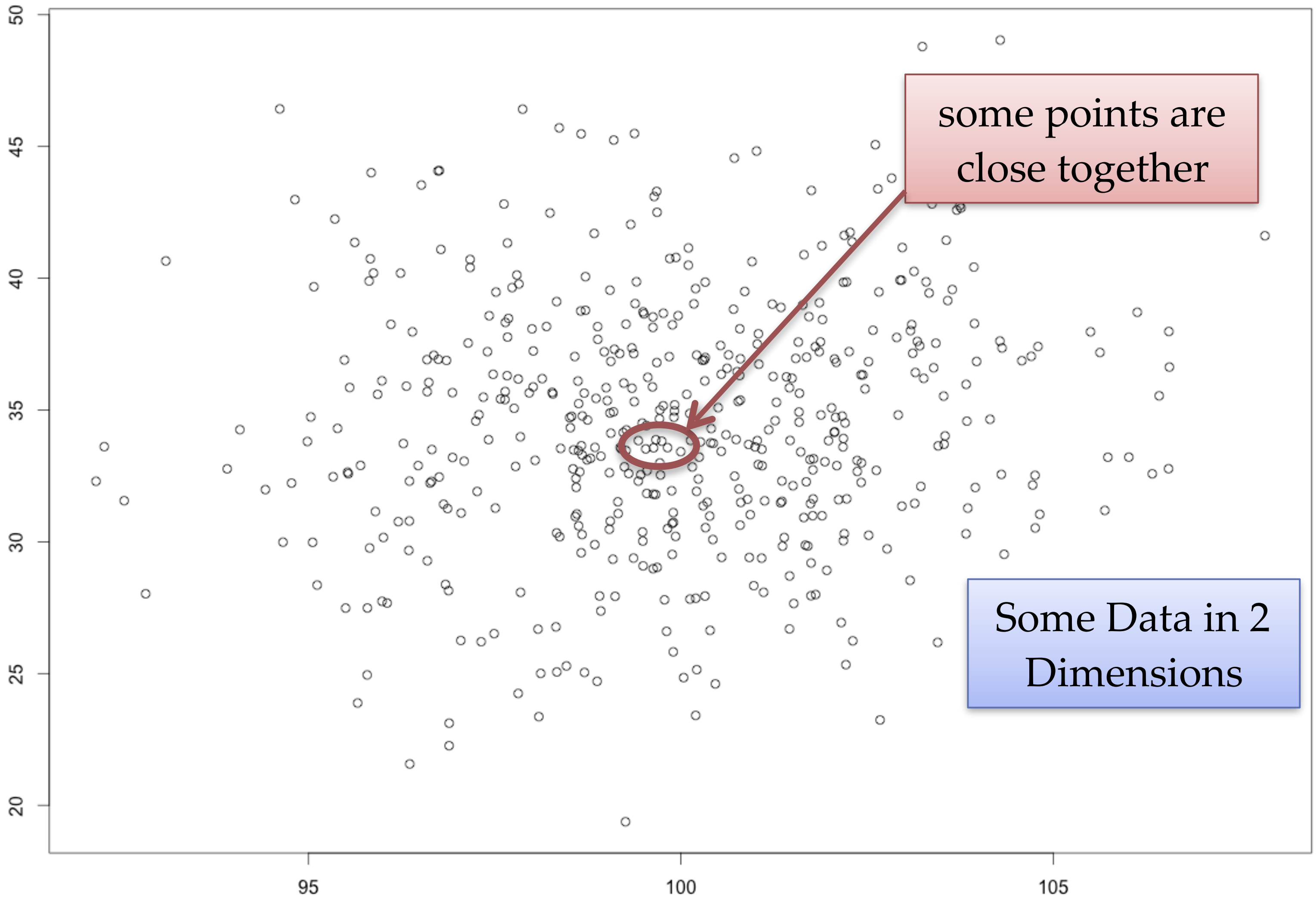
Dimension Reduction

Why and How

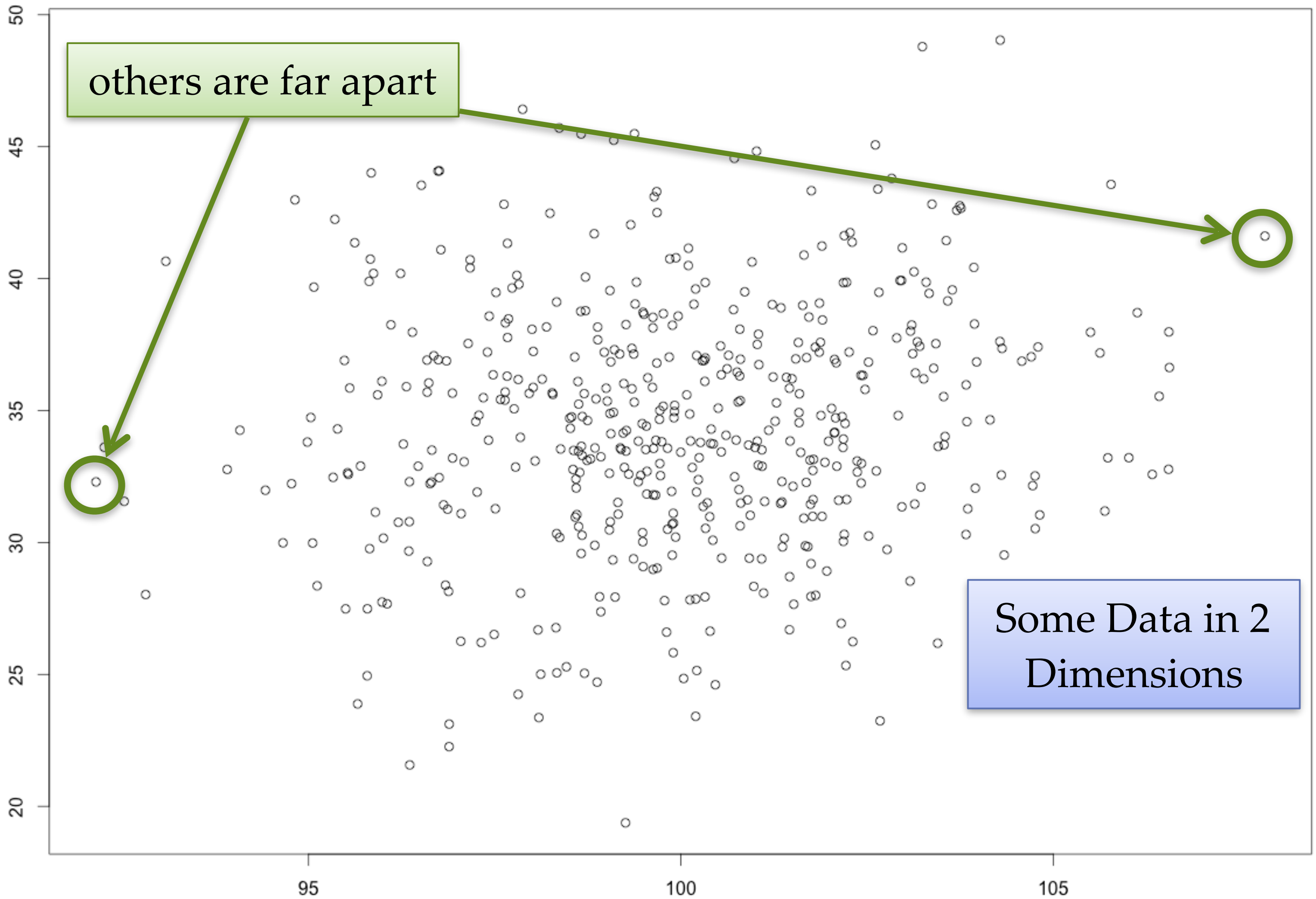
The Curse of Dimensionality

- As the dimensionality (i.e. number of variables) of a space grows, data points become so spread out that the ideas of *distance* and *density* become murky.
- Let’s explore this fact…

