RESOURCE
https://doi.org/10.1038/s41593-021-00962-x

nature neuroscience

A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence

Emily J. Allen[1,2], Ghislain St-Yves[3,17], Yihan Wu[4], Jesse L. Breedlove[3,18], Jacob S. Prince[5,19], Logan T. Dowdle[6,7], Matthias Nau[8], Brad Caron[9,10], Franco Pestilli[11,12,13], Ian Charest[14,15], J. Benjamin Hutchinson[16], Thomas Naselaris[3,17,20] and Kendrick Kay[1,20] ✉

Extensive sampling of neural activity during rich cognitive phenomena is critical for robust understanding of brain function. Here we present the Natural Scenes Dataset (NSD), in which high-resolution functional magnetic resonance imaging responses to tens of thousands of richly annotated natural scenes were measured while participants performed a continuous recognition task. To optimize data quality, we developed and applied novel estimation and denoising techniques. Simple visual inspections of the NSD data reveal clear representational transformations along the ventral visual pathway. Further exemplifying the inferential power of the dataset, we used NSD to build and train deep neural network models that predict brain activity more accurately than state-of-the-art models from computer vision. NSD also includes substantial resting-state and diffusion data, enabling network neuroscience perspectives to constrain and enhance models of perception and memory. Given its unprecedented scale, quality and breadth, NSD opens new avenues of inquiry in cognitive neuroscience and artificial intelligence.

Neuroscience has an insatiable appetite for data. Many ongoing efforts to extensively sample brain activity[1–3] and structure[4–6] are motivated, in part, by the availability of new computational methods that make analysis of massive datasets feasible. Equally as important is the growing desire to understand how the brain coordinates complex sensory and motor behaviors and the realization that the neural networks supporting such behaviors span multiple scales, from single neurons to local circuits to whole systems. Understanding massive, complex networks will inevitably require commensurately massive amounts of data.

The need for massive data is especially acute in visual neuroscience, which is a model system for understanding brain function. The network that mediates our ability to flexibly and efficiently perceive the visual world occupies approximately one-third of human cerebral cortex[7] and interconnects brain areas with profoundly different functional properties[8]. This network both encodes visual stimuli and interfaces visual representations into a cognitive context, including information about what one has already seen[9], might see[10] or is selectively attending[11]. Understanding vision thus means interrogating a high-dimensional, context-dependent neural network.

Given these considerations, it is clear that extensive experimental data providing access to whole-brain responses to complex stimuli are critical in the quest to understand the human visual system. The ideal dataset should include naturalistic stimuli: the visual system is distributed widely across the brain, and natural scenes, in addition to being ecologically relevant, are effective activators of the entire system[12]. Moreover, the ideal dataset should be large: to take full advantage of powerful data analysis and machine learning (ML) techniques that have recently become available, we need considerably more data than are currently available. How much? Modern ML methods used in computer vision to process natural scenes (for example, deep convolutional neural networks (CNNs)) require tens to hundreds of thousands of image samples for training[13,14]. A dataset that sampled brain activity at these scales would raise the exciting possibility of exploiting these methods to develop better models of how the brain processes natural scenes[15–20] and would accelerate efforts to bridge cognitive neuroscience and artificial intelligence[21].

In this paper, we present a dataset that achieves sampling at this ambitious scale. The NSD consists of high-resolution (1.8-mm) whole-brain 7T functional magnetic resonance imaging (fMRI) of eight carefully screened human participants who each viewed 9,000–10,000 color natural scenes (22,000–30,000 trials) during 30–40 scan sessions distributed over the course of 1 year. Aggregated across participants, NSD includes responses to 70,566 distinct natural scene images—this is more than an order of magnitude larger than similar datasets involving fMRI sampling of many images[22–24]. Moreover, as we show, the high quality of the NSD dataset makes

[1]Center for Magnetic Resonance Research (CMRR), Department of Radiology, University of Minnesota, Minneapolis, MN, USA. [2]Department of Psychology, University of Minnesota, Minneapolis, MN, USA. [3]Department of Neuroscience, Medical University of South Carolina, Charleston, SC, USA. [4]Graduate Program in Cognitive Science, University of Minnesota, Minneapolis, MN, USA. [5]Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, USA. [6]Department of Neuroscience, Center for Magnetic Resonance Research (CMRR), University of Minnesota, Minneapolis, MN, USA. [7]Department of Neurosurgery, Center for Magnetic Resonance Research (CMRR), University of Minnesota, Minneapolis, MN, USA. [8]National Institute of Mental Health (NIMH), Bethesda MD, USA. [9]Program in Neuroscience, Indiana University, Bloomington IN, USA. [10]Program in Vision Science, Indiana University, Bloomington IN, USA. [11]Department of Psychology, University of Texas at Austin, Austin, TX, USA. [12]Center for Perceptual Systems, University of Texas at Austin, Austin, TX, USA. [13]Institute for Neuroscience, University of Texas at Austin, Austin, TX, USA. [14]Center for Human Brain Health, School of Psychology, University of Birmingham, Birmingham, UK. [15]cerebrUM, Département de Psychologie, Université de Montréal, Montréal QC, Canada. [16]Department of Psychology, University of Oregon, Eugene, OR, USA. [17]Present address: Department of Neuroscience, University of Minnesota, Minneapolis, MN, USA. [18]Present address: Department of Psychology, University of Minnesota, Minneapolis, MN, USA. [19]Present address: Department of Psychology, Harvard University, Cambridge, MA, USA. [20]These authors jointly supervised this work: Thomas Naselaris, Kendrick Kay. ✉e-mail: kay@umn.edu

it possible to leverage the full power of modern ML methods for developing better models of visual representation. Achieving high data quality was afforded, in part, by the use of ultra-high magnetic field strength (7T) to improve signal-to-noise ratio (SNR) over what is attained at lower field strengths[25].

NSD incorporates several innovations in addition to its unprecedented scale and quality. To reconcile extensive sampling with a practical time commitment, we used an aggressive rapid event-related design. This drove the development of new analysis techniques that accurately compensate for the overlap of hemodynamic responses across successive trials. To ensure participant engagement and control cognitive state, we incorporated a continuous recognition task[26] in which participants were instructed to indicate whether they have seen each presented image at any point in the past. In addition to making the experiment tolerable (and even somewhat interesting) for participants, the inclusion of this task makes the NSD, to our knowledge, the longest-term continuous recognition memory fMRI study in history and, thus, a likely source of new insights into long-term memory formation and the cognitive context of vision. Finally, to ensure the broad reach of the NSD dataset, we incorporated design input from a large network of collaborators with diverse scientific interests (for example, low-level vision, high-level vision, memory, connectivity and neuroanatomy) and technical expertise (for example, mapping, multivariate pattern analysis, encoding models, representational similarity analysis and neural network modeling). This input helped precipitate a carefully curated dataset with extensive auxiliary measures.

This paper provides a comprehensive description of the design, acquisition and preparation of the NSD dataset. In particular, we detail the state-of-the-art acquisition and analysis methods that we developed for the dataset and present comprehensive assessments that evidence the high quality of the data. We also present initial analyses of the NSD dataset, demonstrating the feasibility of using data-driven analyses to reveal insights into vision and memory. We expect that the NSD will serve as a valuable resource with widespread application in neuroscience and its intersection with artificial intelligence.

Results

Sampling thousands of images during continuous recognition. We obtained 73,000 color natural scenes from the richly annotated Microsoft Common Objects in Context (COCO) image dataset[14], a dataset that is heavily used in the computer vision and ML communities. Our experimental design specified that each of eight participants would view 10,000 distinct images, and a special set of 1,000 images would be shared across participants (eight participants × 9,000 unique images + 1,000 shared images = 73,000 images). This sampling strategy was chosen to maximize the number of distinct images in the NSD while also facilitating investigations of similarities and differences in brain representations across individuals[27]. Each image would be presented three times to a given participant. Although this is a low number, we reasoned that three trials would be sufficient to produce robust responses given our use of ultra-high field (7T) fMRI. Furthermore, images would be presented using a rapid event-related design consisting of 4-s trials (Fig. 1a). This was done to maximize statistical power and to create an engaging experience for the participants. In addition, the continuous nature of task engagement—in contrast to slow event-related designs and block designs where engagement is likely to fluctuate—helps avoid unwanted respiratory variations[28] and arousal-related confounds[29].

The NSD experiment was split across 40 scan sessions for each participant (Fig. 1b). To control cognitive state and encourage deep processing of the images, participants were instructed to perform a continuous recognition task in which they reported whether the current image had been presented at any previous point in the experiment. We controlled the distributions of image presentations such that both short-term and long-term repetitions were probed (Extended Data Fig. 1a). Parameters were selected such that, even in the first scan session, images were not always new, and, even in the last scan session, images were not always old (Extended Data Fig. 1b).

Neuroimaging data collection on carefully selected participants. All fMRI data in the NSD were collected at 7T using a whole-brain, 1.8-mm, 1.6-s, gradient-echo, echo-planar imaging (EPI) pulse sequence. After verbally screening several potential participants with respect to basic eligibility criteria, we recruited 14 individuals to participate in an initial 7T fMRI screening session that involved population receptive field (pRF)[30] and category functional localizer (fLoc)[31] experiments. Based on data from this scan session, we ranked the 14 participants with respect to data quality. Specifically, we quantified BOLD variance explained in the pRF and fLoc experiments, behavioral performance in the pRF and fLoc experiments and two metrics of head motion, normalized these six measures and then averaged the measures (for details, see 'Rankings from the 7T fMRI screening session' in the Methods). We then invited the top eight individuals to participate in the full NSD experiment (all individuals accepted). This selection process was conducted to ensure the best possible data quality for the NSD. Analyses conducted after completion of the NSD experiment confirm that the ranking procedure successfully identified individuals who yield high-quality data and that data quality would have suffered substantially had we omitted the selection process (Fig. 2c).

Data were collected from the eight NSD participants over the course of 1 year (Fig. 1c). Participants consistently engaged with the task: the average response rate across scan sessions was above 99% for all participants, and the response rate never dropped below 96% in any single scan session. Moreover, all participants exhibited successful recognition performance (Fig. 1d), issuing 'old' responses at a higher rate for previously presented images (blue and orange lines) than for novel images (yellow lines). The full NSD dataset includes a variety of anatomical neuroimaging measures (including T_1, T_2, diffusion, venogram and angiogram), functional neuroimaging measures (including the pRF and fLoc experiments, the NSD experiment, resting-state data and two additional experiments involving synthetic stimuli and visual imagery) and behavioral measures (Fig. 2a,b). In some fMRI sessions, physiological data (ten sessions per participant) and eye-tracking data (2–4 sessions per participant) were also collected. Analysis of the eye-tracking data indicates that participants were able to successfully maintain central fixation most of the time, with some variability in fixation performance across participants (Extended Data Fig. 4). Regarding the core NSD experiment, we completed the full set of 40 NSD scan sessions for four of the participants, but, owing to unforeseen summer absences and scheduled decommissioning of the 7T scanner, we completed 30–32 NSD scan sessions for each of the other participants. A full breakdown of data collection and analysis procedures is provided in Extended Data Figs. 2 and 3.

Stable high-resolution imaging across scan sessions. In our experience, although visual inspection is non-quantitative and somewhat subjective, it is still the most effective way to assess many common aspects of fMRI pre-processing[32]. Accordingly, we generated a comprehensive set of visualizations that detail the excellent quality of the raw and pre-processed NSD data. These include detailed inspections of raw time series data to confirm the presence of stimulus-evoked signals (Supplementary Fig. 3); movies that assess the co-registration of the different imaging modalities (for example, T_1, T_2 and EPI; Supplementary Video 1); movies that assess the manually edited cortical surface reconstructions generated using FreeSurfer (Supplementary Video 2); movies that assess the registration of the NSD participants to the fsaverage (Supplementary Video 3)

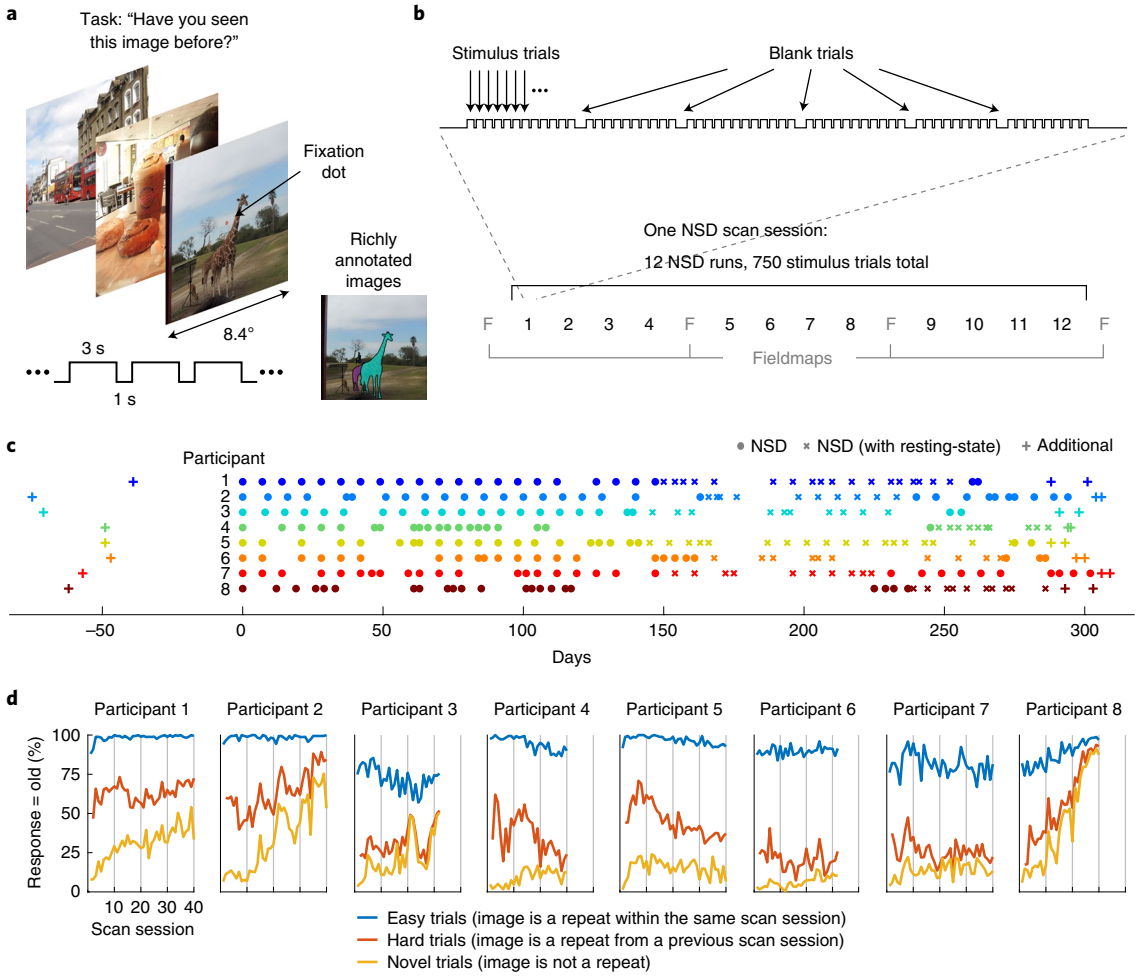


Fig. 1 | Design of the NSD experiment. a, Trial design. While maintaining central fixation, participants viewed sequences of color natural scenes and judged whether each image had been previously shown at any point in the past. The scenes, taken from Microsoft's COCO[14], are richly annotated with object information (as depicted). **b**, Run and session design. Each run lasted 5 min and consisted of 62 or 63 stimulus trials with occasional interspersed blank trials. Each scan session consisted of 12 runs (750 stimulus trials). **c**, Timeline of 7T fMRI scan sessions. Each individual participated in an initial screening session (prffloc), 30–40 NSD core sessions and two final sessions (nsdsynthetic and nsdimagery). The first NSD core session corresponds to day 0. **d**, Behavioral performance. For each of three trial types, we quantify the percentage of trials on which the participant indicated an 'old' response.

and MNI (Supplementary Video 4) group spaces; movies that inspect raw and pre-processed EPI volumes (Supplementary Video 5); and movies that provide volume and surface visualizations of the stability of mean EPI intensity across sessions (Supplementary Videos 6 and 7 and Supplementary Fig. 4) and the stability of BOLD responses across sessions (Supplementary Videos 8 and 9). All movies are readily viewable online (https://osf.io/zyb3t/). The visualizations—in particular, Supplementary Video 9—indicate that the quality of the NSD data enable precision functional mapping[33]: activity patterns are fine-scale and highly reliable within individual participants, and these patterns are distinct across participants.

In addition to visual inspection, quantitative data quality metrics were computed for each NSD scan session. This was in fact done on a rolling basis as the data were acquired, allowing us to monitor data quality and provide performance bonuses to the participants. Inspecting the metrics, we see that temporal signal-to-noise ratio (tSNR) is stable across scan sessions for each participant (Fig. 2d, left). One participant, participant 8, exhibited low tSNR compared to the other participants; this can be attributed to higher levels of head motion for this participant (Fig. 2d, middle). We also observe that BOLD responses (quantified as median variance explained across voxels and runs by a simple ON–OFF general linear model (GLM)) are stable across scan sessions for each participant, although there is substantial variation in the strength of BOLD responses across participants (Fig. 2d, right).

One feature that we implemented in the pre-processing of the fMRI data was to interpolate the data on a fine temporal grid and a fine spatial grid in the same steps used to correct for slice timing differences and spatial displacements (for example, head motion). This upsampling strategy preserves fine-scale detail that is present in the raw fMRI data due to the temporal jitter of the acquired fMRI volumes relative to the experimental paradigm and the spatial jitter of the acquired fMRI volumes relative to the anatomy of the brain[32,34]. An illustration of the benefits of upsampling is provided in Extended Data Fig. 5. This example highlights the existence of fine-scale detail in fMRI image intensities (Extended Data Fig. 5b, top row) as well as in BOLD responses extracted from the fMRI data (Extended Data Fig. 5b, bottom row, and Extended Data Fig. 5c). Notably, this fine-scale detail is replicable across different scan sessions (Extended Data Fig. 5c, bottom, and Extended Data Fig. 5d), indicating that the upsampled preparation reveals meaningful detail that is lost under a non-upsampled approach.

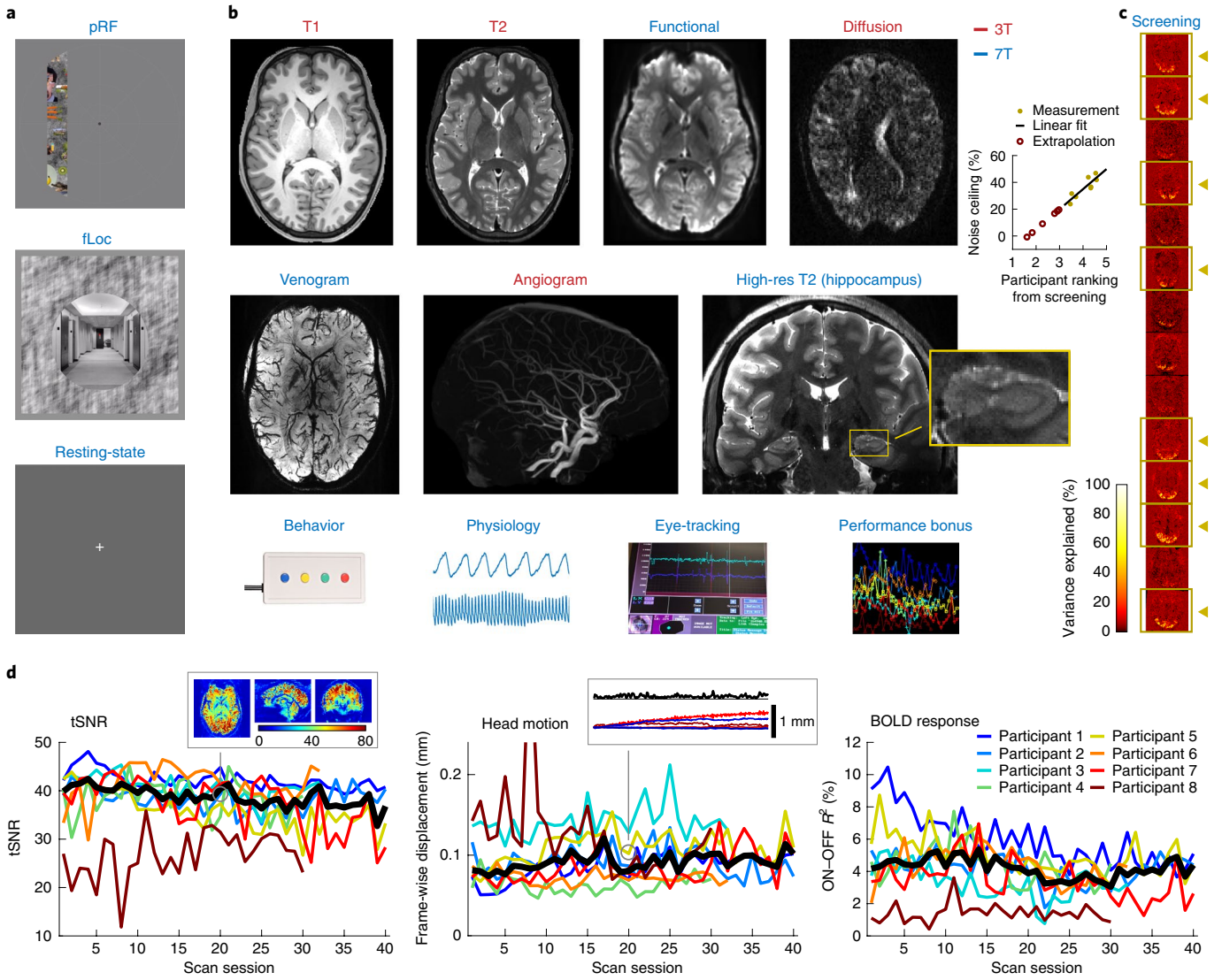


Fig. 2 | Overview of acquired data. a, Auxiliary fMRI experiments. Data from the pRF and fLoc experiments were used to define retinotopic visual areas and category-selective regions, respectively. Resting-state data were collected before and after the NSD runs in a subset of the NSD core sessions (totaling 100 or 180 min per participant). **b**, Available measures. Examples of the actual data are depicted. **c**, Participant selection. Data quality from the initial screening session was used to rank a set of 14 participants. On the right is an illustration of one measure contributing to the ranking—specifically, variance explained in the fLoc experiment (one slice per participant; identical color range). The inset compares the participant ranking against the b3 noise ceiling calculated on the full NSD dataset (Fig. 3). A line fit to the eight NSD participants (gold dots) is extrapolated to predict noise ceilings for the individuals who were not selected for participation in the NSD (red circles). **d**, Metrics of data quality (for details, see 'Data quality metrics' in the Methods). Results for individual participants (thin colored lines) and the median across participants (thick black line) are shown. The insets show detail on tSNR and head motion for one sample run (see Supplementary Figs. 1 and 2 for more information).

