Food for thought: selectivity for food in human ventral visual cortex

Nidhi Jain[1], Aria Wang[5,6], Margaret M. Henderson[5,6], Ruogu Lin[3], Jacob S. Prince[2,4], Michael J. Tarr[2,5], and Leila Wehbe[*,5,6]

[1]Computer Science Department, Carnegie Mellon University
[2]Department of Psychology, Carnegie Mellon University
[3]Computational Biology Department, Carnegie Mellon University
[4]Department of Psychology, Harvard University
[5]Neuroscience Institute, Carnegie Mellon University
[6]Machine Learning Department, Carnegie Mellon University

[*]To whom correspondence should be addressed. E-mail: lwehbe@cmu.edu

Abstract

Ventral visual cortex contains regions of selectivity for domains of ecological importance. Food is an ecologically and evolutionarily important category whose high degree of visual variability may make the identification of selectivity more challenging. First, we investigated neural responsiveness to food using natural images combined with large-scale human fMRI. Leveraging the improved sensitivity of modern designs and statistical analysis methods, we identify two food-selective regions in the ventral visual cortex. Our results were robust across 8 subjects from the Natural Scenes Dataset (NSD), multiple independent sets of images and multiple analysis methods. Second, we tested our findings regarding visual food selectivity by designing and running an fMRI "localizer" experiment that included grayscale food images. Our independent localizer results confirm the existence of food selectivity in human ventral visual cortex and help illuminate why earlier studies may have failed to do so. The identification of food-selective regions stands alongside prior findings of functional selectivity and provides an important addition to our understanding of the organization of knowledge within the human visual system.

Keywords: food, visual selectivity, fMRI,natural images, objects in context, ventral visual cortex, single-subject analysis

Introduction

The representation of high-level visual information in the human brain has been marked by the phenomenon of selectivity for visual categories or properties of high ecological importance. Focusing on ventral visual cortex, there are multiple brain regions that show preferential responses to categories such as faces[1,2], bodies[3], places[4], and words[5], and to broad organizational principles such as animacy[6], real-world size[6], and "reach space"[7]. Independent of any particular theory on the origins and specificity of these functional brain regions[8,9], the prevailing view is that the likely role of these regions is to instantiate processes and representations for categories and properties that are highly relevant for common and important day-to-day behaviors. In a similar vein, food is a category that is relevant to evolution – the need to find nourishment is more ancient than social interaction and, arguably, more fundamental to survival. It is therefore surprising that food has not been consistently identified as a visual category for which localized, selective neural responses are observed.

The visual presentation of food images is known to prompt a range of brain responses[10–13], including affective, sensory, and cognitive effects. However, agreement on neuroanatomical locations of food-related activation across studies using food images has been low to moderate[13]. In one meta analysis of relevant studies, only 41% of 17 experiments contributed to food-related clusters in the bilateral fusiform gyrus and left orbitofrontal cortex[13]. Another study of selectivity across a range of proposed categories found no robust selectivity for either fruits or vegetables in occipitotemporal cortex[14]. In the cases where statistically significant responses to food have been observed, they have typically been attributed to increased attention to food images arising from subjects' mental states and/or physiological factors[11,13,15] rather than to visual category representations *per se*. For example, supporting the idea that it is the value of particular foods that drives responses, Huerta and colleagues[11] performed a meta analysis across 11 studies specifically focused on eating behavior, where they compared high caloric food pictures (e.g., hamburgers, cake, waffles, fries, etc.) to non-food pictures (e.g., rocks, bricks, trees, houses, etc.) and found the most consistent group-average activation in the right fusiform gyrus[11]. Additionally, in the study most relevant to our present work, Adamson and Troiani[16] considered the connection between a subject's body mass index (BMI) and neural responses to food in a paradigm that compared 80 food images to an equal number of faces, places, and clocks. Interestingly, independent of any interaction with BMI, they found evidence for left-lateralized food selectivity, overlapping with the fusiform face area (FFA), and interpreted this as an indication that fusiform activation may be driven by motivation and valence factors that are common to both food and faces. This earlier finding of food selectivity in the FFA was further interpreted as a counter-example to the theory that FFA selectivity is a consequence of "expertise" – high proficiency at individuating exemplars within a visually-similar category[9] (in that food images are relatively dissimilar from one another). However, their conclusion was based primarily on group average responses and focused on establishing overlap between food selectivity and the FFA, rather than parsing the fine-grained anatomical relationship between food- and face-selective populations. Thus, while it is known that food images elicit neural responses in a variety of brain regions, including the fusiform gyrus, it is not yet clear whether selectivity for food images is instantiated as a distinct category-selective region within ventral visual cortex.

Using more sensitive experimental designs and statistical methods across two experiments, we were able to spatially localize food selective regions at a much more fine-grained level within individuals, thereby providing a strong test of the relationship between food-selective and other category-selective regions. Notably, two other studies[17,18] based on the same Natural Scenes Dataset (NSD)[19] we used in Experiment 1, both identify distinct food-selective regions consistent with our results (although relying on somewhat different analysis methods). We will return to these studies in the Discussion. Both our study and these two similar studies depart somewhat from most prior investigations of food-related neural responses in that, in contrast to past studies, we do not include any physiological variables (e.g., BMI or hunger level) as covariates in our analyses, and we do not restrict our image set to high-calorie, appetizing stimuli. Rather, our study explicitly aimed to identify the brain regions that represent and process the visual properties of food in a more general context; that is, without explicitly or implicitly attempting to recruit circuits involved in reward, motivation, or valence.

A variety of factors may have impacted the results (or lack of results) in many prior studies investigating food-selectivity in visual cortex (e.g., P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster). One possibility is that some of the apparent inconsistency in detecting food-selective responses is, in part, due to relying on isolated, somewhat unrealistic food and non-food images (e.g., Downing et al.[14]). However, as we discuss below, our Experiment 2 identifies food selectivity using grayscale images of food. As such, while the naturalness of the COCO images used in NSD may enhance food-selective responses, it seems unlikely that naturalness alone (nor the absence of color) can account for prior failures. At the same time, it is worth noting that both Adamson and Troiani[16] and Tsourides et al.[20] used naturalistic food images and were able to successfully identify food-related neural responses as measured by functional MRI (fMRI) and magnetoencephalography (MEG), respectively.

A second factor contributing to earlier null results may be that prior studies used an insufficient number of food images, thereby failing to capture the large variety of visual properties of food or of the natural contexts in which food appears. Unlike faces, bodies, or word stimuli, food images vary widely in low- to mid-level visual characteristics such as curvature, shape, texture, color or the organization of the parts into a whole. Thus, greater numbers of food stimuli not only increase experimental power in and of itself, but lead to better coverage of “food appearance space” as it may be mentally and neurally represented.

A third factor which may have made identifying food-selective regions more challenging is potential variability across individuals in the neural localization of food-related responses – a prediction supported by the individual variability seen in the results of our Experiment 2 (and in [16]). One possible reason for this variability may be that voxels processing food are interleaved with voxels processing other object related properties.[18] Another possible reason for individual variability is the complexity of building a food processing area due to the visual heterogeneity of food. Indeed, this latter point has been raised as one reason why a food-selective visual region seemed unlikely – in contrast to visually-homogeneous categories such as faces and written words, foods vary dramatically in shape, texture, and color. Consequently, it is unclear how a single visual mechanism might learn across this appearance diversity. One possible solution has been articulated in modern machine learning where building a classifier for a complex class comprised of multiple sub-classes with an inherent organization (such as food) is a problem referred to as hierarchical classification[21,22]. One way of the common ways this problem is solved is by combining the predictions of specialized classifiers for each of the different sub-classes into a single prediction[23]. Thus, one can conceive of food-selective responses as a set of specialized classifiers for different food sub-types. Given the complexity of such representations, as well as their potential interactions with culture, taste,[16] and experience,[11] spatial variability in food-selective responses would not be surprising.

Fourth, as just discussed, the visual heterogeneity of food may lead to food-selective regions that are more spatially distributed as compared to other category-selectivity responses, possibly including multiple sub-regions. However, to boost statistical power, standard neuroimaging analyses often forgo individual-level statistical tests in favor of across-subject tests that are biased to "blur" localized responses.[16,24] These analyses are more likely to identify regions that are well aligned across subjects[25]. In contrast, in both of our experiments, we rely on within-subject analyses that are better able to pull out category-specific neural responses for individual brains. Moreover, in Experiment 1, our fine-grained analyses reveal that the top images in food regions overwhelmingly depict food. Thus, while it is possible that the neural representations of other categories are intermingled with the representation of food, our results favor distinct, but perhaps distributed, food-selective regions within ventral visual cortex. As a coda to reliance on more sensitive data analysis tools, we also note that modern fMRI measurements are much improved over earlier experiments. For example, Experiment 1 used NSD which was collected using a 7T scanner and high resolution temporal and spatial sampling, while Experiment 2 used a state-of-the-art 3T scanner and 64 channel head coil.

Our study addresses these issues across two experiments relying on very different designs. Experiment 1 uses a large-scale, "hypothesis-free" approach in which fMRI data was collected at a massive scale as part of NSD[19], thereby improving our ability to detect effects across *post-hoc* defined conditions. Real-world images, drawn from the the Microsoft COCO dataset[26], were used for both the food and non-food conditions. To preview our most important result, we reliably identify two distinct regions in ventral high-level visual cortex that are preferentially responsive to food images. These two strips surround the Fusiform Face Area (FFA) and are aligned on the anterior to posterior axis. We replicate these regions across subjects while controlling for other aspects of images that are thought to be coded in the ventral visual system, such as image perspective. We also provide exploratory analyses that probe the more fine-grained structure of conceptual representations within food-selective cortex.

Experiment 2 validates the finding of food-selective regions in a hypothesis-driven manner by collecting new fMRI data. We designed a visual "food localizer" by adding a food condition to the existing fLoc localizer by Stigliani et al.[27] As in the other conditions of the fLoc localizer, we composited grayscale food images on scrambled backgrounds. Our analysis identified food-selective regions in each subject, with the location being consistently adjacent to the FFA. The results of Experiment 2 provide direct evidence supporting the hypothesis that food-selective regions in the ventral visual system represent a new domain of category selectivity similar to faces, places, bodies, and words. Our results also directly exclude color and image context from being the major drivers of the visual responses to food observed in Experiment 1. Of particular note, the localization of the food region was consistent across individuals when defined according to a functional landmark (e.g., proximity to the FFA), but when averaging spatially across individuals (e.g., when their brains were aligned), the neuroanatomical overlap of the food region across subjects was less pronounced than other

