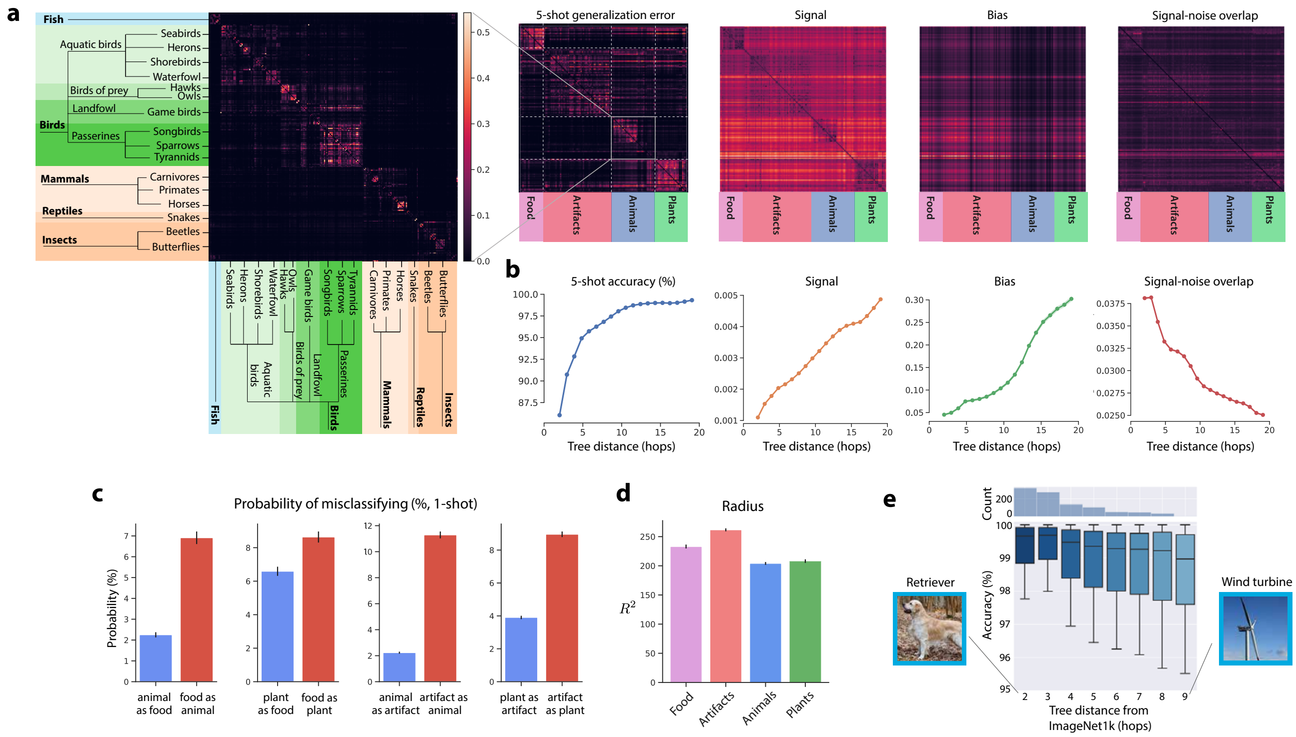
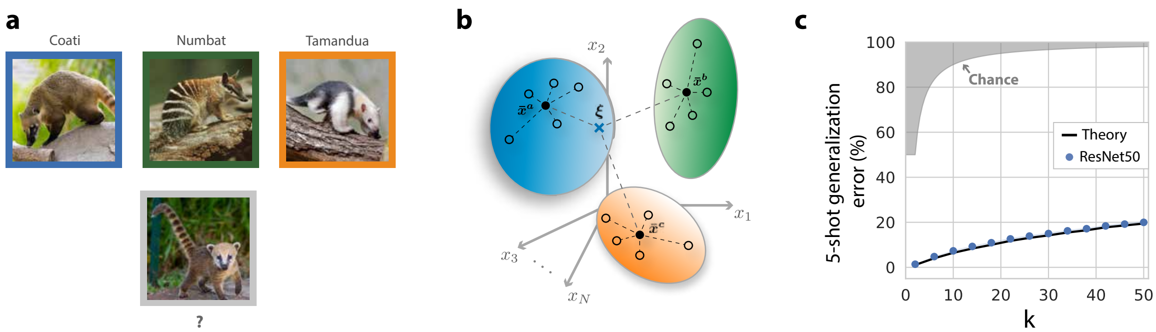
Supplementary materials: Neural representational geometry underlies few-shot concept learning