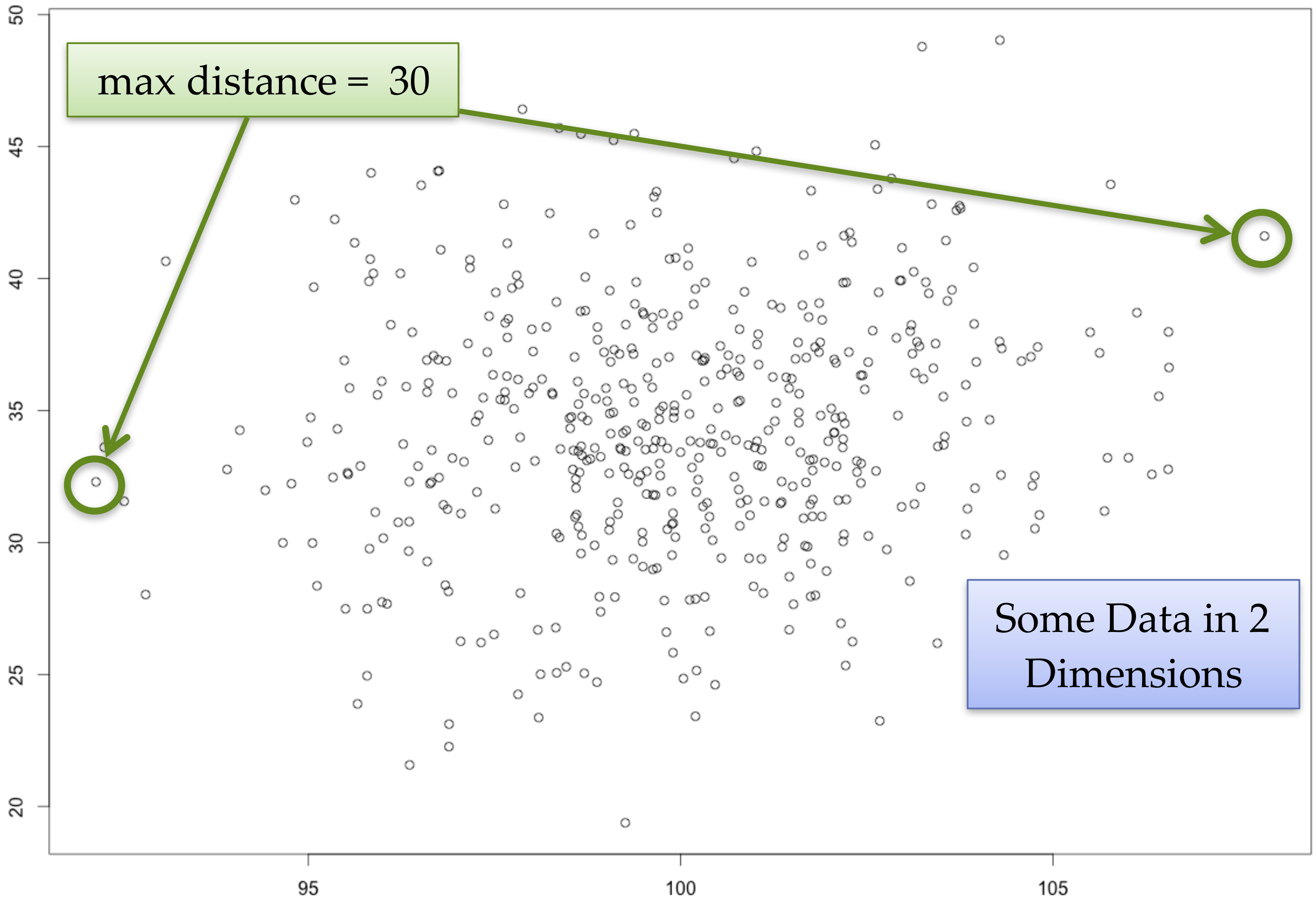
The Curse of Dimensionality



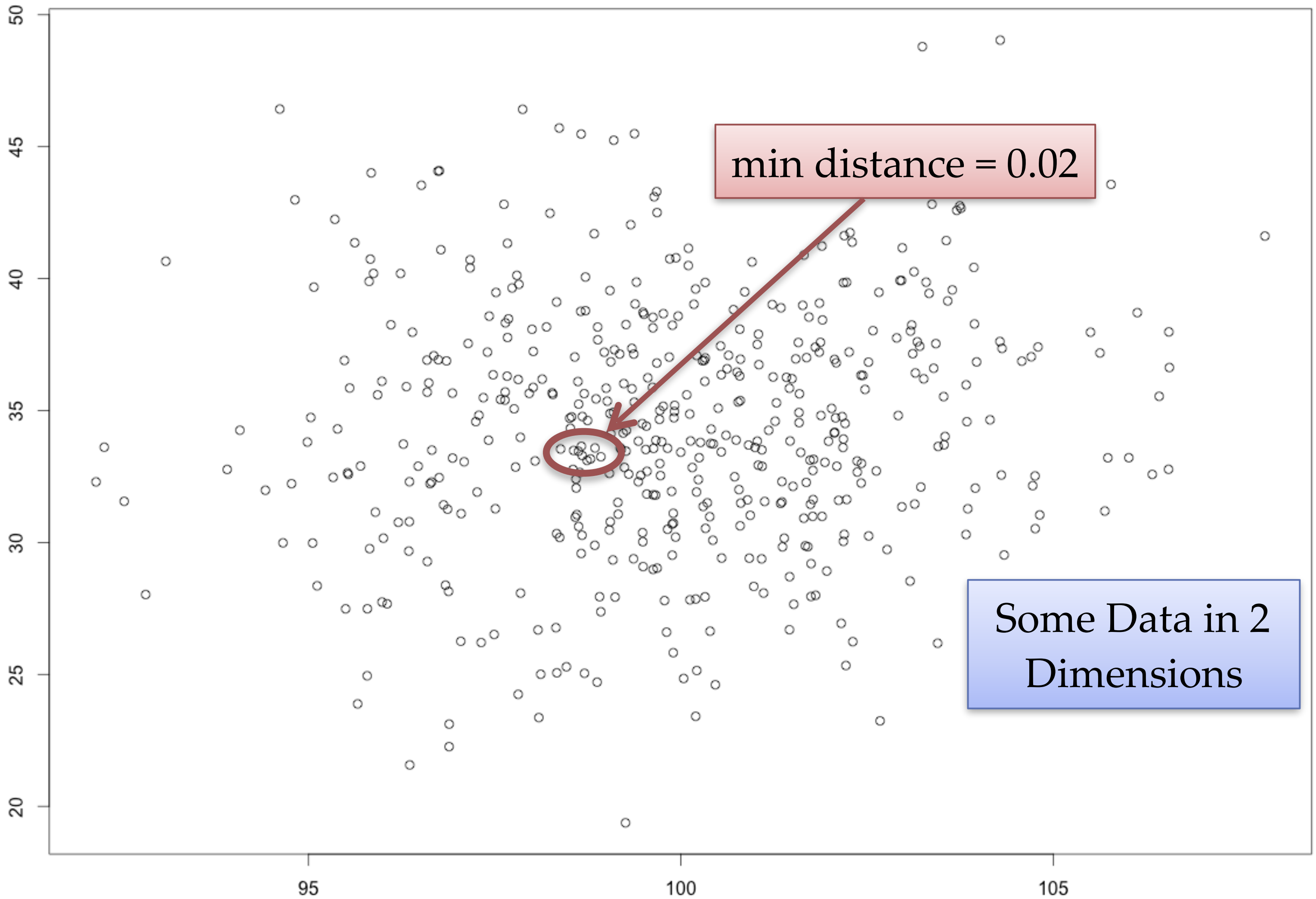
The Curse of Dimensionality



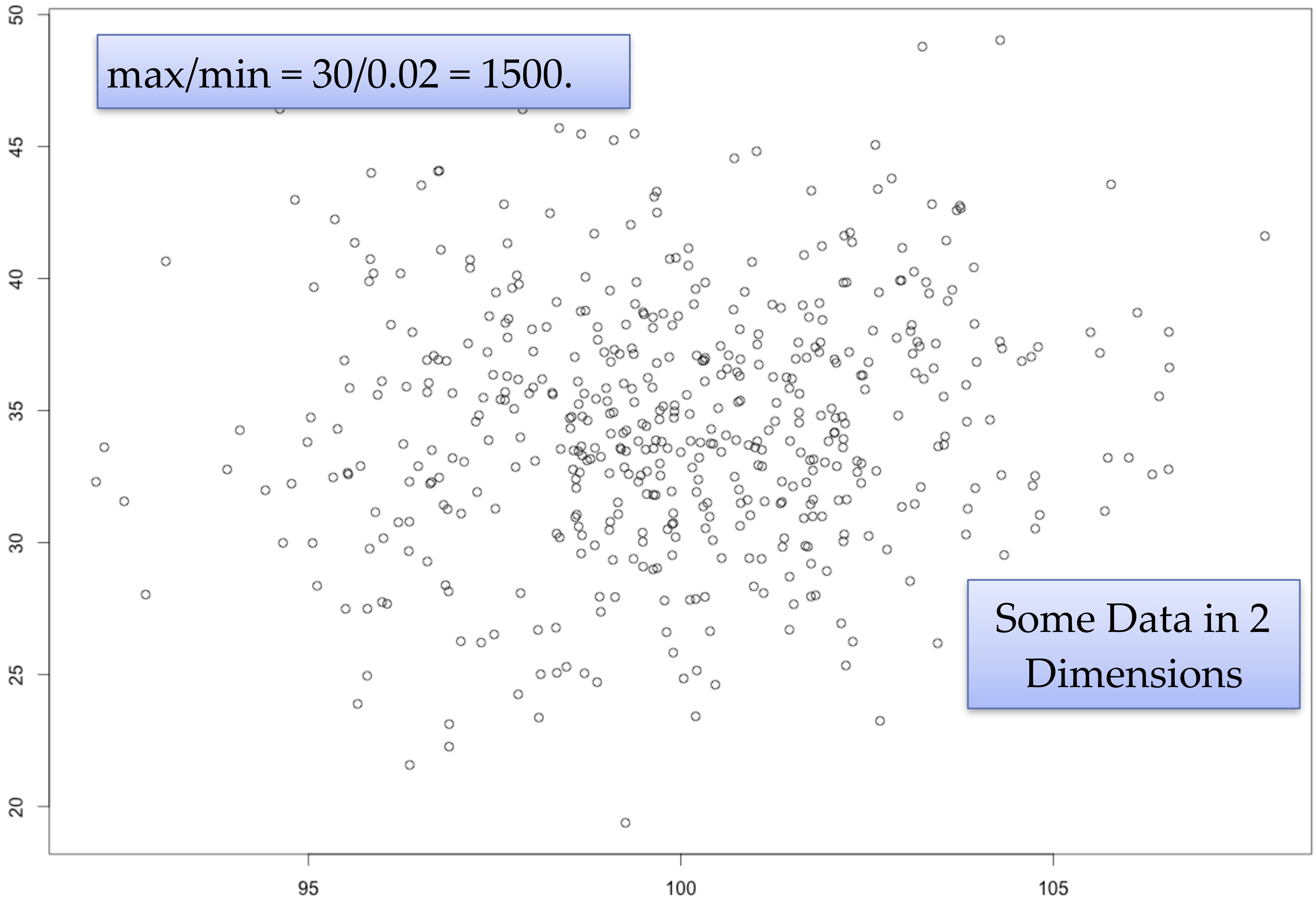
The Curse of Dimensionality



