Published in final edited form as:
J Physiol Paris. 2012 September ; 106(5-6): 239–249. doi:10.1016/j.jphysparis.2012.02.001.

Cortical representation of animate and inanimate objects in complex natural scenes

Thomas Naselaris[1], **Dustin E. Stansbury**[2], and **Jack L. Gallant**[1,2,3]

Thomas Naselaris: tnaselar@berkeley.edu; Dustin E. Stansbury: stan_s_bury@berkeley.edu

[1]Helen Wills Neuroscience Institute, University of California, Berkeley, CA 94720, USA

[2]Vision Science Program, University of California, Berkeley, CA 94720, USA

[3]Department of Psychology, University of California, Berkeley, CA 94720, USA

Abstract

The representations of animate and inanimate objects appear to be anatomically and functionally dissociated in the primate brain. How much of the variation in object-category tuning across cortical locations can be explained in terms of the animate/inanimate distinction? How is the distinction between animate and inanimate reflected in the arrangement of object representations along the cortical surface? To investigate these issues we recorded BOLD activity in visual cortex while subjects viewed streams of natural scenes. We then constructed an explicit model of object-category tuning for each voxel along the cortical surface. We verified that these models accurately predict responses to novel scenes for voxels located in anterior visual areas, and that they can be used to accurately decode multiple objects simultaneously from novel scenes. Finally, we used principal components analysis to characterize the variation in object-category tuning across voxels. Remarkably, we find that the first principal component reflects the distinction between animate and inanimate objects. This dimension accounts for between 50 and 60 percent of the total variation in object-category tuning across voxels in anterior visual areas. The importance of the animate-inanimate distinction is further reflected in the arrangement of voxels on the cortical surface: voxels that prefer animate objects tend to be located anterior to retinotopic visual areas and are flanked by voxels that prefer inanimate objects. Our explicit model of object-category tuning thus explains the anatomical and functional dissociation of animate and inanimate objects.

Keywords

fMRI; decoding; encoding; object representation

1. Introduction

There is considerable evidence that the representations of animate (i.e., humans and animals) and inanimate (i.e., everything else) objects are dissociated in the human brain. Anatomical

Correspondence should be addressed to: Jack L. Gallant, 3210 Tolman Hall #1650, University of California at Berkeley, Berkeley, CA 94720, gallant@berkeley.edu.

Conflict of Interest:
The authors declare no conflict of interest related to this work.

dissociation is supported by evidence that damage to anterior visual areas can selectively impair processing of either animate or inanimate objects, while having no effect on processing of objects from the other category (Caramazza and Shelton, 1998; Hillis and Caramazza, 1991; Warrington and Shallice, 1984). Functional dissociation is supported by a recent study of Kiani et al. (2007) who examined the responses of ~600 neurons in monkey inferior temporal (IT) cortex to images of ~1,000 objects. They reported evidence for a hierarchical clustering of population responses according to animate and inanimate object categories. A subsequent fMRI study of human IT found voxel population vectors that clustered according to animate and inanimate object categories as well (Kriegeskorte et al., 2009).

These results raise several interesting questions about the organization of object representations in visual cortex. First, how much of the variation in object-category tuning across cortical locations can be explained in terms of the animate/inanimate distinction? The results discussed above suggest that the animate/inanimate distinction is important relative to other categorical distinctions, but they do not specify how much of the variation in object-category tuning it accounts for. Second, how is the distinction between animate and inanimate objects reflected in object-category tuning of individual cortical locations? An analysis of the behavioral deficits in patients with brain damage or post hoc inspection of the object categories underlying voxel response clusters (Kiani et al. 2007; Kriegeskorte et al., 2009) can provide only partial answers. The most straightforward way to answer this question is to construct explicit and accurate models of object-category tuning for individual cortical locations. Third, how is the representation of animate and inanimate object categories arranged on the surface of the visual cortex? The results discussed above suggest that cortical locations strongly excited (or suppressed) by objects from either category are near one another, but they do not provide a map that reveals how these locations are arranged along the cortical surface. Finally, does the dissociation between animate and inanimate object categories persist when subjects view multiple objects embedded in a complex natural scene? The visual system evolved to process complex natural scenes with multiple objects, but most studies have used decontextualized objects presented in isolation. It is therefore important to confirm that the animate/inanimate distinction is relevant in naturalistic viewing contexts.

We addressed each of these questions by analyzing data acquired in a single fMRI experiment. BOLD activity was recorded in visual cortex while subjects viewed a large series of complex natural scenes. For each recorded voxel that intersected with the cortical surface, we fit a predictive encoding model (Naselaris et al., 2009) that related the objects in the viewed scenes to evoked responses. The model for each voxel consisted of a set of weights that describe how different object categories affect responses. We refer to the weights estimated for each voxel as an *object-category tuning function*. We confirmed the accuracy of the object-category model fit to each voxel in two ways. First, we used it to to predict the activity elicited by a separate set of images (the validation set) that were not used to fit the model. Second, we used it to decode multiple object categories simultaneously from the activity of those voxels whose responses were predicted accurately. To describe how object-category tuning varies across the visual cortex we applied principal component analysis to the object-category tuning functions of these voxels. We found that the first principal component (PC) described variation in the preference of voxels for animate versus inanimate object categories, and that this PC accounted for 50–60 percent of the total variation in object-category tuning across voxels. We then constructed a cortical surface map of the projection of each voxel's object-category tuning function onto the first PC. This map revealed that voxels with projections onto the animate end of the PC are located predominantly anterior to retinotopic areas, and are flanked by voxels with projections onto the inanimate end of the PC. These findings provide an explanation for the functional and

anatomical dissociation between animate and inanimate objects in terms of an explicit model of object-category tuning.

2. Methods

2.1 MRI parameters

All MRI data were collected at the Brain Imaging Center at UC Berkeley, using a Siemens Tim Trio 3T whole-body MR scanner and a 32 channel phased-array coil. The slice prescriptions varied slightly between subjects. The 32 channel Siemens head coil was used to record data from subject 1. A gradient-echo echo planar imaging sequence, combined with a custom water-specific excitation (fat-shunting) RF pulse, was used for functional data collection. Thirty-one axial slices covered the entire brain. 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; 2004.5 ms TR; 33 ms TE; 74° flip angle; voxel size of 2.24 × 2.24 × 4.13 mm^3). The back half of the Siemens 32 channel head coil was used for subject 2, so for this subject it functioned as a 20 channel surface coil. A gradient-echo echo planar imaging sequence combined with a custom fat-saturating RF pulse was used for functional data collection. Twenty-five axial slices covered occipital, occipito-parietal, and occipito-temporal cortex with greatest receiver coil sensitivity. Each slice had a 234 × 234 mm^2 field of view, 2.59 mm slice thickness, and 0.39 mm slice gap (matrix size 104 × 104;TR = 2009.9 ms; TE = 35 ms;flip angle = 74°; voxel size = 2.25 × 2.25 × 2.99 mm^3).

2.2 Stimuli

Stimuli were 1386 color natural scenes. Some scenes were taken from a collection available for sale from the Lotus Hill Institute (Wuhan, China) and others were selected from Google Images. All scenes were 20° × 20° (500 × 500 pixels). A central fixation square (0.2° × 0.2°; 5×5 pixels) switched color randomly (red, green, blue, yellow, white) at 3 Hz to ensure continued visibility.

2.3 Experimental design

Data for model training and model validation were collected during each scan session. Scan sessions consisted of separate training and validation runs. Training runs (a total of 36 across 5 sessions) lasted 5.23 minutes, and consisted of 36 distinct images presented 2 times each. Validation runs (a total of 21 across 5 scan sessions) lasted 5.23 minutes, and consisted of 6 distinct images presented 12 times each. A total of 1260 distinct scenes were presented across the training runs and a total of 126 distinct scenes were presented across the validation runs. This experimental design is based upon previous research from our laboratory showing that in order to construct optimal encoding models, the estimation data should sample the stimulus space as widely as possible, but the validation data should contain many repeats (David et al. 2005). Scenes were randomly selected for each run but were not repeated across runs. Scenes were presented in successive 4 second periods. In each period, a photo was flashed at 200 millisecond intervals (200 ON, 200 OFF) for 1 second, followed by 3 seconds of gray background. The fixation square was always present.

2.4 Data preprocessing

Functional brain volumes were reconstructed and then coregistered across scan sessions. The time-series data were used to estimate the response timecourse for each voxel separately. Deconvolving this timecourse from the data produced an estimate of the amplitude of the response (a single value) to each scene for each voxel (see Kay et al., 2008 for details). Early and intermediate visual areas (V1–V4, V3a/b and LO) were identified using retinotopic mapping data collected in separate scan sessions (Hansen et al., 2007). The

fusiform face are (FFA; Kanwisher et al., 1997), parahippocampal place area (PPA; Epstein and Kanwisher, 1998), retrosplenial cortex (RSC), extra-striate body area (EBA; Downing et al., 2001), and occipital face area (OFA) were identified in a separate experiment using standard functional localizers (Spiridon et al., 2006).

2.5 Scene labels

Prior to the experiment each of the natural scene stimuli were labeled by the authors. Labels consisted of specific object names chosen to be consistent with conversational English (see Figure 1 for examples). About half of the scenes used in the experiment were labeled originally by the Lotus Hill Institute, and were checked and corrected by the authors as necessary. The remaining scenes were labeled by the authors using a custom-made GUI. The authors then used the labels to assign each object in every scene to one of 19 object categories. The categories were selected based on previous experiments (Naselaris et al., 2009) and on intuition. In particular, we distinguished between “crowd of humans” and “several humans” because in scenes that depict crowds it is difficult to resolve specific features of individual faces. The selection of categories ensured that all objects could be assigned unambiguoulsy to one and only one category. Note that the labels were not present during visual presentation of the stimuli.

