SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of AI &DS

Course Name – 19AD602 DEEP LEARNING

III Year / VI Semester

Unit 3-DIMENSIONALITY REDUCTION
Topic: Linear (PCA, LDA) and manifolds

Case Study: Customer Segmentation in an E-commerce Platform

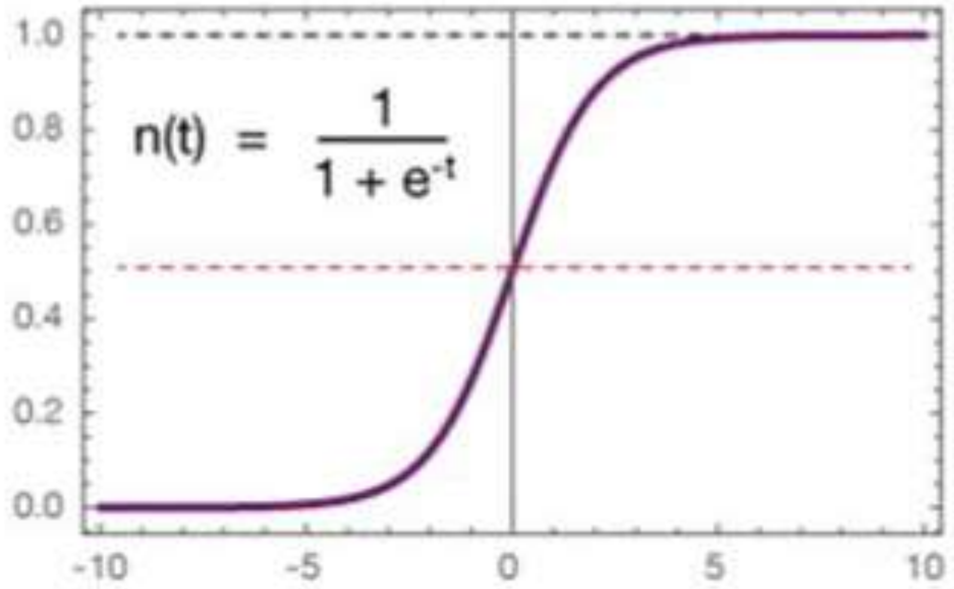
An e-commerce company wants to segment customers based on shopping behavior. PCA is applied to reduce dimensions from a dataset with purchase history, while LDA classifies customers into known groups (e.g., frequent buyers, occasional buyers). Manifold learning methods like t-SNE help visualize clusters to uncover new insights about behavioral patterns.

Limitations of Logistic Regression

Limited to binary classification problems (2 class)

Can be unstable when classes are well separated

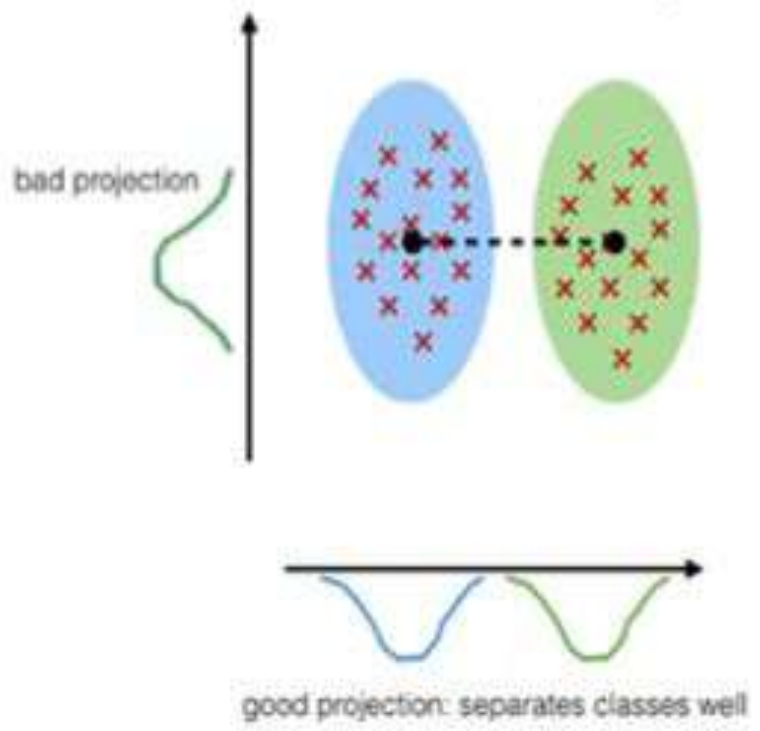
Unstable for low number of examples



Linear Discriminant Analysis (LDA)