Extensive auxiliary measures to complement the NSD data. To enrich the fMRI data from the NSD experiment, we collected and prepared a large set of auxiliary measures. These measures include substantial amounts of resting-state data (minimum 100 min per participant), external physiological measures during the resting-state scan sessions, diffusion data and associated derivatives (white-matter tracts and structural connectivity matrices) and an extensive collection of manually defined regions of interest (ROIs), including retinotopic and category-selective areas as well as subregions of the thalamus and medial temporal lobe. Results and discussion of these resources can be found in Supplementary Note 1, Extended Data Figs. 6 and 7 and Supplementary Fig. 5.

Accurate estimation of single-trial fMRI response amplitudes. We performed a GLM analysis of the data from the NSD experiment to help streamline subsequent analyses of the data. The goal of the GLM was to obtain single-trial betas—that is, estimates of the fMRI response amplitude of each voxel to each trial conducted. Given the low SNR of fMRI and the overlap of the hemodynamic response from trial to trial, estimating accurate betas is a challenging endeavor. We thus developed a novel GLM approach consisting of three components. First, we used a library of hemodynamic response functions (HRFs) derived from an initial analysis of the dataset as an efficient and well-regularized method for estimating voxel-specific HRFs (Fig. 3a–c). Second, we adapted the GLMdenoise technique[35] to the single-trial GLM framework, thereby enabling the use of data-driven nuisance regressors (Fig. 3d). Third, to address the challenge posed by highly correlated single-trial regressors, we developed an efficient implementation of ridge regression[36] and used this to regularize and improve the accuracy of the betas

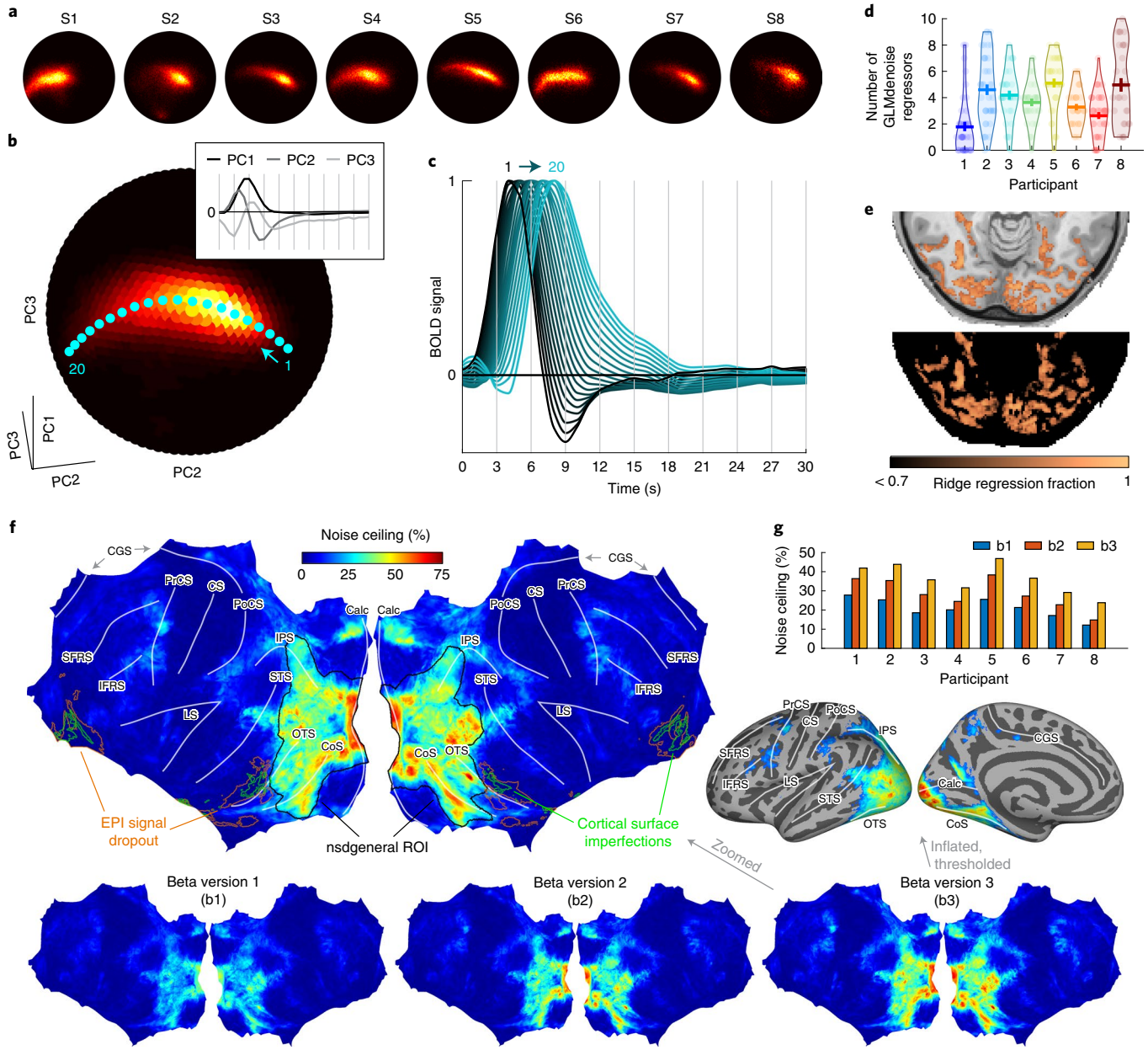


Fig. 3 | Improving SNR through novel response estimation and denoising methods. a–c, Library of HRFs. HRFs were estimated within a subspace spanned by three PCs. Distributions of voxel-specific HRFs are shown for individual participants (**a**) and the group average (**b**). These distributions reside on the unit sphere with coordinate axes corresponding to three PC time courses (**b**, inset). We defined a series of points on the unit sphere (cyan dots), and the time courses associated with these points are used as the HRF library (**c**). **d**, GLMdenoise. Horizontal lines indicate the average number of GLMdenoise regressors identified in a scan session (1.8-mm preparation; error bars indicate bootstrapped 68% confidence intervals). **e**, Ridge regression. Optimal ridge regression fractions are shown for an example scan session (participant 5, nsd10, 1-mm preparation). **f**, Noise ceilings for the case where responses are averaged across three trials. Results from individual participants (nativesurface preparation) were mapped to fsaverage and then averaged. Right inset shows results thresholded at 15% on the inflated left hemisphere (see also Supplementary Video 10). **g**, Performance summary. Each bar indicates the median noise ceiling across vertices in the nsdgeneral ROI. Calc, calcarine sulcus; CGS, cingulate sulcus; CoS, collateral sulcus; CS, central sulcus; IFRS, inferior frontal sulcus; IPS, intraparietal sulcus; LS, lateral sulcus; OTS, occipitotemporal sulcus; PoCS, post-central sulcus; PrCS, precentral sulcus; SFRS, superior frontal sulcus; STS, superior temporal sulcus.

(Fig. 3e). To assess the efficacy of these various GLM techniques, we generated three versions of the betas, reflecting increasing sophistication (Extended Data Fig. 8a–c). Beta version 1 (b1) is the result of simply using a canonical HRF for all voxels. Beta version 2 (b2) is the result of fitting an HRF to each voxel using the library-of-HRFs approach. Beta version 3 (b3) uses the library-of-HRFs approach as with b2 but also adds the use of GLMdenoise and ridge regression in an attempt to improve the accuracy of the betas.

We quantified the quality of the different beta versions (b1, b2 and b3) by calculating noise ceilings for individual voxels. The noise ceiling is a measure of trial-to-trial reliability, quantifying the percentage of variance in a voxel's responses that can be attributed to the stimulus and not to measurement noise (Methods). Surface maps of noise ceiling results reveal locations of reliable responses to the NSD stimuli: high noise ceilings are present in occipital cortex and extend into temporal and parietal cortex (Fig. 3f and

Supplementary Video 10). Notably, the maps reveal very large increases in noise ceilings from b1 to b2 to b3, indicating that the additional GLM techniques incorporated into b2 and b3 improve reliability of responses. Detailed quantifications show that these improvements are highly consistent across voxels and participants (Fig. 3g and Supplementary Fig. 6a) and that noise ceiling estimates are highly reliable (Supplementary Fig. 6b). For b3, the noise ceiling levels in visual cortex are, on average, 36% (calculated by computing the median across the nsdgeneral ROI and then averaging across participants). This means that a typical visual cortex voxel in the NSD dataset has associated with it a set of 10,000 responses (30,000 trials divided by 3 trials per image = 10,000 images), and a large percentage, 36%, of the variance in these 10,000 values is a signal that is, in theory, predictable. Expressed in terms of Pearson's correlation (r), this is equivalent to a prediction accuracy of $r = 0.60$. Complementing the noise ceiling analysis, we also performed simple univariate analyses of the NSD betas (Extended Data Fig. 8d,e); these analyses show that the NSD dataset contains high response reliability across trials within a participant as well as high response reliability across participants.

A massive increase in equivalent trials. To put the quality of the NSD data into perspective, we propose the concept of 'equivalent trials', which allows comparison of different datasets that vary in SNR and trial distribution (see Methods for details). The next largest data collection effort that is similar in nature to NSD is BOLD5000 (ref. [22]). Using the same GLM analysis methods on both NSD and BOLD5000, we found that the SNR per trial is approximately 0.260 for the NSD and 0.187 for BOLD5000. Combining these values with the number of trials conducted in each dataset, we estimate that the total size of the NSD dataset is 213,000 trials $\times$ $(0.260)^2 = 14{,}399$ equivalent trials, whereas the total size of BOLD5000 is 18,870 trials $\times$ $(0.187)^2 = 660$ equivalent trials. Thus, using the metric of equivalent trials, the NSD can be viewed as 14,399/660 = ~22 times as large as the BOLD5000 dataset. This is a massive increase in statistical power. Note that even if we do not take into account the higher SNR per trial in the NSD dataset, the NSD still has substantially more participants (eight versus four), more trials per participant (26,625 versus 4,718, on average) and more hours of fMRI per participant (35.5 versus 13.7, on average) than BOLD5000.

Successful recovery of retinotopy. Having demonstrated the quality of the NSD data, we now turn to example analyses that illustrate the rich scientific insights that can be derived from the data. As a simple starting example, we fit a voxel-wise pRF model that uses local contrast in the NSD images to account for the NSD betas. This simple model is expected to recover spatial tuning in early visual cortex where responses co-vary with stimulus energy[37]. Indeed, in all eight participants, high-quality maps of angle and eccentricity estimates are obtained in early visual cortex, and these estimates extend all the way to the fovea (Extended Data Fig. 9 and Supplementary Modeling Note 1). These results provide a check of the validity of the NSD betas. They also show that participants were able to maintain central fixation reliably enough to support detailed mapping of visual space. This finding is consistent with our analysis of the eye-tracking data (Extended Data Fig. 4).

Reliable and long-term recognition memory effects. The use of a continuous recognition task establishes the NSD as one of the largest datasets relevant to human memory. Despite the challenging nature of the task, we found that participants were able to successfully discriminate old images from new images (average d' across participants: 1.28, maximum: 1.47, minimum: 0.94). Furthermore, recognition memory remained above chance even at long time scales between repetitions (Fig. 4a). Specifically, for each session, we calculated a measure of recognition accuracy accounting for guessing (adjusted hit rate: hit rate minus false alarm rate) and binned this measure by the time since last exposure (considering only those trials involving a previously shown image). At the group level, participants exhibited performance levels greater than chance (adjusted hit rate > 0) in all measured intervals, ranging from 1 s to 1 year. At the level of individuals, all participants showed a positive adjusted hit rate in the longest time bin for which data are available for every participant (when binning on a log scale; seven of eight participants when binning on a linear scale). These results indicate that, from its behavioral component alone, NSD is powered to address questions concerning human memory spanning short (seconds) to relatively long (months) time scales.

But what about neural effects? To assess whether recognition effects are present in the fMRI data, we performed two-sample t-tests contrasting NSD betas observed for hits with NSD betas observed for correct rejections (the so-called 'old/new effect'[38]). We found highly consistent old/new effects at the level of individual scan sessions (Fig. 4b, top; see also Supplementary Fig. 7). Moreover, these effects occur in expected frontal and parietal regions[39] and persist at the group level (Fig. 4b, bottom). The scale and statistical power afforded by the NSD dataset also provide additional insight. Whereas old/new effects are typically studied using group-level analyses, the quality of the NSD dataset reveals highly statistically significant results at the level of individual participants. Indeed, when pooling trials across all NSD scan sessions, several participants exhibited statistically significant activity differentiating hits and correct rejections in nearly the entire cerebral cortex (see results for a representative participant in Fig. 4b, top). Reminiscent of past datasets employing extensive sampling of individuals[40], the current results suggest that the extent of cortex engaged by basic memory processes is much more widespread than previously appreciated, although a careful consideration of effect sizes would be important for a full understanding of the effect.

Rich stimulus sampling for probing brain representations. The NSD samples a large variety of natural scenes. To gain insight into the breadth of stimulus sampling available, we constructed representational dissimilarity matrices (RDMs) from the NSD betas and performed t-distributed stochastic neighbor embedding[41] (t-SNE) to visualize the underlying representations. We computed t-SNE embeddings in different regions along the ventral visual pathway for an example participant (Fig. 5a). These embeddings reflect arrangements of stimuli that are driven by the overall similarity of multi-voxel activity patterns in the brain, independent of their anatomical organization within a given ROI. Visualizing the data in this way reveals intriguing patterns of semantic representation that are clearly visible by eye. For example, by color-coding the resulting embeddings according to animacy attributes (Fig. 5b), we found that, in posterior ventral temporal cortex (pVTC), there is a clear large-scale pattern progressing from images containing people (gray dots, lower left), images containing animals (red dots, middle) and images containing inanimate objects (blue dots, upper right), whereas the pattern is not present in early visual areas V1, V2 and V3. This aspect of semantic representation is consistent with previous studies[42,43].

Other intriguing patterns are also visible. In anterior ventral temporal cortex (aVTC), the animacy progression is present to some extent, but a different, more clustered representation emerges that presumably reflects more complex categorical and semantic clusters. Indeed, zooming in on small sections of the t-SNE embedding for aVTC reveals that these clusters contain images with relatively homogeneous semantic content (Fig. 5c): the blue cluster is dominated by images of round edible objects, whereas the gray cluster is dominated by images of people interacting with objects. Note that the clustering of semantically related images does not necessarily mean that these representations are truly semantic

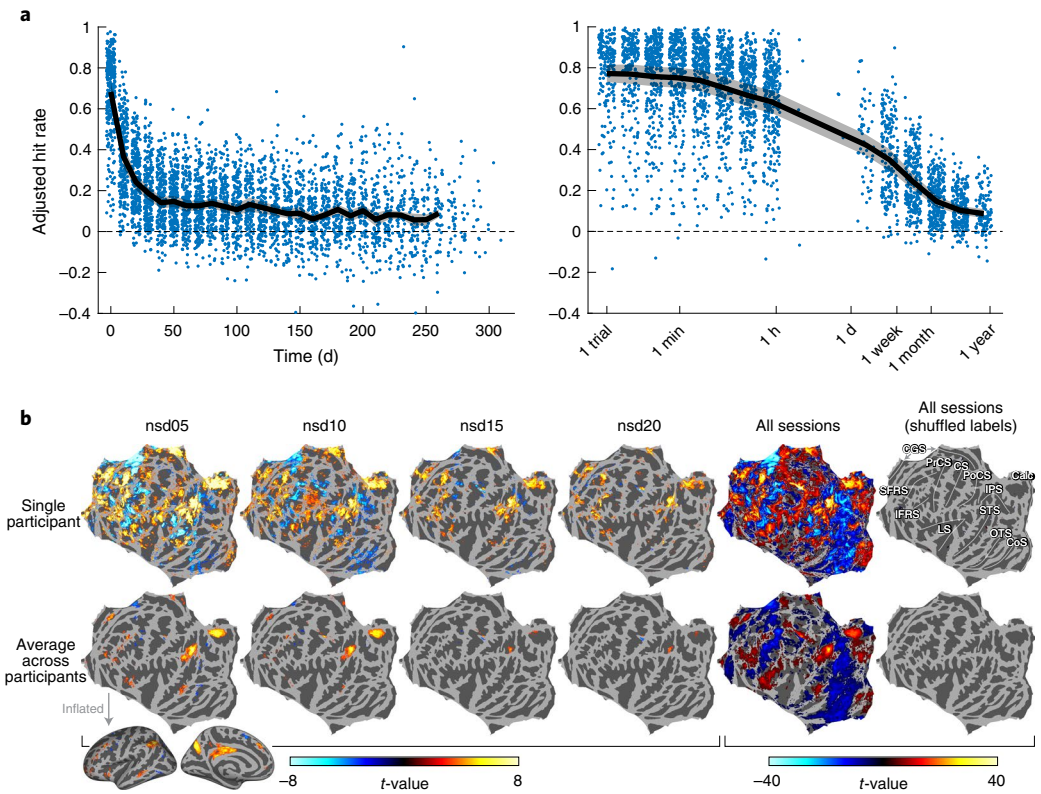


Fig. 4 | Reliable and long-term recognition memory effects. a, Behavioral recognition effects. Adjusted hit rate indicates recognition accuracy accounting for guessing (hit rate minus false alarm rate) and is binned by time between repetitions on a linear scale (left) or a log scale (right). Dashed line indicates chance performance. Each dot in each bin summarizes relevant trials from one scan session. Black line indicates the mean across participants, with the ribbon indicating ± 1 s.e.m. **b**, Neural recognition effects. We performed two-sample *t*-tests on NSD betas contrasting 'hits' > 'correct rejections'. All results are shown on a flattened left hemisphere fsaverage surface and thresholded at $|t| > 3$ (inset shows inflated surface). Tests were performed for trials taken from individual NSD scan sessions (columns 1–4) as well as for trials pooled across all NSD scan sessions (column 5). In addition, we performed a control in which trial labels in the pooled analysis were shuffled (column 6). Results for participant 1 (top row) and a simple average of results across participants (bottom row) are shown. Calc, calcarine sulcus; CGS, cingulate sulcus; CoS, collateral sulcus; CS, central sulcus; IFRS, inferior frontal sulcus; IPS, intraparietal sulcus; LS, lateral sulcus; OTS, occipitotemporal sulcus; PoCS, post-central sulcus; PrCS, precentral sulcus; SFRS, superior frontal sulcus; STS, superior temporal sulcus.

in the sense of being invariant or independent of visual features; the clustering could be driven by certain visual features that are diagnostic of object categories[44]. To tease apart these possibilities, additional detailed analyses would be necessary. Overall, these findings show how simple visual inspections of the NSD dataset can be used to generate hypotheses about visual representations in the human brain.

To further characterize brain representations using a quantitative analysis, we calculated how well brain RDMs are captured by a model RDM constructed from category labels in the COCO image dataset. Consistent with the clustering observed in the *t*-SNE embeddings, we found that categorical structure is pronounced in VTC compared to early visual areas (Fig. 5d). Finally, to assess the utility of the NSD for investigating similarities of brain representations across participants, we isolated images that were common across participants and created a second-order RDM that quantifies the similarity of brain RDMs across ROIs and participants (Fig. 5e). In this second-order RDM, we observed high levels of consistency in each ROI's representation across participants (red outlines). We also observed distinct representations across ROIs, with the largest distinctions occurring between early visual areas and VTC. One noticeable finding is the existence of strong off-diagonal elements (white arrows); these elements indicate spatial noise correlations that are typical in fMRI and other neural measurement techniques. To counteract these noise correlations, one simple approach is to compare representations across ROIs using data from distinct trials[45]. To further summarize the second-order RDM, we computed the average correlation of brain RDMs across all ROI pairs, restricting this calculation to distinct participants to avoid the effects of spatial noise correlations (Fig. 5f). We observe that correlations are highest for brain RDMs from the same ROI (for example, a given participant's V1 RDM is more correlated with other participants' V1 RDMs compared to other ROIs), confirming consistencies in brain representations across participants (for a complementary univariate analysis of across-participant consistency, see Extended Data Fig. 8d,e).

A brain-optimized neural network model of the visual system. One of the main motivations for the NSD was to amass sufficient sampling of brain activity to be able to drive data-hungry ML techniques. As an intriguing test case, we specifically investigated whether we could successfully use the scale of the NSD to train, from scratch, a deep CNN to accurately predict brain activity[17]. Adopting the framework of encoding models[46], we took NSD betas from visual areas V1–hV4, divided these data into a training set (used for parameter tuning) and a validation set (used to assess prediction performance) and evaluated how accurately different computational models predict brain responses in the validation set based

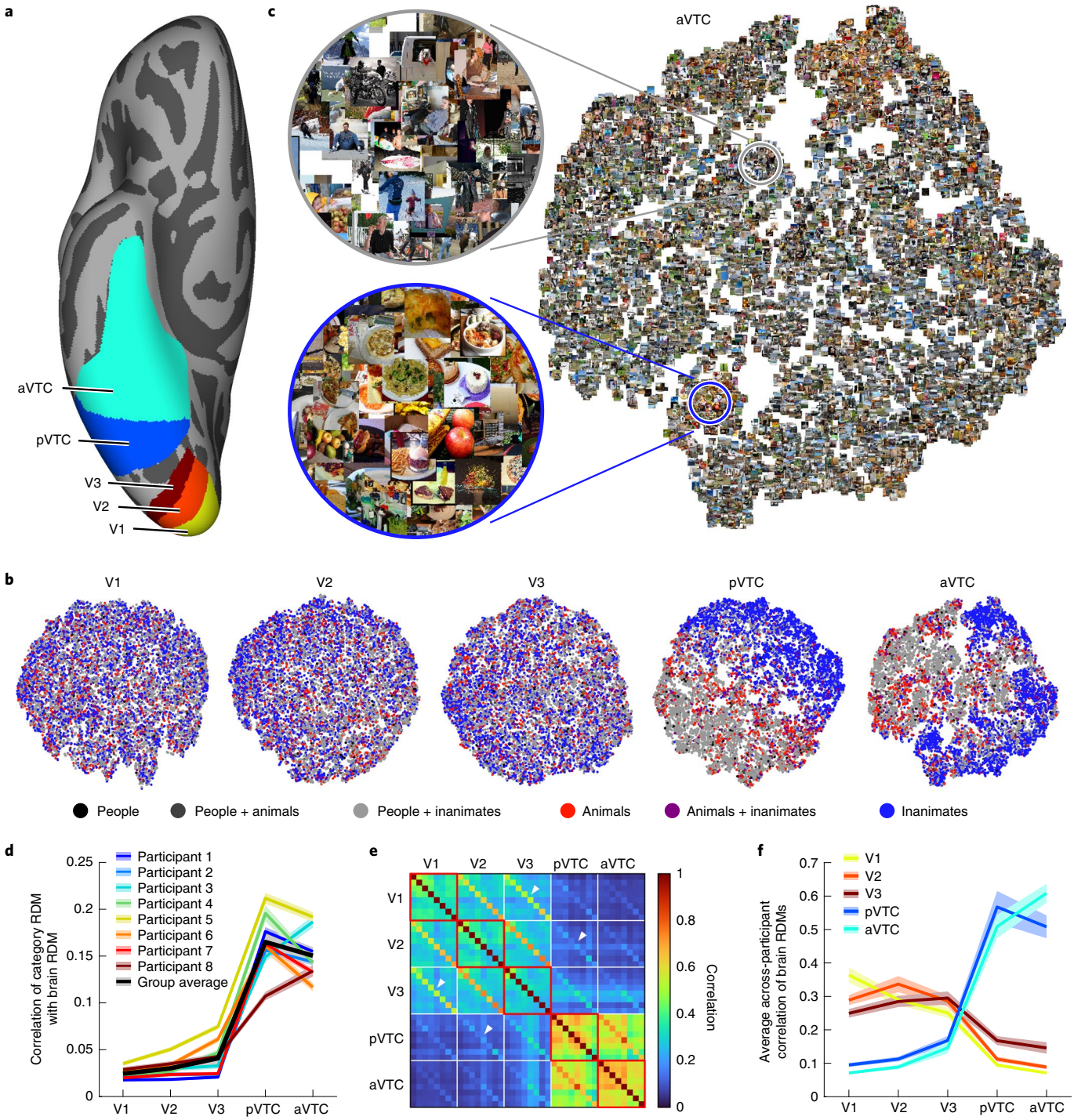


Fig. 5 | Representational similarity analysis reveals transformations of representations along the ventral visual stream. a, Illustration of fsaverage ROIs used for the representational similarity analysis. **b**, *t*-SNE embedding for each ROI in an example participant (participant 1). Each dot represents a distinct image (total, 10,000). Using category labels from the COCO image dataset, we color each dot according to whether the associated image contains particular combinations of people, animals and inanimates. **c**, *t*-SNE embedding for aVTC with actual images depicted. Insets highlight an inanimate cluster (blue inset) and a cluster of people with inanimate objects (gray inset). **d**, Categorical brain repesentations. We plot the correlation between brain RDMs and a model RDM constructed from category labels in the COCO dataset. Color-shaded regions indicate within-participant error (mean and standard error across distinct groups of images), whereas the gray-shaded region indicates across-participant error (mean and standard error across participants). **e**, Similarities of brain representations across ROIs and participants. Depicted are correlations across brain RDMs obtained for different ROIs and participants. Thin white lines separate groups of eight participants. **f**, Quantitative summary. We summarize the results of **e** by averaging the upper triangle of each group of 8 × 8 participants, reflecting the correlation of RDMs from different participants. Shaded regions indicate standard errors estimated by bootstrapping participants with replacement.