Contents

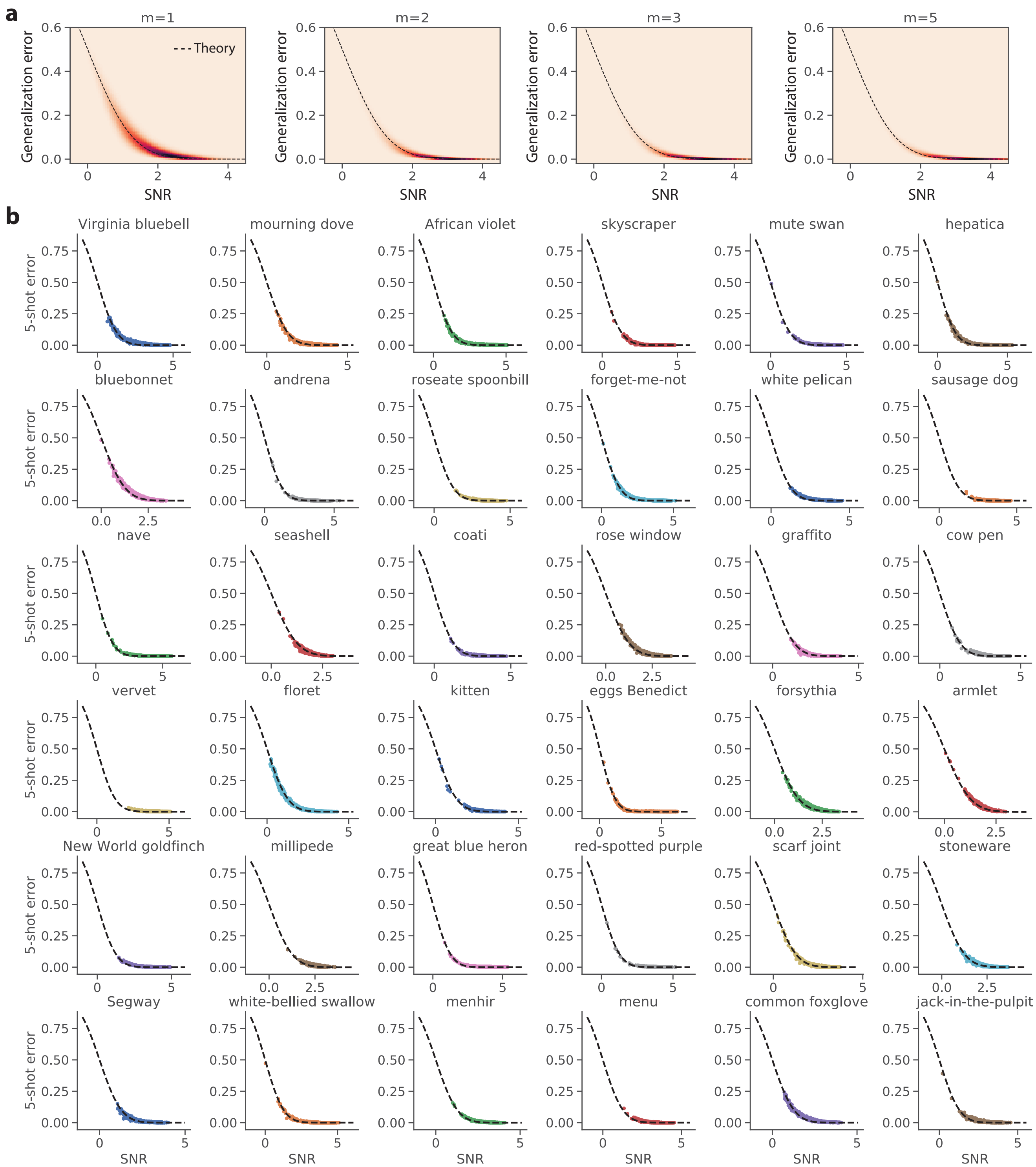
1. Supplementary figures



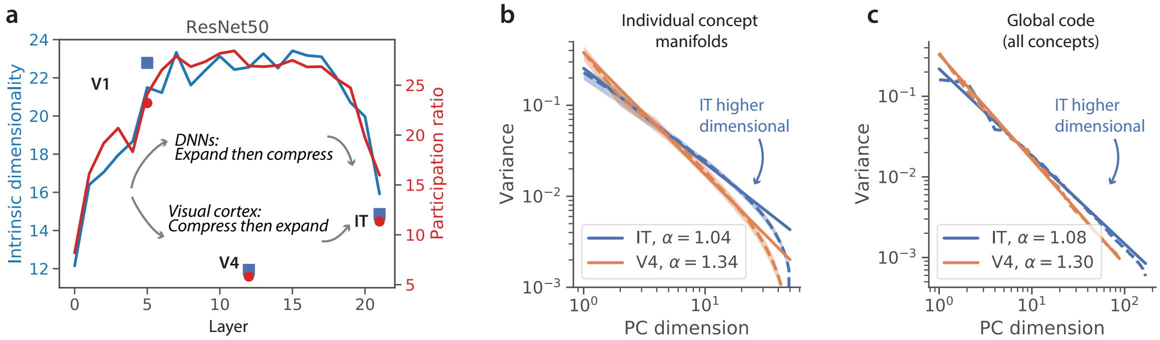
Supplementary Figure 1. Geometry of DNN concept manifolds encodes a rich semantic structure. See SI 7. **a**, We sort the generalization error pattern of prototype learning using concept manifolds from a trained ResNet50 to obey the hierarchical semantic structure of the ImageNet21k dataset. The sorted 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.



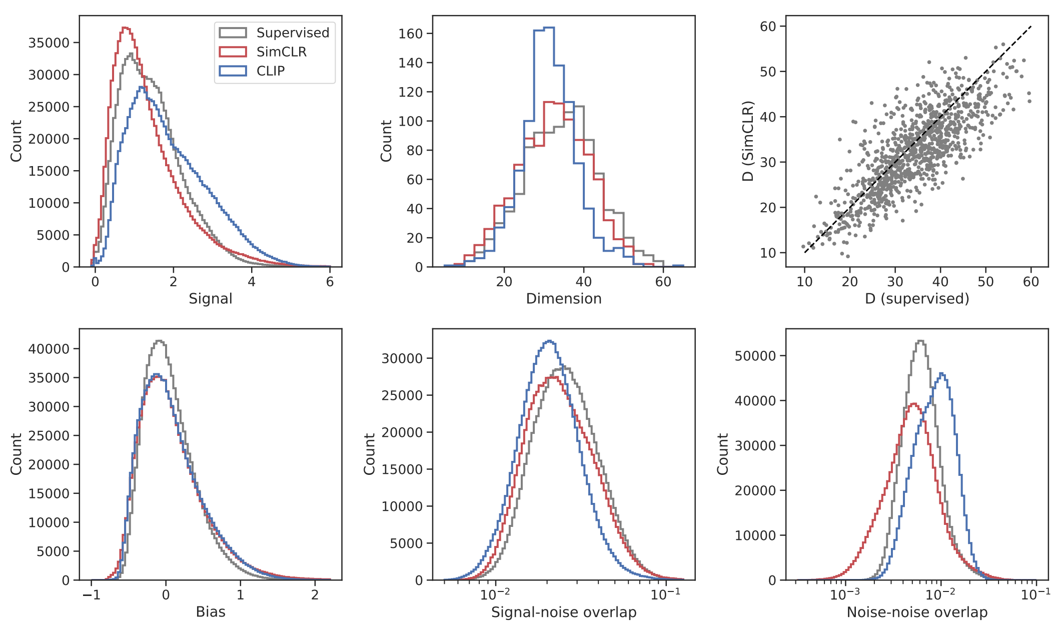
Supplementary Figure 2. Learning many novel concepts from few examples. Concept learning often involves categorizing more than two novel concepts. In SI D we extend our theory to model few-shot learning of k novel concepts. **a**, An example one-shot learning task for $k = 3$: does the test image in 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 ($\bar{\boldsymbol{x}}^1, \ldots, \bar{\boldsymbol{x}}^k$; solid circles). A test example ($\boldsymbol{\xi}$, 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. **b**, 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%).



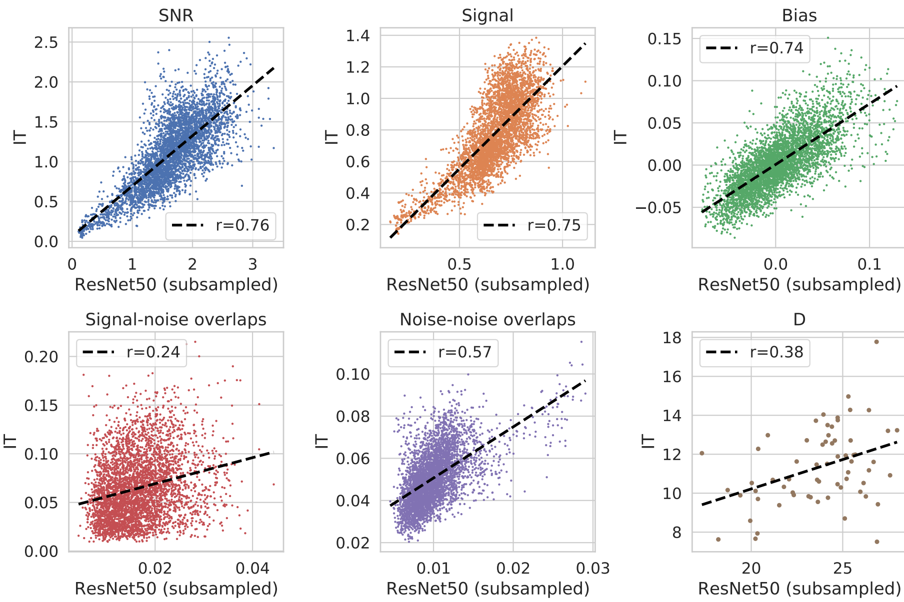
Supplementary Figure 3. geometric theory and few-shot learning experiments on a variety of novel concepts. a, We compare the empirical generalization error in 1–, 2–, 3–, and 5-shot learning experiments to the prediction from our geometric theory (Eq. SI.34) on all $1,000 \times 999$ pairs of visual concepts from the ImageNet21k datset, 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..