Linear method for multi-class classification problems

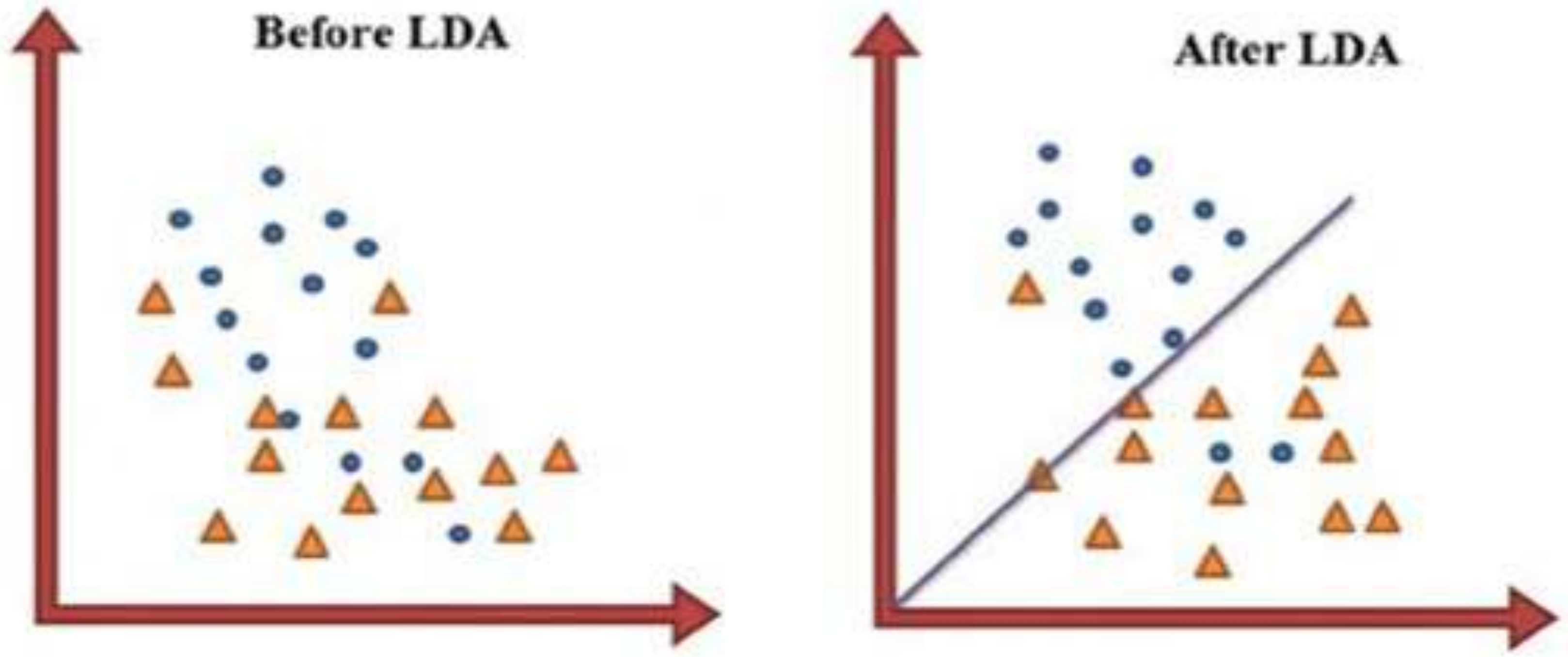
Project the features in higher dimension space into a lower dimension space.