2.6 The object-category model

The object-category model is the basis for all the analyses presented in this paper. To build the object-category model, the 19 object categories discussed above were encoded using indicator variables. For each natural scene, an *object indicator vector* $\mathbf{o}^{\mathrm{T}} = (o_1, \ldots, o_{19})$, $o_i \in [0, 1]$ specifies which of the nineteen object categories listed in Figure 1 are present (o_i=1) or absent (o_i=0) in the scene. The object-category model consists of a set of weights $\mathbf{w}^T = (w_1, \ldots, w_{19})$ applied to each of these object indicators. A unique set of weights is fit (section 2.8) for each voxel and is referred to in the main text as the *object-category tuning function* for the voxel. Once the weights are fit, the predicted response, r, of a voxel to a natural scene with objects specified by $\mathbf{o}$ is given by $r = \mathbf{w}^T\mathbf{o}$.

2.7 Gabor wavelet model

As a control, we compared the prediction accuracy of the object-category model to the prediction accuracy of a purely structural model based on Gabor wavelets. The Gabor wavelet model consists of a bank of 928 complex Gabor wavelets. The Gabor wavelets occur at four spatial frequencies: 2, 3.6, 6.6, and 12 cycles per field of view (FOV = 20°; images were presented at a resolution of 500×500 pixels, but were downsampled to 32×32 pixels for this analysis), and two orientations: 0° and 90° (we used only two orientations because in preliminary modeling we found that orientation has a very weak effect on model prediction accuracy). Each wavelet is multiplied with a Gaussian envelope where the ratio between standard deviation and the wavelet spatial period is 0.55. Wavelets are positioned on a fixed square grid that covers the field of view. Grid spacing is established such that wavelets at each spatial frequency are separated by 2 standard deviations of their respective Gaussian envelopes. As with the the object-category model, the Gabor wavelet model includes a set of weights, $\mathbf{w}$, that are applied to the outputs of the Gabor wavelets. A separate set of weights is fit to each voxel. Once the weights are fit, the predicted response, r, of a voxel to a natural scene $\mathbf{s}$ is $r = \mathbf{w}^{\mathrm{T}}\boldsymbol{f}(\mathbf{s})$. Here, $\boldsymbol{f}(\mathbf{s})$ describes Gabor wavelet filtering and an additional nonlinear transformation: $\boldsymbol{f}(\mathbf{s}) = log(|G\mathbf{s}|)$, where G is a matrix with rows that contain the complex Gabor wavelets. Phase-invariance is achieved by taking the magnitude |.| of the filtered scenes. The log transform was applied because in other studies we have found that compressive output nonlinearities can improve model accuracy

(Nishimoto et al., 2011). Although the log transform is not completely optimal (Vu et al., 2011) it was sufficient for the purposes of this study.

2.8 Fitting procedure for the object-category and Gabor wavelet models

For each model, coordinate descent with early stopping (Naselaris et al., 2009) was used to find the set of weights **w** that minimized the sum-of-square-error between the actual and predicted responses on the training set. We have found that this regression procedure is far more robust than conventional linear regression, even when the observations outnumber the model parameters (Naselaris et al., 2009). For each voxel, this minimization was performed on three sets of M-l training samples (M=1260, $l = M$*0.1), selected randomly without replacement. Weight estimates were updated until the error on the held-out samples (the *early-stopping set*) began to increase consistently. Each set produced a separate estimate $\mathbf{w}_j$ (j=[1, 2, 3]). **w** was set equal to the arithmetic mean of ($\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$). This fitting procedure was implemented using STRFlab (Naselaris et al., 2009) a free MATLAB (Mathworks. Natick, MA) toolbox.

2.9 The object-category decoding algorithm

We also used the object-category model to decode multiple object categories from the voxel responses. Let $\boldsymbol{r}$ denote the collected responses of N separate voxels in an Nx1 voxel response vector. For each scene, we find the object indicator vector **o*** that maximizes the multi-voxel likelihood (Naselaris et al, 2009) corresponding to the object-category encoding model. Assuming that voxel responses are affected by Gaussian additive noise, the multi-voxel likelihood is a multivariate Gaussian distribution:

$$p(\boldsymbol{r}|\mathbf{o}) \sim exp[(\boldsymbol{r}-W\mathbf{o})^{\mathrm{T}}\mathrm{S}^{-1}(\boldsymbol{r}-W\mathbf{o})]$$

where W (N × 19) is a matrix with rows specified by encoding model weights (i.e., the object-category tuning functions) of the selected voxels, and S (N × N) is the noise covariance matrix (section 2.10). This function specifies the likelihood that a natural scene with object indicator vector **o** evoked the observed response $\boldsymbol{r}$.

The object indicator vector **o*** that maximizes the multi-voxel likelihood was found by brute search: all possible object indicator vectors (excluding the null vector with no objects; 2^{19} -1 distinct binary vectors) were generated and evaluated under the multi-voxel likelihood function, and the one with the highest likelihood was selected.

2.10 Estimating the noise covariance matrix

Decoding object categories requires evaluating the multi-voxel likelihood function specified above (section 2.9). In order to evaluate the multi-voxel likelihood function we must first estimate the noise covariance matrix S:

$$S_{\mathrm{ij}} = \langle(r_{\mathrm{i}}-\mathbf{w}_{\mathrm{i}}{}^{\mathrm{T}}\mathbf{o})(r_{\mathrm{j}}-\mathbf{w}_{\mathrm{j}}{}^{\mathrm{T}}\mathbf{o})\rangle$$

where <> denotes averaging across all samples in the training set. The inverse of S is typically unstable. Therefore, we used Tikhonov regularization (or ridge regularization) to regularize the inverse operation (Tikhonov and Arsenin, 1977). With this method, we evaluate the multi-voxel likelihood by replacing S^{-1} with $S^{+} = (S+\alpha I)^{-1}$. Here I is the identity matrix, and α is a regularization parameter that is optimized separately for each individual decoding trial using a leave-one-out procedure. To decode the objects in the j^{th} scene, decoding accuracy was evaluated on the remaining 125 decoding trials using a range of 8 different α values log-spaced between .01 and 100. The optimal α value was then used for decoding on the left-out j^{th} trial.

2.11 Voxel selection procedure for object category decoding

In order to minimize potential selection bias, voxels used for object-category decoding were selected using the training data only. The validation data were not used to select voxels. For each voxel, prediction accuracy (i.e., correlation coefficient) was evaluated on each of the early-stopping sets and averaged across sets. Voxels whose average prediction accuracy on the early-stopping sets was significant ($p < .01$) were selected for use in object-category decoding. For subject 1 the total number of selected voxels was 596. For subject 2 the total number of selected voxels was 653. We emphasize that this selection procedure did not use any of the data that was subsequently used for the decoding analysis.

2.12 Decoding performance as a function of the number of object categories

To determine whether decoding performance was affected by the number of object categories in the scenes, we compared the multi-voxel likelihood (see 2.9 above) of the true object indicator vector for each scene with the multi-voxel likelihood of every other possible object indicator. Thus, we determined the fraction of instances in which $p(\mathbf{r}|\mathbf{o}') > p(\mathbf{r}|\mathbf{o})$, where $\mathbf{o}'$ is the true object indicator vector for a scene, and $\mathbf{o}$ is one of the other possible object indicator vectors. This fraction can be interpreted as the probability that the decoder will correctly discriminate between the true object indicator vector and the object indicator vector for another randomly selected scene. This fraction is plotted as a function of the number of object categories (i.e., the sum of the elements of $\mathbf{o}'$) in Figure 6.

2.13 ROC analysis of object category decoding

The accuracy of object category decoding (Figure 6) was quantified by receiver operating characteristic (ROC) analysis (see Fawcett, 2006 for an introduction). For each object category, the true positive rate (TPR) was defined as the fraction of times the object category was correctly decoded as present in the scene. The false positive rate (FPR) was defined as the fraction of times the object category was incorrectly decoded as present. TPR's and FPR's were calculated separately for each object category. Note that if all object categories occurred with equal probability and object categories were selected randomly, the TPR would equal the FPR for each object category. Therefore objectcategory decoding accuracy was defined as the distance of the point (x,y)=(FPR,TPR) from the line at unity. Here we call this the average *discriminability distance*.

To control for the fact that object categories had unequal probabilities of occurrence, significance of decoding accuracy for each object category was determined using a permutation test. The decoded object categories and the true object categories were permuted across validation trials and the discriminability distance for each of the 19 object categories was re-calculated. This permutation procedure was repeated 1000 times to create a distribution of discriminability distances consistent with the null model. Accuracy for each specific object category was considered significant if the average discriminability distance was greater than or equal to 95% of the values in this distribution.

2.14 Principal component analysis

Principal components analysis (PCA; see Figures 7, 8 and 9) was first applied to the object-category tuning functions of the same voxels selected for the decoding analysis (section 2.11). Let P^1 (19 × 1) be the first principal component. The projection of the object-category tuning function for each voxel onto the first principal component presented in Figure 7 is $\mathbf{p} = WP_1$. Here, W (N × 19) is the matrix of object-category tuning functions for *all* N voxels on the cortical surface (not just those selected for the decoding analysis).

Significance of the principal components was determined using a permutation test. The weights of the object-category tuning functions were permuted independently for each voxel and PCA was then re-applied. This permutation procedure was repeated 10,000 times to create a null distribution of variation explained by each PC. A PC was considered significant if the variation it explained in the actual data was greater than 99% of the values in the null distribution for the corresponding PC. The fist PC was considered significantly greater than the second PC if the difference in variation explained by each was greater than 99% of the values in the null distribution for the corresponding differences.