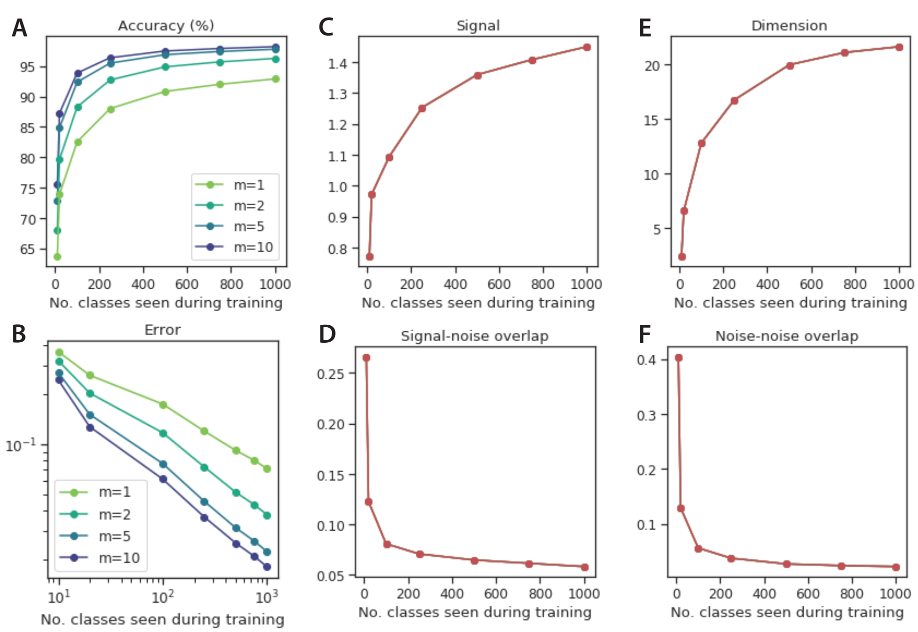
Supplementary Figure 4. Dimensionality diverges between trained DNNs and the primate visual pathway. a, To verify that the mismatch in concept manifold dimensionality between DNNs and visual cortex observed in Fig. 5**d** is not simply due to our choice to measure dimensionality using the participation 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 $D_{\text{ID}} = 0.53 \times D_{\text{SVD}} + 0.87$, 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.



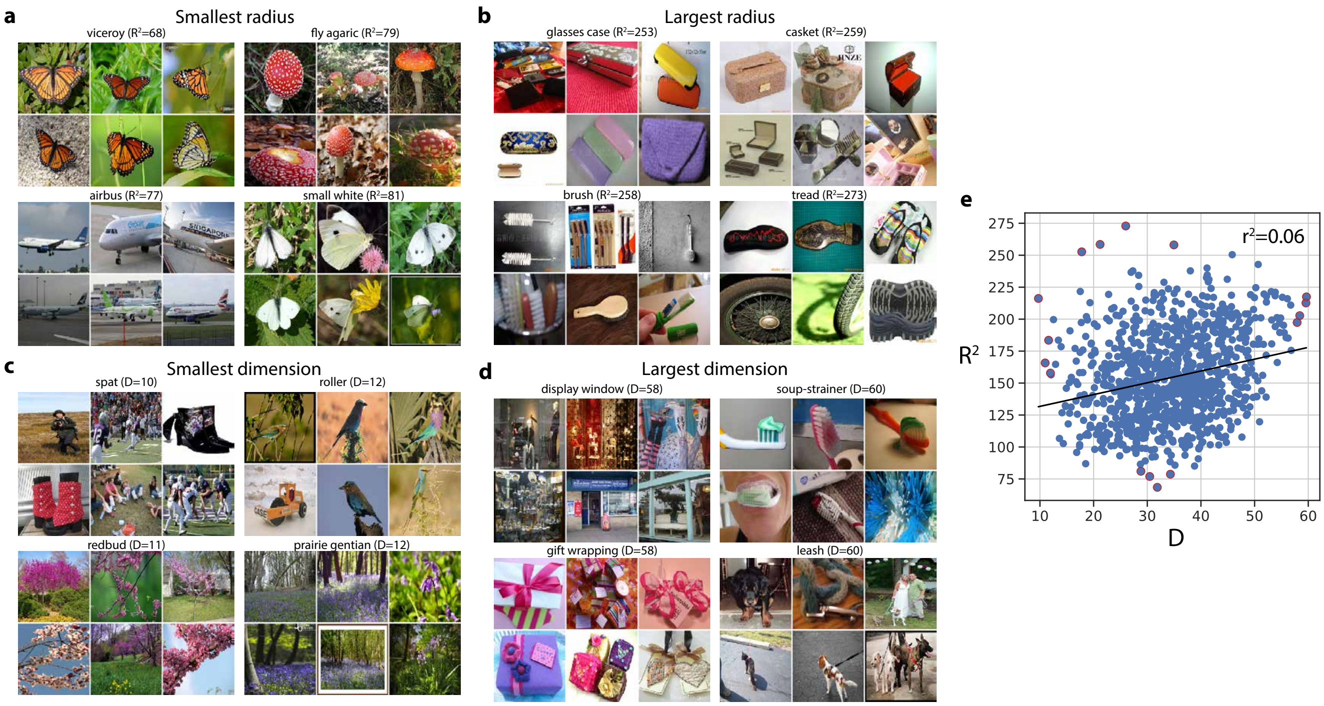
Supplementary Figure 5. Comparing concept manifold geometry across supervised, self-supervised, and unsupervised models. We compare the geometry of 1000 concept manifolds derived from a ResNet50 trained in either a supervised (grey) or self-supervised (SimCLR (3), red) manner, as well as manifolds derived 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).



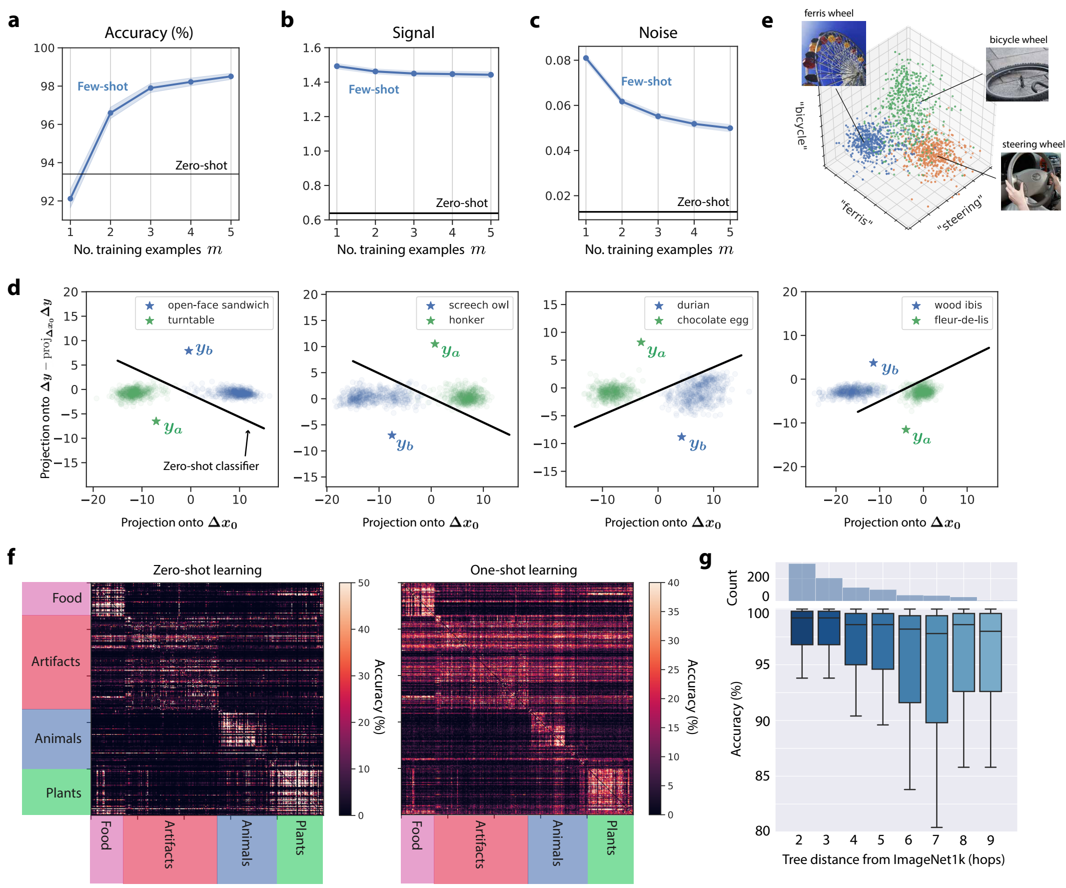
Supplementary Figure 6. Concept manifold geometry is correlated across primate IT cortex and trained DNNs. We estimate the geometry of visual concept manifolds in primate IT cortex and in trained DNNs in response to the same 64 naturalistic visual concepts (5). We then compute the correlation between 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 ($r = 0.38$, $p < 0.005$).