Learning LDA

Assuming data is Gaussian (bell-shaped) and consistent variance …

LDA estimates the mean and variance

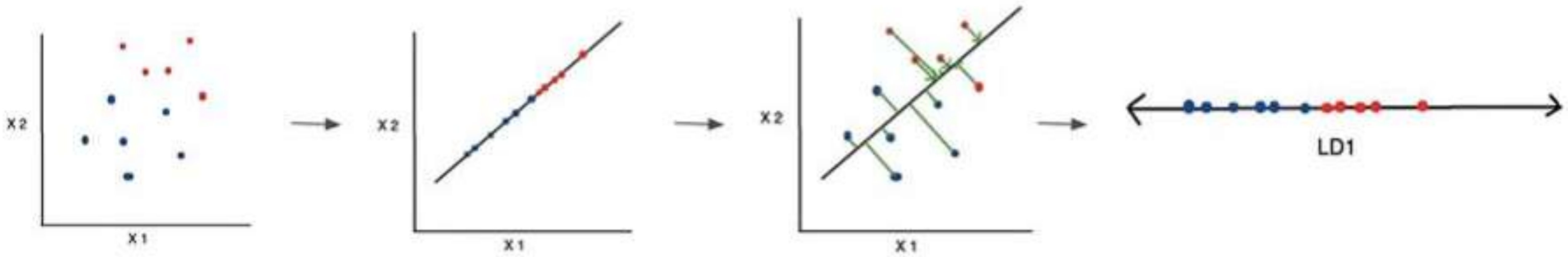


Dimensionality Reduction for LDA

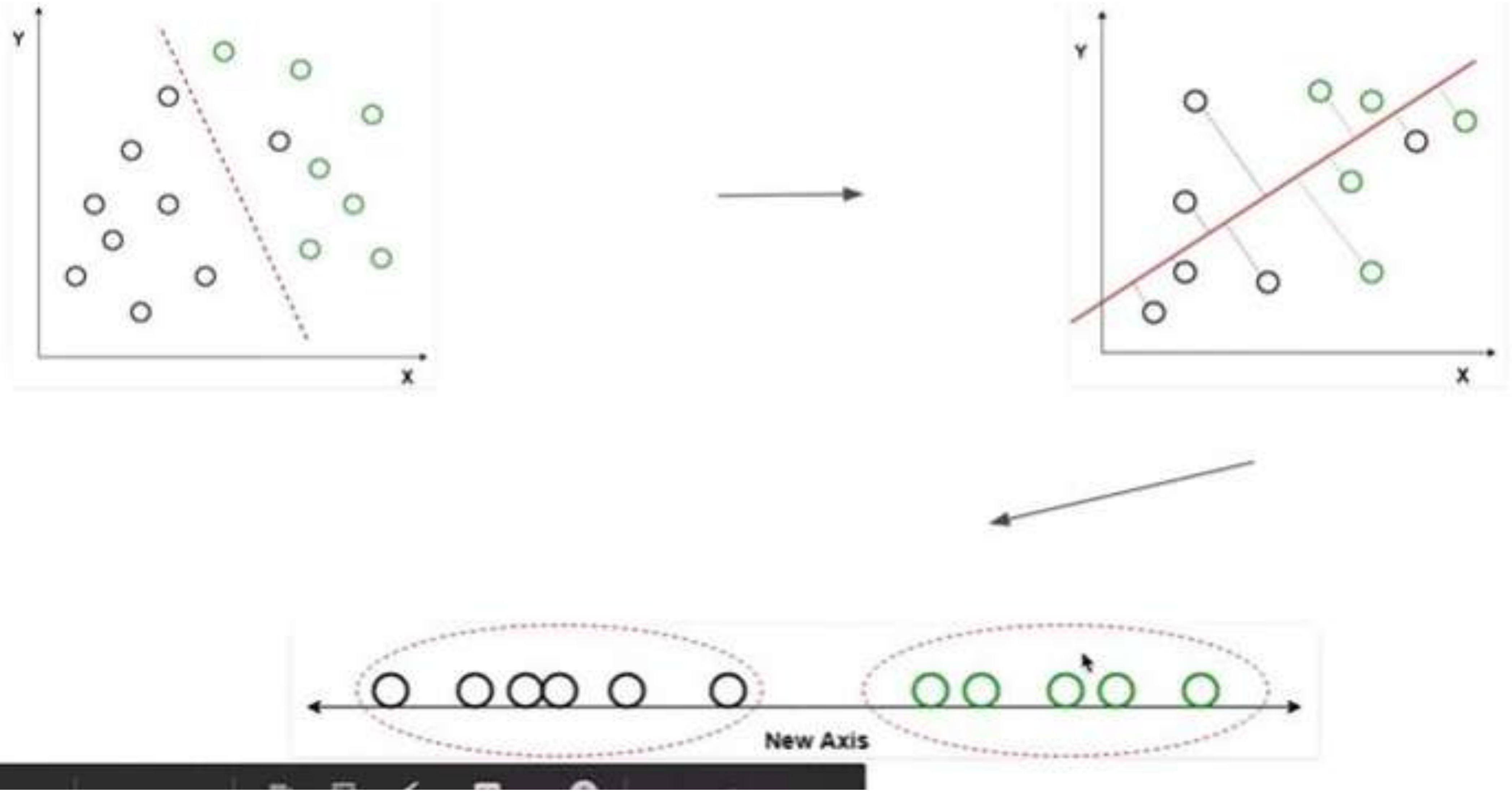
Project data into lower dimension

Creates a new axis and projects the data on to the new axis

Criteria: Minimize the variance and maximize the distance between the means

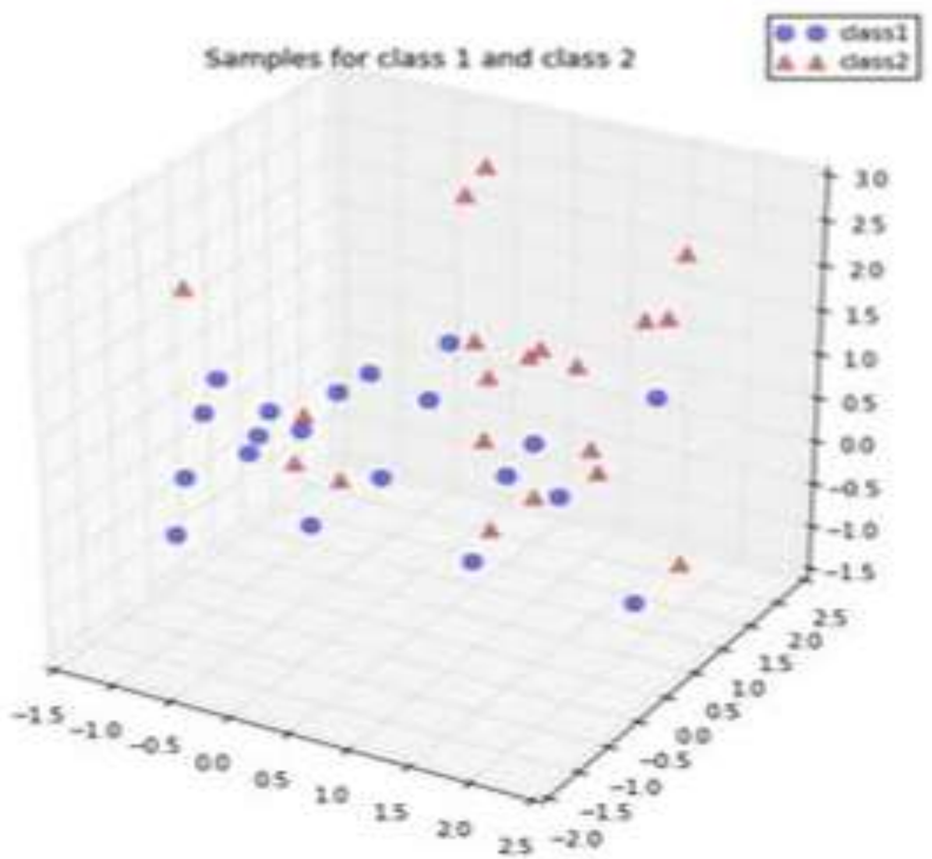


