

Visualizing the geometry of labeled high-dimensional data with spheres

Andrew D Zaharia^{1,*}, Anish S Potnis¹, Alexander Walther², and Nikolaus Kriegeskorte^{1,3}

¹Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027

²Realeyes, 2 Riding House St, London W1W 7FA, UK

³Departments of Psychology, Neuroscience, and Electrical Engineering, Columbia University, New York, NY 10027

*Address correspondence to: andrew.z@columbia.edu

July 5, 2021

Abstract

Data visualizations summarize high-dimensional distributions in two or three dimensions. Dimensionality reduction entails a loss of information, and what is preserved differs between methods. Existing methods preserve the local or the global geometry of the points, and most techniques do not consider labels. Here we introduce “hypersphere2sphere” (H2S), a new method that aims to visualize not the points, but the relationships between the labeled distributions. H2S fits a hypersphere to each labeled set of points in a high-dimensional space and visualizes each hypersphere as a sphere in 3D (or circle in 2D). H2S perfectly captures the geometry of up to 4 hyperspheres in 3D (or 3 in 2D), and approximates the geometry for larger numbers of distributions, matching the sizes (radii), and the pairwise separations (between-center distances) and overlaps (along the center-connection line). The resulting visualizations are robust to sampling imbalances. Leveraging labels and the sphere as the simplest geometrical primitive, H2S provides an important addition to the toolbox of visualization techniques.

Introduction

Data visualization engages our visual system’s capacity for parallel processing, enabling us to understand a complex set of relationships at a glance. Visualization of high-dimensional data typically involves embedding the data in a 2- or 3-dimensional space, such that every high-dimensional data point is displayed in the visualization. The embedding, however, requires dimensionality reduction, which inevitably entails a loss of information. Different methods make different trade-offs in terms of what information to preserve and what to discard. The most commonly used visualization methods reduce dimensionality without making use of data labels.

Several popular dimensionality reduction techniques attempt to preserve the geometry of the data (Figure 1a). Principal Component Analysis (PCA) preserves the geometry of the high-dimensional data exactly, but only for the two or three dimen-

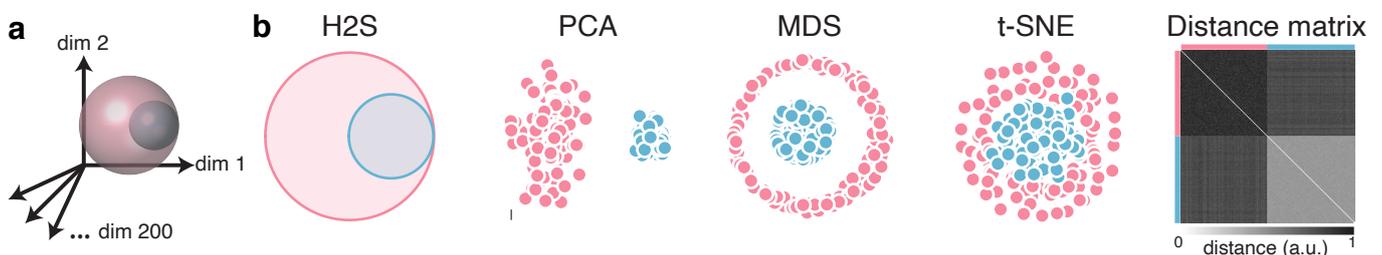


Figure 1. How should labeled high-dimensional distributions be visualized? Visualization of 200-dimensional simulated data with the proposed method hypersphere-to-sphere (H2S) and traditional methods (PCA, MDS, t-SNE). (a) The ground-truth data-generating distributions here are uniform balls in 200 dimensions. The larger 200D ball contains the smaller one entirely. Their bounding hyperspheres touch in a single point. The data set consists of 100 points sampled independently from each 200D-ball. (b) Different visualization methods. Hypersphere2sphere (H2S, proposed here) visualizes the 200D balls as circles, revealing the enclosure relationship, the radii of the balls, and the offset separating their centers. Principal component analysis (PCA), multidimensional scaling using metric stress (MDS), and t-distributed stochastic neighbor embedding (t-SNE) do not reveal the global geometry in this example. The matrix of pairwise point distances defines the geometry completely, but does not convey it intuitively.

sions explaining the most variance (Figure 1b). Multidimensional Scaling (MDS [1–3]) finds an embedding of the points in the visualization space that best matches the high-dimensional distances (Figure 1b).

Other techniques preserve local neighborhood relationships rather than global geometry. Sammon Mapping [4] is a variant of MDS that prioritizes getting local relationships right. Isomap [5] and Locally Linear Embedding [6] preserve distances on a manifold defined by local neighborhood relationships in the data. The t-distributed stochastic neighbor embedding (t-SNE [7], Figure 1b) and the Uniform Manifold Approximation and Projection (UMAP, [8]) attempt to find transformations that preserve stochastically defined neighborhood relationships. Despite their frequent application to labeled data, each of these techniques embeds the individual data points in visualization space without using labels.

Often we have labels for the data we wish to visualize, and we are interested in the relationships among the labeled distributions. We can render visualized points with colors signifying their class labels. Using a label-blind (unsupervised) embedding and coloring by class label has the advantage that the embedding is not biased by the labels. Label-aware (supervised) analyses can reveal complementary information. Here we introduce a new label-aware visualization method with an alternative aim: to use data and their labels to represent each class as a single visual object, directly conveying the global geometric relationships between the labeled distributions.

We develop a novel method, “hypersphere2sphere” (H2S¹, Figure 2), for labeled high-dimensional data that does not embed individual data points in the visualization space, but instead represents each distribution as a solid object in the visualization space. Each labeled distribution is first modeled as a hypersphere in the original high-dimensional data space. The set of hyperspheres and their pairwise relationships are then conveyed by a corresponding ensemble of spheres in 3D (or circles in 2D).

A hypersphere is the bounding surface of a ball in N dimensions, where a ball is defined as all points within a radius r of a central point. N -dimensional balls constitute the simplest generic model of the location and expanse of a distribution. Each hypersphere encloses all (or almost all) data points with a given label (Figure 2a). The radii, pairwise distances between centers, and pairwise overlaps of the hyperspheres are then expressed in the radii, distances between centers, and overlaps of the spheres in the 3D embedding (or the circles in a 2D one, Figure 2b). We also visualize the values of the summary statistics (radii, inter-center distances, and overlaps) using variants of Hinton diagrams [9], alongside statistical comparisons among them.

H2S is useful for representing labeled, high-dimensional data, as are common in many fields of science. To demonstrate and validate H2S, we applied it to simulated data, to human brain activity patterns elicited by visual images, and to internal representations in a neural network model. For comparison, we present equivalent analyses with PCA, MDS, and t-SNE. H2S provides a useful complement to these methods for conveying the relationships among the enclosing hyperspheres of the labeled distributions. In contrast to alternative methods, H2S is robust to imbalances in the numbers of samples between classes. It is well suited to provide a global characterization of class relationships even when the number of data points is much smaller than the dimensionality of the data.

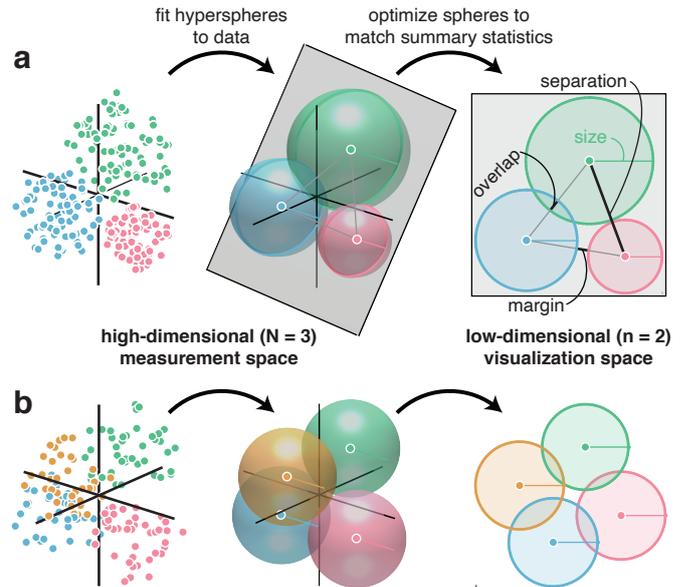


Figure 2. Hypersphere2sphere (H2S) fits hyperspheres enclosing high-dimensional labeled distributions and visualizes them as spheres. In two example cases (a, b), the true distributions (not shown) are uniform within hyperspheres in $N = 3$ dimensions (three labeled distributions in a, four in b). The data (left column) are drawn i.i.d. from these distributions and fit with hyperspheres (middle column) in the original space (3D here). The visualization (right column) consists of a circle in $n = 2$ dimensions for each distribution (a, right column). The H2S visualization is a planar slice defined by the three centers of the spheres. Each hypersphere is represented by a circle marking the intersection of the hypersphere and the visualization plane. The resulting 2D set of circles *perfectly* represents the pairwise distances between the hypersphere centers and the hypersphere radii and overlaps. In general, H2S can perfectly represent $n + 1$ classes in an n -dimensional visualization space (4 distributions in a 3D or 3 distributions in a 2D visualization), since an n -dimensional hyperplane can perfectly capture the radii and center separations for all $n + 1$ hyperspheres. (b) When there are more classes, H2S optimizes the ensemble of visualization spheres such that their summary statistics most closely match the hypersphere summary statistics: sphere radii match hypersphere radii, sphere center separations match hypersphere center separations, and sphere overlaps match hypersphere overlaps (defined for each pair of classes as the length along the line connecting the centers).

¹Available for MATLAB, Python version forthcoming.

Results

The simplest model of a high-dimensional distribution. A uniform distribution within a hypersphere is arguably the simplest probability density model for a high-dimensional distribution. A hypersphere has few parameters to estimate: a radius and N center coordinates, where N is the dimensionality of the original data space. Spheres in 3D can efficiently and intuitively convey the summary statistics: the sizes, separations, and overlaps of the hyperspheres. For a 3D visualization, the language of spheres can perfectly express the summary statistics of up to four hyperspheres (Figure 2a). Likewise for a 2D visualization, circles can perfectly express the summary statistics of up to three hyperspheres. For larger numbers of distributions, the summary statistics are expressed approximately.

There are additional motivations for using hyperspheres to represent high-dimensional distributions. When $P \leq N + 1$ for P points drawn from some distribution in N dimensions, then all data points lie on a hypersphere. In many fields, we frequently have fewer samples than dimensions. In neuroscience, for example, we may study the responses of N neurons to P stimuli, where $P < N$. In that case, it is not possible to estimate more complex density models. Even a nonisotropic Gaussian model would require a greater number of samples (or prior assumptions) for a non-singular covariance estimate. A hypersphere can still capture where and how spread out the data points are.

Even when we have many samples drawn from a distribution that is not uniform within a hypersphere, it is still useful to characterize its location and expanse with an enclosing hypersphere. Moreover, many nonuniform distributions converge to a hypersphere in high dimensions. Consider an isotropic Gaussian distribution. In low dimensions, the high density at the center and long tail of a Gaussian represent marked differences from the uniform N -ball distribution. In high dimensions, however, N -ball and Gaussian distributions converge. For both, most points fall within a small margin of a hypersphere as dimensionality grows (Supplementary Figure S2). Other density models also converge to a hypersphere in high dimensions. This is true, for example, for a uniform hypercubic distribution, despite its anisotropy. More importantly for practical applications, these distributions are well approximated by a hypersphere in N dimensions when N is high.

Hypersphere parameter estimation. The first step of H2S is to estimate the center \hat{c}_i and radius \hat{r}_i of the hypersphere that best describes the i th labeled distribution in the original high-dimensional space. A number of estimators have been proposed for the minimum enclosing ball of a set of points [10–13]. Alternative estimators for the center and radius differ in their assumptions and in their statistical and computational efficiency (Supplementary Table S1). An estimator is suitable if it is within our computational budget and its assumptions hold approximately for the data.

We implemented and validated a range of estimators. Assuming an N -ball uniform distribution model, the maximum likelihood (ML) estimate is the smallest N -ball that encloses all points. The smallest N -ball has the highest density, thus maximizing the likelihood. The parameters of the smallest enclosing N -ball can be efficiently approximated by optimization [14, 15]. Because the the ML estimate here is the smallest hypersphere possible, given the data, it will underestimate the radius of the true distribution for finite samples. The negative bias of the ML estimate of the radius can be severe when the number of samples is small relative to the number of dimensions. To infer the joint posterior over the center and radius, we implemented a Markov-Chain Monte Carlo (MCMC) estimator. This estimator is ideal, but requires much computation.

As we discussed above, the N -ball uniform and the Gaussian distributions converge in high dimensions. This motivates estimating the center as we would for a Gaussian, where the ML estimate is simply the mean of all points. The radius is then estimated as the mode of the distances of the data points from the center estimate. This approach is motivated by the fact that different distributions converge toward the surface of a hypersphere in high dimensions (Supplementary Figure S2, see Methods for details on the estimate of the mode). Our implementation of H2S can use any estimator of the hypersphere parameters, but defaults to an estimator that is based on the empirical distance-to-center distribution (Equation 13 in Methods). This estimator combines speed and accuracy across a range of simulated distributions, dimensionalities, and numbers of samples (Supplementary Figure S3).