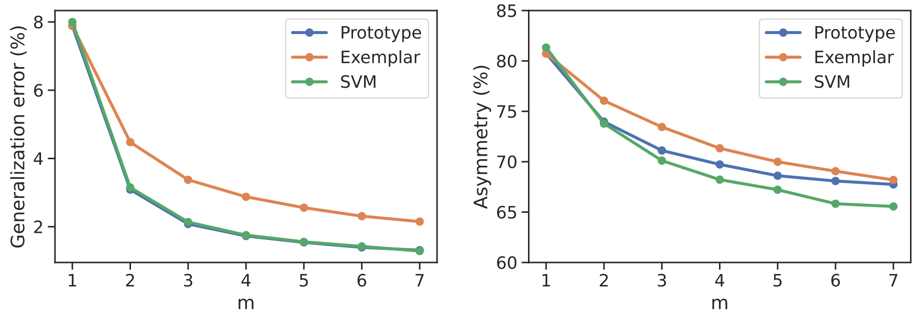
Supplementary Figure 7. Few-shot learning performance improves consistently with the number of concepts seen during training. To investigate the effect of training dataset size on novel concept learning and manifold geometry, we train DNNs (ResNet18) on random subsets of the ImageNet1k dataset with smaller 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.



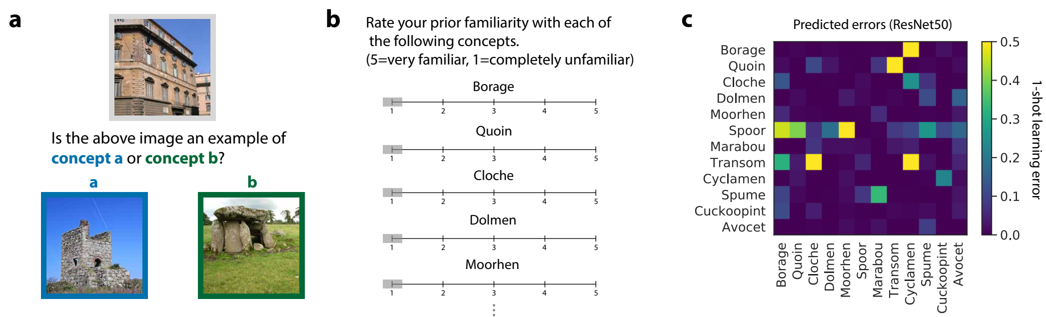
Supplementary Figure 8. Visual examples of concept manifolds with small and large dimension and radius. Among the $1,000$ novel visual concepts in our heldout set, we collect examples of the visual concepts whose manifolds in a trained ResNet50 have, **a**, smallest radius, **b**, largest radius, **c**, 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.



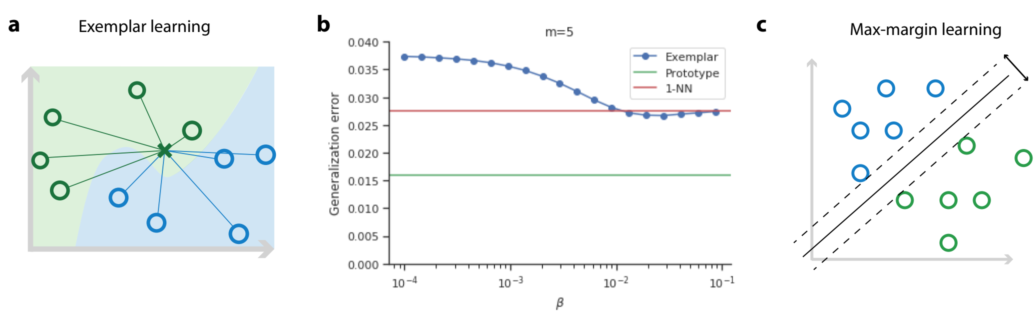
Supplementary Figure 9. How many words is a picture worth? Comparing prototypes derived from language and vision. See SI B. **a**, We compare the performance of prototype learning using prototypes derived from language representations (*zero-shot learning*, Sec. G) to those derived from one or a few visual examples (*few-shot learing*, 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, $\boldsymbol{\Delta x_0}$, and the language prototype readout direction, $\boldsymbol{\Delta y}$. 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 $\boldsymbol{\Delta x_0}$ direction, while the language prototypes pick out readout directions $\boldsymbol{\Delta y}$ 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. 1**e**, but the effect is not dramatic.



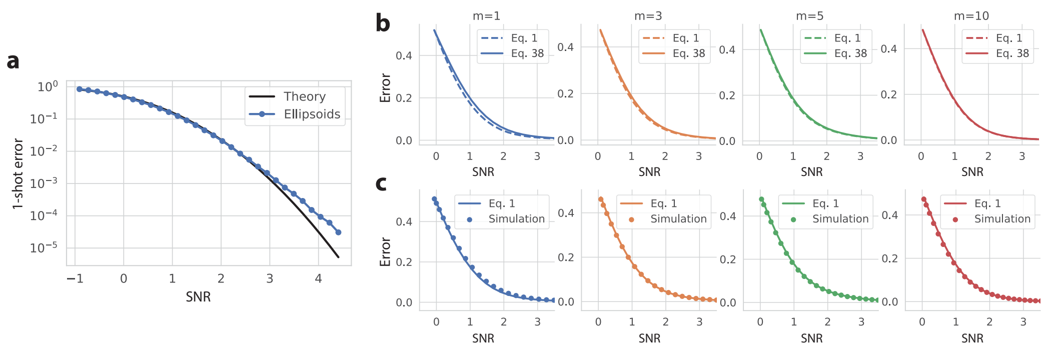
Supplementary Figure 10. Diverse decision rules exhibit asymmetry. We perform few-shot learning simulations using three different decision rules (prototype, exemplar, and SVM) in the same setting studied throughout the main text: feature layer representations from a ResNet50 pretrained on ImageNet1k. *Left*, Few-shot generalization error as a function of the number of the training examples m (reproduces Fig. 8**d**). *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 $\beta \to \infty$ 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 $\|\boldsymbol{\Delta x_0}\|^2$ 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 $\boldsymbol{x}_0^a, \boldsymbol{x}_0^b$, and its radius R_a, R_b, along a set of orthonormal axes $\boldsymbol{u}_i^a, \boldsymbol{u}_i^b, i = 1, \ldots, D$, 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, $\boldsymbol{u}_i^a \cdot \boldsymbol{u}_j^b = 0$, and orthogonal to the centroids, $(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b) \cdot \boldsymbol{u}_i^a = (\boldsymbol{x}_0^a - \boldsymbol{x}_0^b) \cdot \boldsymbol{u}_i^b = 0$, so that the signal-noise overlaps are zero. Thus a random example from each manifold can be written as,

