Supplemental Material; Neuron Volume 60

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

---

Supplemental Figures

Overview

Fig.	Title
S1	Stimuli
S2	Interspecies correlation of IT object dissimilarities related to single stimuli
S3	Statistical analyses of single-stimulus interspecies dissimilarity correlations
S4	Species-specific face representation
S5	Representation in human early visual cortex defined at 224-5000 voxels
S6	Model representations (1)
S7	Model representations (2)
S8	Unsupervised stimulus-quartet arrangements for monkey and human IT
S9	Representation in left and right human IT
S10	Representation in human IT defined at 100-10,000 voxels
S11	Representation in human IT without FFA and PPA
S12	Linear discriminant analysis for human IT and early visual cortex
S13	Scatterplot of stimulus-pairs relating monkey- and human-IT representations
S14	Scatterplot of stimulus-pairs relating monkey- and human-IT representations (continued)

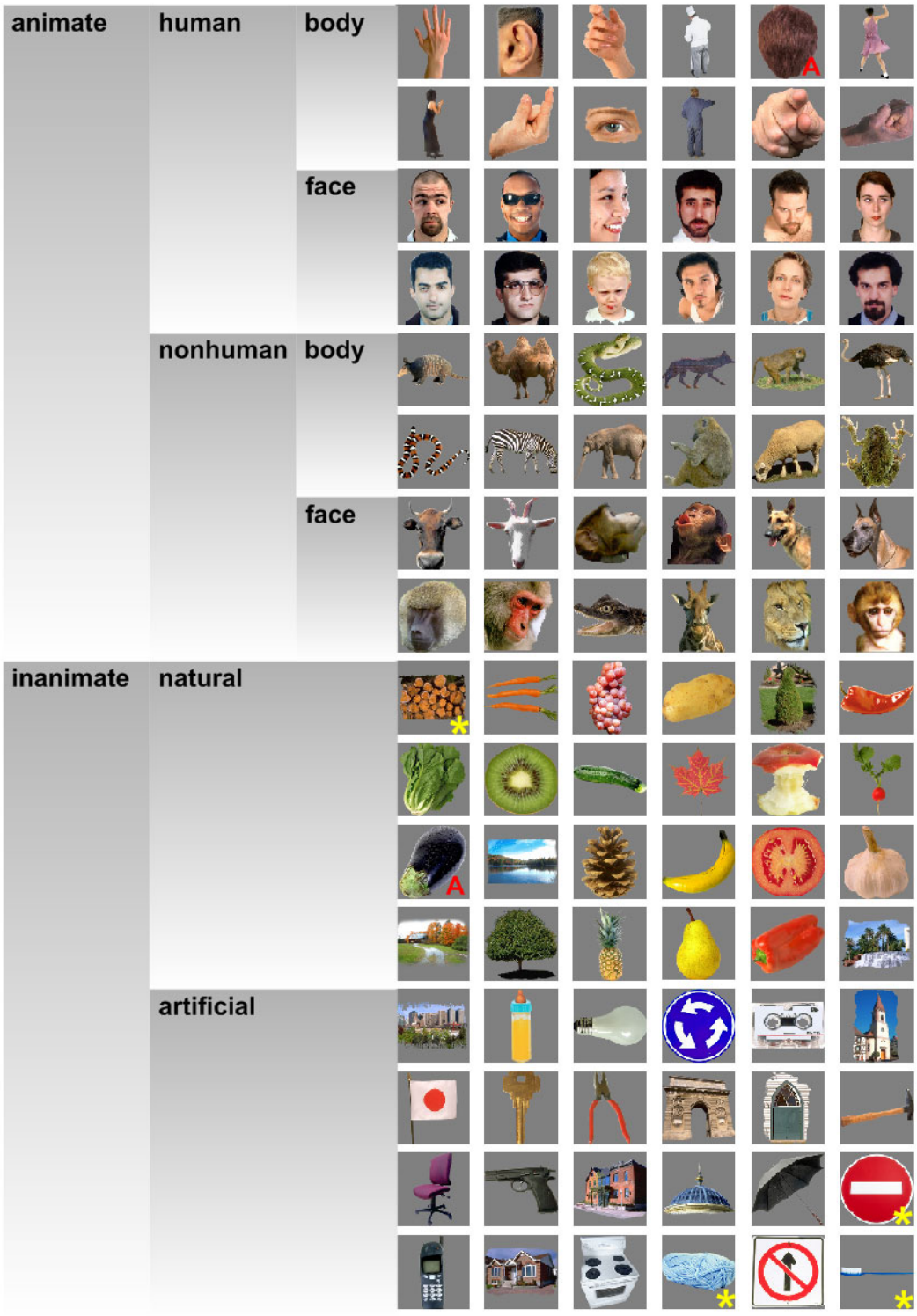


Fig. S1. Stimuli. The object images presented to monkeys and humans. The four images marked by yellow stars were excluded from the analysis because of insufficient data in the monkey experiments. Responses to the remaining 92 form the basis of all analyses. Several of our human subjects described two of the stimuli as ambiguous during debriefing. These two stimuli (egg plant, back of a human head) are marked by a red “A”.

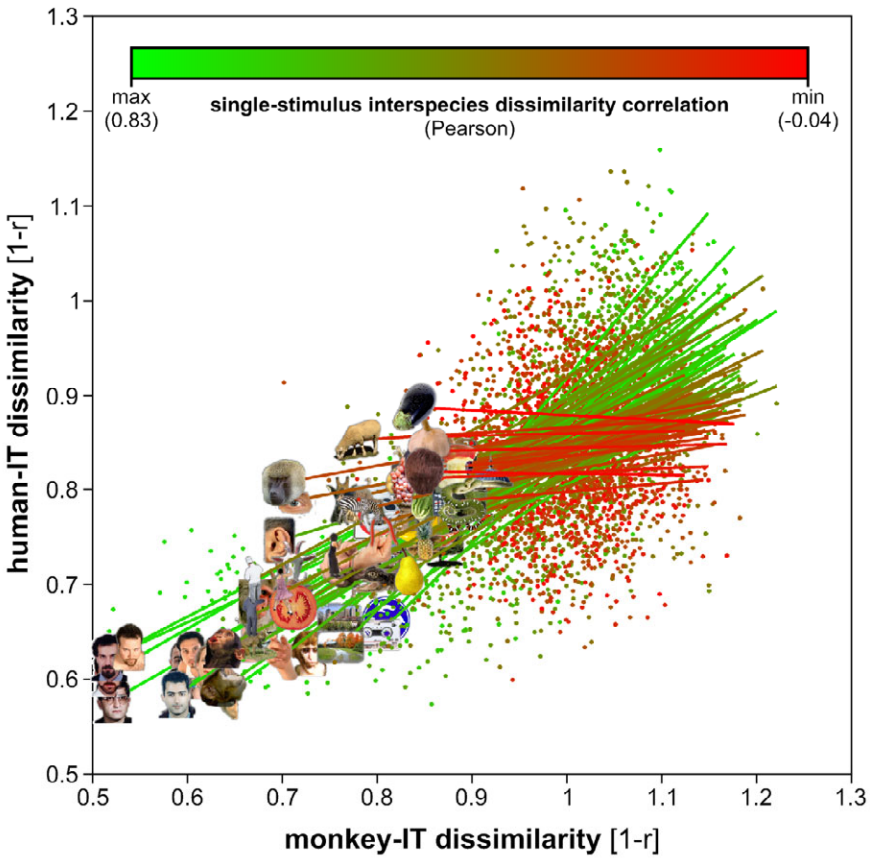
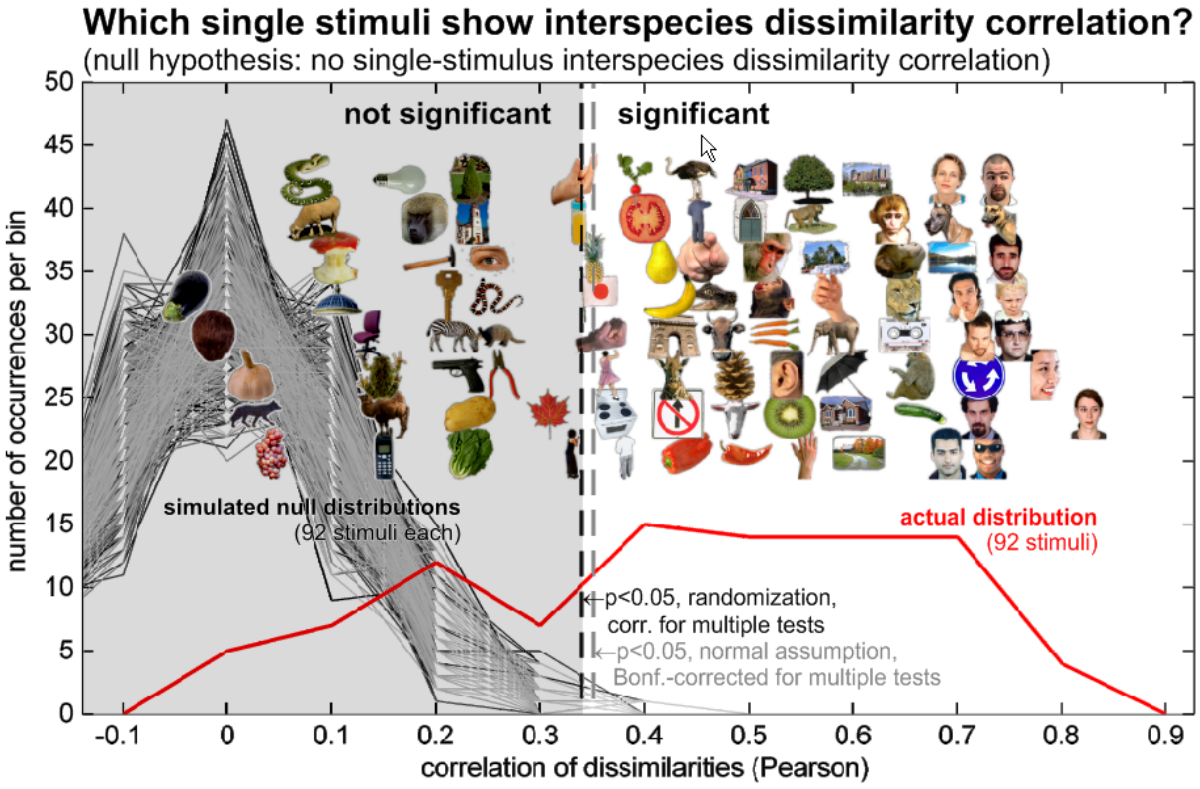


