



Neural representational geometry underlies few-shot concept learning

Ben Sorscher^a, Surya Ganguli^{a,b}, and Haim Sompolinsky^{c,d,e,1}

Edited by Terrence Sejnowski, Salk Institute for Biological Studies, La Jolla, CA; received January 29, 2022; accepted July 18, 2022

Understanding the neural basis of the remarkable human cognitive capacity to learn novel concepts from just one or a few sensory experiences constitutes a fundamental problem. We propose a simple, biologically plausible, mathematically tractable, and computationally powerful neural mechanism for few-shot learning of naturalistic concepts. We posit that the concepts that can be learned from few examples are defined by tightly circumscribed manifolds in the neural firing-rate space of higher-order sensory areas. We further posit that a single plastic downstream readout neuron learns to discriminate new concepts based on few examples using a simple plasticity rule. We demonstrate the computational power of our proposal by showing that it can achieve high few-shot learning accuracy on natural visual concepts using both macaque inferotemporal cortex representations and deep neural network (DNN) models of these representations and can even learn novel visual concepts specified only through linguistic descriptors. Moreover, we develop a mathematical theory of few-shot learning that links neurophysiology to predictions about behavioral outcomes by delineating several fundamental and measurable geometric properties of neural representations that can accurately predict the few-shot learning performance of naturalistic concepts across all our numerical simulations. This theory reveals, for instance, that high-dimensional manifolds enhance the ability to learn new concepts from few examples. Intriguingly, we observe striking mismatches between the geometry of manifolds in the primate visual pathway and in trained DNNs. We discuss testable predictions of our theory for psychophysics and neurophysiological experiments.

few-shot learning | neural networks | ventral visual stream | population coding

A hallmark of human intelligence is the remarkable ability to rapidly learn new concepts. Humans can effortlessly learn new visual concepts from only one or a few visual examples (1–4). Several studies have revealed the ability of humans to classify novel visual concepts with 90 to 95% accuracy, even after only a few exposures (5–7). We can even acquire visual concepts without any visual examples, by relying on cross-modal transfer from language descriptions to vision. The theoretical basis for how neural circuits can mediate this remarkable capacity for “few-shot” learning of general concepts remains mysterious, despite many years of research in concept learning across philosophy, psychology, and neuroscience (8–12).

Theories of human concept learning are at least as old as Aristotle, who proposed that concepts are represented in the mind by a set of strict definitions (13). Modern cognitive theories propose that concepts are mentally represented instead by a set of features, learned by exposure to examples of the concept. Two such foundational theories include prototype learning, which posits that features of previously encountered examples are averaged into a set of “prototypical” features (14), and exemplar learning, which posits that the features of all previously encountered examples are simply stored in memory (8). However, neither theory suggests how these features might be represented in the brain. In laboratory experiments, these features are either constructed by hand by generating synthetic stimuli that vary along a predefined set of latent features (15, 16) or are indirectly inferred from human similarity judgements (17–19).

In this work, we introduce a theory of concept learning in neural circuits based on the hypothesis that the concepts we can learn from a few examples are defined by tight geometric regions in the space of high-dimensional neural population representations in higher-level sensory brain areas. Indeed, in the case of vision, decades of research have revealed a series of representations of visual stimuli in neural population responses along the ventral visual pathway, including V1, V2, and V4, culminating in a rich object representation in the inferotemporal (IT) cortex (20–22), allowing a putative downstream neuron to infer the identity of an object based on the pattern of IT activity it elicits (23, 24).

Significance

Humans have an extraordinary ability to learn new concepts from just a few examples, but the cognitive mechanism behind this ability remains mysterious. A long line of work in systems neuroscience has revealed that familiar concepts can be discriminated by the patterns of activity they elicit across neurons in higher-order cortical layers. How might downstream brain areas use this same neural code to learn new concepts? Here, we show that this ability is governed by key geometric properties of the neural code. We find that these geometric quantities undergo orchestrated transformations along the primate visual pathway and along artificial neural networks, so that in higher-order layers, a simple plasticity rule may allow novel concepts to be learned from few examples.

Author affiliations: ^aDepartment of Applied Physics, Stanford University, Stanford, CA 94305; ^bStanford Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA 94305; ^cCenter for Brain Science, Harvard University, Cambridge, MA 02138; ^dRacah Institute of Physics, Hebrew University, Jerusalem 9190501, Israel; and ^eEdmond and Lily Safra Center for Brain Sciences, Hebrew University, Jerusalem 9190501, Israel

Author contributions: B.S., S.G., and H.S. designed research; B.S. and H.S. performed research; and B.S., S.G., and H.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2022 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nd/4.0/).

¹To whom correspondence may be addressed. Email: haim.sompolinsky@mail.huji.ac.il.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2200800119/-DCSupplemental>.

Published October 17, 2022.

We also hypothesize that sensory representations in IT, acquired through a lifetime of experience, are sufficient to enable rapid learning of novel visual concepts based on just a few examples, without any further plasticity in the representations themselves, by a downstream population of neurons with afferent plastic synapses that integrate a subset of nonplastic IT neural responses.

Our approach is corroborated by recent work demonstrating that artificial deep neural networks (DNNs) pretrained on a diverse array of concepts can be taught to classify new concepts from few examples by simply training linear readouts of their hidden unit representations (25–27). To understand this, we address the questions: What makes a representation good for rapidly learning novel concepts? And does the brain contain such representations? We introduce a mathematical theory, which reveals that few-shot learning performance is governed by four simple and readily measurable geometric quantities. We test our theory using neurophysiological recordings in macaques, as well as representations extracted from DNN models of the primate visual hierarchy, which have been shown to predict neural population responses in V1, V4, and IT (28, 29), the similarity structure of object representations in IT (30), and human performance at categorizing familiar objects (19, 31). We find that in both DNNs and the primate visual hierarchy, neural activity patterns undergo orchestrated geometric transformations so that few-shot learning performance increases from each layer to the next and from the retina to V1 to V4 to IT. Intriguingly, despite these common patterns in performance, our theory reveals fundamental differences in neural geometry between primate and DNN hierarchies, thereby providing key targets for improving models.

We further leverage our theory to investigate neurally plausible models of multimodal concept learning, allowing neural representations of linguistic descriptors to inform the visual system about visually novel concepts. Our theory also reveals that long-held

intuitions and results about the relative performance of prototype and exemplar learning are completely reversed when moving from low-dimensional concepts with many examples, characteristic of most laboratory settings, to high-dimensional naturalistic concepts with very few examples. Finally, we make testable predictions not only about overall performance levels, but, more importantly, about salient patterns of errors, as a function of neural population geometry, that can reveal insights into the specific strategies used by humans and nonhuman primates in concept learning.

Results

A. Accurate Few-Shot Learning with Nonplastic Representations.

To investigate how the geometry of neural representations affects few-shot learning performance, we obtain neural representations from a DNN that has been shown to be a good model of object representations in IT cortex (32) (Fig. 1, ResNet50; Section N). The DNN is trained to classify 1,000 naturalistic visual concepts from the ImageNet1k dataset (e.g., “bee” and “acorn”). To study novel-concept learning, we select a set of 1,000 new visual concepts, never seen during training, from the ImageNet21k dataset (e.g., “coati” and “numbat”; Fig. 1A; Section J). We examine the ability to learn to discriminate each pair of new concepts, given only a few training images, by learning to classify the activity patterns these training images elicit across IT-like neurons in the feature layer of the DNN (Fig. 1B). Many decision rules are possible, but we find that in the few-shot regime, typical choices perform no better than a particularly simple, biologically plausible classification rule: prototype learning (Section H). Prototype learning is performed by averaging the activity patterns elicited by the training examples into concept prototypes, \bar{x}^a , \bar{x}^b (Fig. 1B). A test image is then classified by comparing the activity pattern it elicits to each of the prototypes and identifying it with the closest prototype. This classification

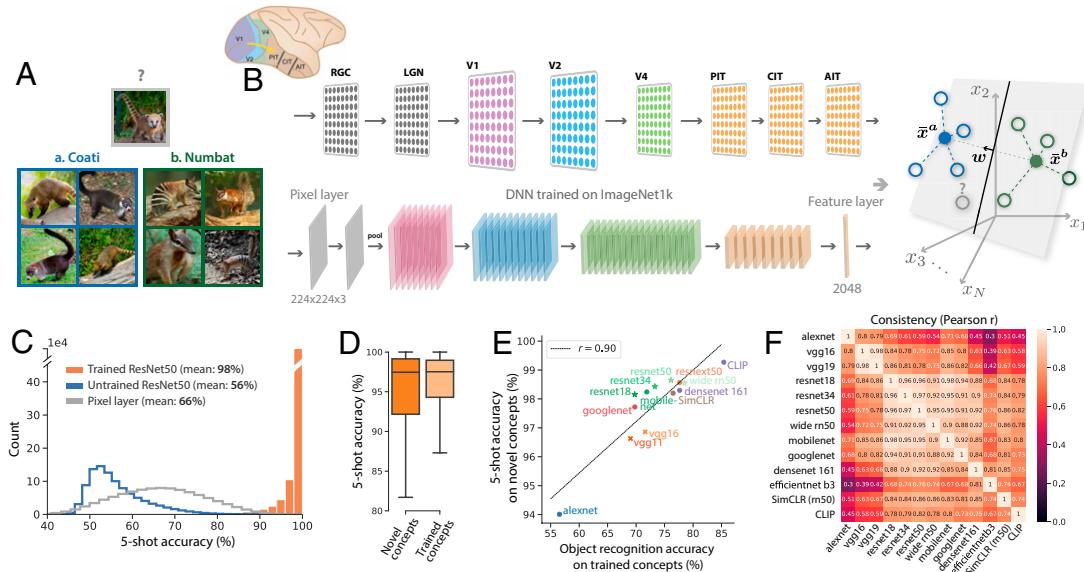


Fig. 1. Concept-learning framework and model performance. (A) Example four-shot learning task: Does the test image in the gray box contain a “coati” (blue) or a “numbat” (green), given four training examples of each? (B) Each training example is presented to the ventral visual pathway (*Upper*), modeled by a trained DNN (*Lower*), eliciting a pattern of activity across IT-like neurons in the feature layer. We model concept learning as learning a linear readout w to classify these activity patterns, which can be thought of as points in a high-dimensional activity space (open circles, *Right*). In the case of prototype learning, activity patterns are averaged into prototypes \bar{x}^a , \bar{x}^b (filled circles), and w is pointed along the difference between the prototypes $w = \bar{x}^a - \bar{x}^b$, passing through their midpoint. To evaluate generalization accuracy, we present a test image and determine whether its neural representation (gray open circle) is correctly classified. (C) Generalization accuracy is very high across 1,000 × 999 pairs of novel visual concepts from the ImageNet21k dataset (orange). In comparison, test accuracy is poor when using a randomly initialized DNN (blue) or when learning a linear classifier in the pixel space of input images (gray). (D) Test accuracy on novel concepts (dark orange) is only slightly lower than test accuracy on familiar concepts seen during training (light orange). (E) Performance on the object-recognition task used to train DNNs correlates with their ability to generalize to novel concepts, given few examples ($r = 0.92$, $P < 1 \times 10^{-4}$), across a variety of DNN architectures. (F) Different DNN architectures are consistent in the pattern of errors they make across the 1,000 × 999 novel concepts ($P < 1 \times 10^{-10}$).

rule can be performed by a single downstream neuron that adjusts its synaptic weights so that its weight vector w points along the difference between the prototypes, $w = \bar{x}^a - \bar{x}^b$ (Fig. 1B). Despite its simplicity, prototype learning has been shown to achieve good few-shot learning performance (33), and we find that on our evaluation set of 1,000 novel concepts from the ImageNet dataset, it achieves an average test accuracy of 98.6% given only 5 training images of each concept (Fig. 1C), only slightly lower than the accuracy obtained by prototype learning on the same 1,000 familiar concepts that were used to train the DNN (Fig. 1D). When only one training example of each concept is provided (one-shot learning), prototype learning achieves a test accuracy of 92.0%. In contrast, the performance of prototype learning applied to representations in the retina-like pixel layer to the feature layer of an untrained, randomly initialized DNN is around 50 to 65% (Fig. 1C), indicating that the neural representations learned over the course of training the DNN to classify 1,000 concepts are powerful enough to facilitate highly accurate few-shot learning of novel concepts, without any further plasticity. Consistent with this, we find that DNNs that perform better on the ImageNet1k classification task also perform better at few-shot learning of novel concepts (Fig. 1E).

For a finer-resolution analysis, we examine the pattern of errors each model makes across all pairs of novel concepts. We find that different models are consistent in which novel concepts they find easy/hard to learn (Fig. 1F; $P < 1 \times 10^{-10}$; chance is zero), although the consistency is noticeably lower for older architectures (e.g., Alexnet or VGG). We also observe that good few-shot

learning performance does not require pretraining on a large dataset of labeled examples, as models trained in a self-supervised or unsupervised [e.g., SimCLR (34) or CLIP (27)] manner also achieve high performance (Fig. 1E), are consistent in the patterns of errors they make (Fig. 1F), and exhibit similar representational geometry (*SI Appendix*, Fig. 5). Intriguingly, across models tested, the few-shot learning-error patterns reveal a pronounced asymmetry on many pairs of concepts (*SI Appendix*, section 7). For instance, models may be much more likely to classify a test example of a “coati” as a “numbat” than a “numbat” as a “coati” (*SI Appendix*, Fig. 1). Importantly, this asymmetry is not only a feature of prototype learning; diverse decision rules (e.g., exemplar learning or support vector machines) exhibit similar asymmetry in the few-shot regime (*SI Appendix*, Fig. 10).

These results raise several fundamental theoretical questions. Why does a DNN trained on an image-classification task also perform so well on few-shot learning? What properties of the derived neural representations empower high few-shot learning performance? Furthermore, why are some concepts easier than others to learn (Fig. 1C), and why is the pairwise classification error asymmetric (*SI Appendix*, Fig. 1)? We answer these questions by introducing an analytical theory that predicts the generalization error of prototype learning, based on the geometry of neural population responses.

B. A Geometric Theory of Prototype Learning. Patterns of activity evoked by examples of any particular concept define a concept manifold (Fig. 2). Although their shapes are likely

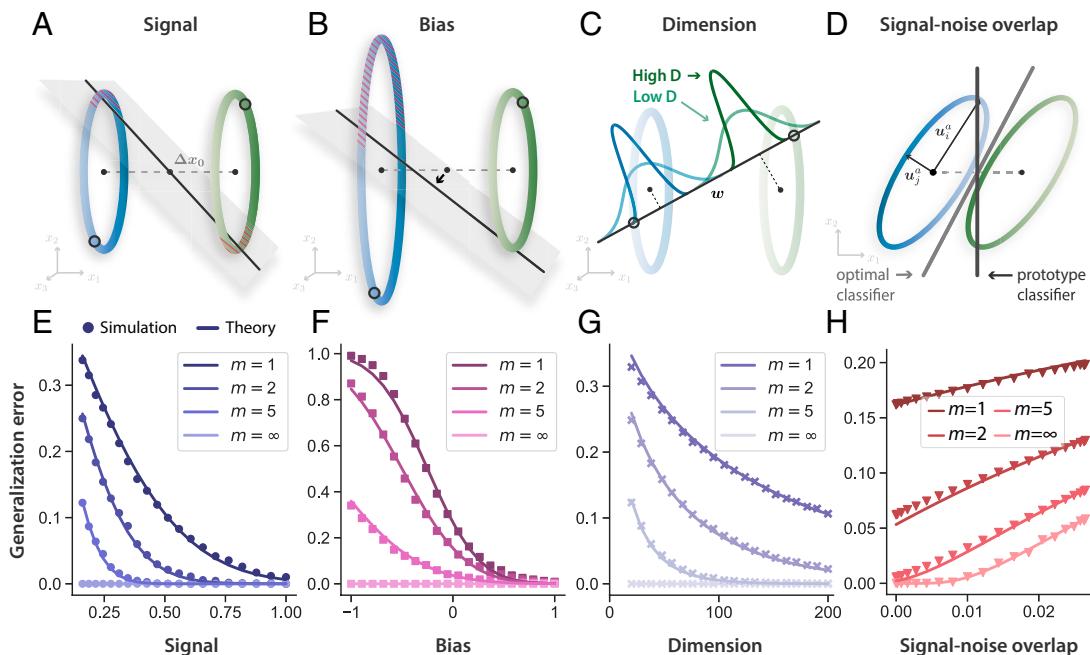


Fig. 2. Geometry of few-shot learning. Patterns of activity evoked by examples of each concept define a concept manifold, which we approximate as a high-dimensional ellipsoid. The generalization error (red hashed area) of a prototype classifier w is governed by four geometric properties of these manifolds (Eq. 1), shown schematically in A–D. (A) Signal refers to the pairwise separation between concept manifolds, $\|\Delta\mathbf{x}_0\|^2$. Manifolds that are better separated are more easily discriminated by a linear classifier (gray hyperplane) given few training examples (open circles). (B) Bias: As the radius of one manifold grows relative to the other, the decision hyperplane shifts toward the manifold with the larger radius. Hence, the generalization error on the larger manifold increases, while the error on the smaller manifold decreases. Although the tilt of the decision hyperplane relative to the signal direction averages out over different draws of the training examples, this bias does not. (C) Dimension: In high dimensions, the projection of each concept manifold onto the linear readout direction w concentrates around its mean. Hence, high-dimensional manifolds are easier to discriminate by a linear readout. (D) Signal–noise overlap: Pairs of manifolds whose noise directions \mathbf{u}_i overlap significantly with the signal direction $\Delta\mathbf{x}_0$ have higher generalization error. Even in the limit of infinite training examples, $m \rightarrow \infty$, the prototype classifier (dark gray) cannot overcome signal–noise overlaps, as it has access only to the manifold centroids. An optimal linear classifier (light gray), in contrast, can overcome signal–noise overlaps using knowledge of the variability around the manifold centroids. (E–H) Behavior of generalization error (with $m = 1, 2, 5, \infty$) in relation to each geometric quantity in A–D. Theoretical predictions are shown as dark lines, and few-shot learning experiments on synthetic ellipsoidal manifolds are shown as points. (F) When bias is large and negative, generalization error can be greater than chance (0.5). (G) Generalization error decreases when manifold variability is spread across more dimensions. (H) In the presence of signal–noise overlaps, generalization error remains nonzero, even for arbitrarily large m . Details on simulation parameters are given in Section M.

complex, we find that when concept manifolds are high-dimensional, the generalization error of prototype learning can be accurately predicted based only on each manifold's centroid, \mathbf{x}_0 , and radii, R_i , along a set of orthonormal basis directions \mathbf{u}_i , $i = 1, \dots, N$ (*SI Appendix, section 3*), capturing the extent of natural variation of examples belonging to the same concept. A useful measure of the overall size of these variations is the mean squared radius $R^2 \equiv \frac{1}{N} \sum_{i=1}^N R_i^2$.

Our theory of prototype learning to discriminate pairs of new concepts (a, b) predicts that the average error of m -shot learning on test examples of concept a is given by $\varepsilon_a = H(\text{SNR}_a)$, where $H(\cdot)$ is the Gaussian tail function $H(x) = \int_x^\infty dt e^{-t^2/2}/\sqrt{2\pi}$ (*SI Appendix, section 3C*). The quantity SNR_a is the signal-to-noise ratio (SNR) for manifold a , whose dominant terms are given by,

$$\text{SNR}_a = \frac{1}{2} \frac{\|\Delta \mathbf{x}_0\|^2 + (R_b^2 R_a^{-2} - 1)/m}{\sqrt{D_a^{-1}/m + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2/m + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2}}. \quad [1]$$

A full expression and derivation are given in *SI Appendix, section 3C*. The SNR depends on four interpretable geometric properties, depicted schematically in Fig. 2 A–D. Their effect on the generalization error is shown in Fig. 2 E–H. The generalization error for tasks involving discriminating more than two novel concepts, derived in *SI Appendix, section 3D*, is governed by the same geometric properties. We now explain each of these properties.

- Signal.** $\|\Delta \mathbf{x}_0\|^2 \equiv \|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2/R_a^2$ represents the pairwise distance between the manifolds' centroids, \mathbf{x}_0^a and \mathbf{x}_0^b , normalized by R_a^2 (Fig. 2A). As the pairwise distance between manifolds increases, they become easier to separate, leading to higher SNR and lower error (Fig. 2E). We denote $\|\Delta \mathbf{x}_0\|^2$ as the signal and $\Delta \mathbf{x}_0$ as the signal direction.
- Bias.** $R_b^2 R_a^{-2} - 1$ represents the average bias of the linear classifier (Fig. 2B). Importantly, when manifold a is larger than manifold b , the bias term is negative, predicting a lower SNR for manifold a . This asymmetry results from the fact that the classification hyperplane is biased toward the larger manifold (Fig. 2F) and causes the asymmetry in generalization error observed above (Section A). This bias is unique to few-shot learning. As can be seen in Eq. 1, its effect on SNR diminishes for large numbers of training examples m .
- Dimension.** In our theory, $D_a \equiv (R_a^2)^2 / \sum_{i=1}^N (R_i^a)^4$, known as the “participation ratio” (35), quantifies the number of dimensions along which the concept manifold varies significantly, which is often much smaller than the number of neurons, N (Section L). Previous works have shown that low-dimensional manifolds have desirable properties for object recognition, allowing a greater number of familiar object to be perfectly classified (36, 37) and conferring robustness (38). Intriguingly, Eq. 1 reveals that for few-shot learning, high-dimensional manifolds are preferred. This enhanced performance is due to the fact that few-shot learning involves comparing a test example to the training examples of each novel concept. In low dimensions, the distance from the test example to each of the training examples varies significantly, contributing a noise term to the SNR. But in high dimensions, these distances concentrate around their typical value, and, hence, the noise term is suppressed as D_a^{-1} . Note that this benefit of high-dimensional representations is unique to few-shot learning, since the noise term can also be suppressed by averaging over many training examples (by a factor of $1/m$; Eq. 1).

4 Signal–noise overlap. $\|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2$ and $\|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2$ quantify the overlap between the signal direction $\Delta \mathbf{x}_0$ and the manifold axes of variation $\mathbf{U}_a \equiv [\mathbf{u}_1^a R_1^a, \dots, \mathbf{u}_N^a R_N^a]/\sqrt{R_a^2}$ and $\mathbf{U}_b \equiv [\mathbf{u}_1^b R_1^b, \dots, \mathbf{u}_N^b R_N^b]/\sqrt{R_b^2}$ (see Fig. 2D and Section L for details). Generalization error increases as the overlap between the signal and noise directions increases (Fig. 2H). We note that signal–noise overlaps decrease as the dimensionality D_a increases.

Effect of number of training examples. As the number of training examples, m , increases, the prototypes more closely match the true manifold centroids; hence, the bias and the first two noise terms in Eq. 1 decay as $1/m$. However, the last noise term, $\|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2$, does not vanish, even when the centroids are perfectly estimated, as it originates from variability in the test examples along the signal direction (Fig. 2H; $m = \infty$). Thus, in the limit of a large number of examples m , the generalization error does not vanish, even if the manifolds are linearly separable. The SNR instead approaches the finite limit $\text{SNR}_a(m \rightarrow \infty) = \frac{1}{2} \|\Delta \mathbf{x}_0\|^2 / \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|_2$, highlighting the failure of prototype learning at large- m compared to an optimal linear classifier (Fig. 2D). However, in the few-shot (small- m) regime, prototype learning is close to the optimal linear classifier (Section H).

To evaluate our theory, we perform prototype learning experiments on synthetic concept manifolds, constructed as high-dimensional ellipsoids. We find good agreement between theory and experiment for the dependence of generalization error on each of the four geometric quantities and on the number of examples (see Fig. 2 E–H and Section M for details).

C. Geometric Theory Predicts the Error of Few-Shot Learning in DNNs. We next test whether our geometric theory accurately predicts few-shot learning performance on naturalistic visual concept manifolds, using the neural representations derived from DNNs studied in Section A. For each pair of concept manifolds, we estimate the four geometric quantities defined above (Section K) and predict generalization error via Eq. 1, finding excellent agreement between theory and numerical simulations (Fig. 3A and *SI Appendix, Fig. 3*), despite the obviously complex shape of these manifolds.

Our theory further allows us to dissect the specific contribution of each of the four geometric properties of the concept manifolds to the SNR, elucidating whether errors arise from small signal, negative bias, low dimension, or large signal–noise overlap. We dissect the contributions of signal and bias in Fig. 3B and the contributions of dimension and signal–noise overlap in Fig. 3C, along with specific illustrative examples highlighting, for instance, the strong asymmetry of few-shot learning (e.g., “Prairie gentian” vs. “Plumbago”).

DNNs perform complex nonlinear transformations to map images from the retina-like pixel layer to the feature layer. In Fig. 3D, we examine the effect these transformations have on SNR, signal, dimension, and signal–noise overlap for each pair of concepts. We find that almost all concepts become better-separated (i.e., higher signal), higher-dimensional, and more orthogonal (i.e., lower signal–noise overlap) as they are mapped from the pixel layer to the feature layer. In the following section, we study the evolution of these geometric features along the layers of DNNs.