3 Results

3.1 The object-category model

BOLD signals (hereafter referred to as *voxel responses*) were measured in visually responsive cortex of two subjects (section 2.1). During the scans each subject viewed a total of 1386 unique color natural scenes (~20 deg. field of view). Each scan produced continuously varying BOLD time-series for each voxel. A parametric model of the haemodynamic response function was fit separately to the BOLD time-series data recorded from each voxel. This was used in turn to extract the response of each voxel to each natural scene (for complete details see Kay et al., 2008, supplementary materials). The stimulus/response pairs were then split into separate model training and model validation sets. The training set was used to estimate the weights of the encoding model for each voxel and to select voxels for subsequent decoding analysis; the validation set was used to measure the prediction accuracy of the fit models, and to measure decoding accuracy.

Typical natural scene stimuli are shown at right in Figure 1. The objects in each scene were labeled by the authors and then assigned to one of 19 object categories (section 2.5). These categories were based on a model developed in our previous work (Naselaris et al., 2009). For example, the objects “bear”, “salmon” and “stream” in the top left scene were assigned to the categories “land mammal”, “fish” and “water” respectively. The object categories in each scene were represented by a vector of 19 indicator variables. In the case of the top left scene, the indicator variables for “land mammal”, “fish” and “water” were set to 1. The indicator variables for the remaining 16 object categories were set to 0. Regularized regression (section 2.8) was then used to estimate a separate encoding model for each voxel, using the indicator variables and responses obtained in the model training set (Figure 2A). The resulting *object category* model for each voxel consisted of a set of 19 weights that reflect how each specific category affects voxel responses. We refer to the set of weights estimated for a single voxel as its *object-category tuning function* (see Figure 2B for examples).

3.2 Accuracy of the object-category model

In order to validate any conclusions drawn from the object-category model, we first performed two independent tests to confirm its accuracy. First, we used the fit model for each voxel to generate predicted responses to the natural scenes in the model validation set, and we compared these predictions to the observed responses. We find that the object-category model produces accurate predictions ($p < 0.01$, uncorrected; test of correlation between predicted and measured responses) of voxel responses across a wide band of visual cortex (see Figure 3A). The band extends from the parietal cortex to the ventral temporal cortex and encompasses many category-selective functional ROIs (e.g, FFA and PPA).

The object-category model produces mostly poor predictions in more posterior, retinotopic visual areas (i.e., V1, V2, V3 and V4). This suggests that the object-category model is not simply capturing responses evoked by simple visual features that are correlated with object

category. To confirm this, we compared prediction accuracy of the object-category model to prediction accuracy of a Gabor wavelet model that describes how each voxel is tuned for simple visual features (i.e., spatial frequency, orientation and retinotopic location). This comparison (Figure 3B) shows that voxel responses accurately predicted by the object-category model are often poorly predicted by the Gabor wavelet model. Thus, the object-category model accurately predicts object-related responses that cannot be explained in terms of simple visual features.

The fact that the object-category model accurately predicts responses of many voxels does not necessarily imply that voxel responses encode information about all of the 19 categories in the model. For example, significant prediction accuracy might be achieved if voxel responses encoded information about humans and buildings but none of the other object categories. Yet if this were the case it would not be possible to decode any information about the other object categories from the voxel responses. To control for this case, we used the object-category model to decode each of the 19 object categories from voxel population responses. First, the object category model was used to predict the responses to every possible combination of 19 object categories (excluding the null combination containing no objects). To maximize decoding accuracy this analysis only included voxels for which the object-category model produced accurate predictions (596 voxels for subject 1; 653 voxels for subject 2; in this case prediction accuracy was measured using training data only, see section 2.11). For each scene in the validation set, we then selected the object categories whose predicted population response best matched the actual measured response (see sections 2.9 and 2.10 for details; Kay et al., 2008, Naselaris et al., 2009).

The decoded object categories for two scenes are shown in Figure 4. These examples suggest that most of the object categories present in each scene can be decoded correctly. However, decoding performance could be achieved even by a trivial decoder that simply guesses that all object categories are present in every scene. Thus, we compared the true positive rate (TPR; the fraction of times an object category is correctly identified as present in a scene) to the false positive rate (FPR; the fraction of times an object category is incorrectly identified as present in a scene) for each of the decoded object categories. For most animate and inanimate object categories the TPR is significantly greater than the FPR (see Figure 5). Thus, most object categories can be accurately decoded using the object-category model.

Accurate decoding might also be possible if the decoder accurately identified the object categories in simple scenes but failed to identify the object categories in complex scenes. Therefore, we analyzed decoding performance as a function of the number of object categories present in a scene (Figure 6). We found that the decoder could accurately discriminate between the true indicator variables for a scene and any other set of indicator variables, regardless of the number of objects present in a scene (see section 2.12 for details). Taken together with the results on prediction accuracy our decoding results verify that the object-category model accurately describes how each object category increases or decreases voxel responses.

3.3 Object-category tuning related to the animate/inanimate distinction

Having confirmed the accuracy of the object-category model we used it to examine how the animate/inanimate distinction is reflected in the object-category tuning of voxels. We applied principal components analysis (PCA) to the object-category tuning functions of all voxels for which the object category model provided accurate predictions (i.e., the same voxels selected for the decoding analysis discussed in section 3.2; see section 11 for details). PCA decomposed the object-tuning functions in this population into a set of 19 principal components (PCs). Each PC can be interpreted as an axis in the space of object-category

tuning functions, and is ranked by the amount of variation in object-category tuning that it explains. The first PC (PC 1) thus reflects the most important source of variation in object-category tuning.

Remarkably, the animate/inanimate distinction was reflected by PC 1 (Figure 7): the coefficients for the animate object categories all have the same sign (positive) but are opposite in sign to most of the coefficients for inanimate objects (which are almost all negative). This push-pull relationship between animate and inanimate objects indicates that along the first PC, object-category tuning varies between a strong preference for animate objects and a strong preference for inanimate objects. Voxels with object-category tuning functions that have a large positive projection onto PC 1 will be excited by any kind of animate object and suppressed by any kind of inanimate object. Voxels that have a large negative projection will be excited by any kind of inanimate object and suppressed by any kind of animate object. PC 1 explained 50% – 60% of the variation in object-category tuning across voxels ($p < .01$, permutation test). PC 2 was also significant for both subjects, but it only explained 8 – 10% of the total variation in category tuning across voxels ($p < .01$, permutation test). PC 1 explained significantly more variation than PC 2 ($p < .01$, permutation test) for both subjects. Thus, the functional dissociation between animate and inanimate object categories is reflected in the primary source of variation in object-category tuning.

3.4 Arrangement of animate and inanimate object representations on the cortical surface

How is the variation in object-category tuning along PC 1 mapped on the cortical surface? To answer this, we mapped the projection of the object-category tuning function for each voxel onto PC 1 (see Figure 8). These maps revealed an interesting mirror-symmetric arrangement. Voxels with the strongest preference for animate objects occupy a large density anterior to retinotopic visual areas, and encompass much of EBA, OFA, FFA and the surrounding territory. On the maps, voxels with the strongest preference for inanimate objects are located either above or below the center of this density. Voxels below occupy much of PPA; voxels above extend dorsally from anterior V3a/b to the region occupied by RSC. This mirror-symmetric arrangement is consistent with suggestions from a previous study (Hasson et al., 2003), and with the arrangement of known category-specific ROIs. For example, PPA and RSC are strongly activated by landscapes and buildings. Our data show that voxels within the PPA and RSC are strongly skewed toward a preference for inanimate objects more generally (Figures 8 and 9). In contrast, FFA, OFA and EBA are strongly activated by faces and bodies, respectively. On our maps, voxels within these areas are skewed toward preference for animate objects more generally (Figures 8 and 9). Our maps also reveal many voxels located beyond the borders of these three known category-specific ROIs that strongly prefer either animate or inanimate objects. Thus, the variation in category specificity across these three ROIs is part of a more general and spatially extensive pattern of variation spanning the animate-inanimate distinction.

4. Discussion

We have shown that the most important source of variation in object-category tuning is the variation in preference for animate vs. inanimate object categories. Voxels with a strong preference for inanimate objects flank the superior and inferior aspects of a large density of voxels with a strong preference for animate objects. The respective locations of these voxels are consistent with the locations of known category-specific ROIs, but extend well beyond their boundaries.

4.1 Comparison to other data-driven approaches

Many previous experiments have used a targeted approach to investigate coding for specific animate and inanimate categories (e.g. faces versus places). In contrast, our experiment was very general and was not optimized to test any specific hypothesis about the representation of animate and inanimate categories. The fundamental distinction between animate and inanimate objects emerged from the data after appropriate analysis. Kiani et al (2007) and Kriegeskorte et al. (2009) also used a data-driven approach and arrived at the same central conclusion. However, those earlier studies used a multi-voxel pattern-analysis approach, while we used an encoding model approach (Naselaris et al., 2011).

Our encoding model approach allowed us to estimate how much of the variation in object-category tuning across voxels is accounted for by the animate/inanimate distinction. We find that 50–60% of the total variation is accounted for by this distinction. That is, much of the variation in object-category tuning of individual voxels is related to the degree of preference for animate vs. inanimate objects. Given there is enough information in voxel responses to accurately decode specific sub-categories of animate and inanimate objects (e.g., ‘human’, ‘animal’, ‘car’, ‘building’; see Figure 5), the remaining 50–40% of variation in object-category tuning must be related to sub-categories of animate and inanimate objects. A full understanding of how the tuning for specific sub-categories is organized could be obtained by analyzing the PCs beyond first order (i.e., 2^{nd}, 3^{rd}, …). Our preliminary analysis of these higher-order PCs suggests that they are not easily interpreted.