Fig. S2. Interspecies correlation of IT object dissimilarities related to single stimuli. This figure addresses the question to what extent each of the stimuli is similarly represented in both species. A stimulus is considered "similarly represented" if its pattern of representational dissimilarities to the other 91 stimuli is correlated between human and monkey IT. For each stimulus, we consider its row (or, equivalently, its column) in the representational dissimilarity matrix in each species (Fig. 1) and plot the monkey-IT dissimilarities against the human-IT dissimilarities (analogously to Fig. 3). In addition, we plot a straight line for each stimulus, which is obtained as the least-squares fit to the 91 human-IT dissimilarities (vertical axis) of that stimulus and extends horizontally along the range of the corresponding 91 monkey-IT dissimilarities. The stimulus itself is plotted on the left end of the line. In order to highlight the stimuli most inconsistently represented in monkey and human, the scatterplots, fit lines, and stimuli are overplotted in the order of their interspecies representational-dissimilarity correlation, starting from the most highly correlated (green scatterplot and fit line, thin fibers in Fig. 2b) and progressing to the least interspecies-correlated stimulus (red scatterplot and fit line, thick fibers in Fig. 2b). The scatterplots and fit lines for intermediate stimuli are plotted in intermediate colors ranging from green to red, which linearly reflect the interspecies correlation (see colorbar).

a


Which single stimuli show interspecies dissimilarity correlation?
(null hypothesis: no single-stimulus interspecies dissimilarity correlation)
not significant
significant
number of occurrences per bin
simulated null distributions
(92 stimuli each)
actual distribution
(92 stimuli)
p<0.05, randomization,
corr. for multiple tests
p<0.05, normal assumption,
Bonf.-corrected for multiple tests
correlation of dissimilarities (Pearson)

b

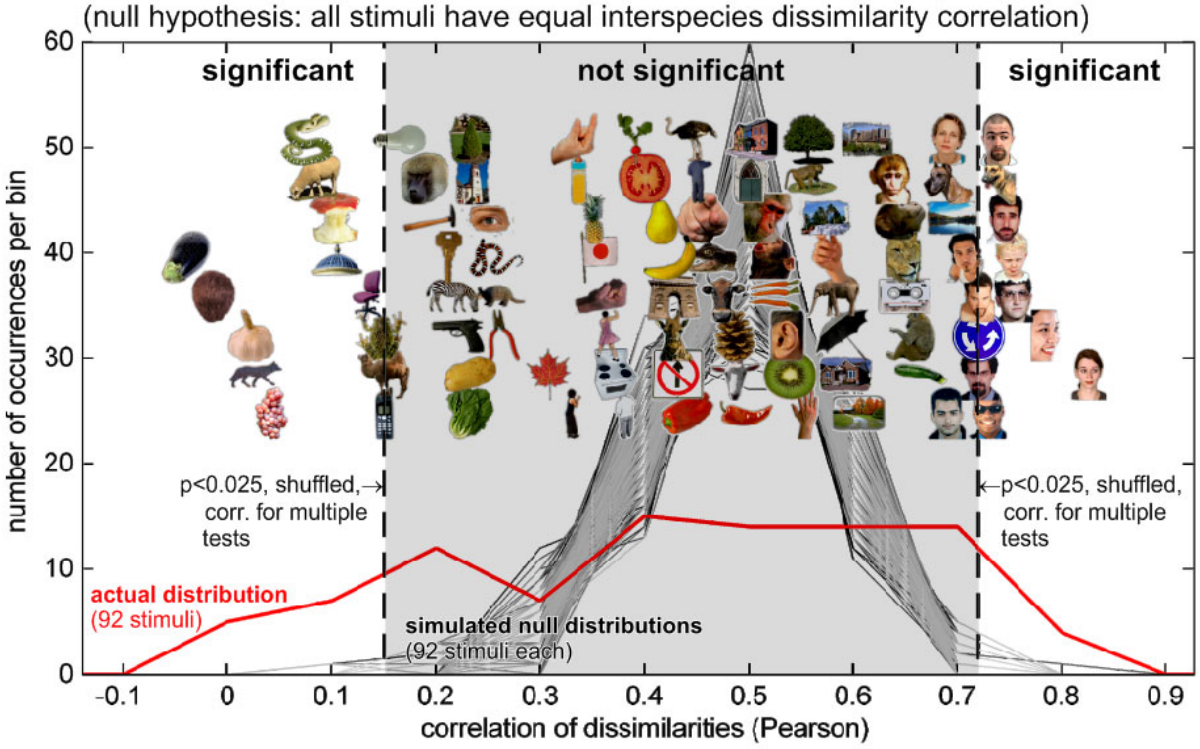

Do stimuli show different interspecies dissimilarity correlation?
(null hypothesis: all stimuli have equal interspecies dissimilarity correlation)
significant
not significant
significant
number of occurrences per bin
p<0.025, shuffled,
corr. for multiple
tests
p<0.025, shuffled,
corr. for multiple
tests
actual distribution
(92 stimuli)
simulated null distributions
(92 stimuli each)
correlation of dissimilarities (Pearson)

Fig. S3. Statistical analyses of single-stimulus interspecies dissimilarity correlations. In both panels (a and b), each stimulus is placed along a horizontal axis according to its interspecies dissimilarity correlation (see previous figure for details). The stimuli are spaced out vertically so that they could be displayed in a larger size. In each panel (a and b), the red line shows the interspecies-correlation histogram for the 92 stimuli. (**a**) Statistical analysis addressing the question, which single stimuli show interspecies dissimilarity correlation. As in Fig. 3, we use randomization of stimulus labels to test the interspecies dissimilarity correlation. Here we assess, which single-stimulus interspecies correlations are significant. Each gray line shows a histogram obtained by randomizing the stimulus labels for one species before computing the interspecies dissimilarity correlations. Each of 1000 such randomizations simulates the null hypothesis that there are no interspecies correlations. We define an interspecies correlation threshold (dashed black line) that is exceeded by even a single stimulus in only 5% of the 1000 null simulations (i.e. by thresholding the randomization distribution of maxima among the 92 interspecies correlations obtained in each null simulation). This threshold limits the family-wise false-positives rate at $p<0.05$. A similar threshold (dashed gray line) is obtained by (incorrectly) assuming normality and independence, and using the Bonferroni method to control the family-wise false-positives rate. Using either method, about 61 of the 92 stimuli, including all faces, exhibit significant interspecies correlation. (**b**) Statistical analysis addressing the question, whether stimuli vary in terms of interspecies correlation. The analysis in (a) highlights some stimuli and not others as significantly consistently represented in IT of both species. However, this does not mean that there are significant differences between single-stimulus interspecies correlations. A mixture of significant and insignificant interspecies correlations as obtained in (a) could result from an interspecies correlation constant across all stimuli in conjunction with noise. We therefore tested the null hypothesis that all single-stimulus interspecies correlations are equal. We simulated the null distribution of equal interspecies correlation across all stimuli by shuffling interspecies pairs of dissimilarites across stimuli (without replacement). This conserves the overall interspecies correlation and yields single-stimulus interspecies correlations that differ only because of the noise and limited data points (91 interspecies dissimilarity pairs for each stimulus). Each of 1000 null simulations yielded an interspecies correlation for each stimulus (1000 histograms shown in gray). We define a lower interspecies correlation threshold (dashed black line on the left), such that lower interspecies correlations occur for even a single stimulus in only 2.5% of the 1000 null simulations (i.e. by thresholding the randomization distributions of minima). Analogously, we define an upper threshold (dashed black line on the right), such that higher interspecies correlations occur for even a single stimulus in only 2.5% of the 1000 null simulations (i.e. by thresholding the randomization distributions of maxima). These two thresholds limit the family-wise false-positives rate at $p<0.05$. Results show that human faces exhibit significantly higher interspecies correlations than the stimulus set as a whole and several stimuli (including images of animate and inanimate objects) exhibit significantly lower interspecies correlations. The two stimuli with the lowest interspecies correlation (eggplant, back-view of human head) were the only two stimuli described as ambiguous by human subjects during debriefing (Fig. S1). Their significantly low interspecies correlation is consistent with the idea that the IT representation reflects not only the visual appearance, but also the conceptual interpretation of a stimulus.

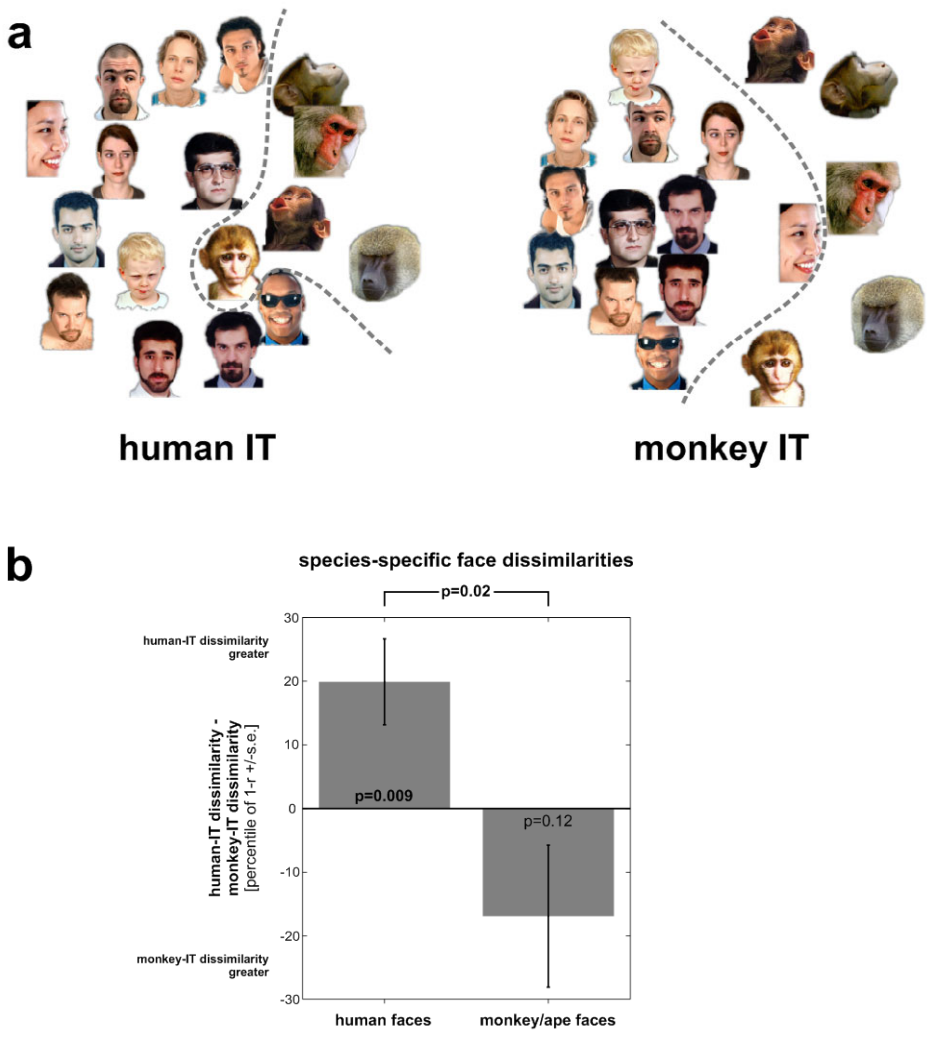


Fig. S4. Species-specific face representation. Here we selectively analyzed the representation of monkey, ape, and human faces in monkey and human IT. (**a**) The face stimuli have been arranged such that their pairwise distances approximately reflect response-pattern similarity. The arrangement was computed by multidimensional scaling with the same settings as in Fig. 2 and elsewhere in this paper (dissimilarity: 1 - Pearson r, criterion: metric stress, arrangements scaled to match the areas of their convex hulls and rigidly aligned for easier comparison with the Procrustes method). A line (dashed gray) separating the monkey/ape faces from the human faces has been manually added. Visual inspection suggests that human IT may better discriminate the human faces than the monkey faces and that the converse may hold for monkey IT. (**b**) Statistical analysis comparing human- and monkey-IT mean dissimilarities for human faces (left) and for monkey/ape faces (right). The left bar shows that dissimilarities among human faces are significantly larger in human IT than in monkey IT (p=0.009). The right bar shows that dissimilarities among monkey/ape faces are larger in monkey IT than in human IT in our data, although the effect is not significant (p=0.12). The difference between the two effects is significant (p=0.02). Because the dissimilarities are not independent or normal, the statistical tests and error bars (indicate +/- 1 standard error) are based on bootstrap resampling of the stimulus set. Note that our stimulus set is not well-suited for comparing the representation of human and monkey/ape faces, because faces were a small subset of our stimuli and because the monkey/ape faces were few and varied in species, pose, and view more than the human faces. The comparison in (b) of the representations of a given set of stimuli (either human faces or monkey faces) between human and monkey IT nevertheless provides an interesting lead for future studies designed to address this question.

