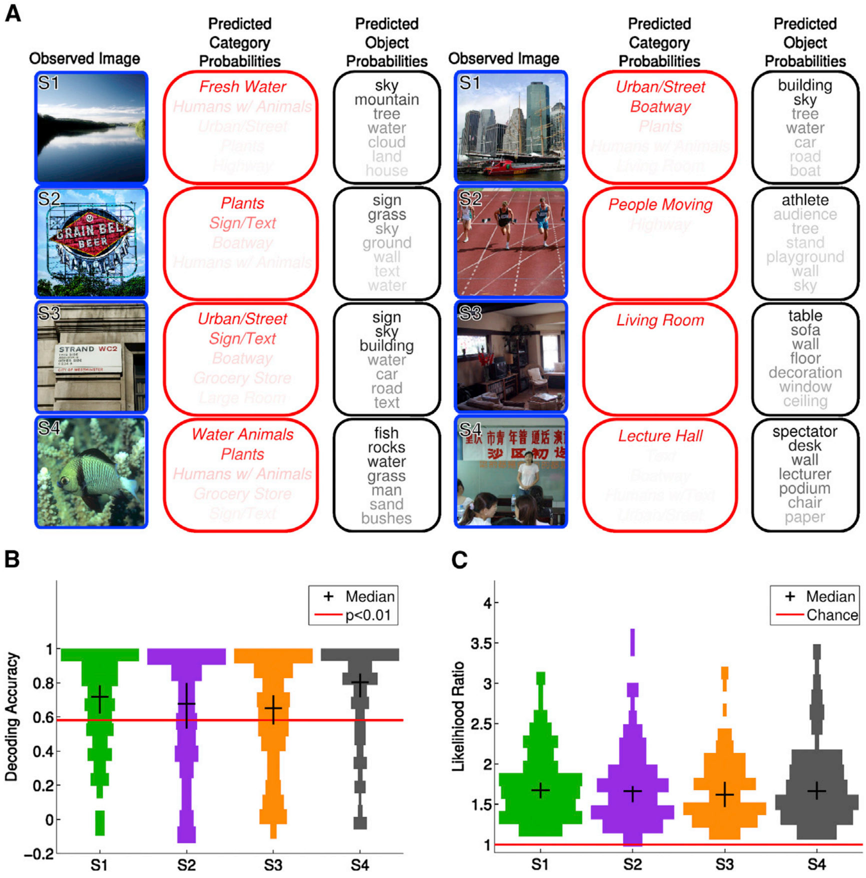


Figure 4. Scene Categories and Objects Decoded from Evoked BOLD Activity

(A) Examples of scene category and object probabilities decoded from evoked BOLD activity. Blue boxes (columns 1 and 4) display novel stimulus scenes observed by subjects S1 (top row) through S4 (bottom row). Each red box (columns 2 and 5) encloses the top category probabilities predicted by the decoder for the corresponding scene to the left. The saturation of each category name within the red boxes represents the predicted probability that the observed scene belongs to the corresponding category. Black boxes (columns 3 and 6) enclose the objects with the highest estimated probability of occurring in the observed scene to the left. The saturation of each label within the black boxes represents the estimated probability of the corresponding object occurring in the scene. See also Figures S16–S19.

(B) Decoding accuracy for predicted category probabilities. Category decoding accuracy for a scene is the correlation coefficient between the category probabilities predicted by the decoder and the category probabilities inferred directly using LDA. Category probabilities were decoded for 126 novel scenes. Each plot shows the (horizontally mirrored) histogram of decoding accuracies for a single subject. Median decoding accuracy and 95% confidence interval (CI) calculated across all decoded scenes is represented by black cross-hairs overlayed on each plot. For subjects S1–S4, median decoding accuracy was 0.72 (CI: [0.62, 0.78]), 0.68 (CI: [0.53, 0.80]), 0.65 (CI: [0.55, 0.72]), and 0.80 (CI: [0.72, 0.85]), respectively. For a given image, decoding accuracy greater than 0.58 was considered statistically significant ($p < 0.01$) and is indicated by the red line. A large majority of the decoded scenes are statistically significant, including all examples shown in (A).

(C) Decoding accuracy for predicted object probabilities. Object decoding accuracy is the ratio of the likelihood of the objects labeled in each scene given the decoded category probabilities, to the likelihood of the labeled objects in each scene if all were selected with equal probability (chance). A likelihood ratio greater than one (red line) indicates that the objects in a scene are better predicted by the decoded object probabilities than by selecting objects randomly. Each plot shows the (horizontally mirrored) histogram of likelihood ratios for a single subject. Median likelihood ratios and 95% CI are represented by the black cross-hairs. For subjects S1–S4, the median likelihood ratio was 1.67 (CI: [1.57, 1.76]), 1.66 (CI: [1.52, 1.72]), 1.62 (CI: [1.45, 1.78]), and 1.66 (CI: [1.56, 1.78]) for subjects S1–S4, respectively.

across all scenes exhibited systematically greater-than-chance performance for all subjects ($p < 0.02$ for all subjects, Wilcox rank-sum test; subject S1: $W(126) = 18{,}585$; subject S2: $W(126) = 17{,}274$; subject S3: $W(126) = 17{,}018$; subject S4: $W(126) = 19{,}214$. The voxels selected for the decoding analysis summarized in Figure 4 were located throughout the visual cortex. However, we also find that accurate decoding can be obtained using the responses of subsets of voxels located within specific ROIs (see Figures S16–S19).

Predicting the Objects that Occur in Decoded Natural Scenes

Our results suggest that the visual system represents scene categories that capture the co-occurrence statistics of objects in the natural world. This suggests that we should be able to predict accurately the likely objects in a scene based on the scene category probabilities decoded from evoked brain activity.

To investigate this issue, we estimated the probability that each of the 850 objects in the vocabulary for the single best set of scene categories identified across subjects occurred in each of the 126 decoded validation set scenes. The probabilities were estimated by combining the decoded category probabilities with the probabilistic relationship between categories and objects established by the LDA learning algorithm during category learning (see Experimental Procedures for details). The resulting probabilities give an estimate of the likelihood that each of the 850 objects occurs in each of the 126 decoded scenes.

In Figure 4A, labels in the black boxes indicate the most likely objects estimated for the corresponding decoded scene. For the harbor and skyline scene at upper right, the most probable objects predicted for the scene are “building,” “sky,” “tree,” “water,” “car,” “road,” and “boat.” All of these objects either occur in the scene or are consistent with the scene context. Inspection of the other examples in the figure suggests that the most probable objects are generally consistent with the scene category.

To quantify how accurately the objects were decoded, we used the distribution of object probabilities estimated for each scene to calculate the likelihood of the labeled objects in the scene. We then calculated the likelihood of the labeled objects from a naive distribution that assumes all 850 objects are equally likely to occur. The ratio of these likelihoods provides a measure of accuracy for the estimated object probabilities. Likelihood ratios greater than one indicate that the estimated object probabilities better predict the labeled objects in the scene than by picking objects at random (see Experimental Procedures for details).

Figure 4C shows the distribution of likelihood ratios for each subject, calculated for all 126 decoded scenes. The medians and 95% confidence intervals of the median estimates are indicated by the black cross-hairs. Object prediction accuracy across all scenes indicates systematically greater-than-chance performance for all subjects ($p < 1 \times 10^{-15}$ for all subjects, Wilcox rank-sum test; subject S1: $W(126)$ = 9,983; subject S2: $W(126)$ = 11,375; subject S3: $W(126)$ = 11,103; subject S4: $W(126)$ = 10,715).

The estimated object probabilities and the likelihood ratio analysis both show that the objects that are likely to occur in a scene can be predicted probabilistically from natural scene categories that are encoded in human brain activity. This suggests that humans might use a probabilistic strategy to help infer the likely objects in a scene from fragmentary information available at any point in time.

DISCUSSION

This study provides compelling evidence that the human visual system encodes scene categories that reflect the co-occurrence statistics of objects in the natural world. First, categories that capture co-occurrence statistics are consistent with our intuitive interpretations of natural scenes. Second, voxelwise encoding models based on these categories accurately predict visually evoked BOLD activity across much of anterior visual cortex, including within several conventional functional ROIs. Finally, the category of a scene and its constituent objects can be decoded from BOLD activity evoked by viewing the scene.

Previous studies of scene representation in the human brain used subjective categories that were selected by the experimenters. In contrast, our study used a data-driven, statistical algorithm (LDA) to learn the intrinsic categorical structure of natural scenes from object labels. These learned, intrinsic scene categories provide a more objective foundation for scene perception research than is possible using subjective categories.

One previous computer vision study used a similar statistical learning approach to investigate the intrinsic category structure of natural scenes (Fei-Fei and Perona, 2005). In that study, the input to the learning algorithm was visual features of intermediate spatial complexity. Because our goal was to determine whether the brain represents the object co-occurrence statistics of natural scenes, we used object labels of natural scenes as input to the learning algorithm rather than intermediate visual features.

The voxelwise modeling and decoding framework employed here (Kay et al., 2008b; Mitchell et al., 2008; Naselaris et al., 2009, 2012; Nishimoto et al., 2011; Thirion et al., 2006) provides a powerful alternative to conventional methods based on statistical parametric mapping (Friston et al., 1996) or multivariate pattern analysis (MVPA; Norman et al., 2006). Studies based on statistical mapping or MVPA do not aim to produce explicit predictive models of voxel tuning, so it is difficult to generalize their results beyond the specific stimuli or task conditions used in each study. In contrast, the goal of voxelwise modeling is to produce models that can accurately predict responses to arbitrary, novel stimuli or task conditions. A key strategy for developing theoretical models of natural systems has been to validate model predictions under novel conditions (Hastie et al., 2008). We believe that this strategy is also critically important for developing theories of representation in the human brain.

Our results generally corroborate the many previous reports of object selectivity in anterior visual cortex. However, we find that tuning properties in this part of visual cortex are more complex than reported in previous studies (see Figures S7, S8–S11, and S16–S19 for supporting results). This difference probably reflects the sensitivity afforded by the voxelwise modeling and decoding framework. Still, much work remains before we can claim a complete understanding of what and how information is represented in anterior visual cortex (Huth et al., 2012; Naselaris et al., 2012).

Several recent studies (Kim and Biederman, 2011; MacEvoy and Epstein, 2011; Peelen et al., 2009) have suggested that the lateral occipital complex (LO) represents, in part, the identity of scene categories based on the objects therein. Taken together, these studies suggest that some subregions within LO should be accurately predicted by models that link objects with scene categories. Our study employs one such model. We find that the encoding models based on natural scene categories provide accurate predictions of activity in anterior portions of LO (Figures 3A and 3B). Note, however, that our results do not necessarily imply that LO represents scene categories explicitly (see Figures S16–S19 for further analyses).

fMRI provides only a coarse proxy of neural activity and has a low SNR. In order to correctly interpret the results of fMRI experiments, it is important to quantify how much information can be recovered from these data. Here we addressed this problem by testing many candidate models in order to determine a single set of scene categories that can be recovered reliably from the BOLD activity measured across all of our subjects (Figure 2A). This test places a clear empirical limit on the number of scene categories and objects that can be recovered from our data. These numbers are larger than what has typically been assumed in previous fMRI studies of scene perception (Epstein and Kanwisher, 1998; Peelen et al., 2009; Walther et al., 2009; MacEvoy and Epstein, 2011), but they are still far smaller than the likely representational capacity of the human visual system.

Theoreticians have argued that the simple statistical properties of natural scenes explain selectivity to low-level features in peripheral sensory areas (Olshausen and Field, 1996; Smith and Lewicki, 2006). Behavioral data suggest that low-level natural scene statistics also influence the perception of scene categories (Oliva and Torralba, 2001; Torralba and Oliva, 2003). Though several qualitative theories have been proposed that link the object statistics of natural scenes with human scene perception (Biederman, 1981; Palmer, 1975), none have provided an objective, quantitative framework to support this link. The current study provides such a framework. Our data-driven, model-based approach shows that scene categories encoded in the human brain can be derived from the co-occurrence statistics of objects in natural scenes. This further suggests that the brain exploits natural scene statistics at multiple levels of abstraction. If this is true, then natural scene statistics might be used as a principled means to develop quantitative models of representation throughout the visual hierarchy.

The work reported here could be extended in several ways. For example, although the spatial distribution of objects within a scene appears to influence the representation of the scene (Biederman et al., 1982; Green and Hummel, 2006; Kim and Biederman 2011), the modeling framework used here makes no assumptions about the spatial distribution of objects within scenes. More sophisticated models that incorporate spatial statistics or other mediating factors such as attention may provide further information about the representation of scenes and scene categories in the human brain.

EXPERIMENTAL PROCEDURES

fMRI Data Acquisition

The experimental protocol used was approved by the UC Berkeley Committee for the Protection of Human Subjects. All fMRI data were collected at the UC Berkeley Brain Imaging Center using a 3 Tesla Siemens Tim Trio MR scanner (Siemens, Germany). For subjects S1, S3, and S4, a gradient-echo echo planar imaging sequence, combined with a custom fat saturation RF pulse, was used for functional data collection. Twenty-five axial slices covered occipital, occipitoparietal, and occipitotemporal cortex. Each slice had a 234 × 234 mm^2 field of view, 2.60 mm slice thickness, and 0.39 mm slice gap (matrix size = 104 × 104; TR = 2,009.9 ms; TE = 35 ms; flip angle = 74°; voxel size = 2.25 × 2.25 × 2.99 mm^3).

For subject S2 only, a gradient-echo echo planar imaging sequence, combined with a custom water-specific excitation (fat-shunting) RF pulse was used for functional data collection. In this case, 31 axial slices covered the entire brain, and each slice had a 224 × 224 mm^2 field of view, 3.50 mm slice thickness, and 0.63 mm slice gap (matrix size = 100 × 100; TR = 2,004.5 ms; TE = 33 ms; flip angle = 74°; voxel size = 2.24 × 2.24 × 4.13 mm^3).

Subject S1 experienced severe visual occlusion of the stimuli when the whole head coil was used. Therefore, for subject S1 the back portion (20 channels) of the Siemens 32 channel quadrature receive head coil was used as a surface coil. The full 32 channel head coil was used for subjects S2, S3, and S4.

Stimuli

All stimuli consisted of color images selected from a large database of natural scenes collected from various sources. Each image was presented on an isoluminant gray background and subtended the central 20° × 20° square of the visual field. Images were presented in successive 4 s trials. On each trial, a photo was flashed for 1 s at 5 Hz, followed by a 3 s period in which only the gray background was present. A central fixation square was superimposed at the center of the display, subtending 0.2° × 0.2° of the visual field. To facilitate fixation, we randomly permuted the fixation square in color (red, green, blue, white) at a rate of 3 Hz. No eye tracking was performed during stimulus presentation. However, all subjects in the study were highly trained psychophysical observers having extensive experience with fixation tasks, and preliminary data collected during an identical visual task showed that the subject cohort maintained stable fixation. Note also that the visual stimuli contained no object labels.

Experimental Design

fMRI experiments consisted of interleaved runs that contained images from separate model estimation and validation sets. Data were collected over six sessions for subjects S1 and S4, and seven sessions for subjects S2 and S3. Each of the 35 estimation set runs was 5.23 min in duration and consisted of 36 distinct images presented two times each. Evoked responses to these 1,260 images were used during model estimation. Each of 21 5.23-min-long validation set runs consisted of six distinct images presented 12 times each. The evoked responses to these 126 images were used during model validation. All images were randomly selected for each run with no repeated images across runs.

fMRI Data Processing

The SPM8 package (University College, London, UK) was used to perform motion correction, coregistration, and reslicing of functional images. All other preprocessing of functional data was performed using custom software (MATLAB, R2010a, MathWorks). Preprocessing was conducted across all sessions for each subject, using the first run of the first session as the reference. For each voxel, the preprocessed time series was used to estimate the hemodynamic response function (Kay et al., 2008a). Deconvolving each voxel time course from the stimulus design matrix produced an estimate of the response amplitude—a single value—evoked by each image, for each voxel. These response amplitude values were used in both model estimation and validation stages of data analysis. Retinotopic visual cortex was identified in separate scan sessions using conventional methods (Hansen et al., 2007). Standard functional localizers (Spiridon et al., 2006) were also collected in separate scan sessions and were used to identify the anatomical boundaries of conventional ROIs.

Learning Database and Stimulus Data Sets

Natural scene categories were learned using Latent Dirichlet Allocation (Blei et al., 2003; see Figure S1 for more details). The LDA algorithm was applied to the object labels of a learning database of 4,116 natural scenes compiled from two image data sets. The first image data set (Lotus Hill; Yao et al., 2007) provided 2,903 (71%) of the learning database scenes. The remaining scenes were sampled from an image data set that was created in house. In both data sets, all objects within the visible area of each image were outlined and labeled. Each in-house image was labeled by one of 15 naive labelers. Since each image was labeled by a single labeler, no labels were combined when compiling the databases. In a supplemental analysis, we verify that scene context created negligible bias in the statistics of the object labels (Figure S2). Ambiguous labels, misspelled labels, and rare labels having synonyms within the learning database were edited accordingly (see Supplemental Experimental Procedure 1). Note that the 1,260 stimulus scenes in the estimation set were sampled from the learning database. The validation set consisted of an independent set of 126 natural scenes labeled in house.

Voxelwise Encoding Modeling Analysis

Encoding models were estimated separately for each voxel using 80% of the responses to the estimation set stimuli selected at random. The model weights were estimated using regularized linear regression in order to best map the scene category probabilities for a stimulus scene onto the voxel responses evoked when viewing that scene. The category probabilities for a stimulus scene were calculated from the posterior distribution of the LDA inference procedure, conditioned on the labeled objects in the scene (see Supplemental Experimental Procedure 6 for details). Half of the remaining 20% of the estimation data was used to determine model regularization parameters and the other half of the estimation data was used to estimate model prediction accuracy (see Supplemental Experimental Procedure 7 for more details on encoding model parameter estimation).

Prediction accuracy estimates were used to determine the single best set of categories across subjects. For each of 760 different scene category settings (defining the number of distinct categories and vocabulary size assumed by LDA during learning), we calculated the number of voxels with prediction accuracy above a statistical significance threshold (correlation coefficient > 0.21; $p < 0.01$; see Supplemental Experimental Procedure 8 for details on defining statistically significant prediction accuracy). This resulted in a vector of 760 values for each subject, where each entry in the vector provided an estimate of the amount of cortical territory that was accurately predicted by encoding models based on each category setting. To combine the cortical territory estimates across subjects, we normalized the vector for each subject to sum to 1 (normalization was done to control for differences in brain size and signal-to-noise ratios across subjects) and the Hadamard (element-wise) product of the normalized vectors was calculated. This resulted in a combined distribution of 760 values (see Figure 2A). The peak of the combined distribution gave the single best set of categories across subjects. For more details on this issue, see Supplemental Experimental Procedure 9.

When calculating the proportion of response variance explained in each ROI by the encoding models, statistical significance was determined by permutation. Specifically, the proportion of variance explained was estimated using the responses to the validation set for each voxelwise encoding model. These explained variance estimates were then permuted across all cortical locations and the average was estimated within each functional ROI. Thus, each permutation produced a random sample of average explained variance within the boundaries of each functional ROI. Statistical significance was defined as the upper 99th percentile of the distribution of average explained variance estimates calculated within each ROI after 1,000 voxel permutations. For more details on this procedure, see Supplemental Experimental Procedure 10.

Decoding Analysis

Voxels were selected for the decoding analysis based on the predictive accuracy of their corresponding encoding models on the held-out estimation data set. To control for multiple comparisons during voxel selection, we defined the predictive accuracy threshold as a correlation coefficient greater than 0.34; $p < 5 \times 10^{-5}$, which is roughly the inverse of the number of cortical voxels in each subject. Using this criterion, 512 voxels were selected for subject S1, 158 for S2, 147 for S3, and 93 for S4.