One might suspect that because the DNN was explicitly optimized to map the 1,000 concepts on which it was trained to linearly separable regions in feature space, the geometry of these 1,000 trained concepts would differ substantially from the geometry of the 1,000 novel concepts we study in this work. Nevertheless, although the signal is slightly higher for trained

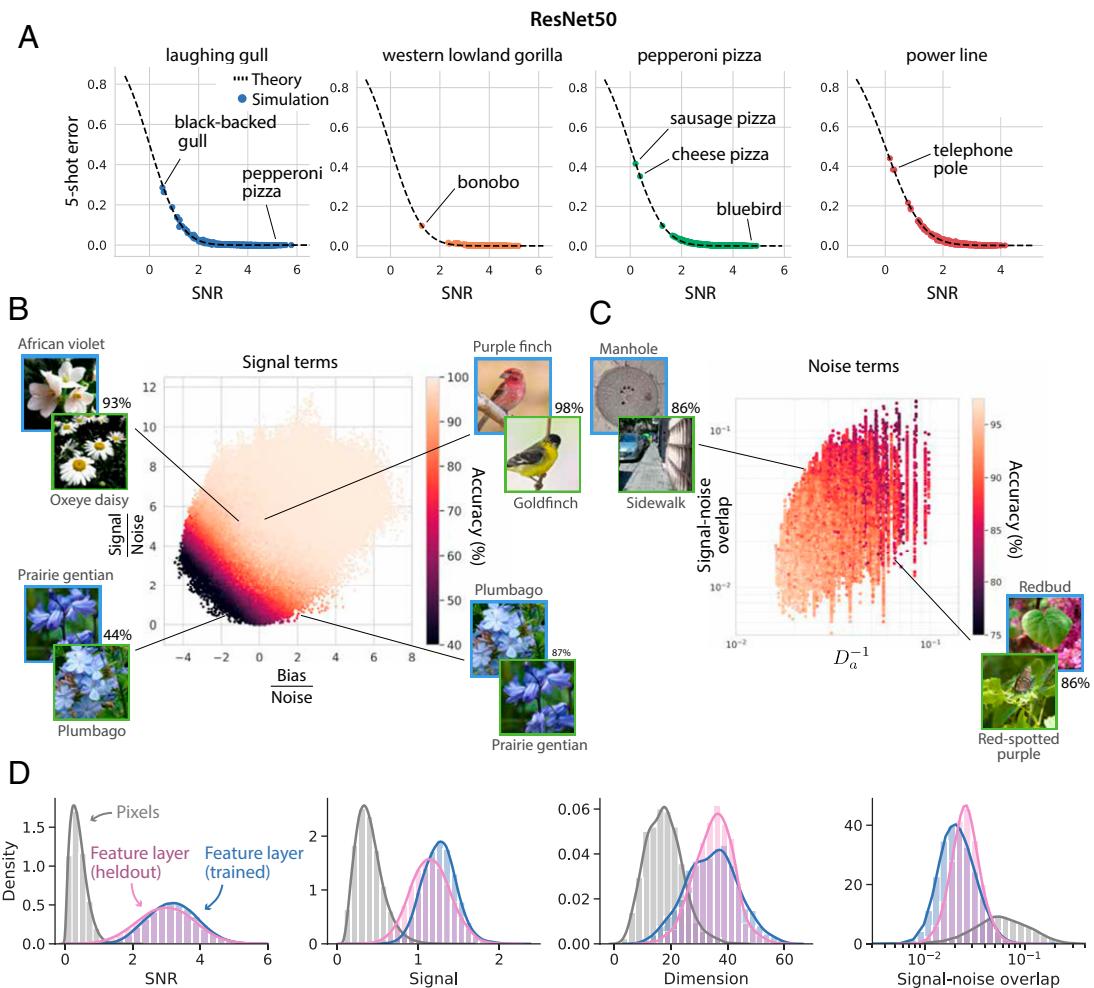


Fig. 3. Geometric theory predicts the generalization error of few-shot learning in DNNs. (A) We compare the predictions of our theory to the few-shot learning experiments performed in Fig. 1. Each panel plots the generalization error of one novel visual concept (e.g., “laughing gull”) against all 999 other novel visual concepts (e.g., “black-backed gull” or “pepperoni pizza”). Each point represents the average generalization error on one such pair of concepts. x axis: SNR (Eq. 1) obtained by estimating neural manifold geometry. y axis: Empirical generalization error measured in few-shot learning experiments. Theoretical prediction (dashed line) shows a good match with experiments. Error bars, computed over many draws of the training and test examples, are smaller than the symbol size. We annotate each panel with specific examples of novel concept pairs, indicating their generalization error. Additional examples are included in *SI Appendix*, Fig. 3. (B) Signal terms: We dissect the generalization accuracy on each pair of novel concepts into differential contributions from the signal and bias (Eq. 1). We plot each pair of visual concepts in the resulting signal-bias plane, where both signal and bias are normalized by the noise so that one-shot learning accuracy (color, dark to light) varies smoothly across the plot. Specific examples of concept pairs are included to highlight the behavior of generalization accuracy with respect to each quantity. For example, the pair “Purple finch” vs. “Goldfinch” has a large signal and a bias close to zero and, hence, a very high one-shot accuracy (98%). The pair “African violet” vs. “Oxeye daisy”, in contrast, has a large signal, but a large negative bias; hence, its accuracy is lower (93%). Pairs with large negative bias and small signal may have very asymmetric generalization accuracy. For instance, “Prairie gentian” vs. “Plumbago” has an accuracy of 87%, while “Plumbago” vs. “Prairie gentian” has an accuracy of 44%. For each pair of concepts, test examples are drawn from the upper left concept in blue. (C) Noise terms: We dissect the contributions of dimensionality and signal-noise overlap to generalization error. Because the variability of the signal terms is much larger than that of the noise terms, we include only pairs of concepts whose signal falls within a narrow range, so that we can visually assess whether one-shot accuracy (color, dark to light) is due to large dimensionality, small signal-noise overlaps, or both. (D) We compare histograms of SNR, signal, dimension, and signal-noise overlap across all pairs of novel concepts in the retina-like pixel layer (gray) to those in the feature layer (pink, “heldout”), revealing dramatic changes in manifold geometry empowering high few-shot learning performance. We further compare to the geometry of the 1,000 concepts used to train the network (blue histograms, “trained”), indicating that the geometry of trained and heldout concepts is largely similar.

concepts and the signal–noise overlaps slightly lower (Fig. 3D), the geometry is largely similar between trained and novel concepts, indicating that training on 1,000 concepts has furnished the DNN with a fairly general-purpose feature extractor. We show that novel-concept learning performance improves consistently as the number of concepts seen during training increases, due to smooth changes in manifold geometry in *SI Appendix*, Fig. 7.

Finally, we find that the geometry of concept manifolds in the DNN encodes a rich semantic structure, including a hierarchical organization, which reflects the hierarchical organization of visual concepts in the ImageNet dataset (*SI Appendix*, section 7). We examine the relationship between semantic structure and geometry in *SI Appendix*, Fig. 1 and provide examples of

concepts with small and large manifold dimension and radius in *SI Appendix*, Fig. 8.

D. Concept Learning along the Visual Hierarchy. How the ventral visual hierarchy converts low-level retinal representations into higher-level IT representations useful for downstream tasks constitutes a fundamental question in neuroscience. We first examine this transformation in models of the ventral visual hierarchy and, in the next section, compare it to the macaque hierarchy. Our theory enables us to not only investigate the performance of few-shot concept learning along successive layers of the visual hierarchy, but also obtain a finer-resolution decomposition of this performance into the geometric properties of the concept manifolds in each layer.

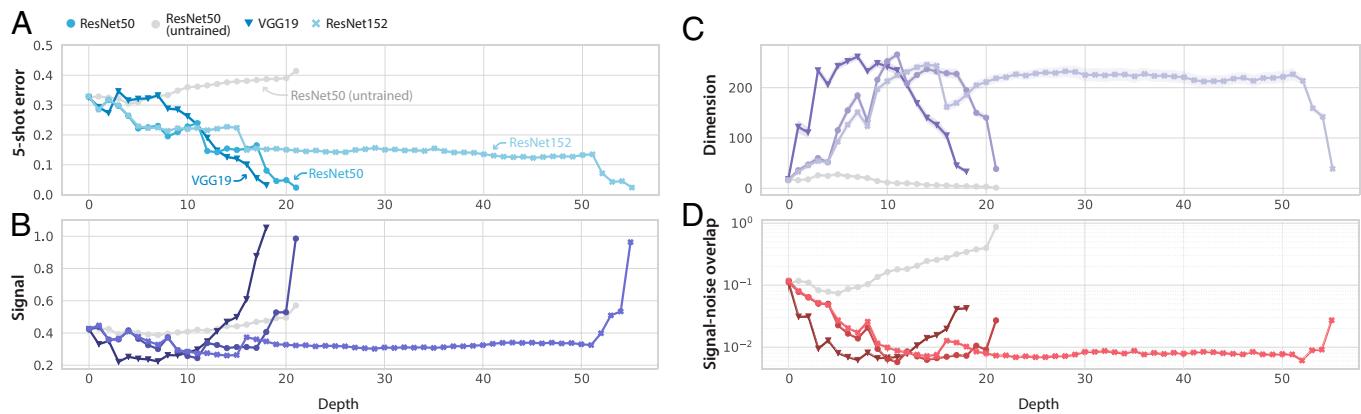


Fig. 4. Few-shot learning improves along the layers of a trained DNN, due to orchestrated transformations of concept manifold geometry. Each panel shows the layerwise behavior of one quantity in three different trained DNNs, as well as an untrained ResNet50 (gray). Lines and markers indicate mean value over 100×99 pairs of objects; surrounding shaded regions indicate 95% CIs (often smaller than the symbol size). (A) Few-shot generalization error decreases roughly monotonically along the layers of each trained DNN. (B) Signal is roughly constant across the early layers of the trained network and increases sharply in the later layers of the DNN. (C) Dimension grows dramatically in the early layers of the network and shrinks again in the late layers of network. (D) Signal-noise overlap decreases from pixel layer to feature layer, first decreasing sharply in the early layers of the network and then increasing slightly in the later layers of the network.

We find that, across a range of networks, the ability to learn novel concepts from few examples improves consistently along the layers of the network (Fig. 4*A*). In the last few layers of the network, the improvement in performance is due to a sharp increase in the signal (pairwise separability) of concept manifolds (Fig. 4*B*). In the earlier layers, the improvement in performance is due to a massive expansion in manifold dimensionality (Fig. 4*C*) and a suppression of signal–noise overlaps (Fig. 4*D*). We note that as the signal increases in the last few layers of the network, the dimensionality drops back down, and the signal–noise overlaps increase, perhaps reflecting a tradeoff between pairwise separability of manifolds and their dimensionality. Nonmonotonicity of dimension has also been observed by using capacity-related measures of dimension (36) and nonlinear measure of intrinsic dimensionality (39, 40). Interestingly, Fig. 4 indicates that ResNets exhibit long stretches of layers with apparently little to no change in manifold geometry.

E. Concept Learning Using Primate Neural Representations.

We next turn to the primate visual hierarchy and investigate the geometry of concept manifolds obtained via recordings of macaque V4 and IT in response to 64 synthetic visual concepts, designed to mimic naturalistic stimuli (24). The results in Fig. 5*A* show that our geometric theory is able to correctly predict the generalization error of few-shot learning simulations using the activation patterns of real neurons in the V4 and IT cortex of behaving animals. Furthermore, prototype learning achieves an average five-shot accuracy of 84% across all 64×63 pairs of visual concepts; slightly better than the performance achieved by the AlexNet DNN (80%) on the same set of visual concepts and worse than ResNet50 (93%), consistent with previous experiments on object-recognition performance in primates (24). We predict that performance based on IT neurons will improve when evaluated on more naturalistic visual stimuli than the grayscale, synthetic images used here (as suggested by the increased performance of DNNs, e.g., from 80 to 94% for AlexNet when using novel stimuli from ImageNet). Beyond overall performance, we find that the error patterns in IT and DNNs are consistent (Fig. 5*B* and *C*; $P < 1 \times 10^{-10}$), all exhibiting a prominent block diagonal structure, reflecting the semantic structure of the visual concepts (24).

How does concept manifold geometry evolve along the visual hierarchy? We compute the SNR in IT, V4, and a simulated population of V1-like neurons (Section O) and find that

few-shot learning performance increases along the visual hierarchy from just above chance in the pixel layer and V1 to 69% in V4 and 84% in IT, approximately matching the corresponding layers in trained DNNs that are most predictive of each cortical area under the BrainScore metric (32) (Fig. 5*D*). Despite this similarity, when we decompose the SNR into its finer-grained geometric components, we find a striking mismatch between visual cortex and trained DNNs. In particular, while manifold dimensionality in DNNs undergoes a massive expansion in the early to intermediate layers, reaching its maximum in the layer most predictive of V4 (32), dimensionality in the visual hierarchy reaches its minimum in V4 (Fig. 5*D, Upper Right*). A similar mismatch can be observed for signal–noise overlaps (Fig. 5*D, Lower Right*). To ensure that this mismatch is not simply due to our choice to measure dimensionality using the participation ratio, we repeated these analyses using nonlinear methods for estimating intrinsic dimensionality and found consistent results (*SI Appendix*, Fig. 4). The lower dimensionality and larger signal–noise overlaps in V4 compared to DNNs are compensated by an enhanced signal (Fig. 5*D, Lower Left*), so that overall performance is similar, despite these differences in representational geometry.

F. How Many Neurons Are Required for Concept Learning?

Until now, we have assumed that a downstream cortical neuron has access to the entire neural representation. Can a more realistic neuron, which only receives inputs from a small fraction of IT-like neurons, still perform accurate few-shot learning? Similarly, can a neuroscientist who only records a few hundred neurons reliably estimate the geometry of concept manifolds (as we have attempted above)?

Here, we answer these questions by drawing on the theory of random projections (41–43) to estimate the effect of subsampling a small number of M neurons (see *SI Appendix*, section 5 for details). We find that subsampling causes distortions in the manifold geometry that decrease both the SNR and the estimated dimensionality, as a function of the number of recorded neurons M ,

$$\text{SNR}(M) = \frac{\text{SNR}_\infty}{\sqrt{1 + D_\infty/M}}, \quad D^{-1}(M) = D_\infty^{-1} + M^{-1}, \quad [2]$$

where SNR_∞ and D_∞ are the asymptotic SNR and dimensionality, given access to arbitrarily many neurons (*SI Appendix*, section 5). However, these distortions are negligible when M is large compared to the asymptotic dimensionality D_∞ . Indeed,

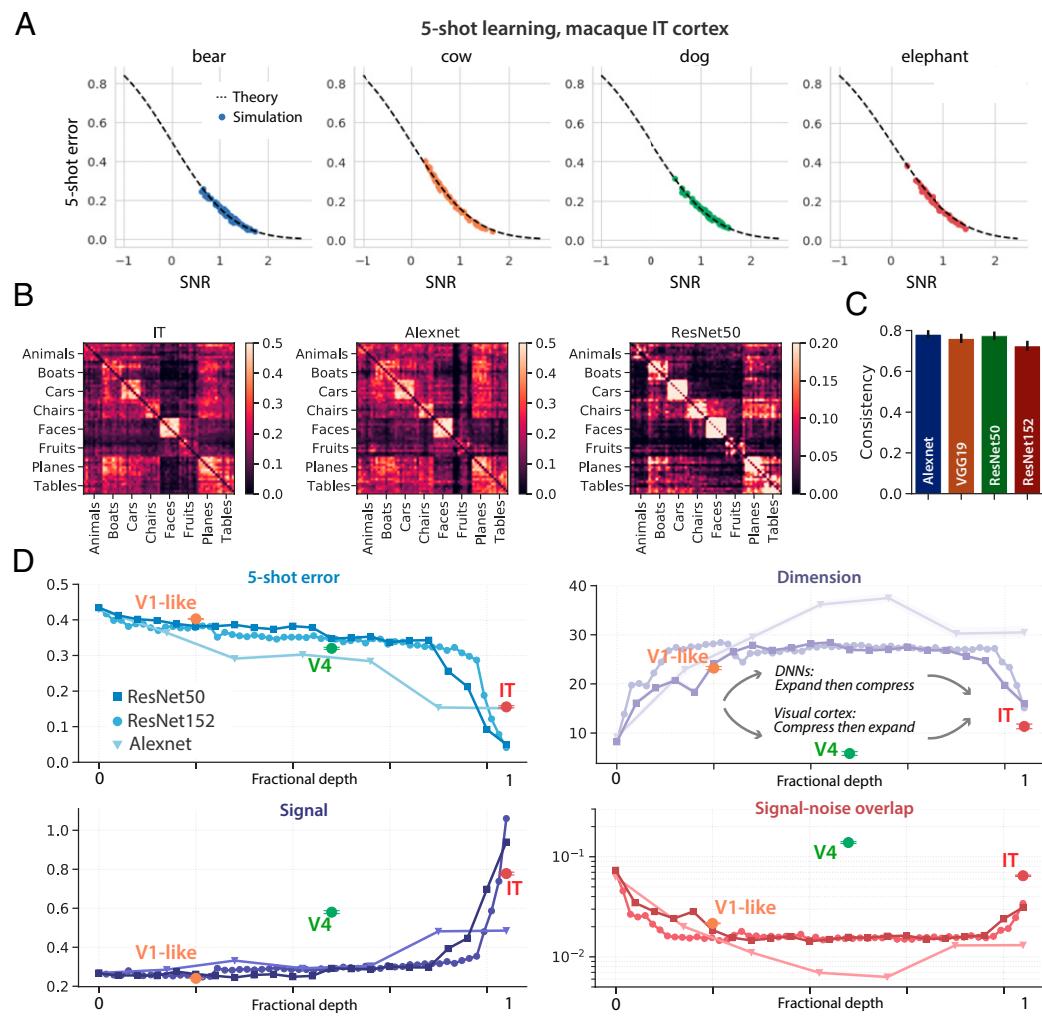


Fig. 5. Concept learning and manifold geometry in macaque IT. (A) Five-shot prototype learning experiments performed on neural population responses of 168 recorded neurons in IT (24). Each panel shows the generalization error of one visual concept (e.g., “bear”) against all 63 other visual concepts (e.g., “cow” or “dog”). Each point represents the average generalization error on one such pair of concepts. x axis: SNR (Eq. 1) obtained by estimating manifold geometry. y axis: Empirical generalization error measured in few-shot learning experiments. The result shows a good fit to the theory (dashed line). Error bars, computed over many draws of the training and test examples, are smaller than the symbol size. (B) Error pattern of five-shot learning in IT reveals a clear block diagonal structure, similar to the error patterns in AlexNet and ResNet50. (C) Error patterns in DNNs are consistent with those in IT (Pearson r , $P < 1 \times 10^{-10}$). Error bars are computed by measuring consistency over random subsets of 500 concept pairs. (D) Few-shot learning improves along the ventral visual hierarchy from pixels to V1 to V4 to IT, due to orchestrated transformations of concept manifold geometry. We compute the geometry of concept manifolds in IT, V4, and a simulated population of V1-like neurons (see Section E for details). Each panel shows the evolution of one geometric quantity along the visual hierarchy. The layerwise behavior of a trained ResNet50 (blue), Alexnet (light blue), and an untrained ResNet50 (gray) are included for comparison. We align V1, V4, and IT to the most similar ResNet layer under the BrainScore metric (32) (see Section E for details). Overall performance exhibits a good match between the primate visual pathway and trained DNNs, but individual geometric quantities display a marked difference. In particular, in DNNs, dimensionality expands dramatically in the early layers and contracts in the later layers, while in the primate visual pathway, dimensionality contracts from the V1-like layer to V4, then expands from V4 to IT. Similarly, in DNNs, signal-noise overlaps are suppressed in the early layers, then grow in the late layers, while in the primate visual pathway, signal-noise overlaps grow from the V1-like layer to V4, then are suppressed from V4 to IT. Lines and markers indicate mean value over 64×63 pairs of concepts; surrounding error bars and shaded regions indicate 95% CIs.

in both macaque IT and a trained DNN model (Fig. 6), a downstream neuron receiving inputs from only about 200 neurons performs essentially similarly to a downstream neuron receiving inputs from all available neurons (Fig. 6 A and B), and with recordings of about 200 IT neurons, the estimated dimensionality approaches its asymptotic value (Fig. 6 C and D; $D_\infty \approx 35$ in the trained DNN and $D_\infty \approx 12$ in the macaque IT).

G. Visual Concept Learning Without Visual Examples. Humans also possess the remarkable ability to learn new visual concepts using only linguistic descriptions, a phenomenon known as zero-shot learning (Fig. 7A). The rich semantic structure encoded in the geometry of visual concept manifolds (*SI Appendix*, Fig. 1) suggests that a simple neural mechanism might underlie this capacity, namely, learning visual concept prototypes from

nonvisual language representations. To test this hypothesis, we obtain language representations for the names of the 1,000 familiar visual concepts used to train our DNN, from a standard word-vector embedding model (44) trained to produce a neural representation of all words in English based on cooccurrence statistics in a large corpus (Section P).

Drawing inspiration from previous works that have trained mappings between visual embeddings and word embeddings (33, 45) or between word embeddings in multiple languages (46), we seek to learn a simple mapping between the language and vision domains (Fig. 7B). Remarkably, despite being optimized on independent objectives (word cooccurrence and object recognition), we find that the language and vision representations can be aligned by a simple linear isometry (rotation, translation, and overall scaling). Furthermore, this alignment generalizes to

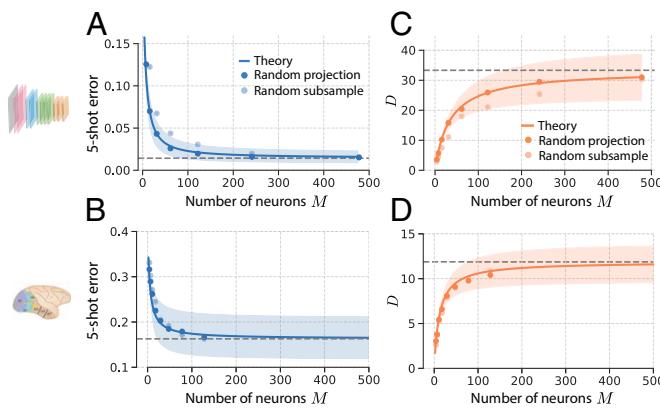


Fig. 6. Effect of number of sampled neurons on concept learning and manifold geometry. Five-shot learning experiments in ResNet50 on $1,000 \times 999$ pairs of concepts from the ImageNet21k dataset (A) and in macaque IT on 64×63 pairs of novel visual concepts (24) (B), given access to only M neurons (light blue points) and given access to only M random linear combinations of the N available neurons (dark blue points). The blue curve represents the prediction from the theory of random projections (Eq. 2); the dashed line is its predicted asymptotic value, SNR_∞ ; and the shaded region represents the SD over all pairs of concepts. In each case, the one-shot learning error remains close to its asymptotic value, provided that the number of recorded neurons M is large compared to the asymptotic manifold dimension D_∞ . (C and D) The estimated manifold dimension $D(M)$ as a function of M randomly sampled neurons (light orange points) and M random linear combinations (dark orange points) of the N neurons in the ResNet50 feature layer (C) and in macaque IT (D). The orange curve represents the prediction from the theory of random projections; the dashed line is its predicted asymptotic value, D_∞ ; and the shaded region represents the SD over all pairs of concepts.

novel concepts, allowing us to construct prototypes for novel visual concepts by simply passing their names into the word-vector embedding model and applying the linear isometry. We use these language-derived prototypes to classify visual stimuli from

pairs of novel visual concepts, achieving a 93.4% test accuracy (Fig. 7C). Intriguingly, this performance is slightly better than the performance of one-shot learning (92.0%), indicating that the name of a concept can provide at least as much information as a single visual example, for the purpose of classifying novel visual concepts (*SI Appendix*, Fig. 9).

Our geometric theory for few-shot learning extends naturally to zero-shot learning, allowing us to derive an analogous zero-shot learning SNR (*SI Appendix*, section 4),

$$\text{SNR}_a^{\text{zero-shot}} = \frac{1}{2} \frac{\|\mathbf{x}_0^a - \mathbf{y}^b\|^2 - \|\mathbf{x}_0^a - \mathbf{y}^a\|^2}{\|\Delta \mathbf{y} \cdot \mathbf{U}_a\|_2}, \quad [3]$$

where $\mathbf{y}^a, \mathbf{y}^b$ are the language-derived prototypes, and $\Delta \mathbf{y} \equiv (\mathbf{y}^a - \mathbf{y}^b)/R_a^2$. This zero-shot learning SNR takes a different form from the few-shot learning SNR since language-derived prototypes are fixed, rather than varying from trial to trial. But note that the two expressions (Eqs. 1 and 3) agree in the limit where $m \rightarrow \infty$, and the language-derived prototypes perfectly match the concept manifold centroids. The signal (numerator) captures how well the language-derived prototypes match the true visual manifold centroids, while the noise (denominator) captures the extent of variation of the visual manifolds along the readout direction $\Delta \mathbf{y}$. Fig. 7D indicates a close match between this theory and zero-shot experiments.