Hypersphere summary statistics. The hypersphere summary statistics measure the size of each hypersphere and the separation and overlap for each pair of hyperspheres. The size of the i th hypersphere is defined as its radius r_i . The separation between the i th and j th hyperspheres is defined as the distance d_{ij} between their centers. The distances are estimated as follows:

$$\hat{d}_{ij} = \|\hat{c}_i - \hat{c}_j\|_2, \quad (1)$$

where \hat{c}_i and \hat{c}_j are estimates of the centers c_i and c_j , usually the means of the empirical samples for the two distributions. The margin m_{ij} between two hyperspheres is defined as the distance between the centers minus the radii. It is estimated from the distance and radii estimates:

$$\hat{m}_{ij} = \hat{d}_{ij} - \hat{r}_i - \hat{r}_j \quad (2)$$

The overlap between two hyperspheres is defined as the negative margin: $-m_{ij}$.

Note that the sizes, separations, and overlaps are one-dimensional quantities that have units of length in the original space. Size and overlap could alternatively be measured by volumes. However, units of volume depend on dimensionality

and so cannot be compared between the data and the visualization space. Measuring lengths in the original space and conveying them with lengths in the visualization space is highly interpretable.

Low-dimensional sphere embedding. We jointly optimize the low-dimensional visualized sphere (or circle) parameters to minimize the deviation between the summary statistics of the hyperspheres (data) and the summary statistics of the spheres (visualization). Specifically, the optimization procedure finds the low-dimensional center positions $\tilde{c}_1, \dots, \tilde{c}_T$ and radii $\tilde{r}_1, \dots, \tilde{r}_T$ that minimize the error E , where T is the number of classes. The error E is a function of all the distances \tilde{d}_{ij} , margins \tilde{m}_{ij} , and radii \tilde{r}_i of the visualization spheres and the corresponding summary statistics estimated from the high-dimensional data (\hat{d}_{ij} , \hat{m}_{ij} , and \hat{r}_i):

$$E(\tilde{c}_1, \dots, \tilde{c}_T, \tilde{r}_1, \dots, \tilde{r}_T) = \sum_{i=1}^{T-1} \sum_{j=i+1}^T (\tilde{d}_{ij} - \hat{d}_{ij})^2 + \sum_{i=1}^{T-1} \sum_{j=i+1}^T (\tilde{m}_{ij} - \hat{m}_{ij})^2 + \sum_{i=1}^T (\tilde{r}_i - \hat{r}_i)^2 \quad (3)$$

The error E is not convex with respect to the visualization centers and radii. We minimize it by gradient descent to find a local minimum. First, we initialize the radii of the visualization spheres to be the same as the estimated high-dimensional radii and use MDS with metric stress [1, 2] to arrange the centers. When the number of classes T is at most $n + 1$ (at most one greater than the visualization dimension $n \in \{2, 3\}$), the distances between centers are perfectly expressed. The radii and margins are then also perfectly expressed, and the error vanishes: $E = 0$. When T is greater than $n + 1$, we use gradient descent to find a local minimum of E , using an interior-point algorithm (MATLAB's *fmincon*) to optimize the radii and low-dimensional center positions (see [Methods](#)).

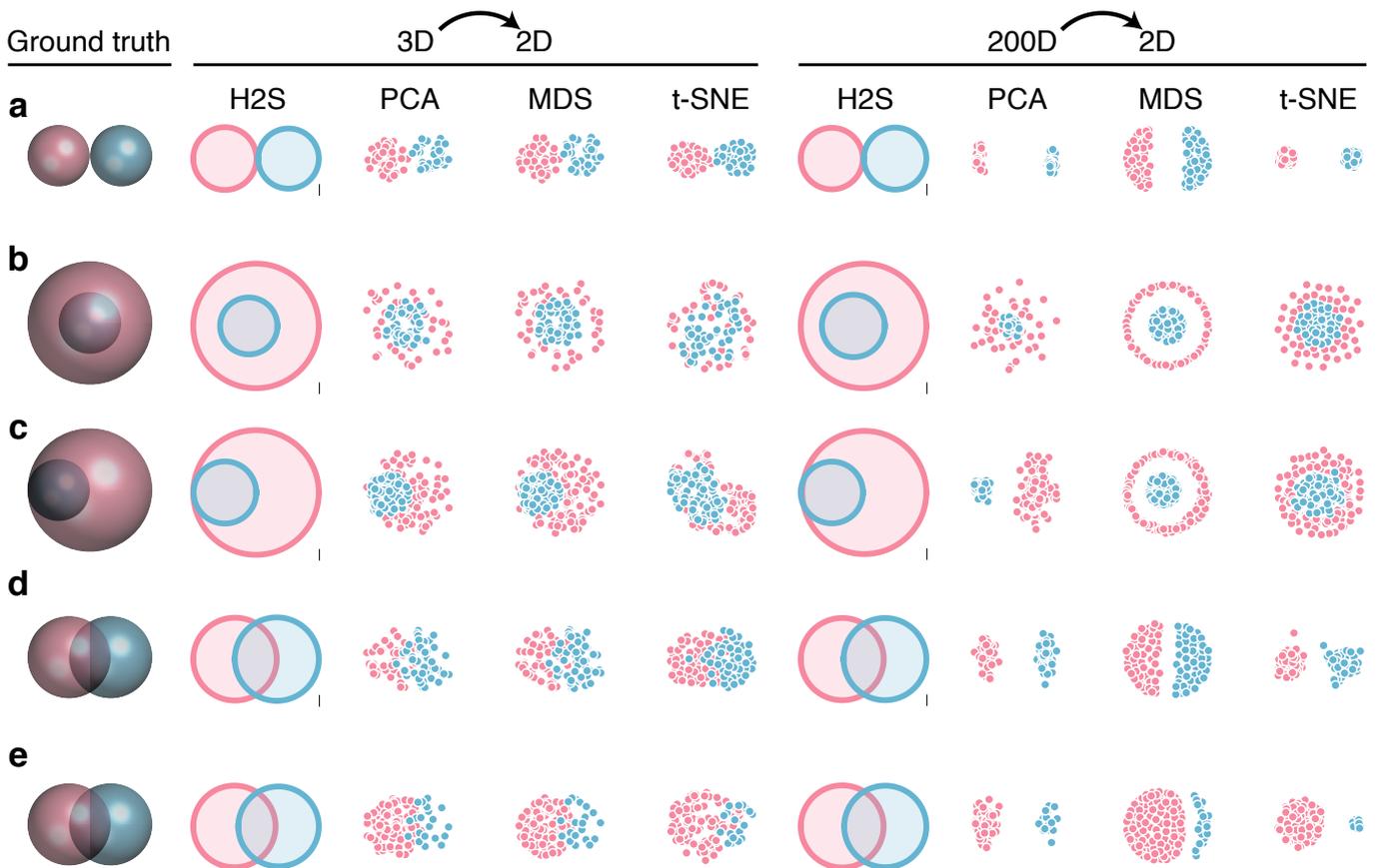


Figure 3. H2S is a more consistent visualization than PCA, MDS, and t-SNE on 3D and 200D synthetic examples. The leftmost column shows 3D renderings of uniform 3-ball distributions generating the points to the right. The next four columns from the left show 2D embeddings of the 3D distributions, and the four rightmost columns show 2D embeddings of equivalent 200-dimensional distributions. (a) Two touching equal-radius balls. (b) Two concentric balls with one having half the radius of the other. (c) One larger ball enclosing a smaller one, touching at one surface point. (d,e) Two intersecting balls with equal radii and the same number of points (d), or a different number of points (e, 100 red, 20 blue).

Validation of H2S using synthetic data. To demonstrate the fundamental properties of H2S, we used it to create 2D visualizations of simulated data in 3D and 200D where ground truth is known. We compared H2S to other visualization techniques: PCA, MDS, and t-SNE. For equal radius N -balls touching at one point (Figure 3a), all methods give similar results when $N = 3$ (left). When $N = 200$ (right), however, all but H2S exaggerate the spatial separation between the two distributions, and PCA and MDS distort the distribution shapes. For concentric balls with different radii (Figure 3b), all visualizations reveal the smaller (blue) distribution surrounded by the larger (red) one. However, the configuration of the distributions changes substantially as the dimensionality grows for all methods except H2S, which accurately represents the geometry regardless of dimensionality.

When the smaller inner ball is touching the larger ball at one point (Figure 3c), H2S reflects this in both the 3D and 200D cases, whereas the other visualizations do not provide an accurate picture of the global geometry of the two hyperspheres in 200D. Similarly, when two equal-radius hyperspheres intersect (Figure 3d), the overlap is correctly reflected, regardless of dimensionality in H2S, but the other visualizations suggest no overlap in 200D.

Finally, the embedding geometry is strongly dependent on the relative number of samples for each class in MDS and t-SNE, but H2S (Figure 3e, 100 red points, 20 blue). H2S is not biased by having different numbers of samples for each class.

Cognitive neuroscience application. We applied H2S to a neuroscientific dataset. Twenty-four human subjects viewed 48 images of objects belonging to four different categories (faces, bodies, objects, and places, Figure 4a) while their brain activity was measured with functional magnetic resonance imaging (fMRI) [16]. For each image, a detailed brain response pattern was estimated. The patterns were composed of 2mm-wide cubic measurement regions, called voxels, each reflecting the blood-oxygen-level-dependent response, which is correlated with local neural population activity. Analyses focused on six visual areas, defined as 320-voxel regions of interest (Figure 4b).

Given the large number of responses for each region (320) and the much smaller number of images (48), estimating the complex high-dimensional category-defined distributions of activity patterns is unrealistic. However, H2S can reveal the expanse of the distribution for each category in the 320-dimensional space of response patterns, as well as the relationships among the different distributions. This can reveal how distinctly the brain represents different images, within and between categories, at different stages of visual processing.

Neural networks (both biological and artificial) are thought to perform object recognition by progressively transforming representations such that the distributions of response patterns corresponding to categories become linearly separable [17]. In early visual areas, such as primary visual cortex (V1), different categories of image are “entangled” in the same territory of the multivariate response space. Although each image may elicit a unique response pattern, implying that the category distributions do not locally overlap, their global entanglement precludes a simple readout by a linear or radial-basis-function classifier [18]. In this dataset, H2S reveals the entanglement of the category distributions within the hyperspherical enclosing regions. The enclosing hyperspheres overlap significantly in early visual areas (all overlaps in V1 and V2 are significant: $P < 0.05$ for scenes vs faces and scenes vs bodies, $P < 0.001$ for all other pairs of categories). In V3, the categories are

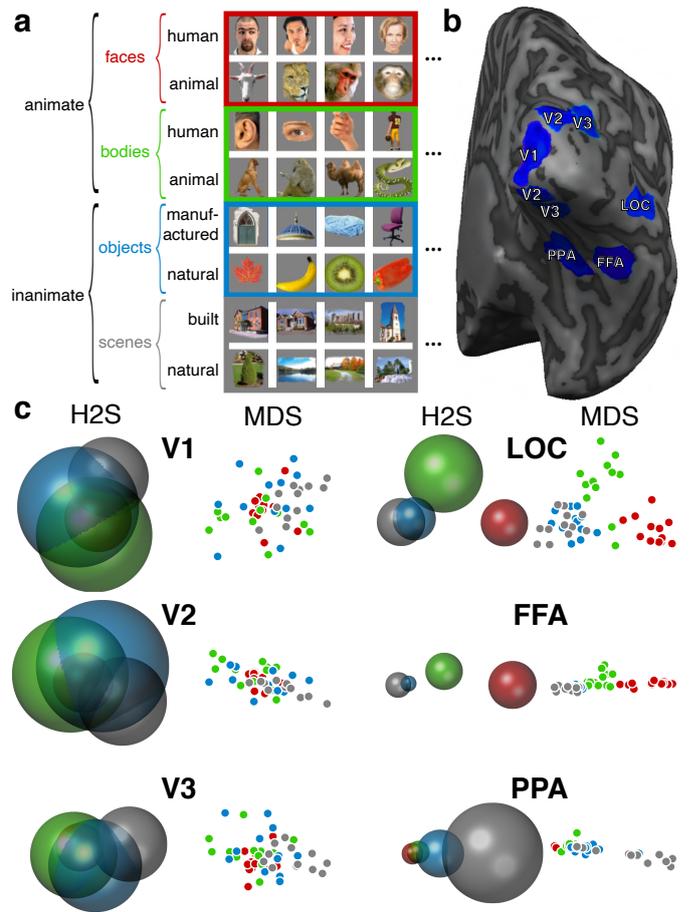


Figure 4. Brain activity pattern distributions associated with visual object categories are confined to nonoverlapping hyperspheres at later stages of visual processing.

(a) Visual stimuli of different categories were presented to human subjects during a functional brain imaging experiment [16]. (b) Inflated cortical surface reconstruction of the right hemisphere of one subject with six visual regions of interest (blue) whose representations of the images were separately analyzed. For each region, a 320-dimensional response pattern was estimated for each image. (c) H2S (left) and MDS (right) representations of faces (red), bodies (green), objects (blue), and scenes (gray), for the six visual areas in (b). At early stages (V1-3) the response patterns associated with different categories are entangled [17], falling into overlapping enclosing hyperspheres. As the visual information is re-represented at later stages of the visual processing hierarchy (LOC, FFA, PPA), each category has its own territory in response pattern space, and the enclosing hyperspheres no longer overlap.

somewhat less entangled and the overlap for scenes vs faces and scenes vs bodies are not significant (other overlaps in V3 are still significant $P < 0.001$, see [Figure 4c](#) and [Supplementary Figure S5](#)).

In the lateral occipital complex (LOC), by contrast, bodies and faces elicit response patterns falling into nonoverlapping hyperspheres, consistent with an explicit representation of object category [19]. In the fusiform face area (FFA) [20], the margins between the hyperspheres further increase, with faces being particularly well isolated from the other categories. The parahippocampal place area (PPA) responds selectively to scenes, only weakly to objects, and not at all to faces [21].