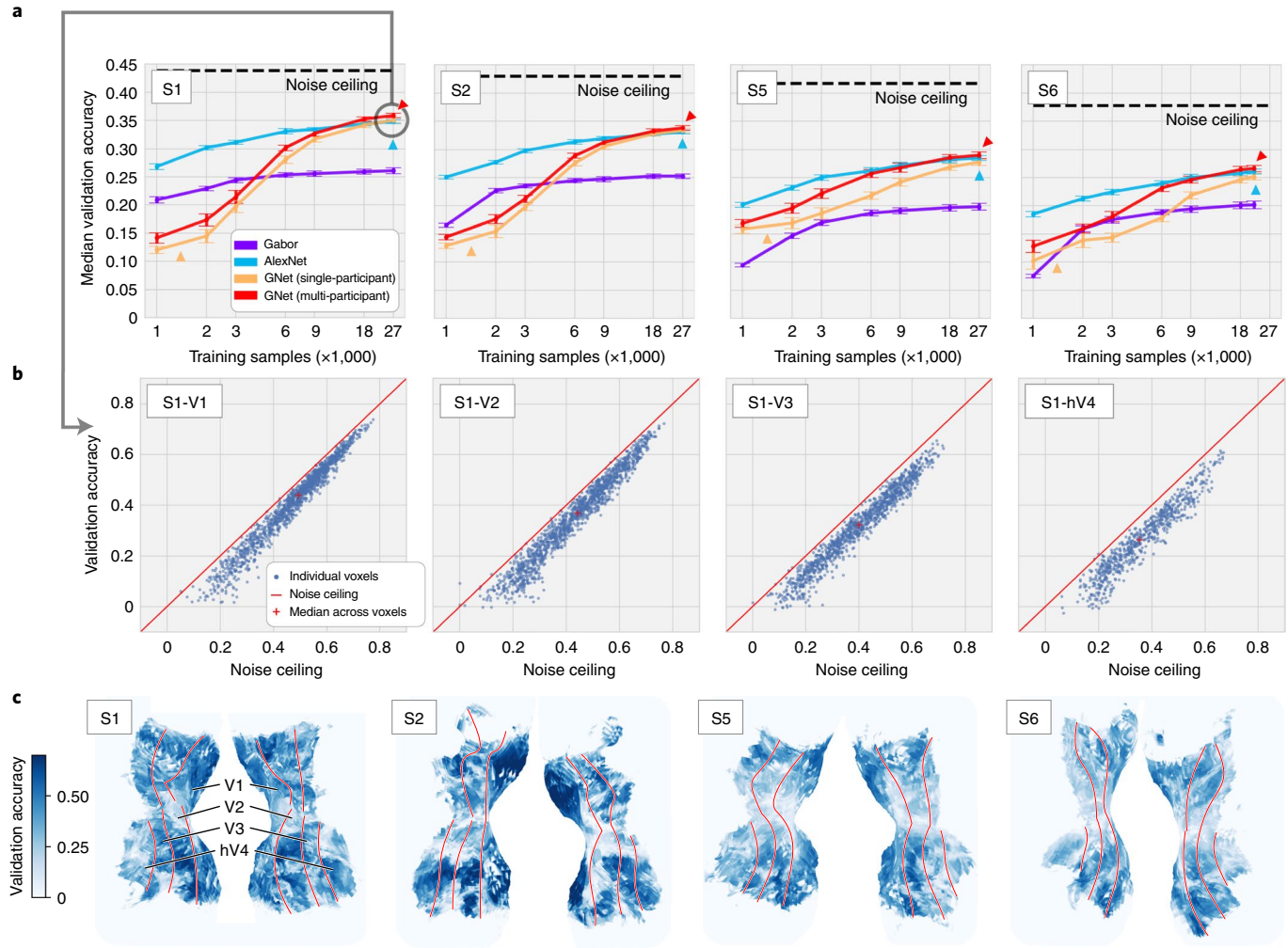


Fig. 6 | Prediction of brain activity using a brain-optimized neural network. We used encoding models[46] to predict voxel activity in V1–hV4. NSD betas were divided into a training set (consisting of up to 9,000 images × 3 trials = 27,000 training samples per participant) and a validation set (consisting of up to 1,000 images × 3 trials = 3,000 validation samples per participant), and the accuracy of different encoding models was quantified as the voxel-wise correlation between model predictions and responses observed in the validation set. **a**, Performance as a function of amount of training data used. Models include an encoding model based on AlexNet, which is a task-optimized neural network (blue); encoding models based on GNet, which is a brain-optimized neural network trained using data from single participants (orange) or data from multiple participants (red); and a V1-like control model based on Gabor filters (purple). Plotted lines and error bars indicate mean and standard deviation across results obtained from different bootstrap samples of the data. **b**, Detailed view of the performance of the multi-participant GNet model for a representative participant. **c**, Surface maps depicting spatial distribution of validation accuracy for the multi-participant GNet model.

on the presented image. The primary encoding model of interest is based on a new network that we refer to as 'GNet', a brain-optimized CNN whose parameters are trained using image–response pairings observed in the training set. For comparison, we also evaluated an encoding model based on AlexNet[47], a task-optimized CNN whose parameters are pre-trained using explicit labels of objects taken from an image database. AlexNet has been previously shown to provide state-of-the-art performance in modeling visual responses[15,19]. Finally, we included a simple V1-like control model based on oriented Gabor filters[24]. Details of modeling procedures are provided in Supplementary Modeling Note 2 and Extended Data Fig. 10.

Varying the amount of training data provided to the models, we found that the performance of the GNet-based encoding model is relatively poor when only small amounts of training data are available (Fig. 6a, orange arrows). This is expected because the feature extractors in GNet are not pre-trained and thus require data for tuning. However, when large amounts of training data are available, the GNet model exhibits an impressive increase in performance, achieving approximate parity with the AlexNet-based encoding model (Fig. 6a, blue arrows). Interestingly, when we trained a single GNet model using brain activity from multiple participants, we found that the model was able to outperform the AlexNet model (two-tailed paired *t*-test across participants, $P = 0.013$), albeit modestly (Fig. 6a, red arrows). Noticeably, the simple Gabor model accounts for substantial variance in the responses; nonetheless, the more complex CNN-based models provide additional predictive power, consistent with previous observations[48]. For additional insight into model performance, we compared voxel-wise performance levels of the GNet model to noise ceiling estimates (Fig. 6b). Across voxels, prediction accuracy is tightly correlated with the noise ceiling, suggesting that voxel-wise differences in prediction accuracy simply reflect differences in SNR. In addition, performance levels are close to, but do not reach, the noise ceiling. Finally, cortical surface maps indicate that voxel-wise performance levels vary across foveal and peripheral representations (Fig. 6c).

The demonstration that an encoding model based on a brain-optimized CNN (GNet) outperforms an encoding model based on a task-optimized CNN (AlexNet) is important for two reasons. First, it indicates that the NSD is large enough to successfully train a complex neural network architecture. Had the NSD dataset been smaller in scale or lower in quality, qualitatively different patterns of model performance would have been obtained (in Fig. 6a, compare orange arrows reflecting a few thousand trials to red arrows reflecting tens of thousands of trials). Second, the successful training of a brain-optimized CNN opens the possibility of new avenues of investigation into the nature of the features used in CNNs. It is an interesting open question whether the features learned by task-optimized networks like AlexNet are similar to, or diverge from, the features present in brain-optimized networks like GNet. In general, brain-optimized networks[17] are a useful alternative to task-optimized networks[16,20], as the narrowly defined tasks that task-optimized networks are typically trained to solve do not necessarily respect the diversity of functions supported by the human visual system[49] nor necessarily match properties found in biological visual systems[50].

Discussion

In the last several years, several large-scale neuroimaging datasets have been made publicly available for re-use (for example, refs. [5,33,51–53]). Several distinguishing aspects of the present work sets the NSD apart from past datasets. One is the unprecedented scale of the dataset. The NSD shares the motivation of recent 'deep' (or 'precision') neuroimaging efforts[33,54–57] that are seeking to amass large amounts of data from individual subjects, as opposed to modest amounts of data from a large number of subjects. In this context of deep neuroimaging, the NSD is, to our knowledge, the most extensive fMRI data collection effort that has been performed to date. This can be gauged not only in terms of the number of hours of fMRI data acquisition per participant (30–40 h of data for each of eight participants on the core NSD experiment) and the high spatial resolution of the acquired data (1.8 mm) but also the wealth of additional measures beyond the core experiment, including substantial amounts of resting-state and diffusion data, physiological data and functional localizers. The availability of extensive measures provides the opportunity to build complete models of how individual brains support vision and memory[58]. Of course, the emphasis on depth in individuals comes at the cost of sampling fewer individuals; datasets emphasizing large numbers of individuals, such as the Human Connectome Project[5], are better suited for studying variability in the general population and how psychological traits broadly relate to brain structure and function.

A second aspect is the unusually high quality of the data. Although the quality of neuroimaging data is more complex to assess than quantity, assessment of data quality is essential because MRI data have relatively low sensitivity and are prone to errors and artifacts. In particular, when acquiring massive datasets, there is a risk of accumulating unknown sources of noise and artifact. The work presented in this paper (and in the accompanying files in the data release) guards against this possibility by crafting a customized and highly optimized approach to pre-processing the NSD data and providing comprehensive documentation of the high data quality (see also Supplementary Note 2). Several factors likely contributed to the high data quality. These include (1) the use of ultra-high magnetic field strength (7T), which enhances BOLD contrast-to-noise ratio; (2) the screening, training and incentivization of participants; (3) the detailed inspection and supervision of data processing; and (4) the large network of collaborators who helped guide the design and trajectory of the dataset.

A third aspect of the present work lies in the novel analysis techniques developed for improved GLM analysis of fMRI time series data. These include (1) an efficient and robust method to estimate voxel-specific HRFs; (2) adaptation of the GLMdenoise technique[35] to a single-trial GLM framework; and (3) development of ridge regression as an effective method for regularizing single-trial response estimates. These three techniques have been integrated into a toolbox that can be applied to other neuroimaging datasets and are the subject of a forthcoming paper. An important lesson stemming from our results is that well-executed data collection is important but not the only factor to consider: data preparation methods exert a major influence on the quality of a dataset and, hence, its scientific value. One can view improvements in data quality as equivalent to increases in data quantity, in the sense that analysis methods that reduce unwanted variability (noise) can be interpreted as increasing the effective amount of data collected[35]. Thus, by improving data quality, the methods introduced with the NSD are contributing to the massive scale of the dataset.

The NSD dataset has many potential applications. Given its extensive sampling of natural scenes (70,566 distinct images aggregated across eight participants) and high SNR, the dataset will be useful for investigating a variety of phenomena in low-, mid- and high-level vision. In addition, the memory component of the NSD experiment provides a unique opportunity to study the neural mechanisms of both short-term and long-term memory (ranging from seconds to many months) as well as potential interactions between vision and memory. From a methodological perspective, the repeated scanning of individuals using a consistent experimental manipulation (up to 40 scan sessions of the NSD experiment per participant) provides a unique opportunity for development and evaluation of neuroimaging pipelines. Finally, perhaps the most exciting use of the NSD is as a common dataset to bridge the disciplines of cognitive science, neuroscience and artificial intelligence[21]. As we have shown in the context of deep neural network modeling (Fig. 6), there are sufficient data in the NSD to successfully drive the training of neural network models with thousands of free parameters. This demonstration exemplifies how the NSD—with its large amounts of carefully curated fMRI data collected during a rich cognitive paradigm—enables data-driven approaches toward understanding the complexities of information processing in the brain.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-021-00962-x.

Received: 1 March 2021; Accepted: 12 October 2021;
Published online: 16 December 2021

References

1. de Vries, S. E. J. et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.* **23**, 138–151 (2020).
2. Siegle, J. H. et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592**, 86–92 (2021).
3. Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M. & Harris, K. D. High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
4. Markram, H. et al. Reconstruction and simulation of neocortical microcircuitry. *Cell* **163**, 456–492 (2015).
5. Van Essen, D. C. et al. The WU-Minn human connectome project: an overview. *Neuroimage* **80**, 62–79 (2013).
6. Zheng, Z. et al. A complete electron microscopy volume of the brain of adult *Drosophila melanogaster*. *Cell* **174**, 730–743 (2018).
7. Van Essen, D. C. et al. Mapping visual cortex in monkeys and humans using surface-based atlases. *Vis. Res.* **41**, 1359–1378 (2001).
8. Grill-Spector, K. & Malach, R. The human visual cortex. *Annu. Rev. Neurosci.* **27**, 649–677 (2004).

9. Wheeler, M. E., Petersen, S. E. & Buckner, R. L. Memory's echo: vivid remembering reactivates sensory-specific cortex. *Proc. Natl Acad. Sci. USA* **97**, 11125–11129 (2000).
10. Breedlove, J. L., St-Yves, G., Olman, C. A. & Naselaris, T. Generative feedback explains distinct brain activity codes for seen and mental images. *Curr. Biol.* **30**, 2211–2224 (2020).
11. Kay, K. N., Weiner, K. S. & Grill-Spector, K. Attention reduces spatial uncertainty in human ventral temporal cortex. *Curr. Biol.* **25**, 595–600 (2015).
12. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
13. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf (University of Toronto, 2009).
14. Lin, T.-Y. et al. Microsoft COCO: Common Objects in Context. European Conference on Computer Vision. https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48, 740–755 (Springer, 2014).
15. Güçlü, U. & van Gerven, M. A. J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).
16. Khaligh-Razavi, S.-M. & Kriegeskorte, N. Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput. Biol.* **10**, e1003915 (2014).
17. Seeliger, K. et al. End-to-end neural system identification with neural information flow. *PLoS Comput. Biol.* **17**, e1008558 (2021).
18. Stansbury, D. E., Naselaris, T. & Gallant, J. L. Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron* **79**, 1025–1034 (2013).
19. St-Yves, G. & Naselaris, T. The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *Neuroimage* **180**, 188–202 (2018).
20. Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
21. Naselaris, T. et al. Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends Cogn. Sci.* **22**, 365–367 (2018).
22. Chang, N. et al. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci. Data* **6**, 49 (2019).
23. Horikawa, T. & Kamitani, Y. Generic decoding of seen and imagined objects using hierarchical visual features. *Nat. Commun.* **8**, 15037 (2017).
24. Kay, K. N., Naselaris, T., Prenger, R. J. & Gallant, J. L. Identifying natural images from human brain activity. *Nature* **452**, 352–355 (2008).
25. Triantafyllou, C. et al. Comparison of physiological noise at 1.5 T, 3 T and 7 T and optimization of fMRI acquisition parameters. *Neuroimage* **26**, 243–250 (2005).
26. Brady, T. F., Konkle, T., Alvarez, G. A. & Oliva, A. Visual long-term memory has a massive storage capacity for object details. *Proc. Natl Acad. Sci. USA* **105**, 14325–14329 (2008).
27. Haxby, J. V., Guntupalli, J. S., Nastase, S. A. & Feilong, M. Hyperalignment: modeling shared information encoded in idiosyncratic cortical topographies. *eLife* **9**, e56601 (2020).
28. Power, J. D., Lynch, C. J., Adeyemo, B. & Petersen, S. E. A critical, event-related appraisal of denoising in resting-state fMRI studies. *Cereb. Cortex* **30**, 5544–5559 (2020).
29. Roth, Z. N., Ryoo, M. & Merriam, E. P. Task-related activity in human visual cortex. *PLoS Biol.* **18**, e3000921 (2020).
30. Benson, N. C. et al. The human connectome project 7 Tesla retinotopy dataset: description and population receptive field analysis. *J. Vis.* **18**, 23 (2018).
31. Stigliani, A., Weiner, K. S. & Grill-Spector, K. Temporal processing capacity in high-level visual cortex is domain specific. *J. Neurosci.* **35**, 12412–12424 (2015).
32. Kay, K. et al. A critical assessment of data quality and venous effects in sub-millimeter fMRI. *Neuroimage* **189**, 847–869 (2019).
33. Gordon, E. M. et al. Precision functional mapping of individual human brains. *Neuron* **95**, 791–807 (2017).
34. Kang, X., Yund, E. W., Herron, T. J. & Woods, D. L. Improving the resolution of functional brain imaging: analyzing functional data in anatomical space. *Magn. Reson. Imaging* **25**, 1070–1078 (2007).
35. Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F. & Wandell, B. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. *Front. Neurosci.* **7**, 247 (2013).
36. Rokem, A. & Kay, K. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *Gigascience* **9**, giaa133 (2020).
37. Albrecht, D. G. & Hamilton, D. B. Striate cortex of monkey and cat: contrast response function. *J. Neurophysiol.* **48**, 217–237 (1982).
38. Wagner, A. D., Shannon, B. J., Kahn, I. & Buckner, R. L. Parietal lobe contributions to episodic memory retrieval. *Trends Cogn. Sci.* **9**, 445–453 (2005).
39. Spaniol, J. et al. Event-related fMRI studies of episodic encoding and retrieval: meta-analyses using activation likelihood estimation. *Neuropsychologia* **47**, 1765–1779 (2009).
40. Gonzalez-Castillo, J. et al. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proc. Natl Acad. Sci. USA* **109**, 5487–5492 (2012).
41. Maaten, Lvander & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. Connolly, A. C. et al. The representation of biological classes in the human brain. *J. Neurosci.* **32**, 2608–2618 (2012).
43. Naselaris, T., Stansbury, D. E. & Gallant, J. L. Cortical representation of animate and inanimate objects in complex natural scenes. *J. Physiol. Paris* **106**, 239–249 (2012).
44. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl Acad. Sci. USA* **115**, E9015–E9024 (2018).
45. Henriksson, L., Khaligh-Razavi, S.-M., Kay, K. & Kriegeskorte, N. Visual representations are dominated by intrinsic fluctuations correlated between areas. *Neuroimage* **114**, 275–286 (2015).
46. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *Neuroimage* **56**, 400–410 (2011).
47. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems 25 https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html, 1097–1105 (2012).
48. Cadena, S. A. et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Comput. Biol.* **15**, e1006897 (2019).
49. Wang, A., Tarr, M. & Wehbe, L. Neural Taskonomy: Inferring the Similarity of Task-Derived Representations from Brain Activity. In *Advances in Neural Information Processing Systems 32* https://papers.nips.cc/paper/2019/hash/f490c742cd8318b8ee6dca10af2a163f-Abstract.html, 15475–15485 (2019).
50. Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M. & Tolias, A. S. Engineering a less artificial intelligence. *Neuron* **103**, 967–979 (2019).
51. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Sci. Data* **7**, 347 (2020).
52. Nastase, S. A., Liu, Y.-F., Hillman, H., Norman, K. A. & Hasson, U. Leveraging shared connectivity to aggregate heterogeneous datasets into a common response space. *Neuroimage* **217**, 116865 (2020).
53. Taylor, J. R. et al. The cambridge centre for ageing and neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* **144**, 262–269 (2017).
54. Bellec, P. & Boyle, J. A. Bridging the gap between perception and action: the case for neuroimaging, AI and video games. Preprint at https://psyarxiv.com/3epws (2019).
55. Pinho, A. L. et al. Individual Brain Charting, a high-resolution fMRI dataset for cognitive mapping. *Sci. Data* **5**, 180105 (2018).
56. Poldrack, R. A. et al. Long-term neural and physiological phenotyping of a single human. *Nat. Commun.* **6**, 8885 (2015).
57. Seeliger, K., Sommers, R. P., Güçlü, U., Bosch, S. E. & van Gerven, M. A. J. A large single-participant fMRI dataset for probing brain responses to naturalistic stimuli in space and time. Preprint at https://www.biorxiv.org/content/10.1101/687681v1 (2019).
58. Naselaris, T., Allen, E. & Kay, K. Extensive sampling for complete models of individual brains. *Curr. Opin. Behav. Sci.* **40**, 45–51 (2021).

Methods

Participant recruitment. The NSD study was advertised to the University of Minnesota community. We sought to recruit right-handed individuals (18–65 years old) with no known cognitive deficits or color blindness and with normal or corrected-to-normal vision. Those who were interested in participating were contacted for a phone interview to explain the nature of the study and to screen them for eligibility. We discussed the long-term nature of the study, the time commitment that it would involve and the feasibility of traveling to the scanner on a regular basis. We paid attention to the communicativeness of potential participants and their general attitude toward study participation. Selecting participants whom we were confident would provide high-quality data was more important to us than obtaining a random sample of the general population. Based on the phone interviews, we invited 14 individuals whom we thought were strong candidates to participate in an initial 7T fMRI screening session. Of these, eight were selected to participate in the full NSD experiment.

Participants. Eight participants (two males and six females; age range, 19–32 years) participated in the NSD dataset (subj01–subj08). There were six additional participants (four males and two females; age range, 20–53 years) who participated in the initial 7T fMRI screening session but not in the remainder of data collection. No statistical methods were used to pre-determine the sample size; rather, our experimental approach[58] emphasizes collecting extensive data from each participant, which enables the demonstration and replication of effects in individual participants. Participants were naive to the design of the NSD dataset. All participants had normal or corrected-to-normal visual acuity. Informed written consent was obtained from all participants, and the experimental protocol was approved by the University of Minnesota institutional review board. Participants were compensated at a rate of $30 per hour, plus performance bonuses. Additional participant information, including height, weight, handedness and visual acuity, was logged and is available online.