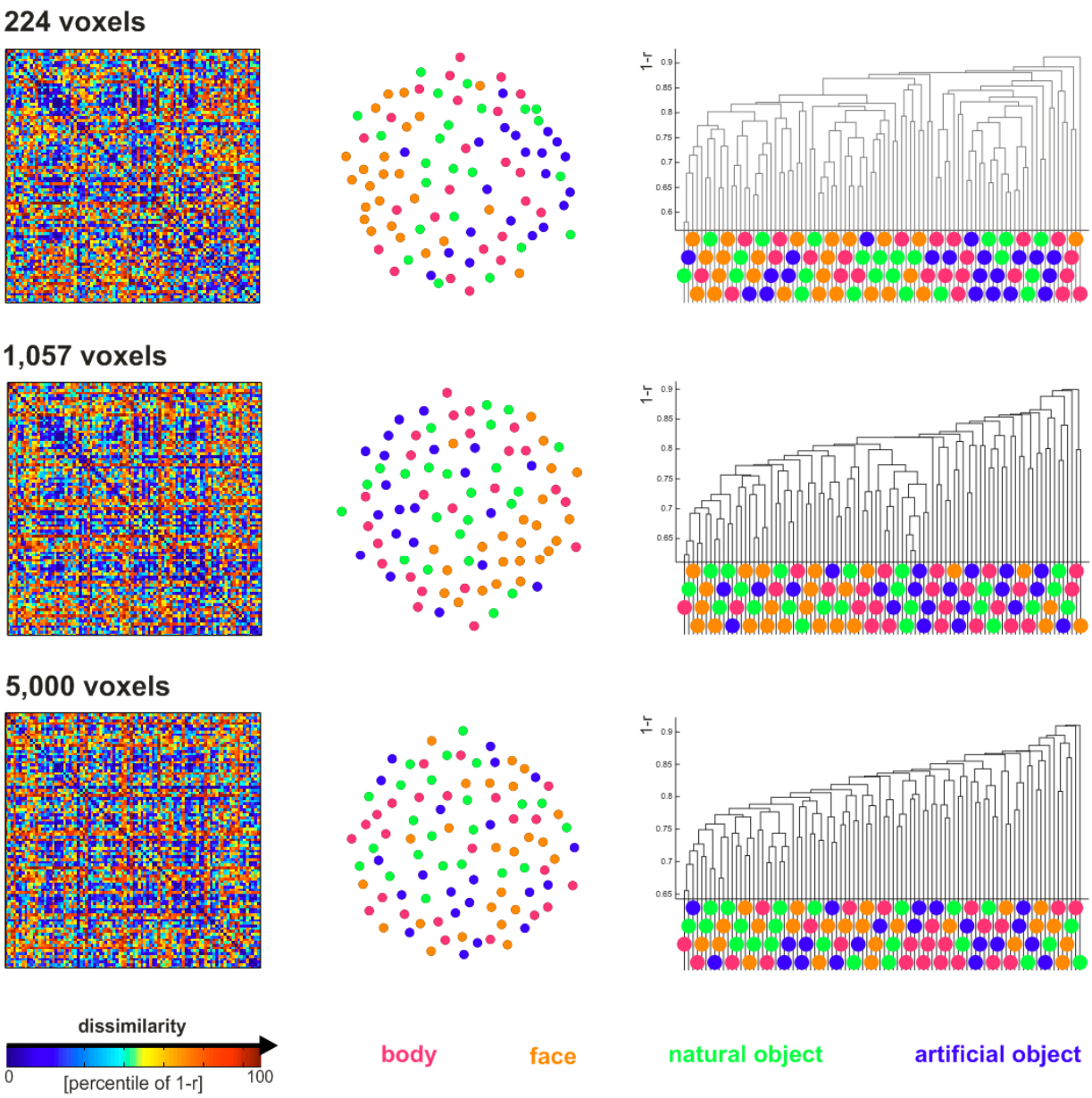


Fig. S5. Representation in human early visual cortex defined at 224-5000 voxels. Human early visual cortex shows no evidence of categorical clustering in the fMRI data. This result is independent of the number of voxels included in the region of interest (rows). Early visual cortex was defined by selecting the most visually responsive voxels within a manually drawn anatomical mask in each subject. As for human IT, independent data were used for voxel selection. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for IT (Figs. 1, 2, 4, respectively).

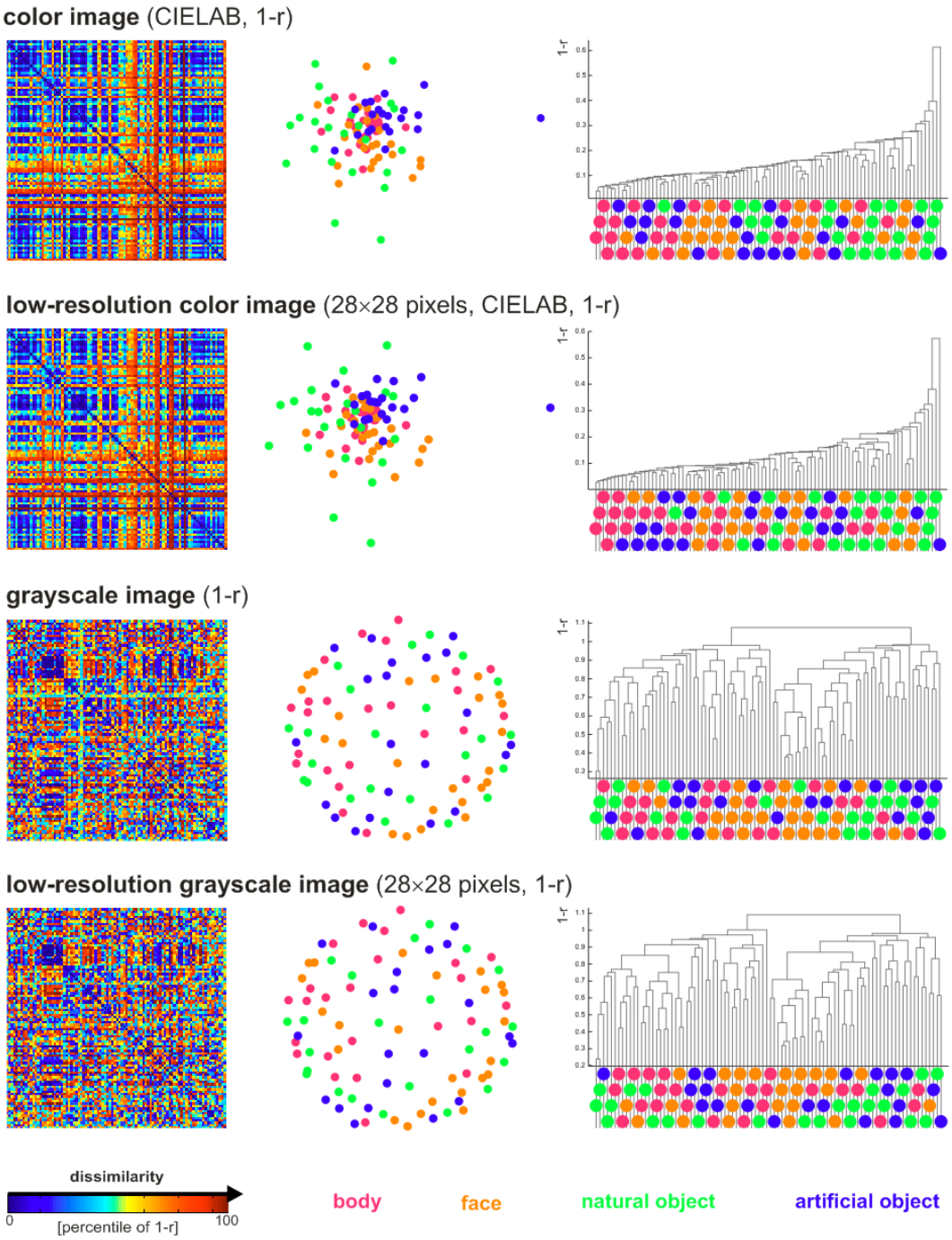


Fig. S6. Model representations (1). We processed our stimuli to obtain their representations in a number of low-level models (rows, continued in Fig. S7). We analyzed these model representations in the same way as the brain-activity data from early visual cortex and IT. None of the models could account for the categorical clustering found in monkey and human IT. The models are described in the section *Model representations* in the Supplemental Material. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for IT (Figs. 1, 2, 4, respectively).