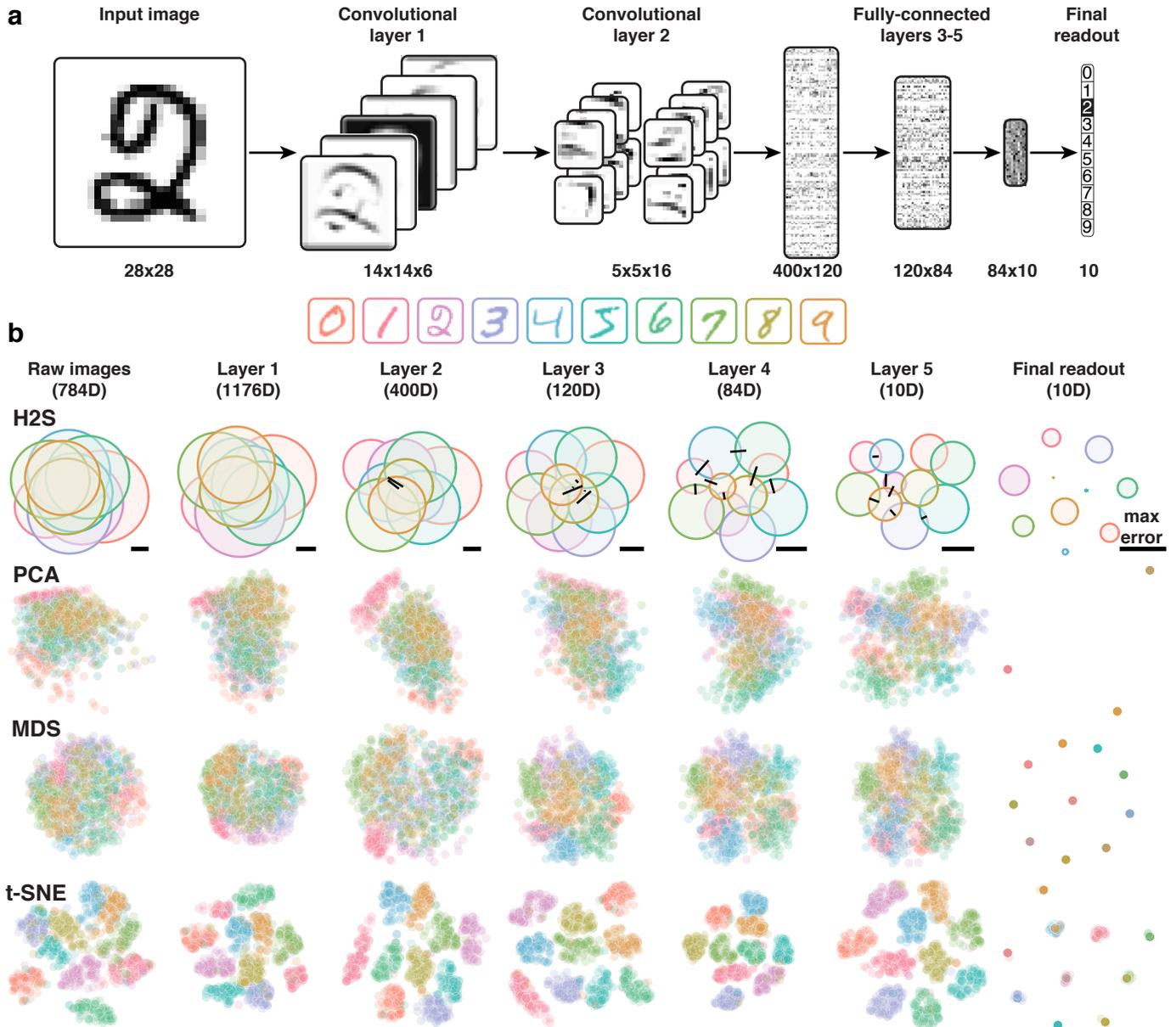


Figure 5. H2S best captures the gradual separation of handwritten digit representations in successive layers of a neural network. (a) We trained a simple variant of the LeNet-5 neural network to distinguish handwritten digits 0 through 9 in the MNIST dataset. Activations in layers 1-4 are followed by linear rectification, and layers 1 and 2 are further followed by 2x2 max-pooling (see methods for details). Pixel intensities represent activations in each layer of the neural network to the example ‘2’ input image shown on the left. Each layer’s activation pattern shown represents a single datapoint in the visualizations below. Numbers below each layer indicate the size of the activations for convolutional layers (after max-pooling), and the size of the connection matrices for fully connected layers (their vector outputs comprise the activations visualized below). (b) Hypersphere2sphere, PCA, MDS, and t-SNE visualizations of either the raw data (leftmost column), the activations of all neurons in a given layer (middle 5 columns), and the final readout of the last layer’s output (rightmost column). The black lines in the H2S visualizations that appear over the circles indicate overlaps in the visualization that have flipped sign: they are in fact statistically significant margins in the original high-dimensional data (and vice versa for lines in visualized margins). Each of these lines’ lengths indicate the size of that particular margin/overlap error. For PCA, MDS, and t-SNE, the same random selection of 1,000 datapoints out of the full 10,000-datapoint validation set are plotted.

H2S shows that PPA response patterns fall into only slightly overlapping hyperspheres. The scene hypersphere (grey) is largest, suggesting that distinctions among scenes are more prominently represented than distinctions among images in other categories ($P < 0.001$ for scenes versus objects, and $P < 0.01$ otherwise). Similarly, body images occupy the largest hypersphere (green) in the LOC, and face images do in the FFA, suggesting that each of these areas most distinctly represents images of a preferred category.

Deep neural network (DNN) application. DNNs are a leading technology in computer vision and provide the best current model of primate object recognition [22,23]. DNNs learn a large number of parameters through extensive training and acquire internal representations that are poorly understood. We used H2S to analyze how layers in a classic DNN, LeNet-5 [24], successively “disentangle” categories. The network was trained to recognize handwritten digits in the MNIST data set [25] (Figure 5a), and achieves high accuracy on this task (99.03% correct on 20% held-out data).

Since the network recognizes digits with high accuracy, all layers must contain the information of what digit is present in the image. This implies that the response pattern distributions corresponding to different digits are separable by some nonlinear classifier, and are thus minimally overlapping locally in the response space. The t-SNE visualization emphasizes nonlinear, local neighborhood relationships and accurately reveals the separate distributions in the first and all succeeding layers (Figure 5b, bottom row). All layers appear similar in the t-SNE visualization, and it is not apparent to what extent the distributions are globally entangled in the representational space of each layer. In PCA and MDS, the distributions are highly overlapping in all but the final readout layer, suggesting similar entanglement in the representations in all layers. As a result, t-SNE, PCA and MDS fail to reveal how the representation is progressively transformed across stages of processing. H2S clearly visualizes how the different digits’ territories in the model’s representational spaces become increasingly separate from each layer to the next [26].

Nonparametric inference. In addition to visualizing the size, separation, and overlap summary statistics, H2S tests these quantities and their differences for significance. The summary statistics and their differences are each evaluated using frequentist methods. They give confidence intervals and measures of significance using permutation and bootstrap methods

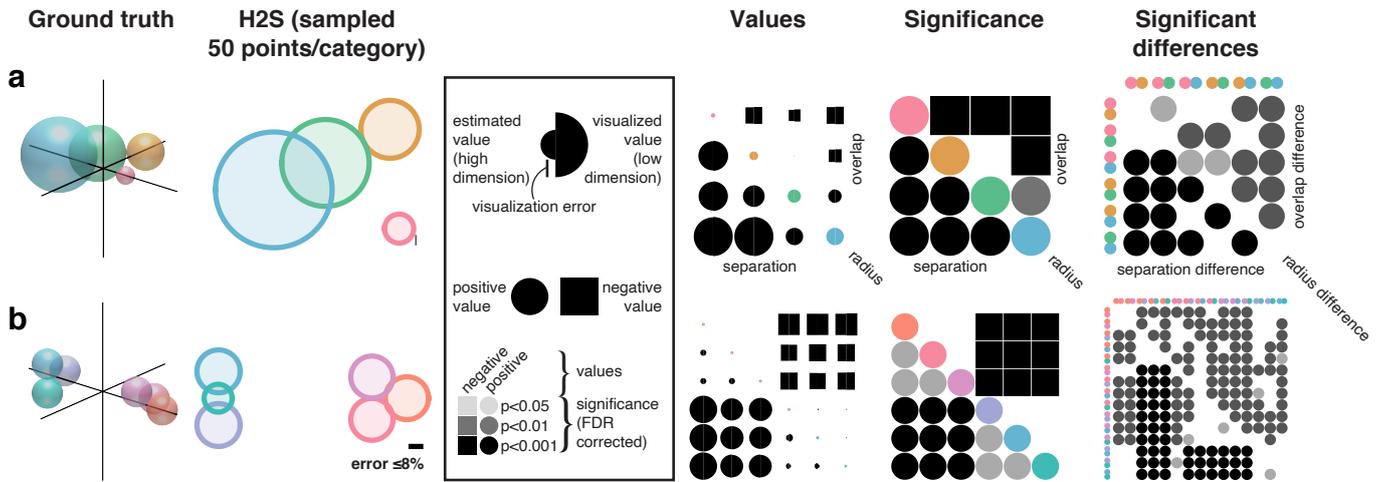


Figure 6. Visualizing values and statistical comparisons of sizes, separations, and overlaps.

Ground truth and H2S visualizations are shown to the left. On the right are the corresponding *H2S inference diagrams*, which use elaborated forms of Hinton diagrams to visualize the summary statistics, their representation in the H2S visualization (revealing distortions due to dimensionality reduction) and inferential results. For both examples (a and b), three diagrams are shown, reporting the summary statistics (left), their significance (compared to 0; middle), and the significance of their pairwise differences (right). Significant separations appear in the lower triangle, overlaps/margins as circles/squares in the upper triangle, and sizes (radii) in the diagonal. To identify a specific overlap or separation, find the off-diagonal entry that is in the same row and column as the two colored circles of interest in the diagonal. Squares indicate negative values (negative overlaps, or margins), and circles indicate positive values. In the value diagram, they are split in half, with the radius of the left half indicating the estimated value in the high-dimensional space, and the right half the visualized value. The difference in the left and right half indicates the visualization error, which in these two cases, is very low (higher visualization errors can be seen in Supplementary Figure S6). The radii of all summary statistic values are on the same scale. In the significance diagrams to the right, the darker circles indicate stronger statistical significance; the absence of a circle indicates a lack of statistical significance. (a) Four spheres, each a different size, all aligned to a single plane. Their summary statistics are perfectly captured by H2S, so there is no significant overall error to report (and thus the error bar is invisible). (b) Six equally-sized spheres, forming two groups of three equidistant spheres. In order to account for the inter-group distances, H2S introduces visualization error into some overlaps.

(see [Methods Non-parametric inference](#) for details). All significance tests use false discovery rate (FDR) correction to account for multiple comparisons [27]. Furthermore, H2S computes the significance of inter-class pairwise differences between summary statistics, answering questions such as: is the distance between classes A and B significantly greater than the distance between classes B and C?

The summary statistics and inferential results are compactly visualized in *H2S inference diagrams*, which use matrix-shaped, elaborated versions Hinton diagrams [9] (Figure 6). One $(T \times T)$ matrix conveys the values of the summary statistics and their representation in the H2S visualization, revealing any distortions due to dimensionality reduction. A corresponding $(T \times T)$ inferential matrix conveys which summary statistics deviate significantly from 0. A third matrix $((T^2 - T)/2 \times (T^2 - T)/2)$ gives inferential results for pairwise comparisons among the summary statistics with the lower triangle representing the pairwise separation comparisons, the upper triangle representing the overlap comparisons, and the diagonal representing the radius comparisons (Figure 6). Supplementary Figure S6 represents data from Figure 5 in this format.

Discussion

Visualization of high-dimensional data requires us to choose what information is to be conveyed. H2S is useful when we have class labels and are interested in the global relationships among the labeled distributions. The visualization conveys the relationships among a set of hyperspheres, each containing one of the labeled distributions. The enclosing hyperspheres capture the general regions occupied by the distributions, and are therefore often of interest even when the distributions are not uniform within the hyperspheres. Spheres in 3D (or circles in 2D) provide a natural visual language for conveying the hypersphere sizes, center separations, and overlaps, either exactly (for up to 4 classes in 3D or up to 3 classes in 2D) or approximately.

Three motivations led us to choose enclosing hyperspheres to represent data distributions. First, hyperspheres are a simple model. A hypersphere provides the simplest possible model of a high-dimensional distribution beyond a point estimate of the center of mass, adding only one more parameter: the radius. This makes the hypersphere an attractive initial model to consider to capture the location and extent of each distribution, even if the distributions are not N -ball uniform. In general, we do not want to assume N -ball distributions. Overlaps among the visualization spheres then indicate that the distributions occupy overlapping hyperspherical territories, though they may be locally separable within their shared territory.

Second, for many compact distributions, including Gaussians and hypercube uniform distributions, most points fall within a small margin of a hypersphere. An isotropic Gaussian, for example, can be characterized by its center and the radius of a hypersphere enclosing most of the distribution. The choice of the fraction of the distribution to be enclosed only slightly changes the radius in high dimensions, because the distances to center have little variance.

Third, whenever the number of samples per class is smaller than the number of dimensions, it is not realistic to characterize the high-dimensional distributions in detail. This is often the case in practice. For example, the cognitive neuroscience application (Figure 4) had only 12 examples per category and a 320-dimensional brain activity pattern space. In this scenario, it is impossible to even estimate an anisotropic Gaussian model of the density. An isotropic model, then, is the only realistic option.