Individuals participated in several neuroimaging and behavioral data collection sessions (a full breakdown is provided in Extended Data Fig. 2). Neuroimaging included 3T structural scan sessions and 7T functional scan sessions. The 7T functional scan sessions included an initial screening session termed 'prffloc', referring to the pRF and fLoc experiments conducted in that session. The 7T sessions also included, for each participant, 30–40 sessions in which the main NSD experiment was conducted ('nsd01–nsd40'). These sessions are collectively termed the 'NSD core'. In some of these sessions, resting-state data were acquired before and after the NSD experiment. Finally, the 7T sessions also included two sessions conducted after completion of the NSD core; these sessions, termed 'nsdsynthetic' and 'nsdimagery', involved measuring responses to synthetic stimuli and cognitive task manipulations (including mental imagery), respectively. The total number of 7T fMRI scan sessions was 43, 43, 35, 33, 43, 35, 43 and 33 for subj01–subj08, respectively. The average number of hours of resting-state fMRI conducted for each participant was 2.0 h, and the average number of hours of task-based fMRI conducted for each participant was 38.5 h. Each individual also participated in several behavioral assessments after scanning was complete. These included a variety of behavioral measures ('nsdpostbehavior'), a final memory test ('nsdmemory') and an image-similarity assessment ('nsdmeadows').

MRI data acquisition. MRI data were collected at the Center for Magnetic Resonance Research at the University of Minnesota. Some data were collected using a combination of a 3T Siemens Prisma scanner and a standard Siemens 32-channel RF head coil. Most data were collected using a combination of a 7T Siemens Magnetom passively shielded scanner and a single-channel-transmit, 32-channel-receive RF head coil (Nova Medical). Illustrations of the different types of MRI data acquired are provided in Fig. 2b. Below, we summarize the scanning protocols (full protocol printouts are available online).

At 3T, we collected several anatomical measures (T_1, T_2, diffusion and angiogram). The motivation for collecting data at 3T was to ensure acquisition of T_1 volumes with good gray-matter/white-matter contrast and homogeneity, which is difficult to achieve at ultra-high field[59]. To increase contrast-to-noise ratio and enable the ability to assess reliability, we acquired several repetitions of T_1-weighted and T_2-weighted volumes. For each participant, we collected between six and ten scans of a whole-brain T_1-weighted MPRAGE sequence (0.8-mm isotropic resolution, TR = 2,400 ms, TE = 2.22 ms, TI = 1,000 ms, flip angle 8°, bandwidth 220 Hz per pixel, no partial Fourier, in-plane acceleration factor (iPAT) 2, TA = 6.6 min per scan) and 2–3 scans of a whole-brain T_2-weighted SPACE sequence (0.8-mm isotropic resolution, TR = 3,200 ms, TE = 563 ms, bandwidth 744 Hz per pixel, no partial Fourier, iPAT 2, TA = 6.0 min per scan). In addition to T_1 and T_2 data, we also acquired four high-angular-resolution, diffusion-weighted spin-echo EPI scans, using protocols from the Lifespan Human Connectome Project Development effort[60]. These protocols involved varying the number of diffusion directions and the phase encode direction (1.5-mm isotropic resolution, TR = 3,230 ms, TE = 89.20 ms, flip angle 78°, refocusing flip angle 160°, bandwidth 1,700 Hz per pixel, echo spacing 0.69 ms, partial Fourier 6/8, no iPAT, multi-band slice acceleration factor 4, TA = 5.6 min per scan for 99 directions, TA = 5.7 min per scan for 100 directions). The four scans included 99 directions AP (anterior-to-posterior phase encode direction), 99 directions PA (posterior-to-anterior phase encode direction), 100 directions AP and 100 directions PA. Diffusion volumes were acquired at b values of 0, 1,500 or 3,000 s mm^{-2}. We also acquired an angiogram using a time-of-flight multi-slab 3D sequence (0.39 mm × 0.39 mm × 0.5 mm resolution, TR = 19.0 ms, TE = 2.91 ms, flip angle 18°, bandwidth 186 Hz per pixel, phase partial Fourier 6/8, slice partial Fourier 6/8, iPAT 2, TA = 5.5 min).

At 7T, we collected functional data and associated fieldmaps and a few additional anatomical measures (venogram and high-resolution T_2). Functional data were collected using gradient-echo EPI at 1.8-mm isotropic resolution with whole-brain (including cerebellum) coverage (84 axial slices, slice thickness 1.8 mm, slice gap 0 mm, field-of-view 216 mm (FE) × 216 mm (PE), phase encode direction anterior-to-posterior, matrix size 120 × 120, TR = 1,600 ms, TE = 22.0 ms, flip angle 62°, echo spacing 0.66 ms, bandwidth 1,736 Hz per pixel, partial Fourier 7/8, iPAT 2, multi-band slice acceleration factor 3). The use of moderate spatial resolution capitalizes on the SNR benefits provided by ultra-high magnetic field strength. At the beginning of each 7T session, we acquired a short test EPI scan and adjusted the gain factor (FFT scale factor) accordingly to ensure large dynamic range while avoiding clipping. Empirical measurements indicate that the acoustic noise caused by the EPI sequence is 112 dBA; assuming a conservative noise reduction estimate of 26 dB for the earplugs that we used, the resulting noise level is 86 dBA, which can be safely endured for approximately 8–16 continuous hours according to guidelines from the National Institute for Occupational Safety and Health (1998) and the Occupational Safety and Health Administration (2009).

In addition to the EPI scans, the 7T sessions also included dual-echo fieldmaps for post hoc correction of EPI spatial distortion (same overall slice slab as the EPI data, 2.2 mm × 2.2 mm × 3.6 mm resolution, TR = 510 ms, TE_1 = 8.16 ms, TE_2 = 9.18 ms, flip angle 40°, bandwidth 301 Hz per pixel, partial Fourier 6/8, TA = 1.3 min per scan). Fieldmaps were periodically acquired over the course of each scan session to track changes in the magnetic field (details provided below). In one of the 7T sessions held for each participant, we acquired a venogram using a susceptibility-weighted imaging 3D sequence (0.5625 mm × 0.5625 mm × 0.6 mm resolution, TR = 28 ms, TE = 21 ms, flip angle 17°, bandwidth 120 Hz per pixel, phase partial Fourier 6/8, slice partial Fourier 6/8, iPAT 3, TA = 10.1 min). This venogram could be useful for investigating the effect of vasculature on fMRI signals[32]. In addition, for the purposes of hippocampal segmentation, we acquired in one of the 7T sessions a high-resolution T_2-weighted TSE scan (0.357 mm × 0.357 mm × 1.5 mm resolution, 56 oblique slices oriented perpendicular to the long axis of the hippocampus, field-of-view 160 mm (FE) × 156.4 mm (PE), TR = 16,000 ms, TE = 53 ms, bandwidth 100 Hz per pixel, no partial Fourier, iPAT 2, turbo factor 15, TA = 4.5 min).

In the prffloc 7T fMRI session, the acquisition structure was [F BWLL F BWLL F BWLL F], where F indicates a fieldmap, B indicates a multibar run of the pRF experiment (188 TRs), W indicates a wedgering run of the pRF experiment (188 TRs) and L indicates a run of the fLoc experiment (195 TRs). In the NSD 7T fMRI sessions, the acquisition structure was either [F NNNN F NNNN F NNNN F] or [F RNNNN F NNNN F NNNNR F], where F indicates a fieldmap, N indicates a run of the NSD experiment (188 TRs) and R indicates a resting-state run (188 TRs).

Stimulus display and scanner peripherals. Ear plugs were used to reduce acoustic noise experienced by the participants. To minimize head motion, we acquired a headcase[61] for each of the eight NSD participants (Caseforge, http://caseforge.co) and deployed the headcases starting from the second NSD core scan session (nsd02). To ensure maximal participant comfort, only the posterior half of the headcases was used (omitting the anterior half). Standard foam padding was used to mitigate head motion before that point (prffloc and nsd01).

Stimuli were presented using a Cambridge Research Systems BOLDscreen 32 LCD monitor positioned at the head of the 7T scanner bed, placed flush against the scanner bore. We chose to use an LCD monitor because it delivers a sharp, high-quality image, in contrast to typical scanner setups involving projectors and backprojection screens. The monitor operated at a resolution of 1,920 pixels × 1,080 pixels at 120 Hz. The size of the full monitor image was 69.84 cm (width) × 39.29 cm (height). Participants viewed the monitor via a mirror mounted on the RF coil. The viewing distance was 5 cm from the participants' eyes to the mirror + 171.5 cm from the mirror to the monitor image = 176.5 cm total. Measurements of the display spectral power density were obtained using a PR-655 spectroradiometer (Photo Research). The BOLDscreen is designed by the manufacturer to behave as a linear display device, and our measurements confirmed this to be the case.

We determined the maximum square extent visible in both eyes given the constraints of the RF coil to be 8.4° × 8.4° (714 pixels × 714 pixels). Thus, stimuli from the various experiments (for example, pRF, fLoc and NSD) were adjusted to fill 8.4° of visual angle (details provided below). At the beginning of each scan session, we made an effort to position the monitor in the same location relative to the scanner and to position the participant's head and RF coil in the same location relative to the scanner. We also used a calibration square (8.4° in size) to determine any incidental horizontal or vertical offsets needed in that session for the participant to see the entire square in each eye, unobstructed. Given these efforts, we think that consistent and high-quality visual stimulation was achieved across scan sessions. Nonetheless, we caution that, due to limitations in

positioning and/or potential drift over the course of a scan session, some slight occlusion of the corners of the 8.4° × 8.4° square extent might have occurred some of the time.

A Mac Pro computer controlled stimulus presentation using code based on Psychophysics Toolbox 3.0.14 (refs.[62,63]). Behavioral responses were recorded using a button box (Current Designs). In some scan sessions (nsd21–nsd30, the same sessions in which the primary set of resting-state data were acquired), physiological data were collected using a pulse oximeter and a respiratory belt (stock Siemens equipment). Care was taken to secure the oximeter with tape to the left index finger of the participant and to secure the respiratory belt snugly to the participant's torso. Physiological data were carefully synchronized with the fMRI data and cropped but are not further analyzed in this paper.

In several scan sessions (see Extended Data Fig. 2 for details), eye-tracking was performed using an EyeLink 1000 system (SR Research) combined with a custom infrared illuminator mounted on the RF coil. Eye-tracking was performed for the left eye, and eye-tracking data were obtained at 2,000 Hz using the Pupil-CR centroid mode. We caution that the eye-tracking data are variable in quality, as achieving sufficient pupil contrast was often difficult given the constraints of the scanner setup. For information complementary to the eye-tracking data, we also captured video recordings of the eye-tracker computer display (Fig. 2b) using a cell phone secured to a mount. These video recordings are useful for checking the accuracy of the eye-tracker and provide information in scan sessions where pupil tracking and data acquisition failed completely. Details of pre-processing and analysis of eye-tracking data are provided in Supplementary Note 3.

Day-to-day acquisition procedures. Participants were scanned approximately once a week, with attempts to keep a regular weekly scan time. At the beginning of each session (starting at approximately nsd07), participants were asked to rate on a five-point scale how well they slept the night before, their mood, how hungry they were and their stress level. We also asked whether they had had caffeine in the past 3 h. At the end of each scan session, participants were asked to rate how comfortable they were during the session and to provide any general feedback they had about the session. These various measures, as well as any technical issues that arose during the session, were logged onto a spreadsheet (available online).

In the first several scan sessions, we emphasized the importance of fixation and performed simple tests before scanning in which we watched the participant's eyes while they attempted to fixate and while they deliberately broke fixation. This was done to help participants understand what good fixation feels like. In every scan session, we reminded participants about the importance of fixation and about the correct mapping between buttons and responses.

During data collection, we monitored aspects of data quality, including overall image quality, head motion, quality of physiological data and behavioral performance. Between functional runs, we checked in with the participant to assess their energy level, enthusiasm and compliance. If we noticed any substantial drops in response rate, we politely notified the participant and offered short breaks before continuing.

To promote participant engagement and retention, participants were given the opportunity to earn monetary bonuses that gradually increased in size over the course of the NSD study. These bonuses were contingent on achieving certain performance levels on data quality metrics, such as head motion and response rate (details available online). Information regarding performance was supplied to participants in the form of a continually updated 'leaderboard' figure. We found that this figure greatly helped to motivate participants.

The NSD experiment. *Basic design.* In the NSD experiment, participants performed a long-term continuous recognition task while viewing a large number of color natural scenes. We chose this recognition task because it engages and challenges the observer and is unbiased with respect to the specific content of the images (unlike other tasks such as animacy judgment). In addition, it infuses the experiment with a rich memory dimension that is likely of interest to memory researchers. In total, 73,000 distinct images were prepared. We intended that the eight NSD participants would each view 10,000 distinct images presented three times each over the course of 40 scan sessions. We designated a special set of 1,000 images (chosen randomly from the full set of prepared images) as shared images that would be seen by all participants (referred to as the 'shared1000'); all other images would be mutually exclusive across participants. The distribution of the three presentations of each image was tightly controlled, and participants were naive to both the number and distribution of the presentations. Note that, because some NSD participants completed only 30 of the 40 prescribed scan sessions, there are ultimately 515 images, out of the shared 1,000 images, that were viewed all three times by all eight participants (referred to as the 'shared515').

Images were presented using a 3-s ON/1-s OFF trial structure (Fig. 1a). In informal piloting, we found that this pacing made the recognition task feasible and not overly taxing. In addition, we reasoned that the relatively long stimulus duration would increase neural activity and that the rapidity of the design would allow more trials to be collected and, thereby, increase overall experimental power. Finally, we speculated that the 3/1 trial structure would yield a pleasant experience for participants, at least compared to slow event-related designs where most experimental time is spent viewing a blank screen.

Image preparation. The NSD stimuli are prepared as a single brick of RGB images with dimensionality 425 pixels × 425 pixels × 3 RGB channels × 73,000 images and unsigned 8-bit integer format.

Images were taken from Microsoft's COCO image database[14]. COCO images are photographs collected from online repositories; each image is supplemented by a rich set of annotations (for example, boundary polygons around objects, natural language captions and body pose estimates). Of the 90 original COCO categories, a total of 80 COCO categories exist in the 73,000 NSD images. We used COCO images in the 2017 train/val split[14] and restricted selection to the subset of images for which pixel-level annotations of 'stuff'[64] (for example, sky, land, wall and road) in addition to 'things' (for example, car, skateboard and hat) were available.

We selected only images whose smaller dimension (height or width) was at least 425 pixels. Where necessary, we squared image dimensions by cropping out pixels along the largest dimension. For example, if the original image was 425×585, we cropped away 160 pixels from the larger dimension, resulting in an image that is 425×425. The median number of pixels cropped per image was 160. After cropping, images were downsampled, if needed, to 425×425.

Cropping an image can change the way the viewer interprets it. We refer to this effect of cropping as 'semantic loss'. To be able to take full advantage of the rich annotations available for the COCO images, we attempted to minimize semantic loss when cropping images. For landscape-oriented images, we selected among a center, left or right crop. For portrait-oriented images, we selected among a center, top or bottom crop (finer grids of cropping options had little effect on results). Selection of crops was carefully performed based on quantitative analysis and visual inspection (details provided in the NSD Data Manual).

In addition to screening to minimize semantic loss, we implemented a screening procedure to remove duplicate images. Some of the COCO images are extremely similar to each other, differing only by a post-processing operation (that is, grayscaling or sharpening) or by a few frames in a motion-capture sequence. To remove these near-duplicates, we downsampled all images to 40×40 and then computed the correlation of grayscale pixel intensities between all image pairs. We manually inspected the image pairs with the 500 highest correlation values. Of these, 38 image pairs were observed to be near-duplicates. We randomly selected another image from the COCO dataset to replace one image in each near-duplicate pair. Finally, we screened captions for all images for indications of violent or salacious content. No images were deemed too offensive to include in the experiment.

The distribution of 'thing' categories across the final images selected for the NSD was nearly identical to distribution in the full COCO dataset. As a result, the 'person' category was over-represented; however, with a few exceptions, all 80 COCO object categories were displayed in at least 100 images to each participant. Note that images tend to depict more than one category, so that a given object category frequently appeared in the same image with other categories. For each participant's images, at least 90% of the images contained two or more of the 80 COCO categories.

Distribution of image presentations. We determined the ordering of the 10,000 images × 3 trials = 30,000 trials in advance and kept the ordering fixed across participants. The idea is that these 10,000 images are actually treated as slots into which different NSD images are inserted. We designated the first 1,000 slots as corresponding to the special shared1000 images; the remaining 9,000 slots were filled with unique images for each participant. Note that because the trial ordering and repetition structure are identical across participants, the difficulty of the recognition task is similar across participants (up to the fact that some images might be more difficult to remember than others).

We controlled the distribution of image presentations to prevent the recognition task from becoming too difficult (and risking loss of participant morale). In the procedure, we conceptualized the task of determining the trial ordering as equivalent to placing image presentations on a circle that would eventually be cut and unraveled. The rationale for this circular design is to minimize the extent to which certain points in the experiment differ from others; of course, because the circle eventually becomes a line, there is some imperfection (see discussion below regarding 'burn-in' and 'dead' time). To determine presentation times, we created a circular probability distribution by mixing a von Mises distribution and a uniform distribution (Extended Data Fig. 1a). Using random draws from the resulting distribution (positioning the distribution at a random location on the circle for each image), we determined three presentation times for each of the 10,000 images. After completing the placement of all 30,000 trials, we then cut the circle, unraveled it into a linear sequence of image presentations and divided this sequence into 40 consecutive segments corresponding to the 40 NSD scan sessions (750 trials per session).

To determine presentation times, we created a circular probability distribution by mixing a von Mises distribution and a uniform distribution (Extended Data Fig. 1a). For each image, we positioned the peak of the von Mises distribution at a random position on the circle (that is, we randomly sampled the mean parameter from −180° to 180°) and then randomly sampled presentation times for each of the three image repetitions from the mixture distribution. We chose specific parameters for the probability distribution: we used a von Mises distribution with a concentration parameter of 729 and a mixing ratio of 60% and 40% for the von

Mises and uniform distributions, respectively. This choice of parameters yields appealing properties. First, the distribution is relatively narrow (Extended Data Fig. 1a) and, therefore, ensures that there will be many trials involving an image that has been presented in the recent past (thus making the trials easy) while still allowing the probing of more distant memory events. Second, there is minimal 'burn-in' time at the beginning of the experiment: even in the first scan session, there is still a substantial number of trials involving old images (Extended Data Fig. 1b, blue line). Third, there is minimal 'dead' time at the end of the experiment: even in the last scan session, there is still a substantial number of trials involving new images (Extended Data Fig. 1b, blue line).

To provide a sense of the overall experimental design, we computed basic statistics on each NSD scan session. For a typical session, the total number of distinct images shown once, twice and all three times within that session is 437, 106 and 34, respectively (these numbers reflect the mean across scan sessions, rounding to the nearest integer).

Trial and run design. Each trial lasted 4 s and consisted of the presentation of an image for 3 s, followed by a 1-s gap. In total, 75 trials were conducted in a run; thus, each run lasted 300 s. The first three trials (12 s) and the last four trials (16 s) were blank trials. The remaining 68 trials were divided into 63 stimulus trials and five blank trials. The blank trials were randomly positioned in each run such that the minimum and maximum continuous number of stimulus trials was nine trials (36 s) and 14 trials (56 s), respectively (Fig. 1b). For even-numbered runs, the 63rd stimulus trial was designated to be a blank trial. In total, 12 NSD runs were collected in one NSD session, yielding a total of $(63+62) \times 6 = 750$ stimulus trials. Moreover, this design was repeated for all 40 NSD sessions: 750 stimulus trials × 40 sessions = 30,000 stimulus trials. The temporal ordering of stimulus and blank trials was generated once and kept fixed across participants.

Note that the experimental design involves minimal trial jittering: for the most part, the time interval separating consecutive stimulus images is fixed at 1 s, although occasionally, due to blank trials, the time interval is 5 s. This design was intended to maximize statistical power and differs from conventional fMRI practice where intervals are often chosen randomly from a fixed range.

Stimulus presentation and task. Because the BOLDscreen is calibrated to behave as a linear display device, we used a squaring luminance response when presenting the NSD experiment to simulate the typical viewing of digital images. At the time of presentation, the prepared NSD images were resized using linear interpolation from their native resolution of 425 pixels × 425 pixels to 714 pixels × 714 pixels to occupy 8.4° × 8.4° on the display. Throughout each run (including blank trials), a small semi-transparent red fixation dot with a black border (0.2° × 0.2°, 50% opacity) was present at the center of the stimuli (Fig. 1a). Stimuli were shown against a gray background with an RGB value of 127, 127 and 127.

Participants were instructed to fixate the central dot and to press button 1 using the index finger of their right hand if the presented image was new—that is, if the image had never been presented before—or button 2 using the middle finger of their right hand if the presented image was old—that is, the image was identical to one that had been presented before, either in the current scan session or any previous scan session. Participants were additionally instructed to continue to fixate and wait for the next image in the event of blank trials.

Before the start of the NSD experiment, we showed the participants a version of the experiment involving cartoon images, for them to become familiarized with the feel and timing of the task. During the NSD experiment, minimal feedback was provided to the participants regarding their performance on the recognition task. Participants were blinded to the precise details of the NSD experiment (for example, total number of images and total number of presentations per image). Participants were informed only about their response rate (fraction of trials on which they successfully made a response) and a vague 'performance metric', which, unbeknownst to them, quantified their percent correct for easy trials (trials that involved the presentation of an image that had occurred earlier in the same scan session). We revealed the nature of the design in a debriefing session after the completion of the NSD experiment (details below).

Details on experiment timing. Stimulus presentation was locked to the refresh rate of the BOLDscreen monitor. Empirical measurements confirmed that the monitor refresh rate was nearly exactly 120 Hz: duration of runs was highly reliable, ranging from 299.95 s to 299.98 s. To compensate for the slight offset from 300 s, the fMRI data were pre-processed to achieve a sampling rate of 0.999878 s (high-resolution preparation) or $0.999878 \text{ s} \times (4/3) = 1.333171$ s (standard-resolution preparation). For brevity, we refer to these numbers as 1.000 s and 1.333 s. Experimental runs were started by a trigger issued by the MR scanner. Due to input polling and monitor refresh, there was slight variability in the delay between trigger detection and the presentation of the first stimulus frame, ranging from 3 ms to 22 ms. We did not attempt to compensate for this delay.

Acquisition. Due to constraints on participant availability (including unplanned out-of-town absences in the summer of 2019) and due to constraints on scanner availability (the 7T scanner was decommissioned in November 2019), we did not complete the full NSD experiment for every participant. Fortunately, we were able to collect a sizable amount of data: 40, 40, 32, 30, 40, 32, 40 and 30 NSD sessions for subj01–subj08, respectively. In these collected data, each participant viewed 9,209–10,000 distinct images and participated in 22,500–30,000 trials. Aggregated across participants, the total number of distinct images shown was 70,566, and the total number of trials was 213,000.

Debriefing. After completion of the final memory test (details below), participants filled out a post-NSD questionnaire. This questionnaire probed topics such as strategies used for performing the NSD task and estimates for the number of images viewed and the number of image repetitions. After filling out this questionnaire, the design of the NSD experiment was then revealed to the participants.