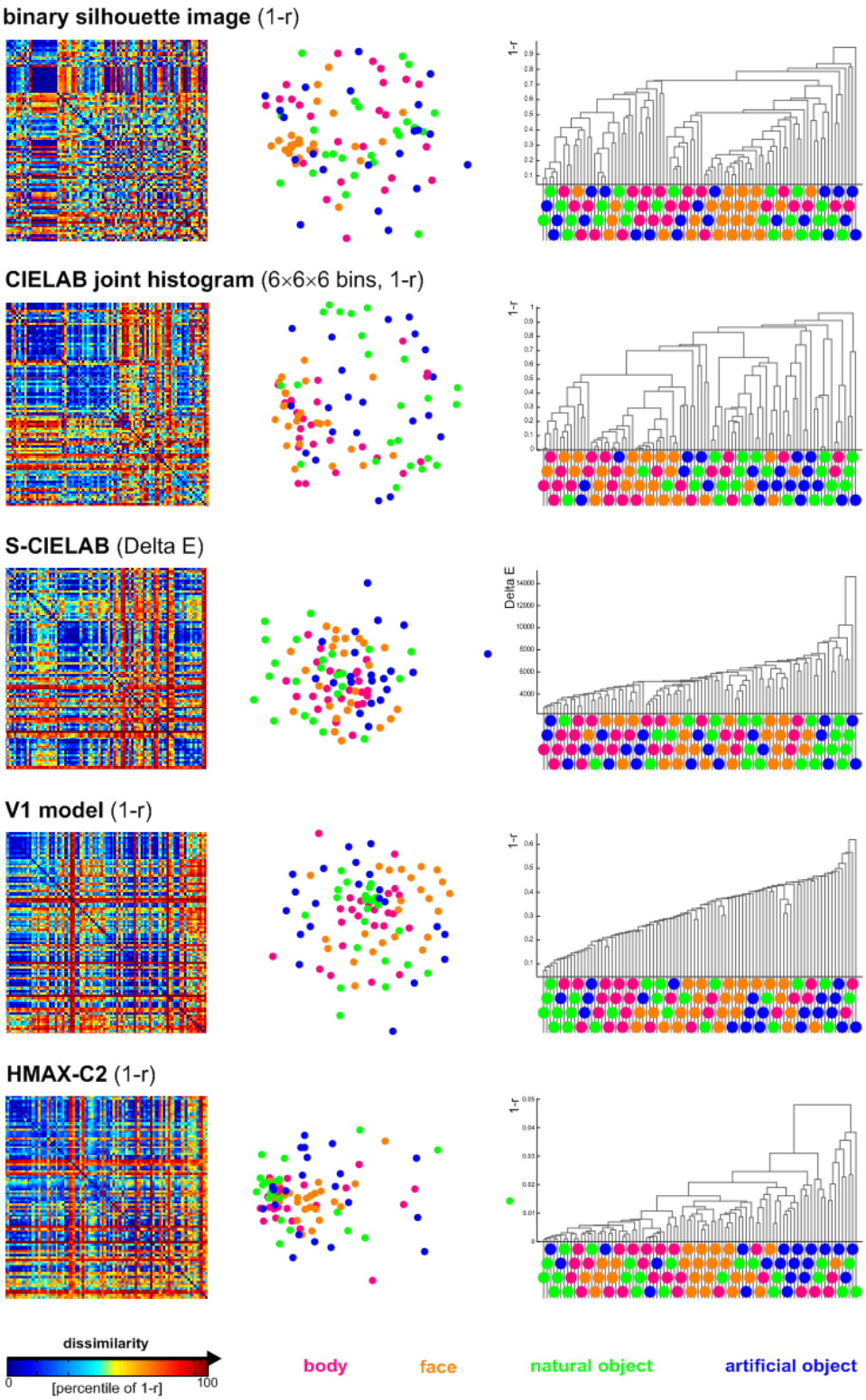


Fig. S7. Model representations (2). Continuation of Fig. S6. See legend of Fig. S6.

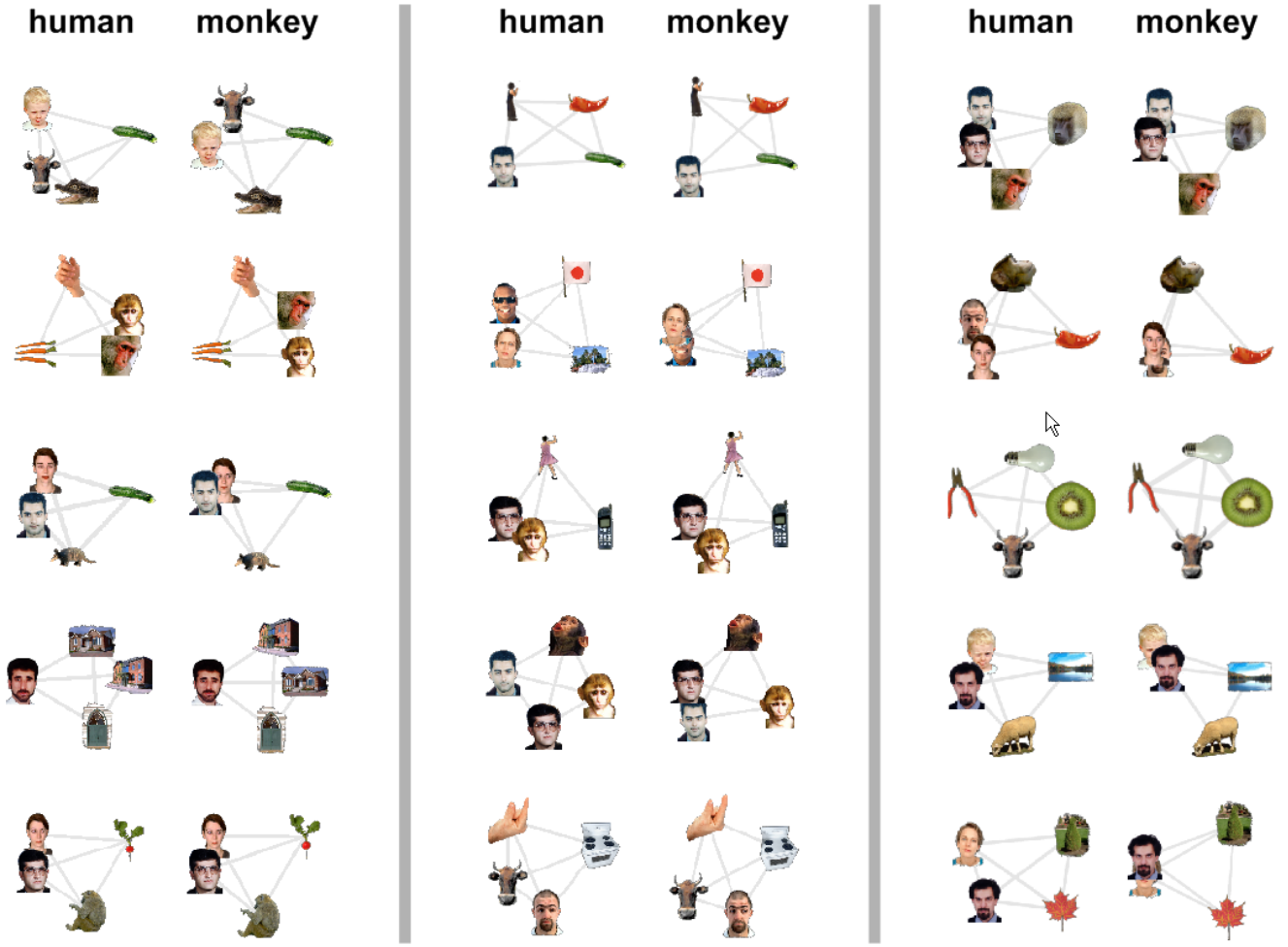


Fig. S8. Unsupervised stimulus-quartet arrangements for monkey and human IT. For 15 representative stimulus quartets, this figure depicts the representation in human and monkey IT in terms of unsupervised arrangements reflecting response-pattern similarity. As in Fig. 2, images placed close together elicited similar response patterns; images placed far apart elicited dissimilar response patterns. There is no special significance to the choice of 4 as the number of stimuli. However, considering quartets allows us to appreciate the underlying dissimilarity relationships at a glance. Moreover, the inevitable distortion of the original dissimilarities in the 2-dimensional arrangement is very small when only 4 stimuli are considered at a time. The arrangements were computed using multidimensional scaling with the same settings as for Fig. 2 (dissimilarity: 1-Pearson correlation, criterion: metric stress). For each stimulus quartet, the two arrangements (human, monkey) have been rigidly aligned for easier comparison (Procrustes alignment) and scaled to the same approximate size. We introduce “rubberband graphs” (gray lines connecting the stimuli) to depict the residual distortion: the gray lines behave like rubberbands, thinning when stretched beyond the length they are to represent and thickening when compressed. More precisely, the actual dissimilarity equals the area of the rubberband connection (dissimilarity = line thickness × line length).

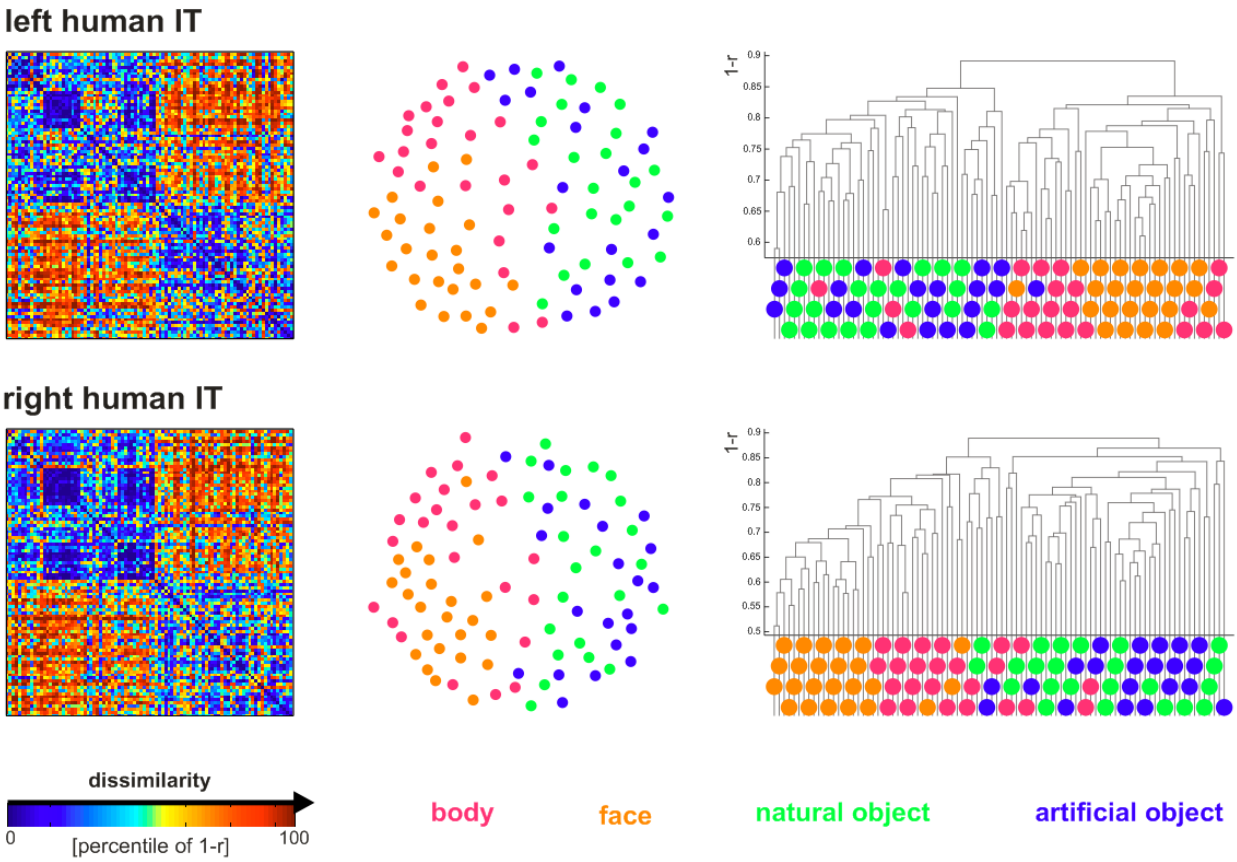


Fig. S9. Representation in left and right human IT. The categorical clustering in human IT is only weakly dependent on the cortical hemisphere (left human IT in top row, right human IT in bottom row). Here we selected 266 voxels according to their visual responsiveness (independent data) within each hemisphere's manually defined IT mask. Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for Figs. 1, 2, 4.

.