We demonstrated H2S on synthetic data (Figure 3), revealing its robustness to changes of data dimensionality and to imbalances in the number of data points available for different classes. PCA, MDS, and t-SNE are all sensitive to both dimensionality and sampling imbalances because they attempt to embed the individual points, rather than to capture the relationships between the distributions (Figure 3). Applied to representations in the human brain and in a neural network model, H2S revealed how categories become successively disentangled across stages of representation, while PCA, MDS, and t-SNE did not (Figure 5). Each method provides a different perspective on the data and H2S uniquely reveals the relationships among the enclosing hyperspheres.

H2S is attractive when we are interested in the relationships among a set of distributions, of which the data points are samples. It is not appropriate when each data point has a unique meaning and we are interested in the relationships among individual data points within each class. By summarizing the individual data points in each class with a single visual primitive, H2S achieves robustness to the particular sample of points provided for each class and to imbalances in the number of samples available for each class. It reveals the global relationships among distributions, rather than idiosyncrasies of a particular data set.

H2S introduces a novel approach to visualizing data: a set of summary statistics of interest is communicated using a visual language of two- or three-dimensional shapes, each representing a labeled, high-dimensional distribution. The language of spheres can express the sufficient statistics of hyperspheres, providing a natural starting point for understanding the relationships among high-dimensional labeled distributions. Future work could explore expanding the set of summary statistics (to include, for example, covariance information). A richer set of summary statistics will require a more expressive visual language, such as ellipsoids. Once researchers define the statistics of interest, they can choose a visual language of shapes to optimize, creating a bespoke visualization tailored to the analysis question. H2S establishes this novel approach, providing a first approximation to sets of labeled high-dimensional distributions that gives us a global sense of their relationships.

References

- [1] Young, G. & Householder, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika* **3**, 19–22, DOI: [10.1007/BF02287916](https://doi.org/10.1007/BF02287916) (1938).
- [2] Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika* **17**, 401–419, DOI: [10.1007/BF02288916](https://doi.org/10.1007/BF02288916) (1952).
- [3] Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* **27**, 125–140, DOI: [10.1007/BF02289630](https://doi.org/10.1007/BF02289630) (1962).
- [4] Sammon, J. W. A Nonlinear Mapping for Data Structure Analysis. *IEEE Transactions on Computers* **C-18**, 401–409, DOI: [10.1109/T-C.1969.222678](https://doi.org/10.1109/T-C.1969.222678) (1969). [10.1111.1669](https://doi.org/10.1111.1669).
- [5] Tenenbaum, J. B., de Silva, V. & Langford, J. C. A global geometric framework for non linear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
- [6] Roweis, S. & Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**, 2323–2326, DOI: [10.1126/science.290.5500.2323](https://doi.org/10.1126/science.290.5500.2323) (2000). [10.1111.1669v3](https://doi.org/10.1111.1669v3).
- [7] van der Maaten, L. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 85, DOI: [10.1007/s10479-011-0841-3](https://doi.org/10.1007/s10479-011-0841-3) (2008).
- [8] McInnes, L. & Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* 1–18, DOI: [10.21105/joss.00861](https://doi.org/10.21105/joss.00861) (2018). [1802.03426](https://doi.org/10.1802.03426).
- [9] Hinton, G. E. & Shallice, T. Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review* **98**, 74–95, DOI: [10.1037/0033-295X.98.1.74](https://doi.org/10.1037/0033-295X.98.1.74) (1991).
- [10] Kumar, P., Mitchell, J. S. & Yildirim, E. A. Approximate minimum enclosing balls in high dimensions using core-sets. *ACM Journal of Experimental Algorithmics* **8**, 1–29, DOI: [10.1145/996546.996548](https://doi.org/10.1145/996546.996548) (2003).
- [11] Bădoiu, M. & Clarkson, K. L. Smaller core-sets for balls. *Proceedings of the Annual ACM-SIAM Symposium on Discrete Algorithms* 801–802 (2003).
- [12] Fischer, K., Gärtner, B. & Kutz, M. Fast smallest-enclosing-ball computation in high dimensions. In *Algorithms - ESA 2003, 11th Annual European Symposium*, 630–641, DOI: [10.1007/978-3-540-39658-1_57](https://doi.org/10.1007/978-3-540-39658-1_57) (Budapest, Hungary, 2003).
- [13] Ding, H. Minimum Enclosing Ball Revisited: Stability and Sub-linear Time Algorithms. *arXiv* (2019). [1904.03796](https://doi.org/10.1904.03796).
- [14] Ritter, J. An Efficient Bounding Sphere. In Glassner, A. S. (ed.) *Graphics Gems*, 301–303, DOI: [10.1016/B978-0-08-050753-8.50063-2](https://doi.org/10.1016/B978-0-08-050753-8.50063-2) (Academic Press Professional, San Diego, CA, 1990).
- [15] Larsson, T. Fast and Tight Fitting Bounding Spheres. *Proceedings of The Annual SIGRAD Conference* 27–30 (2008).
- [16] Walther, A. *Beyond brain decoding: Representational distances and geometries*. Ph.D. thesis, University of Cambridge (2015).
- [17] DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends in Cognitive Sciences* **11**, 333–341, DOI: [10.1016/j.tics.2007.06.010](https://doi.org/10.1016/j.tics.2007.06.010) (2007).
- [18] Poggio, T. & Girosi, F. A theory of networks for learning. Tech. Rep. (1989).
- [19] Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. The lateral occipital complex and its role in object recognition. *Vision Research* **41**, 1409–1422, DOI: [10.1016/S0042-6989\(01\)00073-6](https://doi.org/10.1016/S0042-6989(01)00073-6) (2001).
- [20] Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience* **17**, 4302–4311, DOI: [10.1098/Rstb.2006.1934](https://doi.org/10.1098/Rstb.2006.1934) (1997). [0208024](https://doi.org/10.1006.0208024).
- [21] Epstein, R. A. & Kanwisher, N. G. A cortical representation of the local visual environment. *Nature* **392**, 598–601, DOI: [10.1038/33402](https://doi.org/10.1038/33402) (1998).
- [22] Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences* **111**, 8619–8624, DOI: [10.1073/pnas.1403112111](https://doi.org/10.1073/pnas.1403112111) (2014). [0706.1062v1](https://doi.org/10.1006.0706.1062v1).
- [23] Kriegeskorte, N. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science* **1**, 417–446 (2015).

- [24] LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, vol. 86, 2278 – 2324, DOI: [10.1109/5.726791](https://doi.org/10.1109/5.726791) (1998).
- [25] LeCun, Y. & Cortes, C. The MNIST database of handwritten digits (1998).
- [26] Chung, S., Lee, D. D. & Sompolinsky, H. Linear readout of object manifolds. *Physical Review E* **93**, 1–5, DOI: [10.1103/PhysRevE.93.060301](https://doi.org/10.1103/PhysRevE.93.060301) (2016). [1512.01834](https://arxiv.org/abs/1512.01834).
- [27] Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300 (1995).
- [28] Kriegeskorte, N., Formisano, E., Sorger, B. & Goebel, R. Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20600–20605, DOI: [10.1073/pnas.0705654104](https://doi.org/10.1073/pnas.0705654104) (2007).
- [29] Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology* **10**, DOI: [10.1371/journal.pcbi.1003553](https://doi.org/10.1371/journal.pcbi.1003553) (2014). [1606.05201](https://arxiv.org/abs/1606.05201).
- [30] Walther, A. *et al.* Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* **137**, 188–200, DOI: [10.1016/j.neuroimage.2015.12.012](https://doi.org/10.1016/j.neuroimage.2015.12.012) (2016).
- [31] Efron, B. Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**, 171–185, DOI: [10.1080/01621459.1987.10478410](https://doi.org/10.1080/01621459.1987.10478410) (1987).

Author contributions statement

NK and ADZ conceived and designed the method. ADZ conducted the neural network simulations. AW conducted and ADZ and AW analyzed the functional MRI experiment. ADZ and ASP created software. ADZ and NK wrote the manuscript.

Additional information

The authors declare no competing interests.

Methods

Hypersphere parameter estimation from high-dimensional data. Estimation of the parameters of minimum enclosing balls, the centers and radii, is a well studied problem [10–13]. Any of the estimators offered in the literature can be used in the context of H2S. Here we implemented a maximum-likelihood (ML) estimator, a Bayesian Markov-Chain-Monte-Carlo (MCMC) estimator, and several heuristic estimators that use summary statistics of the distribution of the distances from individual data points of a class to the center (estimated as the mean) and of the distribution of pairwise distances. All estimators are described below, along with an empirical comparison on simulated data. In brief, theoretical considerations and simulations showed the following: The ML estimator overfits and underestimates the radius. The Bayesian MCMC estimator is computationally intensive and not optimally stable in practice. This motivates consideration of the heuristic estimators. The mean provides a good estimate of the center and a good basis for estimating the radius from the distribution of distances between the individual points and the center estimate. Simulations with data generated from Gaussian, uniform N -ball, and uniform hypercubic distributions suggest the use of the adaptive radius estimator \hat{r}_{adapt} (Equation 13), which works robustly across distributions, dimensionalities, and numbers of samples, and is fast to compute Supplementary Figure S3.

Maximum-likelihood estimator. A first estimator is the maximum-likelihood (ML) estimate of a uniform N -ball distribution, from which we assume the data are drawn. This estimator finds the smallest (i.e. highest density) ball that contains all data points. The parameters are estimated by minimizing the maximum distance to center across all points. The minimum hypersphere’s center $\hat{\mathbf{c}}_{ML}$ and radius \hat{r}_{ML} are:

$$\hat{\mathbf{c}}_{ML} = \arg \min_{\mathbf{c}} \max_i \|\mathbf{X}_i - \mathbf{c}\|_2 \quad (4)$$

$$\hat{r}_{ML} = \max_i \|\mathbf{X}_i - \hat{\mathbf{c}}_{ML}\|_2 \quad (5)$$

Finding the smallest hypersphere requires iterative optimization. The ML estimator finds the smallest possible hypersphere and is therefore certain to underestimate the actual radius of the data-generating N -ball distribution.

Bayesian (MCMC) estimator. The shortcomings of the ML estimator motivate a Bayesian approach, where we infer the full posterior over N -ball parameters given the data. We can use Markov Chain Monte Carlo (MCMC) sampling to estimate the joint posterior for center and radius, using infinite, flat (improper) priors for both parameters. The log likelihood is

$$L(\mathbf{X}|\hat{\mathbf{c}}, \hat{r}) = \begin{cases} -\infty & \text{if any } \hat{\delta}_i > \hat{r}, \\ -P \log(\hat{r}^N V_1(N)) & \text{otherwise} \end{cases} \quad (6)$$

where \mathbf{X} is the P -by- N matrix of data points (where P is the number of points), $\hat{\mathbf{c}}$ (N -vector) is the center estimate and \hat{r} (scalar) is the radius estimate. $V_1(N)$ is the volume of a unit-radius N -ball. The distance of each point to center, $\hat{\delta}_i$, is given by:

$$\hat{\delta}_i = \|\mathbf{X}_i - \hat{\mathbf{c}}\|_2 \quad (7)$$

The joint posterior over center and radius is computed by MCMC sampling (Supplementary Figure S1). As a point estimates of the radii and inter-center distances for the H2S visualization, we use the medians of the marginal posteriors of these parameters. The marginal medians have an equal probability of being larger or smaller than the true parameters of the data-generating N -balls.

Distance-to-center-based estimators. We motivated our choice of the hypersphere as a useful representation of high-dimensional distributions because, for a Gaussian or uniform N -ball distribution in high dimensions, most points fall within a small margin of a hypersphere. It is instructive to examine the distribution of the distances of the points to the center (mean) for different high-dimensional distributions. For a uniform N -ball, the distance-to-center (D2C) distribution is uniform for 1 dimension (where the ball reduces to an interval), and tends to a delta function as dimensionality increases (Supplementary Figure S2, left). For a uniform N -ball in a finite number of dimensions greater than 1, the D2C distribution has a single (leftward) tail and the mode is the maximum distance to center, which equals the radius of the ball. Similarly, for a Gaussian or hypercubic uniform distribution, the D2C distribution tends toward a delta function as dimensionality increases. However, for these two distributions, the D2C distribution has two tails (Supplementary Figure S2, center and right).

For the Gaussian and hypercubic uniform distributions, the D2C distributions have modes that grow monotonically with dimensionality. For the Gaussian, the support of the D2C distribution extends to positive infinity. The enclosing hypersphere would thus be infinite and uninterpretable. Since the Gaussian and uniform hypercubic D2C distributions both still tend toward a delta function, using the mode of the D2C distribution as the radius parameter to visualize in H2S appears to be the best choice: in finite dimensions, most D2C fall within a small margin of the mode. As dimensionality grows, the mode converges to the radius of the hypersphere that all data fall onto. The mode of the D2C distribution, thus, appears a good target for our radius estimator, whether the underlying distribution in a uniform ball, an isotropic Gaussian, or even a non-isotropic distribution such as a uniform hypercube.

The question then is how to estimate the mode of the D2C distribution. For the high-dimensional ($N > 100$) case, the D2C distribution is very concentrated, so the mean or the median of the empirical D2C distribution could be used as an estimate of the mode, where the center is estimated as the mean of the data points. This, however, leads to negatively biased estimates of the radius parameter (Supplementary Figure S3), because using the centroid overfits the data and minimizes the sum of the squared D2C. Moreover, we would like our estimator to work even for lower-dimensional data, where the mean or the median of the empirical D2C distribution is not a good estimate of the mode of the population D2C distribution. Since the mode of the D2C distribution depends on the underlying distribution, we developed radius estimators that assume the distributions are of a particular shape. These estimators use summary statistics of the empirical D2C distribution and provide fast and accurate radius estimates (Supplementary Figure S3).