To better understand why the zero-shot method performs so well (even outperforming one-shot learning), we compare the geometry of the language-derived prototypes to the geometry of the vision-derived prototypes used in one-shot learning. We find that although the vision-derived prototypes typically better match the true manifold centroids (Fig. 7E), so that the signal is in Eq. 3 is higher for vision-derived prototypes, language-derived prototypes pick out more informative readout directions, along which the

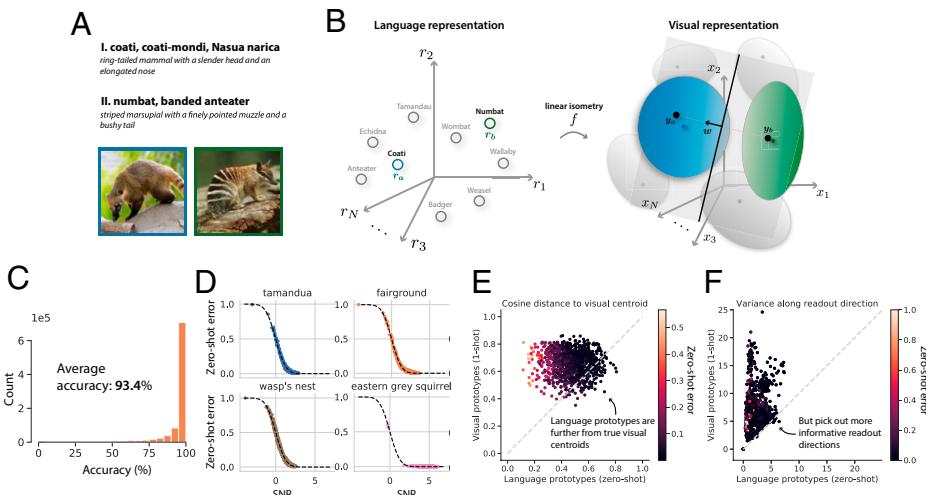


Fig. 7. Visual concept learning without visual examples. (A) Example of a task showing the ability of humans to learn novel visual concepts given only language descriptions (zero-shot learning): Can you identify which image contains the “coati” and which the “numbat”? (B) To simulate this ability, we collect language representations from a word-vector embedding model (Left, gray circles) for the names of the 1,000 visual concepts used to train the DNN (e.g., “anteater” or “badger”), along with corresponding visual concept manifolds from the trained DNN (Right, gray ellipsoids). We learn a linear isometry f to map each language representation as closely as possible to its corresponding visual manifold centroid (Section P). To learn novel visual concepts (e.g., “coati” and “numbat”) without visual examples, we obtain their language representations \mathbf{r}_a from the language model and map them into the visual representation space via the linear isometry, $\mathbf{y}_a = f(\mathbf{r}_a)$. Treating the resulting representations \mathbf{y}_a as visual prototypes, we classify pairs of novel concepts by learning a linear classifier \mathbf{w} (gray hyperplane), which points along the difference between the prototypes, $\mathbf{w} = \mathbf{y}_a - \mathbf{y}_b$, passing through their midpoint. Generalization error (red hashed area) is evaluated by passing test images into the DNN and assessing whether their visual representations are correctly classified. (C) Generalization accuracy is high across all $1,000 \times 999$ pairs of novel visual concepts. (D) Each panel shows the generalization error of 1 visual concept against the 999 others. Each point represents the average generalization error on one such pair of concepts. x axis: SNR (Eq. 3) obtained by estimating neural manifold geometry. y axis: Empirical generalization error measured in zero-shot learning experiments. Theoretical prediction (dashed line) shows a good match with experiments. Error bars, computed over many draws of the training and test examples, are smaller than the symbol size. (E) Visual prototypes from one-shot learning more closely match true concept manifold centroids than do language-derived prototypes, as measured by cosine similarity. (F) However, language prototypes pick out more informative readout directions: Concept manifolds vary less when projected onto readout directions derived from language prototypes than readout directions derived from one-shot visual prototypes.

visual manifolds vary less (Fig. 7*F*), so that the noise in Eq. 3 is smaller for language-derived prototypes. This suppression of noise is greater than the loss in signal, and, hence, language-derived prototypes perform better than one-shot vision-derived prototypes.*

H. Comparing Cognitive Learning Models on Naturalistic Tasks.

Our paradigm of using high-dimensional neural representations of naturalistic concepts as features for few-shot learning presents a unique opportunity to compare different concept-learning models, such as prototype and exemplar learning, in an ethologically relevant setting and explore their joint dependence on concept dimensionality D and the number of examples m . We formalize exemplar learning by a nearest-neighbor decision rule (NN; *SI Appendix, section 6*), which stores in memory the neural representation of all $2m$ training examples and classifies a test image according to the class membership of the nearest training example. We find that exemplar learning outperforms prototype learning when evaluated on low-dimensional concept manifolds with many training examples, consistent with past psychophysics experiments on low-dimensional artificial concepts (15, 47–49) (Fig. 8*A* and Section Q). However, prototype learning outperforms exemplar learning for high-dimensional naturalistic concepts and few training examples. In fact, to outperform prototype learning, exemplar learning requires a number of training examples exponential in the number of dimensions, $m \sim \exp(D/D_0)$ for some constant D_0 (*SI Appendix, section 6*). Hence, for high-dimensional manifolds, like those in trained DNNs and the macaque IT, prototype learning outperforms exemplar learning, given few examples.

A well-known criticism of prototype learning is that averaging the training examples into a single prototype may cause the prototype learner to misclassify some of the training examples themselves (8). However, this phenomenon relies on training examples overlapping significantly along a given direction and almost never happens in high dimensions (Fig. 8*B*), where training examples are approximately orthogonal (Fig. 8*C*).

An intermediate model between prototype and exemplar learning is max-margin learning (50) (Fig. 8*C*). Like prototype learning, max-margin learning involves learning a linear readout; however, rather than pointing between the concept prototypes, its linear readout is chosen to maximize the distance from the decision boundary to the nearest training example of each concept (50) (*SI Appendix, Fig. 12*). Max-margin learning is more sophisticated than prototype learning, in that it incorporates not only the estimated manifold centroid, but also the variation around the centroid. Thus, it is able to achieve zero generalization error for large- m when concept manifolds are linearly separable, overcoming the limitation on the prototype learning SNR due to signal–noise overlaps in the large- m limit (Eq. 1 and Fig. 2*D*). However, like exemplar learning, it requires memory of all training examples. Comparing the three learning models on DNN concept manifolds, we find that prototype and max-margin learning are approximately equivalent for $m \lesssim 8$, and both outperform exemplar learning for m of small to moderate size (Fig. 8*D* and Section Q).

I. Discussion. We have developed a theoretical framework that accounts for the remarkable accuracy of human few-shot learning of novel high-dimensional naturalistic concepts in terms of a simple, biologically plausible neural model of concept learning. Our framework defines the concepts that we can rapidly learn in

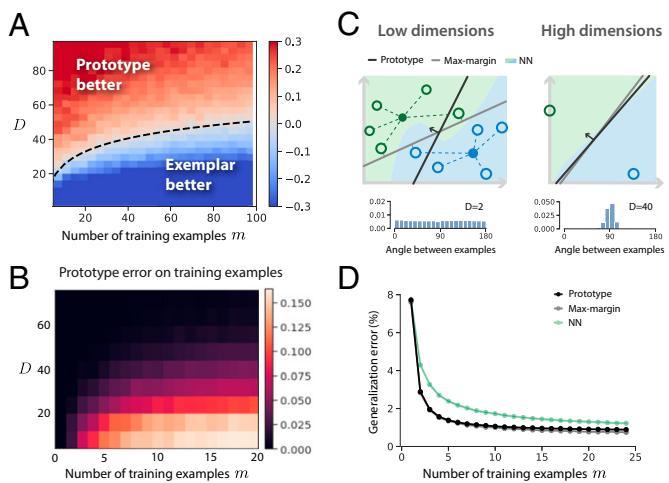


Fig. 8. Comparing cognitive learning models on naturalistic tasks. (A) We compare the performance of prototype and exemplar learning as a joint function of concept manifold dimensionality and the number of training examples, using novel visual concept manifolds from a trained ResNet50. We vary dimensionality by projecting concept manifolds onto their top D principal components. We formalize exemplar learning by NN (*SI Appendix, section 6*). Because the generalization error is very close to zero for m and D large, here, we plot $\text{SNR}^{\text{proto}} - \text{SNR}^{\text{NN}}$. The dashed line demarcating the boundary is given by $m = \exp(D/10)$, reflecting the relationship $\log m \propto D$ predicted by our theory (*SI Appendix, section 6*). The constant 10 is chosen by a one-parameter fit. (B) Prototype learning error as a function of m and D when training examples are reused as test examples. A prototype classifier may misclassify one or more of the training examples themselves when concepts are low-dimensional and the number of training examples is large. However, this rarely happens in high dimensions, as illustrated in C. (C) In low dimensions, multiple training examples may overlap along the same direction (Inset histogram; distribution of angles between examples in DNN concept manifolds for $D = 2$). Hence, averaging these examples (open circles) to yield two prototypes (filled circles) may leave some of the training examples on the wrong side of the prototype classifier's decision boundary (black line). In high dimensions, however, all training examples are approximately orthogonal (C, Inset histogram, $D = 40$), so such mistakes rarely happen. C also shows the decision boundaries of max-margin learning (gray line) and NN learning (green and blue regions), both of which perfectly classify the training examples. In low dimensions, prototype and max-margin learning may learn very different decision boundaries; however, in high dimensions, their decision boundaries are very similar, as quantified in D. (D) Empirical comparison of prototype, max-margin, and NN exemplar learning as a function of the number of training examples m (Section Q). When the number of training examples is small, prototype and SVM learning are approximately equivalent. For larger m , SVM outperforms prototype learning. NN learning performs worse than SVM and prototype learning for both small and intermediate m .

terms of tight manifolds of neural activity patterns in a higher-order brain area. The theory provides readily measurable geometric properties of population responses (Fig. 2) and successfully links them to the few-shot learning performance of a simple prototype learning model. Our model yields remarkably high accuracy on few-shot learning of novel naturalistic concepts using DNN representations for vision (Figs. 1 and 3), which is in excellent quantitative agreement with theoretical prediction. We further show that the four geometric quantities identified by our theory undergo orchestrated transformations along the layers of trained DNNs and along the macaque ventral visual pathway, yielding a consistent improvement in performance along the system's hierarchy (Figs. 4 and 5). We extend our theory to cross-domain learning and demonstrate that comparably powerful visual concept learning is attainable from linguistic descriptors of concepts without visual examples, using a simple map between language and visual domains (Fig. 7). We show analytically and confirm numerically that high few-shot learning performance is possible with as few as 200 IT-like neurons (Fig. 6).

A design tradeoff governing neural dimensionality. A surprising result of our theory is that rapid concept learning is easier when the underlying variability of images belonging to the same object

*But note that given two or more training examples ($m \geq 2$), vision-derived prototypes can average out this noise and achieve better overall performance (see *SI Appendix, section 4B* for details).

is spread across a large number of dimensions (Fig. 2 C and G). Thus, high-dimensional neural representations allow new concepts to be learned from fewer examples, while, as has been shown recently, low-dimensional representations allow for a greater number of familiar concepts to be classified (36, 37). Understanding the theoretical principles by which neural circuits negotiate this tradeoff constitutes an important direction for future research.

Asymmetry in few-shot learning. We found that the pairwise generalization error of simple cognitive models, like prototype, exemplar, and max-margin learning, exhibits a dramatic asymmetry when only one or a few training examples are available (*SI Appendix*, Fig. 1). Our theory attributes this asymmetry to a bias arising from the inability of simple classifiers to estimate the variability of novel concepts from a few examples (Fig. 2 B and F). Investigating whether humans exhibit the same asymmetry and identifying mechanisms by which this bias could be overcome are important future directions. An interesting hypothesis is that humans construct a prior over the variability of novel concepts, based on the variability of previously learned concepts (51).

Connecting language and vision. Remarkably, we found that machine-learning-derived language representations and visual representations, despite being optimized for different tasks across different modalities, can be linked together by an exceedingly simple linear map (Fig. 5), to enable learning novel visual concepts, given only language descriptions. Recent work has revealed that machine-learning-derived neural language representations match those in humans, as measured by both electrocorticography (ECOG) and functional MRI (fMRI) (52). Our results suggest that language and high-level visual representations of concepts in humans may indeed be related through an exceedingly simple map, a prediction that can be tested experimentally. Broadly, our results add plausibility to the tenets of dual-coding theory in human cognition (53).

Comparing brains and machines. Computational neuroscience has recently developed increasingly complex high-dimensional machine-learning-derived models of many brain regions, including the retina (54–57); V1, V4, and IT (28, 58); motor cortex (59); prefrontal cortex (60); and entorhinal cortex (61, 62). Such increased model complexity raises foundational questions about the appropriate comparisons between brains and machine-based models (63). Previous approaches based on behavioral performance (19, 31, 64–66); neuron (54) or circuit (57) matching; linear regression between representations (28); or representational similarity analysis (30) reveal a reasonable match between the two. However, our higher-resolution decomposition of performance into a fundamental set of observable geometric properties reveals significant mismatches (Figs. 4 and 6D). For instance, intermediate representations that have previously been shown to be most predictive of responses in primate V4 are much higher-dimensional than the corresponding representations in V4. These mismatches call for a better understanding of their computational consequences and for more veridical models of the visual pathway.

Finally, our results (*SI Appendix*, Fig. 7) indicate that it is not only the network architecture and overall performance on object-recognition tasks that govern downstream performance and shape neural geometry, but also the number of concepts on which the model is trained. Hence, the size and structure of the dataset on which the model is trained likely play a key role when comparing brains and machines.

Comparing cognitive learning models on naturalistic concepts. Our theory reveals that exemplar learning is superior to prototype learning given many examples of low-dimensional concepts, consistent with past laboratory experiments (15, 47, 48, 67), but is inferior to prototype learning given only a few examples of high-dimensional concepts, like those in DNNs and in primate IT

(Fig. 8), shedding light on a 40-y-old debate (8). These predictions are consistent with a recent demonstration that a prototype-based rule can match the performance of an exemplar model on categorization of familiar high-dimensional stimuli (66). We go beyond prior work by 1) demonstrating that prototype learning achieves superior performance on few-shot learning of novel naturalistic concepts; 2) precisely characterizing the tradeoff as a joint function of concept manifold dimensionality and the number of training examples (Fig. 8); and 3) offering a theoretical explanation of this behavior in terms of the geometry of concept manifolds (*SI Appendix*, section 6).

Proposals for experimental tests of our model. Our theory makes specific predictions that can be tested through behavioral experiments designed to evaluate human performance at learning novel visual concepts from few examples (see *SI Appendix*, Fig. 11 for a proposed experimental design). First, we predict a specific pattern of errors across these novel concepts, shared by neural representations in several trained DNNs (proxies for the human IT cortex), as well as neural representations in the macaque IT (Fig. 5 and *SI Appendix*, Fig. 11C). Second, we predict that humans will exhibit a marked asymmetry in pairwise few-shot learning performance, following the pattern derived in our theory (see *SI Appendix*, Fig. 11 for examples). Third, we predict how performance should scale with the number of training examples m (Fig. 8D). Matches between these predictions and experimental results would indicate that simple classifiers learned atop IT-like representations may be sufficient to account for human concept-learning performance. Deviations from these predictions may suggest that humans leverage more sophisticated higher-order processing to learn new concepts, which may incorporate human biases and priors on object-like concepts (51).

By providing fundamental links between the geometry of concept manifolds in the brain and the performance of few-shot concept learning, our theory lays the foundations for next-generation combined physiology and psychophysics experiments. Simultaneously recording neural activity and measuring behavior would allow us to test our hypothesis that the proposed neural mechanism is sufficient to explain few-shot learning performance and to test whether the four fundamental geometric quantities we identify correctly govern this performance. Furthermore, ECOG or fMRI could be used to investigate whether these four geometric quantities are invariant across primates and humans. Conceivably, our theory could even be used to design visual stimuli to infer the geometry of neural representations in primates and humans, without the need for neural recordings.

In conclusion, this work represents a significant step toward understanding the neural basis of concept learning in humans and proposes theoretically guided psychophysics and physiology experiments to further illuminate the remarkable human capacity to learn new naturalistic concepts from few examples.

Materials and Methods

J. Visual Stimuli. Visual stimuli were selected from the ImageNet dataset, which includes 21,840 unique visual concepts. A subset of 1,000 concepts comprises the standard ImageNet1k training set used in the ImageNet Large Scale Visual Recognition Challenge (68). All DNNs studied throughout this work are trained on these 1,000 concepts alone. To evaluate few-shot learning on novel visual concepts, we gather an evaluation set of 1,000 concepts from the remaining 20,840 concepts not included in the training set, as follows. The ImageNet dataset is organized hierarchically into a semantic tree structure, with each visual concept a node in this tree. We include only the leaves of the tree (e.g., “wildebeest”) in our evaluation set, excluding all superordinate categories (e.g., “mammal”), which have many descendants. We additionally exclude leaves that correspond to abstract concepts, such as “green” and “pet.” Finally, the concepts in the ImageNet dataset vary widely in the number of examples they

contain, with some concepts containing as many as 1,500 examples and others containing only a single example. Among the concepts that meet our criteria above, we select the 1,000 concepts with the greatest number of examples. From each novel concept, we choose a random set of 500 examples. A full list of the 1,000 concepts used in our evaluation set is available at GitHub (<https://github.com/bsorsch/geometry-fewshot-learning>).

K. Estimating the Geometry of Concept Manifolds. Each visual stimulus elicits a pattern of activity $\mathbf{x} \in \mathbb{R}^N$ across the sensory neurons in a high-order sensory layer, such as the IT cortex, or, analogously, the feature layer of a DNN. We collect the population responses to each of all P images in the dataset belonging to a particular visual concept in an $N \times P$ matrix X . To estimate the geometry of the underlying concept manifold, we construct the empirical covariance matrix $C = \frac{1}{P} \mathbf{X} \mathbf{X}^T - \mathbf{x}_0 \mathbf{x}_0^T$, where $\mathbf{x}_0 \in \mathbb{R}^N$ is the manifold centroid. We then diagonalize the covariance matrix,

$$C = \frac{1}{P} \sum_{i=1}^N R_i^2 \mathbf{u}_i \mathbf{u}_i^T, \quad [4]$$

where \mathbf{u}_i are the eigenvectors of C and R_i^2/P the associated eigenvalues. The \mathbf{u}_i each represent unique, potentially interpretable visual features [e.g., animate vs. inanimate, spiky vs. stubby, or short-haired vs. long-haired (22)]. Individual examples of the concept vary from the average along each of these “noise” directions. Some noise directions exhibit more variation than others, as governed by the R_i . Geometrically, the eigenvectors \mathbf{u}_i correspond to the principal axes of a high-dimensional ellipsoid centered at \mathbf{x}_0 , and the R_i correspond to the radii along each axis. A useful measure of the total variation around the centroid is the mean squared radius, $R^2 = \frac{1}{N} \sum_{i=1}^N R_i^2$. R^2 sets a natural length scale, which we use in our theory as a normalization constant to obtain interpretable, dimensionless quantities. Although we do not restrict the maximal number of available axes, which could be as large as the dimensionality of the ambient space, N , the number of directions along which there is significant variation is quantified by an effective dimensionality $D = (\sum_{i=1}^N R_i^2)^2 / \sum_{i=1}^N R_i^4$, called the participation ratio (35), which, in practical situations, is much smaller than N . The participation ratio arises as a key quantity in our theory (*SI Appendix, section 3C*).

L. Prototype Learning. To simulate m -shot learning of novel concepts pairs, we present m randomly selected training examples of each concept to our model of the visual pathway, and collect their neural representations $\mathbf{x}^{a\mu}, \mathbf{x}^{b\mu}$, $\mu = 1, \dots, m$ in a population of IT-like neurons. We perform prototype learning by averaging the representations of each concept into concept prototypes, $\bar{\mathbf{x}}^a = \frac{1}{m} \sum_{\mu=1}^m \mathbf{x}^{a\mu}, \bar{\mathbf{x}}^b = \frac{1}{m} \sum_{\mu=1}^m \mathbf{x}^{b\mu}$. To evaluate the generalization accuracy, we present a randomly selected test example of concept a and determine whether its neural representation is closer in Euclidean distance to the correct prototype $\bar{\mathbf{x}}^a$ than it is to the incorrect prototype $\bar{\mathbf{x}}^b$. This classification rule can be implemented by a single downstream neuron, which adjusts its synaptic weight vector \mathbf{w} to point along the difference between the two concept prototypes, $\mathbf{w} = \bar{\mathbf{x}}^a - \bar{\mathbf{x}}^b$, and adjusts its bias (or firing rate threshold) β to equal the average overlap of \mathbf{w} with each prototype, $\beta = \mathbf{w} \cdot (\bar{\mathbf{x}}^a + \bar{\mathbf{x}}^b)/2$. We derive an analytical theory for the generalization error of prototype learning on concept manifolds in *SI Appendix, section 3C*, and we extend our model and theory to classification tasks involving more than two concepts in *SI Appendix, section 3D*.

M. Prototype Learning Experiments on Synthetic Concept Manifolds.

To evaluate our geometric theory, we perform prototype learning experiments on synthetic concept manifolds constructed with predetermined ellipsoidal geometry (Fig. 2 E–H). By default, we construct manifolds with $R_i^a = R_i^b = 2$, $i = 1, \dots, D$ and $D = 50$, and we sample the centroids $\mathbf{x}_0^a, \mathbf{x}_0^b$ and subspaces $\mathbf{u}_i^a, \mathbf{u}_i^b$ randomly under the constraint that the signal direction $\Delta \mathbf{x}_0$ and the subspace directions are orthonormal, $\|\Delta \mathbf{x}_0\|^2 = \|\mathbf{u}_i^a\|^2 = \|\mathbf{u}_i^b\|^2 = 1$, $\Delta \mathbf{x}_0 \cdot \mathbf{u}_i^a = \Delta \mathbf{x}_0 \cdot \mathbf{u}_i^b = \mathbf{u}_i^a \cdot \mathbf{u}_j^b = 0$, so that the signal-noise overlaps are zero. We then vary each geometric quantity individually, over the ranges reflected in Fig. 2 E–H. To vary the signal, we vary R_i^a, R_i^b from 1 to 2.5. To vary the bias over the interval $(-1, 1)$, we fix $R_i^a = 2$ and vary R_i^b from 0 to $\sqrt{2}R_i^a$. To vary the signal-noise overlaps, we construct ellipsoidal manifolds with one long direction, $R_1^a = R_1^b = 1.5$, and $D - 1$ short directions, $R_i^a = R_i^b = 1$, $i = 2, \dots, D$. We then vary the angle θ between the signal direction $\Delta \mathbf{x}_0$ and the first subspace basis vector \mathbf{u}_1^a by choosing $\Delta \mathbf{x}_0 = \sin(\theta)\mathbf{u}_\perp + \cos(\theta)\mathbf{u}_1^a$, where

$\mathbf{u}_\perp \cdot \mathbf{u}_i^a = 0$, $i = 1, \dots, D$. We vary θ from fully orthogonal ($\theta = \pi/2$) to fully overlapping ($\theta = 0$).

N. DNN Concept Manifolds. All DNNs studied throughout this work are standard architectures available in the PyTorch library (69) and are pretrained on the ImageNet1k dataset. To obtain novel visual concept manifolds, we randomly select $P = 500$ images from each of the $a = 1, \dots, 1,000$ never-before-seen visual concepts in our evaluation set, pass them into the DNN, and obtain their representations in the feature layer (final hidden layer). We collect these representations in an $N \times P$ response matrix X^a . $N = 2,048$ for ResNet architectures. For architectures with $M > 2,048$ neurons in the feature layer, we randomly project the representations down to $N = 2,048$ dimensions using a random matrix $A \in \mathbb{R}^{N \times M}$, $A_{ij} \sim \mathcal{N}(0, 1/\sqrt{N})$. We then compute concept manifold geometry, as described in Section K. To study the layerwise behavior of manifold geometry, we collect the representations at each layer l of the trained DNN into an $N \times P$ matrix X_l^a . For the pixel layer, we unravel raw images into $224 \times 224 \times 3$ -dimensional vectors and randomly project down to N dimensions. Code is available at GitHub (<https://github.com/bsorsch/geometry-fewshot-learning>).

O. Macaque Neural Recordings. Neural recordings of macaque V4 and IT were obtained from the dataset collected in Majaj et al. (24). This dataset contains 168 multiunit recordings in IT and 88 multiunit recordings in V4 in response to 3,200 unique visual stimuli, over ~ 50 presentations of each stimulus. Each visual stimulus is an image of 1 of 64 distinct synthetic three-dimensional objects, randomly rotated, positioned, and scaled atop a random naturalistic background. To obtain the concept manifold for object a , we collect the average response of IT neurons to each of the $P = 50$ unique images of object a in an $N_{IT} \times P$ response matrix X_{IT}^a , where $N_{IT} = 168$, and compute the geometry of the underlying manifold, as described in Section K. We repeat for V4, obtaining a $N_{V4} \times P$ response matrix X_{V4}^a , where $N_{V4} = 88$. We additionally simulate V1 neural responses to the same visual stimuli via a biologically constrained Gabor filter bank, as described in Dapello et al. (70). To compare with trained DNNs, we pass the same set of stimuli into each DNN and obtain a $N_{DNN} \times P$ response matrix X_{DNN}^a , as described in Section N. We then project this response matrix into N_{IT}, N_{V4} , or N_{V1} dimensions to compare with IT, V4, or V1. In order to align V1, V4, and IT to corresponding layers in the trained DNN (as in Fig. 5D), we identify the DNN layers that are most predictive of V1, V4, and IT using partial least-squares regression with 25 components, as in Schrimpf et al. (32).