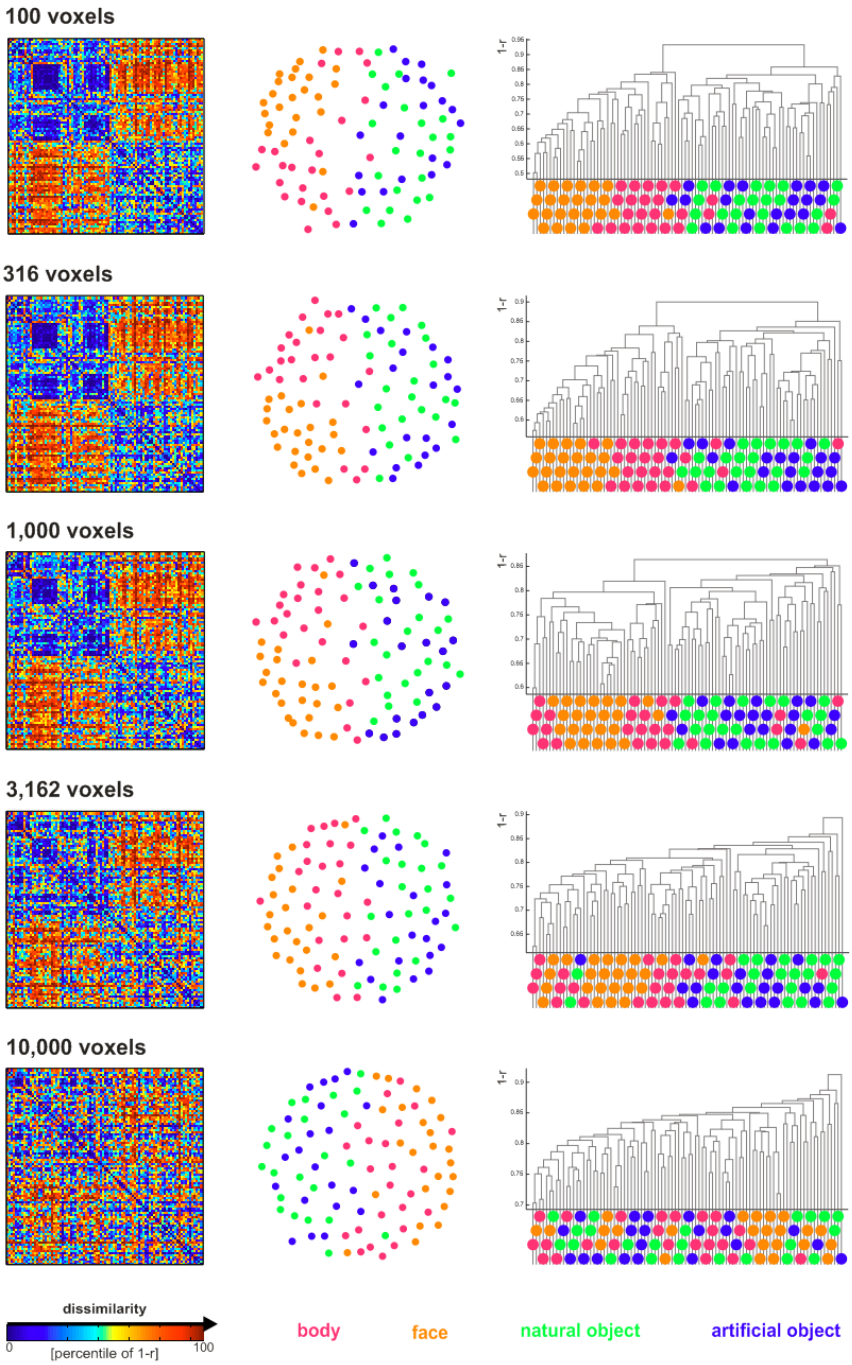


Fig. S10. Representation in human IT defined at 100-10,000 voxels. The similarity structure and categorical clustering characteristic of human IT is only weakly dependent on the number of voxels selected for inclusion in the region of interest. Voxels were selected according to their visual responsiveness as assessed with independent data. The human-IT region shown in the second row (316 voxels) is that used for Figs. 1-6. When thousands of voxels are included in the region of interest, the categorical structure becomes less distinct. Nevertheless multidimensional scaling still separates animate and inanimate objects at 10,000 voxels (bottom row, middle panel). Dissimilarity matrices (left), multidimensional scaling arrangements (middle), and hierarchical clustering trees (right) were computed with the same parameters as for Figs. 1, 2, 4.

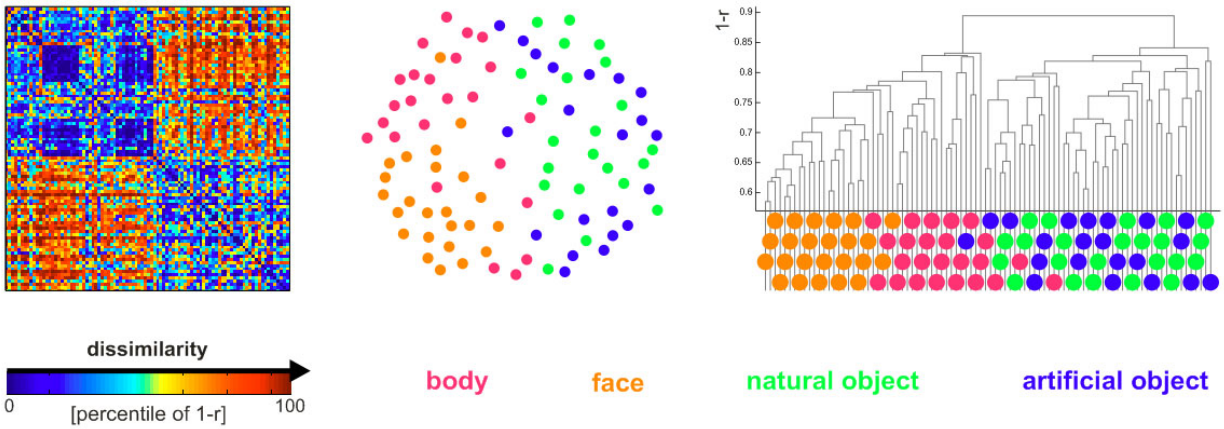


Fig. S11. Representation in human IT without FFA and PPA. Excluding FFA and PPA bilaterally from the voxels selected to define human IT did not qualitatively change the similarity structure or categorical clustering. FFA and PPA were defined in each hemisphere at 1141 mm^3 (150 voxels) and 1521 mm^3 (200 voxels), respectively, by means of an independent block-design localizer experiment. Human IT was defined bilaterally at 316 voxels as for Figs. 1-4, but FFA and PPA were first excluded from the cortex mask in both hemispheres. The dissimilarity matrix (left), multidimensional scaling arrangement (middle), and hierarchical clustering tree (right) were computed with the same parameters as for Figs. 1, 2, 4.

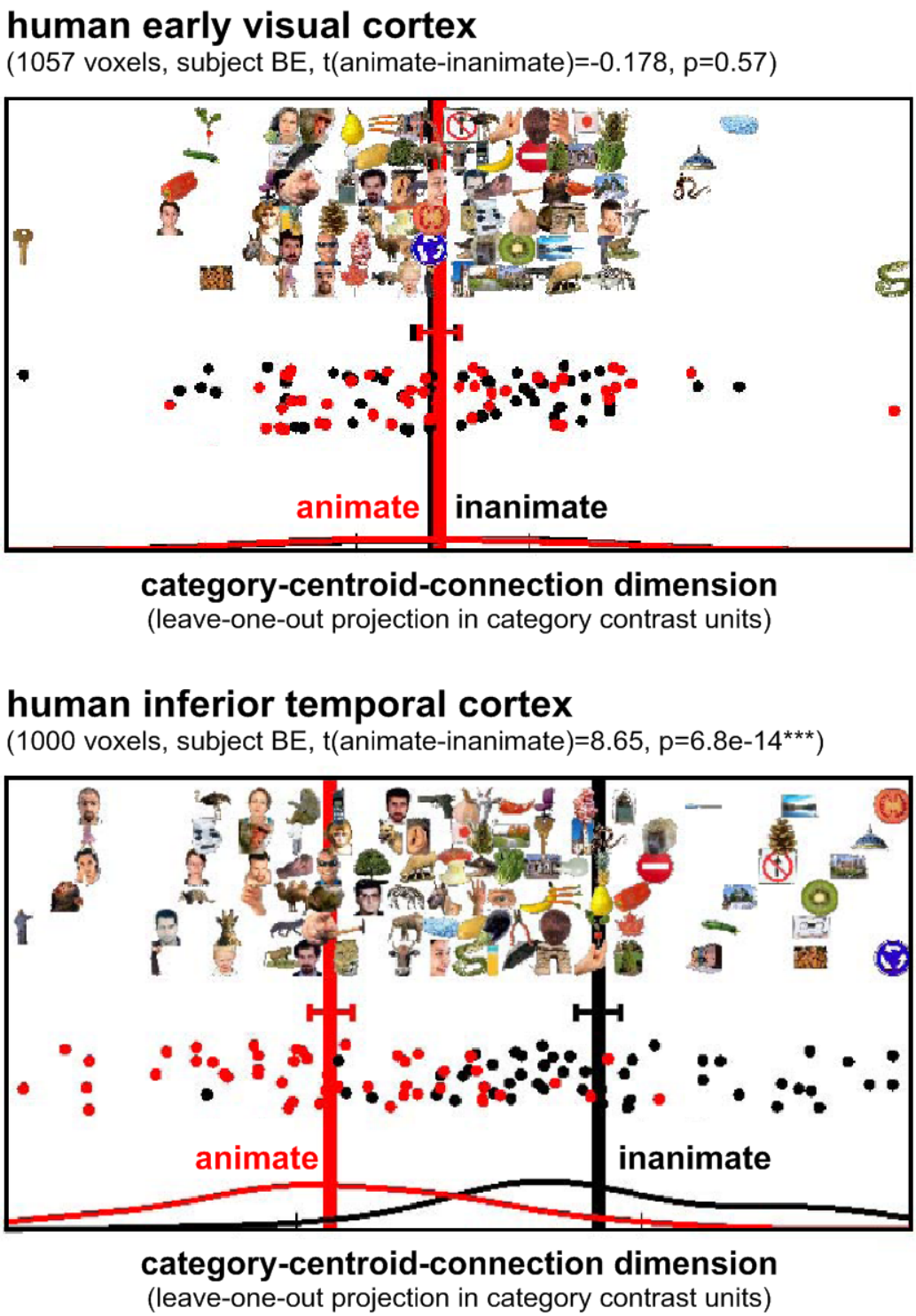


Fig. S12. Response patterns in human IT, but not human early visual cortex, allow linear discrimination between animates and inanimates. Single-subject data show highly significant linear discriminability of animates and inanimates in human IT (t(animate-inanimate)=8.65, p=6.8e-14), whereas linear discriminability of these categories is not evident in human early visual cortex (t(animate-inanimate)=-0.178, p=0.57). The single-image response patterns in early visual cortex (top) and IT (bottom) have been projected onto a dimension defined to discriminate animates from inanimates. The dimension used is the category-centroid-connection dimension (equivalent to the Fisher linear discriminant computed with the assumption of isotropic,

homoscedastic noise). To avoid circularity, the discriminant is computed using a leave-one-out procedure: In order to determine the location of a given single-image response pattern on the discriminant dimension, the other 95 single-image response patterns are used to compute the category centroids defining the discriminant dimension (the thin tick marks indicate the centroid locations defining the discriminant; left for animates, right for inanimates). Note that this approach uses not only independent response estimates, but also different images (the other 95) for defining the discriminant. The thick vertical lines indicate the category means (red for animate, black for inanimate) on the discriminant dimension. Error bars indicate the +/- 1 standard error of the mean. The 96 stimulus images (Fig. S1) have been located along the discriminant (upper portion of each panel) with vertical scattering to allow a larger image size. Colored dots (red for animate, black for inanimate) show the two category distributions (middle portion of each panel), again with vertical random scattering. Probability-density estimates (kernel-smoother method) are shown in the lower portion of each panel (red for animate, black for inanimate).