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

where $\boldsymbol{s}^a, \boldsymbol{s}^b \sim \text{Unif}(\mathbb{S}^{D-1})$ are random vectors sampled uniformly from the D-dimensional unit sphere. We will study 1-shot learning in this section, using $\boldsymbol{x}^a, \boldsymbol{x}^b$ 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, $\boldsymbol{\xi}^a$, based on whether it is more similar to $\boldsymbol{x}^a$ or $\boldsymbol{x}^b$. 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 $\boldsymbol{\xi}^a$ can be written as,

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

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

$$h = \frac{R_a^2}{2}\Big(\|\boldsymbol{\Delta x_0}\|^2 + R_b^2 R_a^{-2} - 1\Big) + R_a^2\ \boldsymbol{s}^a\cdot\boldsymbol{\sigma}^a. \qquad \text{[SI.3]}$$

Where we have defined $\boldsymbol{\Delta x_0} = (\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)/R_a$. Thus we can evaluate the generalization error by computing $\varepsilon_a = \mathbb{P}[h \le 0]$ over all draws of the training and test examples. Defining $\Delta = \frac{1}{2}\big(\|\boldsymbol{\Delta x_0}\|^2 + R_b^2 R_a^{-2} - 1\big)$,

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

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,

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

Writing the delta and step functions using their integral representations,

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

We now perform the Gaussian integral over $\boldsymbol{s}^a$,

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $R_a^2 < \frac{1}{3}(\|\boldsymbol{\Delta x_0}\|^2 + R_b^2)$. 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,

$$\varepsilon_a = H(\sqrt{D}\Delta) = H\left(\frac{1}{2}\frac{\|\boldsymbol{\Delta x_0}\|^2 + R_b^2 R_a^{-2} - 1}{\sqrt{D^{-1}}}\right) \qquad \text{[SI.14]}$$

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 $\|\boldsymbol{\Delta x_0}\|^2$. Second, the SNR increases as the manifold dimensionality D increases. As Fig. 2**c** shows, this is due to the fact that the projection of each manifold onto the linear readout direction $\boldsymbol{w}$ 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 $R_a\sqrt{1 - \|\boldsymbol{\Delta x_0}\|^2}$ 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 $\boldsymbol{w}$ 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. 13**a**. 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 $\mu = 1, \ldots, m$ training examples each from two concept manifolds, a and b,

$$\boldsymbol{x}^{a\mu} = \boldsymbol{x_0}^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \boldsymbol{u}_i^a s_i^{a\mu}, \quad \boldsymbol{x}^{b\mu} = \boldsymbol{x_0}^b + \sum_{i=1}^{D_b^{\text{tot}}} R_i^b \boldsymbol{u}_i^b s_i^{b\mu}, \qquad \text{[SI.15]}$$

Where $\boldsymbol{x_0^a}, \boldsymbol{x_0^b}$ are the manifold centroids, and R_i^a, R_i^b are the radii along each axis, $\boldsymbol{u}_i^a, \boldsymbol{u}_i^b$. $\boldsymbol{s}^{a\mu} \sim \text{Unif}(\mathbb{S}^{D_a^{\text{tot}}-1}), \boldsymbol{s}^{b\mu} \sim \text{Unif}(\mathbb{S}^{D_b^{\text{tot}}-1})$ 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, $\bar{\boldsymbol{x}}^a$ and $\bar{\boldsymbol{x}}^b$,

$$\bar{\boldsymbol{x}}^a = \boldsymbol{x_0}^a + \frac{1}{m}\sum_{i=1}^{m}\sum_{i=1}^{D_a^{\text{tot}}} R_i^a \boldsymbol{u}_i^a s_i^{a\mu}, \quad \bar{\boldsymbol{x}}^b = \boldsymbol{x_0}^b + \frac{1}{m}\sum_{i=1}^{m}\sum_{i=1}^{D_b^{\text{tot}}} R_i^b \boldsymbol{u}_i^b s_i^{b\mu}, \qquad \text{[SI.16]}$$

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

$$\boldsymbol{\xi}^a = \boldsymbol{x_0}^a + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \boldsymbol{u}_i^a \sigma_i^a, \qquad \text{[SI.17]}$$

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

$$\begin{aligned} h = \frac{1}{2}\|\boldsymbol{x_0}^a - \boldsymbol{x_0}^b\|^2 + \frac{1}{m}\sum_{i=1}^{D_a^{\text{tot}}}\sum_{\mu=1}^{m}(R_i^a)^2 s_i^{a\mu}\sigma_i^a + \frac{1}{2m^2}\sum_{i=1}^{D_b^{\text{tot}}}\big(R_i^b\sum_{\mu=1}^{m}s_i^{b\mu}\big)^2 - \frac{1}{2m^2}\sum_{i=1}^{D_a^{\text{tot}}}\big(R_i^a\sum_{\mu=1}^{m}s_i^{a\mu}\big)^2 \\ + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a\sigma_i^a(\boldsymbol{x_0}^a - \boldsymbol{x_0}^b)\cdot\boldsymbol{u}_i^a + \frac{1}{m}\sum_{i=1}^{D_b^{\text{tot}}}\sum_{\mu=1}^{m} R_i^b s_i^{b\mu}(\boldsymbol{x_0}^a - \boldsymbol{x_0}^b)\cdot\boldsymbol{u}_i^b + \frac{1}{m}\sum_{ij}\sum_{\mu=1}^{m} R_i^a R_j^b \sigma_i^a s_i^{b\mu}\boldsymbol{u}_i^a\cdot\boldsymbol{u}_j^b \end{aligned} \qquad \text{[SI.18]}$$

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

Signal. The first term in Eq. SI.18, corresponding to signal, is fixed across different draws of the training and test examples, and so has zero variance. Its mean is given by $\frac{1}{2}\|\boldsymbol{\Delta x_0}\|^2$, where $\boldsymbol{\Delta x_0} = (\boldsymbol{x_0^a} - \boldsymbol{x_0^b})/\sqrt{R_a^2}$, and $R_a^2 \equiv \frac{1}{D_a^{\text{tot}}}\sum_{i=1}^{D_a^{\text{tot}}}(R_i^a)^2$.

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

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

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

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

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

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 concept manifold, quantified by the number of dimensions along which it varies significantly (7). Hence this term reflects the manifold dimensionality, and its variance is suppressed for large D_a.