P. Visual Concept Learning Without Visual Examples by Aligning Visual and Language Domains. To obtain language representations for each of the 1,000 familiar visual concepts from the ImageNet1k training dataset, we collected their embeddings in a standard pretrained word-vector embedding model (GloVe) (44). The word-vector embedding model is pretrained to produce neural representations for each word in the English language based on word cooccurrence statistics. Since concept names in the ImageNet dataset typically involve multiple words (e.g., “tamandua, tamandu, lesser anteater, Tamandua tetradactyla”), we averaged the representations for each word in the class name into a single representation $\mathbf{r}^a \in \mathbb{R}^{N_L}$, where $N_L = 300$ (see *SI Appendix, Fig. 9* for an investigation of this choice). We collected the corresponding visual concept manifolds in a pretrained ResNet50. To align the language and vision domains, we gathered the centroids of the $a = 1, \dots, 1,000$ training manifolds \mathbf{x}_0^a into an $N \times 1,000$ matrix X_0 and gathered the corresponding language representations into an $N_L \times 1,000$ matrix Y . We then learned a scaled linear isometry f from the language domain to the vision domain by solving the generalized Procrustes analysis problem $f = \min_{\alpha, \mathbf{o}, \mathbf{b}} \|f_{\alpha, \mathbf{o}, \mathbf{b}}(Y) - X_0\|^2$, where $f_{\alpha, \mathbf{o}, \mathbf{b}} = \alpha OY - \mathbf{b}$, α is a scalar, $O \in \mathbb{R}^{N_L \times N}$ is an orthogonal matrix, and $\mathbf{b} \in \mathbb{R}^N$ is a translation.

Q. Comparing Cognitive Learning Models on Naturalistic Tasks. In addition to prototype learning, we performed few-shot learning experiments using two other decision rules: a max-margin (SVM) classifier and a nearest-neighbors classifier (Fig. 8). Exemplar learning more generally allows for comparisons to more than just the nearest neighbor and involves the choice of a parameter β , which weights the contribution of each training example to the discrimination function. When $\beta = \infty$, only the nearest neighbor contributes to the discrimination function. When $\beta = 0$, all training examples contribute equally. However, we find that the nearest-neighbors limit $\beta \rightarrow \infty$ is close to optimal in our setting (*SI Appendix, Fig. 12*). Hence, we formalize exemplar learning by a nearest-neighbors decision rule. In order to study the effect of concept manifold

dimensionality D on the performance of each learning rule (Fig. 8 A and B), we vary D by projecting concept manifolds onto their top D principal components.

Data, Materials, and Software Availability. Code and numerical simulations data have been deposited in GitHub (<https://github.com/borsch/geometry-fewshot-learning>) (71).

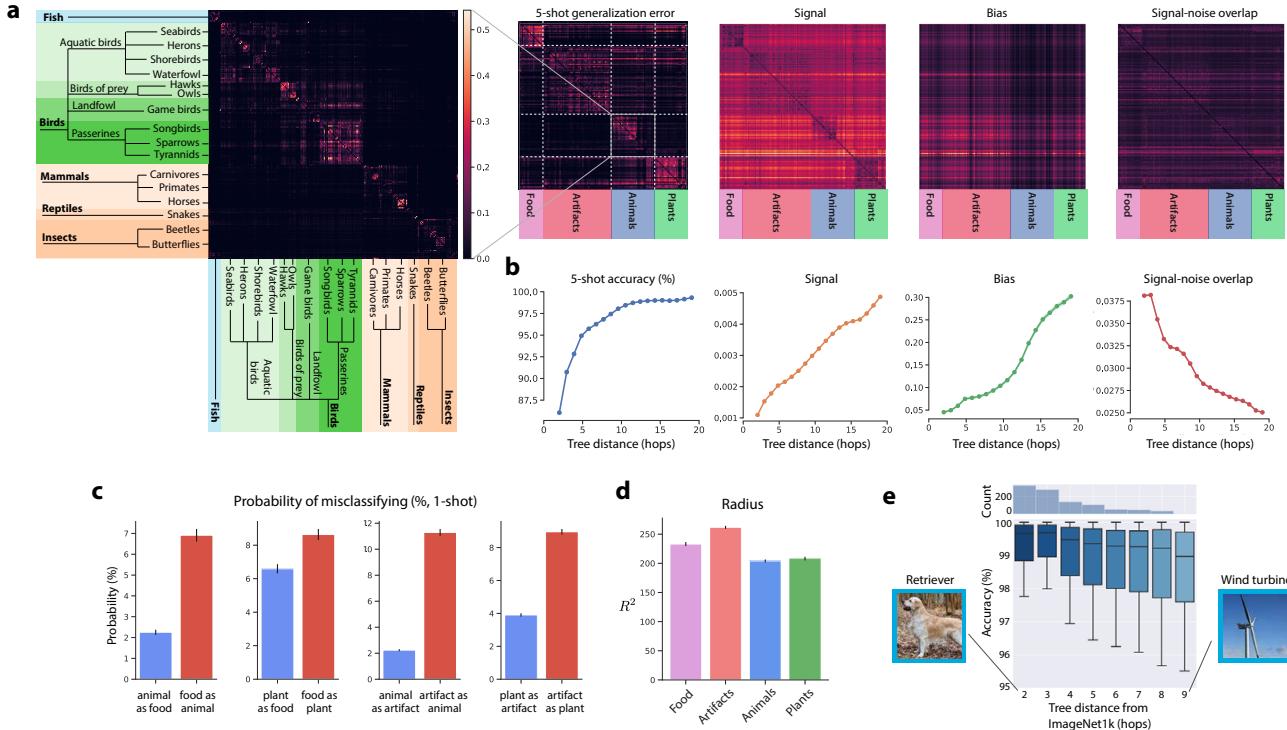
1. L. B. Smith, S. S. Jones, B. Landau, L. Gershoff-Stowe, L. Samuelson, Object name learning provides on-the-job training for attention. *Psychol. Sci.* **13**, 13–19 (2002).
2. P. C. Quinn, P. D. Eimas, S. L. Rosenkrantz, Evidence for representations of perceptually similar natural categories by 3-month-old and 4-month-old infants. *Perception* **22**, 463–475 (1993).
3. G. Behl-Chadha, Basic-level and superordinate-like categorical representations in early infancy. *Cognition* **60**, 105–141 (1996).
4. S. Carey, E. Bartlett, Acquiring a single new word in *Papers and reports on child language development* (1978), vol. 15, pp. 17–29.
5. B. M. Lake, R. Salakhutdinov, J. B. Tenenbaum, Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
6. W. Nie et al., Bongard-Logo: A new benchmark for human-level concept learning and reasoning. *Adv. Neural Inf. Process. Syst.* **33**, 16468–16480 (2020).
7. H. Jiang et al., “Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2022), pp. 19056–19065.
8. G. L. Murphy, *The Big Book of Concepts* (MIT Press, Cambridge, MA, 2004).
9. E. E. Smith, D. L. Medin, Categories and Concepts (Harvard University Press, Cambridge, MA, 1981).
10. D. J. Freedman, M. Riesenhuber, T. Poggio, E. K. Miller, Experience-dependent sharpening of visual shape selectivity in inferior temporal cortex. *Cereb. Cortex* **16**, 1631–1644 (2006).
11. U. R. Karmarkar, Y. Dan, Experience-dependent plasticity in adult visual cortex. *Neuron* **52**, 577–585 (2006).
12. H. P. Op de Beeck, C. I. Baker, J. J. DiCarlo, N. G. Kanwisher, Discrimination training alters object representations in human extrastriate cortex. *J. Neurosci.* **26**, 13025–13036 (2006).
13. Aristotle, *Categories* (Princeton University Press, Princeton, NJ, 1984).
14. E. Rosch, C. B. Mervis, Family resemblances: Studies in the internal structure of categories. *Cognit. Psychol.* **7**, 573–605 (1975).
15. R. M. Nosofsky, Similarity, frequency, and category representations. *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 54–65 (1988).
16. D. L. Medin, E. E. Smith, Strategies and classification learning. *J. Exp. Psychol. Hum. Learn.* **7**, 241–253 (1981).
17. M. N. Hebart, C. Y. Zheng, F. Pereira, C. I. Baker, Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
18. C. A. Sanders, R. M. Nosofsky, Training deep networks to construct a psychological feature space for a natural-object category domain. *Comput. Brain Behav.* **3**, 229–251 (2020).
19. B. M. Lake, W. Zaremba, R. Fergus, T. M. Gureckis, “Deep neural networks predict category typicality ratings for images” in *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*, D. C. Noelle et al., Eds. (Cognitive Science Society, Austin, TX, 2015), pp. 1243–1248.
20. N. K. Logothetis, D. L. Sheinberg, Visual object recognition. *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
21. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).
22. P. Bao, L. She, M. McGill, D. Y. Tsao, A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
23. J. J. DiCarlo, D. Zoccolan, N. C. Rust, How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
24. N. J. Majaj, H. Hong, E. A. Solomon, J. J. DiCarlo, Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
25. G. S. Dhillon, P. Chaudhari, A. Ravichandran, S. Soatto, “A baseline for few-shot image classification” in *International Conference on Learning Representations* (ICLR, 2020).
26. Y. Tian, Y. Wang, D. Krishnan, J. B. Tenenbaum, P. Isola, “Rethinking few-shot image classification: a good embedding is all you need?” in *European Conference on Computer Vision*, A. Vedaldi, H. Bischof, T. Brox, J. M. Frahm, Eds. (Springer, Cham, 2020), pp. 266–282.
27. A. Radford et al., “Learning Transferable Visual Models From Natural Language Supervision” in *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, M. Meila, T. Zhang, Eds. (JMLR 2021), pp. 8784–8763.
28. D. L. Yamins et al., Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 8619–8624 (2014).
29. W. Lotter, G. Kreiman, D. Cox, A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nat. Mach. Intell.* **2**, 210–219 (2020).
30. N. Kriegeskorte et al., Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
31. R. Rajalingham et al., Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *J. Neurosci.* **38**, 7255–7269 (2018).
32. M. Schrimpf et al., Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv* [Preprint] (2018). <https://www.biorxiv.org/content/10.1101/407007v2>.
33. J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **30**, 4078–4088 (2017).
34. T. Chen, S. Kornblith, M. Norouzi, G. Hinton, “A simple framework for contrastive learning of visual representations” in *ICML’20: Proceedings of the International Conference on Machine Learning (PMLR)*, H. Daumé, A. Singh, Eds. (JMLR, 2020), pp. 1597–1607.
35. P. Gao et al., A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* [Preprint] (2017). <https://www.biorxiv.org/content/10.1101/214262v2>. Accessed 15 January 2022.
36. U. Cohen, S. Chung, D. D. Lee, H. Sompolinsky, Separability and geometry of object manifolds in deep neural networks. *Nat. Commun.* **11**, 746 (2020).
37. S. Chung, D. D. Lee, H. Sompolinsky, Classification and geometry of general perceptual manifolds. *Phys. Rev. X* **8**, 031003 (2018).
38. C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, K. D. Harris, High-dimensional geometry of population responses in visual cortex. *Nature* **571**, 361–365 (2019).
39. S. Recanatesi et al., Dimensionality compression and expansion in Deep Neural Networks. *arXiv* [Preprint] (2019). <https://arxiv.org/abs/1906.00443>. Accessed 15 January 2022.
40. A. Ansui, A. Lai, J. H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks. *arXiv* [Preprint] (2019). <https://arxiv.org/abs/1905.12784>. Accessed 15 January 2022.
41. S. Dasgupta, A. Gupta, An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures Algorithms* **22**, 60–65 (2003).
42. S. Ganguli, H. Sompolinsky, Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* **35**, 485–508 (2012).
43. E. M. Trautmann et al., Accurate estimation of neural population dynamics without spike sorting. *Neuron* **103**, 292–308.e4 (2019).
44. J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation. *EMNLP* **14**, 1532–1543 (2014).
45. R. Socher et al., “Zero-shot learning through cross-modal transfer” in *Advances in Neural Information Processing Systems*, F. Pereira, C. J. Burges, L. Bottou, K. Q. Weinberger, Eds. (Curran Associates, Inc., 2013).
46. S. Wu, A. Conneau, H. Li, L. Zettlemoyer, V. Stoyanov, Emerging cross-lingual structure in pretrained language models. *arXiv* [Preprint] (2019). <https://arxiv.org/abs/1911.01464>. Accessed 15 January 2022.
47. S. K. Reed, Pattern recognition and categorization. *Cognit. Psychol.* **3**, 382–407 (1972).
48. T. J. Palmeri, R. M. Nosofsky, Central tendencies, extreme points, and prototype enhancement effects in ill-defined perceptual categorization. *Q. J. Exp. Psychol.* **A** **54**, 197–235 (2001).
49. J. D. Smith, J. P. Minda, Prototypes in the mist: The early epochs of category learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1411–1436 (1998).
50. B. E. Boser, I. M. Guyon, V. N. Vapnik, “A training algorithm for optimal margin classifiers” in *COLT’92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (Association for Computing Machinery, New York, 1992), pp. 144–152.
51. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, How to grow a mind: Statistics, structure, and abstraction. *Science* **331**, 1279–1285 (2011).
52. M. Schrimpf et al., The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv* [Preprint] (2020). <https://www.biorxiv.org/content/10.1101/2020.06.26.174482v2>. Accessed 15 January 2022.
53. A. Paivio, Mental imagery in associative learning and memory. *Psychol. Rev.* **76**, 241–263 (1969).
54. L. McIntosh, N. Maheswaranathan, A. Nayebi, S. Ganguli, S. Baccus, “Deep learning models of the retinal response to natural scenes” in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett, Eds. (Curran Associates, Inc., 2016), vol. 29, pp. 1369–1377.
55. S. A. Ocko, J. Lindsey, S. Ganguli, S. Deny, The emergence of multiple retinal cell types through efficient coding of natural movies. *Adv. Neural Inf. Process. Syst.* **31**, 458737 (2018).
56. J. Lindsey, S. A. Ocko, S. Ganguli, S. Deny, “A Unified Theory of Early Visual Representations from Retina to Cortex through Anatomically Constrained Deep CNNs” in *Proceedings of the International Conference on Learning Representations (ICLR, 2019)*.
57. H. Tanaka et al., From deep learning to mechanistic understanding in neuroscience: The structure of retinal prediction. *Adv. Neural Inf. Process. Syst.* **32**, 8537–8547 (2019).
58. D. L. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
59. D. Sussillo, M. M. Churchland, M. T. Kaufman, K. V. Shenoy, A neural network that finds a naturalistic solution for the production of muscle activity. *Nat. Neurosci.* **18**, 1025–1033 (2015).
60. V. Mante, D. Sussillo, K. V. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
61. B. Sorscher, G. C. Mel, S. Ganguli, S. A. Ocko, A unified theory for the origin of grid cells through the lens of pattern formation. *Adv. Neural Inf. Process. Syst.* **32**, 10003–10013 (2019).
62. B. Sorscher, G. C. Mel, S. A. Ocko, L. Giocomo, S. Ganguli, A unified theory for the computational and mechanistic origins of grid cells. *bioRxiv* [Preprint] (2020). <https://www.biorxiv.org/content/10.1101/2020.12.29.424583v1>. Accessed 15 January 2022.
63. N. Maheswaranathan, A. H. Williams, M. D. Golub, S. Ganguli, D. Sussillo, Universality and individuality in neural dynamics across large populations of recurrent networks. *Adv. Neural Inf. Process. Syst.* **2019**, 15629–15641 (2019).
64. P. Singh, J. C. Peterson, R. M. Battleday, T. L. Griffiths, End-to-end deep prototype and exemplar models for predicting human behavior. *arXiv* [Preprint] (2020). <https://arxiv.org/abs/2007.08723>. Accessed 15 January 2022.
65. C. A. Sanders, R. M. Nosofsky, “Using deep-learning representations of complex natural stimuli as input to psychological models of classification” in *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX, 2018), pp. 1025–1030.
66. R. M. Battleday, J. C. Peterson, T. L. Griffiths, Capturing human categorization of natural images by combining deep networks and cognitive models. *Nat. Commun.* **11**, 5418 (2020).
67. S. C. McKinley, R. M. Nosofsky, Investigations of exemplar and decision bound models in large, ill-defined category structures. *J. Exp. Psychol. Hum. Percept. Perform.* **21**, 128–148 (1995).
68. J. Deng et al., “ImageNet: A large-scale hierarchical image database” in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2009), pp. 248–255.
69. A. Paszke et al., PyTorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8026–8037 (2019).
70. J. Dapello et al., Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. *bioRxiv* [Preprint] (2020). <https://www.biorxiv.org/content/10.1101/2020.06.16.154542v1>. Accessed 15 January 2022.
71. B. Sorscher, Code for: Neural representational geometry underlies few-shot concept learning. GitHub. <https://github.com/borsch/geometry-fewshot-learning>. Deposited 16 June 2021.

ACKNOWLEDGMENTS. We thank Uri Cohen, Gabriel Mel, and Daniel Yamins for helpful discussions. The work is partially supported by the Gatsby Charitable Foundation, the Swartz foundation, and the NIH (Grant 1U19NS104653). B.S. thanks the Stanford Graduate Fellowship for financial support. S.G. thanks the Simons and James S. McDonnell foundations and an NSF CAREER award.

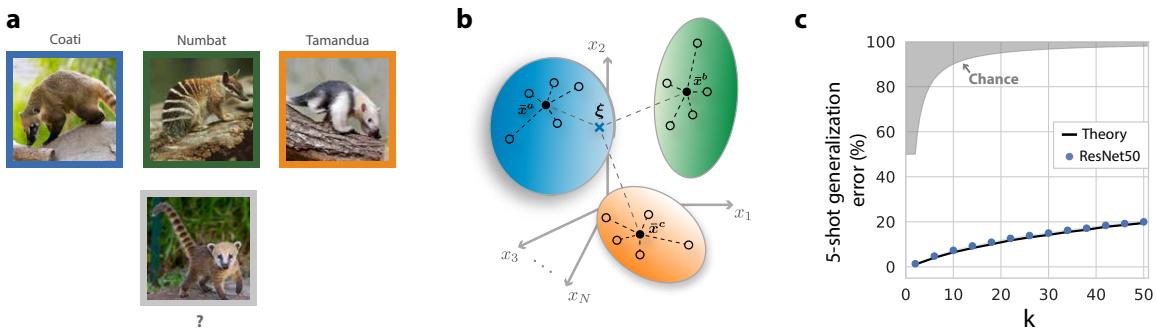
		1
Supplementary materials: Neural representational geometry underlies few-shot concept learning		2
		3
Contents		4
1 Supplementary figures	2	5
2 Introduction	11	6
3 A geometric theory of few-shot learning	11	7
A Prototype learning using neural representations	11	8
B Exact theory for high-dimensional spheres in orthogonal subspaces	11	9
C Full theory: high-dimensional ellipsoids in overlapping subspaces	13	10
D Learning many novel concepts from few examples.	16	11
4 Learning visual concepts without visual examples by aligning language to vision	17	12
A A geometric theory of zero-shot learning	17	13
B How many words is a picture worth? Comparing prototypes derived from language and vision.	18	14
5 How many neurons are required for concept learning?	18	15
A Concept manifold dimensionality under random projections.	18	16
B Few-shot learning requires a number of neurons M greater than the concept manifold dimensionality D	20	17
6 Comparing cognitive learning models in low and high dimensions	20	18
A Identifying the joint role of dimensionality D and number of training examples m	20	19
7 Geometry of DNN concept manifolds encodes a rich semantic structure.	22	20
8 References	22	21

22 1. Supplementary figures

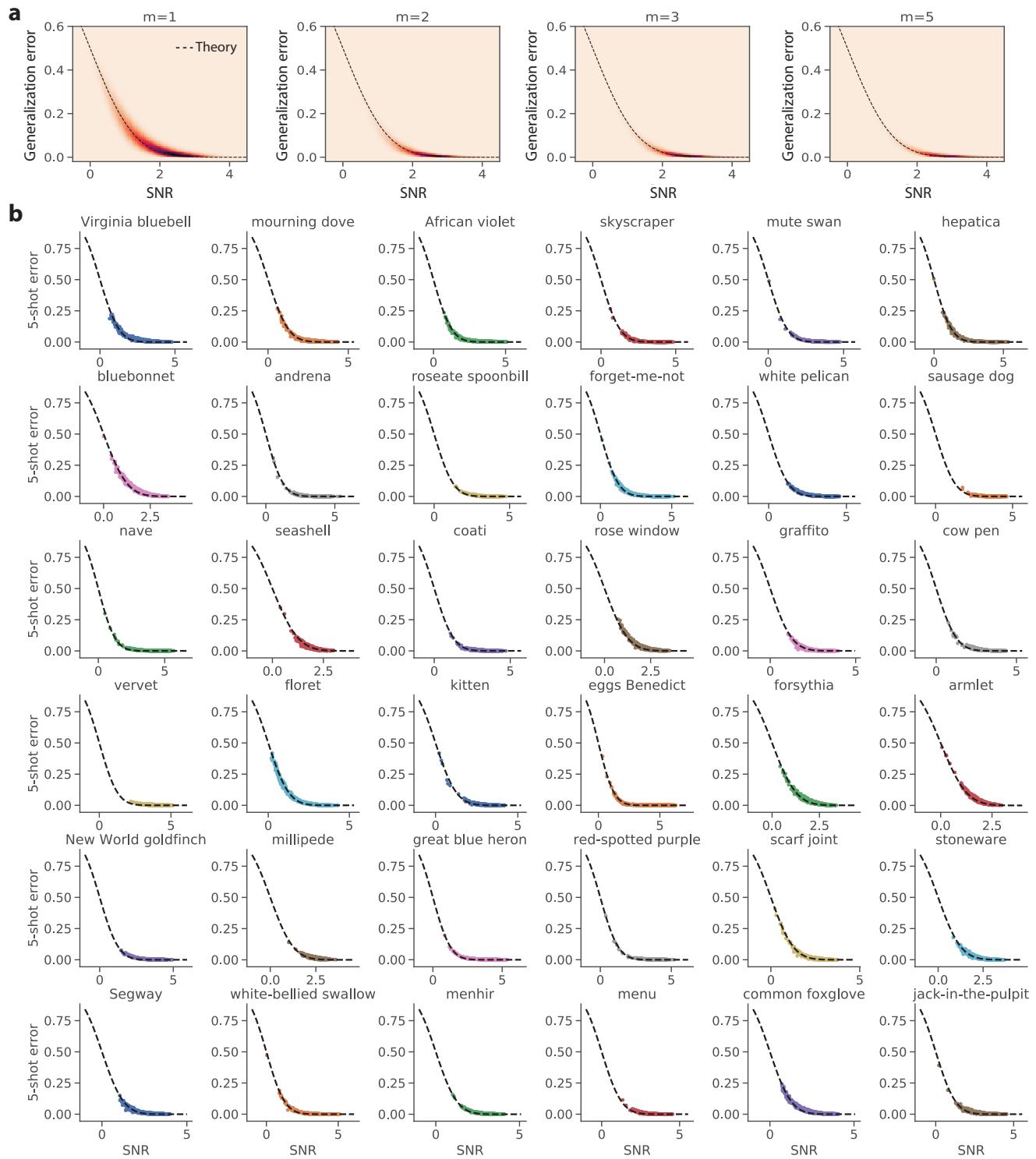
23
24



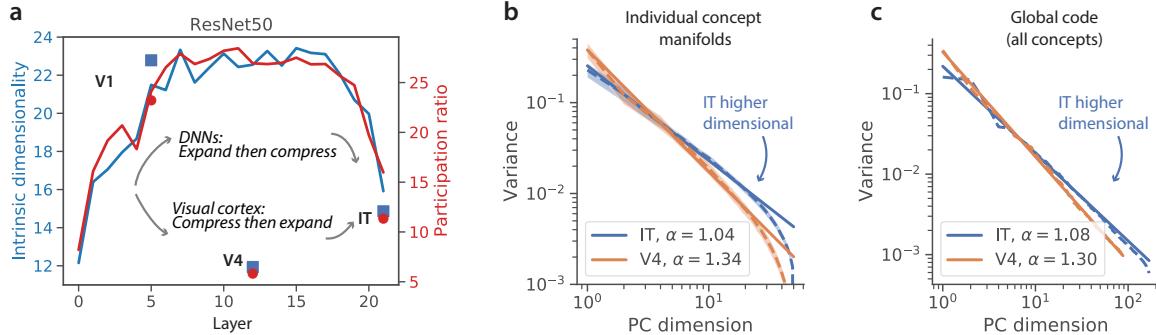
Supplementary Figure 1. Geometry of DNN concept manifolds encodes a rich semantic structure. See SI 7. **a**, We sort the generalization error pattern of prototype learning using concept manifolds from a trained ResNet50 to obey the hierarchical semantic structure of the ImageNet21k dataset. The sorted error matrix exhibits a prominent block diagonal structure, suggesting that most of the errors occur between concepts on the same branch of the semantic tree, and errors between two different branches of the semantic tree are exceedingly unlikely. *Inset:* error pattern across a subset of novel visual concepts, including FISH, BIRDS, MAMMALS, REPTILES, and INSECTS. The full error pattern across all 1,000 novel visual concepts is shown at right. Rows correspond to concepts from which test examples are drawn. This error pattern exhibits a pronounced asymmetry, with much larger errors above the diagonal than below (see panel **c**). We additionally plot the sorted pattern of individual geometric quantities: signal, bias, and signal-noise overlap. Signal exhibits a pronounced block diagonal structure, similar to the error pattern. Bias exhibits a pronounced asymmetry, indicating that plant and animal concept manifolds have significantly smaller radii than artifact and food concept manifolds do (see panel **d**). **b**, We plot the average few-shot accuracy, signal, bias, and signal-noise overlap across all pairs of concepts, as a function of the distance between the two concepts on the semantic tree, defined as the number of hops required to travel from one concept to the other. Few-shot learning accuracy, signal, and bias all increase significantly with semantic distance, while signal-noise overlaps decrease. **c** We quantify the asymmetry of the error pattern in **a**, showing, for instance, that the probability of misclassifying a concept belonging to the category FOOD as an ANIMAL is more than three times as likely as misclassifying an ANIMAL as FOOD. Similar asymmetries are shown for PLANT VS FOOD, ANIMAL VS ARTIFACT, and PLANT VS ARTIFACT. Error bars represent standard error on the mean across all pairs of concepts in the two compared categories. **d, e**, To quantify the effect of distribution shift from the training concepts to the novel concepts, we measure the tree distance from each of the 1k novel concepts to its nearest neighbor among the 1k training concepts. We plot the average few-shot learning accuracy as a function of this distance. Few-shot learning accuracy degrades slightly with distance from the training set, but the effect is not dramatic.