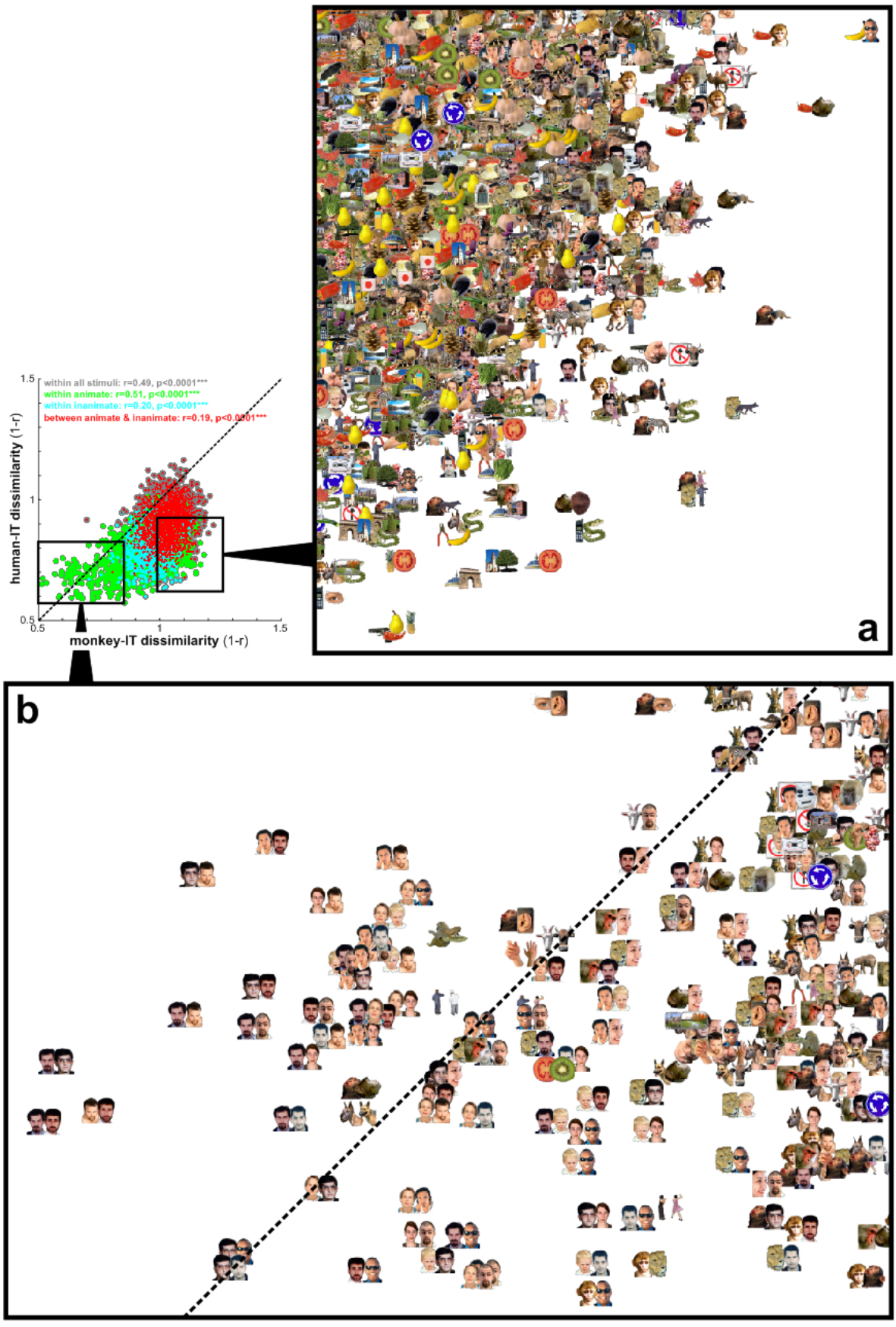

within all stimuli: r=0.49, p<0.0001***
within animate: r=0.51, p<0.0001***
within inanimate: r=0.20, p<0.0001***
between animate & inanimate: r=0.19, p<0.0001***
human-IT dissimilarity (1-r)
monkey-IT dissimilarity (1-r)
0.5
1
1.5
a
b

Fig. S13. Scatterplot of stimulus-pairs relating monkey- and human-IT representations. The scatterplot in Fig. 3a, containing a dot for each stimulus pair, relates monkey- and human-IT dissimilarities. Here we zoom in on two regions of Fig. 3a, in order to make space for plotting each pair of stimuli, whose dissimilarity in monkey and human IT determines the horizontal and vertical location of the corresponding dot in the scatterplot. For each pair, the two stimuli are placed side by side, centered on the location of the dot that represents the pair in Fig. 3a. (**a**) This region contains the stimulus pairs that are furthest from the line of identity. They have the greatest difference between monkey- and human-IT dissimilarity (1 - Pearson r) with the monkey dissimilarity greater than the human dissimilarity. (Note that the corresponding region with greater human- than monkey-IT dissimilarities is unpopulated.) An attractive interpretation would be that these pairs are more similar to humans than to monkeys. Note, however, that the relationship between human and monkey representational dissimilarities may not be linear. Plausible interpretations suggest themselves for many of the pairs, but remain speculative. (**b**) This region contains the stimulus pairs eliciting the most similar activity patterns in both monkey and human IT. This region is dominated by pairs of human faces.

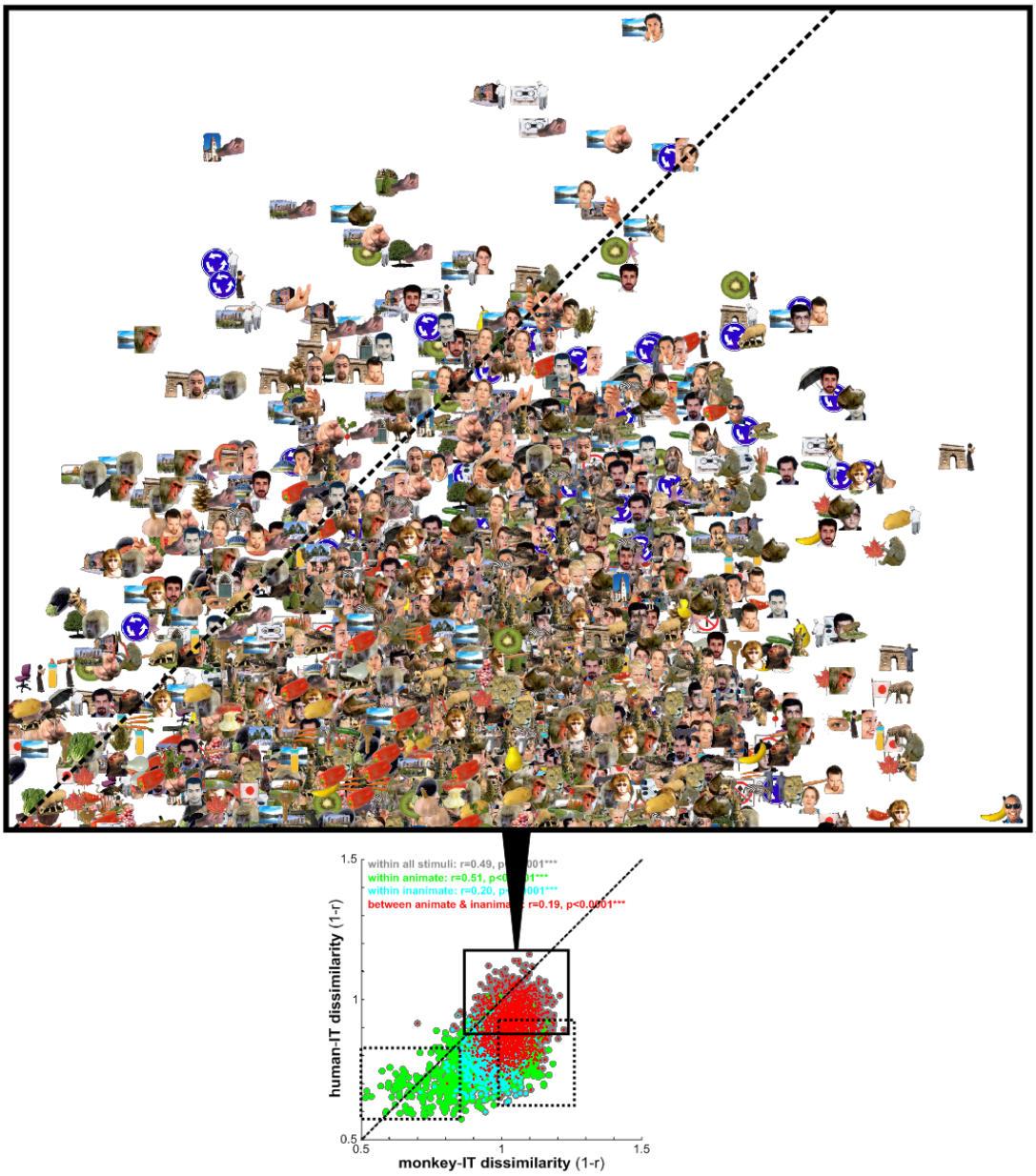


Fig. S14. Scatterplot of stimulus-pairs relating monkey- and human-IT representations (continued). This figure follows the same logic as the previous one, but zooms in on a third region of Fig. 3a. This region contains the stimulus pairs eliciting the most dissimilar activity patterns in both monkey and human IT. This region is dominated by pairs of stimuli crossing the animate-inanimate category boundary. (Dotted rectangles indicate the two regions zoomed in on in the previous figure.)

Supplemental text

Summary of the previous evidence on IT categoricality

Ever since neuropsychology described object-category-related deficits following brain damage (Humphreys and Forde, 2001; Capitani et al., 2003; Martin, 2007), it has been generally accepted that there is some relationship between conventional categories and human-IT representations. Human neuroimaging has investigated category-average responses, showing that IT contains focal regions whose activation is correlated with conventional categories (Puce et al., 1995; Martin et al., 1996; Kanwisher et al., 1997; Aguirre et al., 1998; Epstein and Kanwisher, 1998; Downing et al., 2001; Downing et al., 2006) and that the category can be read out from the IT response pattern with a linear classifier (Haxby et al., 2001; Cox and Savoy, 2003; Carlson et al., 2003). However, these studies investigated responses averaged across many different stimuli within predefined categories. This approach requires the assumption of a particular category structure and therefore cannot address whether the representation is inherently categorical. IT features might respond to natural image fragments that happen to be correlated with categories, without being optimized to distinguish categories.