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

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

$$= 1/m \quad \text{[SI.24]}$$

And the variance is given by,

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

There are three possible pairings of indices which yield even powers of s_i. Due to symmetry, all other pairings integrate to 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 $3m(m-1)$ terms of the form $(s_i^\mu)^2(s_i^\nu)^2$, 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 which integrates to $1/(D_a^{\text{tot}}(D_a^{\text{tot}}+2))$. Thus the integral gives,

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

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

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

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

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 sphere), the variance goes to zero. However, in practical situations the effective dimensionality is much smaller than the total 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,

$$\frac{1}{(R_a^2)^2}\mathrm{Var}\left[\sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \sigma_i^a (\boldsymbol{x_0^a} - \boldsymbol{x_0^b})\cdot \boldsymbol{u}_i^a\right] = \frac{1}{(R_a^2)^2}\sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 \big((\boldsymbol{x_0^a} - \boldsymbol{x_0^b})\cdot \boldsymbol{u}_i^a\big)^2 \int_{\mathbb{S}^{D_a^{\mathrm{tot}}-1}} \frac{d^{D_a^{\mathrm{tot}}}\boldsymbol{\sigma}}{S_{D_a^{\mathrm{tot}}-1}} (\sigma_i^a)^2 \quad \text{[SI.29]}$$

$$= \frac{1}{R_a^2}\sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 \big((\boldsymbol{x_0^a} - \boldsymbol{x_0^b})\cdot \boldsymbol{u}_i^a\big)^2 \quad \text{[SI.30]}$$

We refer to this term as signal-noise overlap because it quantifies the overlap between the noise directions $\boldsymbol{u}_i^a$ and the signal direction $\boldsymbol{\Delta x_0}$, weighted by the radii R_i^a along each noise direction. To make the notation more compact, we define $\boldsymbol{U}_a = [R_1^a u_1^a, \ldots, R_{D_a^{\mathrm{tot}}}^a u_{D_a^{\mathrm{tot}}}^a]/\sqrt{R_a^2}$, so that the signal-noise overlap takes the form,

$$\frac{1}{(R_a^2)^2}\mathrm{Var}\left[\sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \sigma_i^a (\boldsymbol{x_0^a} - \boldsymbol{x_0^b})\cdot \boldsymbol{u}_i^a\right] = \|\boldsymbol{\Delta x_0}\cdot \boldsymbol{U}_a\|^2, \quad \text{[SI.31]}$$

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,

$$\frac{1}{(R_a^2)^2}\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{D_a^{\mathrm{tot}}}\sum_{\mu=1}^{m} R_i^b s_i^{b\mu} (\boldsymbol{x_0^a} - \boldsymbol{x_0^b})\cdot \boldsymbol{u}_i^b\right] = \frac{1}{m}\|\boldsymbol{\Delta x_0}\cdot \boldsymbol{U}_b\|^2, \quad \text{[SI.32]}$$

where we have defined $\boldsymbol{U}_b = [R_1^b u_1^b, \ldots, R_{D_b^{\mathrm{tot}}}^b u_{D_b^{\mathrm{tot}}}^b]/\sqrt{R_a^2}$ in analogy to $\boldsymbol{U}_a$. 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,

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

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

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

$$\begin{aligned} \mu &= \frac{1}{2}\|\boldsymbol{\Delta x_0}\|^2 + \frac{1}{2}(R_b^2 R_a^{-2} - 1)/m, \\ \sigma^2 &= \frac{D_a^{-1}}{m} + \frac{D_a^{-1}}{2m^2}\left(1 - \frac{1}{m}\frac{D_a}{D_a^{\mathrm{tot}}}\right) + \frac{D_b^{-1}}{2m^2}\frac{(R_b^2)^2}{(R_a^2)^2}\left(1 - \frac{1}{m}\frac{D_b}{D_b^{\mathrm{tot}}}\right) \\ &\quad + \|\boldsymbol{\Delta x_0}\cdot \boldsymbol{U_a}\|^2 + \|\boldsymbol{\Delta x_0}\cdot \boldsymbol{U_b}\|^2/m + \|\boldsymbol{U}_a^T \boldsymbol{U}_b\|_F^2/m \end{aligned} \quad \text{[SI.34]}$$

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, $\varepsilon_a = \mathbb{P}[\tilde{h} \le 0] = H(\mathrm{SNR}) \equiv H(\mu/\sigma)$. 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. 13**b,c**, are the two noise terms arising from the bias, $\frac{D_a^{-1}}{2m^2}\left(1 - \frac{1}{m}\frac{D_a}{D_a^{\mathrm{tot}}}\right)$ and $\frac{D_b^{-1}}{2m^2}\frac{(R_b^2)^2}{(R_a^2)^2}\left(1 - \frac{1}{m}\frac{D_b}{D_b^{\mathrm{tot}}}\right)$, and the noise-noise overlaps term $\|\boldsymbol{U}_a^T \boldsymbol{U}_b\|_F^2/m$. 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 $m \ge 3$ (Supp. Fig. 13**b,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. 13**b**), we find that Eq. 1 and Eq. SI.34 are nearly indistinguishable for $m \ge 3$. 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 $m \ge 3$, but very slightly underestimates the generalization error for $m = 1$ (Supp. Fig. 13**c**). 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 $m \le 3$.

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

$$h_b = \frac{1}{2}\|\boldsymbol{\xi}^a - \bar{\boldsymbol{x}}^b\|^2 - \frac{1}{2}\|\boldsymbol{\xi}^a - \bar{\boldsymbol{x}}^a\|^2. \tag{SI.35}$$

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

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,

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

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

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

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

$$\varepsilon_a = 1 - \int_{\mathbb{R}_+^{k-1}} d^{k-1}\boldsymbol{h}\; p(\boldsymbol{h}) \tag{SI.38}$$

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 scaling of h_b, so we will work with the renormalized $\tilde{\boldsymbol{h}} = \boldsymbol{h}/R_a^2$ in order to obtain dimensionless quantities. We have already evaluated the mean $\mu_b = \mathbb{E}[\tilde{h}_b]$ and the diagonal covariance elements $\Sigma_{bb} = \mathrm{Var}[\tilde{h}_b]$ in SI C; these are just the signal and noise, 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 the expression for h_b in Eq. SI.18, we find that when $b \neq c$ three terms contribute,

$$\Sigma_{bc} = \frac{1}{(R_a^2)^2}\mathrm{Var}\left[\frac{1}{2m^2}\sum_{i=1}^{D_a^{\mathrm{tot}}}\big(R_i^a\sum_{\mu=1}^{m}s_i^{a\mu}\big)^2\right] + \frac{1}{(R_a^2)^2}\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{D_a^{\mathrm{tot}}}\sum_{\mu=1}^{m}(R_i^a)^2 s_i^{a\mu}\sigma_i^a\right]$$
$$+ \frac{1}{(R_a^2)^2}\mathrm{Var}\left[\left(\sum_{i=1}^{D_a^{\mathrm{tot}}}R_i^a\sigma_i^a(\boldsymbol{x}_0^a - \boldsymbol{x}_0^b)\cdot\boldsymbol{u}_i^a\right)\left(\sum_{i=1}^{D_a^{\mathrm{tot}}}R_i^a\sigma_i^a(\boldsymbol{x}_0^a - \boldsymbol{x}_0^c)\cdot\boldsymbol{u}_i^a\right)\right] \tag{SI.39}$$