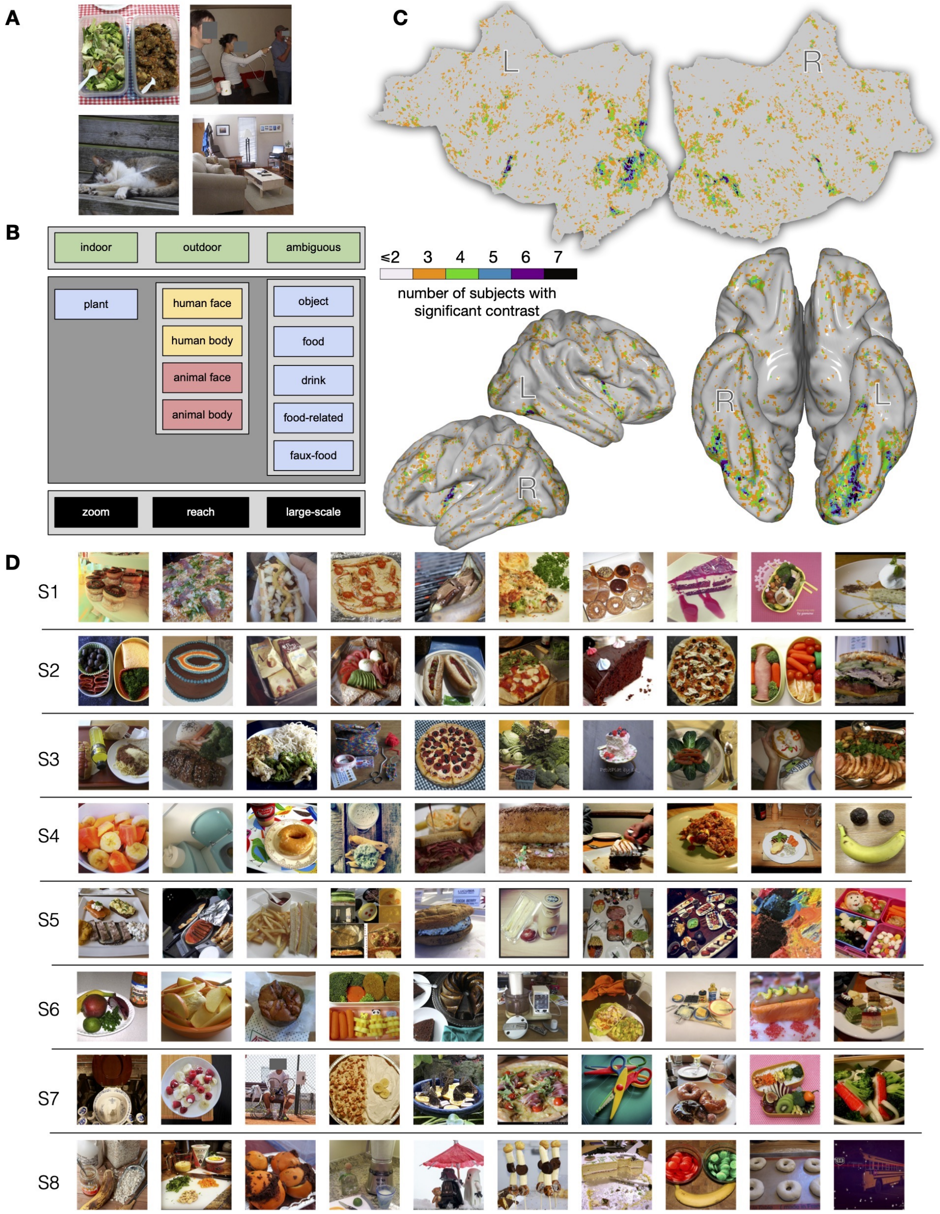


Figure 1: **Experiment 1. The images that could have been potentially viewed by all 8 subjects in NSD** were manually relabeled to investigate responsiveness to naturalistic food images. (A) Example images labeled as (clockwise, from upper left): {outdoor, food, food-related, reach} {indoor, human face, human body, object, large-scale}, {indoor, object, large-scale}, {outdoor, animal face, animal body, object, zoom}. (B) The labeling taxonomy, including attributes of location (top), content (middle), and image perspective (bottom). (C) Flattened, semi-inflated lateral, and semi-inflated bottom views of the MNI surface indicating voxels with higher activity for food than all non-food labels for the shared images. The subject count for a significant contrast was obtained at each MNI voxel. Voxels more responsive to food are found in the frontal, insular, and dorsal visual cortex, with the highest concentration across subjects occurring in the fusiform visual cortex. Both hemispheres show two strips within the fusiform that are separated by a gap that lies on the posterior-to-anterior axis. (D) Top 10 images per subject (S1-S8) leading to the largest responses in the food area. These images, which overwhelmingly depict food, were unique for each subject and were not in the set used to localize the food-selective region.

functional ROIs (and closely replicated the results reported in,[16] thereby accounting for the differences between the results of Experiment 1 and earlier studies). This leads us to consider the third and fourth factors – the spatial heterogenity of food-selective regions and the impact such heterogenity has on traditional localizer designs – as the leading causes for the elusiveness of food selectivity. We release the food localizer code and stimuli as part of this paper.

Naturalistic and hypothesis-driven experimental approaches can be used in a complementary manner that leverages their unique strengths. Here we were able to identify and validate a food-selective region of the human ventral cortex using a naturalistic experiment with complex stimuli to formulate our hypothesis and then use a hypothesis-driven experiment to test that hypothesis. We believe such a combination is a valuable tool in neuroscience that can help advanced the field in the coming years.

From a theoretical standpoint, in that food is incontrovertibly an ecologically critical category, our finding of a food-selective region (confirmed in[17,18]) is consistent with earlier findings of selectivity in the perception of faces, bodies, places, and words. Building on this result, principal component analyses across food-selective voxels provides a finer-grained view into the rich organization of food-relevant information within visual cortex, possibly reflecting gradients along which food is combined with other ecologically relevant categories.

Results

Experiment 1: Large-scale analyses of food representations in a naturalistic setting

To investigate responsiveness to food in a large-scale natural setting, we used the Natural Scenes Dataset (NSD)[19], which consists of high-resolution fMRI responses to naturalistic scenes. NSD contains fMRI data from 8 screened subjects (S1-S8) who each viewed 9,000-10,000 scene images.Of the 70,566 total unique images viewed across subjects, for purposes of consistency we focused on the 1000 images that were shared among subjects (see *Methods* for more details) .

Though COCO images already include labels for many categories, including some types of food, there is important information not captured by these labels, such as whether an image contains human faces. We methodologically relabeled by hand the 1,000 images shared across subjects, based on 3 main attributes: location, content, and image perspective. We used the hierarchical structure shown in Fig. 1B (refer to *Methods* for labeling details, and Fig. 1A for examples). Image perspective was included because there is evidence that objects shown at human-reachable distances have a distinct representational signature in the brain[7,28] and food is often viewed at reachable distances.

Using these labeled images, we constructed a standard linear model that expresses brain activity as a combination of the attributes assigned to each image. This model identified voxels that are more responsive to food than other categories, based on a t-test comparing the weights for food versus all other labels (Fig. 1C). Across the cortex, there are several regions showing significantly higher activation for food than non-food categories ($p <$0.05, false discovery rate (FDR) corrected), including some areas in parietal and frontal cortex, as well as on the ventral surface of the occipital lobe. We focus on ventral visual cortex due to the long history of mapping category-selective responses in this brain region. Across all subjects, we consistently find two food-selective strips in the ventral visual cortex that surround the FFA on the lateral and medial sides. (Fig. 1C shows the count of subjects for whom these contrasts are significant at each MNI voxel (montreal neurological institute coordinate system), and the contrast strength is shown for individual subjects in Figs. 2A and S1A). Note that these identified regions persist even when removing all images with the "reach" (Fig. S3) or "zoom" (Fig. S4) annotations – demonstrating that food-selective responses are not dependent on food being shown at a particular distance[7].

Since this paper focuses on visual food selectivity, we isolated fusiform food-selective voxels using a mask of the ventral visual cortex based on corresponding ROIs from the HCP atlas[29] (see *Methods*). The resulting "food relevant" voxel masks, which were used for the following analyses, are shown in Figures 2B and S1B. We then look at which images maximize the activity in those areas, using a completely separate dataset (the non-shared NSD images). Considering only unique images that were viewed by a given subject, Figure 1D shows the top 10 activating images for the food-selective voxels for that subject. These images overwhelmingly depict food. These images were not used to identify the food regions, and thus reinforce the generality of food selectivity across independent image sets.

Given that food-selective regions appear adjacent to the FFA, we focused on the spatial relationship between food-selective and face-selective populations on the ventral surface. We compared the t-statistics for a contrast of food vs. non-food and t-statistics from a contrast of faces vs. non-faces for S1-S8 individually (Figs. 2A and S1A). The faces vs. non-faces contrast reveals a voxel cluster overlapping with the FFA[1,2] (Figs. 2A and S1A). The FFA was localized for each subject through a separate visual category localizer experiment. (The faces vs. non-faces comparison also makes the methodological point that established category-selective regions can be reliably localized in a large-scale event-related design using stimuli embedded in complex, real-world scenes. This generalizes findings from typical localizer designs and decontextualized images[30]). The regions with higher activity for food are spatially distinct from the ones with higher activity for faces. This pattern persists when comparing food or faces to non-face and non-food images only (Fig. S5), indicating that the regions that have high activity for food and faces have highly independent or non-overlapping spatial extents.

We further investigated how food representations might be distributed across multiple voxels, using searchlight classification[31] (Fig.S2). Training a decoder to classify food versus other categories revealed that food was decodable across a wide area of the ventral surface. The regions from which food information was decodable are a union of the regions that are high for food vs. all and the regions that are high for faces vs. all. This finding is consistent with the idea that voxels primarily selective for other categories, such as faces, may contain information that distinguishes food from other categories[32].

We have focused on identifying food-selective regions through responses to the shared images and our hand-labeled annotations. For the approximately 9,000 remaining images per subject that were not manually labeled, we can still take advantage of COCO annotations[26] (including specific types of food) to further investigate brain responses to food and validate our findings on an independent set of images. We built an encoding model using the 80 object labels provided by COCO and obtained the resulting voxel-wise weights for food labels. We

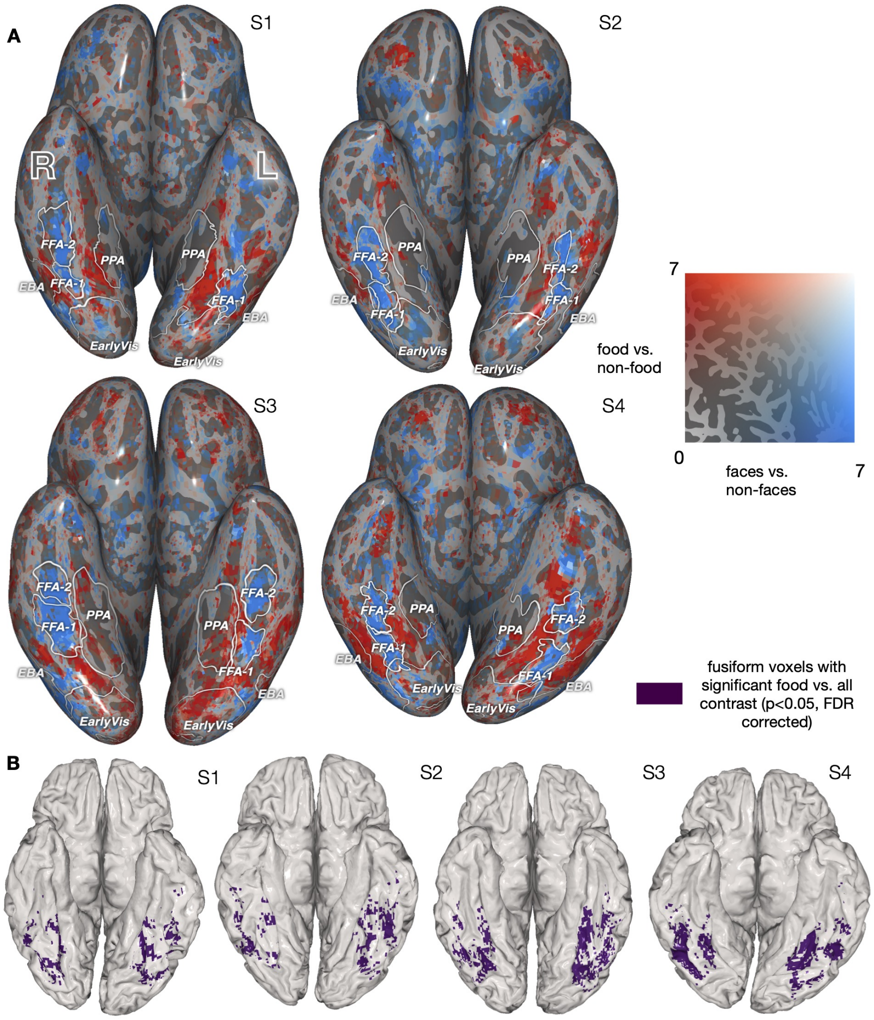


Figure 2: **Experiment 1. Food-selective regions at the individual subject level.** (A) Comparing the spatial localization of food- and face-selective neural populations on the ventral surface, for S1-S4 (see Fig. S1 for S5-S8). Voxels' t-statistics from two 1-sided t-tests comparing food vs. non-food (red) and face vs. non-face (blue). The regions identified by each contrast are largely non-overlapping. This pattern is maintained for food vs. non-(food and face) and face vs. non-(face and food) (Fig. S5). (B) Spatial mask for food-selective regions used in subsequent analyses for S1-S4 (highlighting ventral visual responses). The mask is the overlap between the region that is identified from the t-test for food vs. non-food (panel A, red) at p <0.05 (FDR corrected) and relevant neuroanatomically localized regions using the HCP atlas[29] (see *Methods*).