Visualizing Dimensionality Reduction for LDA



Dimensionality Reduction

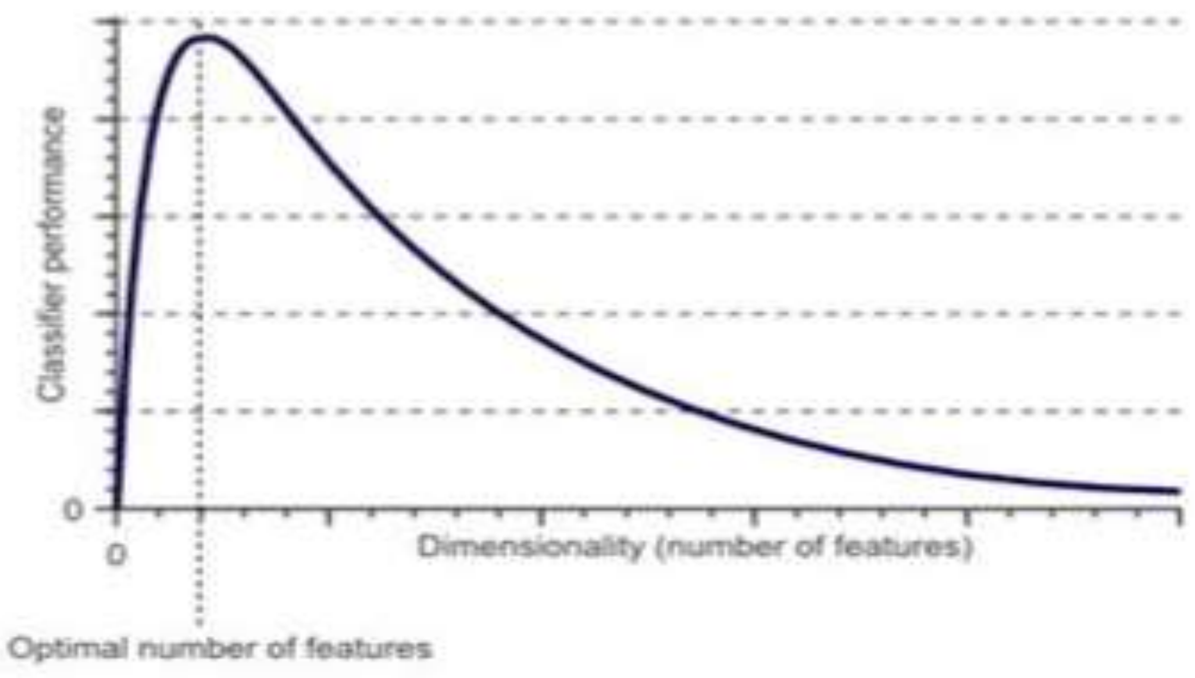
Reducing the dimensions of your feature set.



Issue with Higher Dimension Data

Classifier accuracy becomes saturated upon addition of features

Features correspond to dimensions in higher space



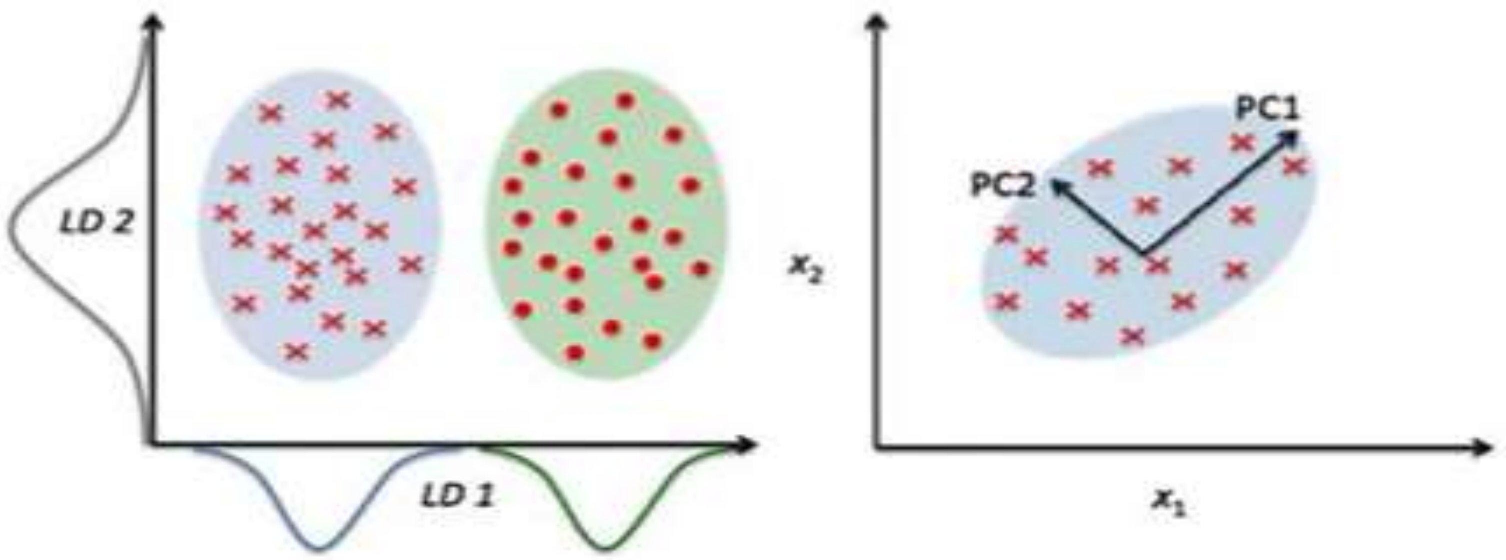
The Curse of Dimensionality

Relationship to overfitting

More features == more likely to overfit

(increasingly dependent on training data)