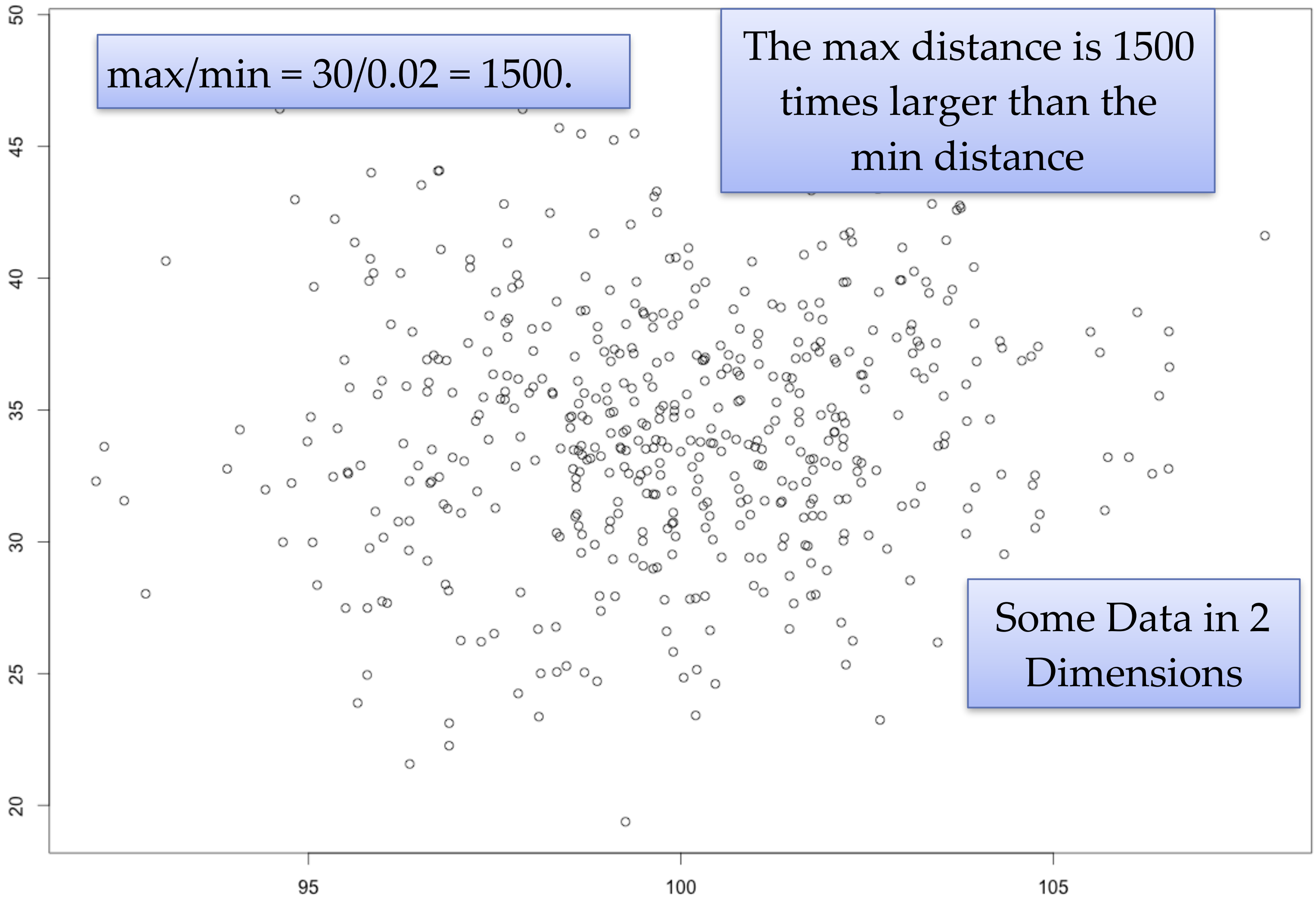
The Curse of Dimensionality



The Curse of Dimensionality



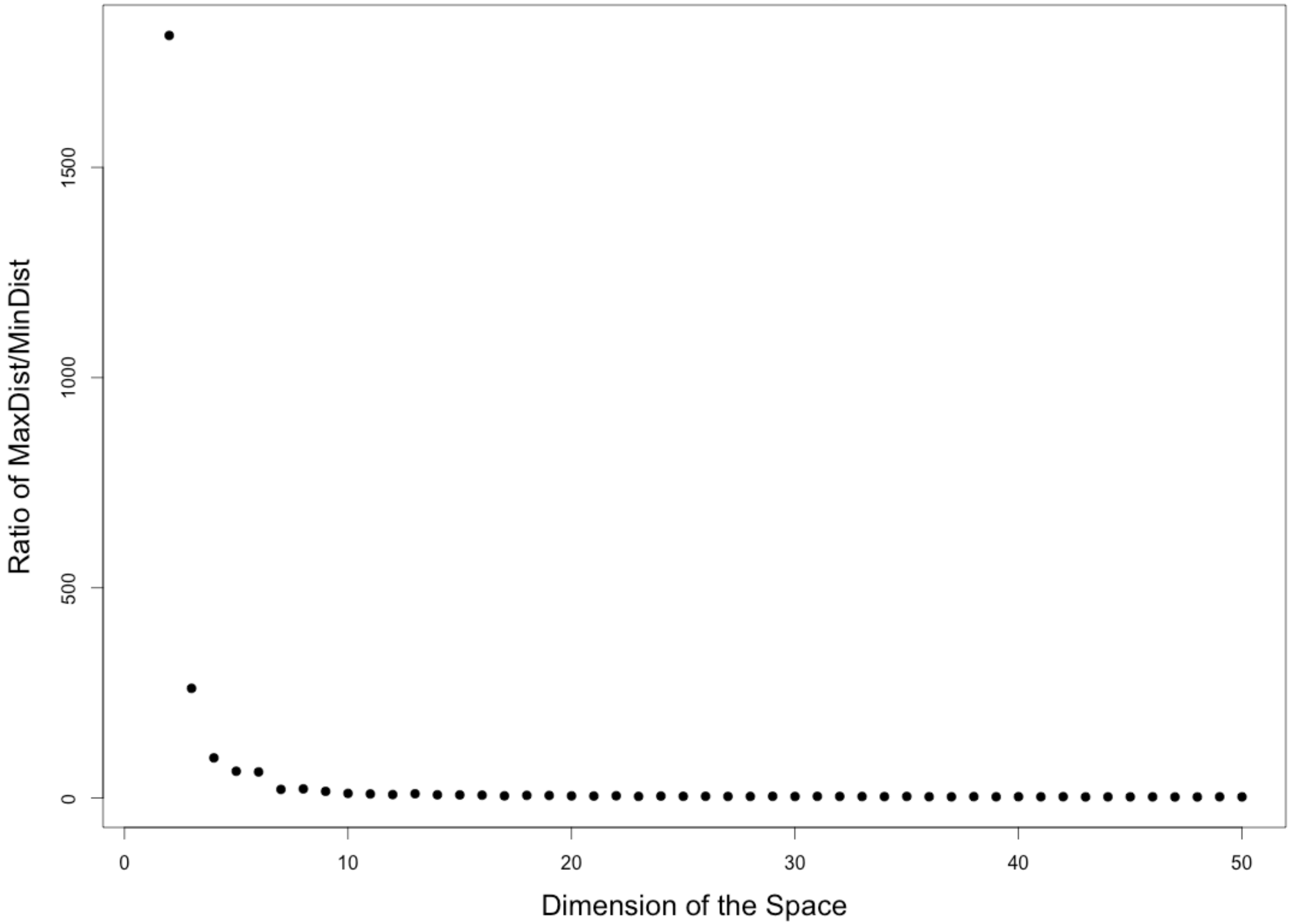
The Curse of Dimensionality



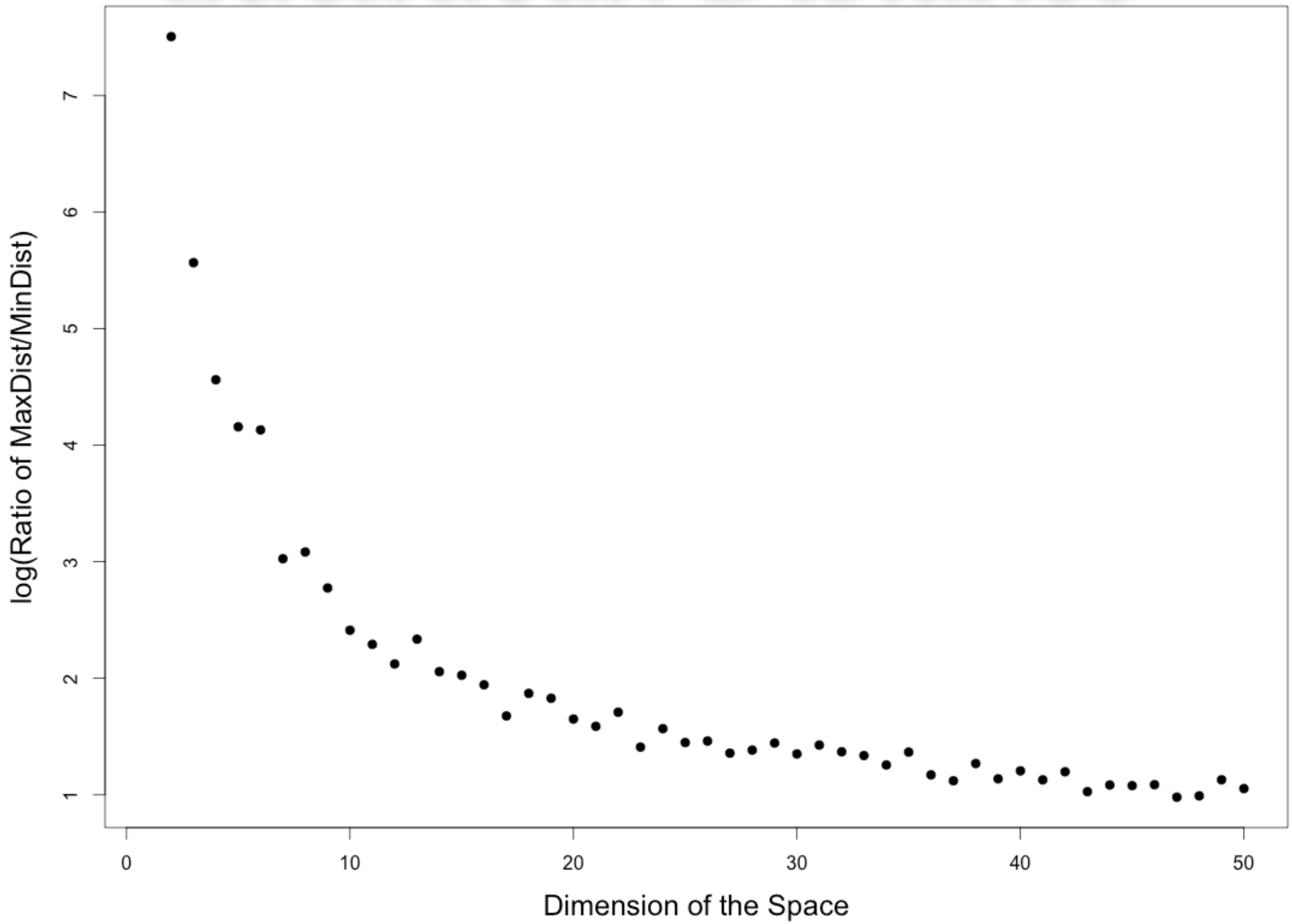
The Curse of Dimensionality

- Now lets generate those 500 points in 3-space, 4-space, … , 50-space.
- We'll compute that same metric, the ratio of the maximum distance to the minimum distance
- See how it changes as the number of dimensions grows…