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

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

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

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

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

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

where $\boldsymbol{\Delta x_0^{ab}} = (\boldsymbol{x_0^a} - \boldsymbol{x_0^b})/\sqrt{R_a^2}$, and $\boldsymbol{\Delta x_0^{ac}} = (\boldsymbol{x_0^a} - \boldsymbol{x_0^c})/\sqrt{R_a^2}$. Combining these terms, and re-inserting the terms for $b = c$ derived in Eq. SI.34, we obtain the full expression for the covariance,

$$\Sigma_{bc} = \frac{D_a^{-1}}{m} + \frac{D_a^{-1}}{2m^2}\left(1 - \frac{1}{m}\frac{D_a^{\text{tot}}}{D_a}\right) + (\boldsymbol{\Delta x_0^{ab}} \cdot \boldsymbol{U}_a)^T(\boldsymbol{\Delta x_0^{ac}} \cdot \boldsymbol{U}_a) + \delta_{bc}\left(\frac{D_b^{-1}}{2m^2}\frac{(R_b^2)^2}{(R_a^2)^2}\left(1 - \frac{D_b^{\text{tot}}}{D_b}\right) + \frac{1}{m}\|\boldsymbol{\Delta x_0^{ab}} \cdot \boldsymbol{U}_b\|^2 + \frac{1}{m}\|\boldsymbol{U}_a^T \boldsymbol{U}_b\|_F^2\right). \quad \text{[SI.43]}$$

Recall from Eq. SI.34 that $\boldsymbol{\mu}$ is given by,

$$\mu_b = \frac{1}{2}\|\boldsymbol{\Delta x_0}\|^2 + \frac{1}{2}(R_b^2 R_a^{-2} - 1)/m \quad \text{[SI.44]}$$

Integrating the multivariate Gaussian with mean $\boldsymbol{\mu}$ 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 $\boldsymbol{y}^a, \boldsymbol{y}^b$ 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 $\boldsymbol{\xi}^a$ it elicits in a population of IT-like neurons. We then classify $\boldsymbol{\xi}^a$ according to which prototype it is closer to. As in few-shot learning, we assume that $\boldsymbol{\xi}^a$ lies along an underlying ellipsoidal manifold,

$$\boldsymbol{\xi}^a = \boldsymbol{x_0^a} + \sum_{i=1}^{D_a^{\text{tot}}} R_i^a \boldsymbol{u}_i^a \sigma_i^a, \quad \text{[SI.45]}$$

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

$$h = \frac{1}{2}\|\boldsymbol{x_0^a} - \boldsymbol{y}^b\|^2 - \frac{1}{2}\|\boldsymbol{x_0^a} - \boldsymbol{y}^a\|^2 - \sum_{i=1}^{D_a^{\text{tot}}}(\boldsymbol{y}^a - \boldsymbol{y}^b) \cdot \boldsymbol{u}_i^a R_i^a \sigma_i^a \quad \text{[SI.46]}$$

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,

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

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

Defining $\boldsymbol{\Delta y} = (\boldsymbol{y}^a - \boldsymbol{y}^b)/\sqrt{R_a^2}$ the variance can be written more compactly as $(R_a^2)^2\|\boldsymbol{\Delta y} \cdot \boldsymbol{U}_a\|^2$. 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,

$$\text{SNR}_a^{\text{zero-shot}} = \frac{1}{2}\frac{\|\boldsymbol{x_0^a} - \boldsymbol{y}^b\|^2 - \|\boldsymbol{x_0^a} - \boldsymbol{y}^a\|^2}{\|\boldsymbol{\Delta y} \cdot \boldsymbol{U}_a\|_2} \quad \text{[SI.49]}$$

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. 7**d**), and affords deeper insight into the performance of zero-shot learning, as we show in Fig. 7**e**, and explore further in the following section.

B. How many words is a picture worth? Comparing prototypes derived from language and vision.. We found that prototypes derived from language yield a better generalization accuracy than those derived from a single visual example (Section G), but not two or more visual examples (Supp. Fig. 9**a**). To better understand this behavior, we use our geometric theory for zero-shot learning, Eq. 3, to decompose the zero-shot learning SNR into a contribution from the 'signal', which quantifies how closely the estimated prototypes match the true manifold centroids, and a contribution from the 'noise', which quantifies the overlap between the readout direction and the noise directions. We use the same theory to examine the prototypes generated by few-shot learning, even though these prototypes vary across different draws of the training examples, by averaging the signal and noise over many different draws of the training examples. This allows us to compare zero-shot learning and few-shot learning in the same framework, to understand whether the enhanced performance of zero-shot learning is due to higher signal (i.e. a closer match between estimated prototypes and true centroids) or lower noise (i.e. less overlap between the readout and noise directions). In Supp. Fig. 9**b,c** we show that both signal and noise are significantly lower for zero-shot learning than for few-shot learning. Therefore, 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 overall generalization accuracy by picking out linear readout directions which overlap significantly less with the concept manifolds' noise directions. We visualize these directions in Supp. Fig. 9**d** by projecting pairs of concept manifolds into the two-dimensional space spanned by the signal direction $\boldsymbol{\Delta x_0}$ and the language prototype readout direction $\boldsymbol{\Delta y}$. In each case, the manifolds' variability is predominantly along the signal direction $\boldsymbol{\Delta x_0}$, while the language prototypes pick out readout directions $\boldsymbol{\Delta y}$ with much lower variability.

5. How many neurons are required for concept learning?

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

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

A. Concept manifold dimensionality under random projections.. Consider randomly projecting each point $\boldsymbol{x} \in \mathbb{R}^N$ on a concept manifold to a lower-dimensional subspace, $A\boldsymbol{x} = \boldsymbol{x'} \in \mathbb{R}^M$ using a random projection matrix $A \in \mathbb{R}^{M\times N}$, $A_{ij} \sim \mathcal{N}(0, 1/M)$. We collect all points on the original concept manifold into an $N \times P$ matrix X, and collect all points on the projected concept manifold into an $M \times P$ matrix $X' = AX$. Recall that the effective dimensionality $D(N)$ of the original concept manifold can be expressed in terms of its $N \times N$ covariance matrix $C_N = \frac{1}{P}XX^T - \boldsymbol{x_0}\boldsymbol{x_0}^T$,

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

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

$$\mathrm{tr}C_M = \mathrm{tr}\Big(\frac{1}{P}X^TA^TAX - A\boldsymbol{x_0}\boldsymbol{x_0}^TA^T\Big) \quad \text{[SI.51]}$$

$$= \mathrm{tr}\Big(A^TA\Big(\frac{1}{P}XX^T - \boldsymbol{x_0}\boldsymbol{x_0}^T\Big)\Big) \quad \text{[SI.52]}$$

$$= \mathrm{tr}(A^TAC_N). \quad \text{[SI.53]}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 $N \to \infty$, 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,

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

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

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

$$\mathrm{SNR}(N) = \frac{1}{2}\frac{\|\boldsymbol{\Delta x_0}\|^2}{\sqrt{D_a^{-1}}} = \frac{1}{2}\frac{\|\boldsymbol{x_0^a} - \boldsymbol{x_0^b}\|^2}{\sqrt{\mathrm{tr}(C_N^2)}} \qquad \text{[SI.69]}$$

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

$$\mathrm{SNR}(M) = \frac{1}{2}\frac{\|A\boldsymbol{x_0^a} - A\boldsymbol{x_0^b}\|^2}{\sqrt{\mathrm{tr}(C_M^2)}} \qquad \text{[SI.70]}$$