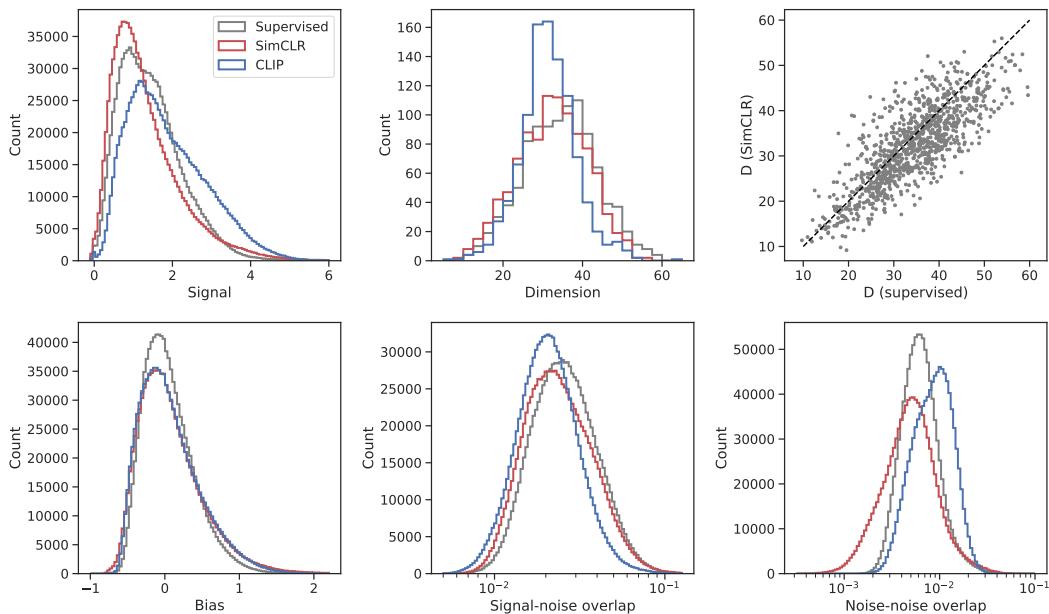
Supplementary Figure 2. Learning many novel concepts from few examples. Concept learning often involves categorizing more than two novel concepts. In SI D we extend our theory to model few-shot learning of k novel concepts. **a**, An example one-shot learning task for $k = 3$: does the test image in the gray box contain a ‘coati’ (blue box), a ‘numbat’ (green box), or a ‘tamandua’ (orange box), given one training example of each? **b**, Illustration of k -concept learning. Training examples of each novel concept (open circles) are averaged into k class prototypes ($\bar{x}^1, \dots, \bar{x}^k$; solid circles). A test example (ξ , blue cross) is classified based on its Euclidean distance to each of the concept prototypes. This classification can be performed by k downstream neurons, one for each novel concept, which adjust their synaptic weights to point along the concept prototypes. **c**, Empirical performance and theoretical predictions. We perform 5-shot learning experiments on visual concept manifolds extracted from a DNN in response to 1,000 novel visual concepts from the ImageNet21k dataset. We compute the generalization error as a function of the number of novel concepts to be learned, k , as well as the prediction from our theory (SI D). Performance is remarkably high, and generalization error stays below 20% even for $k = 50$ (where error at chance is 98%).



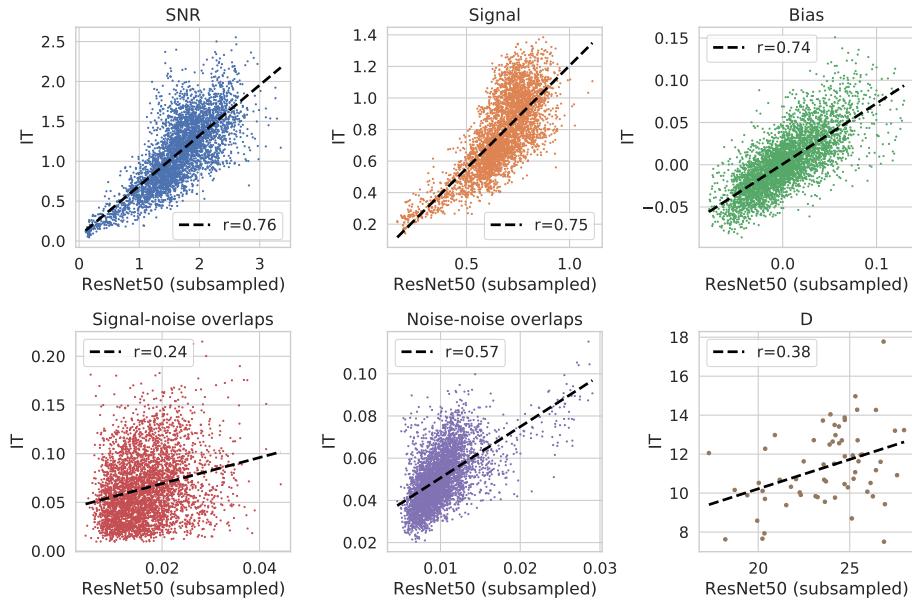
Supplementary Figure 3. geometric theory and few-shot learning experiments on a variety of novel concepts. **a**, We compare the empirical generalization error in 1-, 2-, 3-, and 5-shot learning experiments to the prediction from our geometric theory (Eq. SI.34) on all $1,000 \times 999$ pairs of visual concepts from the ImageNet21k dataset, using concept manifolds derived from a trained ResNet50. We plot a 2d histogram rather than a scatterplot because the number of points is so large. *x-axis*: SNR obtained by estimating neural manifold geometry. *y-axis*: Empirical generalization error measured in few-shot learning experiments. Theoretical prediction (dashed line) shows a good match with experiments. **b**, We provide additional examples of 5-shot prototype learning experiments in a ResNet50 (colored points), along with the prediction from our geometric theory (dashed line), on 36 randomly selected novel visual concepts from the ImageNet21k dataset. Each panel plots the generalization error of one novel visual concept (e.g. ‘Virginia bluebell’) against all 999 other novel visual concepts (e.g. ‘bluebonnet’, ‘African violet’). Each point represents the average generalization error on one such pair of concepts. *x-axis*: SNR (Eq. 1) obtained by estimating neural manifold geometry. *y-axis*: Empirical generalization error measured in few-shot learning experiments. Theoretical prediction (dashed line) shows a good match with experiments. Error bars, computed over many draws of the training and test examples, are smaller than the symbol size..



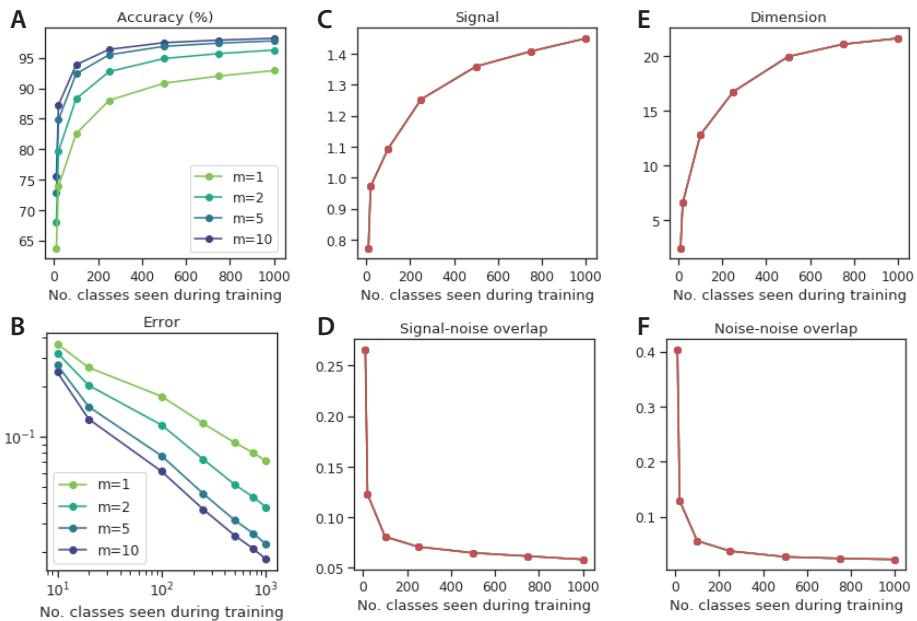
Supplementary Figure 4. Dimensionality diverges between trained DNNs and the primate visual pathway. **a**, To verify that the mismatch in concept manifold dimensionality between DNNs and the primate visual pathway observed in Fig. 5d is not simply due to our choice to measure dimensionality using the participation ratio, we repeat this analysis using a nonlinear estimate of intrinsic dimensionality based on nearest neighbor distances, studied in (1, 2). We find that the intrinsic dimension (blue) evolves similarly to the participation ratio (red) in both DNNs and the ventral visual pathway, corroborating the stark mismatch between trained DNNs and the primate visual pathway. The specific linear transformation used to relate the y-axes is $D_{ID} = 0.53 \times D_{SVD} + 0.87$, where D_{ID} is the intrinsic dimensionality and D_{SVD} is the participation ratio. **b,c**, Eigenspectra in V4 and IT are well described by power laws, both for individual concept manifolds (**b**, shaded area represents standard deviation across manifolds) and for the global population code across all concepts (**c**). The power law is shallower in IT, indicating that representations in IT are higher dimensional.



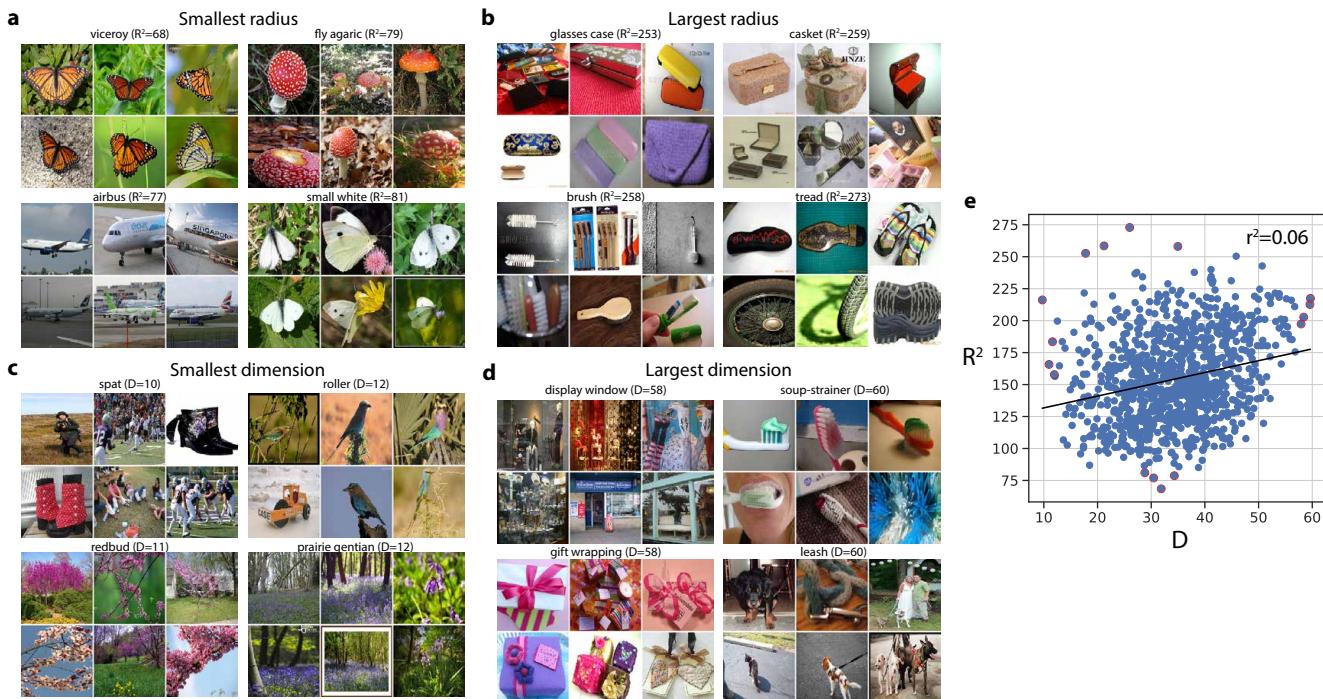
Supplementary Figure 5. Comparing concept manifold geometry across supervised, self-supervised, and unsupervised models. We compare the geometry of 1000 concept manifolds derived from a ResNet50 trained in either a supervised (grey) or self-supervised (SimCLR (3), red) manner, as well as manifolds derived from CLIP (4). SimCLR achieves comparable overall performance to the supervised model, despite its slightly lower average signal, due to its lower signal-noise overlaps. CLIP achieves better overall performance than the other models due to higher average signal, and lower signal-noise overlaps (the contribution from noise-noise overlaps to the few-shot learning error is small). Note also that dimensionality is not only similar between the supervised and self-supervised models (top center), but is also correlated across the 1000 novel concepts (top right).



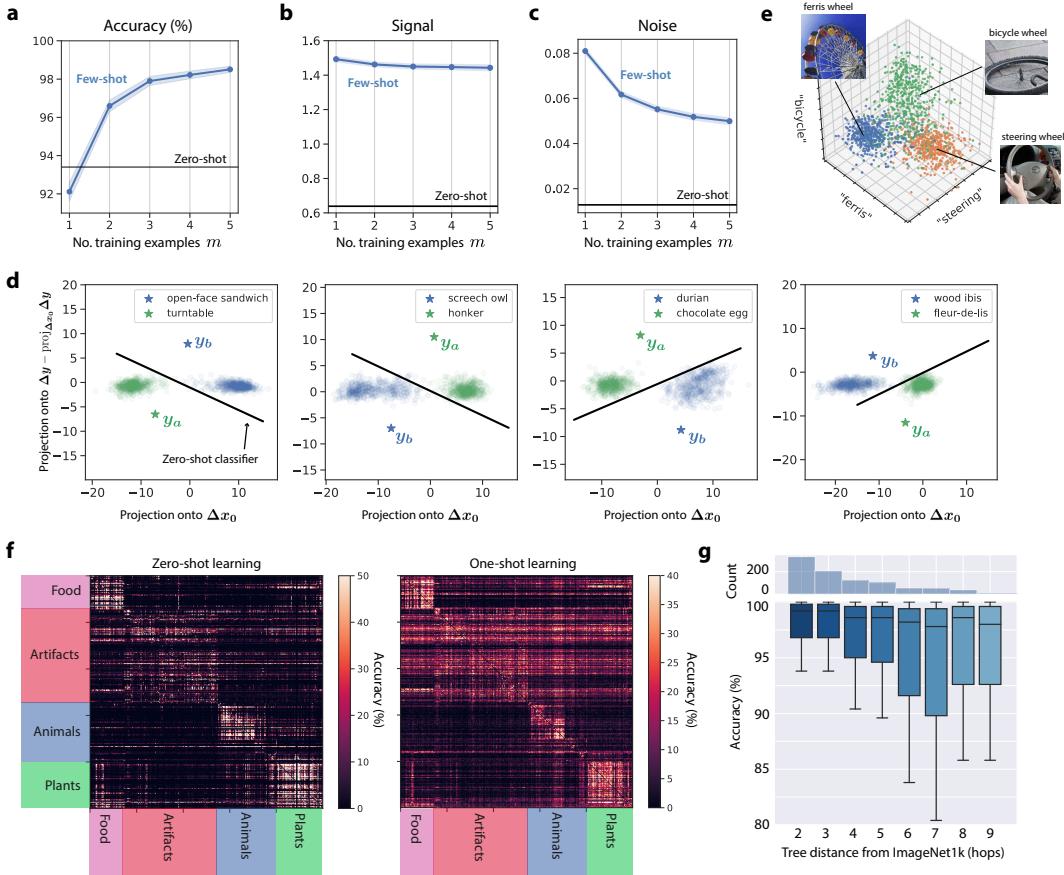
Supplementary Figure 6. Concept manifold geometry is correlated across primate IT cortex and trained DNNs. We estimate the geometry of visual concept manifolds in primate IT cortex and in trained DNNs in response to the same 64 naturalistic visual concepts (5). We then compute the correlation between each quantity in IT cortex and in a trained DNN. Here we use a ResNet50, whose neurons have been randomly subsampled to match the number of recorded neurons in macaque IT (168 neurons). Each panel shows one geometric quantity: SNR ($r=0.76, p < 1 \times 10^{-10}$), signal ($r=0.75, p < 1 \times 10^{-10}$), bias ($r=0.74, p < 1 \times 10^{-10}$), signal-noise overlaps ($r=0.24, p < 1 \times 10^{-10}$), noise-noise overlaps (see SI C; $r=0.57, p < 1 \times 10^{-10}$), and dimension ($r = 0.38, p < 0.005$).



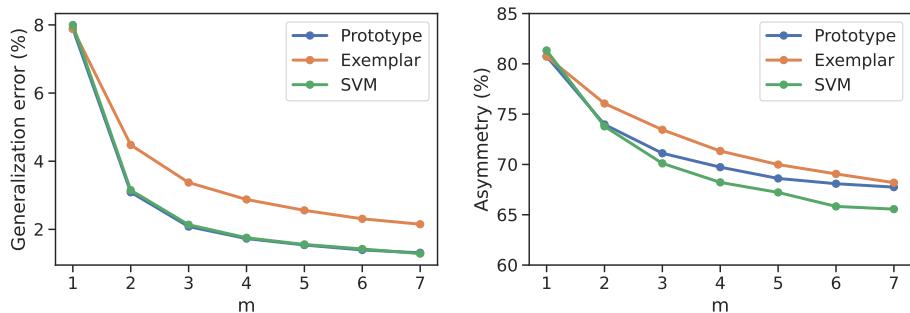
Supplementary Figure 7. Few-shot learning performance improves consistently with the number of concepts seen during training. To investigate the effect of training dataset size on novel concept learning and manifold geometry, we train DNNs (ResNet18) on random subsets of the ImageNet1k dataset with smaller numbers of unique classes. We find that few-shot learning accuracy on novel concepts improves consistently (A), with error decaying roughly like a power law (B) with no indication of saturating before reaching the 1k concepts corresponding to the standard ImageNet1k dataset. Hence we predict that training on even larger subsets of ImageNet21k will yield further improvements in few-shot learning performance. We further observe that the manifold geometry of novel concepts, signal (C), signal-noise overlap (D), dimension (E), and noise-noise overlap (F), evolves smoothly with increasing training dataset size.



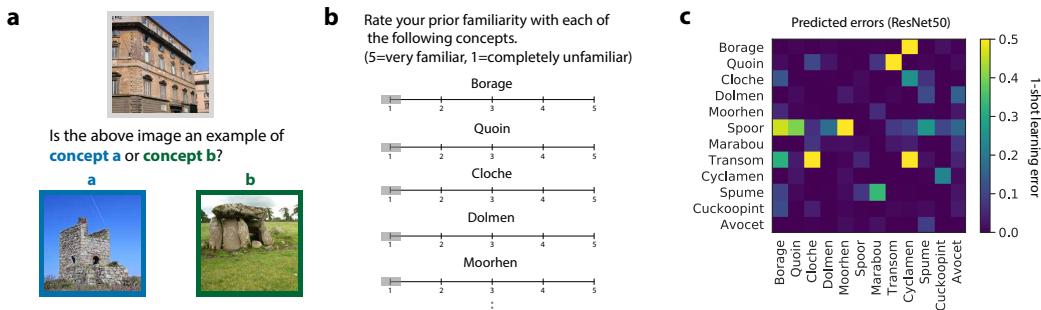
Supplementary Figure 8. Visual examples of concept manifolds with small and large dimension and radius. Among the 1,000 novel visual concepts in our heldout set, we collect examples of the visual concepts whose manifolds in a trained ResNet50 have, **a**, smallest radius, **b**, largest radius, **c**, smallest dimension, and **d**, largest dimension. The salient visual features of concepts with small manifold radius, **a**, appear to exhibit significantly less variation than those of concepts with large manifold radius, **b**. Furthermore, we observe that the visual concepts with smallest manifold radius and dimension are largely comprised of plants and animals **a,c**, while the visual concepts with largest manifold and dimension are largely comprised of human-made objects **b,d**. **e**, A scatterplot of radius and dimension across all 1,000 novel visual concepts reveals very little correlation between R^2 and D ($r^2 = 0.06$, $p < 1 \times 10^{-10}$). The 16 examples in panels **a,b,c,d** are marked with red outlines.



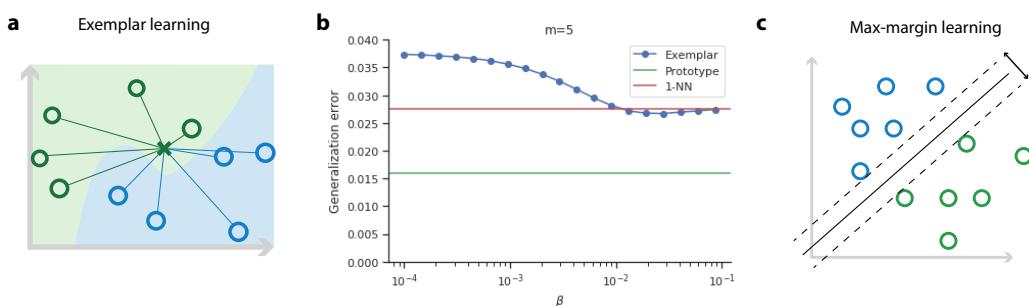
Supplementary Figure 9. How many words is a picture worth? Comparing prototypes derived from language and vision. See SI B. **a**, We compare the performance of prototype learning using prototypes derived from language representations (*zero-shot learning*, Sec. G) to those derived from one or a few visual examples (*few-shot learning*, Sec. A). We find that prototypes derived from language yield a better generalization accuracy than those derived from a single visual example, but not two or more visual examples. **b,c,d**, To better understand this behavior, we use our geometric theory for zero-shot learning, Eq. 3, to decompose the performance of zero- and few-shot learning into a contribution from the ‘signal’, which quantifies how closely the estimated prototypes match the true concept centroids, and a contribution from the ‘noise’, which quantifies the overlap between the readout direction and the noise directions. We find that both signal, **b**, and noise, **c**, are significantly lower for zero-shot learning than for few-shot learning. Hence one-shot learning prototypes more closely match the true concept prototypes on average than language prototypes do. However, language prototypes are able to achieve a higher generalization accuracy by picking out readout directions which overlap significantly less with the concept manifolds’ noise directions. **d**, To visualize this, we project pairs of concept manifolds into the two-dimensional space spanned by the difference between the manifold centroids, Δx_0 , and the language prototype readout direction, Δy . Blue and green stars indicate the language-derived prototypes, and the black boundary indicates the zero-shot learning classifier which points between the two language prototypes. Each panel shows a randomly selected pair of concepts. In each case, the manifolds’ variability is predominantly along the Δx_0 direction, while the language prototypes pick out readout directions Δy with much lower variability. **e**, To obtain a single language representation for visual concepts with multiple word labels (e.g. ‘ferris wheel’, ‘bicycle wheel’, ‘steering wheel’), we chose to simply average the representations of each word. This choice only succeeds if the modifying words (e.g. ‘ferris’, ‘bicycle’, ‘steering’) correspond to meaningful directions when mapped into the visual representation space. We investigate this choice visually by projecting the ‘ferris wheel’, ‘bicycle wheel’, and ‘steering wheel’ visual concept manifolds into the three-dimensional space spanned by the word representations for ‘ferris’, ‘bicycle’, and ‘steering’ mapped into the visual representation space. We find that the three concept manifolds are largely linearly discriminable in this three-dimensional space, indicating that averaging the word representations can be an effective strategy, though likely not the optimal choice. **f** Zero-shot learning (left) exhibits a strikingly similar pattern of errors to one-shot learning (right) across the 1000×1000 novel concepts. **g** Zero-shot learning accuracy degrades slightly with distance from the training set, similar to few-shot learning in Supp. Fig. 1e, but the effect is not dramatic.