For the sake of simplicity in this manuscript we have referred to animate- or inanimate-preferring voxels collectively, but these terms risk oversimplifying the complex spatial arrangement apparent in the maps (Figure 8). In fact our results provide no explicit evidence that voxels are organized into discrete animate and inanimate modules. The first PC is an axis, not a categorical designation. For this reason, our results suggest that the strength of preference for animate or inanimate object categories may vary continuously across the cortical surface (see Figure 8). In fact, the maps appear to be perfectly consistent with a spatially smooth variation of animate / inanimate object representations. Of course, lack of evidence for modularity is not evidence of its absence, particularly when analyzing fMRI data. Therefore, on the basis of our data we can only safely conclude that a spatially smooth variation of object representation is a viable possibility.

4.2 The spatial arrangement of object representations

Hasson et al. (2003) reported a large-scale mirror symmetric organization of preferences for buildings, faces, and artifacts in occipito-temporal cortex anterior to retintotopic visual areas. Dorsally and ventrally they found patches of voxels that preferred buildings. Moving toward the lateral surface from either extremity they found alternating patches of voxels that preferred artifacts and faces. This map of object preference appeared to be registered to the map of spatial eccentricity: voxels at more peripheral eccentricities tended to prefer buildings, while voxels at peri-foveal eccentricities tended to prefer faces or artifacts.

The map of PC 1 shown in Figure 7 has a mirror-symmetric organization similar to that reported in Hasson et al. The band of cortex that is best predicted by the object-category model roughly matches the anterior portion of visual cortex analyzed by Hasson et al. Voxels that prefer animate categories are located on the lateral portion of this band, consistent with the face-preferring patches in Hasson et al. Voxels that prefer inanimate categories are generally located on the dorsal and ventral extremities of this band, consistent with the building-preferring patches in Hasson et al. Thus, our data suggest that the arrangement of preferences for building and faces reported by Hasson et al. might apply more broadly to other categories of inanimate and animate objects.

4.3 Using natural scenes to study object representaiton

Most fMRI experiments on object representation have used decontextualized objects that are presented in isolation (e.g. Downing et al., 2006) or in pairs (MacEvoy and Epstein, 2009) on a neutral background. The use of natural scenes raises important conceptual and methodological issues. For example, studies of object representation frequently contrast responses to “faces” and “places”. Yet most natural scenes depict a place--including most that depict faces. The use of natural scenes thus forces a consideration of “place areas” in terms of the specific object categories they represent. Natural scenes are also more efficient because they provide a natural way to probe multiple objects in a single trial. Finally, natural scenes are clearly more ecologically valid than the simple stimuli typically used in experiments. We have presented the first predictive encoding model that relates voxel activity directly to multiple objects in natural scenes. The model also provides the first means for simultaneously decoding multiple objects in natural scenes. Our results therefore demonstrate the feasibility and power of using labeled natural scenes to study object representations.

- A model of object-category tuning predicts BOLD activity evoked by natural scenes.
- The model is used to decode multiple objects in natural scenes.
- The model reveals the sources of variation in object-category tuning across cortex.
- The primary source of variation is preference for animate or inanimate objects.

Acknowledgments

This work was supported by grants to TN and JLG from the National Eye Institute and the National Institute of Mental Health. Some of the software used for fMRI data processing was written by Kendrick Kay.

References

Caramazza A, Shelton JR. Domain-specific knowledge systems in the brain: The animate-inanimate distinction. J. of Cogn. Neurosci. 1998; 10:1–34. [PubMed: 9526080]

David SV, Gallant JL. Predicting neuronal responses during natural vision. Network. 2005; 16:239–260. [PubMed: 16411498]

Downing PE, Chan AW-Y, Peelen MV, Dodds CM, Kanwisher N. Domain specificity in visual cortex. Cereb. Cortex. 2006; 16:1453–1461. [PubMed: 16339084]

Downing PE, Jiang Y, Shuman M, Kanwisher N. A cortical area selective for visual processing of the human body. Science. 2001; 293:2470–2473. [PubMed: 11577239]

Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature. 1998; 392:598–601. [PubMed: 9560155]

Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters. 2006; 27:861–874.

Hansen KA, David SV, Gallant JL. Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. Neuroimage. 2004; 23:233–241. [PubMed: 15325370]

Hasson U, Michal H, Ifat L, Malach R. Large-scale mirror-symmetry organization of human occipito-temporal object areas. Neuron. 2003; 37:1027–1041. [PubMed: 12670430]

Hillis AE, Caramazza A. Category-specific naming and comprehension impairment: A double dissociation. Brain. 1991; 114:2081–2094. [PubMed: 1933235]

Kanwisher N, McDermott J, Chun MM. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 1997; 17:4302–4311. [PubMed: 9151747]

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452:352–355. [PubMed: 18322462]

Kiani R, Esteky H, Mirpour K, Tanaka K. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J. Neurophys. 2007; 97:4296–4309.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 2008; 60:1126–1141. [PubMed: 19109916]

MacEvoy SP, Epstein RA. Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. Curr. Biol. 2009; 19:943–947. [PubMed: 19446454]

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. Neuroimage. 2011; 56:400–410. [PubMed: 20691790]

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009; 63:902–915. [PubMed: 19778517]

Nishimoto N, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. 2011; 21:1641–1646. [PubMed: 21945275]

Peelen MV, Downing PE. Selectivity for the human body in the fusiform gyrus. J. Neurophysiol. 2005; 93:603–608. [PubMed: 15295012]

Spiridon M, Fischl B, Kanwisher N. Location and spatial profile of category-specific regions in human extrastriate cortex. Hum. Brain Mapp. 2006; 27:77–89. [PubMed: 15966002]

Tikhonov, AN.; Arsenin, VY. Solution of Ill-posed Problems. Washington: Winston & Sons; 1977.

Vu VQ, Ravikumar P, Naselaris T, Kay KN, Gallant JL, Yu B. Encoding and decoding V1 fMRI responses to natural images with sparse nonparametric models. Annals App. Stat. 2011; Vol. 5(No. 2B):1159–1182.

Warrington EK, Shallice T. Category specific semantic impairments. Brain. 1984; 107:829–854. [PubMed: 6206910]

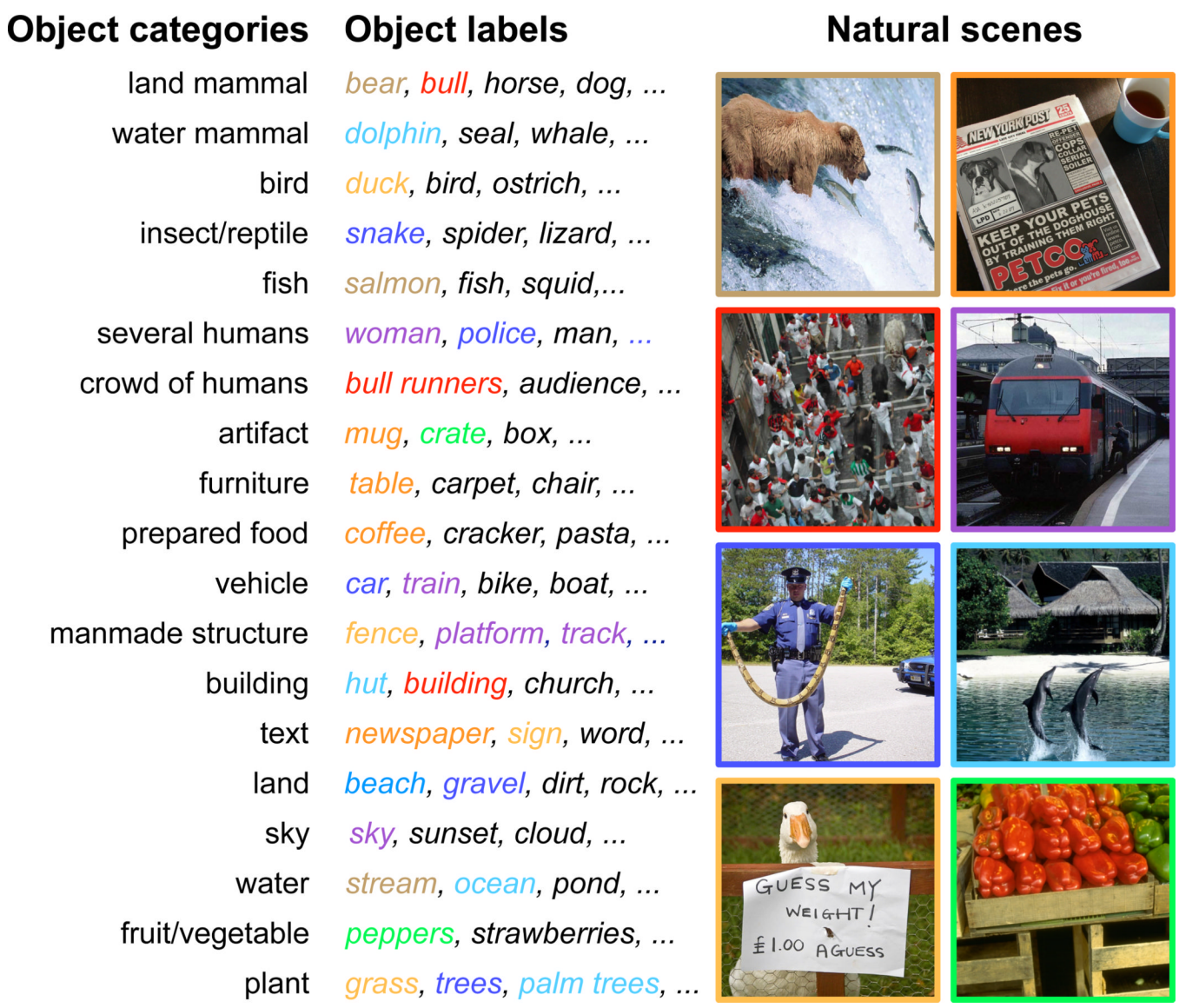

Figure 1. Natural scene stimuli labeled with nineteen object categories

(left) A set of nineteen object categories was selected for our analysis of object representation. (middle) A few example objects that are members of each object category. Objects in colored font appear in the corresponding natural scene at far right. (right) All natural scenes shown here were selected from the 1386 presented during the experiment. Prior to the experiment, the objects in each natural scene were labeled and assigned to the appropriate object category. The entire set of natural scenes contained many examples of objects belonging to each category.