Other experiments. *pRF experiment.* We adapted the experiment used in the Human Connectome Project 7T Retinotopy Dataset[30]. Stimuli consisted of slowly moving apertures filled with a dynamic colorful texture (Fig. 2a). Apertures and textures were updated at a rate of 15 Hz. Two run types were used. The first, termed 'multibar', involves bars sweeping in multiple directions (same as RETBAR in the Human Connectome Project 7T Retinotopy Dataset). The second, termed 'wedgering', involves a combination of rotating wedges and expanding and contracting rings. Both run types included blank periods.

For consistency with the NSD experiment, stimuli were resized to fill a circular region with diameter 8.4°. Each run lasted 300 s (exact empirical timings were highly accurate and ranged between 299.95 s and 300.00 s). Throughout stimulus presentation, a small semi-transparent dot (0.2° × 0.2°) was present at the center of the stimuli. The color of the central dot switched randomly to one of three colors (black, white or red) every 1–5 s. Participants were instructed to maintain fixation on the dot and to press a button whenever the color of the dot changed. To further aid fixation, a semi-transparent fixation grid was superimposed on the stimuli and was present throughout the experiment[65]. A total of six runs (three multibar and three wedgering) were collected in the first 7T fMRI session (prffloc).

fLoc experiment. This experiment was developed by the Grill-Spector laboratory[31] (stimuli and presentation code available at http://vpnl.stanford.edu/fLoc/). The experiment consisted of the presentation of grayscale images depicting different stimulus categories (Fig. 2a). There were ten categories, grouped into five stimulus domains: characters (word and number), bodies (body and limb), faces (adult and child), places (corridor and house) and objects (car and instrument). Stimuli were presented on a scrambled background (different backgrounds for different stimuli). Stimuli were presented in 4-s trials. In a trial, eight images from a given category were sequentially presented (image duration, 0.5 s). Each run included six presentations of each of the ten categories as well as blank trials (also of 4-s duration).

For consistency with the NSD experiment, stimuli were resized to fill a square region filling 8.4° × 8.4° of visual extent. Each run lasted 300 s (exact empirical timings were highly accurate and ranged between 300.000 s and 300.002 s). Throughout stimulus presentation, a small red fixation dot was present at the center of the stimuli. Participants were instructed to maintain fixation on the dot and to press a button whenever they noticed an image in which only the background was present ('oddball' task). In total, six runs were collected in the first 7T fMRI session (prffloc).

Resting-state experiment. Stimuli consisted of a white fixation cross (0.5° × 0.5°) on a gray background (Fig. 2a). Each resting-state run lasted 300 s. In the second resting-state run held within a given scan session, the fixation cross turned red after 12 s had elapsed and remained red for 4 s before returning to white.

Resting-state data were acquired in several NSD core scan sessions: nsd21–nsd38 for subj01 and subj05 and nsd21–nsd30 for all other participants. Thus, a total of 100 min or 180 min of resting-state data were acquired for each participant. In each session, one resting-state run was acquired at the beginning of the session (before the NSD runs), and another resting-state run was acquired at the end of the session (after the NSD runs).

In the first resting-state run, participants were instructed to stay awake and fixate the cross but otherwise rest. In the second resting-state run, participants were additionally instructed to inhale deeply when the fixation cross turned red. This instructed breath was designed to aid analysis of the physiological data collected concomitantly with the resting-state data. Before each resting-state run, participants were asked to report their current sleepiness level using the Stanford Sleepiness Scale[66] (1–7, where 1 is most active and 7 is most sleepy). After each resting-state run, participants were asked to report their sleepiness level during the run that had just completed.

After the last scan session involving resting-state data, participants filled out a post-resting-state questionnaire. This questionnaire queried what the participants were doing during the resting-state runs and whether they thought about the images from the NSD experiment.

Synthetic stimuli experiment (nsdsynthetic). After completion of the NSD experiment, we conducted an additional 7T fMRI scan session in which responses were measured to a variety of carefully controlled synthetic

(non-naturalistic) stimuli while the participant performed either a fixation task or a one-back task. These data will be described and released in a forthcoming manuscript.

Visual imagery experiment (nsdimagery). After completion of the nsdsynthetic experiment, we conducted an additional 7T fMRI scan session in which responses were measured while participants engaged in visual imagery and other cognitive tasks. These data will be described and released in a forthcoming manuscript.

Additional behavioral measures (nsdpostbehavior, nsdmemory and nsdmeadows). Several behavioral assessments were conducted after completion of the NSD experiment. Some of these were relatively brief and included the following (nsdpostbehavior): open-ended questions regarding language ability; the Vividness of Visual Imagery Questionnaire[67]; the Test of Word Reading Efficiency[68], including both Sight Word Efficiency and Phonemic Decoding Efficiency; the Cambridge Memory Test for Faces[69]; ultra-fast measurement of contrast sensitivity[70]; and an assessment of chromatic sensitivity (participants adjusted intensities of red, green and blue channels on the BOLDscreen display until minimal luminance flicker was perceived).

We also conducted a final memory test in which we collected various memory-related measures regarding the images shown to the participants during the NSD experiment (nsdmemory). These data will be described and released in a forthcoming manuscript.

Finally, using the web-based Meadows platform (http://meadows-research.com), we conducted an assessment of how the NSD participants perceive and interpret the NSD images (nsdmeadows). First, we selected a small set of images that maximally span semantic space. This was done by isolating the shared515 images; computing shifted inverse frequency sentence embeddings for the sentence captions provided by the COCO dataset[71]; and using a greedy approach to determine the subset of 100 images that maximize the average distance between each image's embedding and its closest neighbor. We then asked participants to perform a Multiple Arrangements Task[72] in which they arrange using a drag-and-drop interface the 100 images within a white circular arena according to the similarity of their content. Using an adaptive procedure, subsequent arrangements were conducted using subsets of the images to maximize information gain. This was done until 45 min had elapsed. Using a similar interface on Meadows, participants then provided valence and arousal ratings for the 100 images as well as three additional images pulled from the shared515 images. Ratings were performed separately for valence and arousal and were accomplished by freely arranging, using a drag-and-drop interface, the images (delivered in small batches) along a one-dimensional axis ranging from low to high. This assessment took about 15 min.

Overview of data analysis. We designed custom analysis strategies to maximize the quality of derived measures from the NSD data. Several methods are based on recent work[32,73] where further details can be found. Data analysis and visualization were performed using custom code in MATLAB and Python as well as tools from various packages, such as FreeSurfer, SPM, FSL, ANTs[74] and ITK-SNAP[75]. An archive of code used is provided online (https://github.com/cvnlab/nsddatapaper/), and specific code files are referenced in the text below.

A comprehensive schematic outlining the data analysis performed in this paper is provided in Extended Data Fig. 3. The analysis of the NSD data can be divided into three components: (1) pre-processing of the anatomical, diffusion and functional data; (2) time series analysis of the fMRI data to estimate trial-wise betas; and (3) further analyses of the trial-wise betas to answer specific scientific questions. The first two components produce the so-called 'prepared data' that are generally useful to the community, whereas the third component refers to analyses performed for the purposes of this paper (estimation of pRFs from the NSD data, univariate memory analysis, representational similarity analysis and brain-optimized neural network training). Data collection and analysis were not performed blinded to the conditions of the experiments. No data were excluded from analyses, with the exception of a few T_1 volumes (2 of 52 volumes = 4%) and certain portions of the eye-tracking data that were corrupted by noise (11 of 160 eye-tracking runs = 7%).

The pre-processing approach that we designed for the NSD dataset prioritizes accuracy and preservation of information (for example, avoiding spatial smoothing). We avoid 'baking in' unnecessary assumptions (for example, aggressively removing signal fluctuations without careful assessment of validity), and we avoid assuming the accuracy of automated methods; care is taken to manually inspect each pre-processing step to ensure satisfactory results. Although we think our pre-processing is general and likely suitable for most downstream uses of the data, the raw data are also available for those who want to explore other pre-processing approaches, such as fmriprep[76]. We note several aspects of the NSD dataset that might render the dataset challenging from a pre-processing standpoint: the relatively high spatial resolution of the fMRI data (1.8 mm) places higher demands on spatial accuracy; the ultra-high field strength (7T) used for the fMRI data yields higher levels of EPI spatial distortion compared to lower field strengths; and the emphasis on many repeated scans of individuals heightens the importance of achieving consistent imaging results across scan sessions.

Pre-processing of MRI data. Details of the pre-processing of anatomical, functional and diffusion data are provided in Supplementary Notes 4 and 5. Functional data were pre-processed using one temporal resampling to correct for slice time differences and one spatial resampling to correct for head motion within and across scan sessions, EPI distortion and gradient non-linearities. Two versions of the functional data were prepared: a 1.8-mm standard-resolution preparation (temporal-resolution, 1.333 s) and an upsampled 1.0-mm high-resolution preparation (temporal-resolution, 1.000 s). Analyses of the pRF and fLoc experiments were used to define retinotopic and category-selective ROIs, respectively. Other ROIs were also defined, including an 'nsdgeneral' ROI indicating occipital regions generally responsive in the NSD experiment and a 'corticalsulc' ROI collection indicating major cortical sulci and gyri. Annotations for several of the corticalsulc ROIs are shown in Figs. 3f and 4b.

Data quality metrics. Several data quality metrics were calculated (export_runmetrics.m) and summarized in Figs. 1d and 2d. tSNR was computed from raw fMRI volumes (no pre-processing) by first de-trending the time series data from each voxel (quadratic polynomial fit) and then dividing the mean signal intensity by the standard deviation of signal intensity values (autoqc_fmri.m). We calculated the median tSNR across voxels within a simple brain mask (mean volume thresholded at 1/10th of the 99th percentile of values) and then computed the median across runs. Head motion was quantified by calculating frame-wise displacement[77] based on motion parameter estimates (1.8-mm preparation). We calculated the mean frame-wise displacement across volumes in a run and then computed the median across runs. BOLD response was quantified by calculating the percentage of variance explained by a simple ON–OFF GLM model (1.8-mm preparation). We calculated the median variance explained across voxels within the nsdgeneral ROI and then computed the median across runs. (Additional details on the ON–OFF GLM can be found in the 'GLMsingle algorithm' section.) Response rate was quantified by calculating the percentage of trials for which the participant pressed a button and then computing the mean across runs. Behavioral performance was quantified by dividing trials into easy trials (trials for which the presented image had been previously presented in the same scan session), hard trials (trials for which the presented image had been previously presented but in a previous scan session) and novel trials (trials for which the presented image had never been previously presented) and then calculating, for each trial type, the percentage of trials on which the participant indicated an 'old' response.

To identify EPI signal dropout regions (export_signaldropout.m), we divided the T_2 volume (resampled to match the EPI data) by the mean EPI volume (1-mm preparation). The resulting volume is useful as it indicates which voxels have high signal intensity in the T_2 but are corrupted by signal dropout in the EPI. We mapped the volume to the cortical surface (cubic interpolation; mean across depth), transformed the result to fsaverage and then used a data-driven threshold to mark atypically high values. Vertices marked in at least four of the eight participants are indicated in Fig. 3f. To visualize surface imperfections, we took the voxels that were marked in the 0.8-mm anatomical space (during the manual inspection of FreeSurfer surface imperfections), smoothed this binary volume with a 3D Gaussian with full width at half maximum of 2 mm, mapped the result to the cortical surface (cubic interpolation; maximum across depth) and then transformed the result to fsaverage. Vertices exceeding 0.01 in at least one of the eight participants are indicated in Fig. 3f.

Rankings from the 7T fMRI screening session. Six quality measures (pRF BOLD, fLoc BOLD, pRF behavior, fLoc behavior, raw motion and de-trended motion) were computed for each of the 14 individuals who participated in the screening session. BOLD quality was quantified as the percentage of voxels for which variance explained by modeling the fMRI time series data (either pRF model fitting or GLM model fitting) exceeded 20%. Behavior quality was quantified as described above. Motion was quantified by calculating the median voxel displacement relative to the reference volume used for motion correction, computing the median of this quantity across volumes and then computing the mean across runs. This motion quantification was performed using raw motion parameter estimates (thereby providing a measure of global head displacement over the course of the session) as well as using motion parameter estimates that are linearly de-trended within each run (thereby providing a measure of within-run head instability). Each of the six measures was linearly scaled to span the range 1–5, where 1 corresponds to the worst performance and 5 corresponds to the best performance observed across participants. Finally, the normalized measures were averaged to produce an overall ranking for each participant, as depicted in Fig. 2c.

Analysis of behavioral data from the NSD experiment. The behavioral data from the NSD experiment were lightly reformatted for the convenience of subsequent analyses (analyzebehavior_nsd.m). We first checked whether the participant had accidentally positioned their fingers on incorrect buttons on the button box and compensated for this if necessary. (In a few instances, we deliberately instructed participants to use alternative buttons due to hardware malfunction of the button box.) We then recorded, for each stimulus trial, several quantities, including time of image presentation, whether the image presented was new or old, whether the response was correct or incorrect and the reaction time. Button responses were

extracted from a time window extending 250–4,250 ms after image onset. In the case of multiple buttons pressed during a trial, we scored the final button pressed, excluding any redundant presses of that button (participants sometimes repeated button presses for good measure).

GLM analysis of the NSD experiment. *Overview of approach.* We performed a GLM analysis of the pre-processed time series data from the NSD experiment. To maximize flexibility for subsequent analyses, the GLM approach was designed to provide estimates of BOLD response amplitudes ('betas') for single trials. Due to low SNR, single-trial estimation in fMRI is challenging. We, therefore, developed several analysis components to optimize the quality of single-trial betas. These components are packaged into a tool called GLMsingle, which is the subject of a forthcoming manuscript where additional details and discussion can be found.

The first analysis component of GLMsingle is the use of a library of HRFs, whereby the best-fitting HRF from the library is chosen for each voxel. This simple approach for compensating for differences in hemodynamic time courses across voxels[78] has several appealing features: it is efficient and can be executed with little computational cost (and, hence, can accommodate the massive scale of the NSD); and it invariably provides well-regularized HRF estimates. The second analysis component is an adaptation of GLMdenoise to a single-trial GLM framework. GLMdenoise[35] is a technique in which data-derived nuisance regressors are identified and used to remove noise from—and, therefore, improve the accuracy of—beta estimates. The third component is an application of ridge regression[79] as a method for dampening the noise inflation caused by correlated single-trial GLM predictors. To determine the optimal level of regularization for each voxel, we make use of a recently developed efficient re-parameterization of ridge regression called 'fractional ridge regression'[36].

Derivation of the library of HRFs. To generate a library of HRFs that accurately capture empirically occurring time course variation, we performed an initial analysis of data from the first NSD core session (nsd01). This library was fixed and used for the analysis of all subsequent NSD sessions. The first step was to create a comprehensive summary of observed time courses (hrf_derivecanonicalpcs.m). The time series data from each participant's nsd01 session was fit using a finite impulse response model (0–30 s) where all of the stimulus trials are treated as instances of a single experimental condition (this simplification is necessary to make estimation feasible). We identified voxels for which model variance explained (R^2) was greater than 10%, and, from these voxels, we randomly drew 20,000 voxels (with replacement). Pooling across participants, time course estimates from the resulting 160,000 voxels were subjected to singular value decomposition to determine the top three PCs (shown in Fig. 3b, inset). To fine-tune time course estimates, we re-fit the time series data from the nsd01 session using these three PCs as the basis (as opposed to the finite impulse response basis). Finally, adopting the visualization approach of the Temporal Decomposition Method[73], we projected voxel time course estimates onto the unit sphere (using the same voxel selection criterion of $R^2 > 10\%$) and constructed a 2D histogram for each participant (shown in Fig. 3a).

The second step was to define a set of time courses that span the observed time course variation (hrf_constructmanifold.m). To do this, we converted the 2D histograms to units of relative frequency and then averaged the histograms across participants. Inspecting the group average histogram (shown in Fig. 3b), we manually clicked a sequence of points on the unit sphere that follow the data density as closely as possible. We then parameterized the path traced by these points (a simple one-dimensional manifold) by positioning regularly spaced points where successive points are separated by six angular degrees (Fig. 3b, cyan dots). The time courses corresponding to the resulting set of 20 points were cubic interpolated to a sampling rate of 0.1 s and normalized to peak at 1 (Fig. 3c). Finally, we fit each time course using a double-gamma function as implemented in SPM's spm_hrf.m (hrf_fitspmhrftomanifold.m). This yielded a library of 20 canonical HRFs that might be useful for application to other experimental datasets (getcanonicalhrflibrary.m). We note that variation in time course shape is likely due to the influence of macrovasculature on BOLD temporal dynamics[73].

Cross-validation framework for single-trial GLM. The GLMdenoise and ridge regression analysis components of GLMsingle both require tuning of hyperparameters. To determine the optimal setting of hyperparameters, we use a cross-validation approach in which out-of-sample predictions are made for single-trial beta estimates, as opposed to time series data. This simplifies and reduces the computational requirements of the cross-validation procedure. Note that, because of cross-validation, although GLMsingle produces estimates of responses to single trials, it does require the existence of and information regarding repeated trials—that is, trials for which the stimulus is the same.

The first step of the cross-validation procedure is to analyze all of the available data using no regularization. In the case of GLMdenoise, this amounts to the inclusion of zero nuisance regressors; in the case of ridge regression, this amounts to the use of a shrinkage fraction of 1, indicating ordinary least squares regression. In both cases, the analysis produces a full set of unregularized single-trial betas (for example, in one NSD session, there are 750 single-trial betas distributed across 12 runs). The second step of the procedure is to perform a grid search over values of the hyperparameter (for example, number of nuisance regressors and shrinkage fraction). For each value, we assess how well the resulting beta estimates generalize to left-out runs. For example, in leave-one-run-out cross-validation, one run is held out as the validation run; stimuli that occur in both the training runs and the validation run are identified; and squared errors between the regularized beta estimates from the training runs and the unregularized beta estimates from the validation run are calculated. This procedure is iterated with each run serving as the validation run, and errors are summed across iterations.

GLMsingle algorithm. Having described the essential aspects of the estimation framework above, we now turn to the steps in the GLMsingle algorithm. GLMsingle involves fitting several different GLM variants. Each variant includes polynomial regressors to characterize the baseline signal level: for each run, we include polynomials of degrees 0 through round ($L/2$), where L is the duration in minutes of the run.

1. Fit a simple ON–OFF GLM. In this model, stimulus trials are treated as instances of a single experimental condition, and a canonical HRF is used (getcanonicalhrf.m). Thus, there is a single 'ON–OFF' predictor that attempts to capture signals driven by the experiment. The utility of this simple model is to provide variance explained (R^2) values that help indicate which voxels carry experiment-driven signals.
2. Fit a baseline single-trial GLM. In this model, each stimulus trial is modeled separately using the canonical HRF. This model provides a useful baseline for comparison.
3. Identify HRF for each voxel. We fit the data multiple times with a single-trial GLM, each time using a different HRF from the library of HRFs. For each voxel, we identify which HRF provides the best fit to the data (highest variance explained) and inherit the single-trial betas associated with that HRF. Note that the final model for each voxel involves a single chosen HRF from the library (not a weighted sum of HRFs).
4. Use GLMdenoise to determine nuisance regressors to include in the model. We define a pool of noise voxels (brain voxels that have low ON–OFF R^2) and then perform principal component (PC) analysis on the time series data associated with these voxels. The top PCs are added one at a time to the GLM until cross-validation performance is maximized on average across voxels.
5. Use fractional ridge regression to regularize single-trial betas. With the nuisance regressors determined, we use fractional ridge regression (fracridge[36]) to estimate the single-trial betas, systematically evaluating different shrinkage fractions. For each voxel, in the context of a GLM that incorporates the specific HRF chosen for that voxel, cross-validation is used to select an optimal shrinkage fraction for that voxel. To mitigate bias on the overall scale of betas, we apply a post hoc scaling and offset on betas obtained for a given voxel to match, in a least squares sense, the unregularized betas obtained for that voxel.

Application of GLMsingle to the NSD data. We used GLMsingle to analyze the time series data independently for each NSD scan session (glm_nsd.m). Major algorithmic parameters included the following: we evaluated up to ten nuisance regressors; we evaluated shrinkage fractions from 0.05 to 0.90 in increments of 0.05 and from 0.91 to 1 in increments of 0.01 (representing a finer grain for voxels with the best SNR); we performed six-fold cross-validation (consecutive pairs of runs) for Steps 4 and 5; and we used an ON–OFF R^2 threshold of 5% in Step 4.

Three different versions of the single-trial betas were computed and saved. The first beta version (b1, 'betas_assumehrf') is the result of Step 2 and reflects the use of a canonical HRF. The second beta version (b2, 'betas_fithrf') is the result of Step 3 and reflects the result of voxel-wise HRF estimation. The third beta version (b3, 'betas_fithrf_GLMdenoise_RR') is the result of Step 5 and reflects the additional GLMdenoise and ridge regression procedures. Betas were converted to units of percent BOLD signal change by dividing amplitudes by the mean signal intensity observed at each voxel and multiplying by 100. Although we provide betas in units of percent signal change, we suggest that users might want to z-score the responses of each voxel within each scan session to eliminate potential non-stationarities and to equalize units across voxels.

For user convenience, we created preparations of the single-trial betas in additional spaces other than the native 1.8-mm and 1.0-mm functional spaces. For the 'nativesurface' preparation, we performed cubic interpolation of the 1.0-mm betas onto each of the three cortical surface depths and averaged across depths (analysis_transformfsaverage.m). The result was then mapped using nearest neighbor interpolation to fsaverage space to create the 'fsaverage' preparation. For the 'MNI' preparation, we mapped the 1.0-mm betas to MNI space using cubic interpolation (analysis_transformMNI.m).

GLM analysis of the resting-state experiment. As an optional resource, we fit the time series data from the resting-state experiment using methods that parallel those used for the NSD experiment (glm_nsdresting.m). For each scan session involving resting-state, we took the two resting-state runs (first and last run acquired) and analyzed the data using the design matrix of the neighboring NSD runs and the same voxel-wise HRFs determined from analyzing the NSD runs in

that scan session (this is analogous to beta version b2). Although there is no reason to think that spontaneous resting-state activity conforms to the 4-s trial structure of the NSD experiment, these resting-state betas might be useful as a direct comparison for the NSD betas.

Noise ceiling estimation. To obtain a measure of data quality, noise ceilings were estimated for the NSD betas (export_noiseceiling.m). The noise ceiling for a given voxel is defined as the maximum percentage of variance in the voxel's responses that can, in theory, be explained, given the presence of measurement noise. Our method for estimating the noise ceiling follows the general framework laid out in previous studies[80,81]. Several assumptions are made: (1) the signal contained in the voxel's response is determined solely by the presented image; (2) the variability of the signal across different images is Gaussian distributed; (3) the noise is Gaussian distributed with zero mean; and (4) the response to an image is equal to the signal plus noise. Given these assumptions, any observed response is a sample from a sum of Gaussian distributions:

$$\mathrm{RESP} \sim \mathcal{N}\left(\mu_{\mathrm{signal}}, \sigma_{\mathrm{signal}}\right) + \mathcal{N}\left(0, \sigma_{\mathrm{noise}}\right)$$

where RESP indicates the NSD beta observed on a given trial, μ_{signal} is the mean signal across different images, σ_{signal} is the standard deviation of the signal across different images and σ_{noise} is the standard deviation of the noise (for illustration of these concepts, see Extended Data Fig. 8c). Note that the first Gaussian distribution characterizes true signal variability, whereas the second Gaussian characterizes variability due to noise. Also, note that this framework treats response variability unrelated to the stimulus as 'noise', but such variability might, in fact, reflect 'signal' from the perspective of functional connectivity[82].