Decoders were estimated using the selected voxels' responses to the scenes in the estimation set. Decoder weights were estimated using elastic-net-regularized multinomial regression (Friedman et al., 2010) using 80% of the estimation set data. The remaining 10% of the estimation responses were used to determine model regularization parameters. (The 10% of the estimation responses that were used to calculate encoding model prediction accuracies for voxel selection were not used to estimate the decoder.) After weight estimation, the decoders were used to predict the probability that each scene in the validation set belonged to each of the 20 best scene categories identified across subjects from the responses evoked within the selected population of voxels. For more details on the decoding parameter estimation, see Supplemental Experimental Procedure 13.

Decoder prediction accuracy for each scene was defined to be the correlation coefficient (Pearson's *r*) calculated between the category probabilities predicted by the decoder and the category probabilities inferred using LDA and conditioned on the objects that were labeled in each scene. Statistical significance of decoder prediction accuracy across all scenes was determined using a Wilcox rank-sum test comparing the distribution of decoder prediction accuracies to a null distribution of prediction accuracies. For more details, see Supplemental Experimental Procedures 13.

Using the category probabilities predicted by the decoder for each scene in the validation set, we repeatedly picked from the 850 objects comprising the object vocabulary for the 20 best scene categories identified across subjects. Each object was picked by first drawing a category index with probability defined by the decoded scene category probabilities, followed by picking an object label with probability defined by the learned LDA model parameters. The learned LDA model parameters capture the statistical correlations of the objects in the learning database. Thus, the frequency of an object being picked also obeyed this correlation. The frequency distribution resulting from 10,000 independent object label picks was then normalized. The result defined an estimated distribution of occurrence probabilities for the objects in the vocabulary. Statistical significance of object decoding accuracy across all scenes was determined using a Wilcox rank-sum test comparing the distribution of likelihood ratios for the decoder to a null distribution of likelihood ratios. For more details on this issue, see Supplemental Experimental Procedures 14.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and 19 figures and can be found with this article online at http://dx.doi.org/10.1016/j.neuron.2013.06.034.

ACKNOWLEDGMENTS

This work was supported by grants to J.L.G. from the National Eye Institute (EY019684), the National Institute of Mental Health (MH66990, and the National Science Foundation Center for the Science of Information (CCF-0939370). We thank An Vu for data collection assistance and Tom Griffiths, Shinji Nishimoto, Tolga Cukur, Mark Lescoarte, Michael Oliver, Alex Huth, James Gao, Natalia Bilenko, Anwar Nunez, Ben Dichter, and Melanie Miller for helpful discussions and comments.

Accepted: June 12, 2013
Published: August 8, 2013

REFERENCES

Biederman, I. (1972). Perceiving real-world scenes. Science *177*, 77–80.

Biederman, I. (1981). On the semantics of a glance at a scene. In Perceptual Organization, M. Kubovy and J.R. Pomerantz, eds. (Hillsdale: Lawrence Erlbaum), pp. 213–263.

Biederman, I., Glass, A.L., and Stacy, E.W., Jr. (1973). Searching for objects in real-world scences. J. Exp. Psychol. *97*, 22–27.

Biederman, I., Mezzanotte, R.J., and Rabinowitz, J.C. (1982). Scene perception: detecting and judging objects undergoing relational violations. Cognit. Psychol. *14*, 143–177.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). Latent dirichlet allocation. J. Mach. Learn. Res. *3*, 993–1022.

Buxton, R.B. (2002). Introduction to Functional Magnetic Resonance Imaging Book Pack: Principles and Techniques (Cambridge: Cambridge University Press).

Davenport, J.L., and Potter, M.C. (2004). Scene consistency in object and background perception. Psychol. Sci. *15*, 559–564.

Downing, P.E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. Science *293*, 2470–2473.

Epstein, R.A., and Kanwisher, N. (1998). A cortical representation of the local visual environment. Nature *392*, 598–601.

Fei-Fei, L., and Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005, *2*, 524–531.

Friedman, J.H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. *33*, 1–22.

Friston, K.J., Holmes, A., Poline, J.-B., Price, C.J., and Frith, C.D. (1996). Detecting activations in PET and fMRI: levels of inference and power. Neuroimage *4*, 223–235.

Gauthier, I., Tarr, M.J., Moylan, J., Skudlarski, P., Gore, J.C., and Anderson, A.W. (2000). The fusiform "face area" is part of a network that processes faces at the individual level. J. Cogn. Neurosci. *12*, 495–504.

Green, C.B., and Hummel, J.E. (2006). Familiar interacting object pairs are perceptually grouped. J. Exp. Psychol. Hum. Percept. Perform. *32*, 1107–1119.

Grill-Spector, K. (2003). The neural basis of object perception. Curr. Opin. Neurobiol. *13*, 159–166.

Hansen, K.A., Kay, K.N., and Gallant, J.L. (2007). Topographic organization in and near human visual area V4. J. Neurosci. *27*, 11896–11911.

Hasson, U., Harel, M., Levy, I., and Malach, R. (2003). Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron *37*, 1027–1041.

Hastie, T., Tibshirani, R., and Friedman, J.H. (2008). Model assessment and selection. The Elements of Statistical Learning: Data mining, Inference, and Prediction, Second Edition (New York: Springer), pp. 219–260.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.

Joubert, O.R., Rousselet, G.A., Fize, D., and Fabre-Thorpe, M. (2007). Processing scene context: fast categorization and object interference. Vision Res. *47*, 3286–3297.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. *17*, 4302–4311.

Kay, K.N., David, S.V., Prenger, R.J., Hansen, K.A., and Gallant, J.L. (2008a). Modeling low-frequency fluctuation and hemodynamic response timecourse in event-related fMRI. Hum. Brain Mapp. *29*, 142–156.

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

Kim, J., and Biederman, I. (2011). Where do objects become scenes? Cereb. Cortex *21*, 1738–1746.

MacEvoy, S.P., and Epstein, R.A. (2011). Constructing scenes from objects in human occipitotemporal cortex. Nat. Neurosci. *14*, 1323–1329.

Maguire, E.A. (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. Scand. J. Psychol. *42*, 225–238.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc. Natl. Acad. Sci. USA *92*, 8135–8139.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., Ito, K., Fukuda, H., Schormann, T., and Zilles, K. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study. Brain *123*, 1903–1912.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian reconstruction of natural images from human brain activity. Neuron *63*, 902–915.

Naselaris, T., Stansbury, D.E., and Gallant, J.L. (2012). Cortical representation of animate and inanimate objects in complex natural scenes. J. Physiol. Paris *106*, 239–249.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. *21*, 1641–1646.

Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

Oliva, A., and Torralba, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. Int. J. Comput. Vis. *42*, 145–175.

Olshausen, B.A., and Field, D.J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature *381*, 607–609.

Palmer, S.E. (1975). The effects of contextual scenes on the identification of objects. Mem. Cognit. *3*, 519–526.

Peelen, M.V., Fei-Fei, L., and Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature *460*, 94–97.

Potter, M.C. (1975). Meaning in visual search. Science *187*, 965–966.

Smith, E.C., and Lewicki, M.S. (2006). Efficient auditory coding. Nature *439*, 978–982.

Spiridon, M., Fischl, B., and Kanwisher, N. (2006). Location and spatial profile of category-specific regions in human extrastriate cortex. Hum. Brain Mapp. *27*, 77–89.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J.-B., Lebihan, D., and Dehaene, S. (2006). Inverse retinotopy: inferring the visual content of images from brain activation patterns. Neuroimage *33*, 1104–1116.

Torralba, A., and Oliva, A. (2003). Statistics of natural image categories. Network *14*, 391–412.

Walther, D.B., Caddigan, E., Fei-Fei, L., and Beck, D.M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. J. Neurosci. *29*, 10573–10581.

Yao, B., Yang, X., and Zhu, S.C. (2007). Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. A.L. Yuille, S.-C. Zhu, D. Cremers, and Y. Wang., eds. Proceedings of the 6th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, 169–183.

Neuron, Volume *79*

Supplemental Information

Natural Scene Statistics Account for the Representation of Scene Categories in Human Visual Cortex

Dustin E. Stansbury, Thomas Naselaris, and Jack L. Gallant

Supplemental Inventory

Supplemental Figures

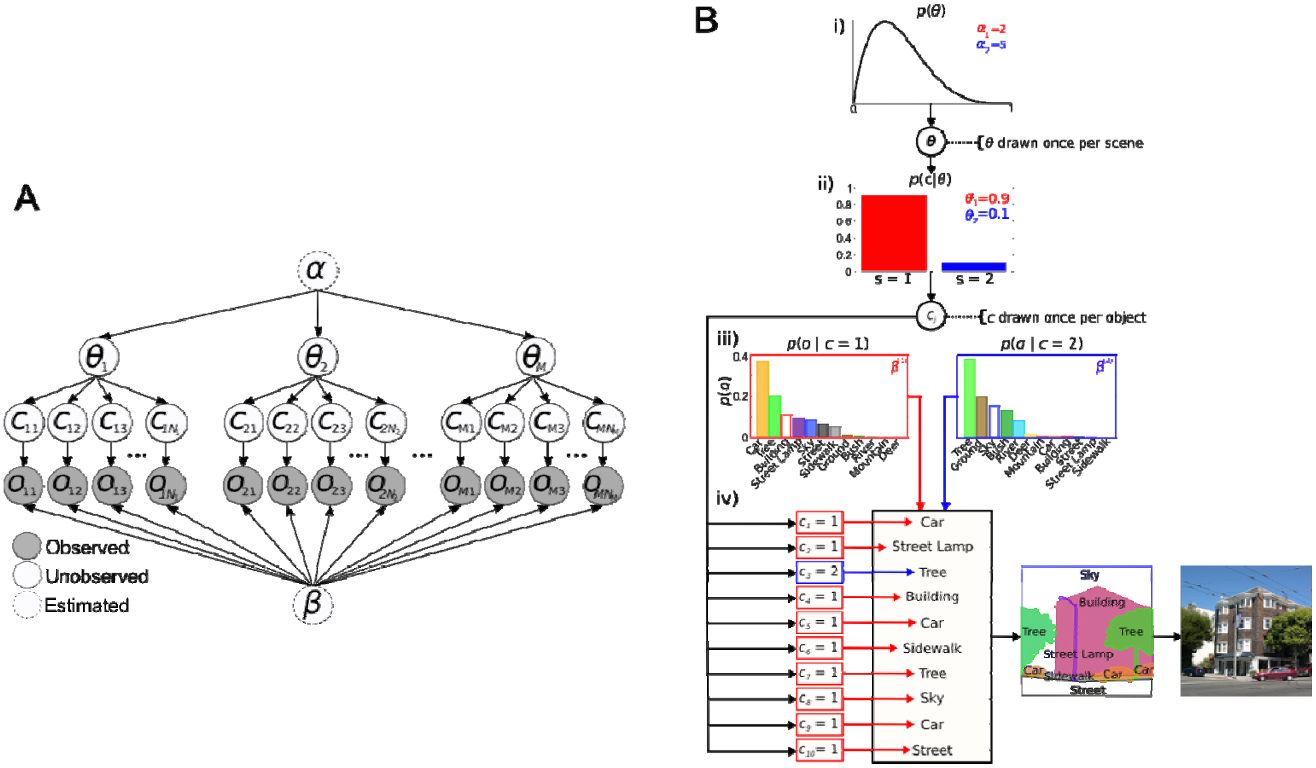


Figure S1. Description of Latent Dirichlet Allocation

(A) Graphical representation of the generative Latent Dirichlet Allocation model. For each of the M scenes that compose the learning database, it is assumed that the observed objects in the i-th scene are generated as follows: First, a parameter vector $\boldsymbol{\theta}_i$ ($i = 1...M$) is drawn independently from a Dirichlet distribution $p(\boldsymbol{\theta}; \boldsymbol{\alpha})$ that is parametrized by the vector $\boldsymbol{\alpha}$. Each $\boldsymbol{\theta}_i$ defines a scene-level multinomial probability distribution over the indices to K possible scene categories $p(\boldsymbol{c} = k \mid \boldsymbol{\theta}_i)$ ($k = 1...K$). Further, LDA assumes that each of N_i objects o_{ij} ($j = 1...N_i$) in the i-th scene are associated with the scene category index c_{ij} drawn from $p(\boldsymbol{c} \mid \boldsymbol{\theta}_i)$. For each of the K possible values that $\boldsymbol{c}$ can take, there is an associated object-level parameter vector $\boldsymbol{\beta}^{(k)}$ that defines a multinomial probability distribution $p(o = \boldsymbol{v}\, ; \boldsymbol{\beta}^{(k)})$ ($v = 1...V$) over V object indices. Here V is the size of the available object label vocabulary. Each of the objects in the i-th scene are then assumed to be generated by drawing an object index from the conditional distribution $p(o_{ij} = \boldsymbol{v} \mid c_{ij}; \boldsymbol{\beta})$. During category learning, the goal is to estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ that would have generated the object labels observed in the learning database.

(B) A toy example demonstrating how LDA would generate a scene comprised of 10 objects. **i)** Sample a scene-level parameter $\boldsymbol{\theta}$, which defines the probability $p(\boldsymbol{c} \mid \boldsymbol{\theta})$ over possible scene category indices. **ii)** Sample $N = 10$ different scene category indices $c_j = k$ (k taking values of either 1 or 2) independently from $p(\boldsymbol{c} \mid \boldsymbol{\theta})$. **iii)** For each of the 10 scene category indices drawn, associate one of two the object-level distributions $p(\boldsymbol{o} \mid c_j; \boldsymbol{\beta}^{(k)})$ with a to-be generated object. **iv)** Generate the 10 objects in the scene by sampling each according to corresponding object distribution $p(o_j = \boldsymbol{v} \mid c_j, \boldsymbol{\beta})$ associated with the object via c_j. The images at the bottom right display a scene generated by this process.

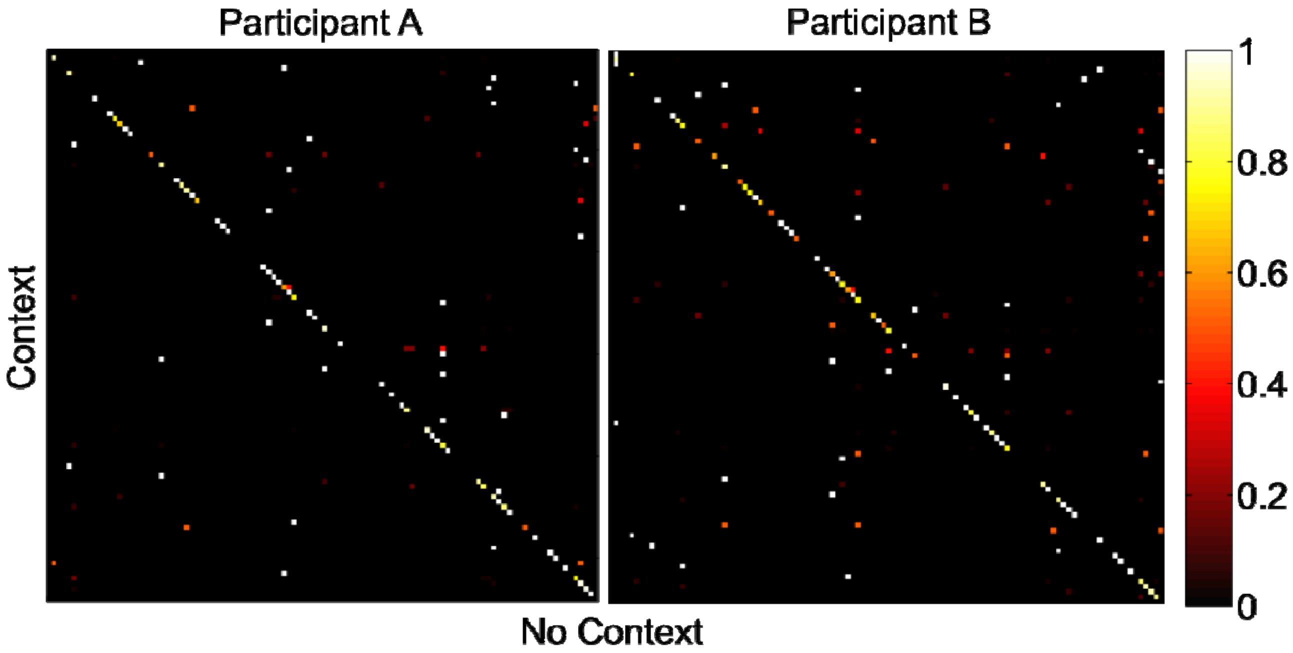


Figure S2. Assessing the importance of scene context on object labeling

Before compiling the learning database we sought to ensure that scene context did not affect the labels assigned to objects, which would have biased the statistics of the database. To address this issue we ran a control experiment in which two naïve labelers relabeled the objects in 106 representative scenes sampled from the learning database, but in which the context had been removed by masking the remainder of each scene. The confusion matrices shown here give the correspondence between labels assigned to the 104 distinct objects across the 106 sampled scenes with (vertical axis) and without (horizontal axis) scene context. The rows of the confusion matrix are normalized to indicate the proportion of the times that each of 104 objects was given a particular label in this control. The strong diagonal in each confusion matrix indicates high correspondence between labels assigned with and without scene context. The correlation coefficient between the object label indicator vectors for the context and no context conditions is 0.78 for labeler A and 0.76 for labeler B ($p < 0.01$, calculated analogously to Supplemental Experimental Procedure 8).