Under the uniform N -ball assumption, the mode of D2C distribution is the maximum distance to center (Supplementary Figure S2). The D2C-based radius estimate for a uniform N -ball is

$$\hat{r}_{DCB1} = (1 + P^{-N}) \max_{i \in \{1..P\}} \hat{\delta}_i \quad (8)$$

where $\max_i \hat{\delta}_i$ is the empirical maximum distance to the mean of the points and the factor preceding it corrects the estimate upward to account for the fact that the maximum empirical distance to centroid will be smaller than the radius.

For a unit-variance Gaussian distribution, the D2C distribution is the chi distribution, and its mean $\gamma(N)$ is a function of the dimension N :

$$\gamma(N) = \sqrt{2} \frac{\Gamma((N+1)/2)}{\Gamma(N/2)} \quad (9)$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$ is the Gamma function. The D2C-based radius estimator for a Gaussian scales the standard deviation of the point coordinates to this mean D2C defined by $\gamma(N)$:

$$\hat{r}_{DCG} = \gamma(N) \sqrt{\frac{1}{N(P-1)} \sum_{i=1}^P \hat{\delta}_i^2}. \quad (10)$$

For a uniform hypercubic distribution, the expected mean distance from the center is

$$\int_{[-\frac{1}{2}, \frac{1}{2}]^N} \|\mathbf{x}\|_2 d\mathbf{x}, \quad (11)$$

which quickly converges to $\sqrt{N/12}$ with increasing N . We could use an estimate of the width w of the hypercubic distribution and estimate the radius as $w\sqrt{N/12}$. However, we instead define the D2C-based estimator for hypercubic distributions, \hat{r}_{DCC} , as the median of the empirical D2C distribution, because simulations show that the median is more stable.

Adaptive radius estimator. The D2C-based estimators require the user to specify an assumed shape of the distribution. Since the purpose of dimensionality-reducing visualization is often exploratory, it is desirable to have an option that works robustly and does not require the user to assume any particular distribution. Based on empirical analyses of the D2C distributions for various multivariate distributions, we developed a fast, general-purpose radius estimator that uses the variance of the empirical D2C distribution to decide whether the distribution is more like a Gaussian (with infinite support and a wider D2C distribution), or more like a ball or hypercube (with finite support and a narrower D2C distribution). If the empirical D2C distribution has high variance, the radius is estimated as the \hat{r}_{DCG} (aiming to estimate the mean of the D2C distribution for a Gaussian). If the empirical D2C distribution has low variance, then the median of the empirical D2C distribution is used as a basis and corrected upward by a fraction $\xi(N)$ of the standard deviation of the D2C distribution:

$$\hat{r}_{DCB2} = \text{median}(\hat{\delta}_i) + \text{std}(\hat{\delta}_i)\xi(N) \quad (12)$$

The scaling factor $\xi(N)$ is based on simulations where data are generated from uniform N -ball distributions:

N	2	4	8	16	32	64	128	256	512	1024	2048	4096
$\xi(N)$	1.2733	1.0115	0.8796	0.8107	0.8384	0.8638	0.9579	1.0403	1.1938	1.4268	1.8384	2.4485

Table 1. Simulation-derived values of $\xi(N)$. For other N , $\xi(N)$ is computed via linear interpolation.

The adaptive radius estimator (\hat{r}_{adapt}) automatically selects whether to use \hat{r}_{DCG} or \hat{r}_{DCB2} on the basis of the observation that the variance of D2C for both distributions follow linearly separable power laws (Supplementary Figure S4):

$$\hat{r}_{adapt} = \begin{cases} \hat{r}_{DCG} & \text{if } \text{var} \frac{\hat{\delta}_i}{\text{median}_i(\hat{\delta}_i)} > 2^{-1-\frac{4}{3} \log_2 N} \quad (\text{Gaussian assumed}) \\ \hat{r}_{DCB2} & \text{otherwise} \end{cases} \quad (13)$$

The D2C variance criterion in Equation 13 distinguishes between Gaussian and N -ball data-generating distributions with maximum margin. This criterion also works well for hypercubic uniform distributions.

Pairwise-distance-based estimator. We can also estimate the radius of a ball distribution on the basis of an estimate of the expected distance between randomly drawn points. For a given dimensionality N , the expected inter-point distance $E[d_{i,j}]$ is proportional to the radius: $r = \zeta(N) \cdot E[d_{i,j}]$, where $\zeta(N)$ converges to $1/\sqrt{2}$ for large N . The expected-distance estimator \hat{r}_{dist} uses the mean empirical inter-point distance as a basis for estimating the expected inter-point distance of the data-generating uniform N -ball distribution. It uses the proportionality constant $\zeta(N)$ to adjust the empirical mean inter-point distance, so as to estimate the radius:

$$\hat{r}_{dist} = \zeta(N) \frac{2}{P(P-1)} \sum_{i=1}^{P-1} \sum_{j=i+1}^P \|\mathbf{X}_i - \mathbf{X}_j\|_2 \quad (14)$$

This estimator performs well and is quite robust to changes of the shape of the distribution (Supplementary Figure S3). It is useful in particular when we are only given the pairwise distances among samples, and not the point locations.

N	1	2	4	8	16	32	64	128	256	512	1024	2048	4096	4097+
$1/\zeta(N)$	0.6673	0.9039	1.1043	1.2407	1.3230	1.3657	1.3898	1.4020	1.4081	1.4111	1.4127	1.4134	1.4138	$\sqrt{2}$

Table 2. Experimentally-derived values of $1/\zeta(N)$. For other N , $\zeta(N)$ is computed via linear interpolation.

Comparison of the estimators on simulated data. We compared the performance and speed of the different estimators on synthetic data—including uniform N -balls and hypercubes, and Gaussians—as a function of dimensions and number of samples (Supplementary Figure S3). In practice, MCMC performs well for uniform N -ball distributions, and increasing the number of MCMC samples only significantly improves performance when it is significantly higher than the data dimensionality.

For both the uniform and Gaussian distributions, however, the respective D2C-based estimators (\hat{r}_{DCB} , \hat{r}_{DCG} , or \hat{r}_{DCC} , depending on the data-generating distribution; orange in Supplementary Figure S3) perform quite well, and are faster to compute. The D2C-based estimators are very accurate and fast to compute when they match the underlying distributions, but fail badly when there is a mismatch (not shown). The adaptive D2C-based estimator \hat{r}_{adapt} (Equation 13) is fast to compute, does not require user input (of distributional assumptions), and performs robustly across the three data-generating distributions. It is therefore the default estimator in our current H2S implementation.

Low-dimensional sphere embedding. For the second step in the H2S algorithm, the radii of the visualization spheres are always initialized to the estimated high-dimensional radii. We initialize the centers using MDS with metric stress [1, 2] as the optimization criterion. For up to 4 categories visualized as spheres in 3D (or up to 3 visualized as circles in 2D), the summary statistics will be perfectly expressed at this point. Otherwise, the low-dimensional center positions and the radii are further adjusted to minimize the sum of squared errors of the visualized distances \tilde{d}_{ij} , margins \tilde{m}_{ij} , and radii \tilde{r}_i (i.e. their deviations from their high-dimensional target estimates \hat{d}_{ij} , \hat{m}_{ij} , and \hat{r}_i):

$$E_{\alpha,\beta} = \sum_{i=1}^{T-1} \sum_{j=i+1}^T (\tilde{d}_{ij} - \hat{d}_{ij})^2 + \alpha \sum_{i=1}^{T-1} \sum_{j=i+1}^T (\tilde{m}_{ij} - \hat{m}_{ij})^2 + \beta \sum_{i=1}^T (\tilde{r}_i - \hat{r}_i)^2 \quad (15)$$

where the distances are a function of the centers, and the margins are a function of the distances and radii, as previously defined in Equation 1 and Equation 2.

Equation 15 represents a generalized version of Equation 3 in which the user can choose hyperparameters α and β to place more weight on representing the margins and radii accurately, relative to the distances between centers. By default, each of these hyperparameters is set to 1. One could for example, set $\beta = 10$, increasing the relative accuracy of the visualized radii, at the expense of the distances and margins. It is important to note that in the limit where $\beta \rightarrow \infty$, $\tilde{\mathbf{r}} \rightarrow \hat{\mathbf{r}}$, which makes $(\tilde{\mathbf{m}} - \hat{\mathbf{m}}) \rightarrow (\tilde{\mathbf{d}} - \hat{\mathbf{d}})$, so the problem exactly reduces to finding the center parameters which minimize the squared error of the distance. This is equivalent to embedding the centers using MDS with metric stress. Since the MDS computation is simple and fast, it serves as the initialization for optimizing Equation 15. For this same reason, this MDS-computed initialization is presented to the user as an option to be used as the final embedding configuration, with the radius values directly copied from the high-dimensional space. This option offers perfect radius accuracy, the best possible distance accuracy, and much faster computation, at the expense of lower margin accuracy.

When optimizing the full objective, we use MATLAB's *fminunc* function for optimization, with the following analytic gradients, where $\tilde{\mathbf{C}} = [\tilde{\mathbf{c}}_1 \dots \tilde{\mathbf{c}}_m]$ is the matrix concatenation of the low-dimensional center coordinate estimates and $\mathcal{T} \setminus i = \{1, \dots, T\} \setminus i$ is the set of integers from 1 to T , excluding i :

$$\frac{\partial E}{\partial \tilde{r}_i} = 2\gamma (\tilde{r}_i - \hat{r}_i) + 2\beta \sum_{j \in \mathcal{T} \setminus i} (\tilde{m}_{ij} - \hat{m}_{ij}) \quad (16)$$

$$\frac{\partial E}{\partial \tilde{\mathbf{C}}} = \frac{\partial E}{\partial \tilde{\mathbf{d}}} \cdot \frac{\partial \tilde{\mathbf{d}}}{\partial \tilde{\mathbf{C}}} \quad (17)$$

$$\frac{\partial E}{\partial \tilde{\mathbf{d}}} = 2\alpha (\tilde{\mathbf{d}} - \hat{\mathbf{d}}) + 2\beta (\tilde{\mathbf{m}} - \hat{\mathbf{m}}) \quad (18)$$

$$\frac{\partial \tilde{d}_{ij}}{\partial \tilde{\mathbf{c}}_i} = \sum_{j \in \mathcal{T} \setminus i} \frac{\tilde{\mathbf{c}}_i - \tilde{\mathbf{c}}_j}{\tilde{d}_{ij}} \quad (19)$$

Human imaging experiment. In an fMRI experiment, 24 human subjects were presented with 48 images from visual categories such as animate/inanimate, face/body, and animal/human face (Figure 4a; see [16] for experimental details). Regions of interest comprising 320 voxels were then defined for six visual areas (Figure 4b) using separate localizer experiments and retinotopic mapping stimuli. For each pair of stimuli, the crossnobis distance estimator [28–30] provided an unbiased estimate of the representational distance. The crossnobis distance matrix can then be visualized using H2S or classical MDS (Figure 4c, left and right respectively, in each column). Supplementary Figure S5 shows these same H2S visualizations with their accompanying inference results visualized in H2S inference diagrams.

Deep neural network analysis. We trained a simplified variant of LeNet-5 [24] on the MNIST handwritten digit dataset [25]. The 28x28 MNIST images were padded to be 32x32, and served as input to the network. The first two layers are convolutional with a stride of 1, with learned 5x5 filter kernels. The first layer has 6 output channels, the second layer takes a 16x16x6 input and outputs an 8x8x16 response. Both convolutional layers are followed by 2x2 max-pooling. The following three layers are fully connected: 400x120, 120x84, and 84x10. The final predicted class label is the argmax of the 10D feature vector. The network was trained with backpropagation using stochastic gradient descent (with momentum = 0.9 and learning rate = 0.001) using cross-entropy loss. Training ran for 20 epochs of the full training set (80% of the entire dataset) using minibatches

of size 30. Classifier accuracy on the validation set (20% of data) was 99.03%. [Figure 5](#) shows H2S visualizations of the layer-by-layer activations of this network, alongside PCA, MDS, and t-SNE. [Supplementary Figure S6](#) shows these same H2S visualizations with their accompanying inference results visualized in H2S inference diagrams.

Non-parametric inference. Statistics of interest optimized in the H2S objective can be optionally assessed for statistical significance via resampling methods. Significance can be first-order (is overlap A significant?) and second-order (is overlap A significantly greater than overlap B?). Nonzero radii are, by definition, always significant on a first-order basis.

We strove to use permutation tests where possible. However, some tests required traditional or accelerated bootstrap confidence intervals (BCa) [31]. All tests, by default, use 5,000 resamplings (except for the BCa test which requires jackknife resampling, adding an additional P resamplings).

Separation significance. For each pair of classes, this one-sided test asks whether the separation between the centers of the enclosing hyperspheres is significantly larger than zero. When a distance (which is by definition nonnegative) is used to estimate the true distance from noise data, the estimate is positively biased. We therefore use an unbiased cross-validated distance estimator: the square root of the inner product of the center-difference vectors in a random split of the data [30]. For a given pair of classes, the null hypothesis is that the two hyperspheres have the same center ([Supplementary Figure S7a](#)). Two example samplings are shown in ([Supplementary Figure S7b-c](#)). The cross-validated separation estimator is symmetrically distributed about 0 under the null hypothesis. We randomly permute the labels of the points between the two classes 5,000 times, re-estimate the hyperspheres, and recompute their separation. Using these permuted separations, the separation is deemed significant if it falls in the top 5% of the permutation distribution.

We empirically validated this test for different radii and sample sizes by means of data simulated under the null hypothesis. For a given dimensionality, number of samples per class, and ratio of the radii of the two hyperspheres, we simulated data under the null hypothesis 1,000 times. The separation estimates from the full (non-permuted) data and their confidence intervals are shown as points and lines in [Supplementary Figure S7d](#), respectively. Gray and black points and lines indicate non-significant and significant values — a value computed on the full (non-permuted) data is significant when it exceeds 95% of the permuted separations.