Monkey studies, as well, have reported IT responses correlated with natural categories (Vogels, 1999; Tsao et al., 2003; Kiani et al., 2005; Hung et al., 2005; Tsao et al., 2006; Afraz et al., 2006) and novel, experimentally defined, categories (Sigala and Logothetis, 2002; Baker et al., 2002; Freedman et al., 2003). However, step-function-like categorical responses as reported for cells in medial temporal (Kreiman et al., 2000; Quiroga et al., 2005), prefrontal (Freedman et al., 2001), and parietal regions (Freedman and Assad, 2006) are not typically observed in either single IT cells (Vogels, 1999; Freedman et al., 2003; Kiani et al., 2007; but see Tsao et al., 2006) or category-sensitive fMRI responses (Haxby et al., 2001), suggesting that IT may have a lesser role in categorization (Freedman et al., 2003). Consistent with this perspective, one influential computational model of primate object recognition (Riesenhuber and Poggio, 2002; Serre et al., 2007) employs categorization training (i.e. supervised learning) to optimize the stage thought to correspond to prefrontal cortex. The model's IT stage is not optimized for categorization.

Representational similarity analysis

Estimation of single-image response patterns in the monkeys. The analyses are based on all cells that could be isolated and for which sufficient data was available across the stimuli. This yielded a total of 674 neurons for both monkeys combined (322 in monkey 1 and 352 in monkey 2). For each stimulus, each neuron's response amplitude was estimated as the average spike rate within a 140-ms window starting 71 ms after stimulus onset (for details, see Kiani et al., 2007).

Estimation of single-image response patterns in the humans. Single-image response patterns were estimated by univariate linear modeling. We concatenated the runs within a session along the temporal dimension. For each voxel, we performed a single univariate linear model fit to obtain a response-amplitude estimate for each of the 96 stimuli. The model included a hemodynamic-response predictor for each of the 96 stimuli. Since each stimulus occurred once in each run, each of the 96 predictors had one hemodynamic response per run and extended across all within-session runs included (i.e. all runs except those used for region-of-interest definition). The predictor time courses were computed using a linear model of the hemodynamic response (Boynton et al., 1996) and assuming an instant-onset rectangular neural response during each condition of visual stimulation. For each run, the design matrix included these stimulus-response predictors along with six head-motion-parameter time courses, a linear-trend predictor, a 6-predictor Fourier basis for nonlinear trends (sines and cosines of up to 3 cycles per run) and a confound-mean predictor. Trends were, thus, modeled by a separate set of predictors for each run. The trend predictors for a particular run had zero entries for all other runs along time. For head-motion models and confound means as well, separate predictors accounted for each run. For each stimulus, we converted the response-amplitude (beta) estimate map into a t map. The resulting t maps (one for each stimulus) were used for representational similarity analysis.

Computation of representational dissimilarity matrices. For each pair of stimuli, the dissimilarity between the associated response patterns is measured as 1 minus the Pearson linear correlation across cells or voxels within a region of interest (0 for perfect correlation, 1

for no correlation, 2 for perfect anticorrelation). The resulting dissimilarities for all pairs of object images are assembled in a representational dissimilarity matrix (RDM; Fig. 1). Each cell of the RDM, thus, compares the response patterns elicited by two stimuli. As a consequence, an RDM is symmetric about a diagonal of zeros.[1]

Testing relatedness of two representational dissimilarity matrices by randomization of condition labels. We use the Pearson correlation coefficient r to assess the relatedness of two RDMs (e.g. Figs. 3, S2). The correlation is restricted to the upper (or equivalently the lower) triangle of each RDM. In order to decide whether two RDMs are related, we perform statistical inference on the RDM correlation. The classical method for testing correlations assumes independent pairs of measurements for the two variables. For RDMs such independence cannot be assumed, because each dissimilarity is dependent on two response patterns, each of which also codetermines the dissimilarities of all its other pairings in the RDM. We therefore test the relatedness of RDMs by randomization (e.g. Nichols and Holmes 2002). In particular, we use randomization of the condition labels to rearrange the rows and columns of an RDM. We choose a random permutation of the conditions (i.e. of the 92 stimuli), reorder rows and columns of one of the two RDMs to be compared according to this permutation, and compute the correlation. By repeating this step 10,000 times, we obtain a distribution of correlations simulating the null hypothesis that the two RDMs are unrelated. If the actual correlation (i.e. the one for consistent labeling between the two RDMs) falls within the top 5% of the simulated null distribution of correlations, we reject the null hypothesis of unrelated RDMs. More generally, we estimate the p value as the percent rank/100 of the actual correlation in the randomization distribution. The percent rank is conservatively estimated, such that $p<0.0001$ indicates that the actual correlation was higher than any of the 10,000 correlations obtained after randomization of the condition labels.

[1] Alternatively, two separate data sets can be used, such that the vertical dimension of the RDM indexes pattern estimates from data set 1 and the horizontal dimension indexes pattern estimates from data set 2. The diagonal then contains dissimilarity estimates for replications of the same condition. Moreover, for each pair of conditions, two entries symmetrical about the diagonal of the RDM contain separate estimates of the pattern dissimilarity. We use the single-pattern-set approach here, because it provides more data along time for each fMRI pattern estimate. The overlapping hemodynamic responses to the stimuli are more precisely estimated when more runs are available not only by a factor of sqrt(n), where n is the number of runs, but by a larger factor, because longer random sequences more closely approximate the ideal of uncorrelated hemodynamic-response predictors.

The condition-label randomization test is justified by the random assignment of conditions (i.e. stimuli) to experimental trials. Under the null hypothesis of no relation between the RDMs, the conditions are exchangeable, i.e. the true labeling and each random relabeling of one RDM yield correlations drawn from the same distribution (i.e. from the null distribution). More generally, condition-label randomization can be used to test various RDM statistics against the null hypothesis that the condition labels are interchangeable (in the sense of not affecting the true test statistic). Note that the central results of this paper all rely on the condition-label randomization test (Figs. 3, 5, 6, S3a), but condition-bootstrap resampling has been used to test additional hypotheses (Figs. 5, S4).

Testing statistics of representational dissimilarity matrices by bootstrap resampling of the set of conditions. Not all hypotheses about RDMs can be tested by randomization. For example, the randomization test cannot be used to assess whether the mean of an RDM is greater than some constant, because the mean will be the same for each relabeling. Moreover, by condition-label randomization we test the null hypothesis that the condition labels are interchangeable. This may not be the desired null hypothesis. A less rigorous, but more versatile approach is bootstrap resampling (Efron and Tibshirani, 1993), which we apply here to the set of experimental conditions (i.e. the stimuli), in order to simulate a distribution of RDMs. Like the randomization test, the bootstrap test is appropriate for RDMs in that it does not rely on either distributional assumptions or the assumption of independence of the dissimilarity estimates. As mentioned above, tests assuming independent data may not be valid for RDMs, because dissimilarities within an RDM have a complex dependency structure. Like the condition-label randomization described above, the bootstrap resampling here operates at the level of the experimental conditions and, thus, simulates the dependency structure of RDMs.

The condition bootstrap test proceeds by resampling the set of conditions with replacement: If there are n_c condition labels, we draw n_c times from the whole set, to obtain a set of n_c labels. The bootstrap set of condition labels may include multiple instances of some labels and exclude others altogether. We construct a bootstrap RDM by resampling the original RDM according to the bootstrap sample of condition labels. We then compute the statistic of interest (e.g. the mean of all dissimilarities). We repeat this process many times (e.g. 10,000 times) to obtain a bootstrap distribution for the statistic.

The bootstrap resampling can be stratified in order to compare sets of conditions. For example, in order to test whether between-category dissimilarities are greater than within-category dissimilarities (Fig. 5), we separately bootstrap resampled the animates set and the inanimates set, recomputing mean B of between-category dissimilarities and the mean W of within-category dissimilarities to obtain the test statistic B minus W.

The bootstrap distribution of the statistic allows us to obtain error bars on arbitrary RDM statistics: the standard deviation of the bootstrap distribution is the standard error of the estimate of the statistic. In addition, we can define a 95% confidence interval by excluding 5% of the extreme values in the distribution (either on one side for a one-sided test or on both sides symmetrically for a two-sided test). If the value assigned to the test statistic by the null hypothesis falls outside the confidence interval, the null hypothesis is rejected. This is a valid test, because the confidence interval will include the null value with 95% probability given that the null hypothesis is true and assuming that the bootstrap resampling is an accurate simulation. The latter assumption is questionable, therefore the bootstrap procedures we describe here should be considered rough, approximate methods of inference.

A further complication of this method is that the bootstrap resampling of the set of condition labels moves zeros from the diagonal into the off-diagonal parts of the RDM whenever a condition is selected multiple times in the bootstrap resampling. (For 96 conditions, this is a small proportion of the entries: on the order of 1%.) In order to prevent these zeros from biasing the statistic, we exclude them before computing the statistic. For the purpose of condition bootstrapping, it may be preferable to use two data sets for computing the RDM (as suggested above), such that each diagonal value reflects a dissimilarity between two replications of the same response pattern.

The rationale for bootstrap resampling of a set of experimental conditions is to simulate the distribution of the statistic of interest that we expect to obtain for repetitions of the experiment performed with the same subjects but with different experimental conditions (e.g. stimuli) drawn from the same population of possible conditions that could have been used for the experiment (e.g. stimuli from the same categories). An interesting feature of this approach is its potential to generalize from the set of conditions actually used in the experiment to the hypothetical population of conditions, of which the actually chosen conditions can be considered a random sample. Note, however, that a sufficient number of conditions is required for this and the accuracy of the simulation is not guaranteed. For caveats and advanced bootstrap methods, see Hesterberg (2007).

Human localizer experiments and definition of regions of interest

Definition of regions of interest. All regions of interest (ROIs) were defined on the basis of independent experimental data and restricted to a cortex mask manually drawn on each subject's fMRI slices. Human IT was defined by selecting a variable number of voxels (316 voxels in Figs. 1-6; 100-10,000 voxels in Fig. S11) within the inferior temporal portion of the bilateral cortex mask according to their visual responsiveness. Visual responsiveness was assessed using the t map for the average response to the 96 object images. The t map was computed on the basis of one third of the runs of the main experiment within each session. The remaining runs were used to perform all further analyses. To define early visual cortex, we selected the most visually responsive voxels, as for IT, but within a manually defined anatomical region around the calcarine sulcus within the cortex mask (Figs. 5, 6, S5). For control analyses (Fig. S11), we defined the FFA (Kanwisher et al., 1997) using the contrast faces minus objects, and the PPA (Epstein and Kanwisher, 1998) using the contrast places minus objects in analyzing the separate localizer block-design experiment described below.