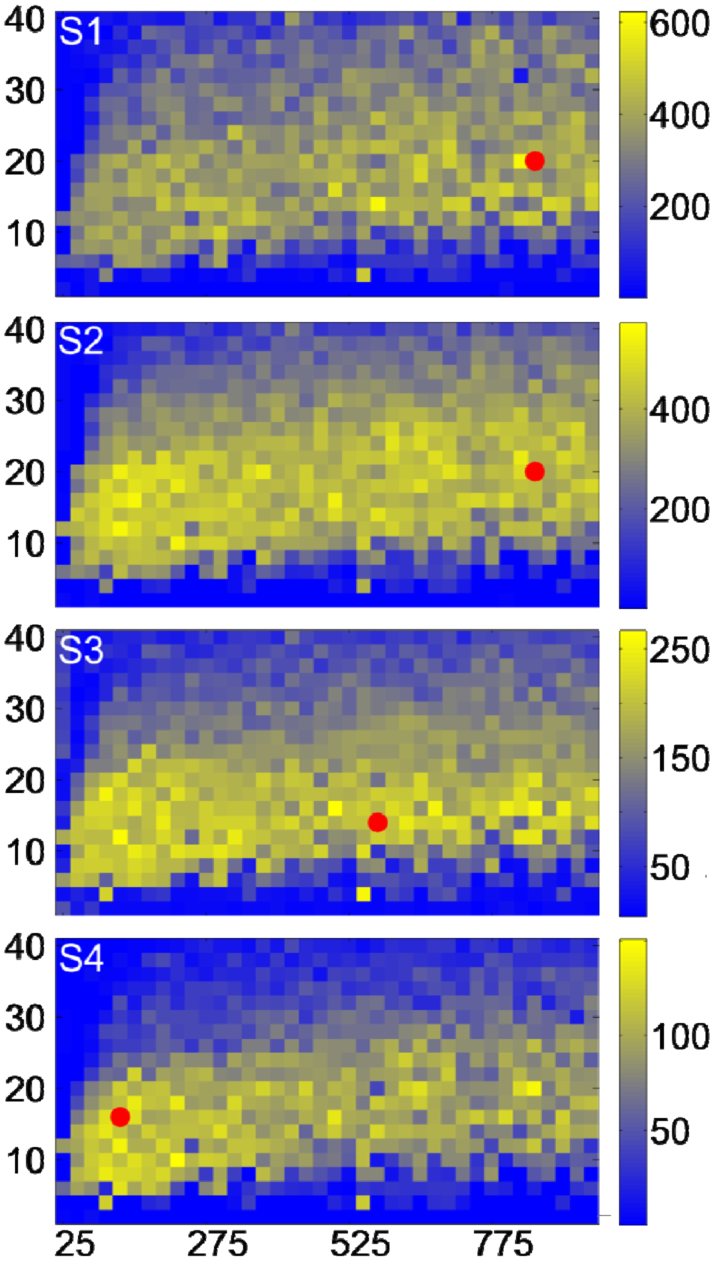
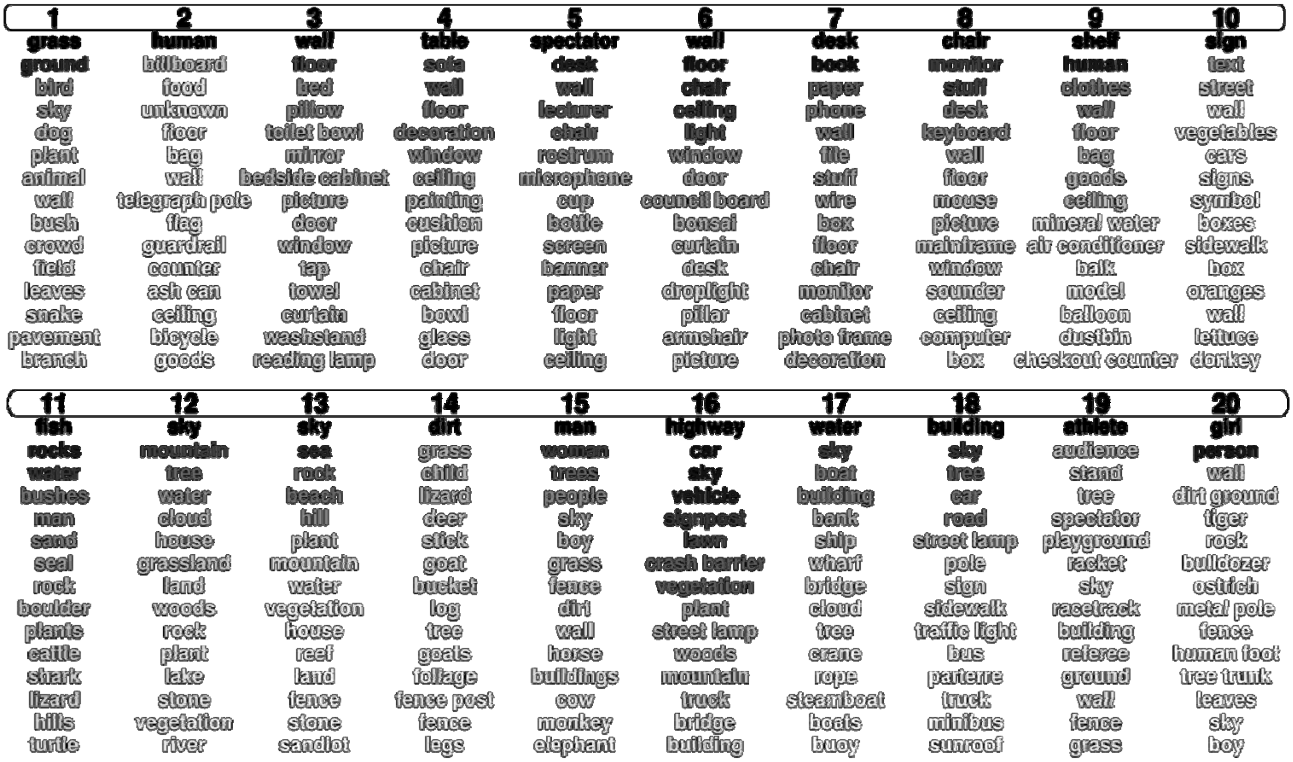


Figure S3. Number of voxel-wise encoding models with significant prediction accuracy, based on each of the candidate scene categories, individual subject results.

LDA was used to learn many candidate sets of scene categories by systematically varying both the number of distinct categories (2-40) and the size of the object vocabulary (25-950) assumed during category learning.

The plots display the number of voxels whose activity was accurately predicted ($p < 0.01$) by encoding models based on a given setting of the number of scene categories (y-axis) and vocabulary size (x-axis). Subjects are identified in white in the upper left of each plot. Red dots indicate the number of categories and vocabulary size that resulted in the largest number of voxels whose activity was predicted significantly.

For subjects S1 and S2 the candidate scene categories that resulted in the largest number of accurately-predicted voxels were the same as the best set of scene categories determined across subjects (Figure 2A in the main text). For subject S3, the candidate categories that resulted in the largest number of significant predictions consisted of 14 distinct categories comprised of a vocabulary of 575 objects. For subject S4, the set of categories producing the largest number of significant predictions was 16 distinct categories composed of a vocabulary of 125 objects.



1	2	3	4	5	6	7	8	9	10
grass	human	wall	table	spectator	wall	desk	chair	shelf	sign
ground	billboard	floor	sofa	desk	floor	book	monitor	human	text
bird	food	bed	wall	wall	chair	paper	stuff	clothes	street
sky	unknown	pillow	floor	lecturer	ceiling	phone	desk	wall	wall
dog	floor	toilet bowl	decoration	chair	light	wall	keyboard	floor	vegetables
plant	bag	mirror	window	rostrum	window	file	wall	bag	cars
animal	wall	bedside cabinet	ceiling	microphone	door	stuff	floor	goods	signs
wall	telegraph pole	picture	painting	cup	council board	wire	mouse	ceiling	symbol
bush	flag	door	cushion	bottle	bonsai	box	picture	mineral water	boxes
crowd	guardrail	window	picture	screen	curtain	floor	mainframe	air conditioner	sidewalk
field	counter	tap	chair	banner	desk	chair	window	balk	box
leaves	ash can	towel	cabinet	paper	droplight	monitor	sounder	model	oranges
snake	ceiling	curtain	bowl	floor	pillar	cabinet	ceiling	balloon	wall
pavement	bicycle	washstand	glass	light	armchair	photo frame	computer	dustbin	lettuce
branch	goods	reading lamp	door	ceiling	picture	decoration	box	checkout counter	donkey

11	12	13	14	15	16	17	18	19	20
fish	sky	sky	dirt	man	highway	water	building	athlete	girl
rocks	mountain	sea	grass	woman	car	sky	sky	audience	person
water	tree	rock	child	trees	sky	boat	tree	stand	wall
bushes	water	beach	lizard	people	vehicle	building	car	tree	dirt ground
man	cloud	hill	deer	sky	signpost	bank	road	spectator	tiger
sand	house	plant	stick	boy	lawn	ship	street lamp	playground	rock
seal	grassland	mountain	goat	grass	crash barrier	wharf	pole	racket	bulldozer
rock	land	water	bucket	fence	vegetation	bridge	sign	sky	ostrich
boulder	woods	vegetation	log	dirt	plant	cloud	sidewalk	racetrack	metal pole
plants	rock	house	tree	wall	street lamp	tree	traffic light	building	fence
cattle	plant	reef	goats	horse	woods	crane	bus	referee	human foot
shark	lake	land	foliage	buildings	mountain	rope	parterre	ground	tree trunk
lizard	stone	fence	fence post	cow	truck	steamboat	truck	wall	leaves
hills	vegetation	stone	fence	monkey	bridge	boats	minibus	fence	sky
turtle	river	sandlot	legs	elephant	building	buoy	sunroof	grass	boy

Figure S4. Best scene categories for modeling data across subjects
As shown in Figure 2A, encoding model performance across subjects was most accurate when the total number of scene categories for the LDA algorithm was set to 20, and the object vocabulary was set to the 850 most frequently occurring words in the learning database. Displayed here are those 20 scene categories represented in a format analogous to Figure 1C in the main text. Each list of object labels corresponds to the objects with the highest probability of occurring in each of the 20 categories. The saturation of each label represents the relative probability of the corresponding object. For further interpretation of these scene categories see Figure S5.

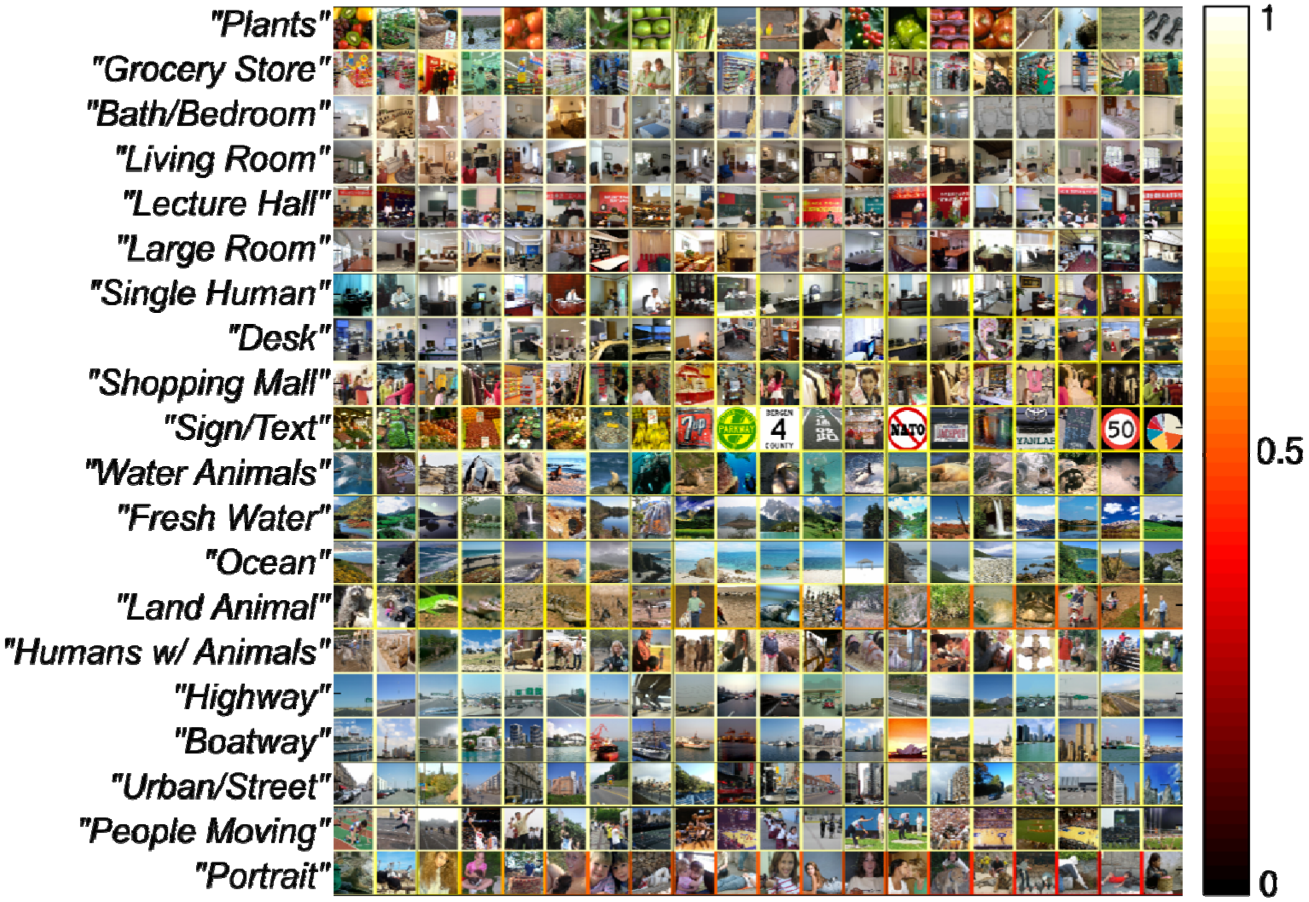


Figure S5. Interpreting the best scene categories for modeling data across subjects

The scene categories investigated in the current study were learned by applying LDA to the object labels in the learning database. To facilitate the analyses we inspected the scenes from the learning database that were assigned the highest posterior probability under each category, and then we assigned an appropriate label to each scene. Each row displays the scenes assigned the 20 largest posterior probabilities under each of the categories estimated by LDA (The top row corresponds to column 1 in Figure S4 and the bottom row column 20 Figure S4.) The relative probability of each image is indicated by the color of the border around the image. The interpretive category labels assigned to each category are displayed in quotes on the left.

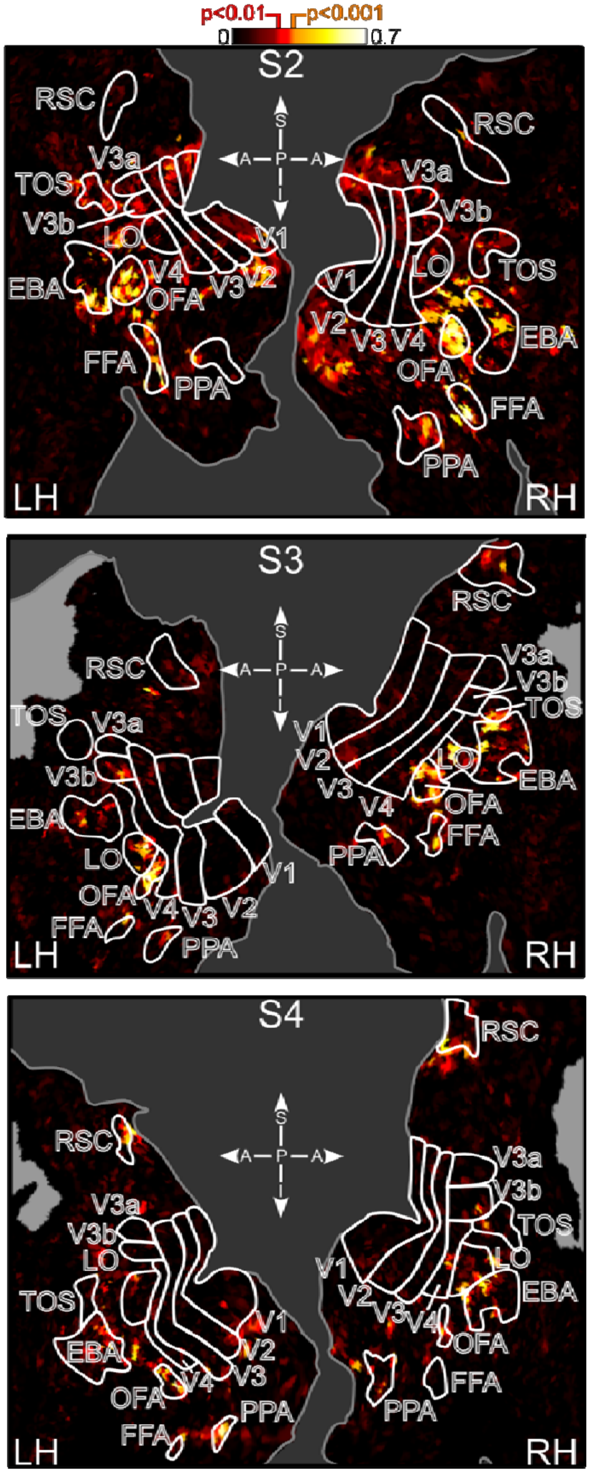


Figure S6. Learned scene categories are encoded in many anterior visual ROIs, subjects S2-S4

Encoding model prediction accuracies plotted on the left (LH) and right (RH) hemisphere cortical surfaces of subjects S2 through S4. Top row: data for subject S2; middle row: data for subject S3; bottom row: data for subject S4. ROIs, color map scale, and spatial organization of each panel is analogous to Figure 3A in the main text. Bright locations indicate voxels that are accurately predicted. Prediction accuracy at two levels of statistical significance, $p < 0.1$ (0.21) and $p < 0.001$ (0.28), are identified above the color bar. Prediction accuracies were calculated on responses to the scenes in the validation set.

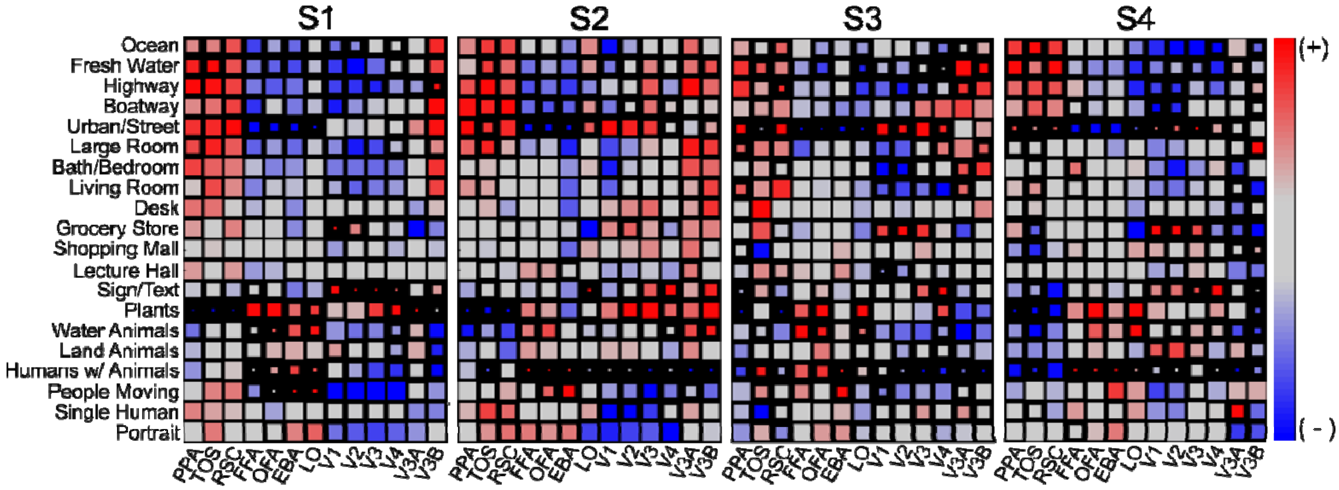


Figure S7. Average encoding model weights for voxels within functional ROIs, individual subject results.

Average encoding model weights for each scene category (rows; category listed at left) are plotted for 12 functional regions of interest (columns; ROIs listed at bottom). Averages are calculated across all voxels within an ROI. The color of each pixel represents the positive (red) or negative (blue) average ROI weight for the corresponding category. The size of each pixel is inversely proportional to the standard error of the mean, so larger pixels indicate more accurate estimates of selectivity. Category tuning for all subjects is consistent with, but more complex than the tuning revealed by conventional methods for identifying ROIs.