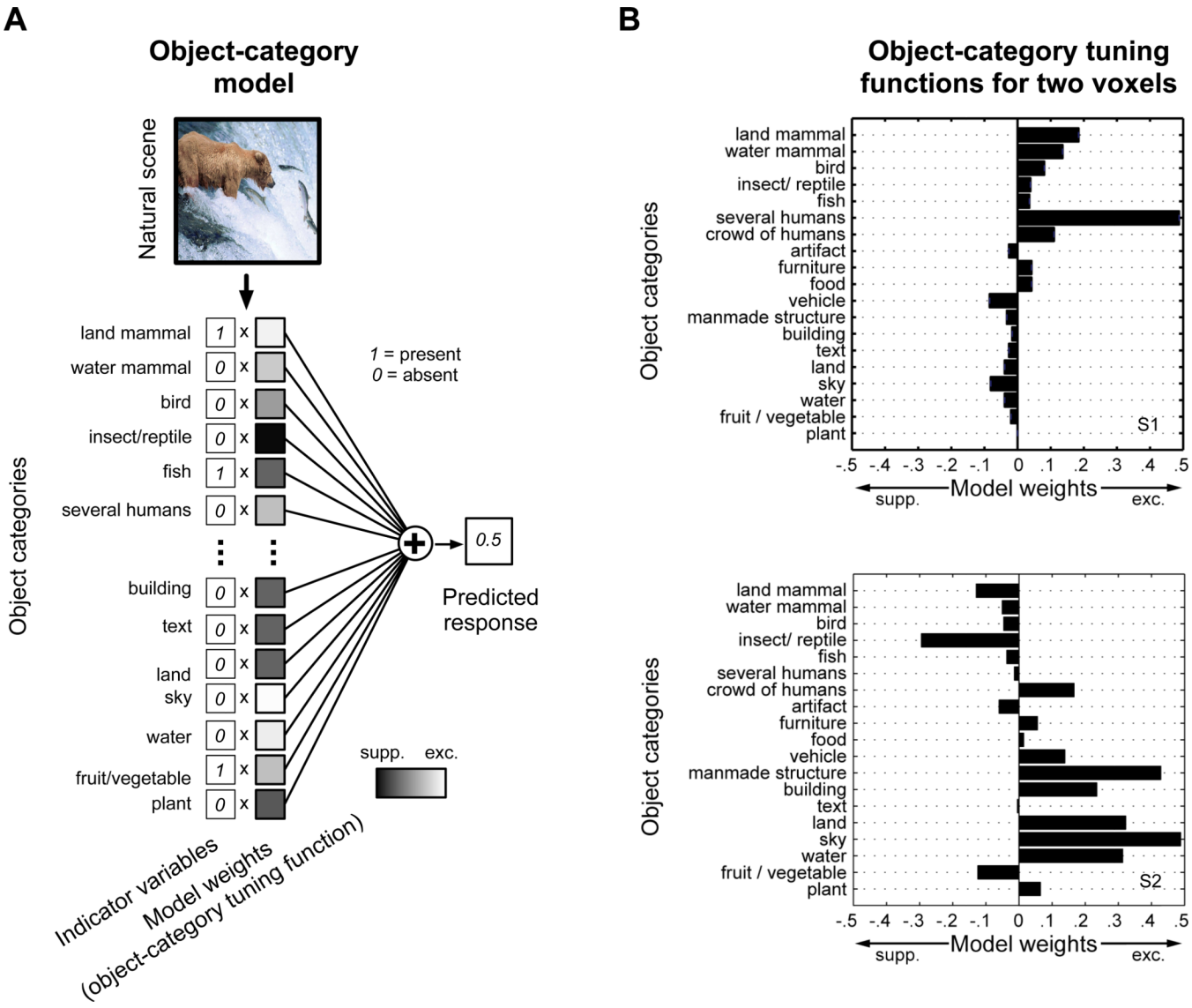

Figure 2. An object-category encoding model based on nineteen object categories

(A) A separate object-category model was constructed for each voxel on the cortical surface. The object category model provides a set of positive (excitatory) or negative (suppressive) weights (plotted as shaded squares) that describe how the presence of each category affects measured BOLD activity. Indicator variables pick out the object categories present in the natural scene; the corresponding weights are summed to generate a predicted voxel response. Voxel responses predicted on a separate data set not used to fit the model are used to verify model accuracy. **(B)** We refer to the set of object-category weights for each voxel as the *object-category tuning function*. Here the object-category tuning functions for two voxels are plotted as bar charts. The voxel shown at top has the highest prediction accuracy across all voxels for subject 1 (voxel # 21240, prediction accuracy $r = 0.697$). This voxel is strongly excited by several humans, though land and water mammals also elicit substantial responses. The voxel at bottom has the highest prediction accuracy across all voxels for subject 2 (voxel # 39097, prediction accuracy $r = 0.733$). This voxel responds to a wide variety of inanimate categories, including sky, water, manmade structures, and buildings. S1 = subject 1. S2 = subject 2.

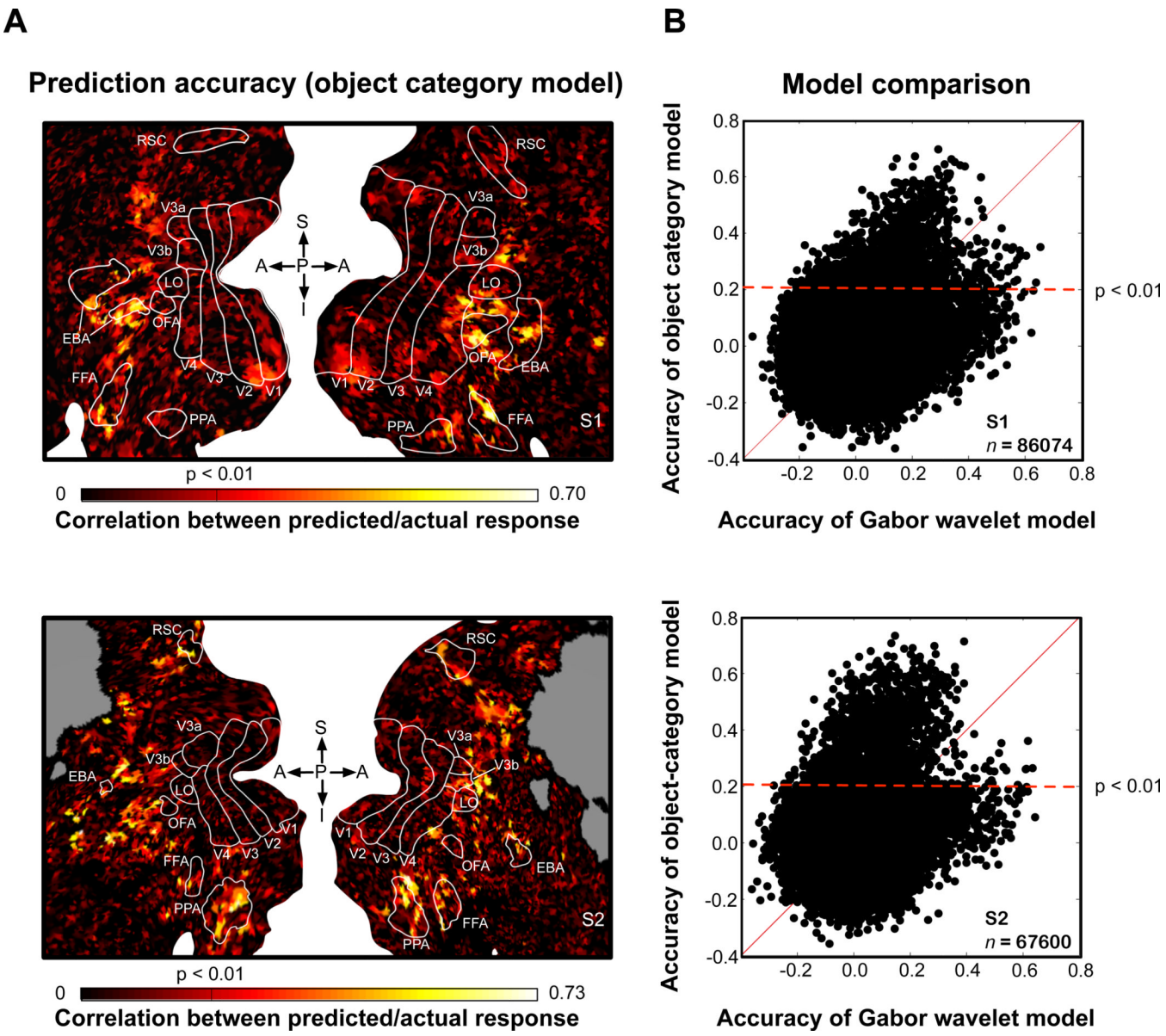

Figure 3. Prediction accuracy of the object category model

(A) Prediction accuracy of the object-category model was estimated separately for each voxel, and these values were projected onto a cortical flat map (top, subject 1; bottom, subject 2). On the map, white space separates the left and right hemispheres; gray indicates locations outside the slice prescription; white lines demarcate functionally-defined ROIs: V1-V4, primary visual cortical areas; LO, lateral occipical complex; OFA, occipital face area; FFA, fuisform face area; PPA, perihippocampal place area; EBA, extrastriate body area; RSC, retrosplenial cortex. Prediction accuracy is represented using a color scale where black represents low accuracy and yellow represents high accuracy. Prediction accuracy is highest for voxels in visual cortex anterior to highly retinotopic visual areas (i.e., V1-V4). **(B)** Prediction accuracy for the object-category model compared to prediction accuracy for a Gabor wavelet model. The Gabor wavelet model depends solely on simple visual features (e.g., spatial frequency and orientation) and does not reference the nineteen object categories included in the object-category model. For each voxel, predicted responses to the validation