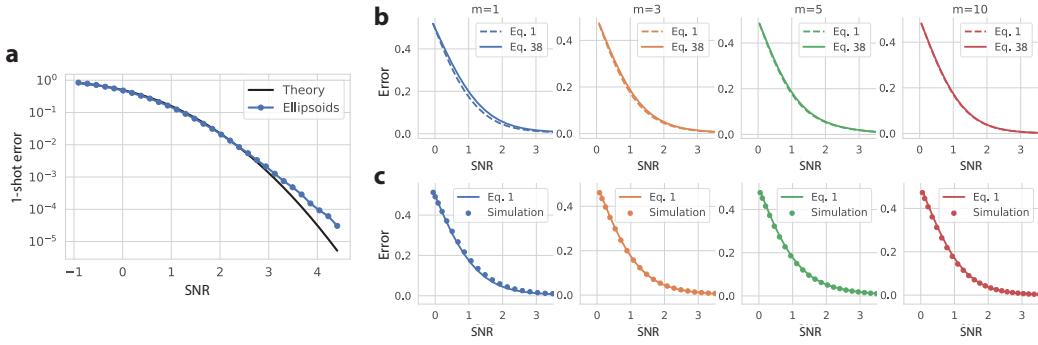
Supplementary Figure 10. Diverse decision rules exhibit asymmetry. We perform few-shot learning simulations using three different decision rules (prototype, exemplar, and SVM) in the same setting studied throughout the main text: feature layer representations from a ResNet50 pretrained on ImageNet1k. *Left*, Few-shot generalization error as a function of the number of the training examples m (reproduces Fig. 8d). *Right*, Asymmetry, defined as $|\varepsilon_a - \varepsilon_b|/(\varepsilon_a + \varepsilon_b)$, is broadly consistent across all three decision rules, decaying with m .



Supplementary Figure 11. Proposed psychophysics experiment to evaluate human few-shot learning on novel naturalistic concepts. **a**, Example one-shot learning task. The participant is asked to correctly identify a novel image (gray box) as an example of either object a (blue box) or object b (green box), given one example of each. **b**, The participant is asked to indicate previous familiarity with each of the visual concepts to be tested. We will use this information to ensure that we are evaluating *novel* concept learning. **c**, We collect the predicted 1-shot learning errors on a proposed set of unfamiliar objects, obtained by performing 1-shot learning experiments on visual concept manifolds in a trained ResNet50. The pattern of errors exhibits a rich structure, and includes a number of visual concept pairs whose errors are dramatically asymmetric.



Supplementary Figure 12. Comparing cognitive learning models. **a**, Under exemplar learning, a test example (green cross) is classified based on its similarity to each of the training examples (green and blue open circles). Hence exemplar learning involves the choice of a parameter β which weights the contribution of each training example to the discrimination function. When $\beta = 0$, all training examples contribute equally. When $\beta = \infty$, only the training example most similar to the test example contributes to the discrimination function. **b**, We perform exemplar learning experiments on concept manifolds in a trained ResNet50, and evaluate the generalization error as a function of β . We find that the optimal choice of β is large, approaching the $\beta \rightarrow \infty$ limit. Furthermore, the optimal generalization error is very close to the $\beta = \infty$ limit, which is equivalent to a nearest neighbors classifier (1-NN), whose generalization error is shown in red. For comparison, the generalization error of a prototype classifier is shown in green. **c**, Illustration of a max-margin classifier. The decision hyperplane (solid black line) of a max-margin classifier is optimized so that its minimum distance to each of the training examples is maximized (6).



Supplementary Figure 13. Numerical evaluation of the approximations used in our theory. **a.** Our theory for the few-shot learning SNR (see SI 3) approximates the projection of concept manifolds onto the linear readout direction as Gaussian-distributed. As discussed in SI B, we expect this approximation to hold well when the SNR is small, and to break down when the SNR is large. To investigate the validity of this approximation, we perform numerical experiments on synthetic ellipsoids constructed to match the geometry of ImageNet21k visual concept manifolds in a trained ResNet50. For each pair of concept manifolds, we vary the signal $\|\Delta\boldsymbol{x}_0\|^2$ over the range 0.01 to 25 and perform 1-shot learning experiments. We compare the generalization error measured in experiments (blue points) to the prediction from our theory (Eq. SI.34; dark line). The theory closely matches experiment over several decades of error, and begins to break down for errors smaller than 10^{-3} . Since errors smaller than 10^{-3} are difficult to resolve experimentally using real visual stimuli—as we have fewer than 1,000 examples of each visual concept, and hence the generalization error may be dominated by one or a few outliers—we judge that this approximation holds well in the regime of interest. The match between theory and experiment for $m > 1$ shot learning (not shown) is as close or closer than for 1-shot learning, due to a law of large numbers-like effect. **b, c.** The few-shot learning SNR in the main text, Eq. 1, differs from the full SNR derived in SI C, Eq. SI.34, which includes several additional terms. In **b** we investigate the difference between the two expressions. The two theoretical curves are nearly indistinguishable for $m \geq 3$, but differ noticeably for $m = 1$. In **c** we compare Eq. 1 to the empirical generalization error measured in few-shot learning experiments on synthetic concept manifolds constructed to match the geometry of ImageNet21k visual concept manifolds in a trained ResNet50. The theory closely matches experiments for $m \geq 3$, but slightly underestimates the generalization error for $m = 1$.

2. Introduction

In this supplementary material we develop our geometric theory for the generalization error of few-shot learning of high-dimensional concepts, we fill in the technical details associated with the main manuscript, and we perform more detailed investigations extending the results we have introduced. The outline of the supplementary material is as follows.

In SI 3 we derive an analytical prediction for the generalization error of prototype learning. We begin with a brief review of prototype learning using neural representations (SI A). We then derive an exact expression for the generalization error of concept learning in a simplified model (SI B), before proceeding to the full theory on pairs of novel concepts (SI C). We then extend our model and theory to capture learning of more than two novel concepts in SI D.

In SI 4 we examine the task of learning novel visual concepts without visual examples (zero-shot learning). We introduce a geometric theory for the generalization error of zero-shot learning in SI A. We then compare the performance of zero-shot learning to few-shot learning, examining the question *how many words is an image worth?*, and identifying intriguing differences between the geometry of language-derived prototypes and vision-derived prototypes that govern the relative performance of the two models (SI B).

In SI 5 we derive analytical predictions, drawing on the theory of random projections, for the number of neurons that must be recorded to reliably measure concept manifold geometry (SI A), as well as the number of IT-like neurons a downstream neuron must listen to in order to achieve high few-shot learning performance (SI B). In SI 6 we compare the performance of two foundational cognitive learning rules: prototype and exemplar learning, and we derive a fundamental relationship between concept dimensionality and the number of training examples that governs the relative performance of the two models.

In SI 7 we investigate the rich semantic structure encoded in the geometry of concept manifolds in trained DNNs. We show that the tree-like semantic organization of visual concepts in the ImageNet dataset is reflected in the geometry of visual concept manifolds, and that few-shot learning accuracy on pairs of novel concepts increases with the distance between the two concepts on the semantic tree, due to changes in each of the four geometric quantities identified in our theory. We additionally quantify the effect of distribution shift between the familiar concepts used to train the DNN, and novel concepts used to evaluate few-shot learning performance.

3. A geometric theory of few-shot learning

A. Prototype learning using neural representations. Our model posits that novel concepts can be learned by learning to discriminate the manifolds of neural activity they elicit in higher order sensory areas, such as IT cortex. We further posit that learning can be accomplished by a population of downstream neurons via a simple plasticity rule. In the following sections we will introduce an analytical theory for the generalization error of concept learning using a particularly simple and biologically plausible plasticity rule: prototype learning. However, we find that this theory also correctly predicts the generalization error of more complex plasticity rules which involve learning a linear readout, such as max-margin learning, when concept manifolds are high-dimensional and the number of training examples is small. Furthermore, when concept manifolds are high-dimensional, their projection onto the linear readout direction is approximately Gaussian, and well characterized by the mean and covariance structure of the concept manifolds. For this reason we approximate concept manifolds as high-dimensional ellipsoids. We find that this approximation predicts the generalization error of few-shot learning remarkably well, despite the obviously complex shape of concept manifolds in the brain and in trained DNNs.

B. Exact theory for high-dimensional spheres in orthogonal subspaces. Before proceeding to the full theory, we begin by studying a toy problem which simplifies the analysis and highlights some of the interesting behavior of few-shot learning in high dimensions. We examine the problem of classifying two novel concepts whose concept manifolds are high-dimensional spheres. Each sphere can be described by its centroid $\mathbf{x}_0^a, \mathbf{x}_0^b$, and its radius R_a, R_b , along a set of orthonormal axes $\mathbf{u}_i^a, \mathbf{u}_i^b, i = 1, \dots, D$, where we assume that each manifold occupies a D -dimensional subspace of the N -dimensional firing rate space. We will further assume that these subspaces are mutually orthogonal, $\mathbf{u}_i^a \cdot \mathbf{u}_j^b = 0$, and orthogonal to the centroids, $(\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a = (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^b = 0$, so that the signal-noise overlaps are zero. Thus a random example from each manifold can be written as,

$$\mathbf{x}^a = \mathbf{x}_0^a + R_a \sum_{i=1}^D \mathbf{u}_i^a s_i^a, \quad \mathbf{x}^b = \mathbf{x}_0^b + R_b \sum_{i=1}^D \mathbf{u}_i^b s_i^b. \quad [\text{SI.1}]$$

where $\mathbf{s}^a, \mathbf{s}^b \sim \text{Unif}(\mathbb{S}^{D-1})$ are random vectors sampled uniformly from the D -dimensional unit sphere. We will study 1-shot learning in this section, using $\mathbf{x}^a, \mathbf{x}^b$ as training examples to learn a decision rule, and proceed to few-shot learning in the next section. Notice that in the 1-shot setting, prototype learning, max-margin learning, and exemplar learning all correspond to the same decision rule, which simply categorizes a test example of concept a , ξ^a , based on whether it is more similar to \mathbf{x}^a or \mathbf{x}^b . Hence the theory we derive in this section is general to prototype learning, max-margin learning, and exemplar learning, as well as a wide range of other learning rules. The test example ξ^a can be written as,

$$\xi^a = \mathbf{x}_0^a + R_a \sum_{i=1}^D \mathbf{u}_i^a \sigma_i^a, \quad [\text{SI.2}]$$

where $\sigma^a \sim \text{Unif}(\mathbb{S}^{D-1})$ is a random vector sampled uniformly from the D -dimensional unit sphere. Using the Euclidean distance metric, ξ^a is classified correctly if $h \equiv -\frac{1}{2}\|\xi^a - \mathbf{x}^a\|^2 + \frac{1}{2}\|\xi^a - \mathbf{x}^b\|^2 > 0$. This decision rule corresponds to a linear classifier, and can be implemented by a downstream neuron which adjusts its synaptic weight vector \mathbf{w} to point along the difference between the training examples, $\mathbf{w} = \mathbf{x}^a - \mathbf{x}^b$, and adjusts its firing threshold (bias) β to equal the average overlap of \mathbf{w} with each training example, $\beta = \mathbf{w} \cdot (\mathbf{x}^a + \mathbf{x}^b)/2$. Then the output of the linear classifier on a test example ξ^a is $\mathbf{w} \cdot \xi^a - \beta = -\frac{1}{2}\|\xi^a - \mathbf{x}^a\|^2 + \frac{1}{2}\|\xi^a - \mathbf{x}^b\|^2 = h$, which can be thought of as the membrane potential of the downstream neuron. The generalization error on concept a , ε_a , is given by the probability that this test example is incorrectly classified, $\varepsilon_a = \mathbb{P}[h \leq 0]$. Evaluating h using our parameterizations for $\mathbf{x}^a, \mathbf{x}^b, \xi^a$ (Eqs. SI.1, SI.2) gives,

$$h = \frac{R_a^2}{2} (\|\Delta \mathbf{x}_0\|^2 + R_b^2 R_a^{-2} - 1) + R_a^2 \mathbf{s}^a \cdot \sigma^a. \quad [\text{SI.3}]$$

Where we have defined $\Delta \mathbf{x}_0 = (\mathbf{x}_0^a - \mathbf{x}_0^b)/R_a$. Thus we can evaluate the generalization error by computing $\varepsilon_a = \mathbb{P}[h \leq 0]$ over all draws of the training and test examples. Defining $\Delta = \frac{1}{2}(\|\Delta \mathbf{x}_0\|^2 + R_b^2 R_a^{-2} - 1)$,

$$\varepsilon_a = \mathbb{P}_{\mathbf{s}^a, \sigma^a}[h \leq 0] = \int_{\mathbb{S}^{D-1}} \frac{d^D \sigma^a}{S_{D-1}} \int_{\mathbb{S}^{D-1}} \frac{d^D s^a}{S_{D-1}} \Theta(-R_a^2 \Delta - R_a^2 s^a \cdot \sigma^a) \quad [\text{SI.4}]$$

where $\Theta(\cdot)$ is the Heaviside step function, and S_{D-1} is the surface area of the D -dimensional unit sphere. Enforcing the spherical constraint via a delta function,

$$= \int_{\mathbb{S}^{D-1}} \frac{d^D \sigma^a}{S_{D-1}} \int_{\mathbb{R}^D} \frac{d^D s^a}{S_{D-1}} \Theta(-R_a^2 \Delta - R_a^2 s^a \cdot \sigma^a) \delta(1 - \|s^a\|^2) \quad [\text{SI.5}]$$

Writing the delta and step functions using their integral representations,

$$= \int_{\mathbb{S}^{D-1}} \frac{d^D \sigma^a}{S_{D-1}} \int_{\mathbb{R}^D} \frac{d^D s^a}{S_{D-1}} \int_{R_a^2 \Delta}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \int \frac{d\hat{\lambda}}{\sqrt{2\pi}} \int \frac{d\alpha}{2\pi} \exp\left(i\hat{\lambda}(\lambda - R_a^2 s^a \cdot \sigma^a)\right) \exp\left(\frac{\alpha}{2} - \frac{\alpha}{2}\|s^a\|^2\right) \quad [\text{SI.6}]$$

We now perform the Gaussian integral over s^a ,

$$= \frac{(2\pi)^{D/2}}{S_{D-1}} \int_{\mathbb{S}^{D-1}} \frac{d^D \sigma^a}{S_{D-1}} \int_{R_a^2 \Delta}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \int \frac{d\hat{\lambda}}{\sqrt{2\pi}} \int \frac{d\alpha}{2\pi} \exp\left(i\hat{\lambda}\lambda - \frac{R_a^4 \|\sigma^a\|^2 \hat{\lambda}^2}{2\alpha} + \frac{\alpha}{2} - \frac{D}{2} \log \alpha\right) \quad [\text{SI.7}]$$

Noting that $\|\sigma^a\|^2$ is constant over the unit sphere, the integral over σ^a drops out,

$$= \frac{(2\pi)^{D/2}}{S_{D-1}} \int_{R_a^2 \Delta}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \int \frac{d\hat{\lambda}}{\sqrt{2\pi}} \int \frac{d\alpha}{2\pi} \exp\left(i\hat{\lambda}\lambda - \frac{R_a^4 \hat{\lambda}^2}{2\alpha} + \frac{\alpha}{2} - \frac{D}{2} \log \alpha\right) \quad [\text{SI.8}]$$

Performing the Gaussian integral over $\hat{\lambda}$,

$$= \frac{(2\pi)^{D/2}}{S_{D-1}} \int_{R_a^2 \Delta}^{\infty} \frac{d\lambda}{\sqrt{2\pi}} \int \frac{d\alpha}{2\pi} \exp\left(-\frac{\lambda^2 \alpha}{2R_a^4} + \frac{\alpha}{2} - \frac{D}{2} \log \alpha\right) \sqrt{\frac{\alpha}{R_a^4}} \quad [\text{SI.9}]$$

Expressing the result in terms of the Gaussian tail function $H(x) = \int_x^{\infty} dt e^{-t^2/2}/\sqrt{2\pi}$,

$$= \frac{(2\pi)^{D/2}}{S_{D-1}} \int \frac{d\alpha}{2\pi} H(\sqrt{\alpha} \Delta) \exp\left(\frac{\alpha}{2} - \frac{D}{2} \log \alpha\right) \quad [\text{SI.10}]$$

We evaluate the integral over α by saddle point. The saddle point condition is,

$$\alpha = D + \frac{\exp(-\alpha \Delta^2/2)}{\sqrt{2\pi}} \frac{\sqrt{\alpha} \Delta}{H(\sqrt{\alpha} \Delta)} \quad [\text{SI.11}]$$

We will begin by studying the case where $\sqrt{\alpha} \Delta \gg 1$, and revisit the case where $\sqrt{\alpha} \Delta = \mathcal{O}(1)$. When $\sqrt{\alpha} \Delta \gg 1$, solving for α gives

$$\alpha = \frac{D}{1 - \Delta^2} \quad [\text{SI.12}]$$

Noting that S_{D-1} is similarly given by $S_{D-1} = \int d\alpha' \exp(\alpha'/2 - D \log(\alpha')/2)(2\pi)^{D/2}$, we obtain the saddle point condition $\alpha' = D$. Using these conditions, we evaluate the integral in Eq. SI.10 at the saddle point, yielding,

$$\varepsilon_a = (1 - \Delta^2)^{D/2} \exp\left(\frac{D}{2} \frac{\Delta^2}{1 - \Delta^2}\right) H\left(\sqrt{\frac{D \Delta^2}{1 - \Delta^2}}\right) \quad [\text{SI.13}]$$

This expression reveals a sharp zero-error threshold at $\Delta = 1$, reflecting a geometric constraint due to the bounded support of each spherical manifold. The generalization error is strictly zero whenever $R_a^2 < \frac{1}{3}(\|\Delta \mathbf{x}_0\|^2 + R_b^2)$. However, when D is

large, the generalization error becomes exponentially small well before this threshold, when $\Delta \ll 1$ and $\sqrt{\alpha}\Delta = \mathcal{O}(1)$. Indeed, the generalization error of prototype learning on concept manifolds in DNNs and macaque IT is better described by the regime where $\sqrt{\alpha}\Delta = \mathcal{O}(1)$. In this regime, the saddle point condition (Eq. SI.11) gives $\alpha = D$, and the generalization error takes the form,

$$\varepsilon_a = H(\sqrt{D}\Delta) = H\left(\frac{1}{2} \frac{\|\Delta \mathbf{x}_0\|^2 + R_b^2 R_a^{-2} - 1}{\sqrt{D-1}}\right) \quad [\text{SI.14}]$$

Hence in this regime the generalization error is governed by a signal-to-noise ratio which highlights some of the key behavior of the full few-shot learning SNR (Eq. 1). First, the SNR increases with the separation between the concept manifolds $\|\Delta \mathbf{x}_0\|^2$. Second, the SNR increases as the manifold dimensionality D increases. As Fig. 2c shows, this is due to the fact that the projection of each manifold onto the linear readout direction \mathbf{w} concentrates around its mean for large D . Remarkably, no matter how close the manifolds are to one another, the generalization error can be made arbitrarily small by making D sufficiently large. Third, the generalization error depends on an asymmetric term arising from the classifier bias, $R_b^2 R_a^{-2} - 1$. Decreasing R_b for fixed R_a increases ε_a , while increasing R_b for fixed R_a decreases ε_a . Interestingly, increasing R_b beyond $R_a\sqrt{1 - \|\Delta \mathbf{x}_0\|^2}$ yields a *negative* SNR, and hence a generalization error worse than chance.

The dependence of Eq. SI.14 on the Gaussian tail function $H(\cdot)$ suggests that the projection of the concept manifold onto the readout direction \mathbf{w} is well approximated by a Gaussian distribution. This approximation holds when the SNR is $\mathcal{O}(1)$, but breaks down when the SNR is large. Motivated by the observation that the few-shot learning SNR for concept manifolds in macaque IT and DNNs is $\mathcal{O}(1)$ (Figs. 4,5), we will use this approximation in the following section to obtain an analytical expression for the generalization error in the more complicated case of ellipsoids in overlapping subspaces, for which no exact closed form solution exists. We investigate the validity of this approximation quantitatively in Supp. Fig. 13a. We perform few-shot learning experiments on synthetic ellipsoids constructed to match the geometry of ResNet50 concept manifolds, and compare the empirical generalization error to the theoretical prediction derived under this approximation. Theory and experiment match closely for errors greater than 10^{-3} . Since errors smaller than 10^{-3} are difficult to resolve experimentally using real visual stimuli –as we have fewer than 1,000 examples of each visual concept, and hence the generalization error may be dominated by one or a few outliers– we judge that this approximation holds well in the regime of interest.

C. Full theory: high-dimensional ellipsoids in overlapping subspaces. We now proceed to the full theory for few-shot learning on pairs of high-dimensional ellipsoids, relaxing the simplifying assumptions in the previous section. We draw $\mu = 1, \dots, m$ training examples each from two concept manifolds, a and b ,

$$\mathbf{x}^{a\mu} = \mathbf{x}_0^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a s_i^{a\mu}, \quad \mathbf{x}^{b\mu} = \mathbf{x}_0^b + \sum_{i=1}^{D_b^{\text{tot}}} R_i^b \mathbf{u}_i^b s_i^{b\mu}, \quad [\text{SI.15}]$$

Where $\mathbf{x}_0^a, \mathbf{x}_0^b$ are the manifold centroids, and R_i^a, R_i^b are the radii along each axis, $\mathbf{u}_i^a, \mathbf{u}_i^b$. $s^{a\mu} \sim \text{Unif}(\mathbb{S}^{D_a^{\text{tot}}-1}), s^{b\mu} \sim \text{Unif}(\mathbb{S}^{D_b^{\text{tot}}-1})$ are random samples from the unit sphere. D_a^{tot} and D_b^{tot} represent the total number of dimensions along which each manifold varies. In practical situations $D_a^{\text{tot}} = D_b^{\text{tot}} = \min\{N, P\}$, where N is the number of recorded neurons and P is the number of examples of each concept. To perform prototype learning, we average these training examples into prototypes, $\bar{\mathbf{x}}^a$ and $\bar{\mathbf{x}}^b$,

$$\bar{\mathbf{x}}^a = \mathbf{x}_0^a + \frac{1}{m} \sum_{i=1}^m \sum_{\mu=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a s_i^{a\mu}, \quad \bar{\mathbf{x}}^b = \mathbf{x}_0^b + \frac{1}{m} \sum_{i=1}^m \sum_{\mu=1}^{D_b^{\text{tot}}} R_i^b \mathbf{u}_i^b s_i^{b\mu}, \quad [\text{SI.16}]$$

To evaluate the generalization error of prototype learning, we draw a test example

$$\xi^a = \mathbf{x}_0^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a \sigma_i^a, \quad [\text{SI.17}]$$

and compute the probability that ξ^a is correctly classified, $\mathbb{P}_{\mathbf{x}^{a\mu}, \mathbf{x}^{b\mu}, \xi^a}[h \leq 0]$, where $h \equiv \frac{1}{2}\|\xi^a - \bar{\mathbf{x}}^b\|^2 - \frac{1}{2}\|\xi^a - \bar{\mathbf{x}}^a\|^2$. Evaluating h using our parameterization gives,

$$h = \frac{1}{2}\|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2 + \frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 s_i^{a\mu} \sigma_i^a + \frac{1}{2m^2} \sum_{i=1}^{D_b^{\text{tot}}} \left(R_i^b \sum_{\mu=1}^m s_i^{b\mu} \right)^2 - \frac{1}{2m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \\ + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a + \frac{1}{m} \sum_{i=1}^{D_b^{\text{tot}}} \sum_{\mu=1}^m R_i^b s_i^{b\mu} (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^b + \frac{1}{m} \sum_{ij} \sum_{\mu=1}^m R_i^a R_j^b \sigma_i^a s_i^{b\mu} \mathbf{u}_i^a \cdot \mathbf{u}_j^b \quad [\text{SI.18}]$$

As we will see, the first term corresponds to the signal, the second to the dimension, the third and fourth terms to the bias, the fifth and sixth to signal-noise overlaps, and the seventh to noise-noise overlaps, which quantify the overlap between manifold

153 subspaces. Each of these terms is independent and, as discussed in the previous section, approximately Gaussian-distributed
 154 when the dimensionality of concept manifolds is high. Hence by computing the mean and variance of each term we can estimate
 155 the full distribution over h . Noting that $\mathbb{P}_{\mathbf{x}^{a\mu}, \mathbf{x}^{b\mu}, \boldsymbol{\xi}^a}[h \leq 0]$ is invariant to an overall scaling of h , we will define the renormalized
 156 $\tilde{h} = h/R_a^2$, which is dimensionless. Computing the generalization error in terms of \tilde{h} , $\varepsilon_a = P_{\mathbf{x}^{a\mu}, \mathbf{x}^{b\mu}, \boldsymbol{\xi}^a}[\tilde{h} \leq 0]$, will allow us to
 157 obtain an expression which depends only on interpretable, dimensionless quantities.