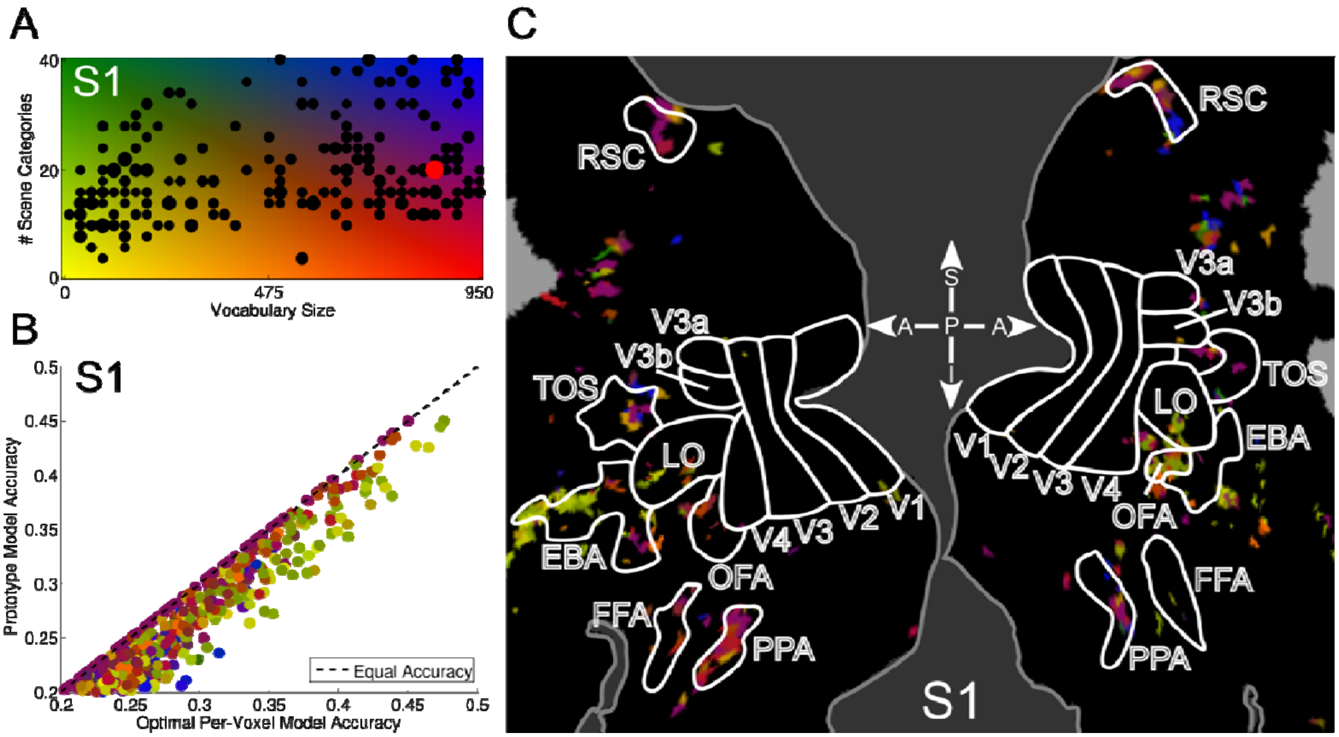


Figure S8. Encoding models based on scene categories and vocabulary optimized for individual voxels, subject S1.

(A) Each black point in the scatter plot represents the number of distinct categories (y-axis) and the vocabulary size (x-axis) that maximized encoding model predictions for individual voxels. The size of each point represents the relative prediction accuracy for the corresponding voxel-wise encoding model; larger points indicate more accurate models. The red dot represents the single best set of scene categories identified across subjects (see Figure 2A in the main text). Scene categories optimized for individual voxels are distributed smoothly across the 2D space of candidate scene categories. Generally, voxels that prefer fewer scene categories also prefer categories composed of smaller object vocabularies. Analyses here are limited to voxels whose responses were predicted significantly (correlation > 0.21; p < 0.01, see Supplemental Experimental Procedure 8 for details).

(B) Comparison of prediction accuracy for encoding models based on the optimal categories for individual voxels (horizontal axis) and the single best (vertical axis) set of scene categories across all subjects (Figures S3-S5). Each point compares the models for one of the voxels represented in panel (A). The color of each point indicates its position in A. Points that lie along the dashed diagonal line indicate voxels that are predicted equally well under both classes of models. The vertical distance below the dashed line indicates the improvement in model prediction accuracy achieved by using scene categories optimized for the individual voxel (note that all points are guaranteed to lie on or below the line at unity). The prediction accuracies for the two classes of models are highly correlated (*r*=0.92.) Thus, optimizing scene categories for each individual voxel confers little advantage over using the single best set across all subjects and voxels. In cases where the optimal number of categories and vocabulary size for an individual voxel are the same as the single best number of categories and vocabulary size there is no advantage at all (see magenta and orange-colored points along the line at unity).

(C) Cortical surface distribution of scene categories optimized for individual voxels. The color at each location corresponds to the color of the background of panel (A). Black areas indicate voxels whose responses were not predicted significantly. Gray areas indicate regions outside of the fMRI acquisition window. The boundaries of functional localizers identified in a separate scan sessions are outlined in white. Flat map spatial orientations and ROI abbreviations are analogous to those described in Figure 3A in the main text. Voxels that encode fewer individual categories and smaller vocabularies, as indicated by yellow and green, generally occupy animate-related functional ROIs FFA, OFA, EBA. The voxels that encode many individual categories and larger object vocabularies, as indicated by blue, magenta, and orange, generally occupy the inanimate-related functional ROIs PPA, RSC, and TOS.

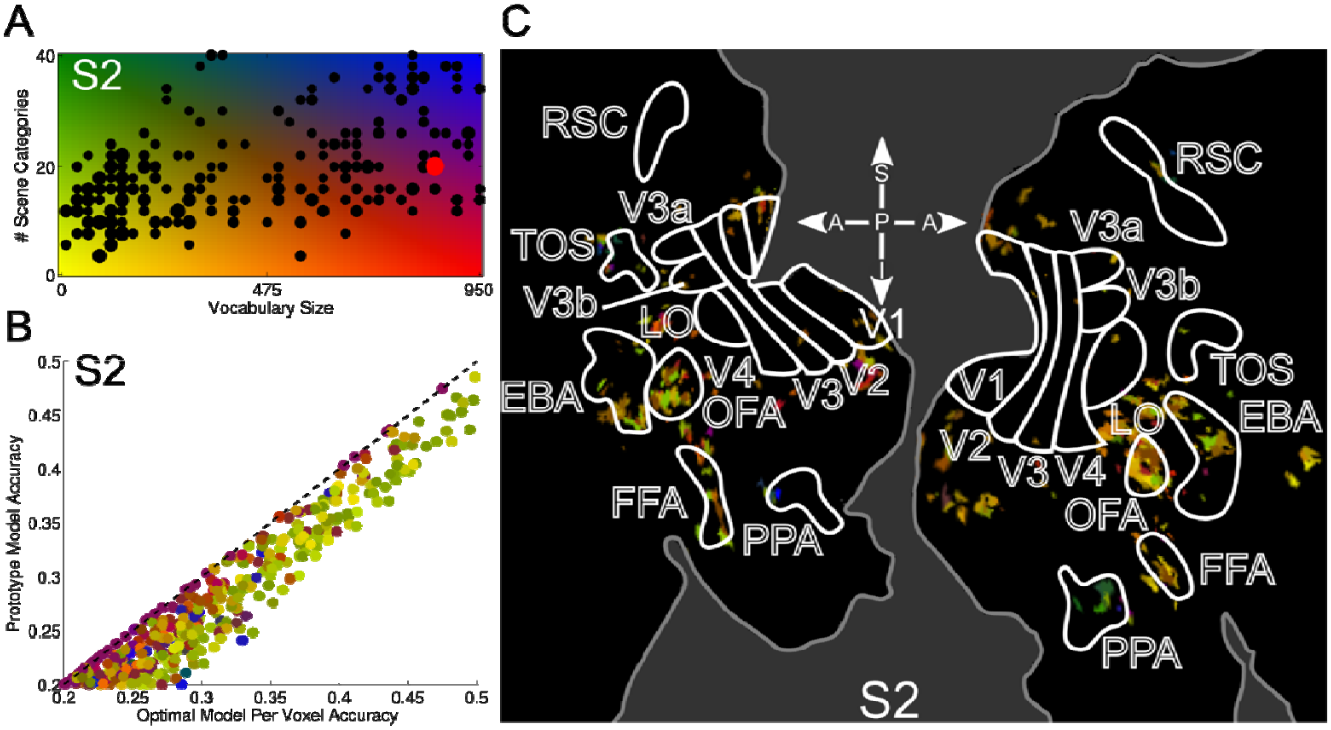


Figure S9. Encoding models based on scene categories and vocabulary optimized for individual voxels, subject S2.

Figure organized analogously to Figure S8. The points in (B) are highly correlated (*r*=0.96). These results indicate comparable performance between the two classes of models. (C) shows that a significant proportion of animate-selective regions FFA, OFA, and EBA populated by yellow and green voxels.

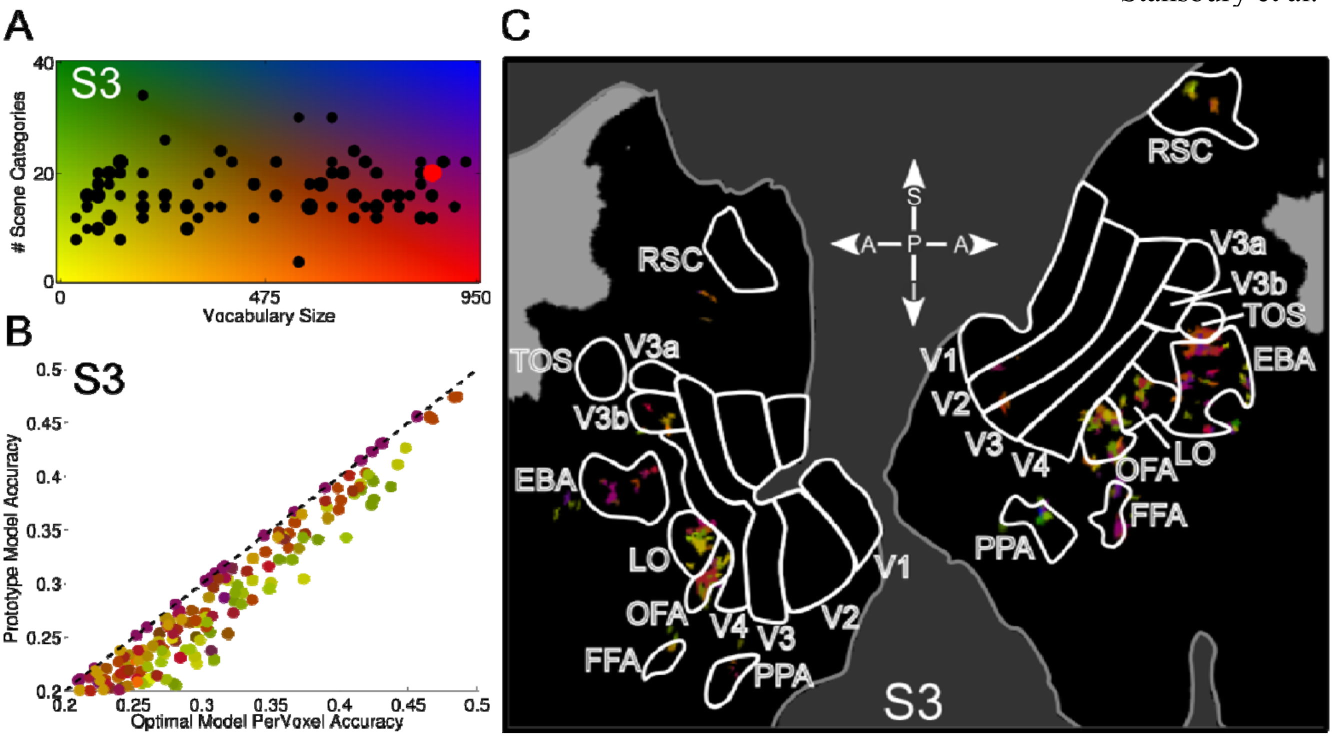


Figure S10. Encoding models based on scene categories and vocabulary optimized for individual voxels, subject S3.

Figure organized analogously to Figure S8. The points in (B) are highly correlated (*r*=0.96). These results indicate comparable performance between the two classes of models. (C) OFA is populated by yellow, green, and magenta-colored voxels.

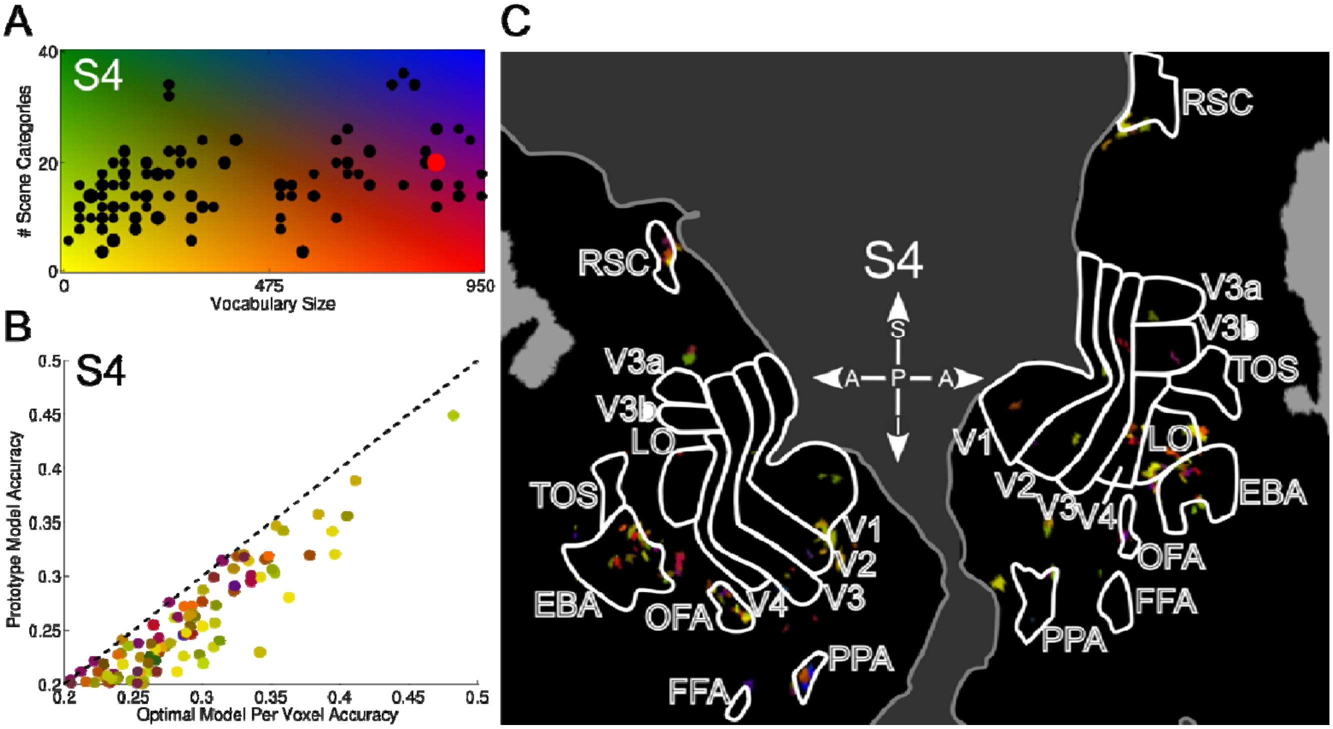


Figure S11. Encoding models based on scene categories and vocabulary optimized for individual voxels, subject S4.

Figure organized analogously to Figure S8. The points in (B) are highly correlated (*r*=0.92). These results indicate comparable performance between the two classes of models.

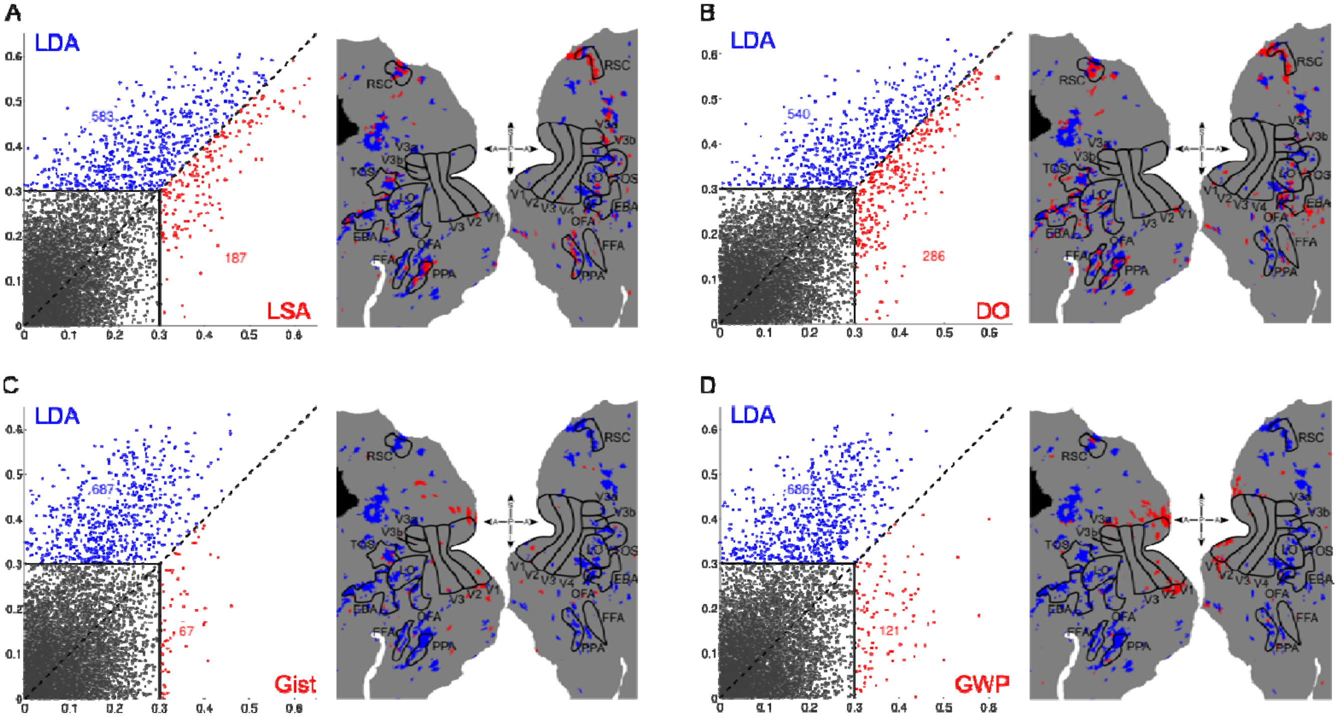


Figure S12. Comparing alternative hypotheses via voxel-wise encoding models, subject S1
Each panel compares prediction accuracies of the LDA-based encoding models to alternative encoding models based on other plausible feature representations. For details on the hypotheses motivating these alternative feature representations, see Supplemental Experimental Procedure 12.

(A) LDA vs. Latent Semantic Analysis (LSA). LSA represents each scene as as a linear projection of the object frequencies in the scene onto a set of orthogonal feature dimensions. (B) LDA vs. Diagnostic Objects (DO). The DO model is derived directly from the LDA model. Under the DO model, each scene is represented as a vector of counts. The counts indicate the number of times that the most probable object for each of the LDA categories occurs in a scene. (C) LDA vs. the Gist image descriptor. Gist descriptors (Oliva & Torralba, 2001) capture global image structure and are widely used in computer vision for scene classification. (D) LDA vs. Gabor Wavelet Pyramid (GWP). The GWP-based encoding models capture spatially-localized contrast energy at various spatial scales and orientations (Kay et al, 2008).

The left scatter plot in each panel plots prediction accuracies of the LDA-based encoding model for multiple voxels against the prediction accuracy of the corresponding alternative encoding models for the same voxels (LDA-based models are based on the single best set of categories identified across subjects). Blue points indicate voxels for which the LDA-based models provide more accurate predictions. Red points indicate voxels for which the alternative models provide more accurate predictions. Gray points indicate voxels with insignificant predictions under either model class (Pearson's r score equal to 0.31, $p < 3\times10^{-4}$, see Supplemental Experimental Procedure 8). The ratio of the number of blue points to the number of red points is used as a metric to compare the general performance of the LDA-based models to the alternative models. Comparison ratios greater than one indicate that the LDA-based models accurately predict the activity of more cortical territory than the alternative models. Ratios less than one indicate that the alternative models are generally more accurate across the cortical surface. The ratios for comparing the LSA-, DO-, Gist-, and GWP-based models are 3.12, 1.89, 10.25, and 5.67, respectively. These results indicate that the LDA-based models generally perform better than the alternative model classes.

The right plot in each panel displays the cortical locations of the voxels represented in the scatter plot in the left of the panel. The color of each voxel corresponds to the color of the same voxel in the scatter plot at left. Black locations represent areas outside of the fMRI acquisition window. The boundaries of function ROIs identified in separate scan sessions are outlined in black. Flatmap orientation and ROI abbreviations are analogous to those described in Figure 3A of the main text. Models based on LSA provide more accurate predictions in areas PPA and RSC, while models based on LDA perform better in other portions of anterior visual cortex. Models based on DO generally predict best in RSC and portions of PPA, FFA, and anterior EBA. Models based on Gist and GWP provide accurate predictions only in early visual cortex.