stimuli were generated separately using both the object-category and Gabor wavelet models. Prediction accuracy of the object-category model is plotted on the y-axis, and accuracy of the Gabor wavelet model is plotted on the x-axis. Many voxels (black dots) whose responses are predicted accurately by the object-category model (black dots above the dashed horizontal line) are predicted poorly by the Gabor wavelet model. Thus, the object-category model accurately predicts object-related responses that cannot be explained by simple visual features.

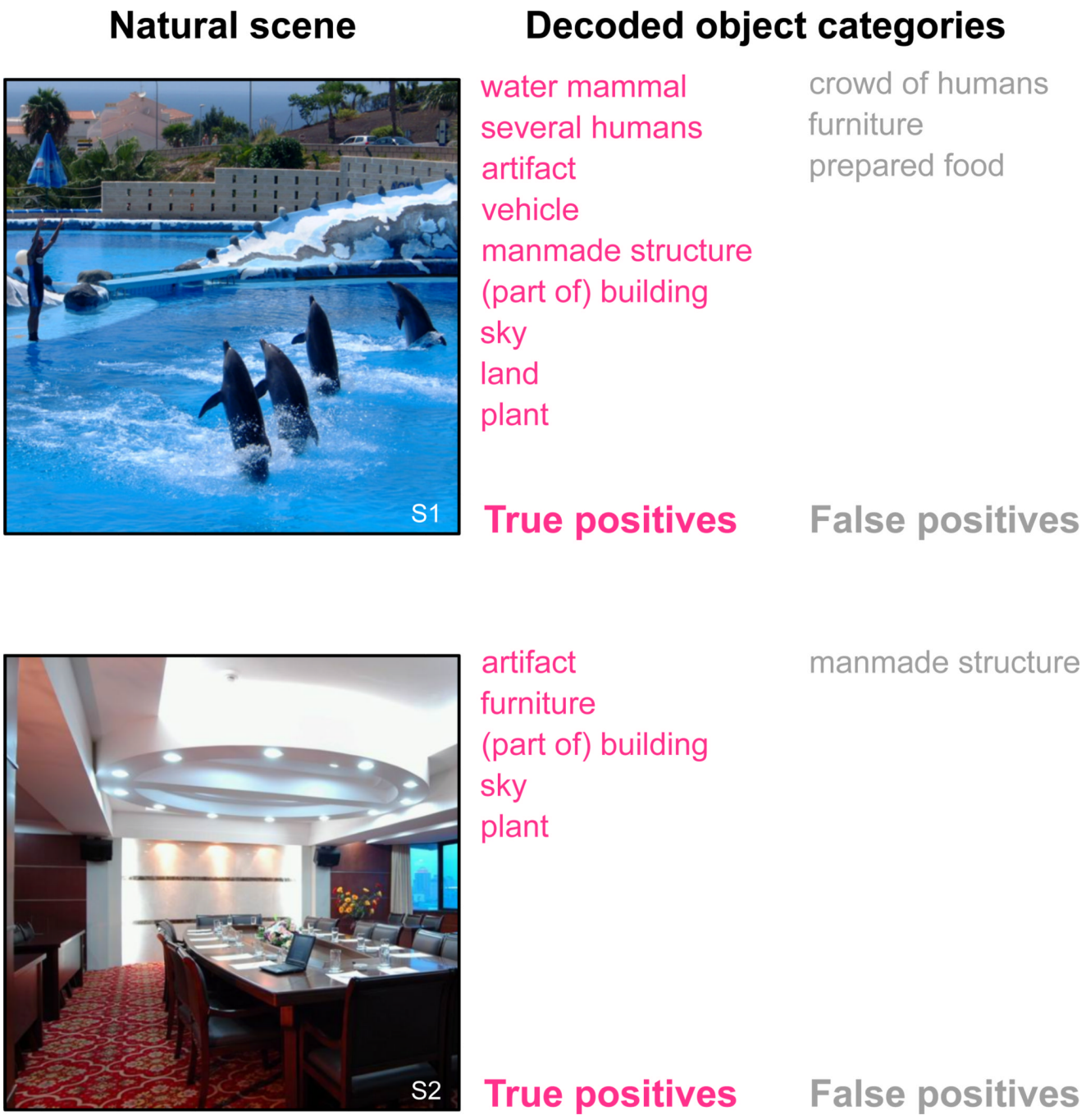

Figure 4. Examples of multiple object categories decoded from complex natural scenes

Decoding can be used to confirm the accuracy of the object-category model. Here the object-category model was used to decode object categories from the responses of voxels for which the object category model provided good predictions (subject 1, n=596; subject 2, n=653). (left) Two natural scene stimuli selected from the validation data set. (right) Object categories that the decoder claims are present in each scene. Object-categories correctly decoded as present (i.e., true positives) are listed in pink, while those incorrectly decoded as present (i.e., false positives) are listed in gray. Decoding is accurate both in heterogeneous scenes that feature objects from many categories (top, subject 1) and in homogeneous scenes that feature objects from fewer categories (bottom, subject 2).

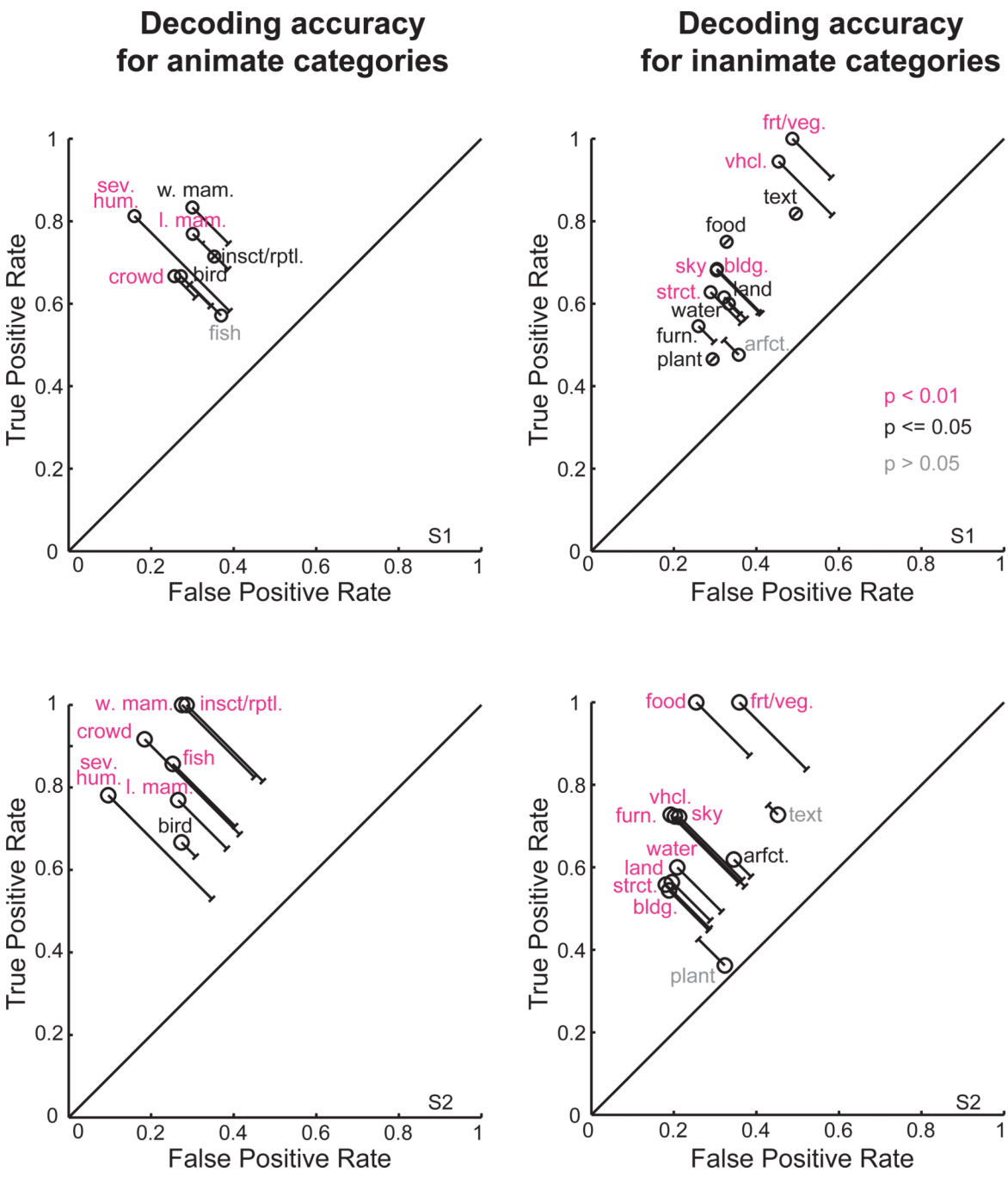

Figure 5. Decoding accuracy for each object category

The object-category model was used to decode the object categories in each image in the validation set, using responses of the same voxels selected in Figure 4. Decoding accuracy for each of the nineteen object categories was analyzed independently. (left) Animate categories. (right) Inanimate categories. (top) Subject 1. (bottom) Subject 2. The vertical axis in each panel gives the true positive rate (TPR), the fraction of scenes in which an object was correctly decoded as present. The horizontal axis in each panel gives the false positive rate (FPR), the fraction incorrectly decoded as present. The solid line at unity represents the TPR and FPR rates that would be expected if the voxel responses provided no decodable information about object category. Object categories farthest from the line at