Localizer block-design experiment. We performed a functional localizer experiment using the same fMRI sequence as for the main experiment and a separate set of stimuli. Subjects viewed grayscale photos of faces, places, and objects (spanning a visual angle of about 5.7°) in category blocks. Each block lasted 30 s (stimulus-onset asynchrony: 1 s; stimulus duration: 700 ms), alternating with 20-s fixation blocks. Three blocks were presented for each stimulus category (face, place, object), resulting in a total run duration of 7 min and 50 s. Stimuli were presented on a constantly visible uniform black background. Subjects continually fixated a central white cross and performed a one-back repetition-detection task on the images, responding with a left-thumb button press for each consecutive repetition (3 to 5 repetitions per block). Each stimulus was only presented once, except for the immediate repetitions to be detected in the one-back task. Stimuli were centered with respect to the fixation cross.

Model representations

We processed our stimuli to obtain their representations in a number of low-level models. We then analyzed these model representations (Figs. S6, S7) in the same way as the brain-activity data from early visual cortex and IT (Figs. 1, 2, 4, S5, S9-S11). Each image was converted to a representational vector as described below for each model. As for the brain-activity data, each representational vector was then compared to each other representational vector by means of 1-r as the dissimilarity measure (where r is the Pearson linear correlation; only for the S-CIELAB representation, the conventional Delta E measure was used instead of 1-r) to obtain a representational dissimilarity matrix, on which further analyses (Figs. S6, S7) were based.

Color image (CIELAB). The RGB color images (175×175 pixels) were converted to the CIELAB color space, which approximates a linear representation of human perceptual color space. Each CIELAB image was then converted to a pixel vector (175×175×3 numbers).

Low-resolution color image (28×28 pixels, CIELAB). The RGB color images (175×175 pixels) were downsampled to 28×28 pixels (with bicubic interpolation) and subsequently converted to the CIELAB color space. Each 28×28 CIELAB image was then converted to a pixel vector (28×28×3 numbers).

Grayscale image. The RGB color images (175×175 pixels) were converted to grayscale. Each grayscale image was then converted to a pixel vector (175×175 numbers).

Low-resolution grayscale image (28×28 pixels). The RGB color images (175×175 pixels) were converted to grayscale and subsequently downsampled to 28×28 pixels (with bicubic interpolation). Each grayscale image was then converted to a pixel vector (28×28 numbers).

Binary silhouette image. The RGB color images (175×175 pixels) were converted to binary silhouette images, in which all background pixels had the value 0 and all figure pixels had the value 1. Each binary silhouette image was then converted to a pixel vector (175×175 binary numbers).

CIELAB joint histogram (6×6×6 bins). The RGB color images (175×175 pixels) were converted to the CIELAB color space. The three CIELAB dimensions (L, a, b), were then divided into 6 bins of equal width. The joint CIELAB histogram was computed by counting the number of figure pixels (gray background left out) falling into each of the 6×6×6 bins. The joint histogram was converted to a vector (6×6×6 numbers).

S-CIELAB (Delta E). The RGB color images (175×175 pixels, 2.9° visual angle) were compared (each to each other, to obtain a representational dissimilarity matrix) by means of the S-CIELAB Delta-E dissimilarity measure (Zhang and Wandell, 1997), which models the perceptual similarity of color images and, unlike CIELAB Delta E, accounts for pattern-color sensitivity results (Poirson and Wandell, 1993) by separating the image into components corresponding to different spatial-frequency bands.

V1 model. The RGB color images (175×175 pixels, 2.9° visual angle) were converted to grayscale and given as input to population of modeled V1 simple and complex cells (Lampl et al., 2004; Riesenhuber and Poggio, 2002; Kiani et al., 2007). The receptive fields (RFs) of simple cells were simulated by Gabor filters of 4 different orientations (0°, 90°, -45° and 45°) and 12 sizes (7-29 pixels). Cell RFs were distributed over the stimulus image at 0.017° intervals in a cartesian grid (for each image pixel there was a simple and a complex cell of each selectivity that had its RF centered on that pixel). Negative values in outputs were rectified to zero. The RFs of complex cells were modeled by the MAX operation performed on outputs of neighboring simple cells with similar orientation selectivity. The MAX operation consists in selecting the strongest (maximum) input to determine the output. This renders the output of a complex cell invariant to the precise location of the stimulus feature that drives it. Simple cells were divided into four groups based on their RF size (7-9 pixels, 11-15 pixels, 17-21 pixels, 23-29 pixels) and each complex cell pooled responses of neighboring simple cells in one of these groups. The spatial range of pooling varied across the four groups (4×4, 6×6, 9×9, and 12×12 pixels for the four groups, respectively). This yielded 4 (orientation selectivities) × 12 (RF sizes) = 48 simple-cell maps and 4 (orientation selectivities) × 4 (sets of simple-cell RF sizes pooled) = 16 complex-cell maps of 175×175 pixels. All maps of simple and complex cell outputs were vectorized and concatenated to obtain a representational vector for each stimulus image.

HMAX-C2 model based on natural image fragments. This model representation developed by Serre et al. (2005) builds on the complex-cell outputs of the V1 model described above (implemented by the same group). The C2 features used in the analysis (Fig. S7) may be comparable to those found in primate V4 and posterior IT. The model has four sequential stages: S1-C1-S2-C2. The first two stages correspond to the simple and complex cells described above, respectively. Stages S2 and C2 use the same pooling mechanisms as stages S1 and C1, respectively. Each unit in stage S2 locally pools information from the C1 stage by a linear filter and behaves as a radial basis function, responding most strongly to a particular prototype input pattern. The prototypes correspond to random fragments extracted from a set of natural images (stimuli independent of those used in the present study). S2 outputs are locally pooled by C2 units utilizing the MAX operation for a degree of position and scale tolerance. A detailed description of the model (including the parameter settings and map sizes we used here) can be found in Serre et al. (2005). The model, including the natural image fragments, was downloaded from the author's website in January 2007 (for the current version, see http://cbcl.mit.edu/software-datasets/standardmodel/index.html).

Additional model representations. In addition to the models described above and analyzed for our stimulus set in Figs. S6 and S7, we tried a number of additional models (not shown). These included (1) low-passed and high-passed grayscale representations, (2) a version of the V1 model described above, in which we averaged all simple and complex cell responses representing the same retinal location (averaging also across orientation selectivities and RF sizes) in order to mimic the effect of downsampling by population averaging within fMRI voxels, and (3) higher-level shape-tuned units created within the HMAX model framework (Riesenhuber and Poggio, 2002) as described in Kiani et al. (2007). None of these model representations exhibited categorical clustering.

References

Afraz, S. R., Kiani, R., and Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. Nature *442*(7103), 692-695.

Aguirre, G. K., Zarahn, E., and D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. Neuron *21,* 373-383.

Baker, C. I., Behrmann, M., and Olson, C. R. (2002). Impact of learning on representation of parts and wholes in monkey inferotemporal cortex. Nat Neurosci *5*(11), 1210-6.

Boynton, G. M., Engel, S. A., Glover, G. H., and Heeger, D. J. (1996). Linear systems analysis of functional magnetic resonance imaging in human V1. J. Neurosci. *16*, 4207-4221.

Capitani, E., Laiacona, M., Mahon, B., and Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. Cogn. Neuropsychol. *20*, 213-261.

Carlson, T. A., Schrater, P., and He, S. (2003). Patterns of activity in the categorical representations of objects. J Cogn Neurosci *15*(5), 704-17.

Cox, D. D., and Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage *19*, 261-270.

Downing, P. E., Chan, A. W.-Y., Peelen, M. V., Dodds, C. M., and Kanwisher, N. (2006). Domain specificity in visual cortex. Cereb. Cortex *16*(10), 1453-1461.

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.

Edelman, S., Grill-Spector, K., Kushnir, T., and Malach, R. (1998). Towards direct visualization of the internal shape space by fMRI. Psychobiology *26*, 309-321.

Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. Chapman and Hall.

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

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2003). A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J. Neurosci. *23*(12), 5235–5246.

Freedman, D. J., Riesenhuber, M., Poggio, T., and Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. Science *291*, 312-316.

Freedman, D. J. and Assad, J. A. (2006). Experience-dependent representation of visual categories in parietal cortex. Nature. *443*(7107), 85-88.

Haxby, J. V. et al. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science *293*, 2425-2430.

Hesterberg, T. C. (2007) Bootstrap, http://home.comcast.net/~timhesterberg/articles/tech-encyclopedia.pdf) under review.

Humphreys, G. W., and Forde, E. M. E. (2001). Hierarchies, similarity, and interactivity in object recognition: “Category-specific” neuropsychological deficits. Behav. Brain Sci. *24*, 453-509.

Hung, C. P., Kreiman, G., Poggio, T., and DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

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.

Kiani, R., Esteky, H., and Tanaka, K. (2005). Differences in onset latency of macaque inferotemporal neural responses to primate and non-primate faces. J. Neurophysiol. *94*, 1587-1596.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J. Neurophysiol. *97*, 4296-4309.

Kreiman, G., Koch, C., and Fried, I. (2000). Category-specific visual responses of single neurons in the human medial temporal lobe. Nat. Neurosci. *3*, 946-953.

Lampl, I., Ferster, D., Poggio, T., and Riesenhuber, M. (2004). Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex. J. Neurophysiol. *92*, 2704–2713.

Martin, A. (2007). The representations of object concepts in the brain. Annu. Rev. Psychol. *58*, 25-45.

Martin, A., Wiggs, C. L., Ungerleider, L. G., and Haxby, J. V. (1996). Neural correlates of category-specific knowledge. Nature *379*(6566), 649-52.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp. *15*(1), 1-25.

Poirson, A. B., and Wandell, B. A. (1993). Appearance of colored patterns: pattern-color separability. J. Opt. Soc. Am. A. *10*, 2458-2470.

Prüssmann, K. P. (2004). Parallel imaging at high field strength: Synergies and joint potential. Top Magn. Reson. Imaging *15*, 237-244.

Puce, A., Allison, T., Gore, J. C., and McCarthy, G. (1995). Face-sensitive regions in human extrastriate cortex studied by functional MRI. J. Neurophysiol. *74*, 1192-1199.

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., and Fried I. (2005). Invariant visual representation by single neurons in the human brain. Nature. *435*(7045), 1102-7.

Riesenhuber, M., and Poggio, T. (2002). Neural mechanisms of object recognition. Curr. Opin. Neurobiol. *12*,162-168.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proc. Natl. Acad. Sci. U.S.A. *104*, 6424-6429.

Serre, T., Wolf, L., and Poggio, T. (2005). Object recognition with features inspired by visual cortex. In: Computer Vision and Pattern Recognition (CVPR 2005), San Diego, USA, June 2005.

Sigala, N., and Logothetis, N. K. (2002). Visual categorization shapes feature selectivity in the primate temporal cortex. Nature *415*, 318-320.

Tsao, D. Y., Freiwald, W. A., Knutsen, T. A., Mandeville, J. B., and Tootell R. B. H. (2003). Faces and objects in macaque cerebral cortex. Nat. Neurosci. *6*, 989-995.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., and Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. Science *311*, 670-674.

Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. Eur. J. Neurosci. *11*, 1239-1255.

Zhang, X., and Wandell, B. A. (1997). A spatial extension of CIELAB for digital color image reproduction. SID Journal.