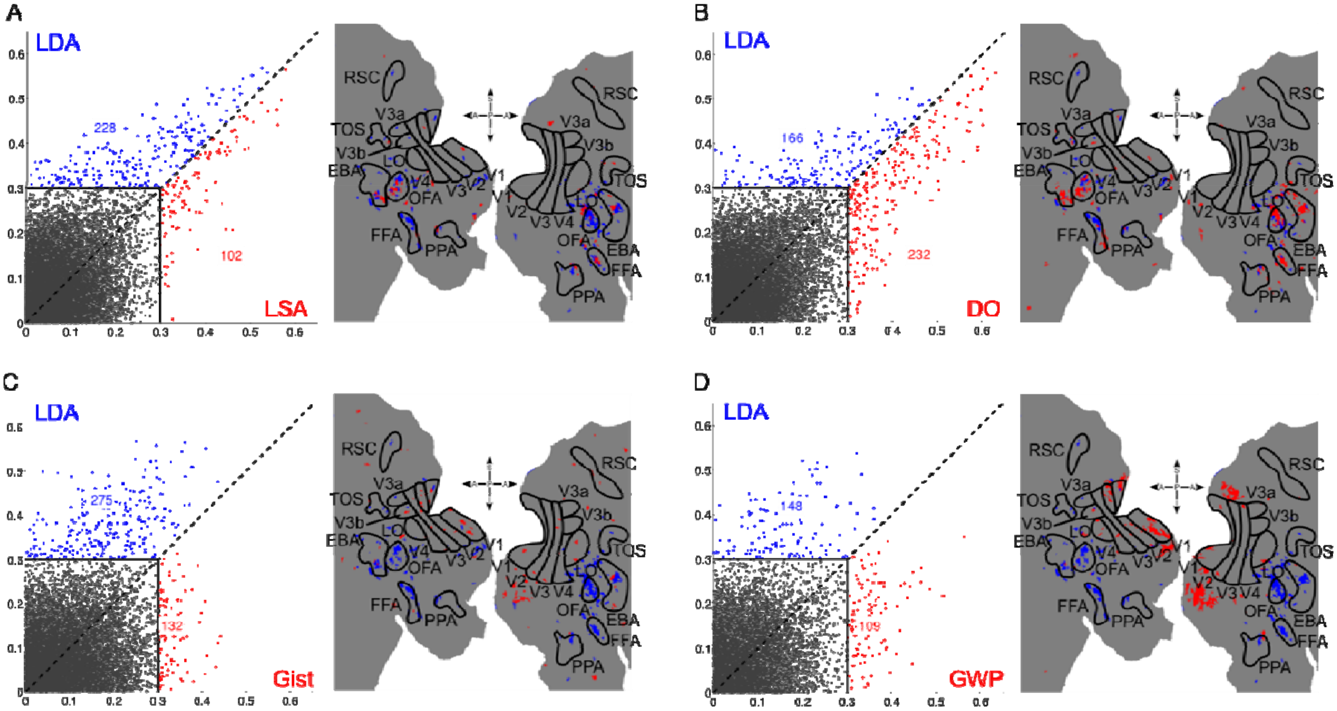


Figure S13. Comparing alternative hypotheses via voxel-wise encoding models, subject S2

Figure organized analogously to Figure S12. The ratios for comparisons between the LSA, DO, Gist, and GWP models are 2.34, 0.72, 2.08, and 1.34, respectively.

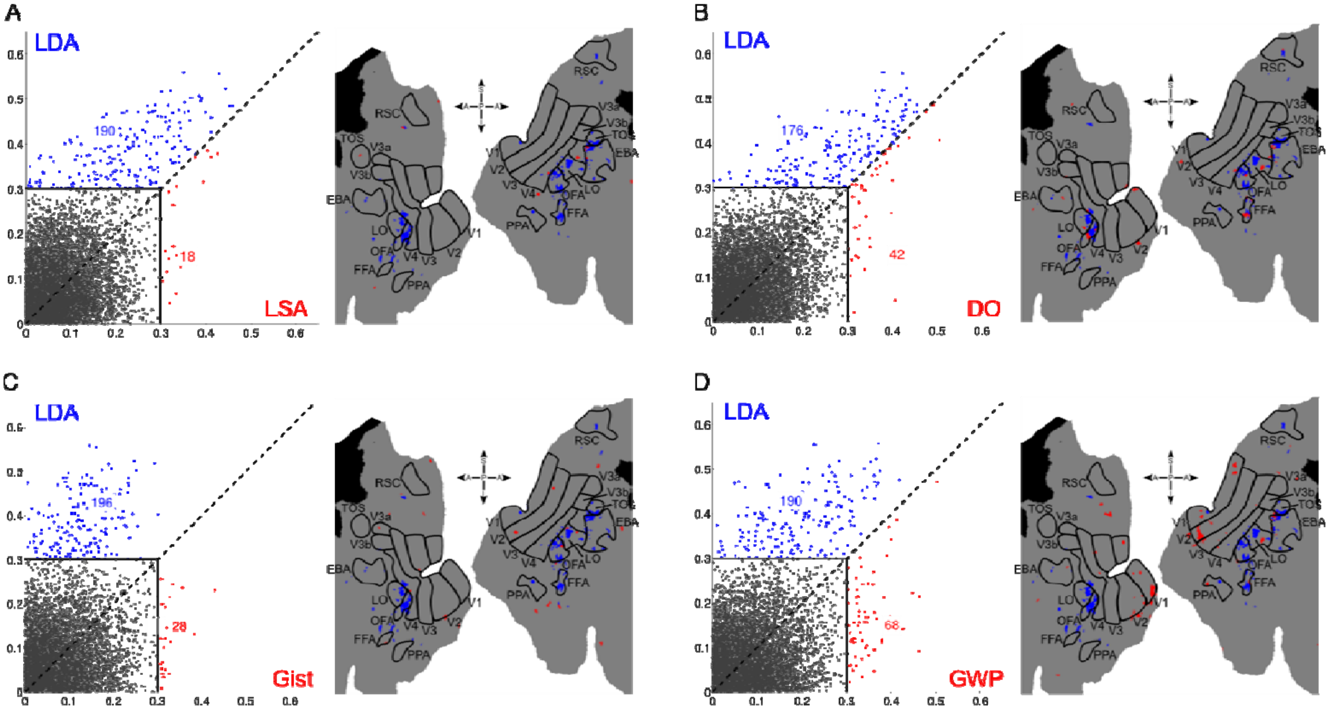


Figure S14. Comparing alternative hypotheses via voxel-wise encoding models, subject S3

Figure organized analogously to Figure S12. The ratios for comparisons between the LSA, DO, Gist, and GWP models are 10.56, 4.19, 7.00, and 2.21, respectively.

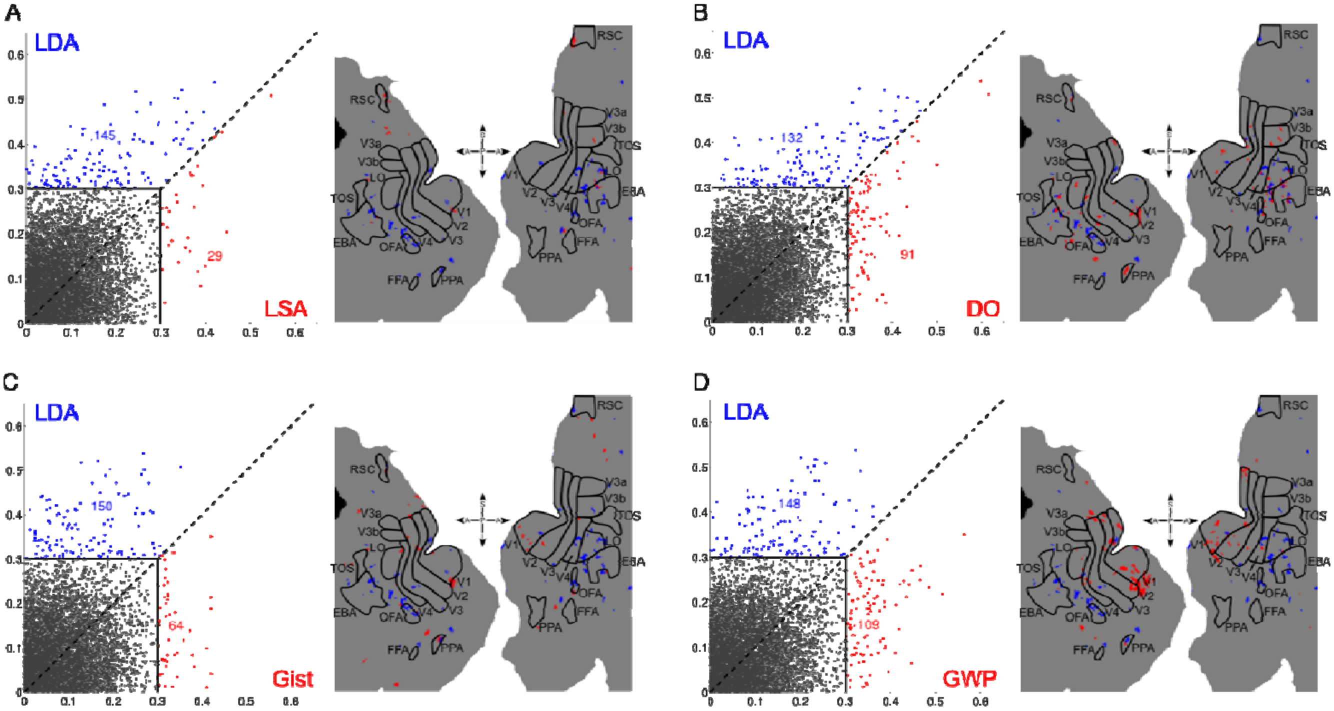


Figure S15. Comparing alternative hypotheses via voxel-wise encoding models, subject S4

Figure organized analogously to Figure S12. The ratios for comparisons between the LSA, DO, Gist, and GWP models are 5.00, 1.45, 2.34, and 1.36, respectively.

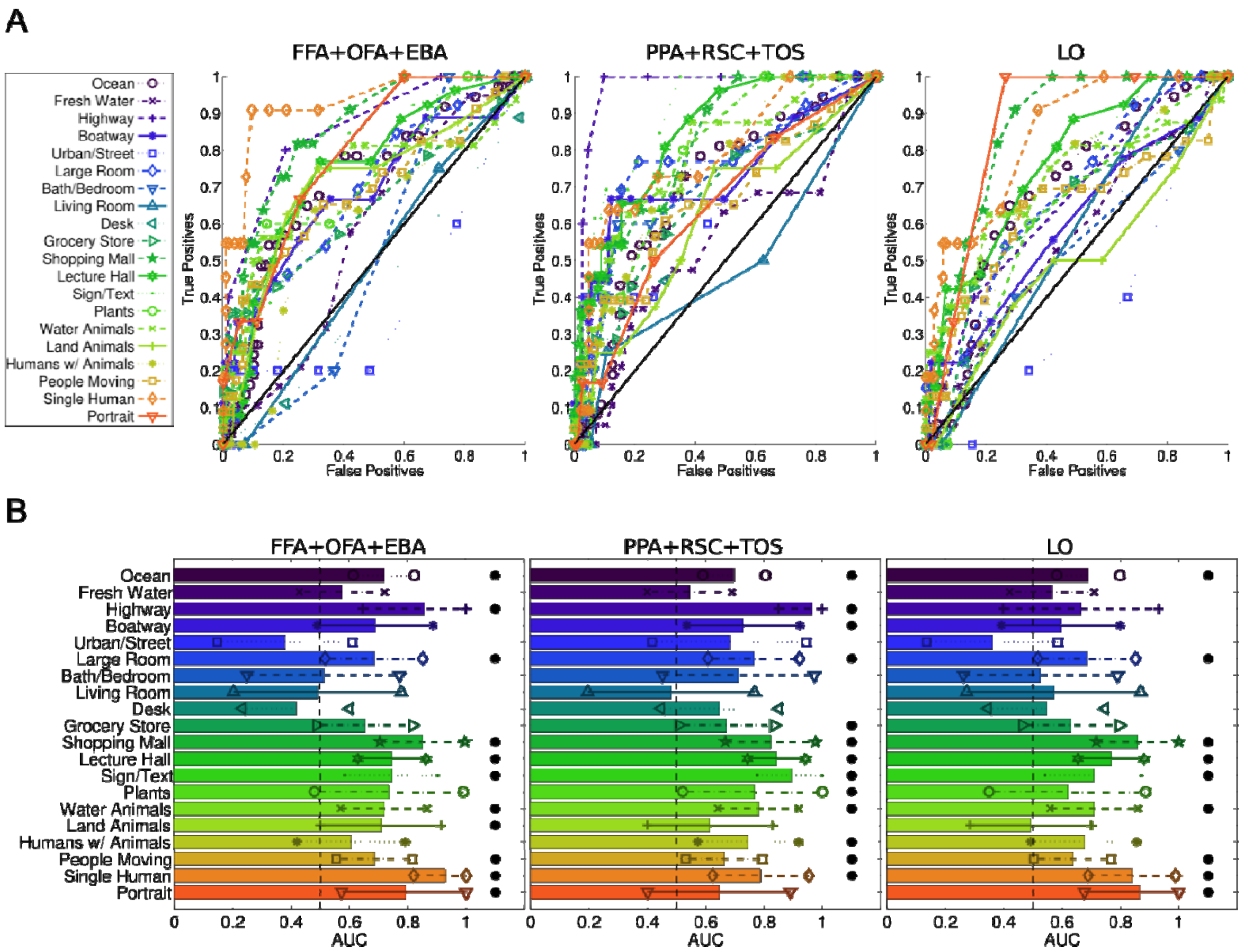


Figure S16. Decoding scene categories from various ROIs, subject S1.

Signal detection analysis was used to assess how well group responses from specific ROIs could decode learned scene categories (*i.e.*, the single best set of categories determined across subjects). Analyses focused on three non-overlapping groups of ROIs. The first group of ROIs (left column) has been associated with selectivity for animate features such as faces or body parts. These include FFA, OFA, EBA (number of voxels, n = 123). The second group of ROIs (center column) has been associated with selectivity for features related to spatial orientation and navigation. These include PPA, RSC, and TOS (n = 82). The third group of ROIs (right column) has been associated with coding of individual objects, and consists only of LO (n = 44). ROI abbreviations are the same as those described in Figure 3 in the main text.

(A) Receiver Operator Characteristic (ROC) curves for each scene category and each ROI-specific decoder. Results for individual categories are specified by line color and marker, as indicated by the key at the far left. ROC curves located far above the diagonal black line indicate that the decoder estimated from the corresponding ROIs accurately detects the corresponding category. **(B)** Area under each ROC curve (AUC). The AUC summarizes decoding performance for each scene category. Each column gives the results for a group of ROIs. Error bars give the standard error on the AUC estimates.

Black dots along the right indicate that the corresponding scene category was detected significantly (AUC > 0.5, dashed black line). Each of the decoders is able to detect a wide range of different categories. Many of the categories that are detected most accurately are consistent with the domain-specific selectivity commonly associated with the ROIs. However, some ROI-specific decoders provide accurate detection of categories that are not commonly associated with those ROIs.

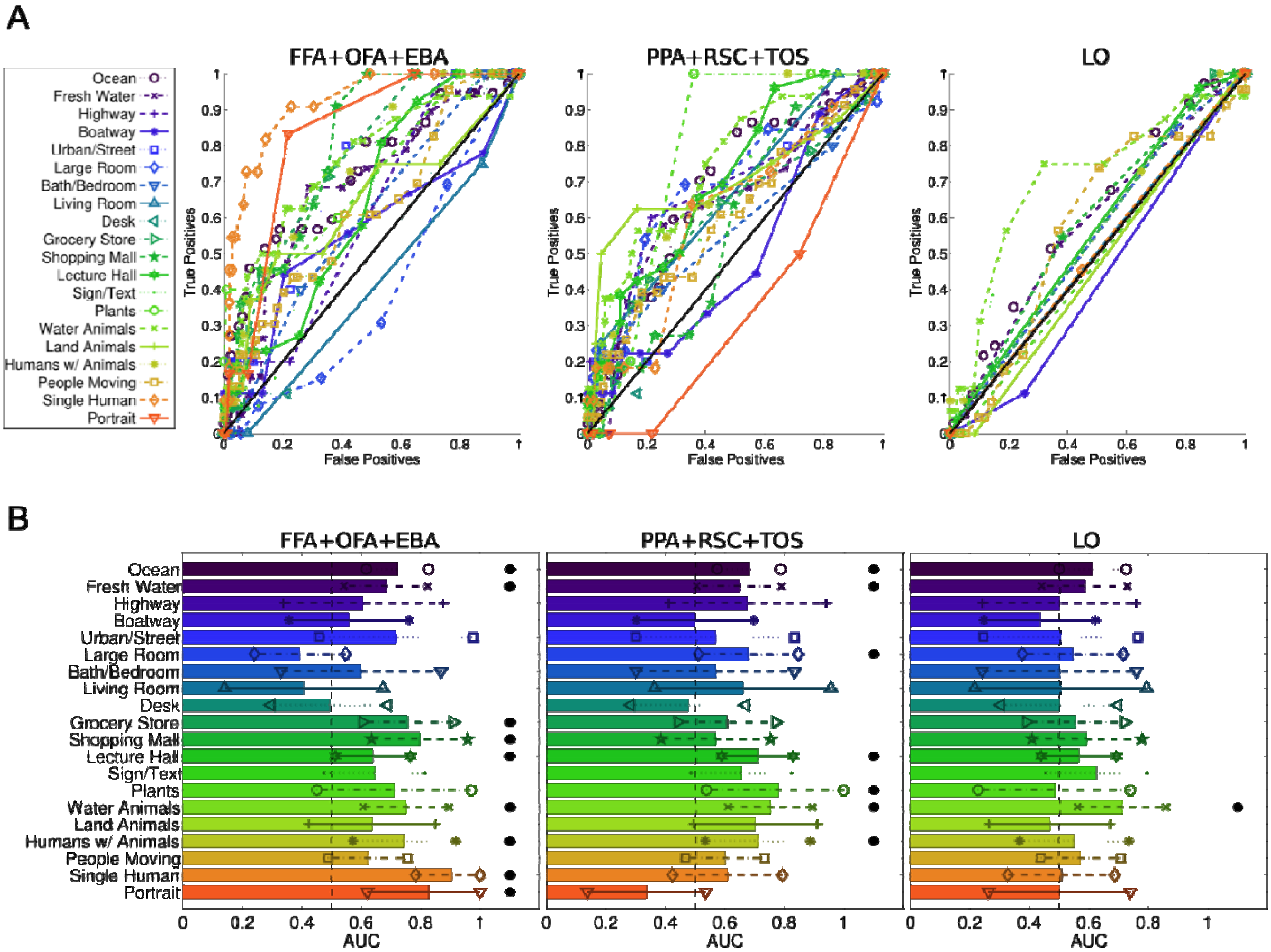


Figure S17. Decoding scene categories from various ROIs, subject S2

Figure organized analogously to Figure S16.The number of voxels in ROIs FFA, OFA, and EBA were *n* = 180. For the ROIs PPA, RSC, and TOS, *n* = 31. For area LO, *n* = 3.