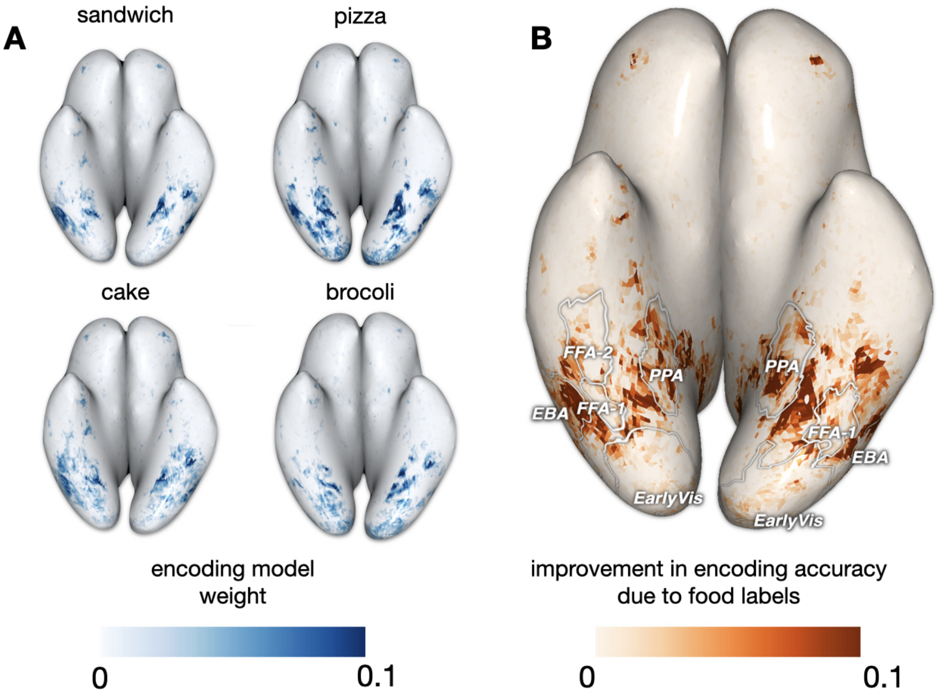


Figure 3: **Experiment 1. A consistent set of food-selective regions can be identified across independent image sets with different labeling schemes.** We used the set of images for each subject that were not included in previous analyses, and an encoding model built from the 80 COCO object labels. (A) Voxel-wise encoding model weights for four food sub-categories from the original COCO dataset, shown for S1. We see variability in the weights, such as (perhaps, not surprisingly) pizza yielding higher weights in some areas than broccoli. (B) We compared predictive accuracy of an encoding model with all COCO labels (including 13 food and 67 non-food labels) to an encoding model with only the 67 non-food COCO labels. On S1's native surface, there is an improvement in validation set R^2 values when including the food labels (R^2 for the full model; R^2 for the model with food removed), with S1-S8 results in Fig. S6. Weights corresponding to individual food labels (A) and the pattern of improvement in R^2 (B) highlight similar food-selective regions. Such consistent results lend further support for these regions being robustly food selective.

find that the voxels having the highest weights for several individual food sub-categories (i.e., *cake, sandwich, pizza, and broccoli*) fall within previously identified food-selective regions (weights for S1 in Fig. 3A). Next, we investigated the specific contribution of food images to these voxel responses by comparing two encoding models: one including the 67 non-food COCO labels, and the other including both food and non-food labels. We compared the R^2 values of the two models on held-out data (Fig. 3B and Fig. S6). Many voxels on the ventral surface show improved prediction performance due to the inclusion of food labels, suggesting that modeling the presence of food beyond other categories was required to accurately predict the voxel responses. These voxels are distributed in roughly the same spatial pattern as the voxels with high-valued weights for individual food categories and our previously identified food regions, further supporting the generality of our results.

To understand the representational structure of these regions, we ran a principal components analysis (PCA) on the responses from all subjects to the shared food images. The PCA

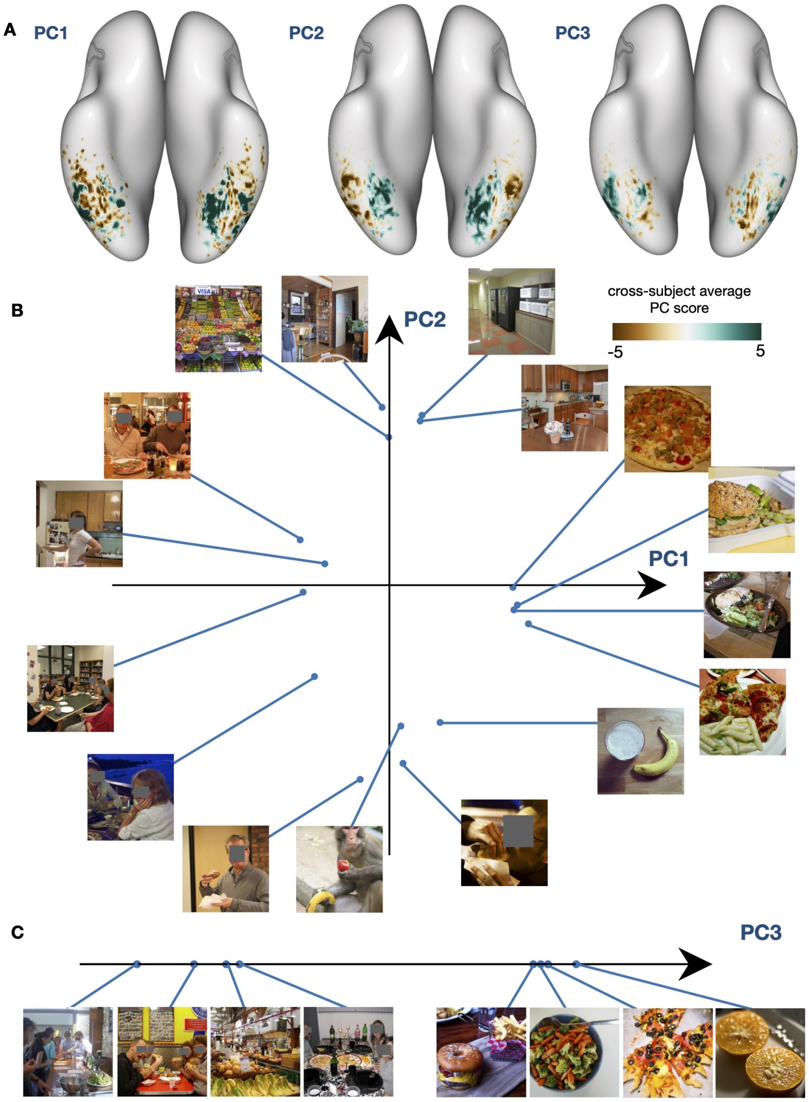


Figure 4: **Experiment 1. PCA of responses from food-selective regions provides insight into their functional structure.** (A) Average principal component score across subjects for PC1, PC2, and PC3, shown on the MNI surface. Blue-green indicates high, brown indicates low PC scores. These top three PCs explain, respectively, 34.31%, 12.68%, and 11.16% of the variance. In (B) and (C), we show the images that lead to the highest and lowest activations in each PC. We include the 4 top and bottom images for ease of visualization. Top images for PC1 and PC2 are plotted in a 2D space (B), with the points connected to each image indicating its position in the space. In (C), we plot the top and bottom images for PC3 along a linear axis. Several patterns emerge here: PC1 scores yield small positive patches around the center of each food-preferring strip with more negative values close to the edges of each strip, and may capture the prominence of food in an image, separating images with focus on food in the foreground from those with food in the background. PC2 scores are higher medially (closer to PPA) and lower laterally, and seem to distinguish large-scale images of food-related places from close-by images of food and people eating food. PC3 scores in the right hemisphere food regions are lower at the center of the two strips, in the areas that border the FFA, while the left hemisphere does not show a clear pattern. PC3 appears to distinguish non-social food settings from social food settings. These results highlight that the combination of food with other ecologically important categories, including people (both faces and bodies) and places, creates a richer co-organization that reveals itself as gradients across cortex.

produces for each voxel a set of principal component scores that capture the projection of its high-dimensional response profile across all images onto a lower dimensional subspace. The axes of this subspace – shared semantic axes – should correspond to the dimensions in food image space that are most strongly reflected in the voxel responses (Fig. 4A). In Figure 4B and C, we visualize the top and bottom images for each PC. The first three PCs are each associated with distinct groups of voxels. PC1 is characterized by small positive patches around the center of each food-preferring strip on the ventral surface, with more negative values close to the edges of each strip. Negative and positive scores for PC2 differentiate the lateral and medial strips of the food-selective region. PC3 scores are generally more spatially diffuse, but in the right hemisphere, PC3 scores are more negative near the FFA (i.e., medial side of the lateral strip, lateral side of the medial strip). Based on inspection of the top and bottom images associated with each PC, PC1 captures the prominence of food in an image, distinguishing images with food as a key focus in the foreground versus those with food as a background element. PC2 distinguishes food images based on overall scale, differentiating close-up images that focus on a few food objects from larger-scale images of food-related scenes (Fig. 4B). This is consistent with the pattern of positive scores for this PC on the medial side of the food-selective area, close to the PPA. PC3 distinguishes food images based on social attributes, separating food images that include few people from images of multiple people eating or preparing food, with social settings being at the end of the spectrum (Fig. 4C). Some amount of person or animacy-related information also appears to be reflected in the first two PCs (top right vs. bottom left images in Fig. 4B). Such results highlight the ecological importance of food as a category, as well as how high-level knowledge structures arise from the interaction between food and other ecologically important categories within the ventral visual cortex.

To further explore what features drive the brain organization of food representations, we clustered food images according to their voxel responses in our food-selective regions. This analysis produces image clusters that are not easily characterized in terms of visual features, viewpoints or semantic attributes (Fig. S7A). We also constructed image clusters using two neural-network models – CLIP[33] and ResNet-18[34] – from which we derived semantic and visual embeddings that did not include the associated brain activity for the images. CLIP is trained on both images and text captions, enabling us to extract features that capture the high-level semantics of the images. ResNet-18, trained solely on images and their associated object labels, yields features with less emphasis on scene semantics. As shown in Figure S7, the clusters arising from CLIP capture semantic classes of food (e.g., fruits, deserts or meals; Fig. S7B) while the clusters arising from ResNet-18 appear more visually organized and more focused on individual elements (e.g., broccoli, pizza; Fig. S7C). Comparing the similarity of the cluster assignments of images for each of the three clustering procedures, neither CLIP or ResNet-18 clusters show any clear correspondence with our voxel-based clusters. The lack of correspondence in our clustering results suggests that the responses in food-selective areas do not organize easily into clusters based on scene semantics or object semantics.

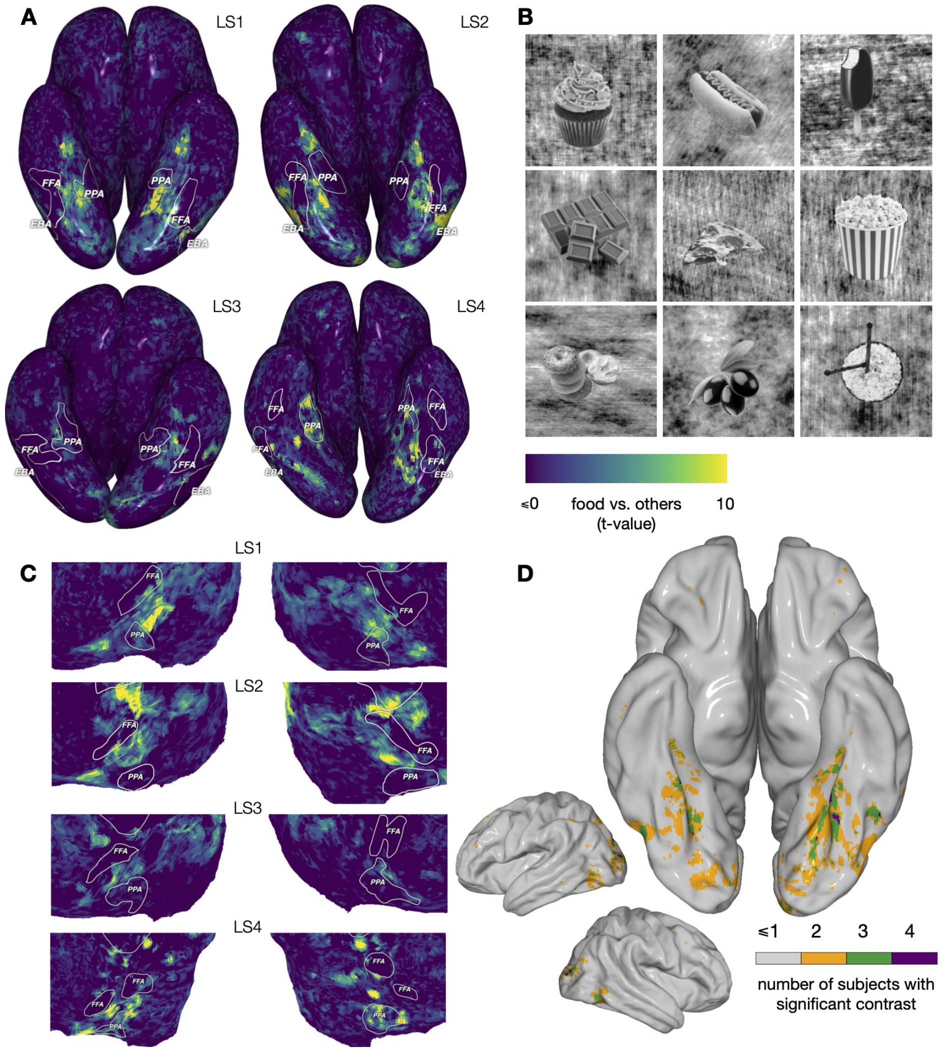


Figure 5: **Experiment 2. Food-selective regions identified in an independent set of subjects using a visual localizer that includes grayscale images.** The fLoc localizer by Stigliani et al.[27] was adapted to include a food condition that was constructed by identifying images of food items from different categories and with different shapes, converting them to grayscale and superposing them on the scrambled images from the fLoc localizer (see *Methods*). Other conditions included faces, bodies, places and written words. (A) *t*-value of the food vs. other contrast shown on the cortical surface (viewed from the bottom) of each localizer subject (LS1-LS4). For each subject, the PPA, FFA and EBA was traced using the corresponding conditions in the localizer. Food-selective regions with a high value for the food vs. other contrast sit between the FFA and PPA of different subjects, with some subjects having high values on both sides of the FFA. See Suppl. Fig. S8 for the significance thresholds. (B) Examples of the stimulus images used in the food condition. (C) A cut-out of the flattened brain of each subject providing a different view of the food regions. There exists some spatial variability between subjects, but the relationship between the ROIs is more stable. (D) Semi-inflated lateral and semi-inflated bottom views of the MNI surface indicating voxels the subject count for a significant food vs. all contrast. Voxels more responsive to food are found in the dorsal visual cortex, with the highest concentration across subjects occurring in the fusiform visual cortex. This result replicates our initial finding with NSD (compare with Fig. 1C). As predicted, the location of the food region is spatially variable across subjects (see Suppl. Fig. S9 to compare with the variability of other classical localizers).