To verify that the permutation test distributions accurately capture the separation distribution, we compute histograms from the 1,000 full and all 5,000,000 permuted data point estimates in [Supplementary Figure S7e](#) (gray and black, respectively) and see that they tightly overlap. While summarizing all permutation distributions in this way is diagnostic, a further test would be to compare each permutation distribution's variance to the variance of the full data estimate distribution. [Supplementary Figure S7f](#) shows that the full data estimate variance is well within the orderly distribution of permutation distribution variances for each simulation. [Supplementary Figure S7g](#) summarizes the variance for permuted and full data estimate distributions across all simulations for parameter choices (lighter shades of red indicate increasing dimensionality — see legend in [Supplementary Figure S7h](#)). Variance is very well captured, regardless of parameter choice and its resultant scale.

To examine the effect of radius ratio, dimensionality, and number of samples per class, mean estimates across all simulations are shown in [Supplementary Figure S7h, i, and j](#), respectively. Unsurprisingly, estimator bias decreases for higher numbers of samples and lower numbers of dimensions. [Supplementary Figure S7k, l, and m](#) show the false positive rate for the permutation tests for all simulation parameter settings. Regardless of parameter choice, the false positive rate is approximately 5%, indicating that the test well captures the statistics of the data.

Overlap significance. For each class, this two-sided test asks whether an overlap is significantly positive or negative (negative overlaps are the same as positive margins). For a given pair of classes, the null hypothesis is that two hyperspheres have zero overlap ([Supplementary Figure S8a](#)). We found while developing this test, using a permutation-based test yielded near-zero false positive rates and a classic bootstrap-based test yielded very high false positive rates, making the test overly conservative or permissive, respectively. This appeared to be due to skewness in the distribution of resampled overlaps. For a small but significant cost in computation, the accelerated bootstrap confidence interval (BCa) yields a more accurate confidence interval, as it is designed to be more robust to skewed distributions and outliers [31]. For each of these 1,000 simulations, we compute bootstrap and jackknife samples of the points within each class 5,000 and P times, respectively, re-estimate the hyperspheres, and recompute their overlap. Using these bootstrapped and jackknifed overlaps, we compute 95% BCa's. The overlap estimates from the full (non-resampled) data and their confidence intervals are shown as points and lines in [Supplementary Figure S8d](#), respectively. Gray and black points and lines indicate non-significant and significant values — a value computed on the full (non-permuted) data is significant when zero falls outside the 95% BCa.

To verify that the BCa test distributions accurately capture the separation distribution, we compute histograms from the 1,000 full and all 5,000,000 bootstrapped data point estimates in [Supplementary Figure S8e](#) (gray and black, respectively) and see that they tightly overlap. Bootstrap distributions from each simulation are centered before being combined into the single distribution shown. [Supplementary Figure S8f](#) shows that the full data estimate variance is well within the orderly distribution of bootstrap distribution variances for each simulation. [Supplementary Figure S8g](#) shows that variance is again well captured, regardless of parameter choice and scale.

Mean estimates show some bias [Supplementary Figure S8h-j](#), due to a tendency for the radius estimator to slightly under-

estimate the true radius value. Estimator bias decreases for higher dimensionality and samples in the 64-256 range. False positive rates are generally reasonably close to 5% (Supplementary Figure S8k-m), with the exception at large numbers of samples (Supplementary Figure S8m). This test should only be used with caution for sample rates above 1,000 points per class.

Radius difference significance. Second-order comparisons are all concerned with whether a difference in two statistics of interest is significantly nonzero. For each class, this two-sided test asks whether one radius is significantly larger or smaller than the other (or equivalently, if the difference between two radii is significantly positive or negative). For a given pair of classes, the null hypothesis is that two hyperspheres have exactly the same radius (Supplementary Figure S9a). For the same reasons as the overlap test, radius differences use BCa. Using 5,000 bootstrapped and P jackknifed radius estimates, we compute the 95% BCa's of their differences. A radius difference is significant when zero falls outside the 95% BCa.

Again, the 1,000 full and all 5,000,000 bootstrapped data point estimates tightly correspond (Supplementary Figure S9e). Bootstrap distributions from each simulation are centered before being combined into the single distribution shown. The distribution of bootstrap distribution variances is clearly non-normal (Supplementary Figure S9f), but still the full data estimate variance is less well-centered but still well within the distribution. Bootstrap variance is slightly overestimated (Supplementary Figure S9g), as was seen in Supplementary Figure S9f. The effect is consistent across parameter choices, indicating a consistent behavior in the radius estimator.

Mean radius difference estimates show very little bias regardless of parameter choice (Supplementary Figure S9h-i), indicating that the bias in the radius estimator is symmetric, and thus cancels out for the purpose of this test. Indeed, false positive rates are generally reasonably close to 5% (Supplementary Figure S9j-k) or are slightly higher.

Separation- and overlap-difference significance. Separation-difference and overlap-difference tests operate in the same manner; they are two-sided bootstrap tests. These tests assess whether two overlaps (or two separations) are significantly different. For a given triplet of classes, the null hypothesis is that two hyperspheres have exactly the same separation (Supplementary Figure S10a) or overlap (Supplementary Figure S11a). In these scenarios permutation testing is not applicable, so these tests use bootstrap confidence intervals. Using 5,000 bootstrapped separation/overlap estimates, we compute the 95% bootstrap confidence intervals of their differences. A separation/overlap difference is significant when zero falls outside the 95% confidence interval.

Again, the 1,000 full and all 5,000,000 bootstrapped data point estimates coincide well (Supplementary Figure S10e and Supplementary Figure S11e). Bootstrap distributions from each simulation are centered before being combined into the single distribution shown. Supplementary Figure S10f and Supplementary Figure S11f show that the full data estimate variance is well within the orderly distribution of bootstrap distribution variances for each simulation. Bootstrap variance is again slightly overestimated (Supplementary Figure S10g and Supplementary Figure S11g), as was seen in Supplementary Figure S9g. The effect is consistent across parameter choices, indicating a consistent behavior.

Mean separation and overlap difference estimates show very little bias regardless of parameter choice (Supplementary Figure S10h-j and Supplementary Figure S11(h-j)). Separation difference false positive rates are generally reasonably close to 5% (Supplementary Figure S10k-m), but increase with increasing radius ratio. Overlap difference false positive rates are generally reasonably close to 5% (Supplementary Figure S11k-m), or are slightly higher for low sample numbers.

Supplementary material

Radius estimator	Meaning of radius	Input format	Distribution support	Low-N Accuracy	High-N Accuracy	Speed
Maximum Likelihood	Expected Ball radius	Points	Ball only	High	Very high	Very slow
Markov Chain Monte Carlo (MCMC)	Expected Ball radius	Points	Ball only	Very high	Very high	Slow
Mean distance-to-center (D2C)	Expected radius	Points	Any	Very low	Very low	Very fast
D2C-based estimators	Expected radius of Ball/ Gaussian/Hypercube	Points	Ball/Gaussian/ Hypercube only	High	High	Fast
Adaptive D2C	Mean of measured center distances	Points	Any	High	High	Fast
Pairwise distance-based empirical	Mean of measured pairwise point distances	Distances	Any	Medium	High	Fast

Table S1. Radius estimators compared. The MCMC and maximum likelihood estimators jointly estimate the center and radius, while the other estimators estimate the radius directly, conditional on the maximum likelihood estimate of the center.

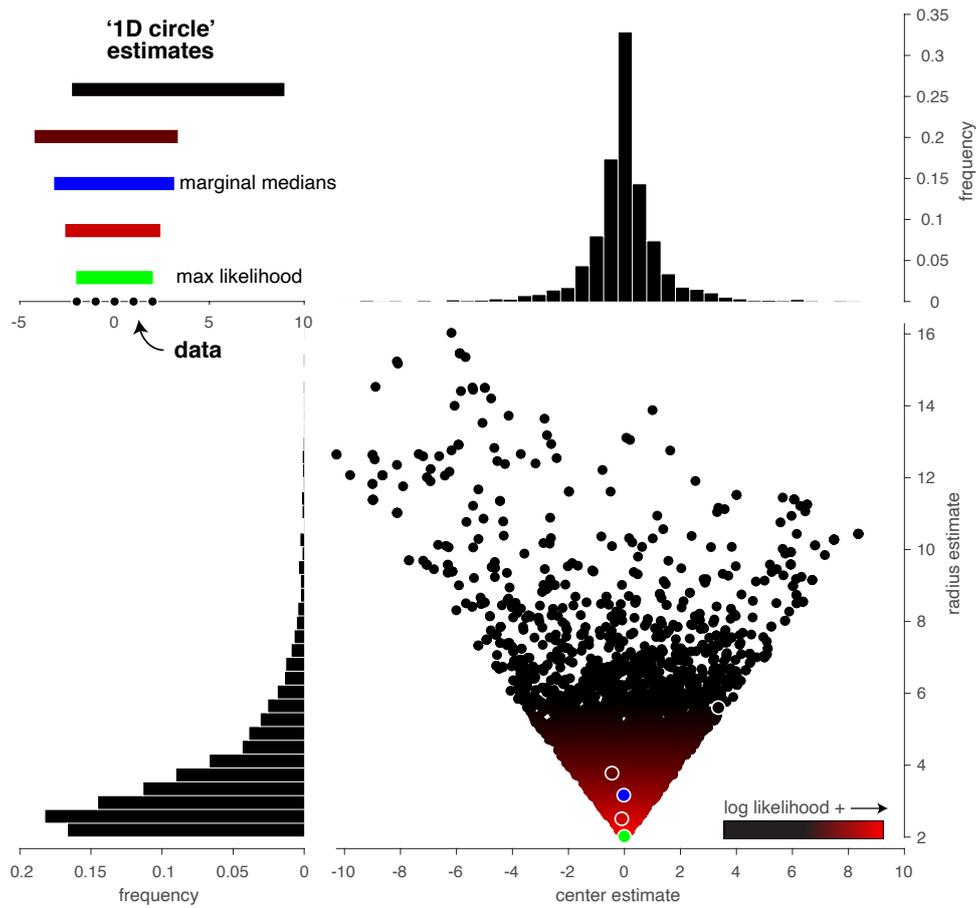


Figure S1. H2S radius estimator performance.

A toy example: “1D circle,” or a closed interval on a line (top left). Black points indicate data points and line segments indicate estimated model 1D circles, colored by their likelihood. (bottom right) An example MCMC simulation with 10,000 samples (closed circles), as a function of the estimated center and radius of the interval, going from black (low log likelihood) to red (high likelihood). The open white circles indicate the sample estimates shown in top left plot. The maximum likelihood solution is shown in green. If an infinitely flat (improper) prior is assumed, the final estimate, shown in blue, is given by the median of the marginal probabilities, shown as frequency histograms.

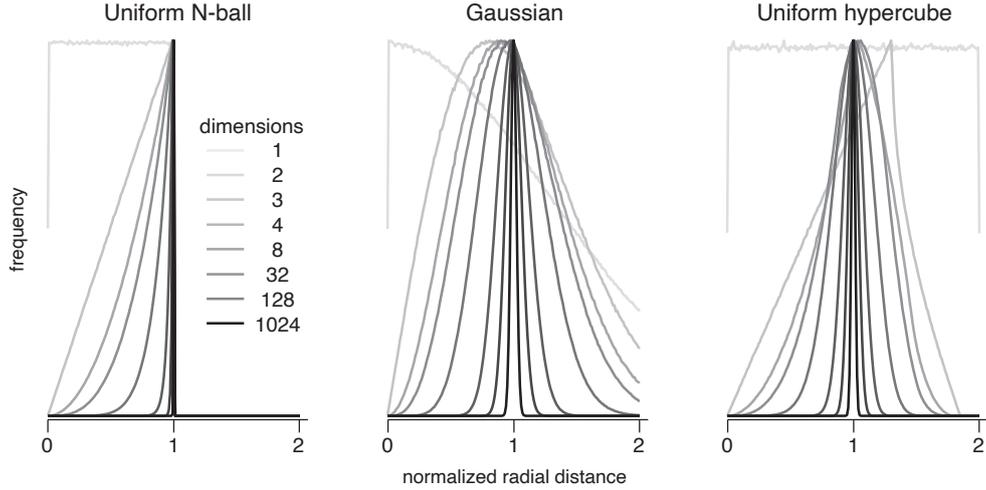


Figure S2. Different distributions tend towards the hyperspherical in high dimensions.

Histograms of 100,000 sample points' distances from the center of a uniform ball (left), Gaussian distribution (center), and uniform hypercube (right) of increasing dimensionality (increasingly dark lines). All histograms are vertically scaled so that their maximum frequencies match. The Gaussian distribution radial distances are normalized to $\gamma(N)$ (see Equation 9). The uniform hypercube radial distances are normalized to the expected mean distance from the center, given by $\int_{[-\frac{1}{2}, \frac{1}{2}]^N} \|\mathbf{x}\|_2 d\mathbf{x}$, which quickly converges to $\sqrt{N/12}$ with increasing N . The distribution types converge in high dimensions to a delta function at 1: the distribution for a uniform unit hypersphere.