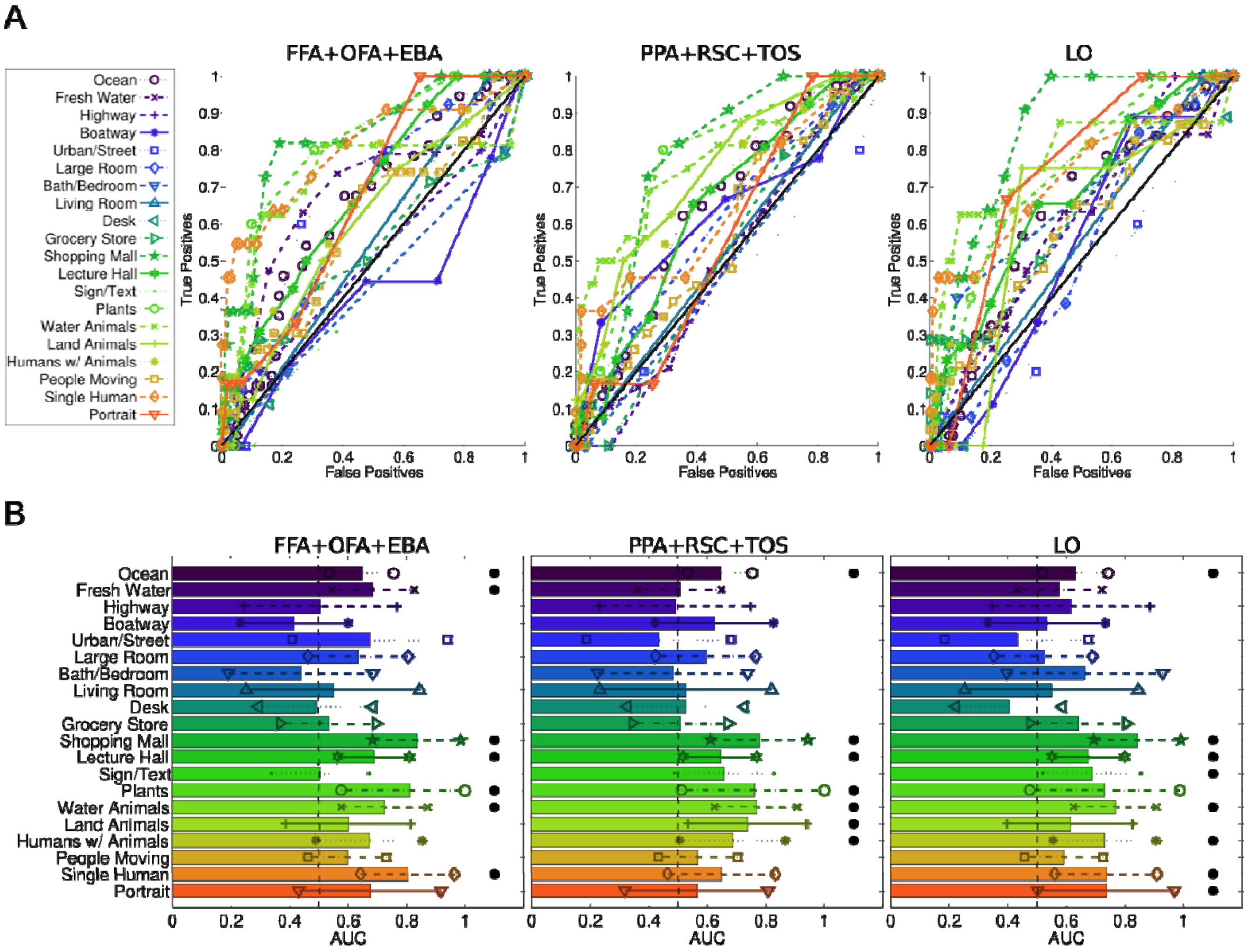


Figure S18. Decoding scene categories from various ROIs, subject S3

Figure organized analogously to Figure S16.The number of voxels in the first group of ROIs consisting of FFA, OFA, and EBA were *n* = 92. For the ROIs PPA, RSC, and TOS, *n* = 13. For area LO, *n* = 32.

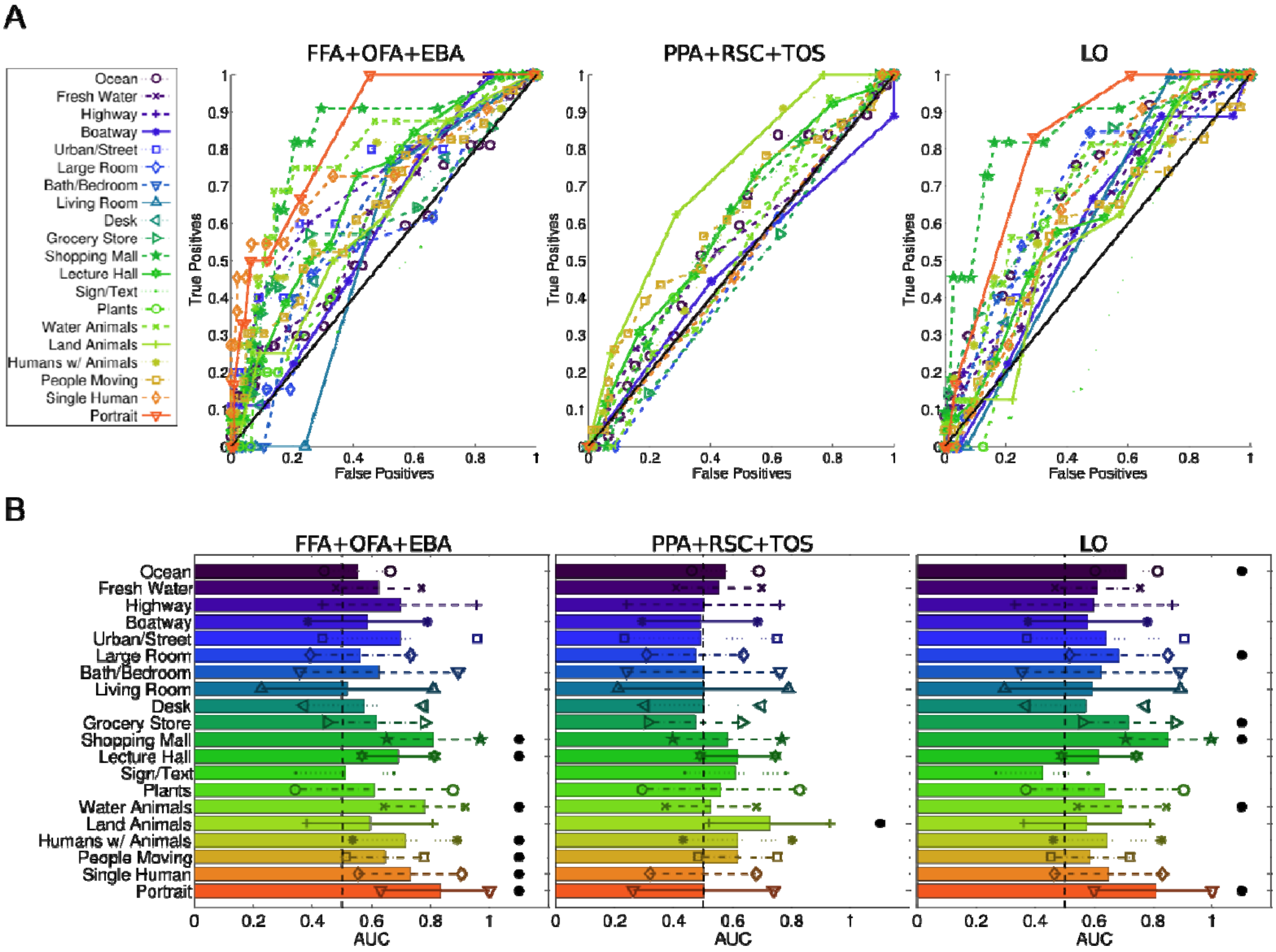


Figure S19. Decoding scene categories from various ROIs, subject S4

Figure organized analogously to Figure S16.The number of voxels in first group of ROIs consisting of FFA, OFA, and EBA were n = 33. For the ROIs PPA, RSC, and TOS, n = 16. For area LO, n = 17.

Supplemental Experimental Procedures

Supplemental Experimental Procedure 1: Compiling the natural and null learning databases

The learning database was composed of the objects that appeared in 4166 photographs of natural scenes. The scenes were compiled from commercial (Yao et al, 2007) and in-house archives. The objects in each scene were labeled by humans who were directed to outline and name all objects in the scene. This process produced 2010 distinct object labels across the entire learning database. To reduce label sparseness, many labels were aggregated into their superordinate category. This process was repeated until each object label had at least two exemplars in the learning database. The final database consisted of 950 distinct labels. Note that the 1260 images in the estimation set were selected from the learning database, but the 126 images in the validation set were drawn independently and were not present in the learning database.

A second null database of object labels was created in order to learn categories used in the null models. To create the null database, the object labels from the learning database were permuted across images while preserving the overall frequency of each distinct object and the number of distinct objects in each scene. This produced 4166 simulated scenes with co-occurrence statistics of the objects that were completely random.

Supplemental Experimental Procedure 2: Investigating the effect of scene context on object labels

When compiling the learning database there was concern that scene context may affect the labels assigned to objects, possibly biasing the statistics of the database. To control for such effects we selected a set of 106 representative images from the learning database and isolated each of the 829 distinct objects depicted by the images in this set. To isolate each object we applied a mask to the remainder of the image area using available object segmentation information. Two naive participants were then asked to re-label the isolated objects. The participants had no previous association with the experiment nor had observed any of the 106 sample scenes. The presentation of the objects was randomized across all 106 images to remove any context that could be inferred by serial presentation of objects from a single image.

Figure S2 displays the results of this control experiment. The bright diagonal line in each of the confusion matrices indicates a strong correspondence between the labels assigned to objects presented in context or in isolation. The correlation coefficient between the label indicator vectors assigned to the objects in context and in isolation was 0.78 and 0.76 for Participants A and B, respectively. Chance performance for this task was 0.09 ($p < 0.01$, see Supplemental Experimental Procedure 8).

Further analysis (data not shown) finds that 32% of the objects that were labeled incorrectly when presented in isolation were either located on the border of the image, or they occupied less than 10% of the image surface area. This indicates that object completion and resolution, not surprisingly, also play a substantial role in labeling. We are therefore convinced that the effects of scene context are not likely to significantly bias the statistics of the learning dataset.

Supplemental Experimental Procedure 3: Description of the Latent Dirichlet Allocation generative model

Under the generative Latent Dirichlet Allocation model, a scene is represented by a length V indicator vector $\boldsymbol{o}$ whose entries are the number of times that the v-th object from an available vocabulary (v=1...V) occurs in the scene. The probability of observing a scene under LDA is

$$p(\mathbf{o} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int \left(\prod_{j}^{N} \sum_{c_j}^{K} p(o_j \mid c_j; \boldsymbol{\beta}) p(c_j \mid \boldsymbol{\theta}) \right) p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) d\boldsymbol{\theta} \tag{1}$$

where

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \sim \text{Dirichlet}(\boldsymbol{\alpha}) \tag{2}$$

$$p(c \mid \boldsymbol{\theta}) \sim \text{Multinomial}(\boldsymbol{\theta}) \tag{3}$$

$$p(o \mid c = k; \boldsymbol{\beta}) \sim \text{Multinomia}\,l\left(\boldsymbol{\beta}^{(k)}\right) \tag{4}$$

The model assumes that the N objects in a scene are generated by the following process: A parameter vector $\boldsymbol{\theta}$ is drawn from a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$ (Equation 2). The parameter $\boldsymbol{\theta}$ defines a multinomial probability distribution $p(c = k| \boldsymbol{\theta})$ (Equation 3) over K discrete scene categories. Given the parameter $\boldsymbol{\theta}$, the j-th object in a scene (j=1....N) is associated with a category index $c_j = k$ that is drawn independently from $p(c = k| \boldsymbol{\theta})$. For each of the K possible values that scene category index can take, there is an associated parameter $\boldsymbol{\beta}^{(k)}$ that defines a multinomial probability distribution over the indices to an available object vocabulary. Given c_j and $\boldsymbol{\beta}^{(k)}$, the index to an object label is drawn from the conditional multinomial distribution in Equation 4.

The parameter $\boldsymbol{\beta}$ is a V x K matrix of probabilities, where V is the number of objects in the assumed vocabulary and K is the number of distinct scene categories assumed to exist in the database of scenes. Each entry of $\boldsymbol{\beta}$ is the average likelihood that the v-th object (row) occurs in (or more exactly, is drawn from) the k-th category (columns). Therefore $\boldsymbol{\beta}^{(k)}$ is the k-th column of $\boldsymbol{\beta}$. The parameter $\boldsymbol{\alpha}$ is a length K vector that characterizes the overall statistical distribution of the K scene categories observed across the entire database. Specifically, $\boldsymbol{\alpha}$ is a list of parameters for a K-dimensional Dirichlet distribution from which category mixtures for each scene are drawn. Learning scene categories under LDA is analogous to estimating the model parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for a given learning database (see Supplemental Experimental Procedure 4).

A toy example demonstrating the generative assumptions of LDA is shown for a known set of parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in Figure S1B. The toy model assumes that there are only two scene categories (k □ [1, 2]), each of which is associated with either the color red (category c = 1) or blue (category c = 2). In the first step (i) the parameter vector $\boldsymbol{\theta}$ is drawn from a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$ = [2, 5]. The parameter $\boldsymbol{\theta}$ defines the multinomial distribution $p(c \mid \boldsymbol{\theta})$ over the two choices of scene category. In particular, the $\boldsymbol{\theta}$ chosen in the example assigns the red category an average probability of being drawn of 0.9 and the blue category an average probability of 0.1. To generate each object a category index k is drawn according to $p(c \mid \boldsymbol{\theta})$ (ii). This index selects a conditional, category-specific multinomial distribution from which to draw an object index. In this particular example there are two conditional multinomial distributions parameterized by $\boldsymbol{\beta}^{(1)}$ and $\boldsymbol{\beta}^{(2)}$, respectively. After selecting either $\boldsymbol{\beta}^{(1)}$ or $\boldsymbol{\beta}^{(2)}$, the object index is drawn with probability $p(o=v|c=k; \boldsymbol{\beta}^{(k)})$ (iii). This process ((ii)-(iii)) is repeated N times to generate N objects (iv). In this example $N = 10$. Nine of the ten objects are sampled from the distribution parameterized by $\boldsymbol{\beta}^{(1)}$ and one is sampled from the distribution parameterized by $\boldsymbol{\beta}^{(2)}$. Note that this generative mechanism is similar to the method used in Supplemental Experimental Procedure 14 to decode object occurrence probabilities from the responses $\boldsymbol{r}$ to a scene. However, in Supplemental Procedure 14, for step (ii) $p(c \mid \boldsymbol{\theta})$ is replaced by the category probabilities predicted by the decoder.

Supplemental Experimental Procedure 4: Learning scene categories using Latent Dirichlet Allocation

Given a database D of M labeled natural scenes, where $\mathbf{D} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots \mathbf{o}_M\}$, the probability of the entire database is given by product of the probabilities of all scenes in database:

$$p(\mathbf{D} \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_i^M p(\mathbf{o}_i \mid \boldsymbol{\alpha},\boldsymbol{\beta}) = \prod_i^M \int \left(\prod_j^{N_i} \sum_{c_j}^K p(o_{ij} \mid c_{ij};\boldsymbol{\beta})\, p(c_{ij} \mid \boldsymbol{\theta}_i) \right) p(\boldsymbol{\theta}_i \mid \boldsymbol{\alpha})\, d\boldsymbol{\theta}_i \tag{5}$$

During category learning, the learning algorithm for LDA estimates the two model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by maximizing Equation (5) with respect to the parameters. Due to the intractability of the integral with respect $\boldsymbol{\theta}$ to in Equation 5, variational methods were employed to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A discussion of these techniques is beyond the scope of this paper; thus we refer the reader to the original manuscript (Blei et al, 2003) for details. Parameter estimation was done using the publicly available MATLAB® package (http://chasen.org/~daiti-m/dist/lda/; Daichi Mochihashi, NTT Communication Science Laboratories).

Supplemental Experimental Procedure 5: Addressing non-stationarity in the LDA learning algorithm

The variational learning algorithm used to estimate the LDA model parameters (**α** and **β**) involves gradient-based optimization from random initial conditions. Consequently, for a given setting of the assumed number categories and object vocabulary (see Supplemental Experimental Procedure 9), it is possible that the learning algorithm will learn different categories depending on different initial conditions. Thus, to determine a single, representative set of categories for each setting of the number of the scene categories and vocabulary size, the LDA algorithm was initialized 10 separate times to learn 10 independent sets of scene categories. Divisive hierarchical clustering (Hastie, Tibshirani, & Friedman, 2009) was then applied to the 10 learned sets of categories, using a distance metric of (1 - *d*), where *d* was the pairwise correlation between the categories learned from each initialization. We defined the representative set of scene categories for a given setting of the number of categories and object vocabulary to be the first set of categories in the larger of the two main clusters recovered by the clustering algorithm.

Supplemental Experimental Procedure 6: Calculating the category probabilities of a stimulus scene

The encoding models and the decoder assume that a scene s is represented as a set of probabilities over K scene categories, conditioned on the objects $\mathbf{o}(s)$ in the scene. We defined these probabilities to be the marginal posterior distribution $p(\boldsymbol{c}|\mathbf{o}(s))$ under the LDA generative model. Intuitively, $p(\boldsymbol{c}|\mathbf{o}(s))$ specifies the probabilities that a scene s, labeled with objects $\mathbf{o}(s)$ belongs to (or more precisely, is drawn from) each of the K categories learned using LDA. Calculating $p(\boldsymbol{c}|\mathbf{o}(s))$ directly is intractable, thus we approximated the posterior using the optimized (and normalized) Dirichlet variational parameters described in (Blei et al, 2003; Equation 4). These parameters were estimated using the optimization routines outlined in (Blei et al, 2003; Equations 5-7). The approximated posterior can be interpreted as the output of a nonlinear transfer function Φ that returns a probabilistic set of dimensions representing the category of a scene conditioned on the objects therein. Thus we refer to the posterior $p(\boldsymbol{c}|\mathbf{o}(s))$ instead as $\Phi(\mathbf{o}(s))$.

Supplemental Experimental Procedure 7: Encoding model parameter estimation

The responses **r** evoked in a set of *T* voxels by a series of *S* scenes **s** was modeled as a linear combination of the scene category probabilities $\Phi(\mathbf{o}(\boldsymbol{s}))$ for those scenes plus isotropic Gaussian noise **ε**.

$$\mathbf{r} = \Phi(\mathbf{o}(\mathbf{s}))\mathbf{W}_E + \boldsymbol{\varepsilon} \tag{6}$$

Here $\Phi(\mathbf{o}(\mathbf{s}))$ is of size $S \times K$, where *S* is the total number of scenes in the training set and *K* is the number of scene categories. The encoding models for all *T* voxels were trained simultaneously by estimating the $K \times T$ encoding weight matrix $\mathbf{W}_E$ via minimization of the regularized least squares objective functional:

$$O(\mathbf{W}_E) = \left\|\Phi(\mathbf{o}(\mathbf{s}))\mathbf{W}_E - \mathbf{r}(\mathbf{s})\right\|^2 + \left\|\boldsymbol{\Gamma}\mathbf{W}_E\right\|^2 \tag{7}$$

with respect to $\mathbf{W}_E$**.** Here the $\left\|\circ\right\|^2$ operator indicates the Euclidean norm, $\boldsymbol{\Gamma} = \lambda\mathbf{I}$ is a regularization matrix that controls the range of values that $\mathbf{W}_E$ can take, and **I** is the $K \times K$ identity matrix. Minimizing $O(\mathbf{W}_E)$ produces the following estimate $\hat{\mathbf{W}}_E$ of the weight matrix:

$$\hat{\mathbf{W}}_E = \left(\Phi(\mathbf{o}(s))^T \Phi(\mathbf{o}(\mathbf{s})) + \lambda^2\mathbf{I}\right)^{-1} \Phi(\mathbf{o}(s))^T \mathbf{r}(\mathbf{s}) \tag{8}$$

The value of the regularization parameter λ was varied on a log scale from zero to 2^{20} and an optimal value for λ was determined by retaining the value that minimized the mean squared error on a held out 10 percent ($n = 126$) of the responses to the estimation set scenes. Equation 8 and the determined value for λ were used to estimate the encoding model weights $\hat{\mathbf{W}}_{\boldsymbol{E}}$ via10-fold cross validation on 80 percent ($n = 1008$) of the responses to the estimation set stimuli.