Experiment 2: Hypothesis-driven analyses of food selectivity with controlled stimuli

Our analyses using the NSD dataset allowed us to form a strong hypothesis on the presence of food-selective areas within the fusiform gyrus neighboring the FFA. Next, we designed a standard food "localizer" and collected new fMRI data to test whether we could replicate our results in a controlled experiment. We selected 82 images of different types of food with transparent backgrounds from the `https://www.stickpng.com/` website. We converted the images to grayscale and superimposed them on images from the scrambled condition in the fLoc localizer[27] (Fig. 5B illustrates some examples). All images are shared in the supplementary materials folder. We included four additional conditions from the fLoc localizers: face (adults), body, place (houses) and words. We used the face vs. others, body vs. others and place vs. others contrasts to trace the FFA, EBA and PPA of each subject. In Figure 5, we show the contrast of food vs. others for each subject on both their inflated and flatted surface (see Suppl. Fig. S8 for maps that include statistical significance).

The results of the localizer replicate the findings of Experiment 1 (and[17,18]) and support the hypothesis that food-selective regions fall within the fusiform gyrus adjacent to the FFA. Our new results also indicate that factors such as color or food appearing in a natural scene are not essential for obtaining selective activation for food. While some spatial variability in food-selectivity exists across subjects, regions with high values for the food vs. other contrast lie between the FFA and PPA of different subjects, with some subjects having high value voxels on both sides of the FFA. After converting the subjects' results to MNI space and counting the number of significant voxels in each MNI location, we see less spatial agreement among subjects in the food vs. other contrast as compared to the face, body, and place contrasts (face vs. all, body vs. all, place vs. all, and words vs. all; see Suppl. Fig. S9). More specifically, for each of these other contrasts there exists a region in which all subjects show a significant effect. However, for the food contrast, we find greater spatial variability: at most 3 subjects have a significant contrast in the same region of the left fusiform, and only a small number of voxels show a significant contrast across all subjects. This result is aligned with our findings using NSD in Experiment 1, where the most consistent region is one in which only 5-6 of the 8 subjects showed a significant contrast (Fig. 1). As discussed above, such spatial variability may be one important reason why earlier studies – particularly those relying on group analyses – may have failed to identify regions selective for food.

Discussion

How are knowledge representations organized in the human brain? Within the visual system, one of the hallmarks of the past several decades has been *category selectivity* for faces, bodies, places, and words[1–5]. Consistent with the ecological importance of these categories, we identified selectivity for another ecologically relevant category, food, within the ventral visual stream. In our present study, we used both data- and hypothesis-driven fMRI methods. Two related studies[17,18] also used data-driven methods applied to the same large-scale natural scenes dataset [19] and confirmed our finding of food selectivity in Experiment 1. Our study also provides a range of analyses not included in these other studies, as well as new and informative data from a second, hypothesis-driven experiment. First, using NSD, we show that the identified food regions are maximally activated by food images. Second, we establish that food-selective responses do not appear to be confounded with image viewpoint (zoom, reach or large-scale). Third, we find that the inclusion of food-related category features in an encoding model leads to improved prediction accuracy in food-selective regions. Fourth, we demonstrate that PCA can be used to uncover both the large-scale topography within food-selective areas and the interaction of food coding with other semantic dimensions. Namely, we find that the representation of food in the food regions appears to be organized in gradients across cortex that relate food to other important information processed nearby (social and place-related information). Fifth, we verify the robustness of food-selectivity by showing consistent food-selective responses across independent sets of NSD images, and we provide the first characterizations of the fine-grained structure of representations within the food category itself. Finally and uniquely, we directly validate these results using hypothesis-driven methods in the form of a standard "localizer" that included grayscale images of food. The results of Experiment 2 replicate our results with NSD and provide direct evidence that color is not a confound in food-selectivity. Equally important, these results also suggest that food-responsive regions are more spatially variable across individuals than other visual functional ROIs, thereby helping us reconcile current findings of food selectivity with previous failures and with claims of overlap between food- and face-selective regions.

Although our focus was on selectivity in the ventral visual system, we note that we also observed food selectivity in the parietal and frontal cortices in Experiment 1; however, the localization of these regions was less consistent over subjects (Fig. 1C) and did not replicate when we used our context-free localizer images (Fig. 5). Other brain regions may also play a role in processing food information, particularly during visually-guided behavior. The dorsal visual areas in particular may process the actions or affordances associated with food (i.e., cooking/eating), as suggested by past work showing that object representations in dorsal visual cortex tend to be action-oriented[35,36]. Activation in frontal cortex appears to overlap roughly with orbitofrontal regions (semi-inflated bottom view map in Fig. 1C), which may reflect the involvement of these areas in processing reward information associated with certain foods[13,16,37,38]. Food selectivity was also observed in a number of subjects in the insular cortex, which has previously been implicated in taste processing[13,38]. While our paper focuses on visual selectivity for food in the fusiform cortex, future work should investigate the interaction of the visual food selective area with these other areas, perhaps using manipulations that vary reward or action representations evoked by food stimuli.

Our approach and results allow us to rule out several alternative explanations for the finding of food selectivity. It is not likely that food selectivity reflects preferential responses to "reachspaces"[7], rather than food *per se*. This is ruled out on the basis that our labeling taxonomy allowed us to control for image perspective (i.e., including *reach* as a label). Specifically, we found that food-selectivity remained stable even after removing the *reach* labeled images. Another possible alternative is that food-selectivity reflects preferential responses to small vs. big real-world object size[6], again, rather than food *per se*. However, the representation of real-world object size manifests as *big* flanking the medial side of the FFA and *small* flanking the lateral side of the FFA. Thus, this explanation can be ruled out in that our observed food selective responses co-locate more with big, as opposed to small, regions, yet food categories, particularly prepared foods, have small real-world size. Another possibility we can reject is that food selectivity can be solely attributed to greater attention or higher intrinsic visual salience for food relative to non-food[39]. Both human faces and bodies are subject to the same kinds of saliency effects[40], yet attentional/saliency differences are not the preferred explanation for face or body selectivity[41]. Moreover, within our study, faces and bodies comprised a reasonable proportion of the non-food contrast images, yet food selectivity was robust across these comparison categories (as is also the case in[16]). Finally, it is not likely that low- or mid-level visual features (i.e., spatial frequency, curvature, texture) underlie our pattern of results. This is supported by the fact that food selectivity was primarily found in higher visual areas, rather than early visual areas (Fig. 2). As discussed previously, the visual variability of food makes it unlikely that there is a set of low- or mid-level visual features or high-level shape structures that consistently correspond to food (in contrast, see[42–44]). Finally, as discussed below, another explanation that we can reject is that our food selective responses are mostly driven by color.

These conclusions are also consistent with a recent MEG study which excluded low-level visual features as an explanation for food selectivity[20]. Similarly, the two recent papers that likewise identified food selectivity using the same dataset we used here included several analyses that help to rule out a variety of low- or mid-level features as the basis for the observed selectivity.[17,18] Of note, both papers reported an intriguing overlap between food selectivity and color-biased brain regions. While Pennock et. al.[17] favor an account in which color is a feature common to food and, as such, food-selective regions may respond to color even in the absence of food inputs, Khosla et. al.[18] explicitly include color in their analyses and conclude that food selectivity cannot be explained by color alone. They do, however, acknowledge that selectivity for food and color-biased responses are "linked". As in Pennock et. al.,[17] they suggest that color is important for the identification,[45] evaluation,[46] and selection of food.[18] Our new data sheds conclusive light on this question. While color may be an important part of learning new food categories, the results of Experiment 2 demonstrate that food-selective responses can arise in the absence of color (Fig. 5). What remains to be determined in future work is whether the functional role of color in food-related behaviors leads to the instantiation of color biases in food-selective regions or whether color biases are present absent food selectivity and, as such, may help facilitate the acquisition of food representations in these regions.

Past work has presented conflicting accounts of the degree of overlap between food-selective and other category-selective visual regions.[16] However, claims of overlap are questionable in that they were based on group-level analyses and any overlap may have been an artifact of the variability in the localization of food-selective regions within individuals (which may arise in part from the high visual variability of food as a category). In particular, Adamson and Troiani[16] claimed that "there is overlap in face and food activation within the fusiform and that this is spatially consistent at the group level." This inference is puzzling in light of the fact that the same study presents a visualization of peak coordinates for both face and food clusters in individual subjects that appears to show separation between the two regions of selectivity (Fig. 2 of[16]). However, Adamson and Troiani focus on across-subject tests in order to support the claim that food selectivity co-localizes with face selectivity. This leads them to conclude, we believe incorrectly, that food and face recognition share a common neural substrate and, presumably, common underlying computational mechanisms.

The difference in our Experiment 1 results versus those of Adamson and Troiani[16] may be due to our use of a more sensitive within-subject, voxel-wise analyses. Across multiple methods and within 8 individual subjects, our results indicate that food and face selectivity do not co-locate (Fig. 2A, Suppl. Fig. S1 and Suppl. Fig. S5). Reinforcing this separation between regions, in our Experiment 2, there is almost no overlap between food- and face-selective areas in individual subjects (Fig. 5A and C). In this same experiment, the cluster of voxels with the greatest consistency across subjects is in the left fusiform directly adjacent to the FFA (Fig. 5D). We note that group averaging – as used in[16] – could potentially blur these significant food- and face-selective areas so as to create the appearance of overlap at the group level within left fusiform (as reported in[16]). Consistent with this interpretation and the results of our Experiment 2, as previously mentioned, Adamson and Troiani[16] report separation in the peak coordinates for face and food clusters for individual subjects. Consequently, there is little evidence to support a claim that food and face representations arise from the same fine-grained principles of visual processing. Rather, for reasons we have already discussed, there is variability in the localization of food-selective regions across subjects; as such, it is critical to assess selectivity on an individual basis.

More generally, why have most previous efforts to localize a food-selective region of ventral cortex failed (e.g., P. Downing and N. Kanwisher, 1999, Cogn. Neurosci. Soc., poster; based also on multiple anecdotal reports of similar failures)? As already discussed, the visual heterogeneity of food, the wide variety of factors that influence food behaviors, and the distributed and interleaved nature of neural representations of food all lead us to expect higher individual variability in the localization of food-related responses as compared to other categories of selectivity. As a result, group-average analyses (Fig. 5D) are unlikely to reveal a consistent, significant cluster of separable food-selective voxels across subjects. However, when one considers individual subject responses, not only do significant food-selective voxel clusters emerge, but we observe *spatially relative* consistency between these clusters and other functionally-localized ROIs (Fig. 5C). For this purpose, we make our food localizer available for the community, along with the code to process it and generate visualization using the pycortex software[47]. Based on this logic we are exploring whether other visually heterogeneous categories with high reward or social significance may, like food, come to be selectively represented – possibly intermixed with food representations – in ventral visual cortex.

Finally, while a finding of food selectivity naturally emerges from considering ecologically important visual categories, this leaves open the question as to how such selectivity arises in the human brain. We speculate that, similar to human language, domain-relevant perceptual inputs related to food can vary widely depending on the cultural and physical environment. Learned representations for food are only loosely constrained at the surface level, but still reflect common underlying mechanisms that have emerged over the course of evolution due to reward and the selection for learning abilities that flexibly responded to variations in inputs (the "Baldwin Effect"[48,49]). Thus, as a core property of knowledge organization, food selectivity is likely to have emerged as a neural preference shaped heavily by semantic associations, context, and reward.

Materials and Methods

Experiment 1

fMRI data We used the Natural Scenes Dataset (NSD)[19], consisting of high-resolution fMRI responses to natural scenes. The detailed experimental procedure are described by Allen et al.[19]. The naturalistic scene images were pulled from the annotated Microsoft Common Objects in Context (COCO) dataset[26]. 8 subjects each viewed between 9,000-10,000 natural scene images over the course of a year, each repeated 3 times. Of the 70,566 total images presented, 1,000 were intended to be viewed by all subjects. However, because some subjects dropped early, they didn't all see the 1000 images 3 times. For the purposes of this paper, we use any of the 1000 images for a subject if it was viewed at least once (515 were seen three times by each subject, 766 were seen at least two times and 907 at least one time). Thus, for subjects S1-S8 we use respectively 1000, 1000, 930, 907, 1000, 930, 1000, and 907 shared images.