158 **Signal.** The first term in Eq. SI.18, corresponding to signal, is fixed across different draws of the training and test examples,
 159 and so has zero variance. Its mean is given by $\frac{1}{2}\|\Delta\mathbf{x}_0\|^2$, where $\Delta\mathbf{x}_0 = (\mathbf{x}_0^a - \mathbf{x}_0^b)/\sqrt{R_a^2}$, and $R_a^2 \equiv \frac{1}{D_a^{\text{tot}}} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2$.

160 **Dimension.** The second term in Eq. SI.18 corresponds to the manifold dimension. Its mean is zero, since by symmetry odd
 161 powers of s_i^a, σ_i^a integrate to zero over the sphere. Quadratic terms integrate to $1/D_a^{\text{tot}}$, $\int_{\mathbb{S}^{D_a^{\text{tot}}-1}} d^{D_a^{\text{tot}}} \mathbf{s} s_i^2 / S_{D_a^{\text{tot}}-1} = 1/D_a^{\text{tot}}$;
 162 hence the variance is given by,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 s_i^{a\mu} \sigma_i^a \right] = \frac{1}{(R_a^2)^2} \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \left(\prod_{\mu=1}^m \frac{d^{D_a^{\text{tot}}} \mathbf{s}^\mu}{S_{D_a^{\text{tot}}-1}} \right) \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \frac{d^{D_a^{\text{tot}}} \boldsymbol{\sigma}}{S_{D_a^{\text{tot}}-1}} \left(\frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 s_i^{a\mu} \sigma_i^a \right)^2 \quad [\text{SI.19}]$$

$$= \frac{1}{(R_a^2)^2} \frac{1}{m^2} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^4 \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \left(\prod_{\mu=1}^m \frac{d^{D_a^{\text{tot}}} \mathbf{s}^\mu}{S_{D_a^{\text{tot}}-1}} \right) (s_i^{a\mu})^2 \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \frac{d^{D_a^{\text{tot}}} \boldsymbol{\sigma}}{S_{D_a^{\text{tot}}-1}} (\sigma_i^a)^2 \quad [\text{SI.20}]$$

$$= \frac{1}{m} \frac{\sum_i (R_i^a)^4}{(\sum_i (R_i^a)^2)^2} \quad [\text{SI.21}]$$

$$= \frac{1}{m D_a} \quad [\text{SI.22}]$$

163 Where $D_a = (\sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2)^2 / \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^4$ is the participation ratio, which measures the effective dimensionality of the
 164 concept manifold, quantified by the number of dimensions along which it varies significantly (7). Hence this term reflects the
 165 manifold dimensionality, and its variance is suppressed for large D_a .

166 **Bias.** We next proceed to the third and fourth terms of Eq. SI.18, which correspond to bias. We show only the calculation
 167 for the first bias term, as the second bias term follows from the same calculation. The mean is given by,

$$\begin{aligned} \frac{1}{R_a^2} \mathbb{E} \left[\frac{1}{m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \right] &= \frac{1}{R_a^2} \frac{1}{m^2} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \left(\prod_{\mu=1}^m \frac{d^{D_a^{\text{tot}}} \mathbf{s}^\mu}{S_{D_a^{\text{tot}}-1}} \right) (s_i^{a\mu})^2 \\ &= 1/m \end{aligned} \quad [\text{SI.23}]$$

168 And the variance is given by,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \right] = \frac{1}{(R_a^2)^2} \frac{1}{m^4} \sum_{ij}^{D_a^{\text{tot}}} (R_i^a)^2 (R_j^a)^2 \sum_{\mu\nu\gamma\delta}^m \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \left(\prod_{\mu=1}^m \frac{d^{D_a^{\text{tot}}} \mathbf{s}^\mu}{S_{D_a^{\text{tot}}-1}} \right) s_i^{a\mu} s_i^{a\nu} s_j^{a\gamma} s_j^{a\delta} - \frac{1}{m^2} \quad [\text{SI.25}]$$

170 There are three possible pairings of indices which yield even powers of s_i . Due to symmetry, all other pairings integrate to
 171 zero. First, there are m terms of the form $(s_i^\mu)^4$, each of which integrates to $3/(D_a^{\text{tot}}(D_a^{\text{tot}}+2))$. Second, there are $3m(m-1)$
 172 terms of the form $(s_i^\mu)^2(s_j^\nu)^2$, each of which integrates to $1/D_a^{\text{tot}}$. Finally, there are m^2 terms of the form $(s_i^\mu)^2(s_j^\nu)^2$, each of
 173 which integrates to $1/(D_a^{\text{tot}}(D_a^{\text{tot}}+2))$. Thus the integral gives,

$$= \frac{1}{(R_a^2)^2} \frac{1}{m^4} \left(\sum_{i=1}^{D_a^{\text{tot}}} \frac{3m(R_i^a)^4}{D_a^{\text{tot}}(D_a^{\text{tot}}+2)} + \frac{3m(m-1)(R_i^a)^4}{D_a^{\text{tot}}(D_a^{\text{tot}}+2)} + \sum_{i \neq j}^{D_a^{\text{tot}}} \frac{m^2(R_i^a)^2(R_j^a)^2}{D_a^{\text{tot}}(D_a^{\text{tot}}+2)} \right) - \frac{1}{m^2} \quad [\text{SI.26}]$$

$$= \frac{1}{(R_a^2)^2} \frac{m D_a^{\text{tot}} + m(m-1)(D_a^{\text{tot}}+2)}{m^4 D_a^{\text{tot}}(D_a^{\text{tot}}+2)} \left(\sum_{i=1}^{D_a^{\text{tot}}} 3(R_i^a)^4 + \sum_{i \neq j}^{D_a^{\text{tot}}} (R_i^a)^2 (R_j^a)^2 \right) - \frac{1}{m^2} \quad [\text{SI.27}]$$

174 Dropping small terms of $\mathcal{O}(m/D_a^{\text{tot}})$, and writing the final expression in terms of the effective dimensionality D_a ,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \right] = \frac{2}{m^2 D_a} \left(1 - \frac{1}{m} \frac{D_a}{D_a^{\text{tot}}} \right) \quad [\text{SI.28}]$$

176 Notice that when $m = 1$ and the radii are spread equally over all dimensions, so that $D_a = D_a^{\text{tot}}$ (i.e. the manifold is a
 177 sphere), the variance goes to zero. However, in practical situations the effective dimensionality is much smaller than the total
 178 number of dimensions, $D_a \ll D_a^{\text{tot}}$, and the variance is given by $2/m^2 D_a$.

Signal-noise overlaps. We now proceed to the signal-noise overlap terms on the second line of Eq. SI.18, each of which has zero mean. The variance of the first signal-noise overlap term is given by,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a \right] = \frac{1}{(R_a^2)^2} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 ((\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a)^2 \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \frac{d^{D_a^{\text{tot}}} \boldsymbol{\sigma}}{S_{D_a^{\text{tot}}-1}} (\sigma_i^a)^2 \quad [\text{SI.29}]$$

$$= \frac{1}{R_a^2} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 ((\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a)^2 \quad [\text{SI.30}]$$

We refer to this term as signal-noise overlap because it quantifies the overlap between the noise directions \mathbf{u}_i^a and the signal direction $\Delta \mathbf{x}_0$, weighted by the radii R_i^a along each noise direction. To make the notation more compact, we define $\mathbf{U}_a = [R_1^a u_1^a, \dots, R_{D_a^{\text{tot}}}^a u_{D_a^{\text{tot}}}^a]/\sqrt{R_a^2}$, so that the signal-noise overlap takes the form,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a \right] = \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2, \quad [\text{SI.31}]$$

Notice that this signal-noise overlap term does not depend on m , since it involves only the test examples. The second signal-overlap term, in contrast, captures the variation of the training examples along the signal direction, and so its variance does depend on m ,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m R_i^b s_i^{b\mu} (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^b \right] = \frac{1}{m} \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2, \quad [\text{SI.32}]$$

where we have defined $\mathbf{U}_b = [R_1^b u_1^b, \dots, R_{D_b^{\text{tot}}}^b u_{D_b^{\text{tot}}}^b]/\sqrt{R_a^2}$ in analogy to \mathbf{U}_a . As the number of training examples increases, the variation of the b prototype along the signal direction decreases, and the contribution of this signal-noise overlap term decays to zero.

Noise-noise overlaps. Finally, we compute the mean and variance of the final term of Eq. SI.18, the noise-noise overlap term, which follows from a similar calculation. The mean is given by zero, and the variance by,

$$\frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m} \sum_{ij} \sum_{\mu=1}^m R_i^a R_j^b \sigma_i^a s_i^{b\mu} \mathbf{u}_i^a \cdot \mathbf{u}_j^b \right] = \frac{1}{m} \|\mathbf{U}_a^T \mathbf{U}_b\|_F^2. \quad [\text{SI.33}]$$

We refer to this term as the noise-noise overlap because it quantifies the overlap between the noise directions of manifold a , \mathbf{U}_a , and the noise directions of manifold b , \mathbf{U}_b .

SNR. Combining the terms computed above, the mean and variance of \tilde{h} are given by,

$$\begin{aligned} \mu &= \frac{1}{2} \|\Delta \mathbf{x}_0\|^2 + \frac{1}{2} (R_b^2 R_a^{-2} - 1)/m, \\ \sigma^2 &= \frac{D_a^{-1}}{m} + \frac{D_a^{-1}}{2m^2} \left(1 - \frac{1}{m} \frac{D_a}{D_a^{\text{tot}}} \right) + \frac{D_b^{-1}}{2m^2} \frac{(R_b^2)^2}{(R_a^2)^2} \left(1 - \frac{1}{m} \frac{D_b}{D_b^{\text{tot}}} \right) \\ &\quad + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2 + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2/m + \|\mathbf{U}_a^T \mathbf{U}_b\|_F^2/m \end{aligned} \quad [\text{SI.34}]$$

We will refer to the mean as the signal, and the standard deviation as the noise. Hence the generalization error can be expressed in terms of the ratio of the signal to the noise, $\varepsilon_a = \mathbb{P}[\tilde{h} \leq 0] = H(\text{SNR}) \equiv H(\mu/\sigma)$. Suppressing terms in Eq. SI.34 which we argue contribute only a small correction yields the few-shot learning SNR in the main text, Eq. 1. These additional terms, whose contribution we quantify in Supp. Fig. 13b,c, are the two noise terms arising from the bias, $\frac{D_a^{-1}}{2m^2} (1 - \frac{1}{m} \frac{D_a}{D_a^{\text{tot}}})$ and $\frac{D_b^{-1}}{2m^2} \frac{(R_b^2)^2}{(R_a^2)^2} (1 - \frac{1}{m} \frac{D_b}{D_b^{\text{tot}}})$, and the noise-noise overlaps term $\|\mathbf{U}_a^T \mathbf{U}_b\|_F^2/m$. We find that for concept manifolds in macaque IT and in DNN concept manifolds, noise-noise overlaps are substantially smaller than signal-noise overlaps and D_a^{-1} , and their contribution to the overall SNR is negligible. The two noise terms arising from the bias fall off quadratically with m , and we find that their contribution is negligible for $m \geq 3$ (Supp. Fig. 13b,c). Indeed, by performing few-shot learning experiments using synthetic ellipsoids constructed to match the geometry of ImageNet21k visual concept manifolds in a trained ResNet50 (Supp. Fig. 13b), we find that Eq. 1 and Eq. SI.34 are nearly indistinguishable for $m \geq 3$. However, for $m = 1$ the additional terms in Eq. SI.34 yield a small but noticeable correction. Consistent with this, we find that Eq. 1 accurately predicts the empirical generalization error measured in few-shot learning experiments for $m \geq 3$, but very slightly underestimates the generalization error for $m = 1$ (Supp. Fig. 13c). For this reason we include only the dominant terms in the main text (Eq. 1), but we use Eq. SI.34 to predict the generalization error in simulations when $m \leq 3$.

213 **D. Learning many novel concepts from few examples.** Concept learning often involves categorizing more than two novel
 214 concepts (Supp. Fig. 2a). Here we extend our model and theory to the case of learning k new concepts, also known as k -way
 215 classification. Prototype learning extends naturally to k -way classification: we simply define k prototypes, $\bar{\mathbf{x}}^1, \dots, \bar{\mathbf{x}}^k$, by
 216 averaging the training examples of each novel concept (Supp. Fig. 2b). A test example ξ^a of concept a is classified correctly if
 217 it is closest in Euclidean distance to the prototype $\bar{\mathbf{x}}^a$ of concept a . That is, if $h_b > 0$ for all $b \neq a$, where

$$218 \quad h_b = \frac{1}{2} \|\xi^a - \bar{\mathbf{x}}^b\|^2 - \frac{1}{2} \|\xi^a - \bar{\mathbf{x}}^a\|^2. \quad [\text{SI.35}]$$

219 Notice that h_b can be rewritten as $h_b = (\bar{\mathbf{x}}^a - \bar{\mathbf{x}}^b) \cdot \xi^a - (\|\bar{\mathbf{x}}^a\|^2 - \|\bar{\mathbf{x}}^b\|^2)/2$. Hence this classification rule is linear, and can
 220 be implemented by k downstream neurons, one for each novel concept. Each downstream neuron adjusts its synaptic weight
 221 vector \mathbf{w}^b to point along the direction of a concept prototype, $\mathbf{w}^b = \bar{\mathbf{x}}^b$, $b = 1, \dots, k$, and adjusts its firing threshold (bias) β to
 222 equal the overlap of \mathbf{w}^b with the prototype, $\beta^b = \mathbf{w}^b \cdot \bar{\mathbf{x}}^b/2$. Then the test example ξ^a of concept a is classified correctly if the
 223 output of neuron a , $\mathbf{w}^a \cdot \xi^a - \beta^a$, is greater than the output of neuron b , $\mathbf{w}^b \cdot \xi^a - \beta^b$, for all $b \neq a$.

224 The generalization error on concept a , ε_a , is given by the probability that at least one $h_b \geq 0$, for all $b \neq a$. Equivalently,

$$225 \quad \varepsilon_a = 1 - \mathbb{P}\left[\prod_{b \neq a} (h_b > 0)\right] \quad [\text{SI.36}]$$

226 To evaluate this probability, we consider the joint distribution of the h_b for $b \neq a$, defining the random variable $\mathbf{h} \equiv$
 227 $[h_1, \dots, h_{a-1}, h_{a+1}, \dots, h_k]$. We have already computed h_b (Eq. SI.18) and seen that it is a Gaussian distributed random
 228 variable when the SNR = $\mathcal{O}(1)$ and the concept manifold is high-dimensional. Hence in this regime \mathbf{h} is distributed as a
 229 multivariate Gaussian random variable,

$$230 \quad p(\mathbf{h}) = \frac{\exp[-\frac{1}{2}(\mathbf{h} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{h} - \boldsymbol{\mu})]}{\sqrt{(2\pi)^{k-1} \det \Sigma}}, \quad [\text{SI.37}]$$

231 with mean $\boldsymbol{\mu}_b \equiv \mathbb{E}[h_b]$, and covariance $\Sigma_{bc} = \mathbb{E}[h_b h_c] - \boldsymbol{\mu}_b \boldsymbol{\mu}_c$. We can therefore obtain the generalization error by integrating
 232 $p(\mathbf{h})$ over the positive orthant, where all $h_b \geq 0$,

$$233 \quad \varepsilon_a = 1 - \int_{\mathbb{R}_+^{k-1}} d^{k-1} \mathbf{h} p(\mathbf{h}) \quad [\text{SI.38}]$$

234 All that is left to do is compute the mean $\boldsymbol{\mu}$ and covariance Σ . As before, $\mathbb{P}[\prod_{b \neq a} (h_b > 0)]$ is invariant to an overall
 235 scaling of h_b , so we will work with the renormalized $\tilde{\mathbf{h}} = \mathbf{h}/R_a^2$ in order to obtain dimensionless quantities. We have already
 236 evaluated the mean $\boldsymbol{\mu}_b = \mathbb{E}[\tilde{h}_b]$ and the diagonal covariance elements $\Sigma_{bb} = \text{Var}[\tilde{h}_b]$ in SI C; these are just the signal and noise,
 237 respectively, from the two-way SNR, Eq. SI.34. So we proceed to the off-diagonal covariances, $\Sigma_{bc} = \mathbb{E}[\tilde{h}_b \tilde{h}_c] - \boldsymbol{\mu}_b \boldsymbol{\mu}_c$. Using
 238 the expression for h_b in Eq. SI.18, we find that when $b \neq c$ three terms contribute,

$$239 \quad \begin{aligned} \Sigma_{bc} &= \frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{2m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \right] + \frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 s_i^{a\mu} \sigma_i^a \right] \\ &\quad + \frac{1}{(R_a^2)^2} \text{Var} \left[\left(\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a \right) \left(\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^c) \cdot \mathbf{u}_i^a \right) \right] \end{aligned} \quad [\text{SI.39}]$$

240 The first term we evaluate in Eq. SI.28,

$$240 \quad \frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{2m^2} \sum_{i=1}^{D_a^{\text{tot}}} \left(R_i^a \sum_{\mu=1}^m s_i^{a\mu} \right)^2 \right] = \frac{1}{2m^2 D_a} \left(1 - \frac{1}{m} \frac{D_a^{\text{tot}}}{D_a} \right) \quad [\text{SI.40}]$$

241 The second term we evaluate in Eq. SI.22,

$$242 \quad \frac{1}{(R_a^2)^2} \text{Var} \left[\frac{1}{m} \sum_{i=1}^{D_a^{\text{tot}}} \sum_{\mu=1}^m (R_i^a)^2 s_i^{a\mu} \sigma_i^a \right] = \frac{1}{m D_a} \quad [\text{SI.41}]$$

243 And for the third term we evaluate an analogous expression in Eq. SI.31, yielding,

$$244 \quad \frac{1}{(R_a^2)^2} \text{Var} \left[\left(\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^b) \cdot \mathbf{u}_i^a \right) \left(\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{x}_0^a - \mathbf{x}_0^c) \cdot \mathbf{u}_i^a \right) \right] = (\Delta \mathbf{x}_0^{ab} \cdot \mathbf{U}_a)^T (\Delta \mathbf{x}_0^{ac} \cdot \mathbf{U}_a) \quad [\text{SI.42}]$$

where $\Delta \mathbf{x}_0^{ab} = (\mathbf{x}_0^a - \mathbf{x}_0^b)/\sqrt{R_a^2}$, and $\Delta \mathbf{x}_0^{ac} = (\mathbf{x}_0^a - \mathbf{x}_0^c)/\sqrt{R_a^2}$. Combining these terms, and re-inserting the terms for $b = c$ derived in Eq. SI.34, we obtain the full expression for the covariance,

$$\Sigma_{bc} = \frac{D_a^{-1}}{m} + \frac{D_a^{-1}}{2m^2} \left(1 - \frac{1}{m} \frac{D_a^{\text{tot}}}{D_a} \right) + (\Delta \mathbf{x}_0^{ab} \cdot \mathbf{U}_a)^T (\Delta \mathbf{x}_0^{ac} \cdot \mathbf{U}_a) \\ + \delta_{bc} \left(\frac{D_b^{-1}}{2m^2} \frac{(R_b^2)^2}{(R_a^2)^2} \left(1 - \frac{D_b^{\text{tot}}}{D_b} \right) + \frac{1}{m} \|\Delta \mathbf{x}_0^{ab} \cdot \mathbf{U}_b\|^2 + \frac{1}{m} \|\mathbf{U}_a^T \mathbf{U}_b\|_F^2 \right). \quad [\text{SI.43}]$$

Recall from Eq. SI.34 that μ is given by,

$$\mu_b = \frac{1}{2} \|\Delta \mathbf{x}_0\|^2 + \frac{1}{2} (R_b^2 R_a^{-2} - 1)/m \quad [\text{SI.44}]$$

Integrating the multivariate Gaussian with mean μ and covariance Σ over the positive orthant, Eq. SI.38, gives the generalization error (Supp. Fig. 2).

4. Learning visual concepts without visual examples by aligning language to vision

A. A geometric theory of zero-shot learning. Prototype learning also extends naturally to the task of learning novel visual concepts without visual examples (*zero-shot* learning), as we demonstrate in Section G by generating visual prototypes from language-derived representations. Moreover, our theory extends straightforwardly to predict the performance of zero-shot learning in terms of the geometry of concept manifolds. Consider the task of learning to classify two novel visual concepts, given concept prototypes $\mathbf{y}^a, \mathbf{y}^b$ derived from language, or from another sensory modality. To classify a test example of concept a , we present the test example to the visual pathway and collect the pattern of activity ξ^a it elicits in a population of IT-like neurons. We then classify ξ^a according to which prototype it is closer to. As in few-shot learning, we assume that ξ^a lies along an underlying ellipsoidal manifold,

$$\xi^a = \mathbf{x}_0^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a \sigma_i^a, \quad [\text{SI.45}]$$

where $\sigma \sim \text{Unif}(\mathbb{S}^{D_a^{\text{tot}}-1})$. We define $h \equiv \frac{1}{2} \|\xi^a - \mathbf{y}^b\|^2 - \frac{1}{2} \|\xi^a - \mathbf{y}^a\|^2$, so that the generalization error is given by the probability that $h \leq 0$, $\varepsilon_a = \mathbb{P}_{\xi^a}[h \leq 0]$. Writing out h ,

$$h = \frac{1}{2} \|\mathbf{x}_0^a - \mathbf{y}^b\|^2 - \frac{1}{2} \|\mathbf{x}_0^a - \mathbf{y}^a\|^2 - \sum_{i=1}^{D_a^{\text{tot}}} (\mathbf{y}^a - \mathbf{y}^b) \cdot \mathbf{u}_i^a R_i^a \sigma_i^a \quad [\text{SI.46}]$$

Hence the error depends only on the distances between the prototypes and the true manifold centroids, and the overlap between the manifold subspace and the difference between the two prototypes. When the concept manifold is high dimensional, the last term is approximately Gaussian-distributed, with zero mean and variance,

$$\text{Var} \left[\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \sigma_i^a (\mathbf{y}^a - \mathbf{y}^b) \cdot \mathbf{u}_i^a \right] = \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 ((\mathbf{y}^a - \mathbf{y}^b) \cdot \mathbf{u}_i^a)^2 \int_{\mathbb{S}^{D_a^{\text{tot}}-1}} \frac{d^{D_a^{\text{tot}}} \sigma}{S_{D_a^{\text{tot}}-1}} (\sigma_i^a)^2 \quad [\text{SI.47}]$$

$$= R_a^2 \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 ((\mathbf{y}^a - \mathbf{y}^b) \cdot \mathbf{u}_i^a)^2 \quad [\text{SI.48}]$$

Defining $\Delta \mathbf{y} = (\mathbf{y}^a - \mathbf{y}^b)/\sqrt{R_a^2}$ the variance can be written more compactly as $(R_a^2)^2 \|\Delta \mathbf{y} \cdot \mathbf{U}_a\|^2$. Hence the generalization error of zero-shot learning is governed by a signal-to-noise ratio, $\varepsilon_a^{\text{zero-shot}} = H(\text{SNR}_a^{\text{zero-shot}})$, where,

$$\text{SNR}_a^{\text{zero-shot}} = \frac{1}{2} \frac{\|\mathbf{x}_0^a - \mathbf{y}^b\|^2 - \|\mathbf{x}_0^a - \mathbf{y}^a\|^2}{\|\Delta \mathbf{y} \cdot \mathbf{U}_a\|_2} \quad [\text{SI.49}]$$

Where we have normalized all quantities by R_a^2 . This theory yields a close match to zero-shot learning experiments performed on concept manifolds in a trained ResNet50 (Fig. 7d), and affords deeper insight into the performance of zero-shot learning, as we show in Fig. 7e, and explore further in the following section.

272 **B. How many words is a picture worth? Comparing prototypes derived from language and vision..** We found that prototypes
 273 derived from language yield a better generalization accuracy than those derived from a single visual example (Section G),
 274 but not two or more visual examples (Supp. Fig. 9a). To better understand this behavior, we use our geometric theory for
 275 zero-shot learning, Eq. 3, to decompose the zero-shot learning SNR into a contribution from the ‘signal’, which quantifies how
 276 closely the estimated prototypes match the true manifold centroids, and a contribution from the ‘noise’, which quantifies the
 277 overlap between the readout direction and the noise directions. We use the same theory to examine the prototypes generated
 278 by few-shot learning, even though these prototypes vary across different draws of the training examples, by averaging the
 279 signal and noise over many different draws of the training examples. This allows us to compare zero-shot learning and few-shot
 280 learning in the same framework, to understand whether the enhanced performance of zero-shot learning is due to higher signal
 281 (i.e. a closer match between estimated prototypes and true centroids) or lower noise (i.e. less overlap between the readout and
 282 noise directions). In Supp. Fig. 9b,c we show that both signal and noise are significantly lower for zero-shot learning than for
 283 few-shot learning. Therefore, one-shot learning prototypes more closely match the true concept prototypes on average than
 284 language prototypes do. However, language prototypes are able to achieve a higher overall generalization accuracy by picking
 285 out linear readout directions which overlap significantly less with the concept manifolds’ noise directions. We visualize these
 286 directions in Supp. Fig. 9d by projecting pairs of concept manifolds into the two-dimensional space spanned by the signal
 287 direction $\Delta\mathbf{x}_0$ and the language prototype readout direction $\Delta\mathbf{y}$. In each case, the manifolds’ variability is predominantly
 288 along the signal direction $\Delta\mathbf{x}_0$, while the language prototypes pick out readout directions $\Delta\mathbf{y}$ with much lower variability.