Dimensionality Reduction is usually done to prevent chances of overfitting

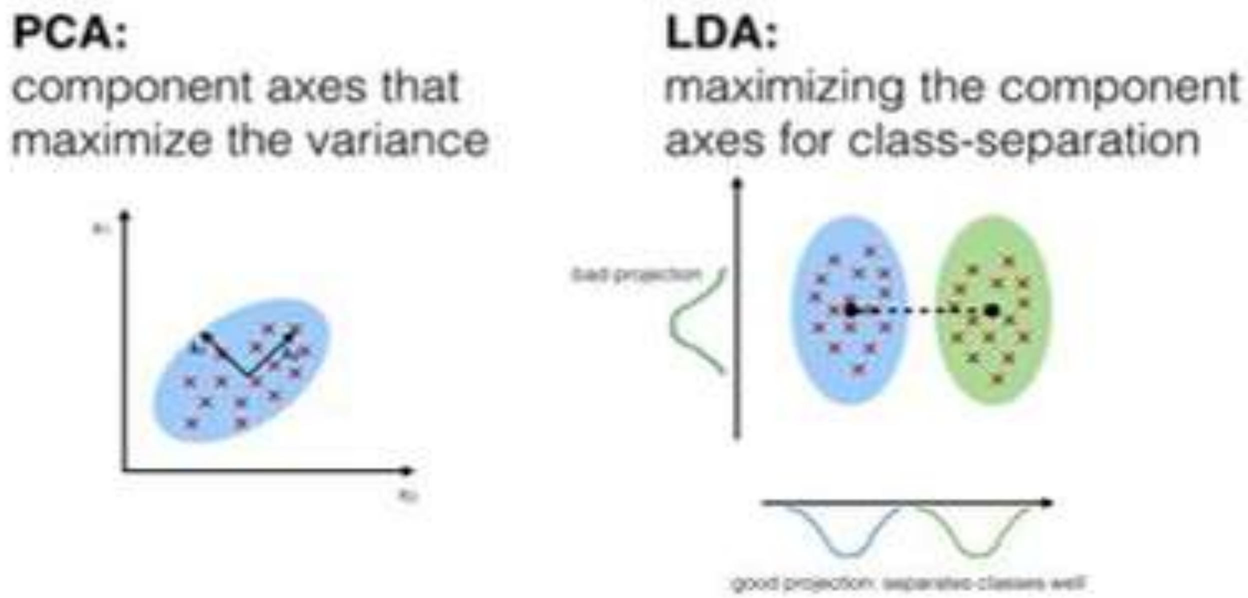


Principal Component Analysis

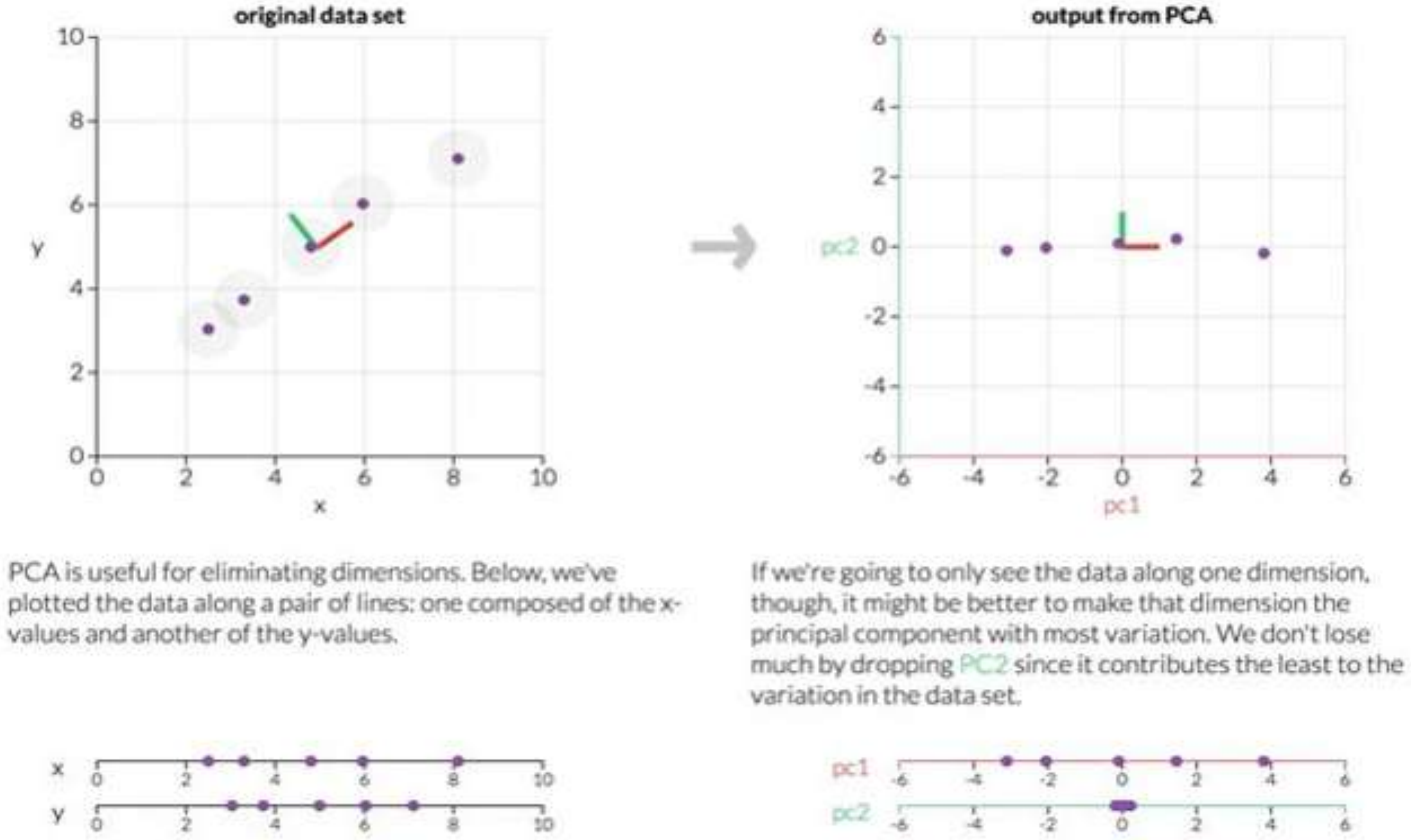
PCA rotates and projects data along the direction of increasing variance.