The data were collected during 30-40 scan sessions. Images were square cropped, presented at a size of 8.4° × 8.4° and for 3 s with 1-s gaps in between images. The subjects were instructed to fixate on a central point and to press a button after each image if they had seen it previously.

The functional MRI data were acquired at 7T using whole-brain gradient-echo EPI at 1.8-mm resolution and 1.6-s repetition time. The preprocessing steps included a temporal interpolation (correcting for slice time differences) and a spatial interpolation (correcting for head motion). Single-trial beta weights were estimated with a general linear model. FreeSurfer[50,51] was used to generate cortical surface reconstructions to which the beta weights were mapped. The beta weights corresponding to each image were averaged across repetitions of the image, resulting in one averaged fMRI response to each image per voxel, in each subject.

The dataset also included several visual ROIs that were identified using separate functional localization experiments. We drew the boundaries of those ROIs for each subject on their native surface for better visualization and interpretation of the results. All brain visualization were produced using the pycortex software[52]. We create flattened, inflated and semi-inflated maps by setting the 'unfold' parameter to 1, 0 and 0.25 respectively. Fig. 1 and Fig. 2 show the left and right hemisphere for each type of view we show (flatmaps and semi-inflated or inflated bottom and lateral views). These conventions are maintained across all brain plots in the manuscript and supplemental materials.

Image labeling The authors and a graduate student in our labs (n=8) performed manual image labeling for the 1,000 potentially shared images based on each image's depicted location, image perspective and content. Location refers to whether the image is indoor or outdoor (or ambiguous), content refers to the categories of objects in the image (including the binary existence of food), and image perspective refers to the approximate scale of the image, discretized into *zoom*, *reach* or *large-scale*. *Zoom* refers to a very close shot, thereby likely concentrated on one object and excluding surrounding information. *Reach* images display objects at a human-reachable distance, and may activate representations related to object affordances[7,28]. *Large-scale* images encompass the remaining images, which include an image of a typical scene as opposed to one or more close-up objects. Images could only be assigned one label for location and perspective, but could be assigned multiple content labels. More details about this image labeling are described in the Figure 1A and B. Labeling was performed using the Computer Vision Annotation Tool[53]. In order to avoid variation in labels and ensure consistency, we performed several rounds of labeling and verification across multiple raters; each image was seen by a least two raters. Disagreements were discussed in the group of raters until unified labeling assignments were reached.

Encoding models We constructed two different encoding models. The first was based on our hand-labeled annotations of the 1,000 potentially shared images (Fig. 2). Encoding all 16 hand-labels into a single binary vector per image, we utilized voxel-wise ordinary least squares (OLS) encoding models to predict each individual voxel response to a given stimulus. Identifying voxels more responsive to category A over other category was done using a 1-sided t-test between the respective learned model coefficients for category A vs. the coefficients for the other categories, as is done in a typical generalized linear model (GLM) analysis. Note that this analysis collapses across the three "attributes" used in our labeling taxonomy (i.e., food is compared against object categories like faces, as well as against location labels like indoor). We used these methods to identify voxels that are more responsive to food than other labels, as well as for face versus other labels. We obtained a p-value from the t-value, then corrected for multiple comparisons across all voxels using the Benjamini-Hochberg False Discovery Rate procedure (FDR)[54], which is appropriate for fMRI results due to the assumption that they show positive dependence[55,56]. The significance of the contrast was computed at the subject level, the results were converted to MNI space, and the sum across subjects was plotted in Fig. 1C. Pycortex was used for transformation to MNI space of each subject's result. It relies on the Flirt tool[57–59] from FSL.[60]

Our second encoding model was based on COCO object category labels, and made use of the set of images that were unique to each subject (Fig. 3). The purpose of this model was to verify that our proposed food region derived from the shared images is consistent across the larger set of images that also includes images not used in the first analysis. We used the 80 COCO object category annotations provided in the dataset, specifically each COCO label's corresponding bounding box proportion relative to the image (i.e., proportion of the image covered by the category of interest), as input to a ridge regression encoding model. We built and tested the model via 10-fold cross-validation, where R^2 was computed on a tenth of the data not used for training at each fold, and the 10 resulting R^2 values were averaged. The penalty parameter for each voxel was chosen independently by nested $10-$fold cross-validation. When determining which images were used to fit the encoding model, we create a set of images that contained half food and half non-food images. We considered images to include food if their maximum food label proportion exceeded a threshold of 0.15. We identified 940 such images, and randomly selected 940 non-food images, together creating a total input set of 1880 images. We built two models, one with all the labels, and one with all the labels that were not food (67 in total). We then computed the voxel-wise R^2 improvement from including food labels in the regression. In addition to helping identify voxels that responded most to inclusion of food, this encoding model also helped us visualize food sub-category activations. We observed the voxel-wise learned weights corresponding to specific COCO food labels (i.e. cake, sandwich) to uncover potential food sub-category patterns.

Decoding models While an encoding model is able to provide some insight into single-voxel selectivity through response predictions, a decoding model can uncover distributed pattern-level representations of visual features. To observe representations at the population level, we used a searchlight decoding method[31]. Specifically, for each voxel in the cortical sheet, we defined a searchlight sphere that consisted of 27 nearby voxels, and we trained a decoder to classify the existence of food based on the pattern of activation across these voxels. We used 5-fold cross validation via Support Vector Classification, with our input image set consisting of 108 food images and 108 randomly selected non-food images from the shared images. High decoding accuracy from this method suggests that an area encodes food-related information at the pattern level, which our model is able to exploit in order to classify the existence of food.

Determining the ventral visual food selective regions To generate a mask that only included the ventral visual food selective region, we first manually selected apparent relevant ROIs via the Glasser HCP Atlas[29]. We use the concatenation of sub-areas TE2p, PH, VVC, v8, PIT, FFC, and VMV3 to create our mask. After converting the mask for this anatomical area into each subject's native space, we identified the intersection of this mask with the identified food region from a food vs non-food significance test (Fig. 2 shows the final mask definition).

Principal Component Analysis (PCA) We ran PCA to better understand possible structure and/or correspondence in these food-selective regions. Using the food mask above that consists only of our proposed food region, we selected 'food-relevant' voxels for each subject. Then, we ran PCA on a matrix of concatenated 'food-relevant' voxels for all subjects (rows) by the activity related to shared food images (columns), reducing along the image dimension (the columns). We extracted the top principal axes of this matrix, and projected our initial data matrix onto the calculated lower-dimensional space to obtain the voxel-wise PC scores on the brain. To compare the voxel-wise PC scores across subjects, we converted the scores for each subject to the MNI template and average the scores across subjects for each MNI voxel. We identified the most positive and negative contributing images to each axis by computing the dot-product between the PC score and the activity related to an image, to assess whether the representations of each principal axis were cohesive or semantically interpretable.

Clustering analyses We ran a K-means clustering analysis to better investigate visual and semantic patterns in the food selective regions. As a point of comparison with the voxel clustering results, we also clustered visual and semantic embeddings of these images derived from deep neural networks. To compute the clusters for one subject, we picked 940 food images. Voxel embeddings were calculated for each individual subject, using responses from voxels within the ventral food mask. To obtain visual and semantic embeddings for these same 940 images we used two trained deep neural networks: CLIP and ResNet-18[33,34]. CLIP, trained on both images and text, allows us to extract features arising from a contrastive learning paradigm with dual semantic and visual constraints. We used the pretrained ViTB32 model, which was trained to align image and text embeddings within a shared space. Within this model, we extracted the features given an input image from the vision module of the model. Given an image, we call these corresponding CLIP features the CLIP embedding.

ResNet-18, trained on solely images, provides a visual feature-based embedding with no language component. Given an image, we ran a ResNet-18 model pretrained on ImageNet to extract the features from the average pool layer immediately preceding the final fully-connected layer[61]. We refer to these extracted features for a given image as the corresponding ResNet embedding of that image.

To cluster embeddings, we used K-means clustering algorithm with Euclidian distance. We consider a range of K values and for each, observe the average Euclidian distance from each data point to their corresponding cluster centroid. Next, we selected the first K value that led to the drop in the average distance for voxel embeddings beyond which the decrease plateaus (the elbow method). This value was 4. We use this same $K = 4$ for all three embedding clusterings.

To compare different clustering assignments, we constructed for each clustering procedure a 940×940 matrix where the rows and columns correspond to the 940 images. Each cell in this matrix is an indicator value where $\text{matrix}_{i,j}$ is 1 if the two images i and j are in the same cluster, and 0 otherwise. We then used Pearson correlation to compute the correlations between two clustering assignments. To visualize each cluster, we chose the closest images to the centroid of that cluster.

Experiment 2

MRI data collection MRI data were acquired on a 3T Siemens Prisma MR scanner at the BRIDGE center at the Carnegie Mellon University campus using a 64-channel phased array head coil.

Functional Images. Functional images were collected using a T2*-weighted gradient recalled echoplanar imaging multi-band pulse sequence (cmrr_mbep2d_bold) from the University of Minnesota Center for Magnetic Resonance Research (CMRR).[62,63] Parameters: 68 oblique axial slices co-planar with the AC/PC; in-plane resolution=2×2mm; 106×106 matrix size; 2mm slice thickness, no gap; interleaved acquisition; field of view=212mm; phase partial Fourier scheme of 6/8; TR=1500 ms; TE=30ms; flip angle=79 degrees; bandwidth=1814 Hz/Px; echo spacing=0.68ms; excite pulse duration=8200 microseconds; multi-band factor=4; phase encoding direction=A to P; fat saturation on; advanced shim mode on. During functional scans, eyetracking was acquired using an EyeLink eye tracker.

Anatomical Images. A T1 weighted MPRAGE scan was collected for each participant. MPRAGE parameters: 208 sagittal slices; 1mm isovoxel resolution; field of view=256mm; TR=2300ms; TE=2.03ms; TI=900ms: flip angle=9 degrees; GRAPPA acceleration factor=2; bandwidth=240Hz/Px.

Subjects Functional data were collected from four subjects (3 female/1 male) aged 21-29. All subjects were healthy and had corrected to normal vision. Written informed consent was obtained from all subjects and the study was approved by the Carnegie Mellon University Institutional Review Board.

Paradigm The data was collected in runs of length 4 minutes. Subject LS1 underwent 4 runs of the localizers, while subjects LS2-LS4 underwent 9 runs. We did not see a difference that appeared to be driven by the amount of data.

We selected 82 images of different types of food with transparent backgrounds from the https://www.stickpng.com/ website. We converted the images to grayscale and superimposed them on images from the scrambled condition in the fLoc localizer[27] (Fig. 5B illustrates some examples). All images can be found at this link:

https://www.cs.cmu.edu/~lwehbe/files/food_images.zip

We used the mini-block design (duration = 6s) proposed by Stigliani et al.[30] Along with the food condition, we also use the adult condition (to define faces), the house condition (to define places), the word condition and the body condition. We use the first 82 images from each condition provided in the localizer to have the same number as the food images. We adapted the Stigliani et al.[30] code to present our stimuli. The code uses Psychtoolbox-3[64–66] and runs on Matlab.

Images were square, presented on a gray background at a size of 11.4° × 11.4° visual angle on a BOLDscreen32 LCD Display and for 0.5 s each. The subjects were instructed to fixate on a central point and to press a button if they see a repeated image (1-back task).

Data preprocessing Each subject's native surface was reconstructed using Freesurfer.[67] Functional scans were motion corrected using SPM12.[68] Through pycortex, alignment of the functional data to the structural data was obtained (using bbregister from Freesurfer). Our code pipeline includes detrending and lightly smoothed with a Gaussian kernel of standard deviation 1mm, using standard functions part of the scipy package.[69] Pycortex was used to mask the cortical data (by relying on maps estimated by Freesurfer). Pycortex was also used for transformation to MNI space. It relies on the Flirt tool[57–59] from FSL.[60]

Encoding models We followed the same procedure used with the shared NSD images to compute a contrast between condition A and other conditions after estimating voxel-wise ordinary least squares (OLS) encoding models. We computed a t-value for each of the "food vs. other", "face vs. other", "body vs. other", "place vs. other" and "word vs. other" contrasts. We used the Benjamini-Hochberg False Discovery Rate procedure (FDR)[54] and $\alpha = 0.05$ to identify significant voxels for each contrast of each subject at each voxel.

We drew the boundaries of the FFA, EBA and PPA for each subject using the "face vs. other", "body vs. other" and "place vs. other" significance maps, respectively. This enables us to better understand the "food vs. other" contrast results. Note that unlike in NSD where the ROIs labeled using separate data, here the data from the "food vs. other" contrast is the same as the one used to draw the other ROIs. The significance of the contrast was computed at the subject level, the results were converted to MNI space, and the sum across subjects was plotted in Fig. 5. The same was repeated for the other contrasts which can be seen in S9.

Acknowledgments Collection of the NSD dataset was supported by NSF IIS-1822683 and NSF IIS-1822929. The authors thank the NSD team for collecting and sharing the dataset. MMH was funded by a Distinguished Postdoctoral Fellowship from the Carnegie Mellon Neuroscience Institute. JP was supported by NSF BCS-1658278. The authors thank the following people for contributing ideas and commentary to this project: Laurie M. Heller, Andrew Luo, Isabel Gauthier, Marlene Behrmann, and Brad Mahon.The authors also thank John Pyles for help with setting up the localizer.