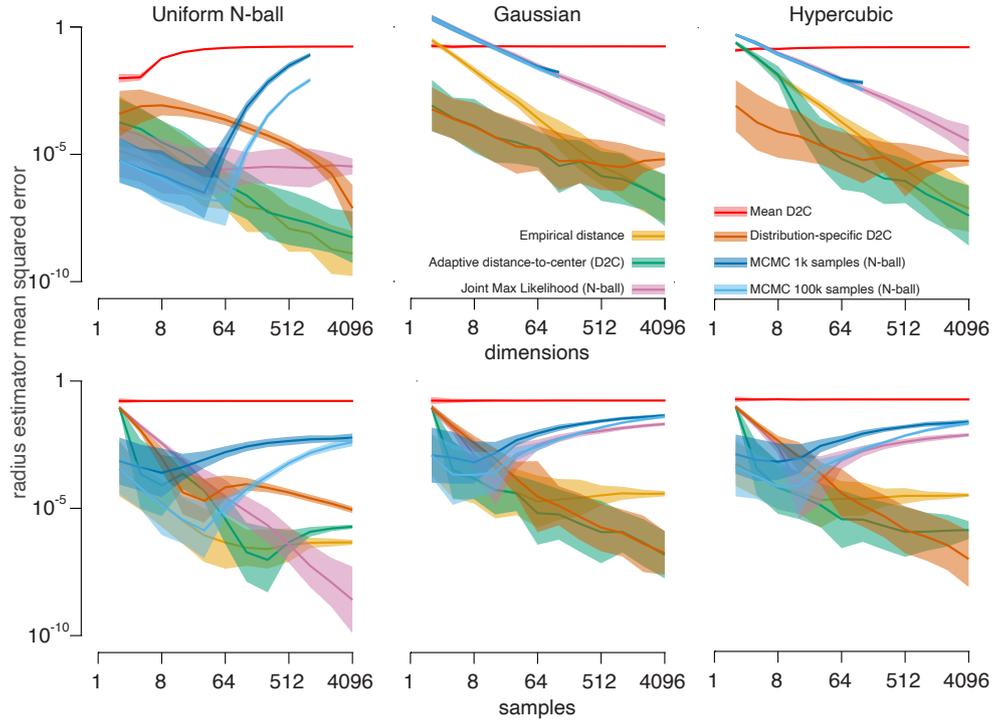


Figure S3. H2S radius estimator performance.

Comparison of estimator accuracy (radius-normalized squared error) for radii, on uniform N -ball (left), Gaussian (center), and uniform hypercube (right) generating distributions. In each case in the top row, 200 points were randomly drawn, and re-drawn 100 times to generate the bootstrapped confidence intervals indicating ± 1 standard deviation. The bottom row shows the radius estimator error as a function of samples, for 200-dimensional data. Estimators that use a N -ball distributional assumption are indicated in the legend. MCMC estimators were not computed for high dimensions due to computational constraints.

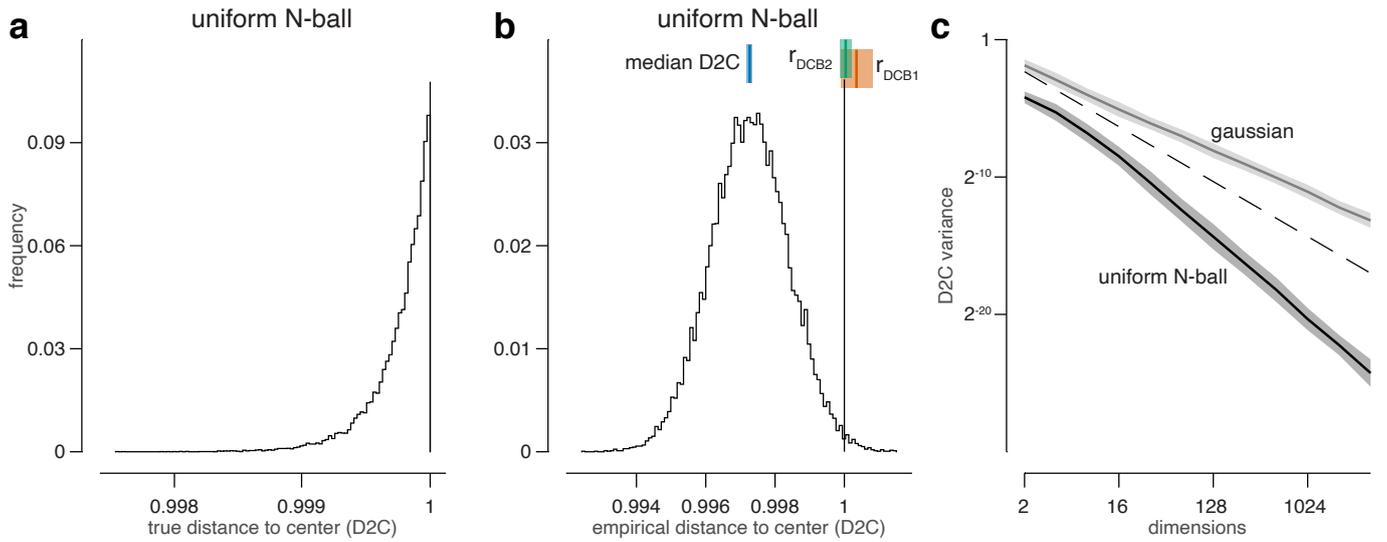


Figure S4. Adaptive D2C estimator motivation.

(a) Histogram of distances-to-center (D2C) for a uniform N -ball ($N = 4096$) relative to the ground-truth center, combined over 100 independent simulations of a data set with 200 sample points. This corresponds to the right-most condition plotted in the top-left of Supplementary Figure S3. (b) Histogram of the D2C for the same simulated data as in (a), but computed relative to the empirical means (centroids) in each of the 100 simulated data sets, as would happen in data analysis, where the center must be estimated. The empirical median D2C and its ± 1 standard error is plotted above in blue. The empirical median D2C is a negatively biased estimate of the radius. The estimators r_{DCB1} and r_{DCB2} aim to reduce the bias. r_{DCB1} slightly overestimates the true radius, whereas r_{DCB2} is more accurate. The r_{DCB2} estimator corrects the negative bias by adding a small multiple $\xi(N)$ (estimated by simulation for each N) of the standard deviation of the empirical D2C distribution (Equation 12). (c) D2C variance decreases with increasing dimensionality. The uniform N -ball and Gaussian distributions are linearly separable in a log-log plot of variance of the D2C distribution versus dimensionality. The dashed line corresponds to the criterion used in r_{adapt} (Equation 13) to discriminate Gaussian and uniform N -ball distributions.

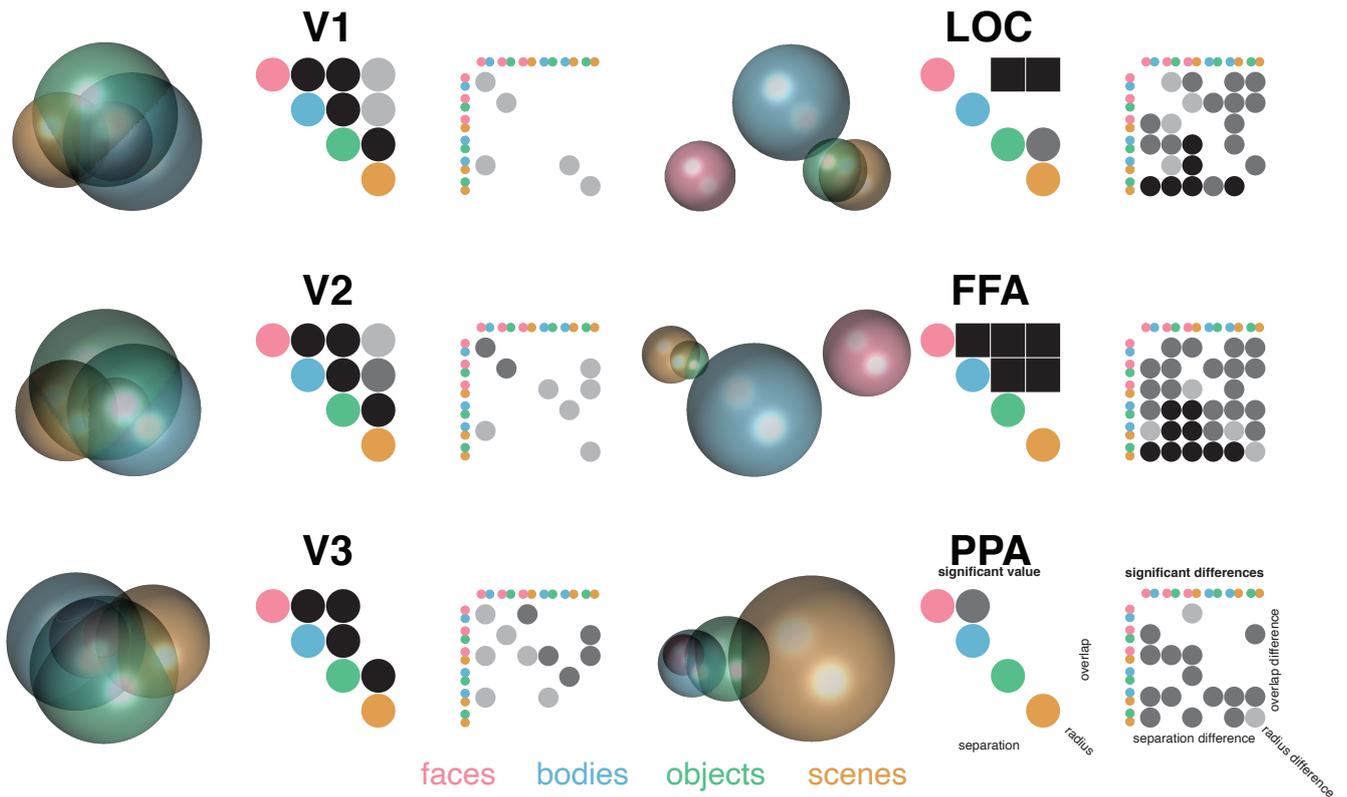


Figure S5. H2S visualization of brain activity pattern distributions associated with visual object categories.

The top row shows the same H2S visualizations for brain regions of interest as in Figure 4c. Modified Hinton-diagrams of the summary statistic values (second columns) and significance (third columns) show how categories are not distinguishable (in terms of overlap and separation) in early visual areas (V1-V3) but are in later areas (LOC, FFA, PPA).

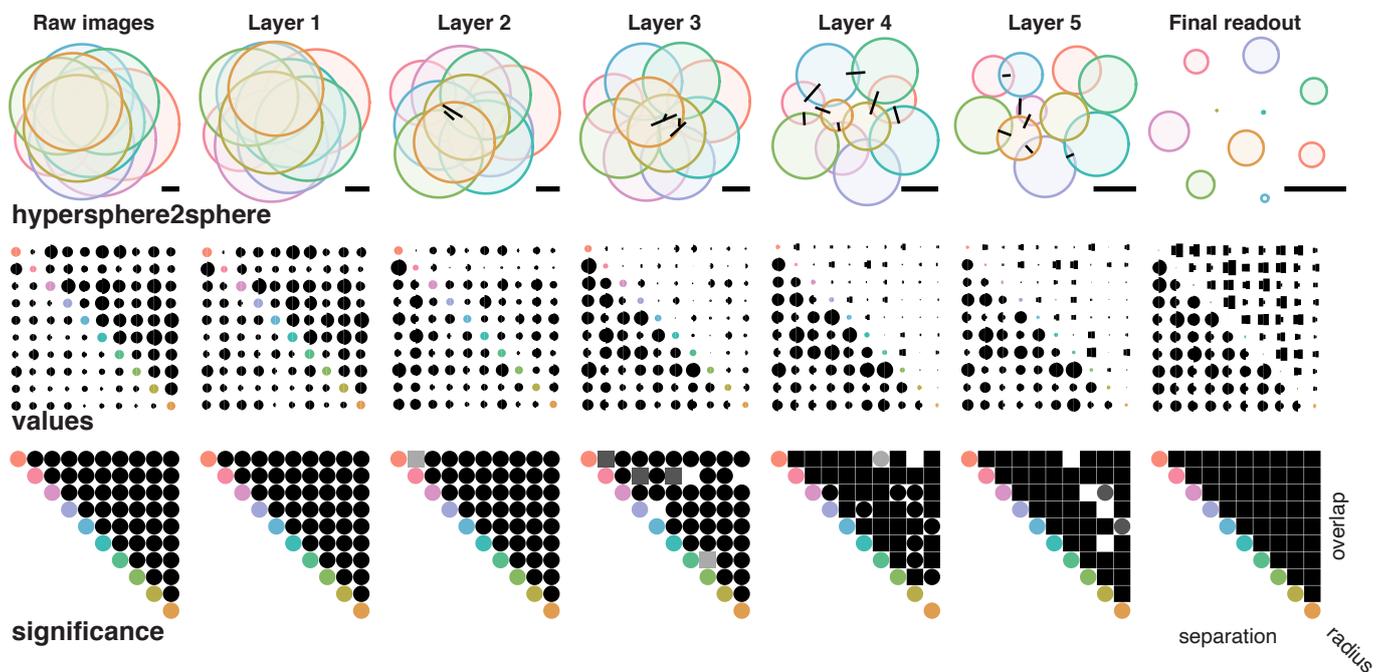


Figure S6. H2S visualization of LeNet-5 neural network layer-wise representations of MNIST digits.

The top row shows the same H2S layer-wise visualizations as in Figure 5b. Significance values computed with 10,000 bootstraps (and 10,000 jackknives for BCa-based tests). Modified Hinton-diagrams of the summary statistic values (middle row) and significance (bottom row) further make clear the gradual shift from positive overlaps to negative ones (positive margins) for representations in increasingly high layers in the network.