To compute the noise ceiling, we first take the trial-wise NSD betas for each voxel and z-score these betas within each scan session. This simple normalization compensates for non-stationarities that might exist across sessions. We then calculate the variance of the betas across the three presentations of each image (using the unbiased estimator that normalizes by $n-1$ where n is the sample size), average this variance across images and then compute the square root of the result. This produces an estimate of the noise standard deviation:

$$\hat{\sigma}_{\mathrm{noise}} = \sqrt{\mathrm{mean}\left(\beta_{\sigma}^{2}\right)}$$

where β_{σ}^{2} indicates the variance across the betas obtained for a given image. Next, given that the variance of the z-scored betas is 1, we estimate the signal standard deviation as follows:

$$\hat{\sigma}_{\mathrm{signal}} = \sqrt{\left|1 - \hat{\sigma}_{\mathrm{noise}}^{2}\right|_{+}}$$

where $||_{+}$ indicates positive half-wave rectification. Finally, we simplify by calculating a single scalar quantity:

$$\mathrm{ncsnr} = \frac{\hat{\sigma}_{\mathrm{signal}}}{\hat{\sigma}_{\mathrm{noise}}}$$

where ncsnr indicates the noise ceiling SNR.

Given the framework described above, the noise ceiling can be calculated as the amount of variance contributed by the signal expressed as a percentage of the total amount of variance in the data:

$$\mathrm{NC} = 100 \times \frac{\sigma_{\mathrm{signal}}^{2}}{\sigma_{\mathrm{signal}}^{2} + \sigma_{\mathrm{noise}}^{2}}$$

where NC indicates the noise ceiling. We would like to be able to calculate the noise ceiling based on the single scalar ncsnr. Moreover, because a researcher might want to average across multiple presentations of each image before attempting to explain the NSD betas, we would like a method for flexibly expressing the noise ceiling for different levels of trial averaging. With some algebra, it can be shown that the noise ceiling can be expressed as follows:

$$\mathrm{NC} = 100 \times \frac{\mathrm{ncsnr}^{2}}{\mathrm{ncsnr}^{2} + \frac{1}{n}}$$

where n indicates the number of trials that are averaged together (see the NSD Data Manual for the derivation and additional details). We note that there is a direct relationship between the commonly used metric of split-half reliability and the noise ceiling: if a voxel has two sets of responses that reflect the same image presentations, then the correlation between the two sets of responses multiplied by 100 is equal to the noise ceiling for single-trial responses expressed in percent variance explained.

Using the above methods, we calculated noise ceilings for each of the beta versions and for each of various spatial preparations (1.8-mm, 1-mm, fsaverage and nativesurface). For simplicity, noise ceiling estimates were calculated using betas associated with images with all three presentations available. To assess stability, we also computed split-half noise ceiling estimates. This was achieved by splitting the available images into two mutually exclusive groups and computing noise ceiling estimates independently for each group. The noise ceiling results shown in Fig. 3f,g and Supplementary Fig. 6 were computed assuming $n = 3$, reflecting the scenario in which trial-wise betas are averaged across three trials for each image. The noise ceiling results shown in Fig. 6a,b were computed assuming $n = 1$ and are expressed in correlation units (square root of percent variance explained).

We include a few important notes as follows. Even though the NSD consists of only up to three trials for a given image, the estimate of response variability for each voxel (that is, the noise standard deviation) is averaged across a very large number of images, thus stabilizing the noise ceiling estimate. Also, note that our noise ceiling metric refers to activity levels in individual voxels in individual participants. It is thus quite different from, for example, noise ceiling metrics computed for group average representational dissimilarity matrices[83]. The latter are more abstracted away from the data given that they summarize properties observed across a collection of voxels, reflect second-order computations on activity levels and not activity levels themselves and probe responses at the group level and not at the individual level.

Calculation of equivalent trials. To provide a common basis for comparing different datasets, we define the number of equivalent trials present in a dataset as $N \times \mathrm{ncsnr}^2$, where N indicates the number of trials conducted and ncsnr is the noise ceiling SNR (as defined above). The assumptions here are that (1) every trial has equal value, irrespective of whether it is used to measure brain responses to an image that has already been shown or a new image (for example, two trials for one image is equivalent to one trial for two distinct images); and (2) increases in SNR are equivalent to the collection of additional trials. For an illustrative example of the second assumption, suppose an experimenter chooses to improve SNR by averaging the response to a given image across p repetitions of that image. This effectively reduces the noise standard deviation by a factor of $\sqrt{p}$, and ncsnr will thus increase by a factor of $\sqrt{p}$. Alternatively, the experimenter could choose to not average and instead use the p trials as is. In the former case, the number of equivalent trials is $1 \times (\sqrt{p} \times \mathrm{ncsnr})^2 = p \times \mathrm{ncsnr}^2$, whereas, in the latter case, the number of equivalent trials is $p \times \mathrm{ncsnr}^2$. Thus, the two cases correspond to the same number of equivalent trials.

We conducted an auxiliary analysis that directly compares the NSD against the BOLD5000 dataset[22]. The goal of this analysis was to calculate a summary ncsnr value for each dataset, so that the number of equivalent trials can be calculated. For fair comparison, both NSD and BOLD5000 were analyzed using the same GLM methods described in this paper (beta version b3). We then defined a common brain region on which data quality can be compared. This was done by transforming the nsdgeneral ROI to MNI space and then mapping the resulting MNI mask to each participant in the two datasets. Finally, we computed the median ncsnr observed across voxels in the mask in each participant.

The median ncsnr, averaged across participants, was 0.260 for the NSD (averaged across the first four NSD participants) and 0.187 for BOLD5000 (averaged across the four participants in BOLD5000). This indicates that, despite the longer time duration allocated per trial in BOLD5000 (10 s) compared to the NSD (4 s), the quality of a single-trial beta in the NSD is higher than that in BOLD5000. Specifically, one NSD trial is approximately equivalent to $(0.260)^2/(0.187)^2 = 1.93$ BOLD5000 trials. This increase in quality is likely due, in part, to the screening of participants and the ultra-high magnetic field strength (7T) used in the NSD. Note that the ncsnr metric quantifies the SNR per trial and is expected to be unbiased with respect to the number of repeated trials used to calculate it. Thus, although the exact number of trials per image is different in the NSD and BOLD5000 datasets, the ncsnr values can still be directly compared.

Univariate analysis of memory recognition. For this analysis (results shown in Fig. 4b), we used version 3 of the NSD betas (b3) in the fsaverage preparation. Betas for each surface vertex were kept in percent signal change units. Using the behavioral responses, we identified trials involving hits (participants responded 'old' to a previously presented image) and trials involving correct rejections (participants responded 'new' to a novel image). Then, for each participant, we calculated two-sample t-values at each surface vertex. This was done both for trials pooled within individual NSD scan sessions as well as for trials pooled across all sessions.

Representational similarity analysis. For this analysis (results shown in Fig. 5), we used version 3 of the NSD betas (b3) in the fsaverage preparation. Betas for each surface vertex were z-scored within each scan session, concatenated across sessions and averaged across repeated trials for each distinct image. To support the representational similarity analysis[84], we defined a set of ROIs (V1, V2, V3, pVTC and aVTC) on the fsaverage surface. This was done by mapping the manually defined V1, V2 and V3 from each participant to fsaverage, averaging across participant and using the result to guide the definition of group-level ROIs. We also defined a posterior and anterior division of ventral temporal cortex (pVTC and aVTC, respectively) based on anatomical criteria. For each participant, we extracted betas for vertices within each ROI (concatenating across hemispheres). We then computed Pearson's correlation between beta patterns across all possible

pairs of images. This yielded RDMs with rows and columns indexing distinct images (for example, the RDMs for participant 1 have dimensionality 10,000 × 10,000 with correlations corresponding to 49,995,000 possible pairs).

To help visualize and interpret these large dissimilarity matrices, we performed *t*-SNE embedding[41,85] using a perplexity level of 100 (Fig. 5b,c). This projects the high-dimensional representations onto a 2D plane such that the distance of a given pair on the plane reflects that pair's distance in the high-dimensional representation as accurately as possible. To verify the strong categorical structure visible in pVTC and aVTC (Fig. 5b), we quantified the similarity of the brain RDMs to a model RDM constructed from the category labels in the COCO dataset. Specifically, we constructed an RDM from a binary matrix indicating the presence or absence of each of the 80 COCO categories (cosine distance metric) and correlated this model RDM with each brain RDM. This process was performed for mutually exclusive groups of 100 images drawn from all images presented three times to a given participant (the number of groups was 100, 100, 62, 54, 100, 62, 100 and 54 for the eight participants, respectively). We calculated the mean and standard error across results obtained for different groups of images (Fig. 5d). Finally, we investigated similarity of brain representations across ROIs and participants. This was done by isolating the shared515 images, constructing brain RDMs for these images and correlating brain RDMs across ROIs and participants. The resulting second-order RDM is shown in Fig. 5e, with further quantification of this matrix shown in Fig. 5f.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The NSD dataset is freely available at http://naturalscenesdataset.org. The data are hosted in the cloud, allowing researchers to exploit high-performance cloud computing to efficiently analyze the dataset. We provide both raw data in BIDS format[86] and prepared data files, along with extensive technical documentation in the NSD Data Manual. To ensure strict validation for an upcoming Algonauts prediction challenge[87], the initial public release will withhold the last three NSD scan sessions from each participant (approximately 8.4% of the NSD data). Images used for the NSD were taken from the Common Objects in Context database[14] (https://cocodataset.org).

Code availability

We provide an archive of code used in this study (https://github.com/cvnlab/nsddatapaper/) as well as utility functions for working with the prepared NSD data (https://github.com/cvnlab/nsdcode/). Custom algorithms developed for this study include GLMsingle (https://github.com/cvnlab/GLMsingle/) and fracridge (https://github.com/nrdg/fracridge/). Example scripts demonstrating scientific analyses of the NSD data are available (https://github.com/cvnlab/nsdexamples/); these scripts might be useful for teaching purposes.

References

59. Polimeni, J. R., Renvall, V., Zaretskaya, N. & Fischl, B. Analysis strategies for high-resolution UHF-fMRI data. *Neuroimage* **168**, 296–320 (2018).
60. Harms, M. P. et al. Extending the Human Connectome Project across ages: imaging protocols for the Lifespan Development and Aging projects. *Neuroimage* **183**, 972–984 (2018).
61. Power, J. D. et al. Customized head molds reduce motion during resting state fMRI scans. *Neuroimage* **189**, 141–149 (2019).
62. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
63. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
64. Caesar, H., Uijlings, J. & Ferrari, V. COCO-Stuff: Thing and Stuff classes in context. In *IEEE/CVF Conf. Computer Vision and Pattern Recognition* https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00132 1209–1218 (2018).
65. Schira, M. M., Tyler, C. W., Breakspear, M. & Spehar, B. The foveal confluence in human visual cortex. *J. Neurosci.* **29**, 9050–9058 (2009).
66. Shahid, A., Wilkinson, K., Marcu, S. & Shapiro, C. M. Stanford Sleepiness Scale (SSS). In: *STOP, THAT and One Hundred Other Sleep Scales* (eds. Shahid, A., Wilkinson, K., Marcu, S. & Shapiro, C. M.) 369–370 (Springer, 2012).
67. Marks, D. F. Visual imagery differences in the recall of pictures. *Br. J. Psychol.* **64**, 17–24 (1973).
68. Torgesen, J. K., Wagner, R. & Rashotte, C. *TOWRE-2: Test of Word Reading Efficiency* (Pearson, 2012).
69. Duchaine, B. & Nakayama, K. The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants. *Neuropsychologia* **44**, 576–585 (2006).
70. Tardif, J., Watson, M., Giaschi, D. & Gosselin, F. Measuring the contrast sensitivity function in just three clicks. *J. Vis.* **16**, 966–966 (2016).
71. Arora, S., Liang, Y. & Ma, T. A simple but tough-to-beat baseline for sentence embeddings. https://openreview.net/pdf?id=SyK00v5xx (2017).
72. Kriegeskorte, N. & Mur, M. Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 245 (2012).
73. Kay, K., Jamison, K. W., Zhang, R.-Y. & Uğurbil, K. A temporal decomposition method for identifying venous effects in task-based fMRI. *Nat. Methods* **17**, 1033–1039 (2020).
74. Avants, B. B. et al. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage* **54**, 2033–2044 (2011).
75. Yushkevich, P. A. et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
76. Esteban, O. et al. fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).
77. Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L. & Petersen, S. E. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage* **59**, 2142–2154 (2012).
78. Handwerker, D. A., Gonzalez-Castillo, J., D'Esposito, M. & Bandettini, P. A. The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *Neuroimage* **62**, 1017–1023 (2012).
79. Hoerl, A. E. & Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
80. Kay, K. N., Winawer, J., Mezer, A. & Wandell, B. Compressive spatial summation in human visual cortex. *J. Neurophysiol.* **110**, 481–494 (2013).
81. Lage-Castellanos, A., Valente, G., Formisano, E. & De Martino, F. Methods for computing the maximum performance of computational models of fMRI responses. *PLoS Comput. Biol.* **15**, e1006397 (2019).
82. Biswal, B., Yetkin, F. Z., Haughton, V. M. & Hyde, J. S. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magn. Reson. Med.* **34**, 537–541 (1995).
83. Nili, H. et al. A toolbox for representational similarity analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
84. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
85. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
86. Gorgolewski, K. J. et al. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* **3**, 1–9 (2016).
87. Cichy, R. M., Roig, G. & Oliva, A. The Algonauts Project. *Nat. Mach. Intell.* **1**, 613 (2019).

Acknowledgements

We thank the NSD participants for their time and endurance; E. Aminoff, J. Pyles, M. Tarr, M. Hebart and C. Baker for advice on experimental design and data collection; J. Power and A. Schapiro for consultation on resting-state and physiological data; V. Carr and R. Olsen for consultation on hippocampal subfield scanning protocols; A. Grant for assistance with scanner peripherals; F. Gosselin and J. Tardif for contrast sensitivity analysis; B. Klimes-Dougan and K. Cullen for designing the valence/arousal assessment; W. Guo for segmentations of the medial temporal lobe; M. Arcaro, A. Bratch, D. Finzi, A. White and J. Winawer for assistance with ROI definition; C. Gorgolewski and R. Poldrack for discussion of BIDS and data sharing; R. Cichy, E. Yacoub, K. Grill-Spector, K. Jamison, A. Rokem, A. Huth, S. Anzellotti, N. Kriegeskorte and J. Winawer for general discussions; and K. Ugurbil for overall project advice. We also thank our NSD collaborators for shaping the trajectory of the project. This work was supported by NSF CRCNS grants IIS-1822683 (K.K.) and IIS-1822929 (T.N.); NIH grants P41 EB015894, P30 NS076408, S10 RR026783 and S10 OD017974-01, the W. M. Keck Foundation and the NIMH Intramural Research Program ZIAMH002909 (M.N.); and NSF BCS-1734853, NIH NIBIB R01EB030896, NIH NIBIB R01EB029272 and NIH IIS-1912270 (F.P.).

Author contributions

E.J.A. collected the neuroimaging data, coordinated the data collection effort and performed manual brain segmentations. G.S.-Y. performed neural network analyses. Y.W. performed participant recruitment, assisted with scanning and prepared eye-tracking videos. J.L.B. assisted in data analysis. J.S.P. performed the equivalent trials analysis on the NSD and BOLD5000. L.T.D. organized and prepared data in BIDS format. M.N. analyzed the eye-tracking data. B.C. and F.P. analyzed the diffusion data. I.C. performed representational similarity analyses. J.B.H. analyzed the behavioral data. K.K. and T.N. conceived of the project and designed the main experiment. J.B.H. and I.C. designed the nsdmeadows and nsdmemory behavioral assessments. K.K. developed analysis methods, analyzed the neuroimaging data and directed the overall project. K.K., T.N., E.J.A., M.N., B.C., F.P., I.C. and J.B.H. wrote the paper. All authors discussed and edited the manuscript.

Competing interests

The authors declare no competing financial interests.

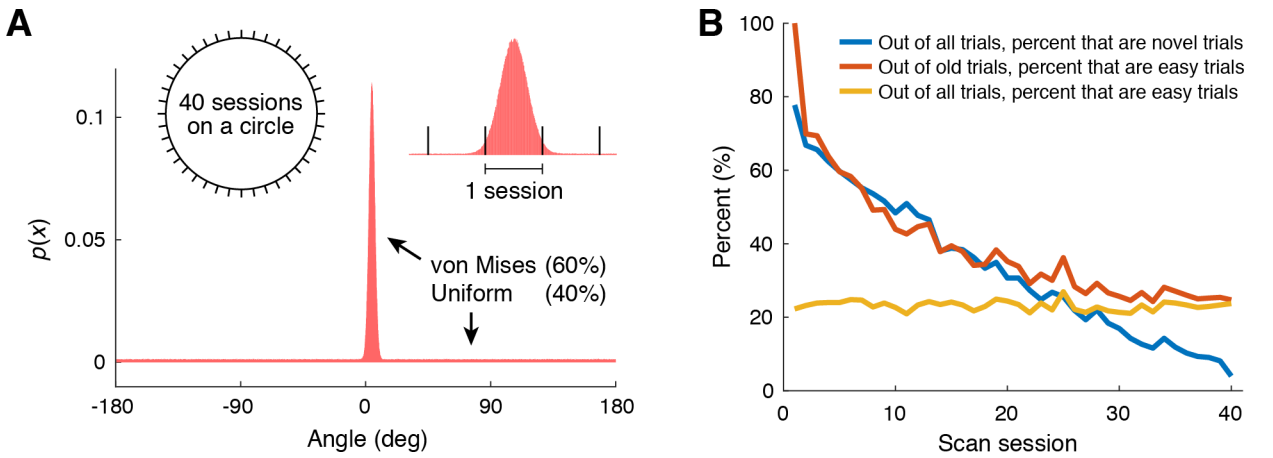
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41593-021-00962-x.

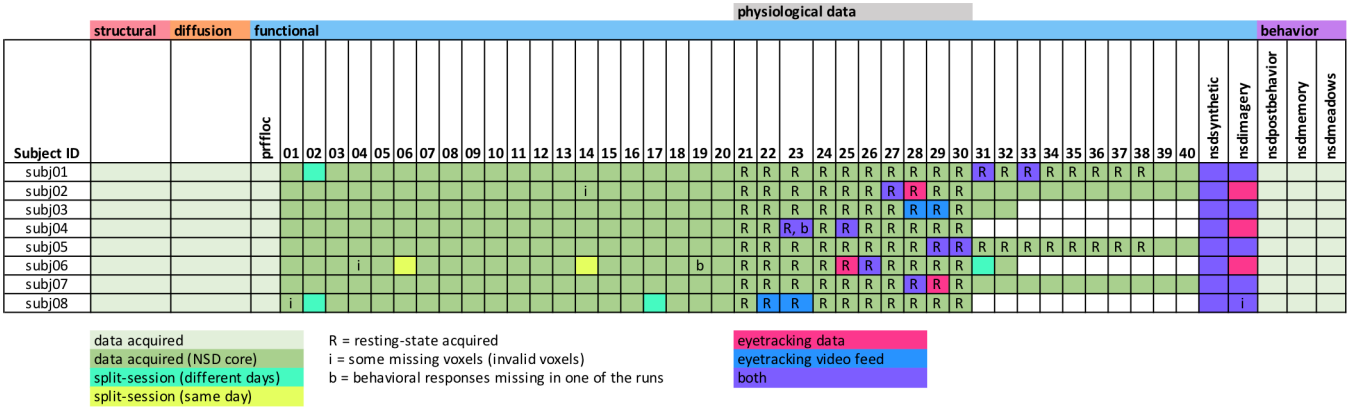
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-021-00962-x.

Correspondence and requests for materials should be addressed to Kendrick Kay.