Author Contributions All authors conceived of the research and provided hand labels for the shared images. N.J. coded and performed all analyses on NSD data. A.W. also provided code. L.W, M.T. and M.H. designed the localizer. L.W. performed and analysed the localizer experiment. All authors wrote and edited the paper.

References

[1] J Sergent, S Ohta, and B MacDonald. Functional neuroanatomy of face and object processing: A positron emission tomography study. *Brain*, 115:15–36, 1992.

[2] N Kanwisher, J McDermott, and M M Chun. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, 17(11):4302–4311, 1997.

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

[4] P E Downing, Y Jiang, M Shuman, and N Kanwisher. A cortical area selective for visual processing of the human body. *Science*, 293(5539):2470–2473, 2001.

[5] Bruce D McCandliss, Laurent Cohen, and Stanislas Dehaene. The visual word form area: expertise for reading in the fusiform gyrus. *Trends Cogn Sci*, 7(7):293–299, 2003.

[6] Talia Konkle and Alfonso Caramazza. Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, 33(25):10235–10242, 2013.

[7] Emilie L. Josephs and Talia Konkle. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, 117(47):29354–29362, 2020.

[8] N Kanwisher. Domain specificity in face perception. *Nat. Neurosci.*, 3(8):759–763, 2000.

[9] M J Tarr and I Gauthier. FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nat Neurosci*, 3(8):764–769, 2000.

[10] Jing Chen, Esther K. Papies, and Lawrence W. Barsalou. A core eating network and its modulations underlie diverse eating phenomena. *Brain and Cognition*, 110:20–42, 2016.

[11] Claudia I Huerta, Pooja R Sarkar, Timothy Q Duong, Angela R Laird, and Peter T Fox. Neural bases of food perception: coordinate-based meta-analyses of neuroimaging studies in multiple modalities. *Obesity (Silver Spring).*, 22(6):1439–1446, 2014.

[12] Raffaella I. Rumiati and Francesco Foroni. We are what we eat: How food is represented in our mind/brain. *Psychon. Bull. Rev.*, 23(4):1043–1054, 2016.

[13] L.N. van der Laan, D.T.D. de Ridder, M.A. Viergever, and P.A.M. Smeets. The first taste is always with the eyes: A meta-analysis on the neural correlates of processing visual food cues. *NeuroImage*, 55(1):296–303, 2011.

[14] P E Downing, A W Chan, M V Peelen, C M Dodds, and N Kanwisher. Domain specificity in visual cortex. *Cerebral cortex (New York, N.Y. : 1991)*, 16:1453–1461, 2006.

[15] Ruud van den Bos and Denise de Ridder. Evolved to satisfy our immediate needs: Self-control and the rewarding properties of food. *Appetite*, 47(1):24–29, 2006.

[16] Kayleigh Adamson and Vanessa Troiani. Distinct and overlapping fusiform activation to faces and food. *NeuroImage*, 174:393–406, 2018.

[17] Ian Morgan Leo Pennock, Chris Racey, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, Anna Franklin, and Jenny Bosten. Color-biased regions in the ventral visual pathway are food-selective. *bioRxiv*, 2022.

[18] Meenakshi Khosla, N. Apurva Ratan Murty, and Nancy Kanwisher. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Current Biology*, 32:1–13, 2022.

[19] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, 2022.

[20] Kleovoulos Tsourides, Shahriar Shariat, Hossein Nejati, Tapan K. Gandhi, Annie Cardinaux, Christopher T. Simons, Ngai-Man Cheung, Vladimir Pavlovic, and Pawan Sinha. Neural correlates of the food/non-food visual distinction. *Biological Psychology*, 115:35–42, 2016.

[21] Allan D Gordon. A review of hierarchical classification. *Journal of the Royal Statistical Society: Series A (General)*, 150(2):119–137, 1987.

[22] Roger Alan Stein, Patricia A Jaques, and Joao Francisco Valiati. An analysis of hierarchical text classification using word embeddings. *Information Sciences*, 471:216–232, 2019.

[23] Carlos N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1):31–72, 2011.

[24] Timothy S. Coalson, David C. Van Essen, and Matthew F. Glasser. The impact of traditional neuroimaging methods on the spatial localization of cortical areas. *Proceedings of the National Academy of Sciences*, 115(27):E6356–E6365, 2018.

[25] Alfonso Nieto-Castañón and Evelina Fedorenko. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage*, 63(3):1646–1669, 2012.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014.

[27] Anthony Stigliani, Kevin S Weiner, and Kalanit Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

[28] Emilie L Josephs, Haoyun Zhao, and Talia Konkle. The world within reach: an image database of reach-relevant environments. *Journal of Vision*, 21(7):14–14, 2021.

[29] Matthew F Glasser, Timothy S Coalson, Emma C Robinson, Carl D Hacker, John Harwell, Essa Yacoub, Kamil Ugurbil, Jesper Andersson, Christian F Beckmann, Mark Jenkinson, Stephen M Smith, and David C Van Essen. A multi-modal parcellation of human cerebral cortex. *Nature*, 536(7615):171–178, 2016.

[30] Anthony Stigliani, Kevin S. Weiner, and Kalanit Grill-Spector. Temporal processing capacity in high-level visual cortex is domain specific. *Journal of Neuroscience*, 35(36):12412–12424, 2015.

[31] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, 103(10):3863–3868, 2006.

[32] James V. Haxby, M. Ida Gobbini, Maura L. Furey, Alumit Ishai, Jennifer L. Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Jorge Almeida, Bradford Z. Mahon, and Alfonso Caramazza. The role of the dorsal visual processing stream in tool identification. *Psychological Science*, 21(6):772–778, 2010.

[36] Maryam Vaziri-Pashkam and Yaoda Xu. Goal-directed visual processing differentially impacts human ventral and dorsal visual representations. *Journal of Neuroscience*, 37(36):8767–8782, 2017.

[37] Edmund T. Rolls. The Orbitofrontal Cortex and Reward. *Cerebral Cortex*, 10(3):284–294, 2000.

[38] Edmund T Rolls. Brain mechanisms underlying flavour and appetite. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 361(1471):1123–1136, 2006.

[39] Sanjay Kumar, Suzanne Higgs, Femke Rutters, and Glyn W. Humphreys. Biased towards food: Electrophysiological evidence for biased attention to food stimuli. *Brain Cogn.*, 110:85–93, 2016.

[40] Stephen R.H. Langton, Anna S. Law, A. Mike Burton, and Stefan R. Schweinberger. Attention capture by faces. *Cognition*, 107(1):330–342, 2008.

[41] Maura L. Furey, Topi Tanskanen, Michael S. Beauchamp, Sari Avikainen, Kimmo Uutela, Riitta Hari, and James V. Haxby. Dissociation of face-selective cortical responses by attention. *Proceedings of the National Academy of Sciences*, 103(4):1065–1070, 2006.

[42] Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

[43] Shahin Nasr and Roger B H Tootell. A cardinal orientation bias in scene-selective visual cortex. *J Neurosci*, 32(43):14921–14926, 2012.

[44] Xiaomin Yue, Irene S Pourladian, Roger B H Tootell, and Leslie G Ungerleider. Curvature-processing network in macaque visual cortex. *Proc Natl Acad Sci U S A*, 111(33):E3467–75, 2014.

[45] G. Naor-Raz, M.J. Tarr, and D. Kersten. Is color an intrinsic property of object representation? *Perception*, 32(6), 2003.

[46] B C Regan, C Julliot, B Simmen, F Vienot, P Charles-Dominique, and J D Mollon. Fruits, foliage and the evolution of primate colour vision. *Philos Trans R Soc L. B Biol Sci*, 356(1407):229–283, 2001.

[47] James S Gao, Alexander G Huth, Mark D Lescroart, and Jack L Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in neuroinformatics*, page 23, 2015.

[48] J. Mark Baldwin. A new factor in evolution. *The American Naturalist*, 30(354):441–451, 1896.

[49] Patrick Bateson. The Active Role of Behaviour in Evolution. *Biology and Philosophy*, 19(2):283–298, 2004.

[50] Anders M. Dale, Bruce Fischl, and Martin I. Sereno. Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999.

[51] Bruce Fischl, Martin I. Sereno, and Anders M. Dale. Cortical surface-based analysis: Ii: Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999.

[52] James S. Gao, Alexander G. Huth, Mark D. Lescroart, and Jack L. Gallant. Pycortex: an interactive surface visualizer for fmri. *Frontiers in Neuroinformatics*, 9, 2015.

[53] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, 2020.

[54] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[55] Christopher R Genovese. A bayesian time-course model for functional magnetic resonance imaging data. *Journal of the American Statistical Association*, 95(451):691–703, 2000.

[56] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[57] Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

[58] Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

[59] Douglas N Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1):63–72, 2009.

[60] Mark Jenkinson, Christian F Beckmann, Timothy EJ Behrens, Mark W Woolrich, and Stephen M Smith. Fsl. *Neuroimage*, 62(2):782–790, 2012.

[61] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[62] Steen Moeller, Essa Yacoub, Cheryl A Olman, Edward Auerbach, John Strupp, Noam Harel, and Kâmil Uğurbil. Multiband multislice ge-epi at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fmri. *Magnetic resonance in medicine*, 63(5):1144–1153, 2010.

[63] David A Feinberg, Steen Moeller, Stephen M Smith, Edward Auerbach, Sudhir Ramanna, Matt F Glasser, Karla L Miller, Kamil Ugurbil, and Essa Yacoub. Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. *PloS one*, 5(12):e15710, 2010.

[64] David H Brainard and Spatial Vision. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997.

[65] Denis G Pelli and Spatial Vision. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10:437–442, 1997.

[66] Mario Kleiner, David Brainard, and Denis Pelli. What's new in psychtoolbox-3? 2007.

[67] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

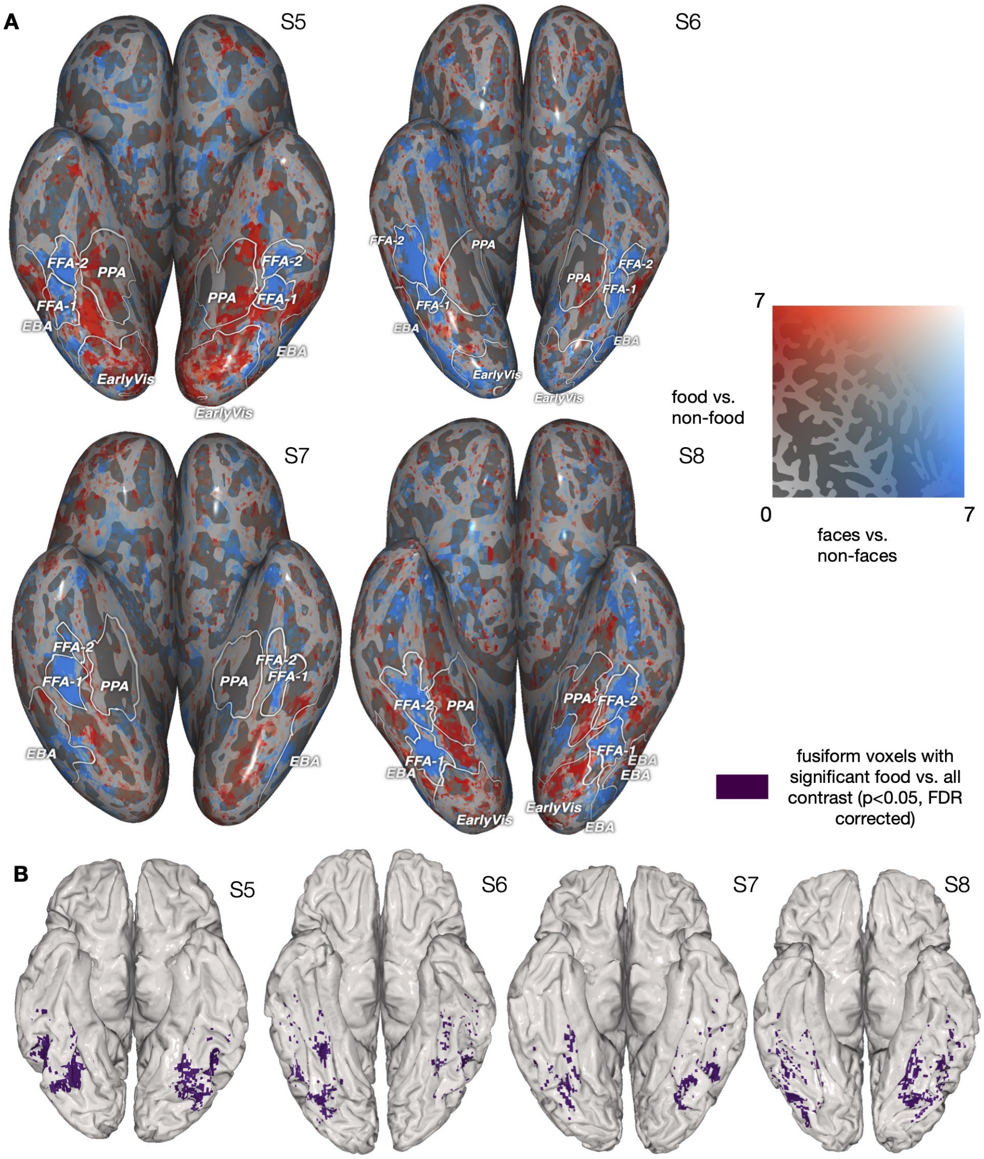
[68] John Ashburner, Gareth Barnes, Chun-Chuan Chen, Jean Daunizeau, Guillaume Flandin, Karl Friston, Stefan Kiebel, James Kilner, Vladimir Litvak, Rosalyn Moran, et al. Spm12 manual. *Wellcome Trust Centre for Neuroimaging, London, UK*, 2464:4, 2014.