Once trained, each encoding model was used to predict the voxel responses $\hat{\mathbf{r}}(\mathbf{s}_{novel})$ evoked by a series of novel scenes $\mathbf{s}_{novel}$ via Equation 6. The prediction accuracy for the encoding model of each voxel was defined to be the correlation coefficient (Pearsons's *r* score) between the observed responses to the validation set scenes $\mathbf{r}(\mathbf{s}_{novel})$ and the predicted responses $\hat{\mathbf{r}}(\mathbf{s}_{novel})$. Initial encoding model prediction accuracy on the estimation set was calculated by setting $\boldsymbol{s}_{novel}$ to a held out 10 percent ($n = 126$) of the estimation responses. These initial prediction accuracies were used to determine the single best set of categories for predicting brain activity across subjects (see Figure 2A in the main text). Once the single best set of categories was determined, models based on these categories were further tested by predicting the responses evoked by a separate novel set of 126 validation scenes. Again, correlation coefficient was used to quantify model accuracy on the validation set.

Supplemental Experimental Procedure 8: Definition of statistically significant prediction accuracy

When calculating correlation coefficients (denoted here as ρ to avoid confusion with responses, **r**) between two vectors of length n, the statistical significance threshold for the correlation $\rho_{threshold}$ at a p-value p < $\alpha_{threshold}$ was calculated from a standard t-test using the following relationship:

$$\rho_{threshold} = \frac{t}{\sqrt{\left(t^2 + n - 2\right)}} \qquad (9)$$

Here, t is the critical value of a one-tailed t-test having (n - 2) degrees of freedom at a significance level of $\alpha_{threshold}$. For calculating significant prediction accuracy for encoding models, n was set to 126, which was both the number of scenes in the held-out 10 percent of the estimation data and the number of scenes in the validation set.

Equation 9 was also used to define the decoding accuracy threshold for the decoder in Supplemental Experimental Procedure 13 and to compare with- and without-context object label indicator vectors in Supplemental Experimental Procedure 2. In the case of calculating significant decoding accuracy, n was set to 20, the number of scene category probabilities (based on the single best set of categories for all subjects) predicted by the decoder. In Supplemental Experimental Procedure 2, n was set to 829, the total length of the object label indicator vectors.

Supplemental Experimental Procedure 9: Identifying the best set of scene categories for modeling data across subjects

Before proceeding with voxel-wise modeling, we first determined the single set of scene categories that provided the best models of brain activity recorded from all subjects. To determine these scene categories, we first used the LDA algorithm to learn many different sets of candidate scene categories from the learning database. Each set of candidate scene categories was learned while systematically varying two parameters: the number of distinct scene categories assumed to exist in the database, which was varied from 2-40 in increments of two; and the number of objects in the vocabulary available to define each set of categories, which was varied from 25-950 in increments of 25. This procedure produced 760 different sets of candidate scene categories, one for each setting of the number of assumed categories and size of the available object vocabulary.

Voxel-wise encoding models were estimated for each subject, based on each of the 760 candidate scene categories, using the regression procedure described in Supplemental Experimental Procedure 7. For each set of voxel-wise models, the number of voxels whose activity was predicted with accuracy above a statistical significance threshold (0.21, $p < 0.01$; for details, see Supplemental Experimental Procedure 8) were then identified. This process produced a distribution of 760 values summarizing the amount of accurately-predicted cortical territory within each subject by encoding models based on each of the candidate scene categories.

The distributions of 760 values for each subject were then combined and the single best set of scene categories for modeling activity across subjects was identified as corresponding to the peak of the combined distribution. To combine the distributions we rescaled each individual distribution to sum to one, then computed the Hadamard product across distributions. The rescaling was done to control for differences in SNR and brain size across subjects.

Supplemental Experimental Procedure 10: Definition of statistically significant proportion of variance explained in an ROI

The average proportion of variance explained within each ROI was used as a metric to quantify the extent to which the voxel-wise encoding models characterized ROI activity. The proportion of variance explained for each voxel within a given ROI was defined as the coefficient of determination (David et al, 2005) calculated on the validation set data. The average proportion of variance explained and standard errors of these estimates were then calculated across voxels within each ROI.

A permutation test was used to determine statistically significant proportion of variance explained within each ROI. First, responses to the validation set stimuli were shuffled across all recorded voxels (individual voxel time courses were preserved). Then, using encoding models based on the globally optimal set of categories for each subject (Figure S3, red dots), the average variance explained in each ROI was calculated as described in the paragraph above. This process (shuffling followed by calculating average variance explained) was repeated 1000 times for each ROI. Statistically significant average variance explained for each ROI was defined to be upper 99th percentile threshold across the distribution of 1000 average variance explained values calculated for each ROI.

Supplemental Experimental Procedure 11: Investigating the optimal scene categories for each voxel

It is possible that the scene categories encoded by functional activity may vary in a meaningful way across subjects and across the cortical surface. To investigate this possibility we identified independently for each voxel the set of scene categories (out of 760 possible sets of scene categories learned; see Supplemental Experimental Procedure 9) that maximized encoding model prediction accuracy.

We then examined how these optimal scene categories varied across voxels within individual subjects. Panel A in Figures S8-S11 show the distribution of optimal scene categories for subjects S1-S4, respectively. For all subjects there is a large positive correlation between the optimal number of scene categories and object vocabulary size (as in Figure 2A in the main text.) This pattern is most likely a reflection of the fact that some brain areas are narrowly tuned to a small number of objects while others are more broadly tuned. While this is an interesting detail, there appears to be little advantage to optimizing scene categories to capture these local differences in the breadth of object tuning. The prediction accuracies of encoding models based on the optimal set of scene categories for each voxel are highly correlation with prediction accuracies of models based on the single best set of scene categories across all subjects and voxels (see Panel B in Figures S8-S11). This finding validates the use of the single best set of scene categories in all subsequent analyses.

Panel C of Figures S8-S11 show the optimal scene categories for each voxel plotted on the visual cortices of subjects S1-S4, respectively. Scene categories are represented homogeneously on a small spatial scale (on the order of a few voxels) but are represented heterogeneously on a larger spatial scale (on the order of of an ROI). Optimal scene categories for voxels within the ROIs FFA and OFA are generally made up of fewer distinct categories and are composed of smaller object vocabularies. In contrast, the optimal scene categories for voxels within the ROIs PPA, RSC, and TOS are generally made up of a larger number of scene categories and are composed of larger object vocabularies

.

Supplemental Experimental Procedure 12: Estimating encoding models based on alternative hypotheses

The results presented here show that voxels located in anterior visual cortex can be modeled accurately in terms of a scene categories estimated from object co-occurrence statistics. However, the question remains whether models based on other features might also provide accurate predictions of activity in these areas. To address this issue we evaluated encoding models based on four alternative hypotheses, as described below.

Models based on Latent Semantic Analysis

Encoding models based on LDA represent scenes in terms of a nonlinear transformation (Bayesian probabilistic inference) of the objects occurring in the scene. However, the brain might represent scenes in terms of a *linear* transformation of the objects in the scene. One linear transformation is provided by Latent Semantic Analysis (LSA; Deerwester et al., 1990). Under LSA, a scene is represented as a point in a latent feature space. The feature space is determined by performing singular value decomposition (SVD; Golub & Kahan, 1965) on a matrix of the object frequencies determined from a database of labeled scenes. Each feature is an eigenvector resulting from the SVD and the features can be ranked by importance according to their corresponding singular values (the squared singular values are proportional to the amount of variance in the frequency matrix accounted for by the corresponding eigenvector). A scene is then represented as a linear projection into this latent feature space, or a subspace of it. Comparing the prediction accuracies of encoding models based on LSA and LDA shows whether a linear or nonlinear transformation of the objects in a scene best describes the representation of scenes in anterior visual cortex.

To derive a feature representation for LSA, SVD was performed on the matrix of object frequencies determined from the experimental learning database. Before SVD, the entries of the matrix were normalized by the entropy of the frequency distribution for each object across the entire learning database. The LSA feature space was then defined as the top (384) eigenvectors produced by the SVD that explained 90% of the variance in the in object frequency matrix. The estimation and validation set stimulus scenes were then projected onto the axes of this latent space based on the objects labeled in each scene. An LSA encoding model was then estimated separately for each voxel by using the LSA feature space projections as features applied to the estimation data (see Supplemental Experimental Procedure 7).

Panel A of Figures S12-S15 compares predictions of the LSA and LDA models on the validation set. For all subjects, the LDA encoding models provide more accurate predictions in anterior visual cortex. These results indicate that the LDA feature representation best describes selectivity across these areas than features obtained using a linear transformation of the objects in a scene.

Models based on the presence of Diagnostic Objects

It is also possible that the scene categories learned by LDA mainly reflect the presence of the most probable object in the category such as a person or a building and that activity in anterior visual cortex is better explained by selectivity for these diagnostic objects. If so, then models based solely on the presence of diagnostic objects should provide equal or better predictions of voxel activity than models obtained using LDA. To investigate this possibility, a class of encoding models was derived based on the presence of only the most probable (*i.e.*, diagnostic) object in each of the best scene categories identified across subjects.

First, the most probable object was identified for each of the 20 best scene categories determined across subjects (Figure 2A, main text). Examples of these diagnostic objects were “human”, “building”, ”wall”, and ”sky.” Each scene in the estimation and validation set was then represented as a vector of 20 entries, one entry for each of the diagnostic objects. Each entry in a scene vector was the number of times a corresponding diagnostic object occurred in each scene. A diagnostic object (DO) encoding model was then estimated separately for each voxel by using the diagnostic object vectors as features applied to the estimation data (see Supplemental Experimental Procedure 7).

Panel B of Figures S12-S15 compares predictions of the DO and LDA models on the validation set. For three out of four subjects, the LDA encoding models provide predictions that are more accurate across anterior visual cortex. For subjects S1, S3, and S4, the LDA feature representation better models selectivity in these areas than a representation based on diagnostic objects. For subject S2, however, the DO model provides more accurate predictions of activity in anterior visual cortex. Note that the diagnostic objects were selected based on the results of a quantitative statistical method (*i.e.*, the LDA algorithm). Thus the DO models provide an objective methodological advance over conventional experiments that employ objects and scene categories selected subjectively. Furthermore, positive results obtained using the DO models also support the main hypothesis of the current work: natural scene statistics of objects are a valid basis of representation for scenes in anterior visual cortex.

Models based on the Gist descriptor

Previous work has shown that human perception of scene categories is directly associated with low-level, spatially global image features. These image features have been coined the spatial envelope or “Gist” of a scene (Oliva & Torralba, 2001). It is possible that the results reported in the current work are due simply to modeling low-level, global stimulus structure that is correlated with the scene categories learned by the LDA algorithm. To investigate this possibility, the prediction accuracies of the LDA encoding models were compared to prediction accuracies of a class of encoding models based on scene Gist.

To derive a representation based on Gist features, each of the stimulus scene images were first resized to 256 x 256 pixels. The Gist descriptor, a computational model of the spatial envelope, was then calculated for each scene in the estimation and validation sets, according to the procedures outlined in (Oliva & Torralba, 2001). MATLAB® functions available for public download http://people.csail.mit.edu/torralba/code/spatialenvelope/ were used to process the scenes. This process produced a vector of 512 Gist descriptors for each scene. A Gist encoding model was then estimated separately for each voxel by using the Gist descriptors as features applied to the estimation data (see Supplemental Experimental Procedure 7).

Panel C of Figures S12-S15 compares predictions of the Gist and LDA models on the validation set. For all subjects, the LDA models provide superior predictions in anterior visual cortex. These results indicate that selectivity in anterior visual cortex is better modeled using LDA features than features based on global image structure.

Models based on the Gabor Wavelet Pyramid

Another possibility is that our results are caused by local, low-level image features that are correlated with the scene categories learned by the LDA algorithm. For example, low spatial frequencies located in the upper portion of the visual field might be correlated with outdoor scenes. If this is true then voxel models based on local image structure should provide predictions in anterior visual cortex as good as those obtained using LDA models.

To capture local image structure, a class of encoding models was derived based on the Gabor wavelet pyramid (GWP) image decomposition (Daugman, 1985; Kay et al, 2008). The GWP decomposition captures local contrast energy at various orientations, spatial frequencies, scales, and locations across the visual field. Each of the scenes in the estimation and validation sets were filtered through the GWP, producing a vector of 1336 Gabor wavelet weights for each scene. A GWP encoding model was then estimated separately for each voxel by using the GWP weights as features applied to the estimation data (see Supplemental Experimental Procedure 7).

Panel D of Figures S12-S15 compares predictions of the GWP and LDA models on the validation set. For all subjects, the LDA-based encoding models provide superior prediction accuracies in anterior visual cortex. These results indicate that the LDA-based feature representation is a better model of selectivity in these areas than features based on localized image structure.

Supplemental Experimental Procedure 13: Decoder parameter estimation

In the decoding analysis the category probabilities of a scene *s* were modeled as a multinomial-distributed random variable. The likelihood for the *l*-th category $\Phi_l(\mathbf{r}(s))$ is

$$\Phi_l(\mathbf{r}(s); \mathbf{W}_D) = \frac{\mathrm{e}^{\mathbf{r}(s)^T \mathbf{W}_D^{(l)}}}{\sum_{k=1}^{K} \mathrm{e}^{\mathbf{r}(s)^T \mathbf{W}_D^{(k)}}} \tag{10}$$

where $\mathbf{W}_D{}^{(k)}$ is the *k*-th column of the decoder weight matrix $\mathbf{W}_D$ (with size $Q \times K$)*,* and $\mathbf{r}(s)$ is a vector (with length *Q*) of population responses measured from a select group of voxels and evoked by stimulus *s*. The population of voxels selected was those having statistically significant prediction accuracy for encoding models based on the single best set of scene categories determined across subjects (Supplemental Experimental Procedures 7-8). Selection bias was avoided by using cross-validated estimates of encoding model prediction accuracy on the estimation data set. The values of $\mathbf{W}_D$ were fit by maximizing the regularized log-likelihood function over the presentation of *S* individual scenes.

$$L_{reg} = \frac{1}{S}\sum_{i=1}^{S} \log \Phi_{l_i}(\mathbf{r}(s_i); \mathbf{W}_D) - \sum_{l=1}^{K} \mathrm{P}_\lambda(\mathbf{W}_D^{(l)}) \tag{11}$$

The second term on the right-hand side of Equation 11 is an "elastic-net" regularization penalty of the form

$$\mathrm{P}_\lambda(\mathbf{W}) = \sum_{x=1}^{X} \frac{1}{2}(1-\lambda)W_x^2 + |W_x| \tag{12}$$

Where **W** is a generic vector of weights of length *X* and |∘| is the absolute value. Varying the value of the regularization parameter *λ* from 0 to 1 controls the degree of sparseness in the values in $\mathbf{W}_D$ ($\lambda = 0$ results in L_2 regularization and $\lambda = 1$ results in L_1 regularization). The value of the regularization parameter *λ* was selected as to minimize the log likelihood of the error between the predicted and inferred (using LDA) category probabilities using a held-out 10 percent of the estimation data. The decoder weights $\hat{\mathbf{W}}_D$ were estimated using coordinate descent optimization methods described in (Friedman et. al., 2010) and implemented using the publicly available glmnet package (http://www-stat.stanford.edu/~tibs/glmnet-matlab/).

After estimating the values of $\hat{\mathbf{W}}_D$, Equation 10 was used (substituting $\hat{\mathbf{W}}_D$ for $\mathbf{W}_D$) to decode the category probabilities $\Phi(\mathbf{r}(s_{val}))$ for each of the validation set scenes $\mathbf{s}_{val}$ from evoked population responses $\mathbf{r}(s_{val})$. Decoding accuracy for each scene was defined to be the correlation coefficient (Pearson's *r*) calculated between the category probabilities predicted by the decoder and the category probabilities inferred using LDA based on the objects that were labeled in each scene (see Supplemental Experimental Procedure 6). Statistically significant decoding accuracy for each image was calculated using Supplemental Experimental Procedure 8, where *n* was replaced by the number of individual category probabilities that were decoded (i.e. 20). Statistical significance of decoding accuracy across all images was determined using a Wilcox rank-sum test comparing the distribution of decoding accuracies to a null distribution of equal size drawn from a Normal distribution with a mean located at the statistical significance threshold (correlation = 0.58, as determined by Supplemental Experimental Procedure 8), and with equal variance to the empirical distribution for each decoder's prediction accuracy. Because the two distributions have different shapes, the rank-sum test determined whether the decoder exhibited systematically greater-than-chance performance.

Supplemental Experimental Procedure 14: Decoding the probable objects in a scene

To predict the probable objects in a scene, we estimated a probability distribution over the object vocabulary associated with the best set of scene categories determined across subjects. This probability distribution was based on the category probabilities $\Phi(\mathbf{r}(s))$ predicted by the decoder (Supplemental Experimental Procedure 13). First, an index k, corresponding to one of the 20 best scene categories identified across subjects, was drawn from a multinomial distribution with mean parameters given by the category probabilities predicted by the decoder. Given k, an object was drawn from object vocabulary (the 850 most frequent objects in the learning database). This was done by drawing an object index from the multinomial distribution parameterized by the vector $\boldsymbol{\beta}^{(k)}$, the k-th column of $\boldsymbol{\beta}$. This process of drawing a scene category index followed by drawing an object index was repeated 10,000 times, resulting in a distribution of object index frequencies. Normalizing this distribution of frequencies to sum to one gave an occurrence probability distribution $\tilde{p}$ over the 850 objects in the available object vocabulary.

The accuracy of the distribution of occurrence probabilities estimated for a scene was assessed in terms how well it accounted for actual objects labeled in the scene. Specifically, the ratio of log likelihoods LR for the labeled objects under two model distributions $\bar{p}$ and $\tilde{p}$ was calculated:

$$LR = \frac{\sum_{p=1}^{P} \log \bar{p}\left(o_p\right)}{\sum_{p=1}^{P} \log \tilde{p}\left(o_p\right)} \tag{13}$$

Here the numerator is the log likelihood of the labeled objects under a naïve distribution $\bar{p}$ hat assumes all objects are equally probable. The denominator is the log-likelihood of the labeled objects under the distribution of occurrence probabilities calculated for the scene. The variable P is the number of objects that were actually labeled in the scene. Values of LR greater than one indicate that the distribution of occurrence probabilities calculated for the scene accounts for the objects in the scene better than picking the 850 objects in the available vocabulary at random. Statistical significance of object decoding accuracy across all images was determined using a Wilcox rank-sum test comparing the distribution of likelihood ratios to a null distribution of equal size drawn from a Normal distribution with a mean of 1 and with equal variance to the empirical distribution of likelihood ratios for each subject. Because the two distributions have different shapes, the rank-sum test determined whether the decoder exhibited systematically greater-than-chance performance.

Supplemental Experimental Procedure 15: Decoding analyses using individual ROIs

The decoding analyses presented in the main text demonstrate that brain activity measured throughout anterior visual cortex can be used to decode a diverse range of scene categories. The conventional view of tuning in anterior visual cortex would suggests that these diverse results are due to decoding from multiple ROIs, each of which is uniquely suited for decoding categories that exist within a specific domain of representation. For example, ROIs tuned for animate objects, such FFA, OFA, and EBA, might be uniquely suited for decoding scene categories that tend to contain faces or body parts. Likewise, ROIs tuned for outdoor scenes, such as PPA and RSC, might be uniquely suited for decoding scene categories that tend to contain buildings.