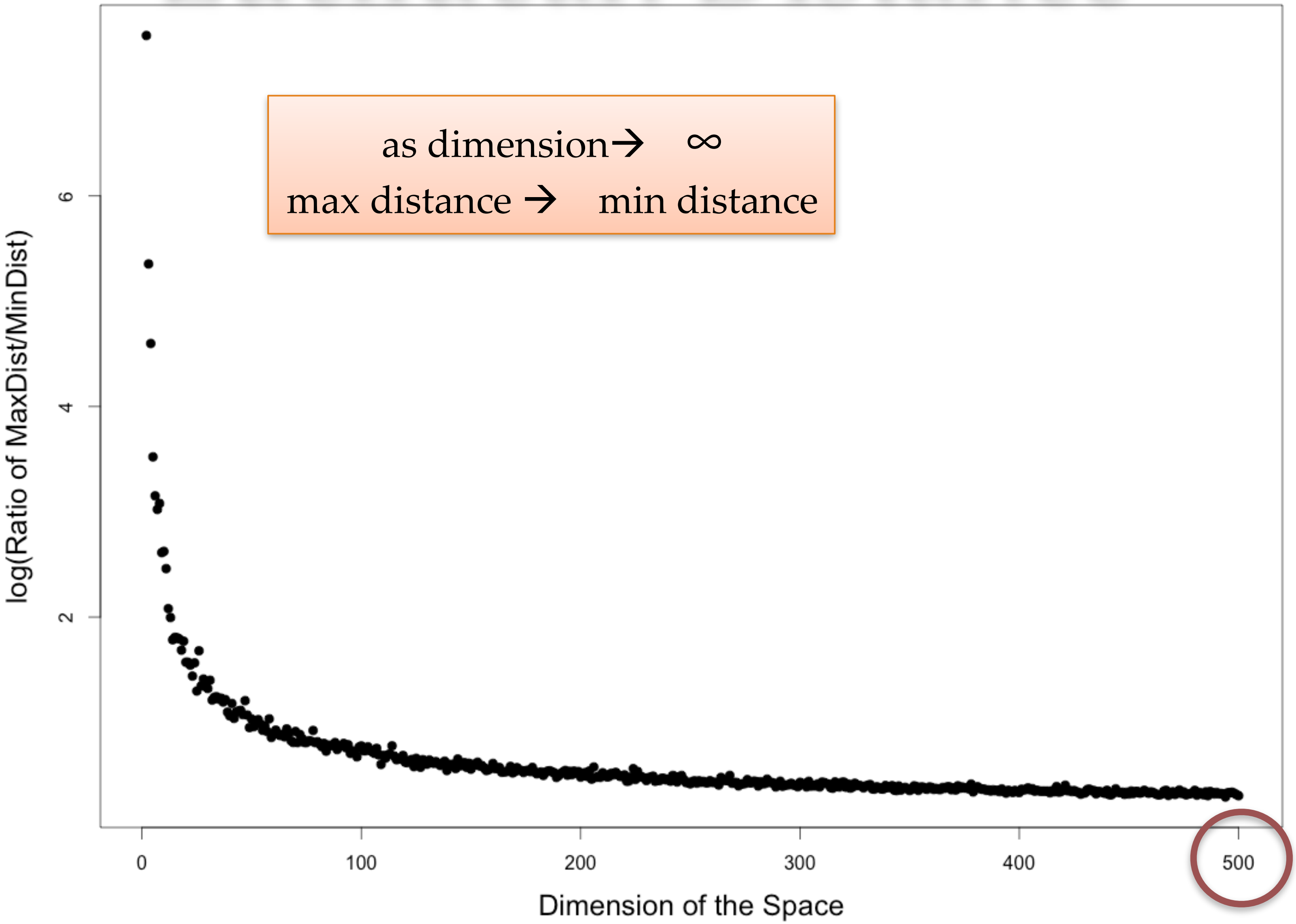
The Curse: Euclidean Distance



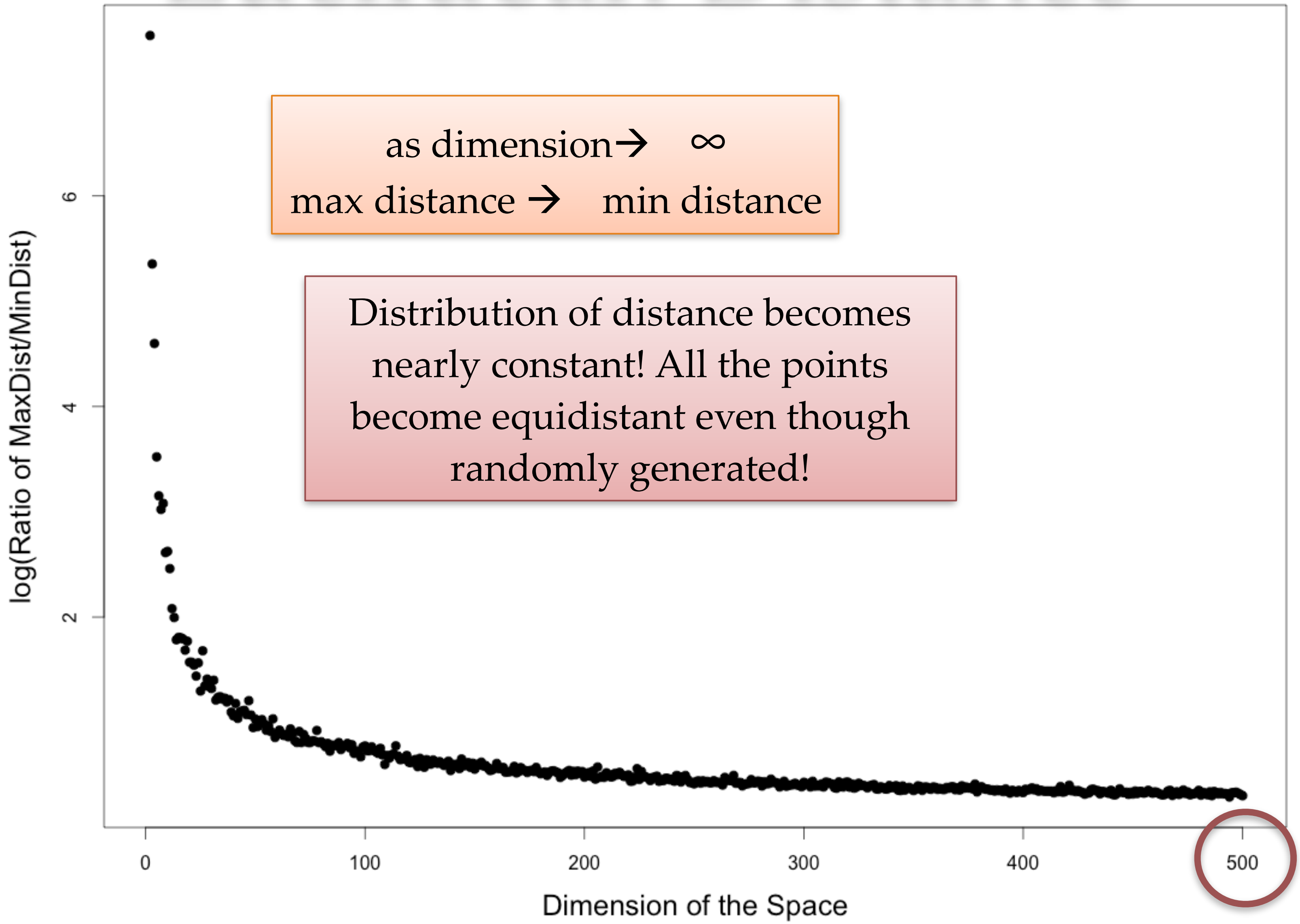
The Curse: Euclidean Distance



The Curse: Euclidean Distance

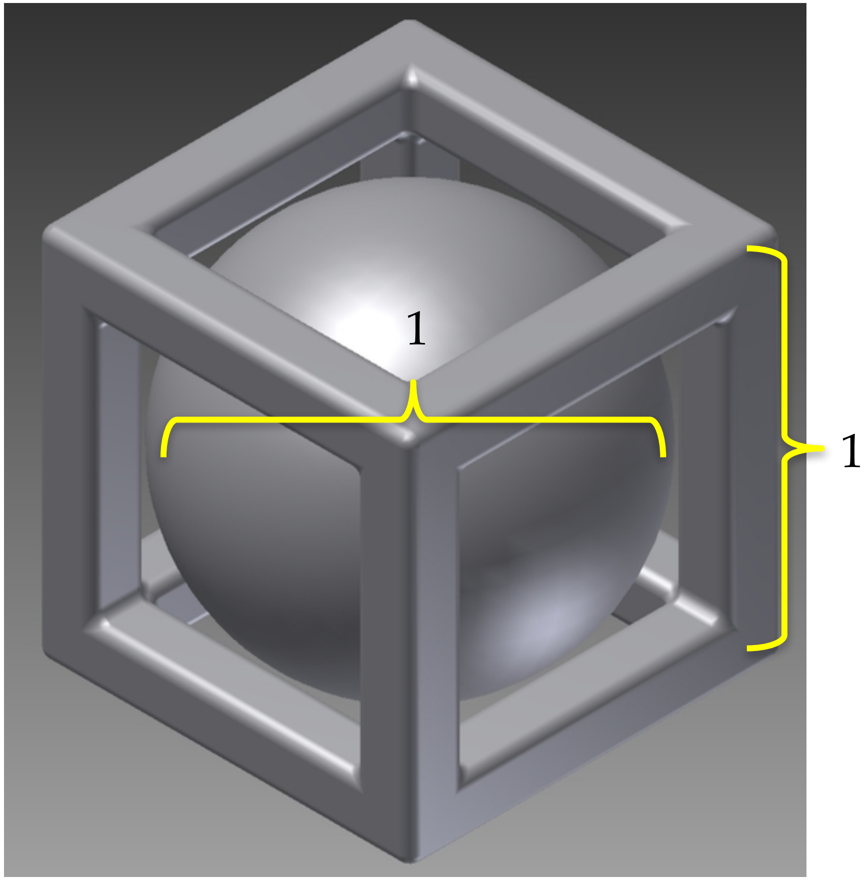


The Curse: Euclidean Distance

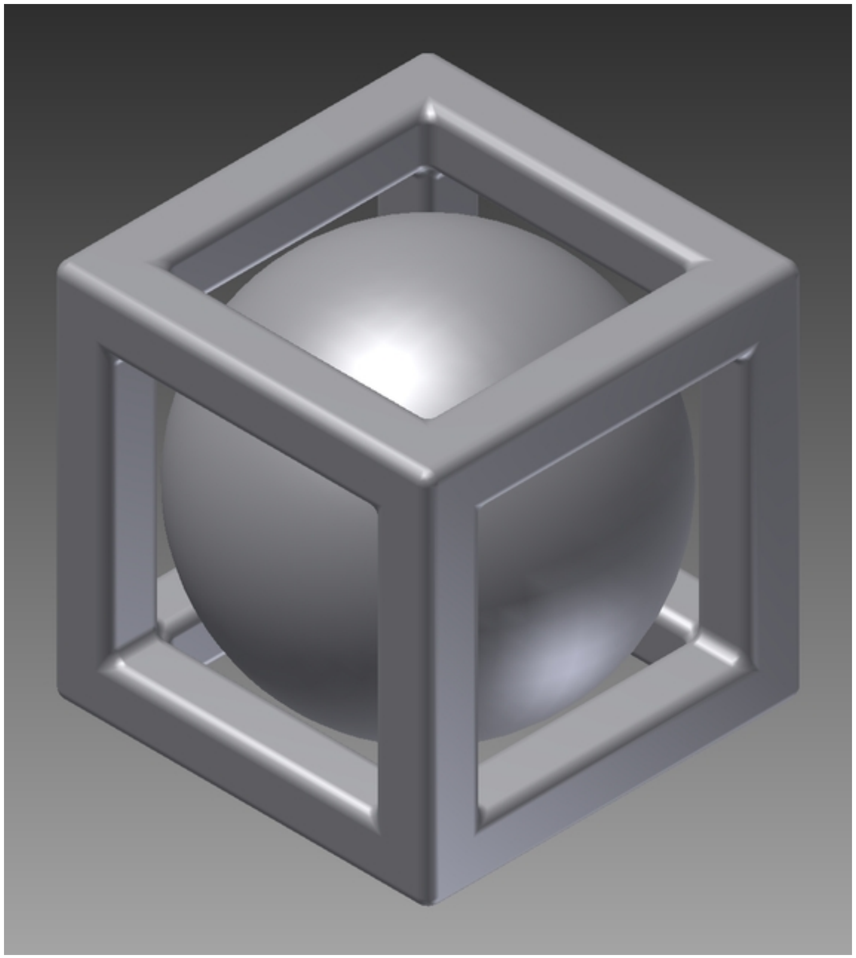