[69] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.

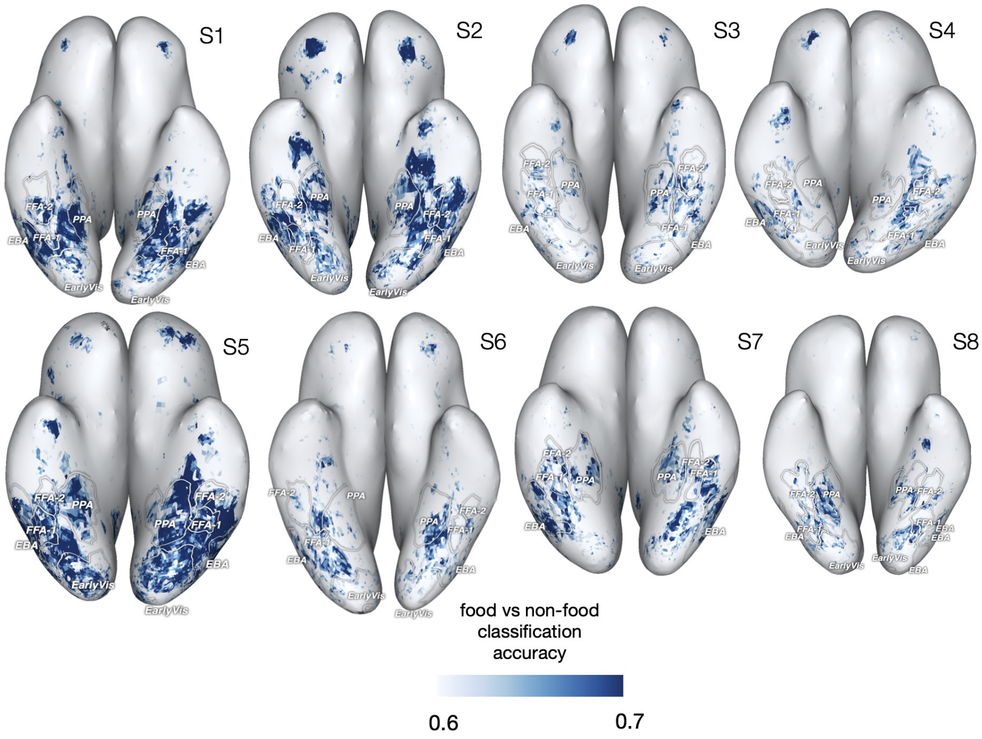
Supplementary Materials

label	count
indoor	251
outdoor	693
ambiguous-location	54
plant	45
human-face	180
human-body	367
animal-face	142
animal-body	246
food	108
drink	25
food-related	130
faux-food	0
zoom	82
reach	80
large-scale-scene	833
object	551

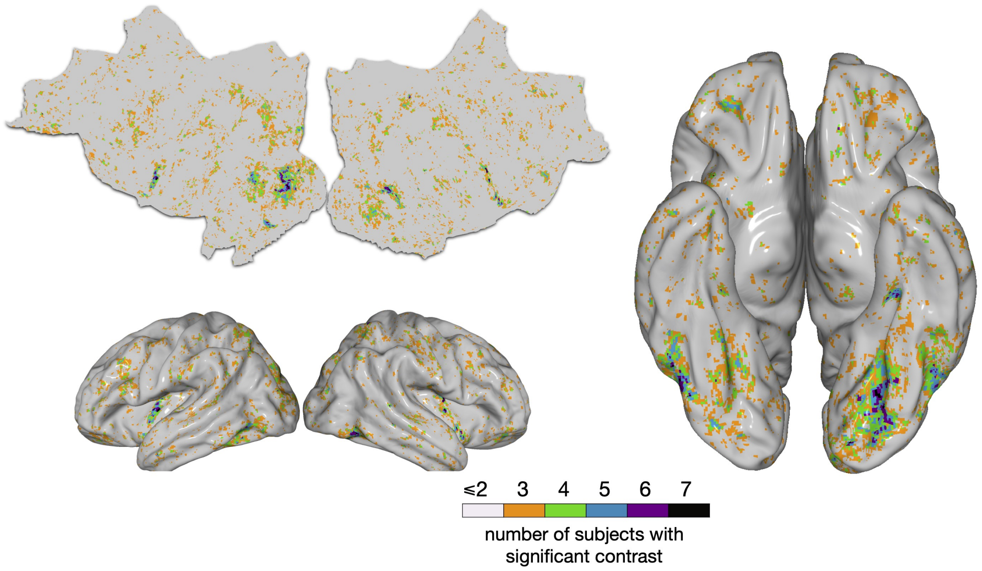
Supplementary Table S1: **Experiment 1.** Count of the occurrence of each label across the 1,000 potentially shared images.



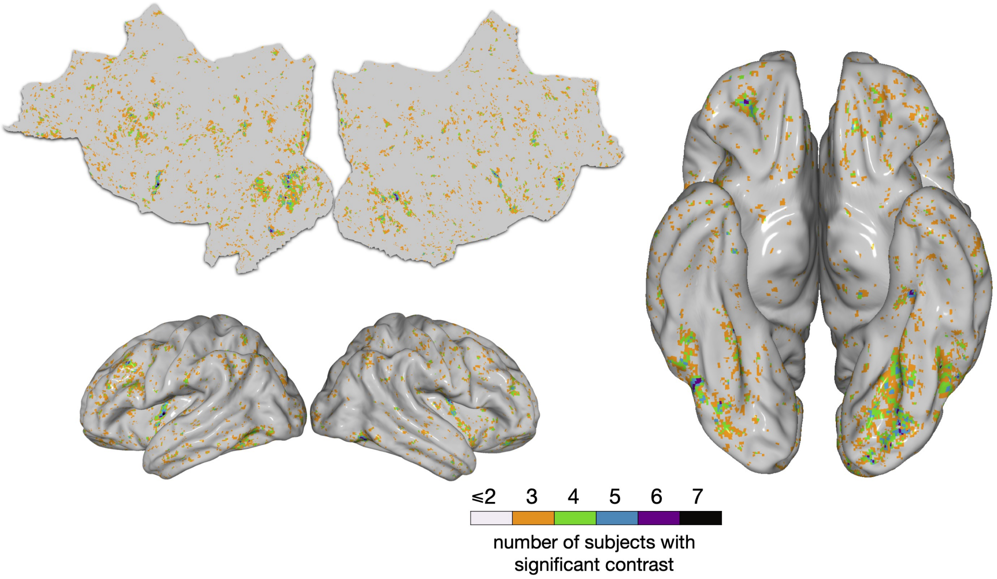
Supplementary Figure S1: ***Experiment 1.*** Voxels identified as selective for food for subjects S5-S8 shown on each subject's native surface with an inflated, bottom view (similar to Figure 2, but for different subjects) (A) Voxels' corresponding t-statistics from two 1-sided *t*-tests comparing food vs. non-food (red) and face vs. non-face (blue). Each t-test was performed on the weights from a trained OLS model, for example comparing the food label's learned weight against non-food labels' learned weights. The two sets of regions identified by each contrast are largely non-overlapping. This pattern is maintained when looking at food vs. non-(food and face) and face vs. non-(face and food) (Fig. S5). These results indicate that the two sets of regions have distinct activity for food and faces. (B) Spatial mask for food-selective regions used in subsequent analyses for S5-S8 (highlighting ventral visual responses). The mask is the overlap between the region that is identified from the *t*-test for food vs. non-food and relevant functionally localized regions using the HCP atlas[29] (see *Methods*).



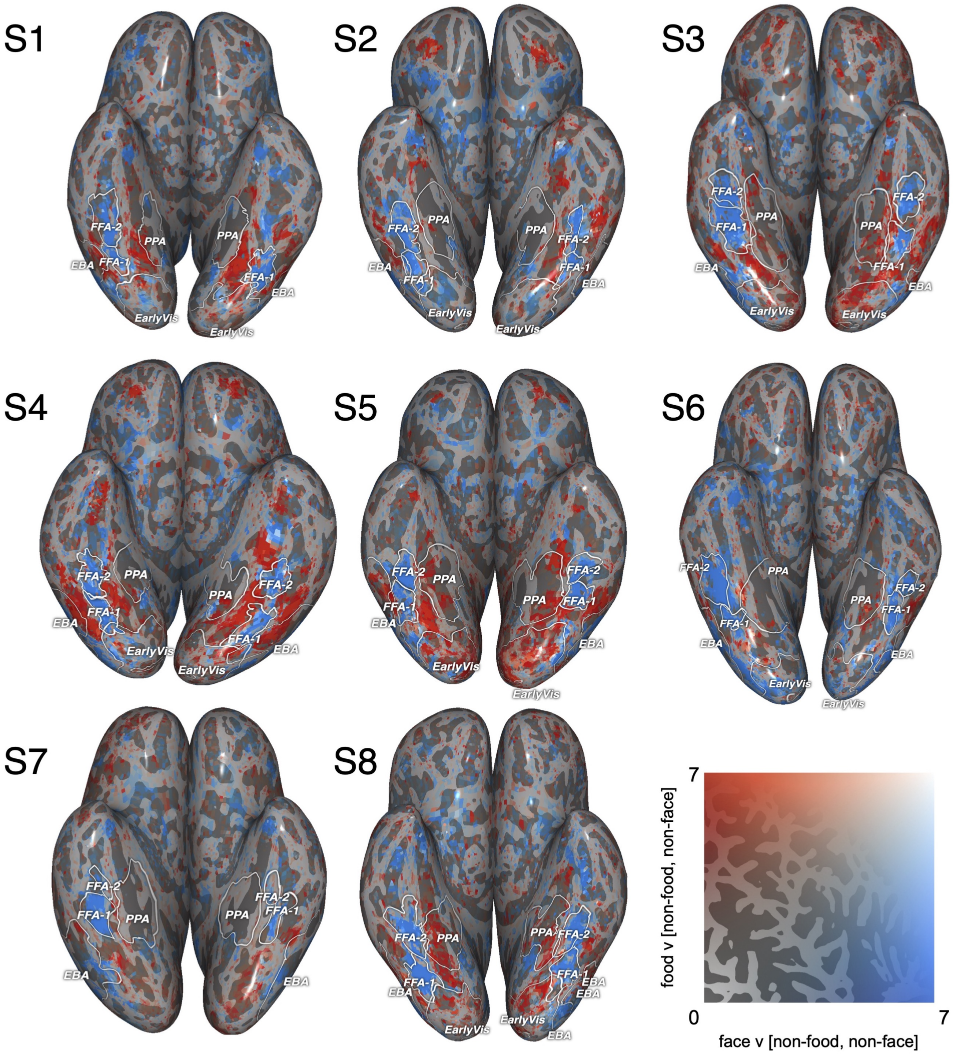
Supplementary Figure S2: ***Experiment 1.*** Classification accuracy for multivariate searchlight decoding food vs. non-food images for S1-S8, with darker voxels signifying higher accuracy. These regions encompass the two sets of regions corresponding to high values for the food vs. non-food and the face vs. non-face contrasts.