An alternative view is provided by recent work from our laboratory (Huth et al., 2012; Naselaris et al., 2012) suggesting that semantic tuning within conventional ROIs is more flexible than reported previously. If this is the case, then it should be possible to decode a wide range of scene categories from any given ROI. To investigate this possibility we decoded scene categories from groups of ROIs conventionally considered to have similar tuning properties. If ROIs are narrowly tuned for scene categories then only a few categories will be decoded accurately. However, if these ROIs are more broadly tuned then many categories will be decoded accurately.

The decoding analyses included those voxels for which the LDA encoding model provided significant predictions (cross-validated on the estimation set; see Supplemental Experimental Procedures 7-8). Three non-overlapping groups of voxels were selected from this pool of voxels, based on tuning properties established by separate functional localizer scans. The first group of voxels was located within the boundaries of FFA, OFA, and EBA. These ROIs are thought to be selective for faces (Kanwisher et al, 1997) or body parts (Downing et al, 2001). The second group consisted of voxels located within the boundaries of PPA, RSC, and TOS. These ROIs are thought to be selective for outdoor scenes and information relevant for navigation (Epstein and Kanwisher, 1998; Maguire, 2001; Nakamura et al, 2000). The third group contained those voxels located within the boundary of LO, which has been implicated in the representation of individual objects (Malach et al., 1995). It has also been suggested that LO represents scene categories based on the objects in a scene (MacEvoy & Epstein, 2011).

The parameters for three separate decoders were estimated from evoked activity measured within the three relevant groups of voxels. The parameters were estimated using the procedures outlined in Supplemental Experimental Procedure 12. Following parameter estimation, each decoder was used to predict the category probabilities for the 126 validation set scenes (based on the set of 20 best scene categories identified across subjects; Figure 2A, main text). The predicted category probabilities were used as a proxy for each decoder to detect the presence (or absence) of each of the 20 best scene categories identified across subjects. The presence or absence of each of the 20 scene categories was calculated while varying the classification threshold from 0.001 to 1. For each classification threshold setting, if the predicted category probability for a category was above threshold the category was predicted to be present, otherwise the category was predicted to be absent.

For each of the three decoders, true positive and false positive occurrences for each scene category and classification threshold setting were calculated based on the results of a ground-truth decoder that was derived from the LDA algorithm. The trained LDA model (based on the 20 best scene categories identified across subjects) was used to infer the category probabilities for each of the 126 validation set scenes, as described in Supplemental Experimental Procedure 6. If any of the category probabilities inferred for a validation set scene exceeded a value of 0.5, the corresponding category was defined as present. Otherwise, the corresponding category was defined as absent for that scene.

Given the results of the ground-truth decoder, true positive and false positive occurrences were determined for each scene category and classification threshold. True positives occurred when an ROI-specific decoder predicted that a category was present while it was also defined as present by the ground-truth decoder. False positives occurred when an ROI-specific decoder predicted that a category was present while the ground-truth decoder defined it as absent. These calculations were performed separately for each of the three ROI-specific decoders. Receiver Operator Characteristic (ROC) curves were calculated for each decoder and scene category by plotting true positives against the corresponding false positives for all classification thresholds (Figures S16-S19).

The area under the ROC curve (AUC) was used as a metric to quantify the accuracy of each decoder at detecting each of the 20 scene categories. AUCs were estimated using trapezoidal integration of each ROC curve. Standard errors of AUC estimates were calculated using the following relationship:

$$SE = \frac{\sqrt{A(1-A)+(N_N-1)(Q_1-A^2)+(N_P-1)(Q_2-A^2)}}{N_P N_N} \tag{14}$$

where A is the AUC estimate and

$$Q_1 = \frac{A}{2-A},\ Q_2 = 2\frac{A^2}{1+A}. \tag{15}$$

In Equation 10 N_P and N_N are the number of true positives and true negatives, respectively (Hanley & McNeil, 1982). Decoders with AUC values (minus one standard error) exceeding a value of 0.5 were defined as predicting the presence of the corresponding scene category above chance level (Figures S16-S19).

Supplemental References

Blei, D. M. Ng, A. Y. Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3(4-5) 993-1022.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169.

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network (Bristol, England)*, *16*(2-3), 239-260.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.

Downing, P. E., Jiang, Y., Shuman, M., and Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470-2473.

Epstein, R. A. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598-601.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software*, *33*(1), 1-22.

Golub, G., & Kahan, W. (1965). Calculating the Singular Values and Pseudo-Inverse of a Matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, *2*(2), 205.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology 143*, 29–36.

Hastie, T. Tibshirani, R., Friedman, J., (2009). Hierarchical Clustering.In *The Elements of Statistical Learning,* 2nd ed. (New York: Springer) pp. 520–528.

Huth A. G., Nishimoto, S., Vu, A. T., Gallant J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210-1224.

Kanwisher, N, McDermott, J., and Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(11), 4302-4311.

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

MacEvoy S. P and Epstein, R. A (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience* 14,1323–1329

Maguire, E. A. (2001). The retrosplenial contribution to human navigation: a review of lesion and neuroimaging findings. *Scandinavian Journal of Psychology*, *42*(3), 225-238.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci U S A *92*, 8135–8139.

Nakamura, K., Kawashima, R., Sato, N., Nakamura, A., Sugiura, M., Kato, T., Hatano, K., et al. (2000). Functional delineation of the human occipito-temporal areas related to face and scene processing. *Brain*, *123*(9), 1903-1912.

Naselaris, T., Stansbury, D.E., and Gallant, J.L. (2012). Cortical representation of animate and inanimate objects in complex natural scenes. *Journal of Physiology-Paris.* 106(5-6) 239-249.

Yao, B., Yang, X., and Zhu, S.-C. (2007). Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. In *Proceedings of the 6th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 169–183.

by which the brain orchestrates region-specific dopamine signaling. Just as importantly, the finding that dopamine neuron responses track cognitive function could prove to be valuable for our understanding of Parkinson's disease, in which dopaminergic medications used for the control of motor symptoms are sometimes accompanied by cognitive side effects. Further work delineating the separate cognitive, motor, and learning signals in the SNc and VTA might eventually lead to better treatments that preferentially target dopamine's role in movement while sparing patients' cognitive abilities. Yet much remains to be done. For a long while yet, it appears, the tiny dopaminergic midbrain will continue to demand a large body of work.

REFERENCES

Berridge, K.C., and Robinson, T.E. (1998). Brain Res. Brain Res. Rev. *28*, 309–369.

Bromberg-Martin, E.S., Matsumoto, M., and Hikosaka, O. (2010). Neuron *68*, 815–834.

Fiorillo, C.D. (2013). Science *341*, 546–549.

Haber, S.N., and Knutson, B. (2010). Neuropsychopharmacology *35*, 4–26.

Horvitz, J.C. (2000). Neuroscience *96*, 651–656.

Li, B.-M., and Mei, Z.-T. (1994). Behav. Neural Biol. *62*, 134–139.

Matsumoto, M., and Hikosaka, O. (2009). Nature *459*, 837–841.

Matsumoto, M., and Takada, M. (2013). Neuron *79*, this issue, 1011–1024.

Noudoost, B., and Moore, T. (2011). Nature *474*, 372–375.

Redgrave, P., and Gurney, K. (2006). Nat. Rev. Neurosci. *7*, 967–975.

Sawaguchi, T., and Goldman-Rakic, P.S. (1991). Science *251*, 947–950.

Sawaguchi, T., and Goldman-Rakic, P.S. (1994). J. Neurophysiol. *71*, 515–528.

Schultz, W., Dayan, P., and Montague, P.R. (1997). Science *275*, 1593–1599.

Watanabe, M., Kodama, T., and Hikosaka, K. (1997). J. Neurophysiol. *78*, 2795–2798.

Williams, G.V., and Goldman-Rakic, P.S. (1995). Nature *376*, 572–575.

The Cerebral Emporium of Benevolent Knowledge

Patrick J. Mineault[1] and Christopher C. Pack[1,*]
[1]Montreal Neurological Institute, McGill University, Montreal, QC H3A 2B4, Canada
*Correspondence: christopher.pack@mcgill.ca
http://dx.doi.org/10.1016/j.neuron.2013.08.012

Visual objects tend to be found in predictable combinations (e.g., pens with paper). How does the brain represent these regularities? In this issue of *Neuron*, Stansbury et al. (2013) use fMRI to study the brain's representation of visual scene categories.

In a 1942 essay, Jorge Luis Borges discusses the categorization of animals, purportedly found in a fictitious Chinese encyclopedia named the "Celestial Empire of Benevolent Knowledge" (Borges, 1942). Animals therein are classified into 14 fanciful categories, including, "fabulous ones," "those that have just broken the flower vase," and "those that look like flies when viewed from a distance." Borges uses this example to suggest that any attempt to categorize the contents of nature is "arbitrary and full of conjectures."

Nevertheless (again quoting Borges), "the impossibility of penetrating the divine scheme of the universe cannot dissuade us from outlining human schemes, even though we are aware that they are provisional." In fact, such schemes can be quite useful in sensory neuroscience. A decade after Borges's essay, Barlow (1953) discovered neurons that respond selectively to stimuli that look like flies when viewed from a distance. These "fly detectors" were found in the retinas of frogs and, hence, were linked to a specific category of behavior (feeding). Subsequently, Hubel and Wiesel (1962) identified visual cortical cells that were described as "simple" and "complex," and these turned out to be useful labels for understanding many aspects of the visual cortex from anatomy to computation.

More recent imaging studies have led to the suggestion that neurons with particular stimulus selectivities are clustered together, forming brain modules responsible for encoding rather abstract categories of stimuli, including faces (Tsao et al., 2006), places (Epstein and Kanwisher, 1998), and buildings (Hasson et al., 2003). Of course, the number of such categories must be far greater than the number of brain regions, which leads to the profound question of how the brain organizes such a vast quantity of visual experience. In this issue of *Neuron*, Stansbury et al. (2013) address this question.

Stansbury et al. (2013) used fMRI imaging of human subjects to study the brain's representation of visual scene categories, defined as classes of images that contain similar co-occurrences of individual objects. For example, a scene that contains a building and a car is more likely to belong to the category "cityscape" than to the category "nautical." Obviously, one object (e.g., a tree) can be found in more than one scene (e.g., cityscape and rural), and one scene (e.g., a harbor) can belong to more than one scene category (e.g., cityscape and nautical). Thus, part of the challenge of understanding the brain's representation of scene categories is in understanding the organization of the categories themselves.

To this end, Stansbury et al. (2013) have adopted an elegant approach that defines the scene categories objectively with an algorithm that detects the presence of certain combinations of objects in a large database of natural scenes. Importantly, the algorithm is not given any prior information about which categories each scene belongs to; it defines categories on the basis of statistical regularities. This approach largely circumvents Borges's problem of the arbitrariness of categories, given that the classification is defined by the images themselves rather than being imposed by the person doing the analysis.

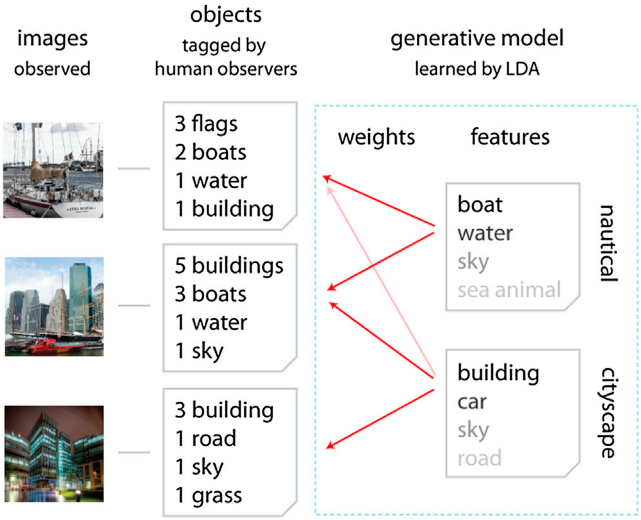


Figure 1. Estimating Categories from Natural Images
Human observers derive lists of objects from natural images (left). A generative model (right) specifies that these lists of objects are generated by weighted mixtures of features, which, in this case, are categories. The parameters of the model—the word probabilities corresponding to each category as well as the category vector corresponding to each image—are learned by the latent Dirichlet allocation algorithm.

In this approach, each scene (Figure 1, left) was tagged with a list of objects (e.g., two boats, one car, one person, etc.; Figure 1, middle) identified by human observers. These descriptors were fed to an unsupervised learning algorithm known as latent Dirichlet allocation (LDA), which inferred the categories represented in the data set on the basis of the pattern of co-occurrences of objects (Blei et al., 2003).

LDA, which has its root in text classification, is one of a number of unsupervised learning techniques that aim to uncover structure in complex data. Typically, they define each example in the data set—e.g., a list of words, an image, or a sound—as being generated by a noisy, weighted mixture of features. Optionally, they define a set of soft constraints, or priors, on the distribution of features and weights. The goal of the learning algorithm is to find a set of features and weights that captures the bulk of the variation in the data set while respecting the prior assumptions of the algorithm.

In LDA, each scene descriptor is assumed to be generated by a mixture of categories—the features (Figure 1, right). LDA assumes that the weights associated with this mixture (Figure 1, red arrows) are sparse—each scene contains only a handful of categories. It also assumes that weights are positive—whereas a scene may belong to a category (positive weight; indicated by a red arrow in Figure 1) or not (zero weight). It is not meaningful to say that a scene belongs negatively to a category (negative weight). The ensemble of weights linking a scene to each scene category is called the scene's category vector.

This sparse, positive encoding scheme allows the algorithm to leverage parts-based or combinatorial coding (e.g., both nautical and cityscape) in order to describe more narrowly defined scenes (e.g., harbor; Figure 1, middle). Each category is itself a sparse, positive mixture of objects (Figure 1, right).

These assumptions are embedded within a hierarchical, probabilistic model; objects contained within each category and the categories contained within each scene are jointly estimated by Bayesian inference. The resulting categories contained a high proportion of related objects. For example, one category assigned the highest weights for highway, car, sky, vehicle, and signpost—most likely corresponding to highways or ground transportation. Furthermore, the model assigned intuitive categories to the scenes in the database, tagging a harbor scene with nautical and cityscape categories. This is not surprising, given that LDA and its extensions have proven widely applicable in an analogous problem, determining categories from text documents (Blei et al., 2003).

The LDA approach taken by Stansbury et al. (2013) has revealed hidden structure in natural images, but does the visual system exploit this structure in its representation of visual scenes? One way to answer this question is to ask whether some aspect of brain activity correlates systematically with scene categories during the viewing of natural images. This would suggest that the brain encodes the scene categories in the same way that previous work has suggested an encoding of faces or orientations.

To tackle this question, Stansbury et al. (2013) had subjects view a variety of different scenes and simultaneously recorded their brain activity with fMRI. Then, the authors attempted to predict the BOLD response in each voxel under the assumption that the response to a scene was given by a weighted sum of the scene's category vector.

Responses in low-level striate and extrastriate visual areas, which are sensitive to elementary features such as orientation and contrast, were poorly modulated by scene category. However, responses in anterior visual areas such as the fusiform face area (FFA) and the parahippocampal place area (PPA) could be accurately predicted by the encoding model. The authors found that the predictions were most accurate when the LDA model contained 20 categories and 850 objects, indicating that there is substantially more categorical information available at the macroscopic fMRI scale than previously appreciated.

Importantly, the number of voxels significantly predicted by the category-encoding model was larger than alternative models relying on elementary visual features, such as orientation or spatial frequency. This was a crucial test of the hypothesis that high-level visual areas actually represent scene categories rather than visual stimuli per se (Malach et al., 1995). Consistent with this idea, the model was also significantly more accurate than others that relied only on the presence of individual objects.

Category preferences in different areas were, to some degree, consistent with previous literature. For example, the FFA showed a relative preference for the portraits category, whereas the PPA was most selective for categories that could be labeled "places." However, the results of this analysis indicated a more complex relationship between brain regions and category selectivity: voxels in several anterior visual areas showed selectivity for other categories. For example, the FFA was selective for the "plants" category in addition to "portraits." These results are consistent with earlier results from the same group, which highlighted the presence of a distributed representation of categories with smooth, overlapping gradients of preferred categories along certain cortical directions (Huth et al., 2012).

A second way to test the idea that scene categories are represented in specific brain regions is to ask whether it is possible to decode the category viewed by the observer on the basis of the BOLD activity alone. This approach is similar to that used by the same group to demonstrate how the brain represents specific images and objects (Naselaris et al., 2011). The authors found that BOLD activity successfully predicted the category membership of individual images. Importantly, these images were of novel scenes that were not used to formulate the encoding model, indicating that the model generalized beyond the specific exemplars on which it was trained.

Then, they used the LDA model to successfully predict the objects present in individual images on the basis of predicted category membership alone. This is quite a remarkable result given that objects are only encoded in the model indirectly through their correlation with scene categories. The success of this decoding approach implies that the distribution of objects in natural scenes contains substantial structure and that this structure can be exploited by the visual system.

These results might help to explain previous psychophysical findings that indicate that, when the gist of a scene is understood, objects within it can be recognized accurately even at extremely low resolutions, in some cases as low as ~6 × 6 pixels (Torralba, 2009). Performance in these tasks becomes worse when objects are isolated from their context. Similarly, human observers can detect an object more efficiently when it is found within a contextually consistent scene than when it is not (Biederman et al., 1973). Evidently, the problem of inferring object identity from low-level visual features is made much easier by context. Much like low-level how vision can make use of prior information to accurately estimate motion direction from noisy observations (Weiss et al., 2002), high-level vision could make use of learned statistical regularities to estimate object identity in ambiguous scenes (Lee and Mumford, 2003).

More generally, the approach developed by Stansbury et al. (2013) may provide an objective way to probe the brain's representation of abstract sensory information. Scene categories are abstract, in that they are largely independent of specific image features, but could they even be independent of vision? Would the sounds of traffic and the smell of baked goods produce the same activation as pictures of a city street? Perhaps sensory stimulation is not necessary at all: could imagining a specific type of scene produce interpretable activation in the relevant brain regions? Such representations might ultimately facilitate the extraction of even more abstract, perhaps semantic, information from brain activity.

REFERENCES

Barlow, H.B. (1953). J. Physiol. *119*, 69–88.

Biederman, I., Glass, A.L., and Stacy, E.W., Jr. (1973). J. Exp. Psychol. *97*, 22–27.

Blei, D.M., Ng, A.Y., and Jordan, M.I. (2003). J. Mach. Learn. Res. *3*, 993–1022.

Borges, J.L. (1942). The analytical language of John Wilkins. In Other Inquisitions (1937-1952), J.F. Solem, B.A. Davidsen, and R. Anderson, eds. (Austin, TX: University of Texas Press), pp. 101–105.

Epstein, R., and Kanwisher, N. (1998). Nature *392*, 598–601.

Hasson, U., Harel, M., Levy, I., and Malach, R. (2003). Neuron *37*, 1027–1041.

Hubel, D.H., and Wiesel, T.N. (1962). J. Physiol. *160*, 106–154.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). Neuron *76*, 1210–1224.

Lee, T.S., and Mumford, D. (2003). J. Opt. Soc. Am. A Opt. Image Sci. Vis. *20*, 1434–1448.

Malach, R., Reppas, J.B., Benson, R.R., Kwong, K.K., Jiang, H., Kennedy, W.A., Ledden, P.J., Brady, T.J., Rosen, B.R., and Tootell, R.B. (1995). Proc. Natl. Acad. Sci. USA *92*, 8135–8139.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Neuroimage *56*, 400–410.

Stansbury, D.E., Naselaris, T., and Gallant, J.L. (2013). Neuron *79*, this issue, 1025–1034.

Torralba, A. (2009). Vis. Neurosci. *26*, 123–131.

Tsao, D.Y., Freiwald, W.A., Tootell, R.B., and Livingstone, M.S. (2006). Science *311*, 670–674.

Weiss, Y., Simoncelli, E.P., and Adelson, E.H. (2002). Nat. Neurosci. *5*, 598–604.