The Curse: Volume of Sphere to Cube

- ➢ Here's another one.
- ➢ Imagine a sphere that sits perfectly (inscribed) inside of a cube.
- ➢ In 3-dimensions, it looks like this:
- ➢ For simplicity, it's a unit cube and unit diameter sphere



The Curse: Volume of Sphere to Cube



Volume of Sphere:

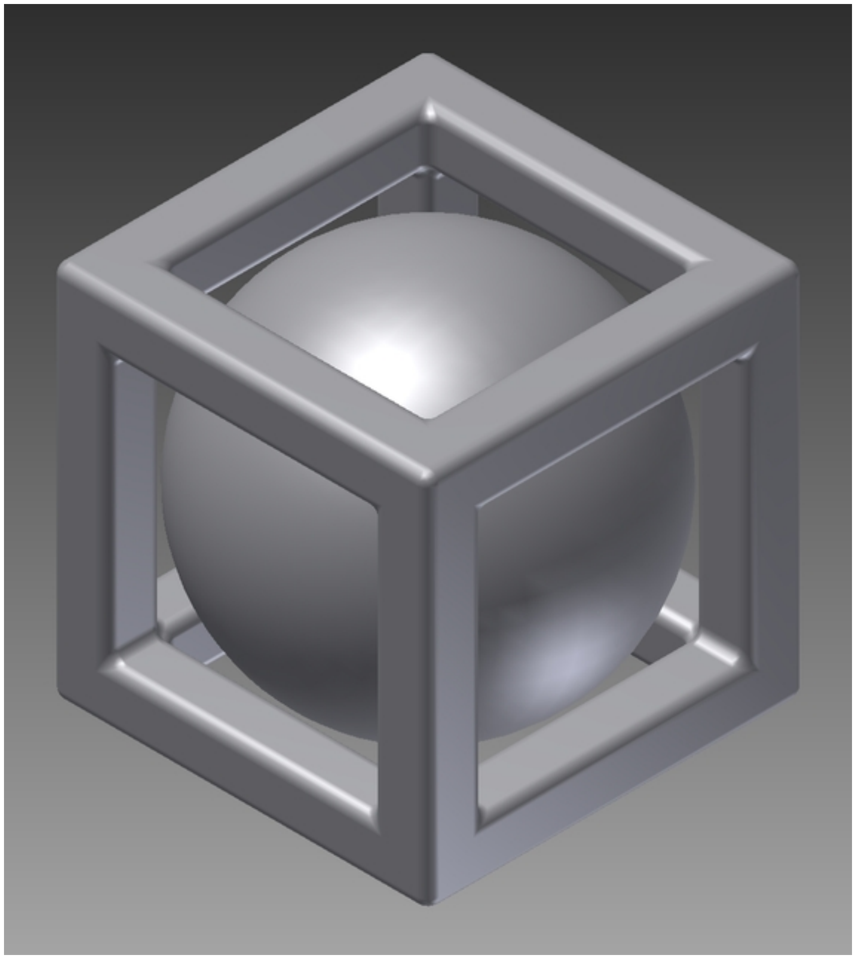