Used for continuous data

Principal components → features with maximum variance



PCA


original data set
output from PCA
x
y
pc1
pc2
PCA is useful for eliminating dimensions. Below, we've plotted the data along a pair of lines: one composed of the x-values and another of the y-values.
If we're going to only see the data along one dimension, though, it might be better to make that dimension the principal component with most variation. We don't lose much by dropping PC2 since it contributes the least to the variation in the data set.

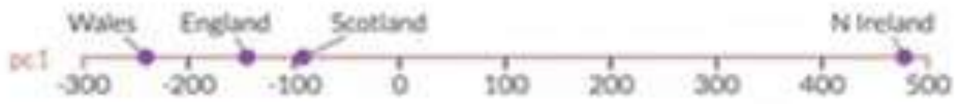
17 Dimension Example

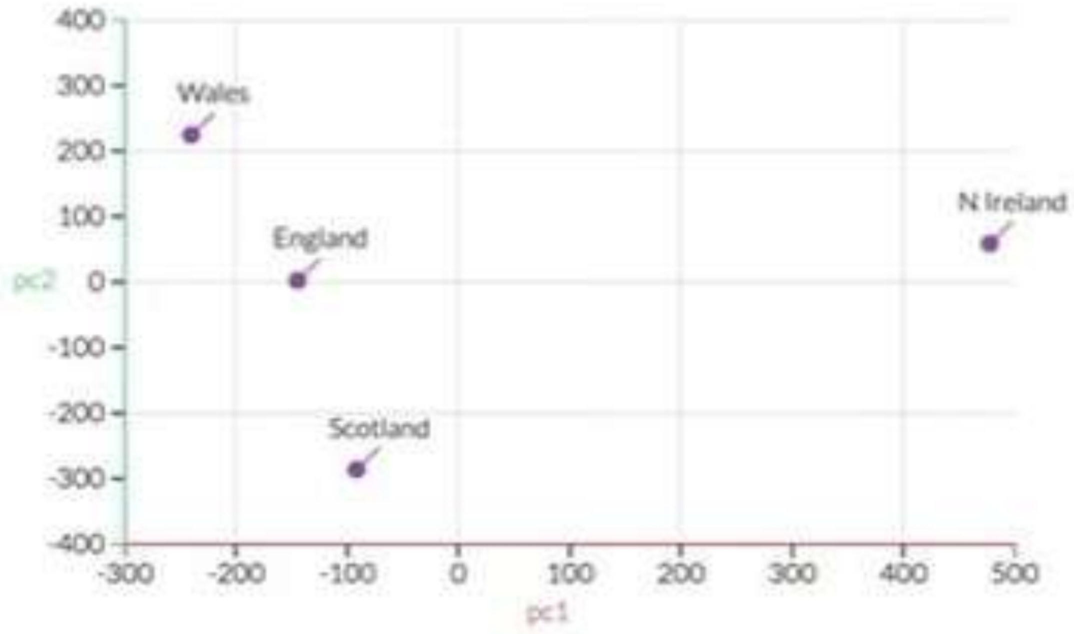
Data on average consumption of 17 types of food in grams per person per week for every country in the UK.

	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

Example (Cont'd)

Here's the plot of the data along the first principal component. Already we can see something is different about Northern Ireland.