We have already found that $\mathbb{E}[\mathrm{tr}(C_M^2)] \approx \mathrm{tr}(C_N^2) + (\mathrm{tr}C_N)^2/M$. Furthermore, random projections are known to preserve the pairwise distances between high-dimensional points under fairly general settings, so that the distance between manifold centroids, $\|\boldsymbol{x_0^a} - \boldsymbol{x_0^b}\|^2$, is preserved under the random projection, $\mathbb{E}[\|A\boldsymbol{x_0^a} - A\boldsymbol{x_0^b}\|^2] = \|\boldsymbol{x_0^a} - \boldsymbol{x_0^b}\|^2$. Deviations from this average are quantified by the Johnson-Lindenstrauss Lemma, a fundamental result in the theory of random projections, which states 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 more than a factor of $(1 \pm \epsilon)$. Combining these results, we have

$$\mathrm{SNR}(M) = \frac{1}{2}\frac{\|\boldsymbol{x_0^a} - \boldsymbol{x_0^b}\|^2}{\sqrt{\mathrm{tr}(C_N^2) + (\mathrm{tr}C_N)^2/M}} = \frac{1}{2}\frac{\|\boldsymbol{\Delta x_0}\|^2}{\sqrt{D(N)^{-1}}\sqrt{1 + D(N)/M}} = \frac{1}{2}\frac{\mathrm{SNR}(N)}{\sqrt{1 + D(N)/M}} \qquad \text{[SI.71]}$$

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

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

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

6. Comparing cognitive learning models in low and high dimensions

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

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

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

$$\boldsymbol{x}^{a\mu} = \boldsymbol{x_0}^a + \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{u}_i^a s_i^{a\mu}, \quad \boldsymbol{x}^{b\mu} = \boldsymbol{x_0}^b + \sum_{i=1}^{D_b^{\mathrm{tot}}} R_i^b \boldsymbol{u}_i^b s_i^{b\mu}, \tag{SI.73}$$

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

$$\boldsymbol{\xi}^a = \boldsymbol{x_0}^a + \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{u}_i^a \sigma_i^a. \tag{SI.74}$$

Where $\boldsymbol{\sigma}^a \sim \mathrm{Unif}(\mathbb{S}^{D_a^{\mathrm{tot}}-1})$. 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,

$$\begin{aligned}
d_a^\mu &\equiv \frac{1}{2}\|\boldsymbol{\xi}^a - \boldsymbol{x}^{a\mu}\|^2 && \text{[SI.75]}\\
&= \frac{1}{2}\|\sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{u}_i^a \sigma_i^a - \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{u}_i^a s_i^{a\mu}\|^2 && \text{[SI.76]}\\
&= \frac{1}{2}\sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 (s_i^{a\mu})^2 + \frac{1}{2}\sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 (\sigma_i^a)^2 - \sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 s_i^{a\mu} \sigma_i^a, && \text{[SI.77]}
\end{aligned}$$

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

$$\begin{aligned}
d_b^\mu &\equiv \frac{1}{2}\|\boldsymbol{\xi}^a - \boldsymbol{x}^{b\mu}\|^2 && \text{[SI.78]}\\
&= \frac{1}{2}\|\boldsymbol{x_0}^a - \boldsymbol{x_0}^b + \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{u}_i^a \sigma_i^a - \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^b \boldsymbol{u}_i^b s_i^{b\mu}\|^2 && \text{[SI.79]}\\
&= \frac{1}{2}\sum_{i=1}^{D_b^{\mathrm{tot}}} (R_i^b)^2 (s_i^{b\mu})^2 + \frac{1}{2}\sum_{i=1}^{D_a^{\mathrm{tot}}} (R_i^a)^2 (\sigma_i^a)^2 + R_a^2 \sum_{i=1}^{D_a^{\mathrm{tot}}} R_i^a \boldsymbol{\Delta x_0} \cdot \boldsymbol{u}_i^a \sigma_i^a && \text{[SI.80]}\\
&\quad - R_a^2 \sum_{i=1}^{D_b^{\mathrm{tot}}} R_i^b \boldsymbol{\Delta x_0} \cdot \boldsymbol{u}_i^b s_i^{b\mu} + \sum_{ij} R_i^a R_j^b \boldsymbol{u}_i^a \cdot \boldsymbol{u_j^b} \sigma_i^a s_i^{b\mu} && \text{[SI.81]}
\end{aligned}$$

Then the generalization error is the probability that $\min_\mu d_a^\mu$ is less than $\min_\mu d_b^\mu$, $\varepsilon_a^{\mathrm{NN}} = \mathbb{P}_{\boldsymbol{s^{a\mu}},\boldsymbol{s^{b\mu}},\boldsymbol{\sigma^a}}[h^{\mathrm{NN}} < 0]$, where $h^{\mathrm{NN}} = -\min_\mu d_a^\mu + \min_\mu d_b^\mu$. 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 $\sigma_a^2 = \mathrm{Var}[\tilde{d}_a^\mu]$, given by

$$\mu_a = \frac{1}{2}, \quad \sigma_a^2 = \frac{1}{2}\frac{1}{D_a}, \tag{SI.82}$$

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

$$\mu_b = \frac{1}{2}\|\boldsymbol{\Delta x_0}\|^2 + \frac{1}{2}R_b^2 R_a^{-2}, \tag{SI.83}$$

$$\sigma_b^2 = \frac{1}{2}\frac{1}{D_a} + \frac{1}{2}\frac{(R_b^2)^2}{(R_a^2)^2}\frac{1}{D_b} + \|\boldsymbol{\Delta x_0} \cdot \boldsymbol{U}_a\|^2 + \|\boldsymbol{\Delta x_0} \cdot \boldsymbol{U}_b\|^2 + \|\boldsymbol{U}_a^T \boldsymbol{U}_b\|_F^2, \tag{SI.84}$$

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 $\tilde{h}^{\mathrm{NN}} = -\min_\mu \tilde{d}_a^\mu + \min_\mu \tilde{d}_b^\mu$,

$$\begin{aligned}
\mathbb{E}[\tilde{h}^{\mathrm{NN}}] &= \mu_b - \mu_a + \sqrt{2\log m}(\sigma_a - \sigma_b) && \text{[SI.85]}\\
&= \frac{1}{2}\|\boldsymbol{\Delta x_0}\|^2 + \frac{1}{2}(R_b^2 R_a^{-2} - 1) + \sqrt{\frac{2\log m}{D_a}}C && \text{[SI.86]}
\end{aligned}$$

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 signal-noise and noise-noise overlaps are much smaller than D_a^{-1}, and therefore $\sigma_a > \sigma_b$. Neglecting the bias term $\frac{1}{2}(R_b^2 R_a^{-2} - 1)$, we have that the signal of prototype learning is given by

$$\text{signal}^{\text{NN}} = \frac{1}{2}\|\mathbf{\Delta x_0}\|^2 + \sqrt{\frac{2\log m}{D_a}}C \tag{SI.87}$$

Compare this to the signal we found for prototype learning,

$$\text{signal}^{\text{proto}} = \frac{1}{2}\|\mathbf{\Delta x_0}\|^2 \tag{SI.88}$$

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

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

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

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

To quantify the extent to which each of these quantities depends on the semantic organization of visual concepts, we compute 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 by the number of hops required to travel from one concept to the other (Supp. Fig. 1**b**). We find that few-shot learning accuracy, signal, and bias all increase significantly with semantic distance, while signal-noise overlaps decrease.

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

8. References

1. A Ansuini, A Laio, JH Macke, D Zoccolan, Intrinsic dimension of data representations in deep neural networks. *arXiv preprint arXiv:1905.12784* (2019).
2. S Recanatesi, et al., Dimensionality compression and expansion in Deep Neural Networks. *arXiv* (2019).
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).
4. A Radford, et al., Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020* (2021).
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).
6. BE Boser, IM Guyon, VN Vapnik, A Training Algorithm for Optimal Margin Classifiers. *Proc. Fifth Annu. Work. on Comput. Learn. Theory* p. 144–152 (1992).
7. P Gao, et al., A theory of multineuronal dimensionality, dynamics and measurement. *bioRxiv* p. 214262 (2017).