$(4/3)\pi(0.5)^3 \approx 0.52$

Volume of Cube:

1

So the sphere takes up over half of the space.

The Curse: Volume of Sphere to Cube



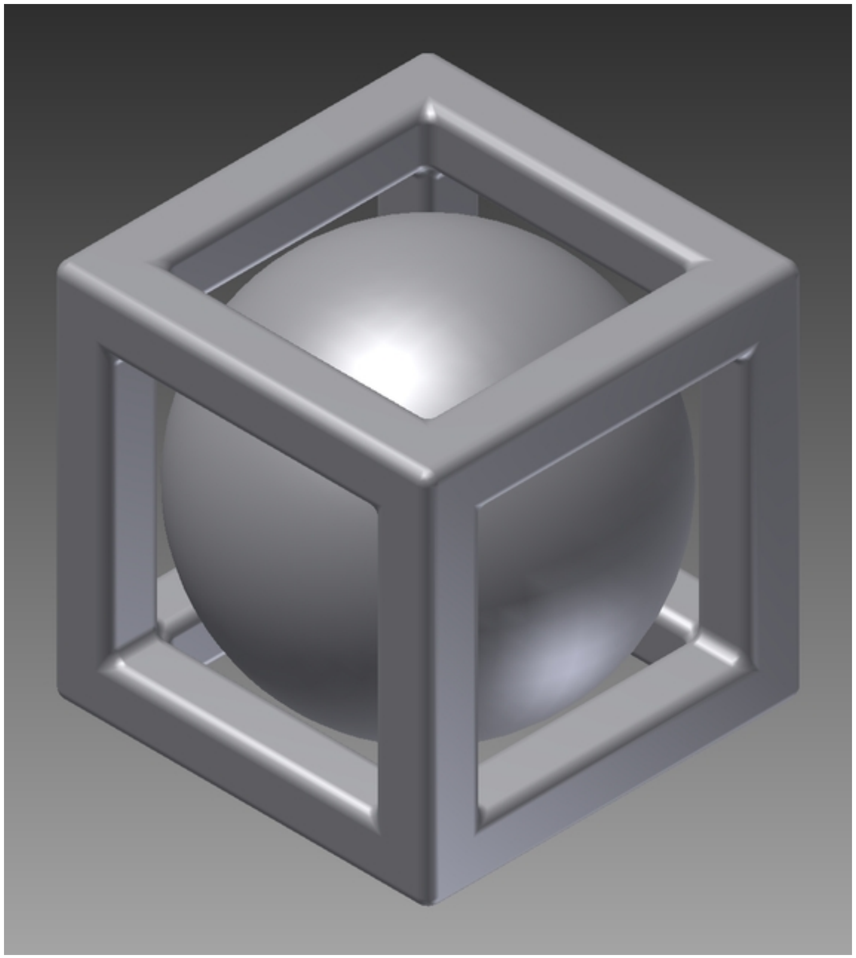
In d-space, the volume of hypersphere:

$$\frac{2r^n \pi^{n/2}}{n\Gamma\left(\frac{n}{2}\right)}$$

Volume of hypercube:

1

The Curse: Volume of Sphere to Cube



As d→∞, the ratio of the volume of the sphere to the cube gets closer and closer to 0.

$$\lim_{d \to \infty} \frac{\text{SphereVolume}}{\text{CubeVolume}} = 0$$

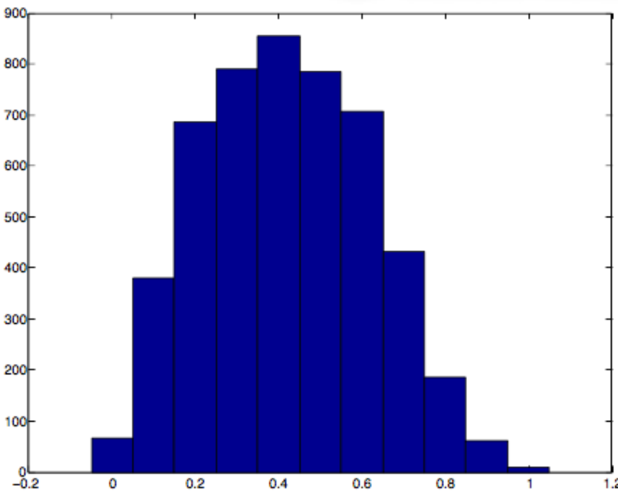
It's as if ALL of the volume of the hypercube is contained in the corners!

(none in the sphere, relatively speaking)

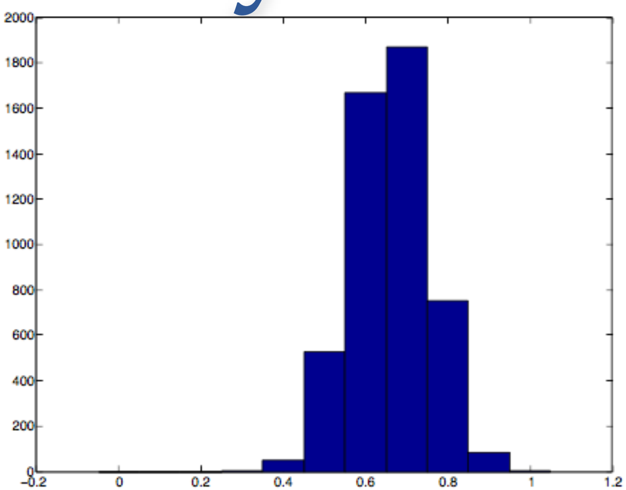
The Curse of Dimensionality

- No distance/similarity metric is immune to the vastness of high dimensional space.
- One more. Let’s look at the distribution (or lack thereof) of cosine similarity.
- Compute the cosine similarity between each pair of points, and divide that similarity by the maximum.

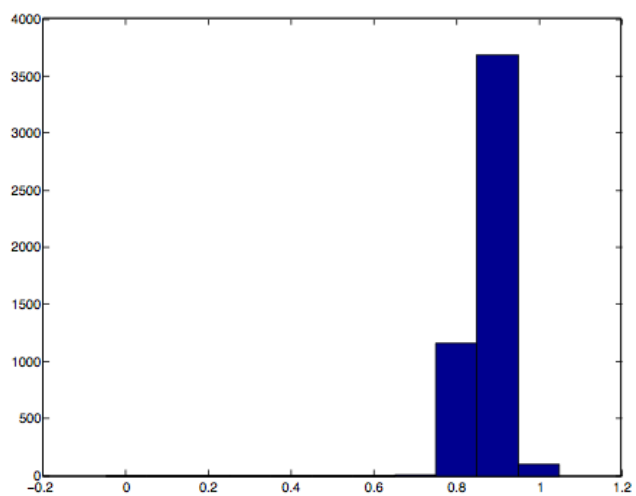
The Curse: Cosine Similarity



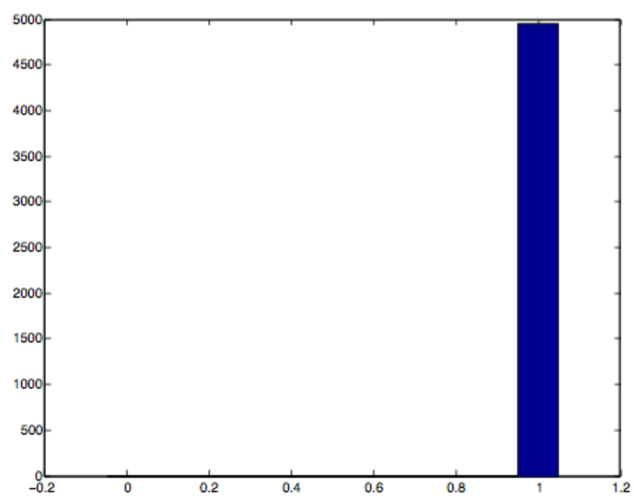
n=2



n=20



n=200



n=20000

When is this a problem?

- *Primarily* when using algorithms which rely on distance or similarity
  - Particularly for clustering and k nearest neighbor methods
- Secondarily on all models due to collinearity and a desire for model simplicity.
- Computational/storage complexity can be problematic in all algorithms.

What can we do about it?

Dimension Reduction

Dimension Reduction Overview

FEATURE SELECTION

Choose subset of existing features

By their relationship to a target (supervised)

By their distribution (unsupervised)

FEATURE EXTRACTION

Create new features

Often linear combinations of existing features (PCA, SVD, NMF)

Often chosen to be uncorrelated

Feature Selection

- Removing features manually
  - Redundant (multicollinearity/VIFs)
  - Irrelevant (Text mining stop words)
  - Poor quality features (>50% missing values)
- Forward/Backward/Stepwise Regression
- Decision Tree
  - Variable Importance Table
  - Can change a little depending on metric
    - Gini/Entropy/Mutual Information/Chi-Square

Feature Extraction: Continuous Variables

- PCA
  - Create a new set of features as linear combinations of your originals
  - These new features are ranked by variance (importance/information)
  - Use the first several PCs in place of original features
- SVD
  - Same as PCA, except the 'variance' interpretation is no longer valid
  - Common for text-mining, since $\mathbf{X}^T\mathbf{X}$ is related to cosine similarity.
- Factor Analysis
  - The principal components are rotated so that our new features are more interpretable.
  - Occasionally other factor analysis algorithms like maximum likelihood are considered.

Feature Extraction: Continuous Variables

- Discretization/Binning
  - While this doesn't reduce the dimensions of your data (it increases them!), it is still a form of feature extraction!

Feature Extraction: Nominal Variables

- Encoding variables with numeric values.

Checking Account Balance	
<u>Original Level</u>	<u>New Value</u>
Negative	-100
No checking account	0
Balance is zero	0
0<Balance<200	100
200<Balance<800	500
Balance>800	900
Balance>800 and IncomeDD	1000

Feature Extraction: Nominal Variables

- Encoding variables with numeric values.
  - In a supervised learning setting, can let the numeric value of level=L be the average target value of all observations that have level = L

- Correspondence analysis
  - Method similar to PCA for categorical data.
  - Uses chi-squared table (contingency table) and chi-squared distance.
  - Can be used to get coordinates of categorical variables in a lower-dimensional space.
  - More often used as exploratory method, potentially for binning purposes.