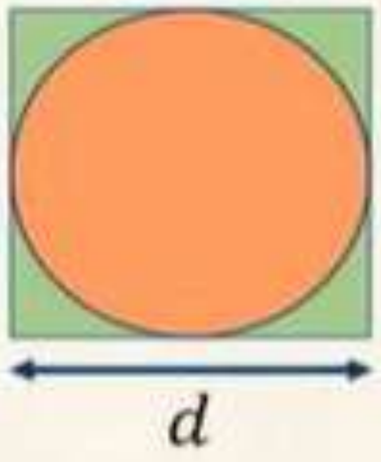




Curse of Dimensionality

Given the dimension D we are interested in the ratio

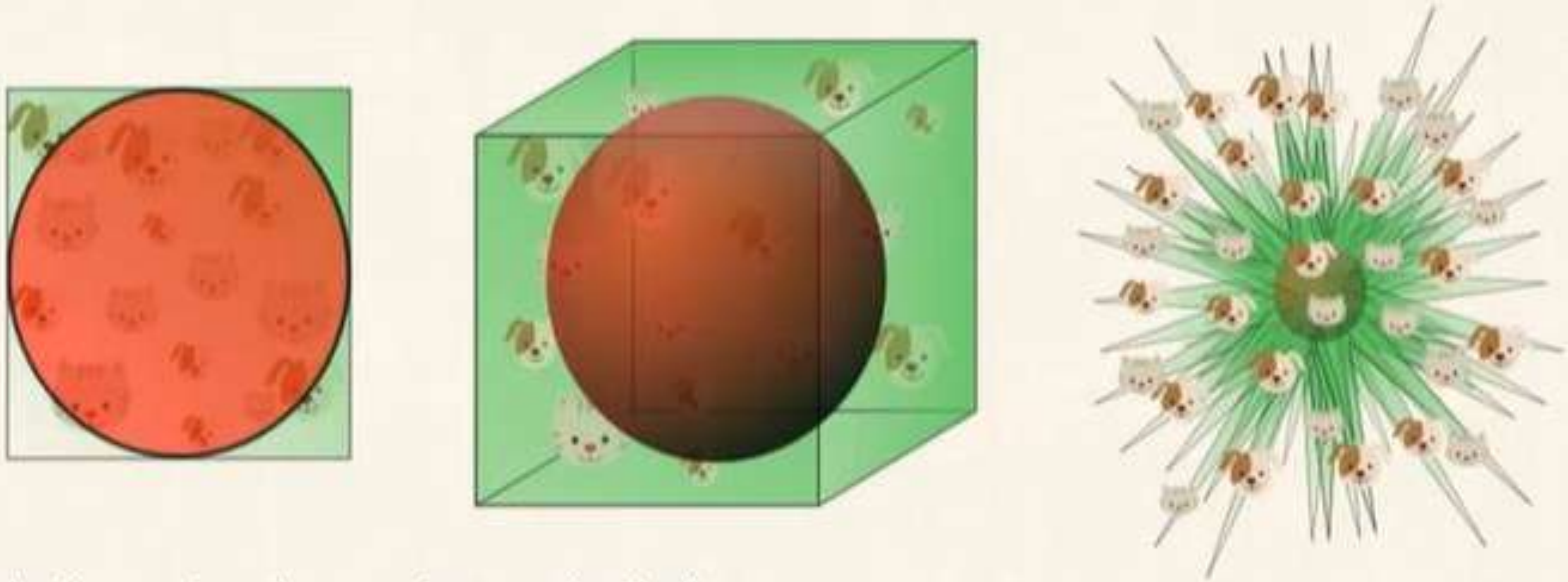
$$\frac{\text{vol. of hypersphere}}{\text{vol. of bounding hypercube}}$$



- $D = 2 \quad \frac{\frac{1}{4}\pi d^2}{d^2} = \frac{1}{4}\pi = 0.785$
- $D = 3 \quad \frac{\frac{1}{6}\pi d^3}{d^3} = \frac{1}{6}\pi = 0.524$
- $D = 4 \quad \frac{\frac{1}{32}\pi^2 d^4}{d^4} = \frac{1}{32}\pi^2 = 0.308$
- $D = 5 \quad \frac{\frac{1}{60}\pi^2 d^5}{d^5} = \frac{1}{60}\pi^2 = 0.164$

} decreases rapidly with increasing no. of dims

Curse of Dimensionality



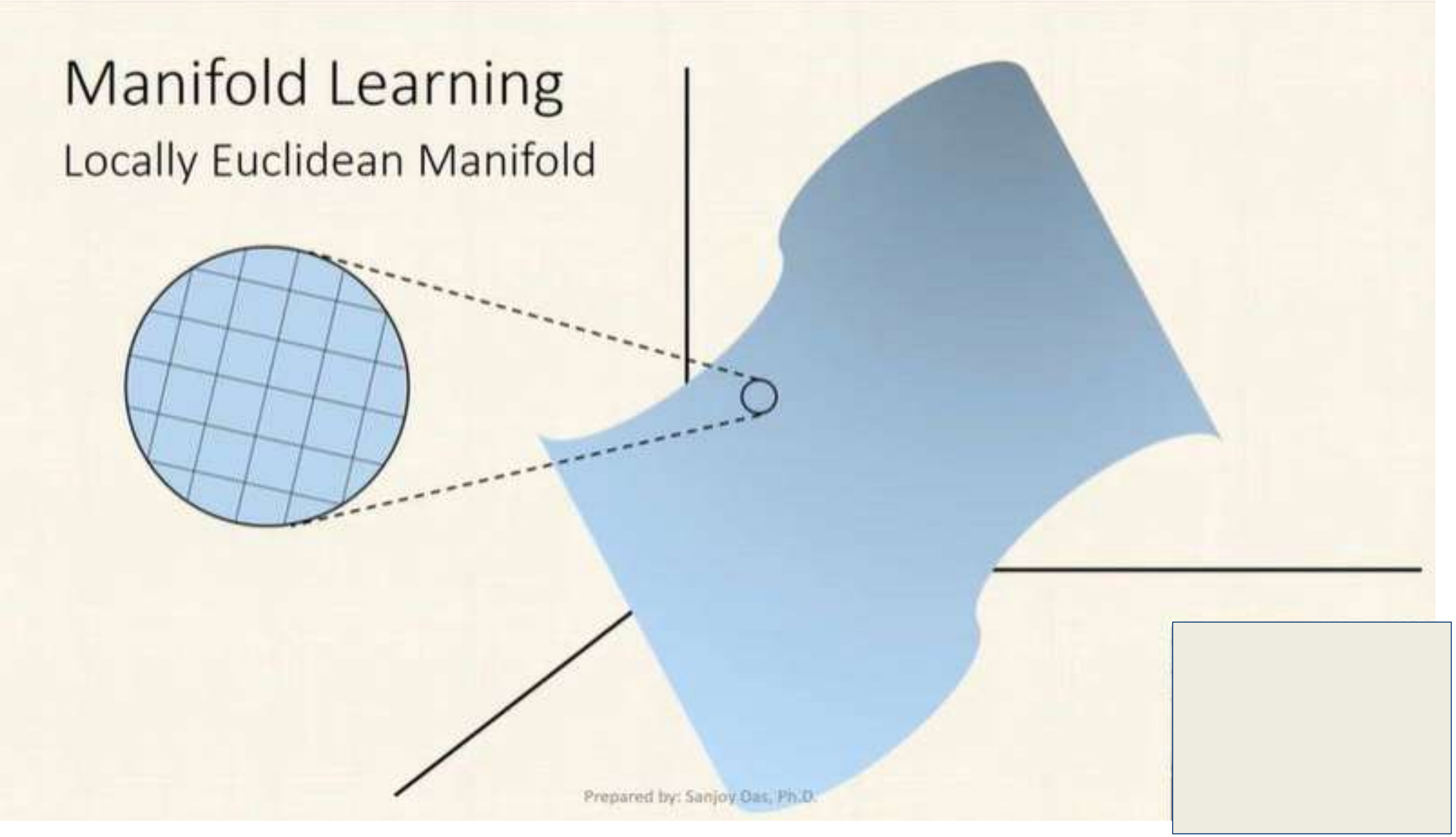
High dimensional space is very "spiky"