289 5. How many neurons are required for concept learning?

290 Neurons downstream of IT cortex receive inputs from only a small fraction of the total number of available neurons in IT. How
 291 does concept learning performance depend on the number of input neurons? Similarly, a neuroscientist seeking to estimate
 292 concept manifold geometry in IT only has access to a few hundred neurons. How is concept manifold geometry distorted when
 293 only a small fraction of neurons is recorded from?

294 In this section we will draw on the theory of random projections to derive analytical answers to both questions. We will
 295 model recording from a small number M of the N available neurons as projecting the N -dimensional activity patterns into an
 296 M -dimensional subspace. When activity patterns are randomly oriented with respect to single neuron axes, selecting a random
 297 subset of neurons to record from is exactly equivalent to randomly projecting the full N -dimensional activity patterns into an
 298 M -dimensional subspace. We will begin by deriving the behavior of concept manifold dimensionality D as a function of the
 299 dimension of the target space M , and use this to derive the behavior of the few-shot learning generalization error.

300 **A. Concept manifold dimensionality under random projections..** Consider randomly projecting each point $\mathbf{x} \in \mathbb{R}^N$ on a concept
 301 manifold to a lower-dimensional subspace, $A\mathbf{x} = \mathbf{x}' \in \mathbb{R}^M$ using a random projection matrix $A \in \mathbb{R}^{M \times N}$, $A_{ij} \sim \mathcal{N}(0, 1/M)$.
 302 We collect all points on the original concept manifold into an $N \times P$ matrix X , and collect all points on the projected concept
 303 manifold into an $M \times P$ matrix $X' = AX$. Recall that the effective dimensionality $D(N)$ of the original concept manifold can
 304 be expressed in terms of its $N \times N$ covariance matrix $C_N = \frac{1}{P}XX^T - \mathbf{x}_0\mathbf{x}_0^T$,

$$305 D(N) = \frac{(\sum_{i=1}^N R_i^2)^2}{\sum_{i=1}^N R_i^4} = \frac{(\text{tr}C_N)^2}{\text{tr}(C_N^2)}. \quad [\text{SI.50}]$$

306 Likewise, the effective dimensionality $D(M)$ of the projected concept manifold can be expressed in terms of its $M \times M$
 307 covariance matrix $C_M = \frac{1}{P}X'X'^T - \mathbf{x}_0'\mathbf{x}_0^T$, $D(M) = (\text{tr}C_M)^2/\text{tr}(C_M^2)$. Notice that

$$308 \text{tr}C_M = \text{tr}\left(\frac{1}{P}X^T A^T AX - A\mathbf{x}_0\mathbf{x}_0^T A^T\right) \quad [\text{SI.51}]$$

$$309 = \text{tr}\left(A^T A\left(\frac{1}{P}XX^T - \mathbf{x}_0\mathbf{x}_0^T\right)\right) \quad [\text{SI.52}]$$

$$310 = \text{tr}(A^T AC_N). \quad [\text{SI.53}]$$

311 Where we have used the cyclic property of the trace. Hence the relationship between $\text{tr}C_N$ and $\text{tr}C_M$ is governed by the
 312 statistics of $\Lambda \equiv A^T A$. Λ is a Wishart random matrix, with mean $\mathbb{E}[\Lambda] = I$ and variance $\text{Var}[\Lambda] = 1/M + I/M$. To estimate
 313 the effective dimensionality $D(M)$ of the projected concept manifold, we can compute the expected value of $(\text{tr}C_M)^2$ and
 314 $\text{tr}(C_M)^2$ over random realizations of Λ .

315 We will start with the denominator of $D(M)$, $\text{tr}(C_M)^2$,

$$316 \mathbb{E}[\text{tr}(C_M)^2] = \mathbb{E}[\text{tr}((\Lambda C_N)^2)] \quad [\text{SI.54}]$$

317 Diagonalizing $C_N = UR^2U^T$,

$$318 = \mathbb{E}[\text{tr}((U^T \Lambda U R^2)^2)] \quad [\text{SI.55}]$$

Defining $\tilde{\Lambda} \equiv U^T \Lambda U$,

$$= \mathbb{E}[\text{tr}((\tilde{\Lambda} R^2)^2)] \quad [\text{SI.56}]$$

$$= \mathbb{E}\left[\sum_{ij=1}^N \tilde{\Lambda}_{ij}^2 R_i^2 R_j^2\right] \quad [\text{SI.57}]$$

Notice that $\tilde{\Lambda}$ has the same statistics as Λ . Hence,

$$= \sum_{i=1}^N R_i^4 + \frac{1}{M} \sum_{i=1}^N R_i^4 + \frac{1}{M} \sum_{ij=1}^N R_i^2 R_j^2 \quad [\text{SI.58}]$$

$$= (1 + 1/M)\text{tr}(C_N^2) + (\text{tr}C_N)^2/M \quad [\text{SI.59}]$$

We now proceed to the numerator $(\text{tr}C_M)^2$,

313

$$\mathbb{E}[(\text{tr}C_M)^2] = \mathbb{E}[(\text{tr}(\tilde{\Lambda} R^2))^2] \quad [\text{SI.60}]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^N \tilde{\Lambda}_{ii} R_i^2\right)^2\right] \quad [\text{SI.61}]$$

$$= \mathbb{E}\left[\sum_{ij=1}^N \tilde{\Lambda}_{ii} \tilde{\Lambda}_{jj} R_i^2 R_j^2\right] \quad [\text{SI.62}]$$

$$= \sum_{ij=1}^N R_i^2 R_j^2 + \frac{2}{M} \sum_{i=1}^N R_i^4 \quad [\text{SI.63}]$$

$$= (\text{tr}C_N)^2 + 2\text{tr}(C_N^2)/M \quad [\text{SI.64}]$$

Combining our expressions for the numerator and the denominator, we obtain an estimate for the expected value of $D(M)$,

314

$$D(M) = \frac{(\text{tr}C_N)^2 + 2\text{tr}(C_N^2)/M}{(1 + 1/M)\text{tr}(C_N^2) + (\text{tr}C_N)^2/M} \quad [\text{SI.65}]$$

$$= \frac{D(N) + 2/M}{(1 + 1/M) + D(N)/M} \quad [\text{SI.66}]$$

Dropping the small terms of order $1/M$,

315

$$D(M) = \frac{D(N)}{1 + D(N)/M} \quad [\text{SI.67}] \quad 316$$

Therefore, provided that M is large compared to D , the random projection will have a negligible effect on the dimensionality. However, when M is on the order of D , distortions induced by the random projection will increase correlations between points on the manifold, significantly decreasing the effective dimensionality. Taking $N \rightarrow \infty$, this expression also allows us to extrapolate the asymptotic dimensionality $D_\infty = D(M)/(1 - D(M)/M)$ we might observe given access to arbitrarily many neurons. When concept manifolds occupy only a small fraction of the M available dimensions given recordings of M neurons, then recording from a few more neurons will have only a marginal effect. But when concept manifolds occupy a large fraction of the M available dimensions, recording from a few more neurons may significantly increase the estimated manifold dimensionality. Using this asymptotic dimensionality D_∞ , we can obtain a single expression for the estimated dimensionality $D(M)$ of concept manifolds given recordings of M neurons,

317

318

319

320

321

322

323

324

325

$$D^{-1}(M) = D_\infty^{-1} + M^{-1} \quad [\text{SI.68}] \quad 326$$

This prediction agrees well with random projections and random subsampling experiments on concept manifolds in IT and in trained DNNs (Fig. 6).

327

328

329 **B. Few-shot learning requires a number of neurons M greater than the concept manifold dimensionality D .** We next ask
 330 how the generalization error of few-shot learning depends on the number of subsampled neurons. We will study the simple case
 331 of 1-shot learning on identical ellipsoids in orthogonal subspaces, and demonstrate empirically that the predictions we derive
 332 hold well for the full case. Recall that the 1-shot learning SNR for identical ellipsoids in orthogonal subspaces (SI C) is given by

$$333 \quad \text{SNR}(N) = \frac{1}{2} \frac{\|\Delta \mathbf{x}_0\|^2}{\sqrt{D_a^{-1}}} = \frac{1}{2} \frac{\|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2}{\sqrt{\text{tr}(C_N^2)}} \quad [\text{SI.69}]$$

334 Then the signal-to-noise ratio in the projected subspace, $\text{SNR}(M)$, is given by

$$335 \quad \text{SNR}(M) = \frac{1}{2} \frac{\|A\mathbf{x}_0^a - A\mathbf{x}_0^b\|^2}{\sqrt{\text{tr}(C_M^2)}} \quad [\text{SI.70}]$$

336 We have already found that $\mathbb{E}[\text{tr}(C_M^2)] \approx \text{tr}(C_N^2) + (\text{tr}C_N)^2/M$. Furthermore, random projections are known to preserve
 337 the pairwise distances between high-dimensional points under fairly general settings, so that the distance between manifold
 338 centroids, $\|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2$, is preserved under the random projection, $\mathbb{E}[\|A\mathbf{x}_0^a - A\mathbf{x}_0^b\|^2] = \|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2$. Deviations from this average
 339 are quantified by the Johnson-Lindenstrauss Lemma, a fundamental result in the theory of random projections, which states
 340 that P points can be embedded in $M = \mathcal{O}(\log P/\epsilon^2)$ dimensions without distorting the distance between any pair of points by
 341 more than a factor of $(1 \pm \epsilon)$. Combining these results, we have

$$342 \quad \text{SNR}(M) = \frac{1}{2} \frac{\|\mathbf{x}_0^a - \mathbf{x}_0^b\|^2}{\sqrt{\text{tr}(C_N^2) + (\text{tr}C_N)^2/M}} = \frac{1}{2} \frac{\|\Delta \mathbf{x}_0\|^2}{\sqrt{D(N)^{-1}} \sqrt{1 + D(N)/M}} = \frac{1}{2} \frac{\text{SNR}(N)}{\sqrt{1 + D(N)/M}} \quad [\text{SI.71}]$$

343 Therefore, few-shot learning performance is unaffected by the random projection, provided that M is large compared to
 344 the concept manifold dimensionality. As before, we can extrapolate an asymptotic SNR given access to arbitrarily many
 345 neurons by taking $N \rightarrow \infty$, $\text{SNR}_\infty = \text{SNR}(M) \sqrt{1 + D_\infty/M}$. When concept manifolds occupy only a small fraction of the M
 346 available dimensions, a downstream neuron improves its few-shot learning performance only marginally by receiving inputs
 347 from a greater number of neurons. However, when concept manifolds occupy a large fraction of the M available dimensions, a
 348 downstream neuron can substantially improve its few-shot learning performance by receiving inputs from a greater number of
 349 neurons. Using this asymptotic signal-to-noise ratio, SNR_∞ , we can obtain a single expression for the few-shot learning SNR
 350 as a function of the number of input neurons, M ,

$$351 \quad \boxed{\text{SNR}(M) = \frac{\text{SNR}_\infty}{\sqrt{1 + D_\infty/M}}} \quad [\text{SI.72}]$$

352 This prediction agrees well with random projections and random subsampling experiments on concept manifolds in IT and
 353 in trained DNNs (Fig. 6).

354 **6. Comparing cognitive learning models in low and high dimensions**

355 A long line of work in the psychology literature has examined the relative advantages and disadvantages of prototype and
 356 exemplar theories of learning. Exemplar learning is performed by storing the representations of all training examples in memory,
 357 and categorizing a test example by comparing it to each stored example (Supp. Fig. 12a). Exemplar learning thus involves a
 358 choice of how to weight the similarity to each of the training examples. In one extreme, all similarities are weighted equally, so
 359 that a test example is categorized as concept a if its average similarity to each of the training examples of concept a is greater
 360 than its average similarity to each of the training examples of concept b . This limit is analytically tractable, and we find that it
 361 performs consistently worse than prototype learning. Indeed, in our experiments the optimal weighting is very close to the
 362 opposite extreme, in which only the most similar training example is counted, and the test example is assigned to whichever
 363 category this most similar training example belongs to (Supp. Fig. 12b). This limit corresponds to a nearest-neighbor (NN)
 364 decision rule. In numerical experiments on visual concept manifolds in trained DNNs (Fig. 8a), we find that prototype learning
 365 outperforms NN when D is large and the number of training examples m is small, while NN outperforms prototype learning in
 366 the opposite regime where D is small and the number of training examples m is large. Here we offer a theoretical justification
 367 for this behavior. We begin with an intuitive summary, and proceed to a more detailed derivation in the following section.

368 **A. Identifying the joint role of dimensionality D and number of training examples m .** The joint role of D and m arises because
 369 NN learning involves taking a minimum over the distances from each training example to the test example. However, as
 370 we have seen, in high dimensions these distances concentrate around their means with variance $1/D$. Under fairly general
 371 conditions, the minimum over m independent random variables with variance $1/D$ scales as $\sim \sqrt{\log m/D}$. When all other
 372 geometric quantities are held constant, the signal of NN learning scales as $\sqrt{\log m/D}$, while the signal of prototype learning
 373 is constant. Hence when $\log m$ is large compared to D , NN learning outperforms prototype learning, and when D is large
 374 compared to $\log m$, prototype learning outperforms NN learning.

375 We now derive the few-shot learning signal for NN learning, analogous to the few-shot learning signal we derived for
 376 prototype learning, Eq. 1. The setup for NN learning is the same as for prototype learning: we draw m training examples
 377 each from two concept manifolds, a and b ,

$$\mathbf{x}^{a\mu} = \mathbf{x}_0^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a s_i^{a\mu}, \quad \mathbf{x}^{b\mu} = \mathbf{x}_0^b + \sum_{i=1}^{D_b^{\text{tot}}} R_i^b \mathbf{u}_i^b s_i^{b\mu}, \quad [\text{SI.73}] \quad 378$$

Where $\mathbf{s}^{a\mu} \sim \text{Unif}(\mathbb{S}^{D_a^{\text{tot}}-1})$, $\mathbf{s}^{b\mu} \sim \text{Unif}(\mathbb{S}^{D_b^{\text{tot}}-1})$. We then draw a test example, 379

$$\boldsymbol{\xi}^a = \mathbf{x}_0^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a \sigma_i^a. \quad [\text{SI.74}] \quad 380$$

Where $\boldsymbol{\sigma}^a \sim \text{Unif}(\mathbb{S}^{D_a^{\text{tot}}-1})$. Rather than averaging the training examples into concept prototypes, to perform NN learning we simply compute the Euclidean distance from the test example to each of the training examples of concept a , 381
382

$$d_a^\mu \equiv \frac{1}{2} \|\boldsymbol{\xi}^a - \mathbf{x}^{a\mu}\|^2 \quad [\text{SI.75}]$$

$$= \frac{1}{2} \left\| \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a \sigma_i^a - \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a s_i^{a\mu} \right\|^2 \quad [\text{SI.76}]$$

$$= \frac{1}{2} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 (s_i^{a\mu})^2 + \frac{1}{2} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 (\sigma_i^a)^2 - \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 s_i^{a\mu} \sigma_i^a, \quad [\text{SI.77}]$$

And the distance to each of the training examples of concept b , 383

$$d_b^\mu \equiv \frac{1}{2} \|\boldsymbol{\xi}^a - \mathbf{x}^{b\mu}\|^2 \quad [\text{SI.78}]$$

$$= \frac{1}{2} \left\| \mathbf{x}_0^a - \mathbf{x}_0^b + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \mathbf{u}_i^a \sigma_i^a - \sum_{i=1}^{D_a^{\text{tot}}} R_i^b \mathbf{u}_i^b s_i^{b\mu} \right\|^2 \quad [\text{SI.79}]$$

$$= \frac{1}{2} \sum_{i=1}^{D_b^{\text{tot}}} (R_i^b)^2 (s_i^{b\mu})^2 + \frac{1}{2} \sum_{i=1}^{D_a^{\text{tot}}} (R_i^a)^2 (\sigma_i^a)^2 + R_a^2 \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \Delta \mathbf{x}_0 \cdot \mathbf{u}_i^a \sigma_i^a \quad [\text{SI.80}]$$

$$- R_a^2 \sum_{i=1}^{D_b^{\text{tot}}} R_i^b \Delta \mathbf{x}_0 \cdot \mathbf{u}_i^b s_i^{b\mu} + \sum_{ij} R_i^a R_j^b \mathbf{u}_i^a \cdot \mathbf{u}_j^b \sigma_i^a s_i^{b\mu} \quad [\text{SI.81}]$$

Then the generalization error is the probability that $\min_\mu d_a^\mu$ is less than $\min_\mu d_b^\mu$, $\varepsilon_a^{\text{NN}} = \mathbb{P}_{\mathbf{s}^{a\mu}, \mathbf{s}^{b\mu}, \boldsymbol{\sigma}^a} [h^{\text{NN}} < 0]$, where $h^{\text{NN}} = -\min_\mu d_a^\mu + \min_\mu d_b^\mu$. As we found in prototype learning, when concept manifolds are high-dimensional, d_a^μ, d_b^μ are approximately Gaussian-distributed. Again, in order to obtain dimensionless quantities we renormalize, $\tilde{d}_a^\mu = d_a^\mu / R_a^2$, $\tilde{d}_b^\mu = d_b^\mu / R_a^2$. We define the mean $\mu_a = \mathbb{E}[\tilde{d}_a^\mu]$ and variance $\sigma_a^2 = \text{Var}[\tilde{d}_a^\mu]$, given by 384
385
386
387

$$\mu_a = \frac{1}{2}, \quad \sigma_a^2 = \frac{1}{2} \frac{1}{D_a}, \quad [\text{SI.82}] \quad 388$$

which follow from eqs. [SI.23](#), [SI.28](#), and [SI.22](#). Similarly, we define $\mu_b = \mathbb{E}[\tilde{d}_b^\mu]$ and $\sigma_b^2 = \text{Var}[\tilde{d}_b^\mu]$, given by 389

$$\mu_b = \frac{1}{2} \|\Delta \mathbf{x}_0\|^2 + \frac{1}{2} R_b^2 R_a^{-2}, \quad [\text{SI.83}] \quad 390$$

$$\sigma_b^2 = \frac{1}{2} \frac{1}{D_a} + \frac{1}{2} \frac{(R_b^2)^2}{(R_a^2)^2} \frac{1}{D_b} + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_a\|^2 + \|\Delta \mathbf{x}_0 \cdot \mathbf{U}_b\|^2 + \|\mathbf{U}_a^T \mathbf{U}_b\|_F^2, \quad [\text{SI.84}] \quad 391$$

which follow from eqs. [SI.23](#), [SI.28](#), [SI.31](#), and [SI.33](#). Now we must evaluate the minimum over μ . The expected value of the minimum of m i.i.d. Gaussian random variables is given by $\mathbb{E}[\min_i X_i] \approx \mu_a - \sqrt{2 \log m} \sigma_a - \gamma$, where $X_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$, $i = 1, \dots, m$ and γ is the Euler-Mascheroni constant. Using this we can obtain the expected value of $\tilde{h}^{\text{NN}} = -\min_\mu \tilde{d}_a^\mu + \min_\mu \tilde{d}_b^\mu$, 392
393
394

$$\mathbb{E}[\tilde{h}^{\text{NN}}] = \mu_b - \mu_a + \sqrt{2 \log m} (\sigma_a - \sigma_b) \quad [\text{SI.85}]$$

$$= \frac{1}{2} \|\Delta \mathbf{x}_0\|^2 + \frac{1}{2} (R_b^2 R_a^{-2} - 1) + \sqrt{\frac{2 \log m}{D_a}} C \quad [\text{SI.86}]$$

395 Where we have pulled the dependence on D_a^{-1} out of σ_a, σ_b to define $C \equiv (\sigma_a - \sigma_b)\sqrt{D_a}$. C is greater than zero, since the
 396 signal-noise and noise-noise overlaps are much smaller than D_a^{-1} , and therefore $\sigma_a > \sigma_b$. Neglecting the bias term $\frac{1}{2}(R_b^2 R_a^{-2} - 1)$,
 397 we have that the signal of prototype learning is given by

$$398 \text{signal}^{\text{NN}} = \frac{1}{2}\|\Delta\mathbf{x}_0\|^2 + \sqrt{\frac{2\log m}{D_a}}C \quad [\text{SI.87}]$$

399 Compare this to the signal we found for prototype learning,

$$400 \text{signal}^{\text{proto}} = \frac{1}{2}\|\Delta\mathbf{x}_0\|^2 \quad [\text{SI.88}]$$

401 The NN signal is larger than the prototype learning signal. However the NN noise is also larger than the prototype learning
 402 noise. Hence when $\log m$ is large compared to D_a , NN outperforms prototype learning, and when D_a is large compared to
 403 $\log m$, prototype learning outperforms NN. We stop short of computing a full SNR for NN, since the random variables $\min_\mu \tilde{h}_a^\mu$
 404 and $\min_\mu \tilde{h}_b^\mu$ are not independent, and computing their correlation is not straightforward. However, the $D \sim \log m$ relationship
 405 we have identified here seems to reliably capture the behavior we observe in experiments on concept manifolds in a trained
 406 DNN (Fig. 8a), where we vary the dimensionality by projecting each concept manifold onto its top D principal components.

407 7. Geometry of DNN concept manifolds encodes a rich semantic structure.

408 The ImageNet21k dataset is organized into a semantic tree, with each of the 1k visual concepts in our evaluation set representing
 409 a leaf on this tree (see Methods J). To investigate the effect of semantic structure on concept learning, we sort the generalization
 410 error pattern of prototype learning in a trained ResNet50 to obey the structure of the semantic tree, so that semantically
 411 related concepts are adjacent, and semantically unrelated concepts are distant. The sorted error matrix (Supp. Fig. 1a)
 412 exhibits a prominent block diagonal structure, suggesting that most of the errors occur between concepts on the same branch of
 413 the semantic tree, and errors between two different branches of the semantic tree are exceedingly unlikely. In other words, the
 414 trained ResNet may confuse two types of Passerine birds, like songbirds and sparrows, but will almost never confuse a sparrow
 415 for a mammal or a fish. The sorted error matrix exhibits structure across many scales: some branches reveal very fine-grained
 416 discriminations (e.g. aquatic birds), while other branches reveal only coarser discriminations (e.g. Passerines). We suspect
 417 that the resolution with which the trained DNN represents different branches of the tree depends on the composition of the
 418 visual concepts seen during training, which we discuss further below. Finally, the sorted error pattern exhibits a pronounced
 419 asymmetry, with much larger errors above the diagonal than below. In particular, food and artifacts are more likely to be
 420 classified as plants and animals than plants and animals are to be classified as food and artifacts.

421 We additionally sort the patterns of individual geometric quantities: signal, bias, and signal-noise overlap, to reflect the
 422 semantic structure of the dataset (Supp. Fig. 1a, right). Signal exhibits a clear block diagonal structure, similar to the
 423 error pattern. Bias reveals a clear asymmetry: plants and animals have significantly higher bias than food and artifacts do,
 424 indicating that the radii of plant and animal concept manifolds are significantly smaller than the radii of food and artifact
 425 concept manifolds. Intriguingly, this suggests that the trained ResNet50 has learned more compact representations for plants
 426 and animals than for food and artifacts.

427 To quantify the extent to which each of these quantities depends on the semantic organization of visual concepts, we compute
 428 the average few-shot accuracy, signal, bias, and signal noise overlap across all pairs of concepts, as a function of the distance
 429 between the two concepts on the semantic tree, defined by the number of hops required to travel from one concept to the other
 430 (Supp. Fig. 1b). We find that few-shot learning accuracy, signal, and bias all increase significantly with semantic distance,
 431 while signal-noise overlaps decrease.

432 A related question is the effect of distribution shift between trained and novel concepts. The composition of the 1,000
 433 heldout visual concepts in our evaluation set is quite different from that of the 1,000 concepts seen during training. For
 434 instance, 10% of the training concepts are different breeds of dogs, while only 0.5% of the novel concepts are breeds of dogs.
 435 To quantify the effect of distribution shift, we measure the tree distance from each of the 1k novel concepts as the distance to
 436 its nearest neighbor among the 1k training concepts in ImageNet1k. In Supp. Fig. 1c we plot the average few-shot learning
 437 accuracy as a function of tree distance to the training set. Few-shot learning accuracy degrades slightly with distance from the
 438 training set, but the effect is not dramatic.

439 8. References

- 440 1. A Ansuini, A Laio, JH Macke, D Zoccolan, Intrinsic dimension of data representations in deep neural networks. *arXiv preprint arXiv:1905.12784* (2019).
- 441 2. S Recanatesi, et al., Dimensionality compression and expansion in Deep Neural Networks. *arXiv* (2019).
- 442 3. T Chen, S Kornblith, M Norouzi, G Hinton, A simple framework for contrastive learning of visual representations in *International conference on machine learning*. (PMLR), pp. 1597–1607 (2020).
- 443 4. A Radford, et al., Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).
- 444 5. NJ Majaj, H Hong, EA Solomon, JJ DiCarlo, Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *J. Neurosci.* **35**, 13402–13418 (2015).
- 445 6. BE Boser, IM Guyon, VN Vapnik, A Training Algorithm for Optimal Margin Classifiers. *Proc. Fifth Ann. Work. on Comput. Learn. Theory* p. 144–152 (1992).
- 446 7. P Gao, et al., A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* p. 214262 (2017).