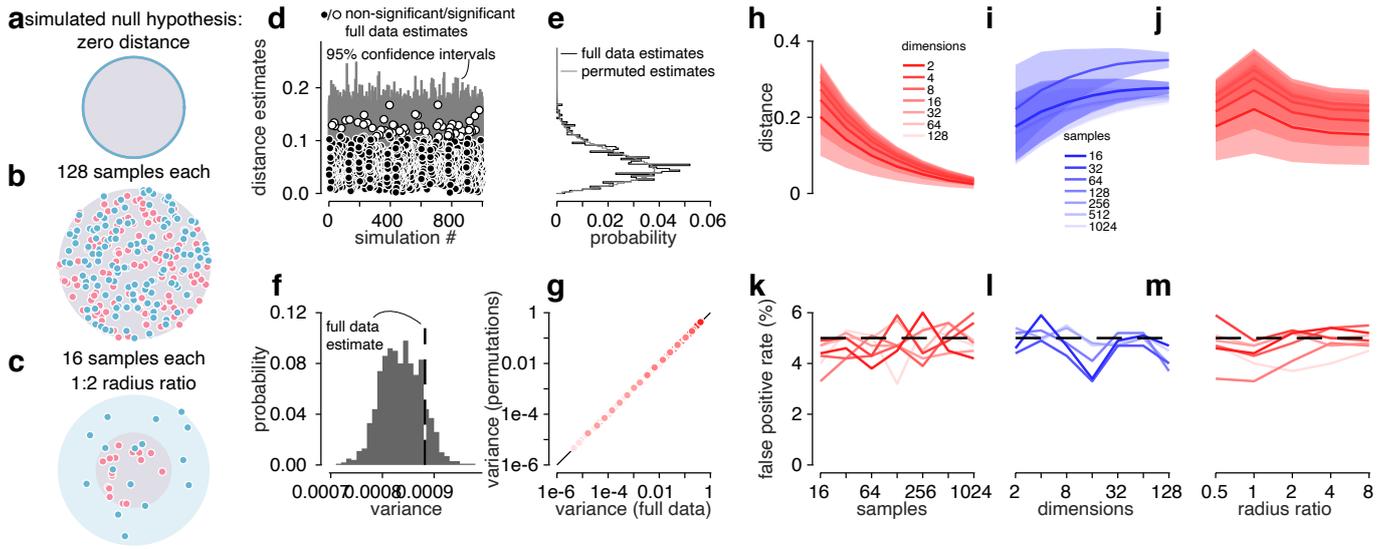


Figure S7. Validating the one-sided permutation test for separation significance.

(a) The null hypothesis is that two hyperspheres have the same center. (b) and (c) show example simulations of the null hypothesis for different parameter choices. (d) 1,000 simulations of this null hypothesis for equal radii, 256 samples per class in 2D. The separation estimates from the full (non-permuted) data and the 0-95% range of the permutation distribution (computed from label permutation) are shown as points and lines, respectively. White points indicate significant values. (e) Histograms from the 1,000 full (black) and all 5,000,000 permuted (gray) data point estimates. (f) The histogram is of each simulation's permutation distribution variance and the dashed line shows the variance of the full data estimates across all simulations. (g) The variance for permuted versus full data estimate distributions across all simulations for all parameter choices tested (lighter shades of red indicate increasing dimensionality — see legend in (h)). Estimator bias is captured by the mean separation estimate across all simulations as a function of number of samples per class (h), dimensionality (i), and radius ratio (j). False positive rate for the permutation tests as a function of number of samples per class (k), dimensionality (l), and radius ratio (m).

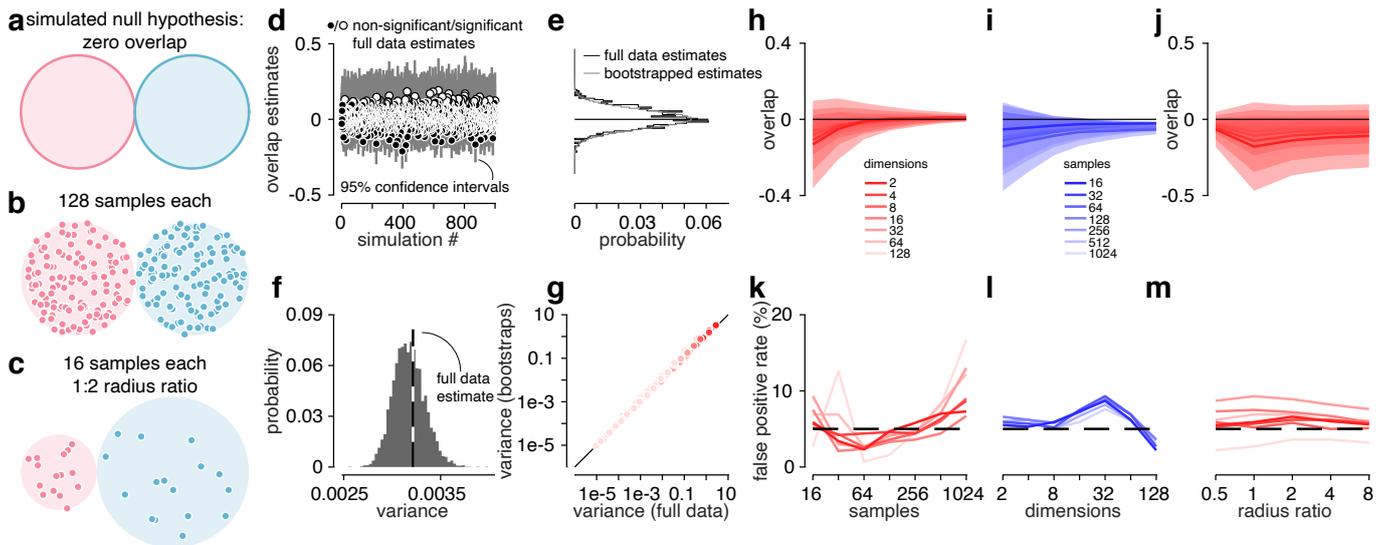


Figure S8. Validating the two-sided accelerated bootstrap confidence interval test for overlap significance.

(a) The null hypothesis is that two hyperspheres have zero overlap. The panels in this figure follow Supplementary Figure S7 and use the accelerated bootstrap confidence interval instead of permutation testing.

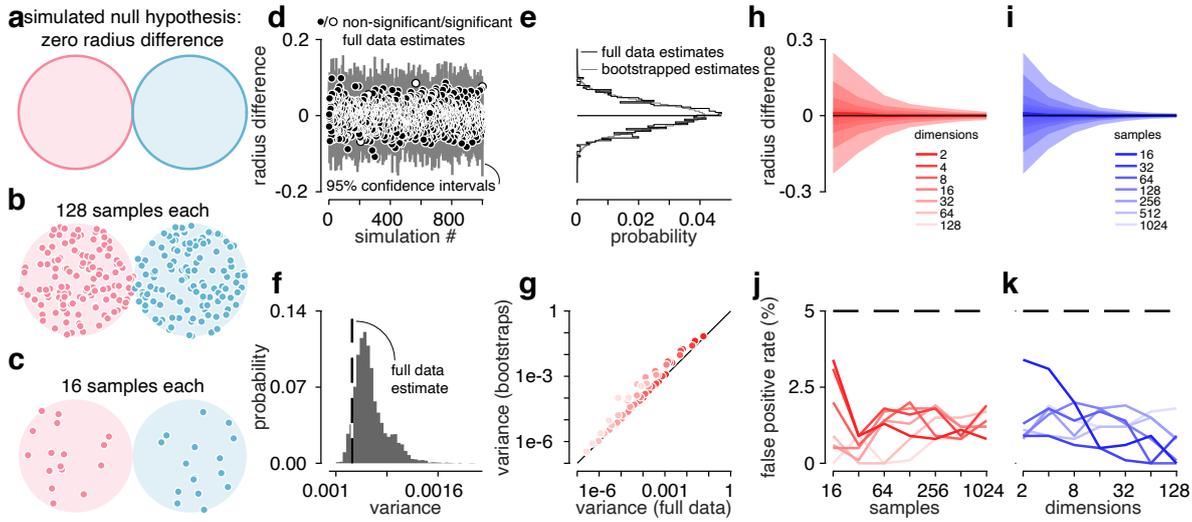


Figure S9. Validating the two-sided bootstrap confidence interval test for radius difference significance.

(a) The null hypothesis is that two hyperspheres have the same radius (and therefore their difference is zero). The panels in this figure follow Supplementary Figure S7 and use the accelerated bootstrap confidence interval instead of permutation testing. Radius ratio is by definition 1, so panels related to radius ratio are omitted.

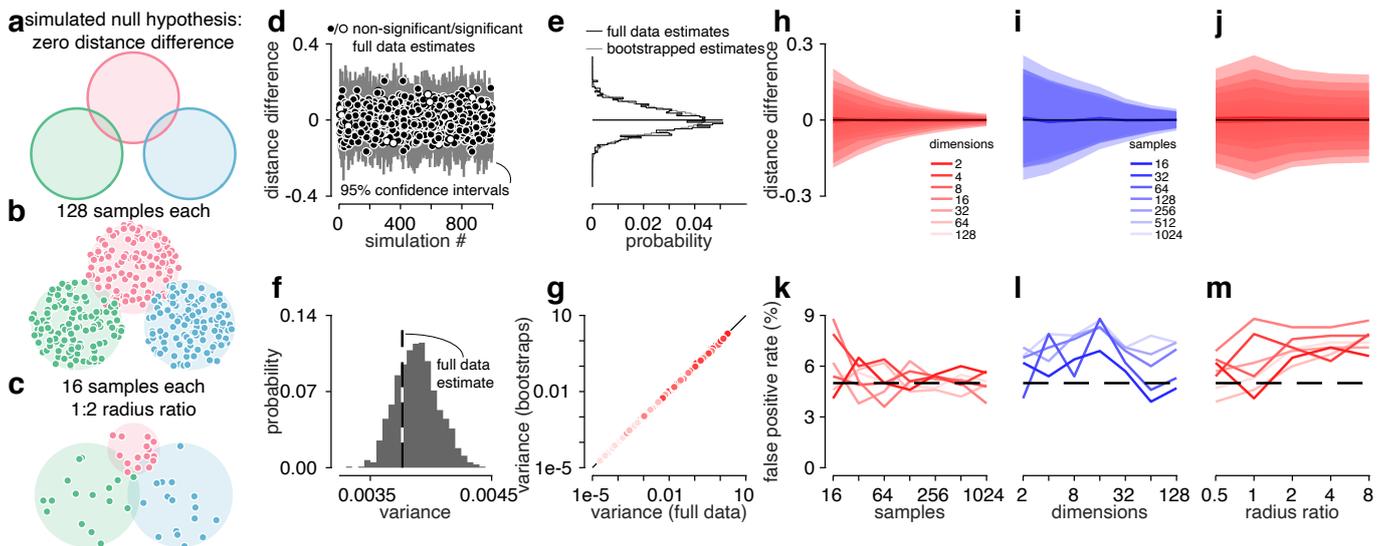


Figure S10. Validating the two-sided bootstrap test for separation difference significance.

(a) The null hypothesis is that of three hyperspheres, two pairs have the same separation (and therefore their difference is zero). The panels in this figure follow Supplementary Figure S7 and use bootstrap instead of permutation testing.

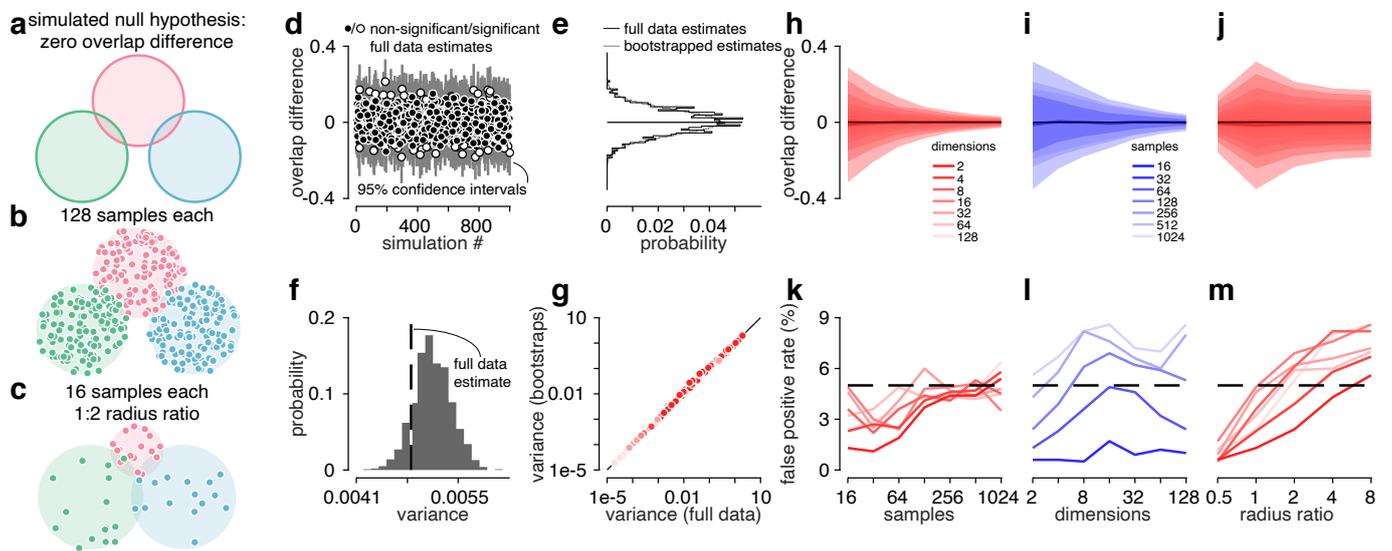


Figure S11. Validating the two-sided bootstrap test for overlap difference significance.

(a) The null hypothesis is that of three hyperspheres, two pairs have the same overlap (and therefore their difference is zero). The panels in this figure follow Supplementary Figure S7 and use bootstrap instead of permutation testing.