unity are those that were decoded most accurately. Object categories in pink and in black can be decoded significantly ($p < 0.01$ and $p <= 0.05$ respectively, permutation test). Note that the object categories have different probabilities of occurrence, so the significant distance from line at unity (capped lines indicate significance at $p < .05$) varies across object category. Most of the animate and inanimate object categories are accurately decoded. Abbreviations: l. mam. = land mammal, w. mam. = water mammal, insct./rptl. = insect/reptile, sev. hum. = several humans, crowd = crowd of humans, arfct. = artifact, furn. = furniture, food = prepared food, vhcl. = vehicle, strct. = manmade structure, bldg. = (part of) building, frt./veg. = fruit/vegetable.

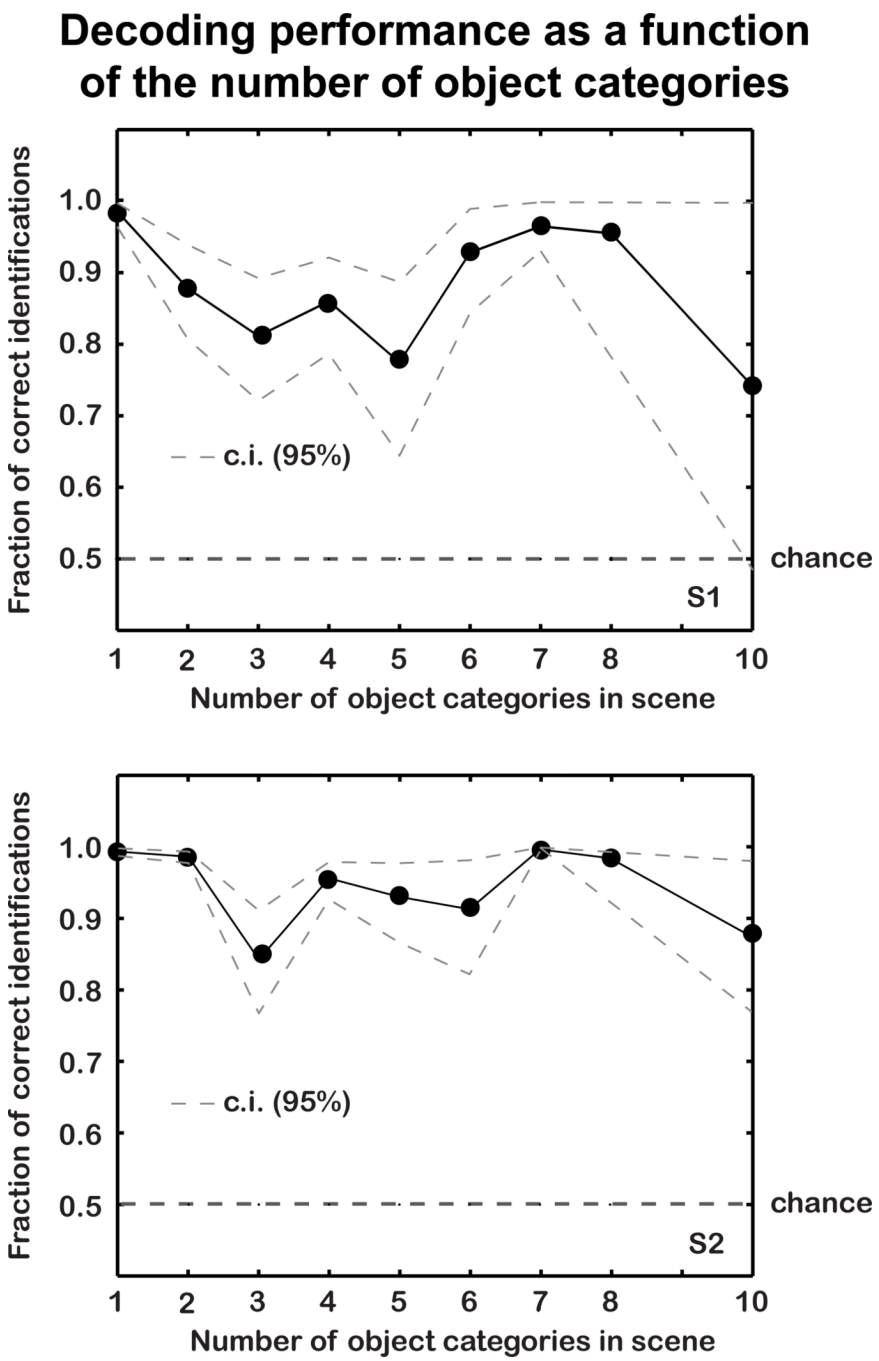

Figure 6. Decoding accuracy as a function of the number of object categories

The horizontal axis gives the number of object categories, and the vertical axis shows the fraction of correct image identifications when the true scene is compared, one-at-a-time, to all other possible scenes. The dashed grey lines indicate bootstrapped 95% confidence intervals. Decoding accuracy shows no systematic relationship to the number of object categories, and the lower bound of the confidence interval is typically above chance (.5) for all numbers of object categories.

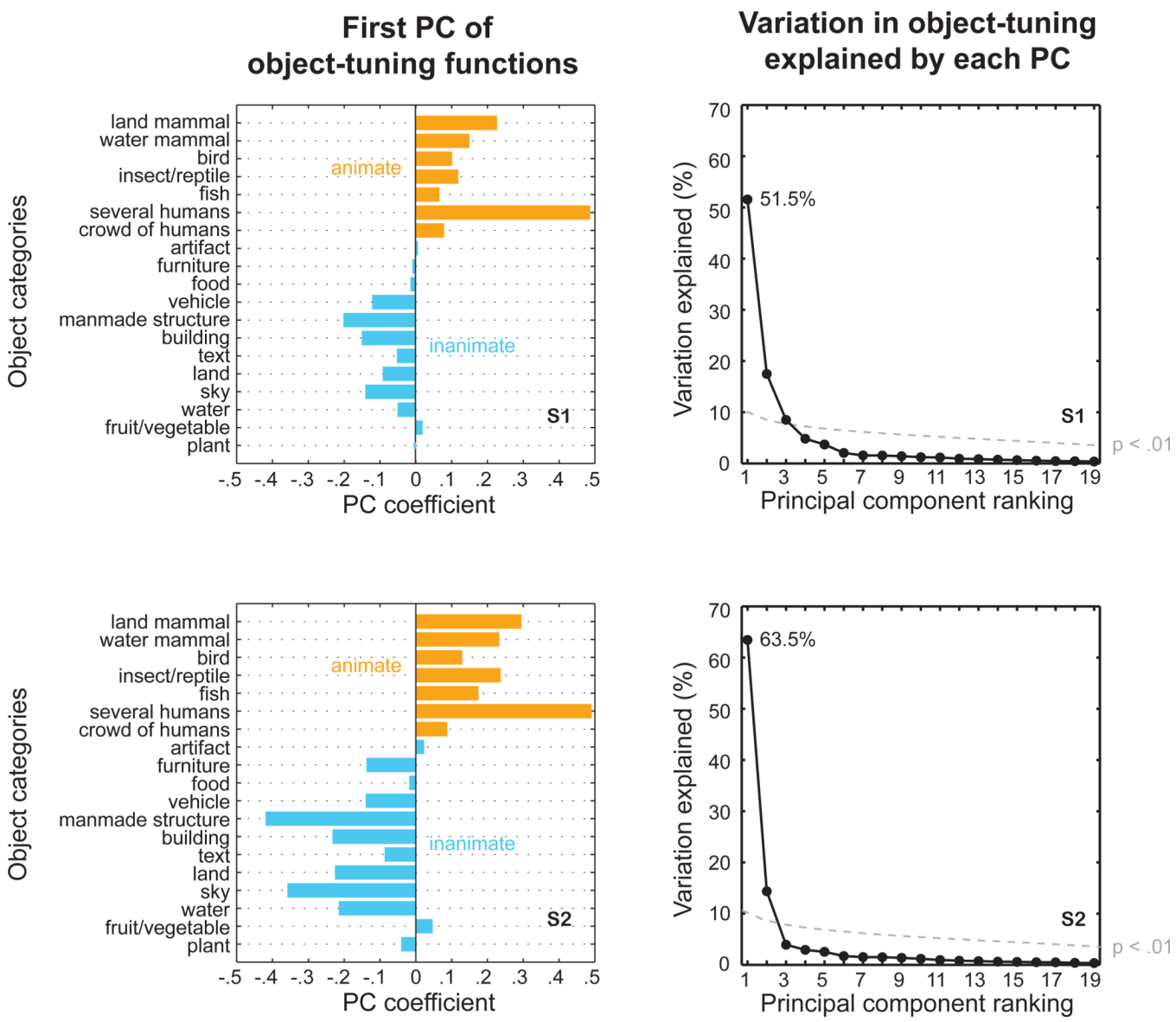

Figure 7. Principal component analysis of object tuning functions

Results of principal component (PC) analysis applied to the object-category tuning functions of the same voxels selected in Figures 4 and 5. (left) Coefficients of the first PC are given on the horizontal axis. The coefficients of the first PC are all of the same sign (positive) for animate categories, and are opposite in sign to the most of the coefficients for inanimate categories. The first PC accounts for 50–60% of the variation (y-axis, right panels) in object-tuning functions across voxels ($p < .01$, permutation test). The significance criterion for each PC (dashed gray line) is the 99th percentile of the histogram of variation explained by the corresponding PC across 10,000 permuted samples. These results suggest that variation in object-category tuning primarily reflects differences in preference for animate and inanimate objects.