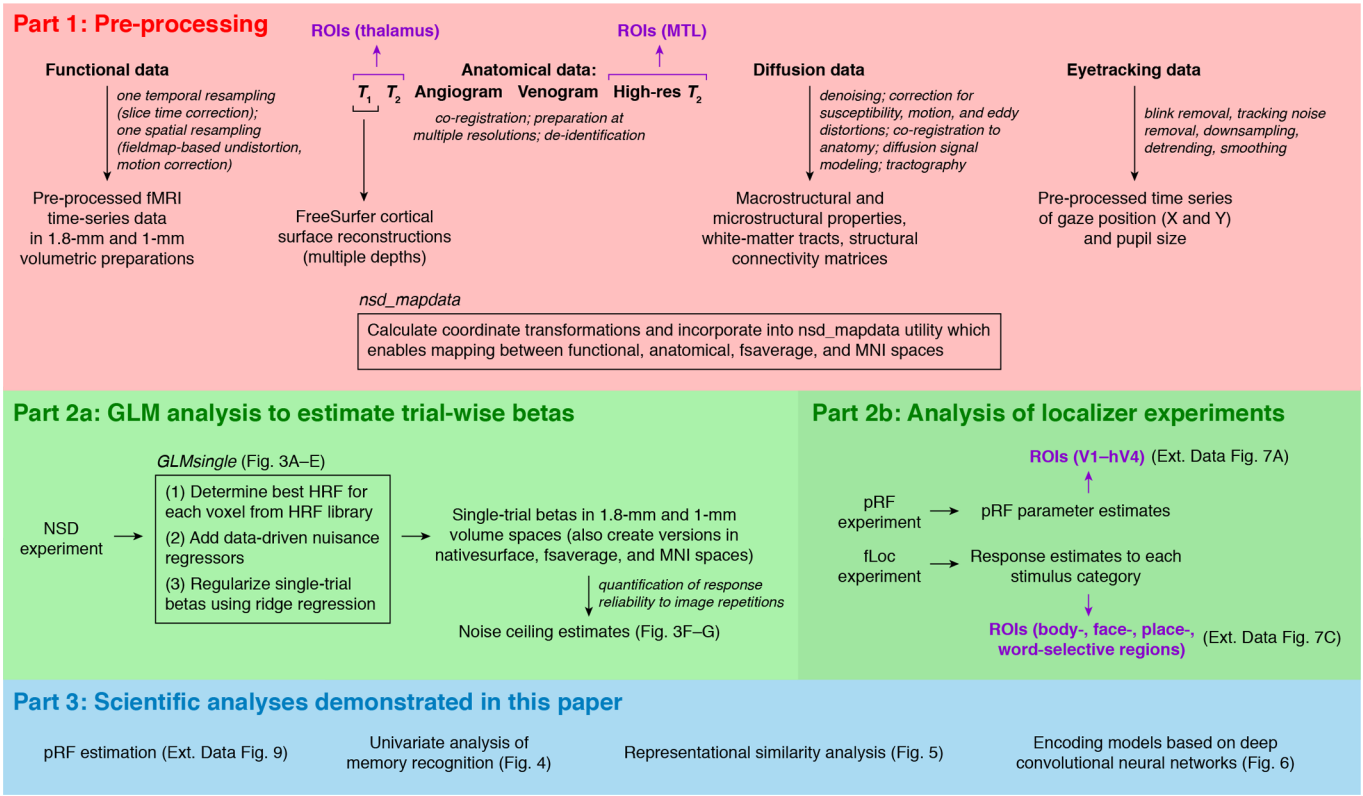
Peer review information *Nature Neuroscience* thanks Evan Gordon, Andrew Zalesky, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



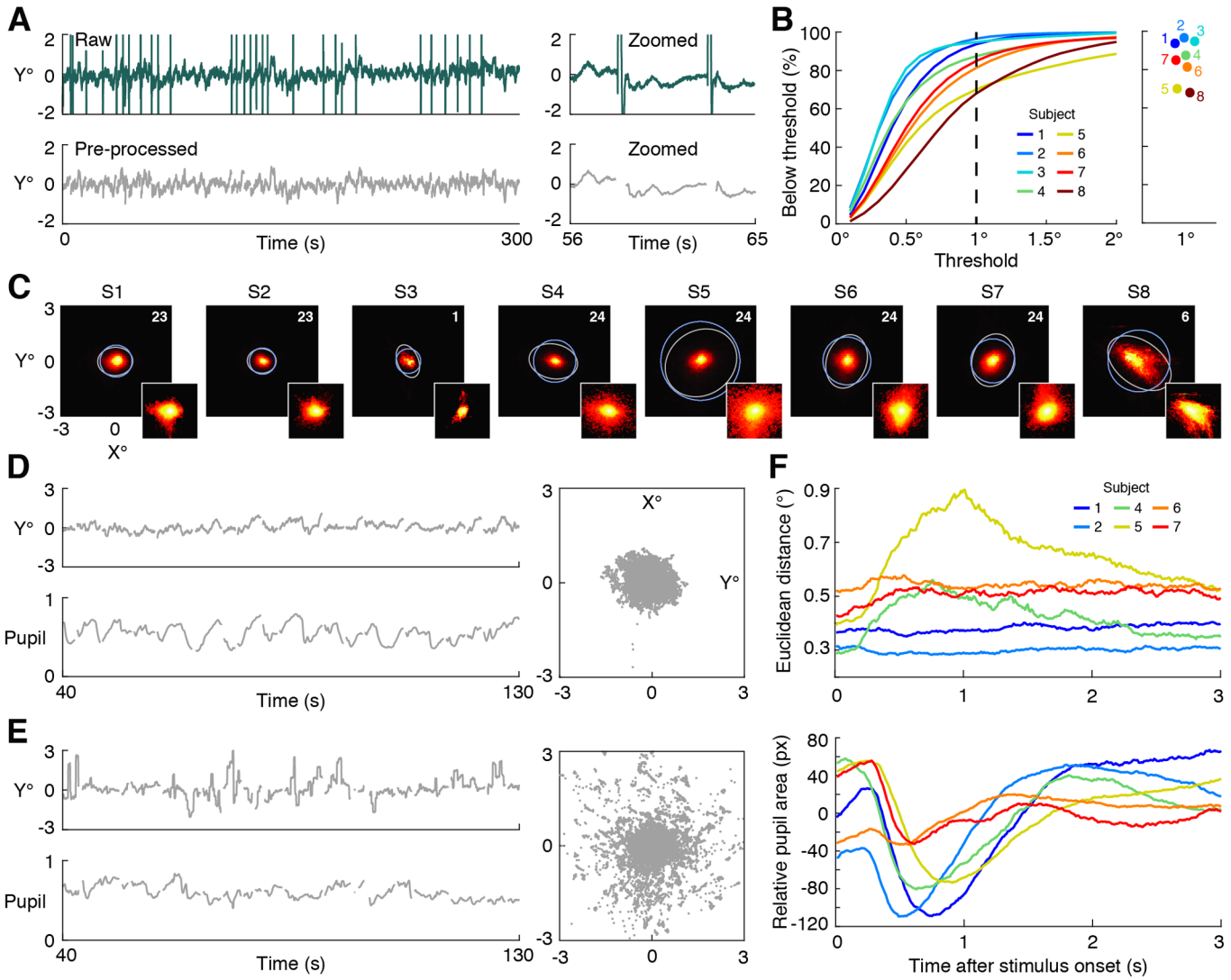
Extended Data Fig. 1 | Design of the NSD experiment. a, Image presentations. Each of 10,000 distinct images was placed 3 times on a circle according to a probability distribution created by mixing a relatively narrow von Mises distribution and a uniform distribution. The resulting image sequence was divided into 40 equally-sized segments for the 40 NSD scan sessions. **b**, Basic statistics of image repetitions. We define *novel trial* as a trial involving an image never shown before, *old trial* as a trial that is not a novel trial, and *easy trial* as an old trial for which the presented image had been shown previously in the same scan session.



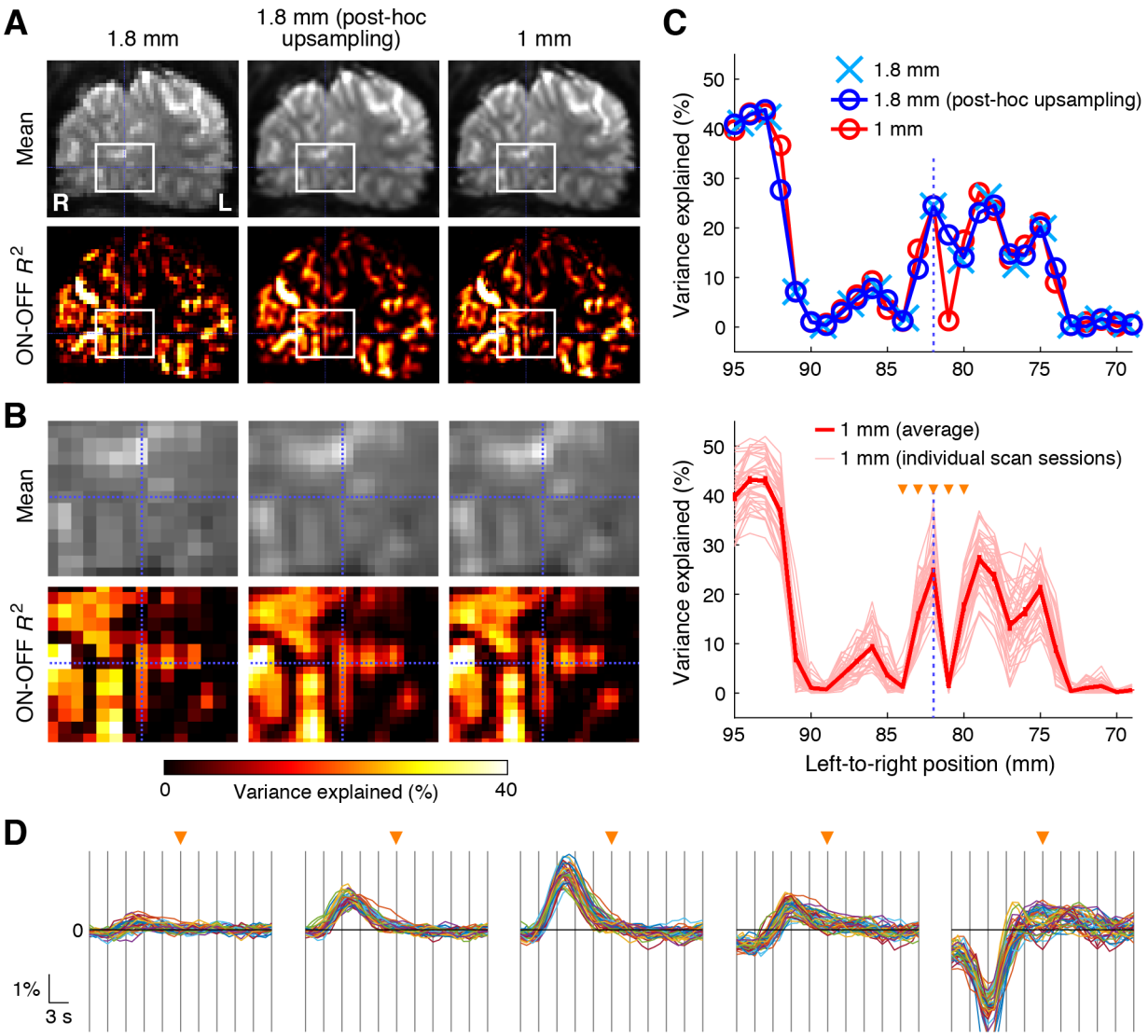
Extended Data Fig. 2 | Overview of data collection. This table summarizes the overall NSD data collection effort. Structural and diffusion MRI data were collected at 3T. Functional MRI data were collected at 7T. The breakdown of the 7T fMRI scan sessions is indicated: for example, subject 2 participated in 1 (prffloc) + 40 (nsd01–nsd40) + 1 (nsdsynthetic) + 1 (nsdimagery) = 43 7T fMRI scan sessions. Additional behavioral data were acquired outside of the scanner (nsdpostbehavior, nsdmemory, nsdmeadows). Note that scan sessions were occasionally split across multiple magnet entries (see aquamarine and yellow cells). For simplicity, we treat these cases as if they represent single scan sessions.



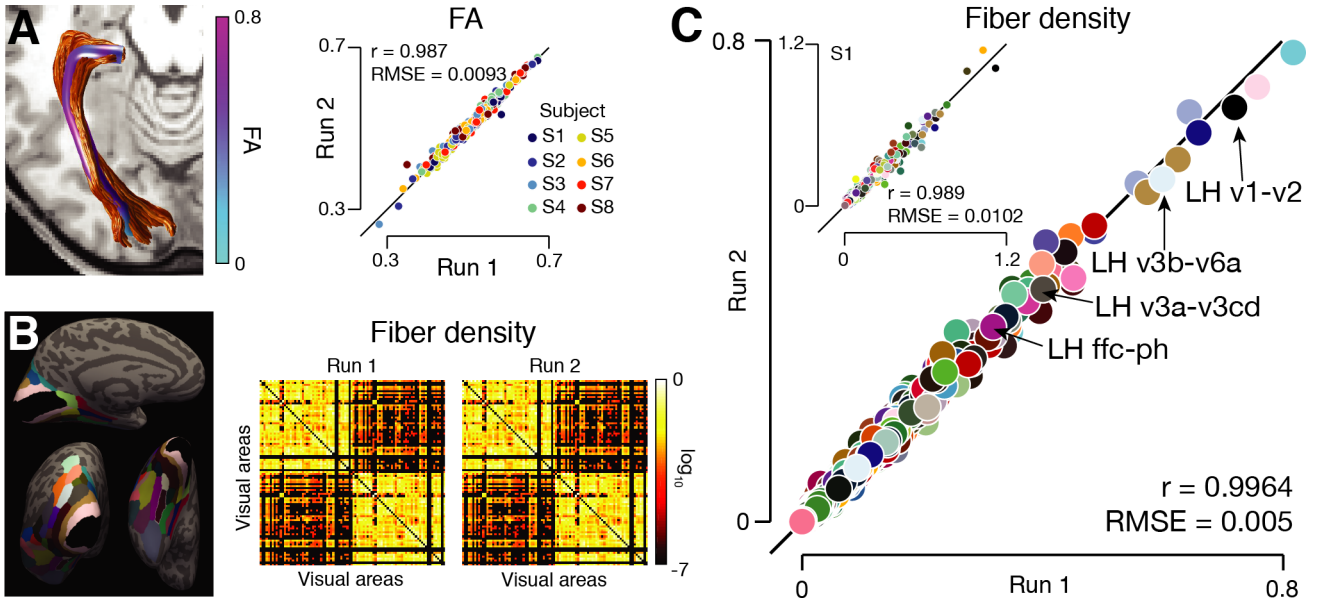
Extended Data Fig. 3 | Overview of data analysis. Analyses conducted in this paper can be divided into three parts. Part 1 consists of pre-processing, in which raw functional, anatomical, diffusion, and eyetracking data are transformed into various useful intermediate outcomes. In addition, coordinate transformations between various spaces are estimated and incorporated into the nsd_mapdata utility. Part 2 consists of analyses of the pre-processed fMRI data. The GLMsingle algorithm introduced in this paper is used to analyze the fMRI data from the NSD experiment (Part 2a), and standard methods are used to analyze the fMRI data from the pRF and fLoc experiments (Part 2b). Part 3 consists of specific scientific analyses demonstrated in this paper that make use of the data prepared in Parts 1 and 2. Given the extensive data preparation procedures (Parts 1–2), it is useful to comment on which aspects are fairly typical in MRI processing and which are more customized or unique to the present work. With respect to the pre-processing steps in Part 1, the general outcomes that these steps achieve are typical in MRI and are necessary for basic interpretation of the data. For example, small shifts in head position over the course of a scan session necessitate some motion compensation in order to interpret the signal from a given voxel in terms of a single brain location. The specific methods by which we execute these pre-processing steps may differ from what is performed in commonly used software packages (for example, SPM, FSL, AFNI). However, the outcomes are similar at a conceptual level: for example, the fMRI data are pre-processed using temporal interpolation of voxel-wise time-series data and spatial interpolation of brain volumes. With respect to the additional preparation procedures in Part 2, the procedures in Part 2b are fairly typical analyses used to functionally localize brain regions. More customized and unique to the present work are the procedures in Part 2a, which are designed to improve the accuracy of single-trial fMRI amplitude estimates. We provide evidence that these procedures do in fact perform as intended (see Fig. 3 and Extended Data Fig. 8).



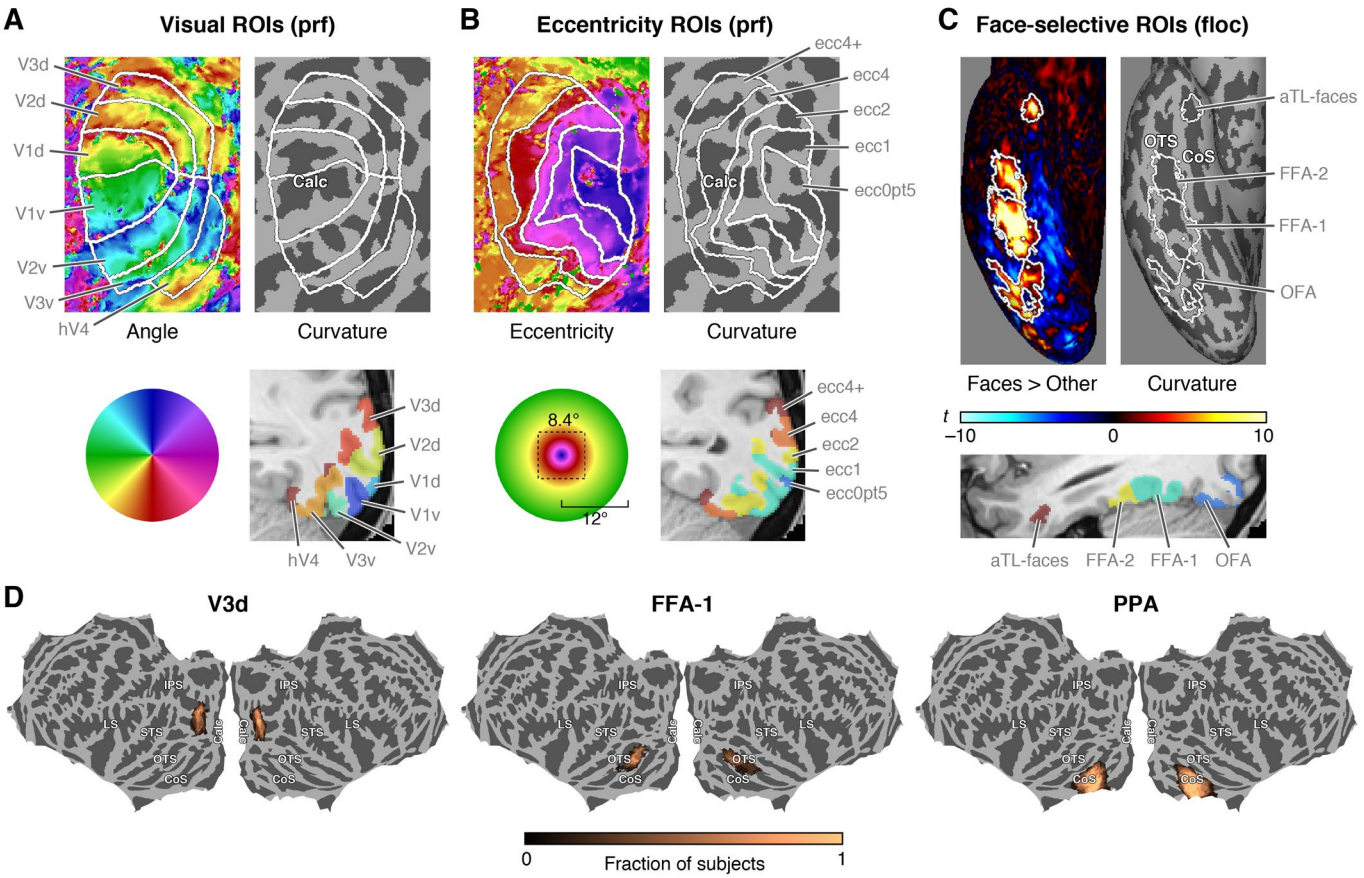
Extended Data Fig. 4 | Eyetracking results. a, Pre-processing of eyetracking data. Blinks and tracking noise were removed, followed by linear detrending, median-centering, downsampling, and smoothing. Runs with less than 1/3 valid samples after these cleaning procedures were excluded from further analysis (see Supplementary Note 5). Shown are results for an example run (subject 1, nsd31 scan session, run 6). Pre-processing reduced noise without obscuring potential eye movements. **b**, Fraction of time during which deviation from central fixation was less than a specific threshold. Results are shown for a range of thresholds (left) and for a threshold of 1° (right). **c**, 2D histograms of gaze positions. The main images show histogram results on a linear scale; the inset images show results on a log scale. To summarize the results, we overlay a gray ellipse marking the central 90% of a multivariate 2D Gaussian distribution that has been fit to the gaze positions, as well as a blue circle containing 90% of the gaze positions. Both the parametric and non-parametric approaches yield similar results and indicate that gaze positions of all subjects clustered around central fixation. The level of precision varied across subjects. The number of usable eyetracking runs for each subject is indicated by the white text. **d**, Example of accurate fixation behavior (subject 1, nsd31 scan session, run 8). Shown are pre-processed vertical gaze coordinates (top left), normalized pupil area (bottom left), and a 2D scatter plot of gaze positions (right). **e**, Example of eye movements (subject 5, nsd29 scan session, run 11). Same format as **d**. Notice that eye movements manifest as staircase structure in the vertical gaze coordinates and as dispersed gaze positions in the scatter plot. **f**, Trial-wise time-resolved analysis. Relative to stimulus trial onsets, we plot the across-trial median deviation from central fixation (top), as well as the across-trial median pupil size after mean-centering the pupil size within each trial (bottom). Results for subjects 3 and 8 are not available for this analysis. Overall, the results show that subjects were able to maintain fixation most of the time: gaze positions were within 1° of central fixation 68–97% of the time (see **b**). Three subjects are worth further discussion. Subject 4 exhibited eye movements after stimulus onset (see **f**, top); however, this is of minor concern given that these movements were small. Subject 5 exhibited more substantial eye movements (see **c**, **e**, and **f**); we suggest exclusion of this subject from analyses of the NSD fMRI data that are contingent on strict central fixation. Finally, while our results indicate fixation instability for subject 8 (see **b** and **c**), careful inspection of the eyetracking video recordings (available online) suggests this reflects pupil tracking noise rather than actual eye movements made by the subject.



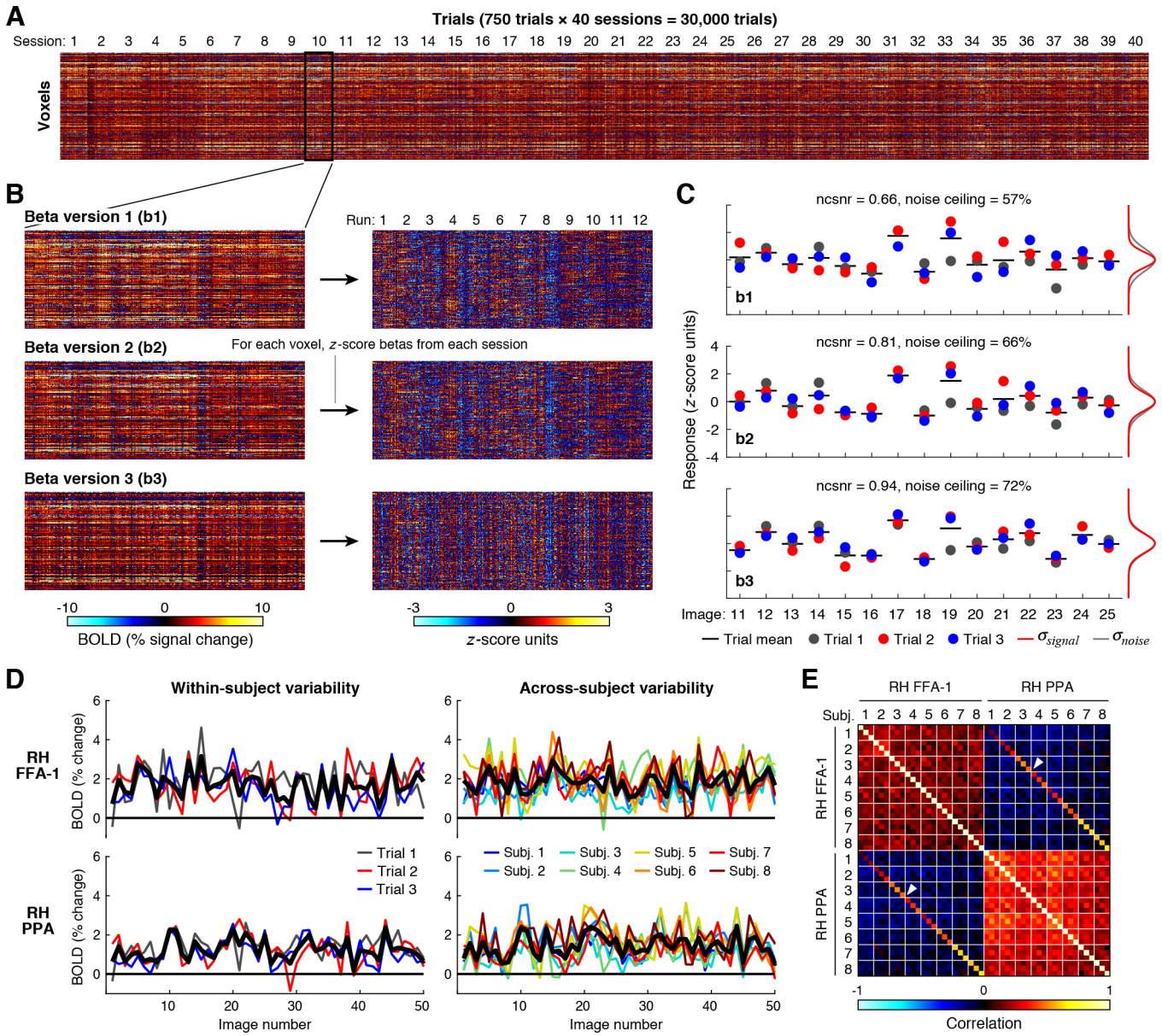
Extended Data Fig. 5 | Improvements in spatial detail through upsampling. a, Comparison of approaches. For an example coronal slice in Subject 1, we compare the non-upsampled 1.8-mm preparation of the data (left), the upsampled 1-mm preparation of the data (right), and a version of the 1.8-mm results that has been post-hoc upsampled to 1-mm resolution to enable direct comparison (middle). Two quantities are shown: mean signal intensity and variance explained by an ON-OFF GLM model. **b**, Zoomed view of white rectangle marked in **a**. **c**, Profile view of blue dotted horizontal line marked in **b**. Error bars in the bottom plot indicate ±1 SEM across 40 scan sessions (error bars are small and nearly invisible). **d**, Timecourse estimates for voxels marked by orange arrowheads at the bottom of **c**. Each colored trace corresponds to an estimate of the hemodynamic timecourse for a single voxel in one NSD scan session from the upsampled 1-mm data preparation. The beginning of the timecourses (first vertical line) corresponds to the onset of the 3-s image presentation. The results shown in this figure support the idea that the upsampled data preparation preserves fine-scale spatial detail that is lost (blurred away) under a non-upsampled data preparation. While the effects are small, preserving as much detail as possible may be critical for certain neuroscientific questions.