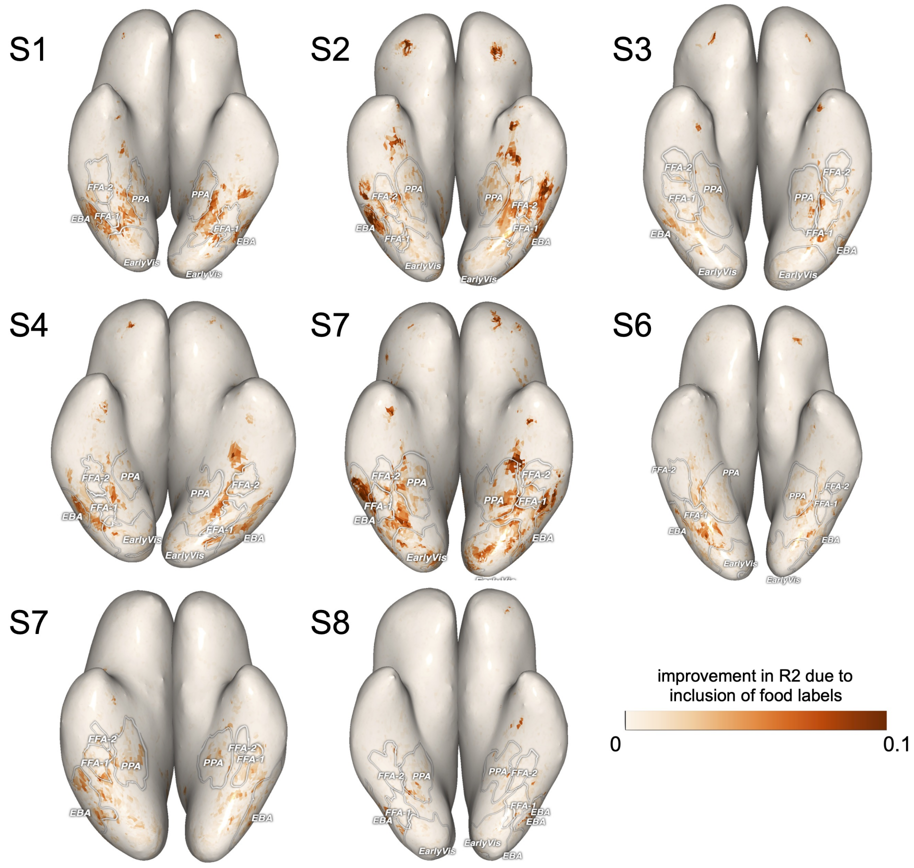
Supplementary Figure S3: ***Experiment 1.*** Semi-inflated bottom view of voxels, summed across S1-S8, that have significantly higher activity for the food than non-food categories, on the MNI surface, *considering only the non-reach images.* Significant voxels were identified similarly to Figure 1C.



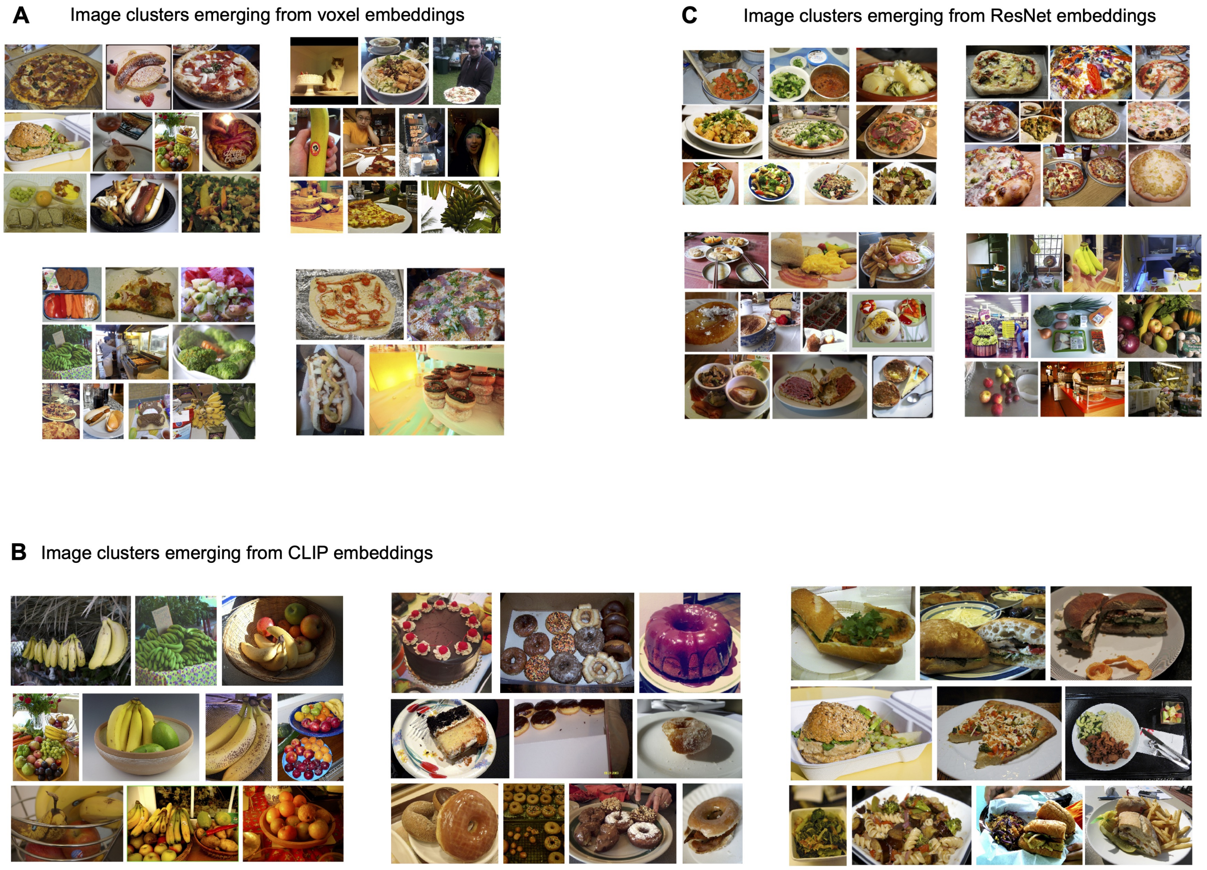
Supplementary Figure S4: ***Experiment 1.*** Semi-inflated bottom view of voxels, summed across S1-S8, that have significantly higher activity for the food than non-food categories, on the MNI surface, *considering only the non-zoom images.* Significant voxels were identified similarly to Figure 1C.



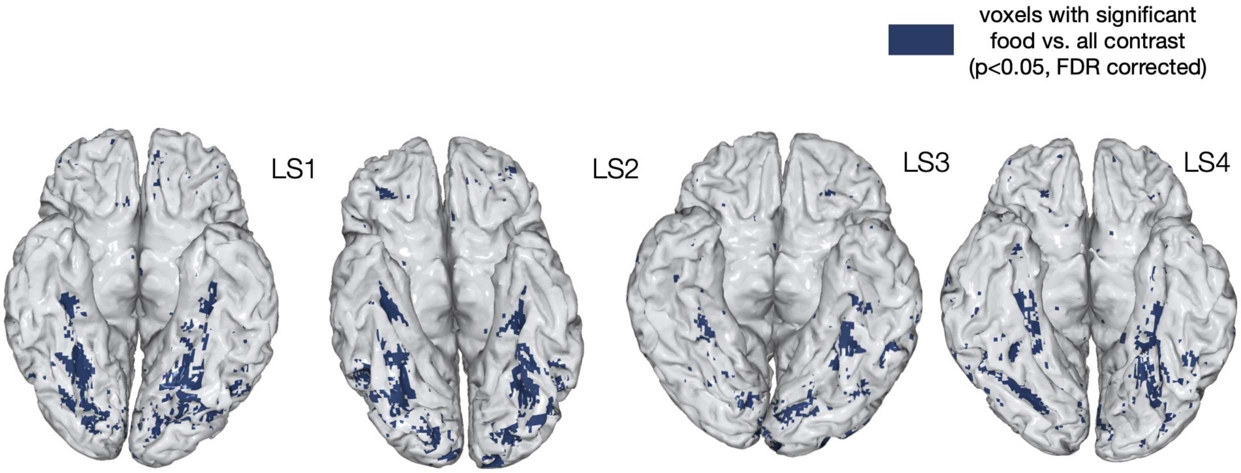
Supplementary Figure S5: ***Experiment 1.*** Voxels identified as selective for food from comparisons between food or faces vs. a baseline with *both* food *and* face removed for S1-S8. As described in Figure 2A, significant voxels were identified using two 1-sided t-tests. Despite a lower-N comparison arising from removing both faces and food from the baseline, there is still clear separability and little overlap between food-selective and face-selective regions.



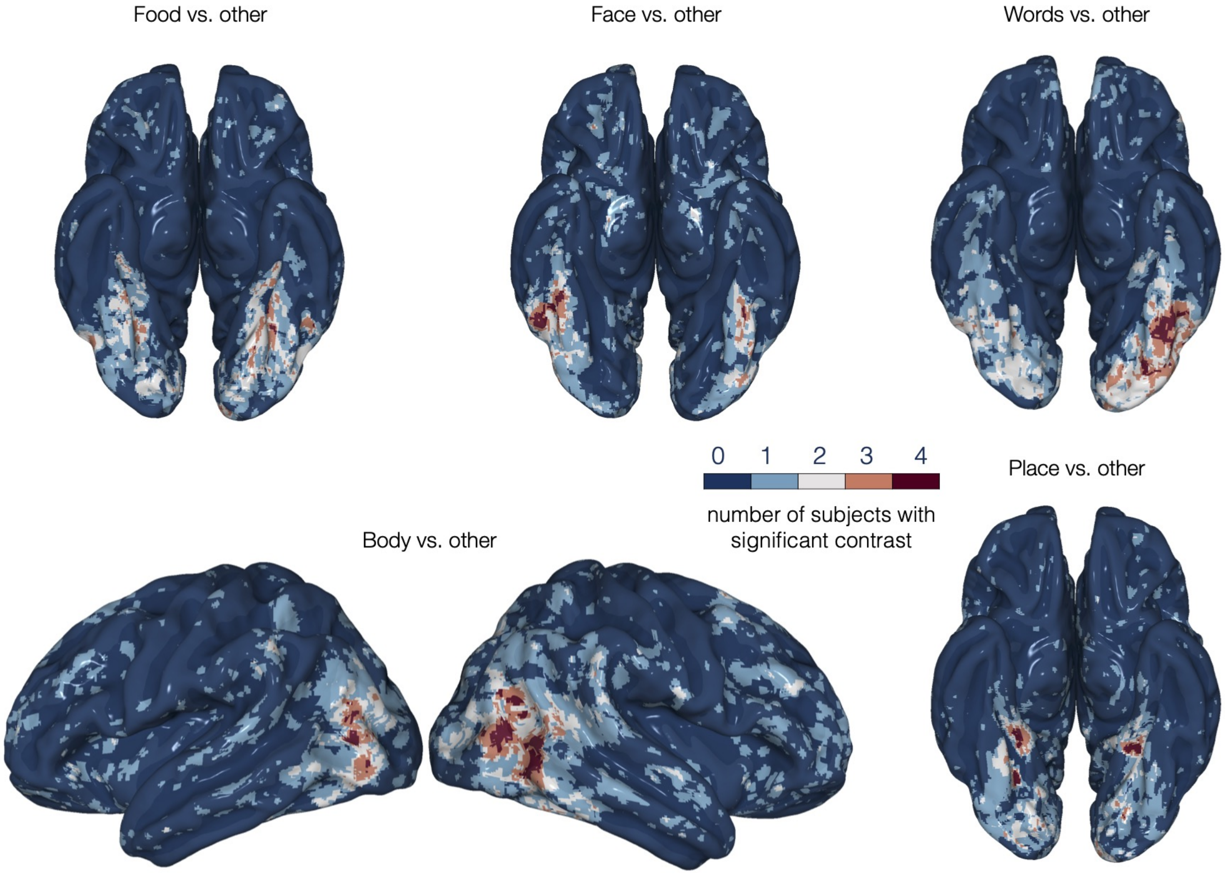
Supplementary Figure S6: ***Experiment 1.*** We compared predictive accuracy of an encoding model with all the COCO labels (including 13 food and 67 non-food labels) to an encoding model with only the 67 non-food COCO labels. The figure shows, for S1-S8, the improvement in validation set R^2 values when including the food labels (R^2 for the full model - R^2 for the model with food removed).


A
Image clusters emerging from voxel embeddings
C
Image clusters emerging from ResNet embeddings
B
Image clusters emerging from CLIP embeddings

Supplementary Figure S7: ***Experiment 1.*** (A) Image clusters based on voxel-response embeddings for S1. (B) Image clusters based on CLIP embeddings. The clusters appear to capture semantic properties such as fruit or baked goods. (C) Image clusters based on ResNet-18 embeddings. The clusters appear to capture visual properties such as color (e.g., green and orange), global shape (e.g., round), or image complexity. The two neural-network-derived clustering patterns show little to no correlation with the brain-derived clusters (the Voxel-CLIP correlation being 0.030; the Voxel-Resnet-18 correlation being 0.026 – both being lower than the CLIP-Resnet-18 correlation of 0.256). This suggests that food-selective regions are organized on the basis of features absent from deep layers of typical high-performing neural networks.


voxels with significant food vs. all contrast (p<0.05, FDR corrected)
LS1
LS2
LS3
LS4

Supplementary Figure S8: ***Experiment 2.*** Voxels identified as selective for food for subjects LS1-LS4 shown on each subject's native surface with a semi-inflated, bottom view. Voxels were identified as selective by testing for the significance of the contrast (p <0.05, FDR corrected).



Supplementary Figure S9: ***Experiment 2.*** Semi-inflated bottom view of voxels, summed across LS1-LS4, that have significantly higher activity for the different contrasts in experiment 2, on the MNI surface. Significant voxels were identified similarly to Fig. 5.