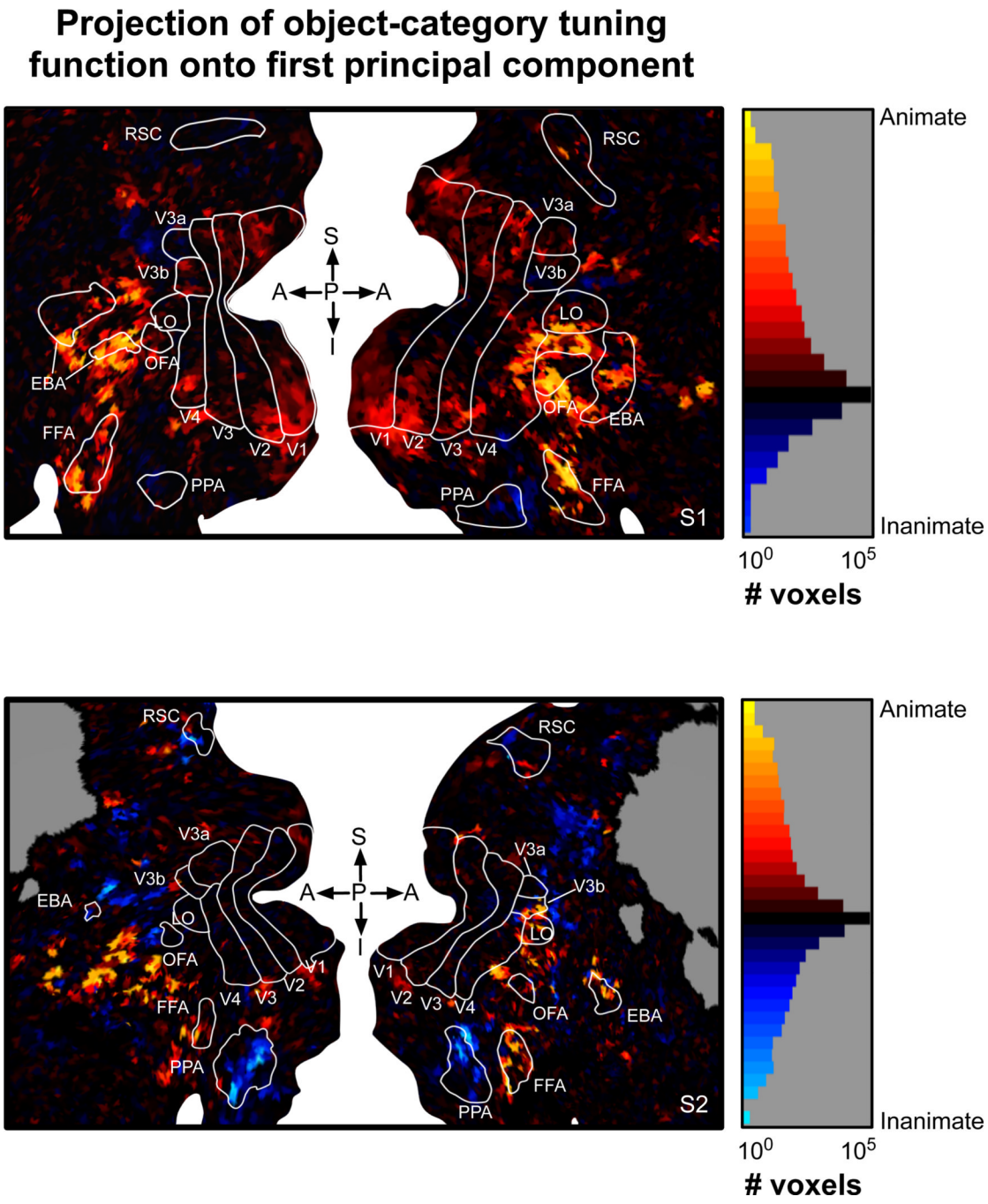

Figure 8. Arrangement of animate and inanimate object representations on the cortical surface

(left) A cortical flat map illustrating the projection of each voxel's object-category tuning function onto the first PC. Details of the maps are the same as in Figure 3. Yellow voxels have large positive projections onto the first PC and generally prefer animate objects. Blue voxels have negative projections onto the first PC and generally prefer inanimate objects. Voxels that prefer animate objects tend to occupy a central density anterior to retinotopic areas. This central density is flanked by voxels with a strong preference for inanimate objects. The locations of these voxels are consistent with the arrangement of category-specific areas (e.g., FFA and PPA), but they extend well beyond the borders of these

classical ROIs. (right) Histogram of the projections of object-category tuning functions onto the first PC (log scale). The color scale is matched to the flatmaps at right.

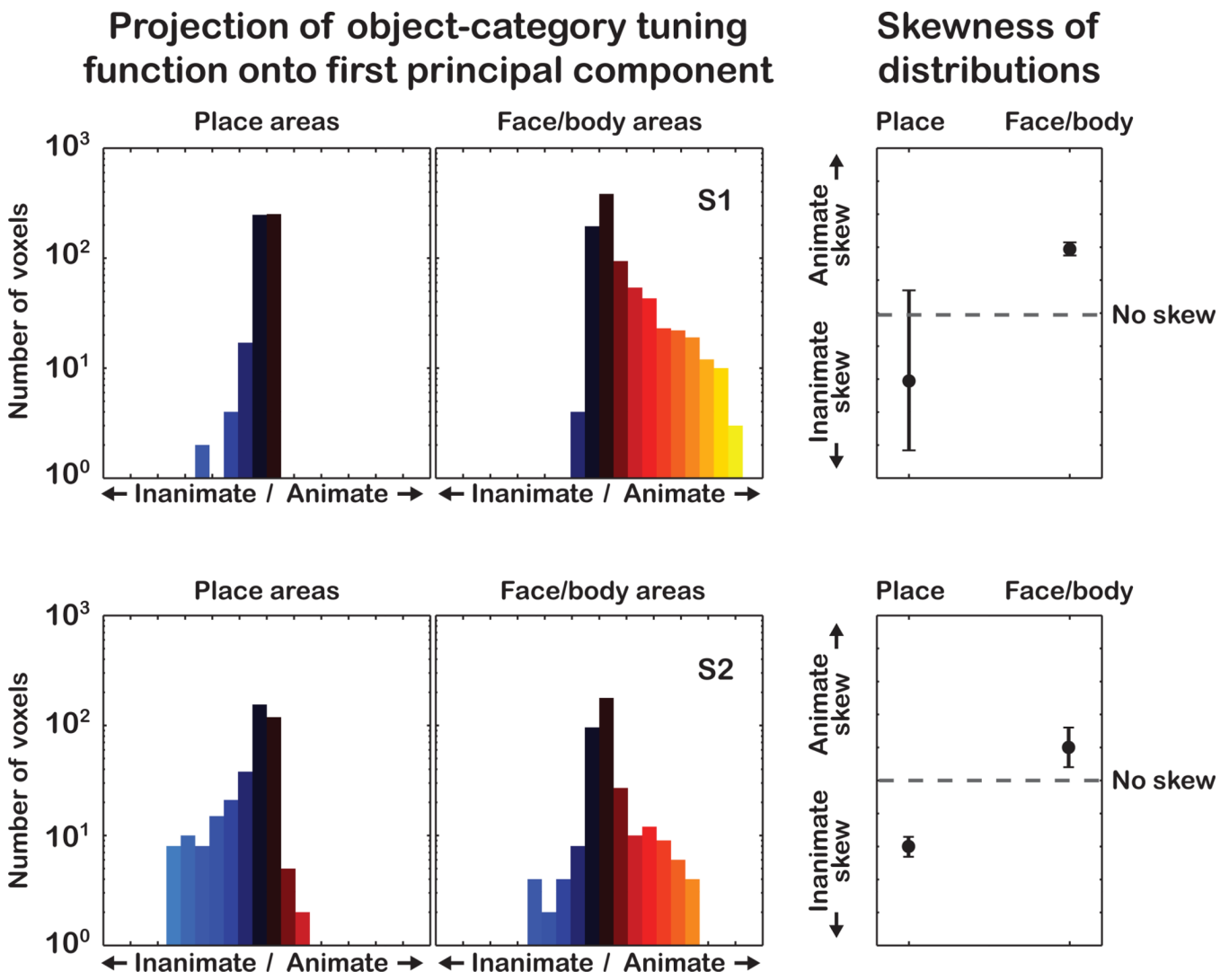

Figure 9. Animate and inanimate object representations in functionally identified regions of interest

(left) Histograms of the projection of the object-tuning functions onto the first PC, for voxels in place-related areas (PPA and RSC) and face or body related areas (FFA, OFA, EBA). (right) Skewness of the histograms with 95% confidence intervals (c.i.). The c.i.'s for skewness were obtained by resampling with replacement the empirical histogram of projection values 10,000 times. The c.i. bounds do not overlap, indicating that place-related areas are skewed toward preference for inanimate categories, and face- and body-related areas are skewed toward preference for animate categories.