Extended Data Fig. 6 | Reliable diffusion derivatives facilitate investigation of white-matter connectivity. a, Fractional anisotropy (FA). The left shows tractography and FA results for the optic radiation identified in subject 7. The right shows reliability of FA results for 61 white-matter tracts identified using the atlas from Bullock et al.[114] For other measures, see Supplementary Fig. 5c–e. **b**, Structural connectivity. Using 43 visual areas × 2 hemispheres = 86 regions from the HCP-MMP1 atlas[109] (left), we construct group-average connectivity matrices indicating the density of fibers connecting pairs of regions (right). **c**, Quantitative summary. Each dot represents fiber density between a pair of regions (as in **b**). Dot colors reflect different region pairs but are otherwise arbitrary. Group-average results (main figure) and results for an individual subject (inset) are shown.

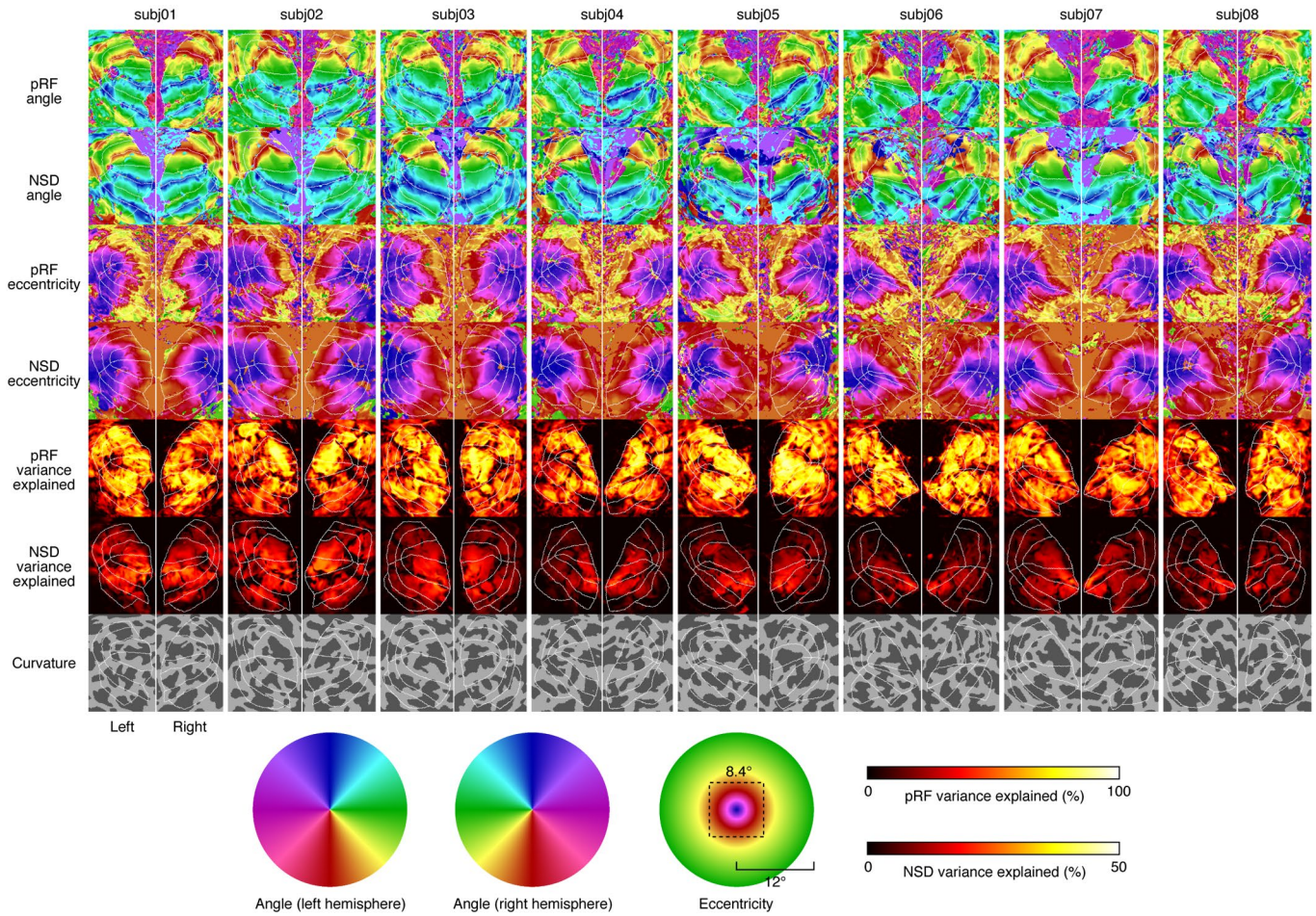


Extended Data Fig. 7 | Regions of interest (ROIs) provided with NSD. A variety of ROIs were defined based on auxiliary fMRI experiments (pRF, fLoc). In **a**–**c**, we show example results for subject 3, right hemisphere. **a**, Early visual areas. Results are shown on FreeSurfer's sphere surface as well as in the 0.8-mm anatomical volume space. **b**, Eccentricity-based regions. Similar format to **a**. Note that the total stimulus extent is 8.4° × 8.4° in the pRF, fLoc, and NSD experiments. **c**, Face-selective regions. Regions were defined based on *t*-values computed for the contrast of faces against all other categories. Results are shown on FreeSurfer's inflated surface as well as in the 0.8-mm anatomical space. **d**, Probabilistic maps of ROI locations. For each of three example ROIs, we map the location of the ROI in each subject to fsaverage and then compute, for each fsaverage vertex, the fraction of subjects labeled at that vertex. Notice there is reasonable consistency across subjects in fsaverage space.

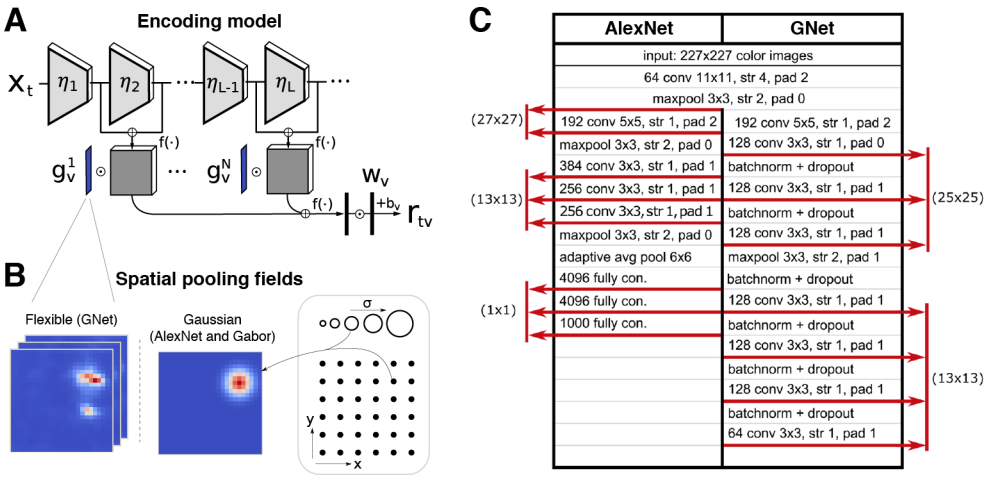


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Detailed visualization of NSD betas. We prepared three beta versions (b1, b2, b3) reflecting GLM analyses of increasing sophistication. **a**, Inspection of NSD betas. The full set of estimated single-trial responses (1.8-mm preparation, beta version b1) is shown for voxels in subject 1 right hemisphere region of interest (ROI) FFA-1 (fusiform face area subdivision 1). We observe horizontal stripes, indicative of gross variation in percent BOLD signal change across voxels. **b**, Zoomed view of one scan session. Shown are all three beta versions, as well as the result of *z*-scoring betas within each scan session (in general, we suggest that users may wish to *z*-score each voxel's responses within each scan session in order to eliminate potential non-stationarities and to equalize units across voxels). The different beta versions generally resemble one another (left column), implying that the variations in GLM methods do not drastically change the data. Vertical stripes visible in the visualizations tend to decrease from b1 to b2, suggesting that fitting voxel-wise HRFs reduces artifacts. Vertical stripes also tend to decrease from b2 to b3, which might reflect the reduction of correlated noise achieved by GLMdenoise. **c**, Detailed inspection of one voxel. To assess the reliability of evoked responses, we group trials according to the image presented. The estimated signal standard deviation (σ_{signal}) and noise standard deviation (σ_{noise}) are illustrated at the right of each subplot. Notice that b2 and b3 reduce variability of betas across the 3 trials associated with each image. **d**, Response reliability. Here we plot single-trial responses observed in two example ROIs (1.8-mm preparation, beta version b2, right hemisphere FFA-1 and PPA (parahippocampal place area), response averaged across voxels in each ROI), showing the first 50 of the shared515 images. The left column shows responses for different trials in subject 1; the right column shows trial-averaged responses in different subjects. Lines connecting consecutive images are used to aid visualization but do not indicate specific temporal relationships between images. Thick black lines indicate the mean across trials (left) or subjects (right). Notice that reliability is reasonably high both within and across subjects. **e**, Quantitative summary. To summarize results shown in **d**, we plot the correlation between responses to the shared515 images across all trials and all subjects. Thin white horizontal and vertical lines separate different subjects (each having 3 trials). Notice there is high reliability within each ROI, and responses are highly dissimilar across ROIs. The strong off-diagonal elements (white arrows) indicate the presence of spatial noise correlations that occur on individual trials, which is typical in fMRI[45]. Noise correlations likely reflect a combination of measurement noise (for example, head motion) and real neural activity variability (for example, arousal effects). In some cases, correlations are larger across subjects than within subjects; one explanation is that there is, to some degree, a common ROI representation and a noisy measurement of this representation obtained in one subject might actually be better correlated with a less noisy measurement of this representation obtained in a different subject. Also, the results indicate the existence of temporal ordering effects (for example, trial 1 in a given subject tends to be more correlated with trial 1 in other subjects as opposed to trials 2 or 3). This likely indicates the presence of adaptation- and/or memory-related effects in the NSD data, given that the temporal ordering of trials was fixed across subjects.



Extended Data Fig. 9 | Angle and eccentricity estimates from the NSD data. Here we show results from the analysis of the pRF experiment and results from an analogous analysis performed on trial-averaged NSD betas (see Supplementary Modeling Note 1 for details). Each panel shows an occipital view of FreeSurfer's sphere surface, and white lines indicate borders of visual areas V1–hV4 (defined based on results of the pRF experiment). Angle and eccentricity estimates are plotted using the same colormaps as in Benson et al.[30] We also plot the amount of time-series variance explained in the pRF data (variance relative to the mean signal level) and the amount of variance explained in the NSD betas (variance relative to 0% BOLD signal change). Clear retinotopic maps in early visual cortex are visible in the NSD results, including robust angle estimates even in foveal regions. In addition, there is high consistency of retinotopic estimates across the pRF and NSD datasets. There is some discrepancy in absolute eccentricity estimates at peripheral locations; this is likely due to technical differences in how modeling procedures behave for voxels near the stimulus edge.



Extended Data Fig. 10 | Design of AlexNet- and GNet-based encoding models. a, Illustration of an encoding model that predicts brain activity in a given voxel (r_{tv}) in response to images (x_t). Images are passed to nonlinear feature extractors, η_l (trapezoids), that output feature maps (grey cuboids). Feature maps are grouped, passed through an element-wise nonlinearity, $f(\cdot)$, and then multiplied pixel-wise by a spatial pooling field ($g^1,\ldots,g^N$ where superscripts index distinct groups of feature maps) that determines the region of visual space that drives voxel activity. The weighted pixel values in each feature map are then summed, reducing each feature map to a scalar value. These scalar values are concatenated across all feature maps, forming a single feature vector that is passed through another element-wise nonlinearity (left black rectangle) and then weighted by a set of feature weights, w (right black rectangle), to yield predicted voxel activity. Note that for each type of encoding model (for example, AlexNet-based encoding model, GNet-based encoding model), the feature extractors are identical for all voxels, but the spatial pooling fields and feature weights are optimized and may vary across voxels. For the AlexNet-based encoding model, the feature extractors were pre-specified, the spatial pooling fields were optimized via line search, and the feature weights w were optimized via ridge regression. For the GNet-based encoding model, stochastic gradient descent with early stopping was used to optimize the parameters of the feature extractors η_l, the spatial pooling fields $g^1,\ldots,g^N$, and the feature weights w. **b**, Illustration of spatial pooling fields. For the AlexNet model, a single isotropic 2D Gaussian pooling field (middle) selected from a set of candidates (right) was applied to all feature maps. For the GNet model, an independent, flexible pooling field (left) was applied to each group of feature maps. Applying flexible pooling fields to AlexNet leads to lower prediction accuracy overall, so we present the version that uses isotropic 2D Gaussian fields. **c**, Comparative architecture of AlexNet and GNet. AlexNet and GNet are both deep convolutional neural networks, but differ in the types and sequencing of layers (rows of the table). The first three layers are the same for both networks and correspond to the first three layers of an AlexNet trained to classify objects in the ImageNet dataset. For both networks, these shared 'pre-filtering' layers are followed by sequences of convolutional layers (rows labeled 'conv'; values indicate feature depth and convolutional filter resolution; 'str' = filter stride, 'pad' = convolutional padding), max-pooling layers ('maxpool'), batch-normalization and weight-dropout layers ('batchnorm + dropout'), adaptive averaging layers ('adaptive avg'), and fully-connected layers ('fully con.'; value indicates number of units). Feature maps in the convolutional or fully connected layers (indicated by red arrows; resolution of the feature maps in parentheses) are used as predictors of brain activity in the context of an encoding model (see **a**).

nature research

Corresponding author(s): Kendrick Kay

Last updated by author(s): Oct 10, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Psychophysics Toolbox 3.0.14, MATLAB R2018a, Meadows web-based platform (http://meadows-research.com)
Data analysis	MATLAB (https://www.mathworks.com/products/matlab.html), Python (https://www.python.org), SPM5 (https://www.fil.ion.ucl.ac.uk/spm/), FSL 5.0.7, 5.0.9, 6.0.3 (https://fsl.fmrib.ox.ac.uk/), FreeSurfer 6.0 and 7.0 (https://surfer.nmr.mgh.harvard.edu), GLMdenoise 1.4 (https://github.com/cvnlab/GLMdenoise), analyzePRF 1.2 (https://github.com/cvnlab/analyzePRF/), GLMsingle 0.9 (https://github.com/cvnlab/GLMsingle), fracridge 1.3 (https://github.com/nrdg/fracridge), ANTs 2.1.0 (http://stnava.github.io/ANTs/), MRTrix 3.0 (https://www.mrtrix.org), Dipy 1.1 (https://dipy.org), Vistasoft master branch (https://github.com/vistalab/vistasoft), Connectome Workbench 1.4.2 (https://github.com/Washington-University/workbench), PyTorch 3 (https://pytorch.org)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The NSD dataset is freely available at http://naturalscenesdataset.org. The data are hosted in the cloud, allowing researchers to exploit high-performance cloud computing to efficiently analyze the dataset. We provide both raw data in BIDS format (Gorgolewski et al., 2016) and prepared data files, along with extensive technical documentation in the NSD Data Manual. To ensure strict validation for an upcoming Algonauts prediction challenge (Cichy et al., 2019), the initial public

release will withhold the last three NSD scan sessions from each participant (about 8.4% of the NSD data). Images used for NSD were taken from the Common Objects in Context database (Lin et al., 2014) (https://cocodataset.org).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	This study collects massive amounts of data in individual subjects. Analyses demonstrated in this paper are conducted primarily at the within-subject level, demonstrating the precision and robustness of the data collected. For group-level analyses, the number of subjects used for NSD (n = 8) is sufficiently large to provide some power for statistical inference. The sample size (n = 8) was chosen based on consideration of guarding against incidental findings that occur only in a few individuals and based on consideration of subject burden (if all images had been presented to a single subject, data collection would have extended for 8 years).
Data exclusions	We implemented a subject-selection procedure in which the best 8 subjects out of a pool of 14 potential subjects (on basis of criteria such as head motion and BOLD signal strength) were selected for full NSD data acquisition. This was done to optimize the quality of the NSD dataset. For the neural network analysis, due to computational memory limitations, we used data from the best 4 out of the 8 NSD subjects in terms of signal-to-noise ratio (SNR); this analysis is intended primarily to demonstrate proof of concept, and the SNR-based selection is not expected to incur significant inferential biases. Due to image artifacts, we excluded 2/52 (4%) of the acquired T1-weighted volumes (excluded volumes were from Subject 8). Due to eyetracking noise, for the eyetracking results shown in Extended Data Figure 4, we excluded 1/24 (4%), 1/24 (4%), 7/8 (88%), 0/24 (0%), 0/24 (0%), 0/24 (0%), 0/24 (0%), and 2/8 (25%) of the acquired eyetracking runs from the 8 subjects, respectively (in aggregate: 11/160 (7%)).
Replication	This resource paper describes extensive quality checks on the data acquired from the 8 NSD subjects. We provide substantial evidence that high-quality data were obtained from all subjects. Sufficient data were obtained such that we were able to demonstrate effects at the level of individual subjects and replicate effects across multiple subjects. Note that some subjects fare better on certain quality metrics (e.g. head motion) than others. In addition, there is some variation in the total amount of data collected across subjects (e.g. between 30–40 core NSD scan sessions were acquired for each subject).
Randomization	All participants engaged in the same set of experiments. However, somewhat non-overlapping sets of stimuli were chosen for each subject. The allocation of stimuli to different subjects was done randomly from a fixed set of images pulled from the Microsoft COCO database. Given the large scale of stimulus sampling (e.g. 9,000–10,000 unique images were shown to each subject), it is likely that although the exact same images are not shown to each subject, the same general types of stimulus features are well sampled for each subject.
Blinding	Blinding is not relevant to this study given that there is little that the investigators could have done to bias the nature of the recorded data and given that the participants do not belong to any discrete groupings.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	A mixture of males and females were used (2m, 6f). All participants were healthy young adults between 19–32 years old, and all provided informed written consent. Participants were compensated at a rate of $30 per hour, plus performance bonuses.

Recruitment	Participants were recruited through advertisements to the local community and were screened based on ability to participate in this long-term neuroimaging study. In addition, we selected participants based on data quality from an initial 7T fMRI session. This selection does induce a bias towards individuals with low head motion, high cognitive performance, and strong BOLD responses. The goal of the NSD dataset is largely to create a massive dataset to inform studies of the basic mechanisms of vision and memory, and does not represent an unbiased sampling of the human population.
Ethics oversight	University of Minnesota Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type	The core NSD experiment is task-based and has an event-related design. The prf experiment is task-based and has a continuous design. The floc experiment is task-based and has an event-related design. We also collected resting-state data, as well as structural and diffusion data.
Design specifications	In the core NSD experiment, images were presented for 3 seconds, and were followed by a minimum of 1 second of gap before the next trial. Many thousands of distinct images were presented over the course of many distinct scan sessions, with a maximum number of presentations per distinct image of 3.
Behavioral performance measures	Button presses and associated reaction times for each trial in the NSD experiment were recorded. To ensure high data quality, we monitored basic response metrics, like response rate. We quantified recognition performance in the NSD experiment using signal detection theory.

Acquisition

Imaging type(s)	Functional, structural, diffusion, venogram, angiogram
Field strength	7T and 3T
Sequence & imaging parameters	The primary fMRI sequence involved gradient-echo EPI, FOV 216 mm x 216 mm, matrix size 120 x 120, slice thickness 1.8 mm, orientation axial, TR 1.6 s, TE 22.0 ms, and flip angle 62°.
Area of acquisition	Whole-brain scans including the cerebellum

Diffusion MRI ☒ Used ☐ Not used

Parameters: 99–100 directions; b-values of 0, 1,500, and 3,000; no cardiac gating

Preprocessing

Preprocessing software	A combination of custom MATLAB and Python code, FreeSurfer 6, and selected tools from SPM, FSL, ANTs, and MRTrix3.
Normalization	The NSD data were prepared in a variety of spaces including subject-native space and atlas spaces (MNI, fsaverage). Some of the data demonstrations in this paper show results in subject-native spaces; some show results that reflecting averaging in atlas spaces.
Normalization template	For preparation of data in atlas spaces, the MNI152 and fsaverage templates were used.
Noise and artifact removal	For the GLM preparation of the NSD data, the data-driven analysis method GLMdenoise and the statistical technique of ridge regression were used. These methods can account for a variety of sources of noise (e.g., physiological, motion, scanner artifacts, effects of collinearity). A version of the GLM results that omit these noise removal methods is also provided.
Volume censoring	No censoring was performed.

Statistical modeling & inference

Model type and settings	Trial-wise fMRI response amplitudes were estimated for individual voxels in individual subjects. A variety of analyses were then performed on these response amplitudes, including univariate, multivariate, RSA, and encoding models.
Effect(s) tested	We conducted rich sampling of the brain's response to a large number of complex natural scenes. The resulting measurements can now be used to test and explore a variety of different scientific hypotheses.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s)	Atlas-based regions of interest were incorporated into this resource for user convenience. In addition, a number of manually defined regions of interest based on both functional and anatomical criteria were created.

Statistic type for inference
(See Eklund et al. 2016)

This paper provides a resource in which data from all voxels are processed and made available. Thus, thresholding and specific inferential claims are largely not applicable here.

Correction

Not applicable, as voxel-wise statistical significance inferences are not a primary focus of this paper.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

For pRF estimation, we used local contrast of NSD images to predict NSD betas through a simple pRF model, using nonlinear optimization to determine parameters for each voxel/vertex. For representational similarity analysis, we constructed representational dissimilarity matrices by correlating multivariate brain activity patterns. For neural network modeling, we used either pre-trained image-computable neural network models (AlexNet, Gabor model) or brain-optimized image-computable neural network models (GNet). These models were trained on a set of training data (the non-shared NSD images) and validated on a separate set of validation data (the shared NSD images).