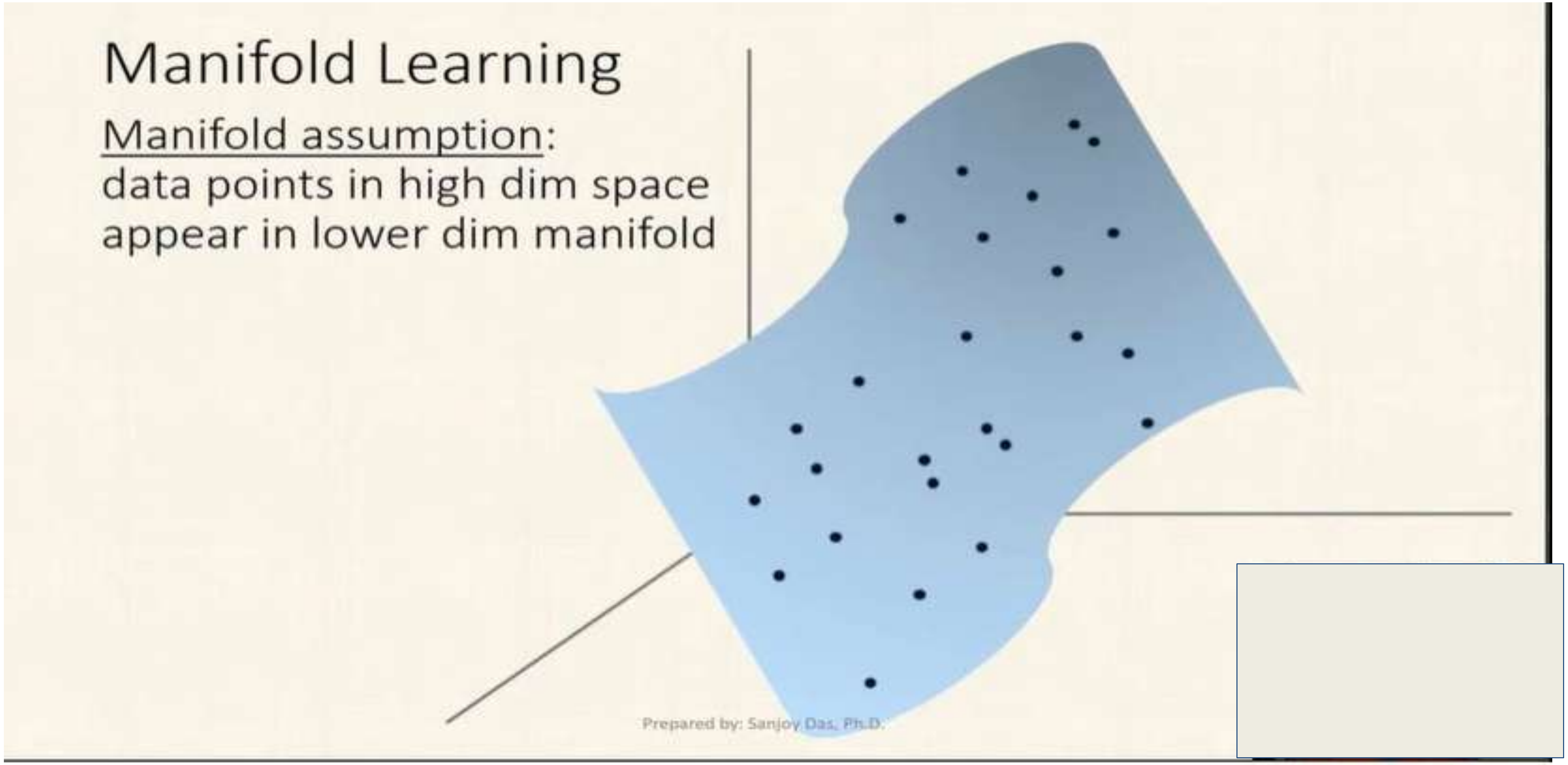
Most of the data points are <u>outliers</u> (i.e. not inside hypersphere)

Curse of Dimensionality

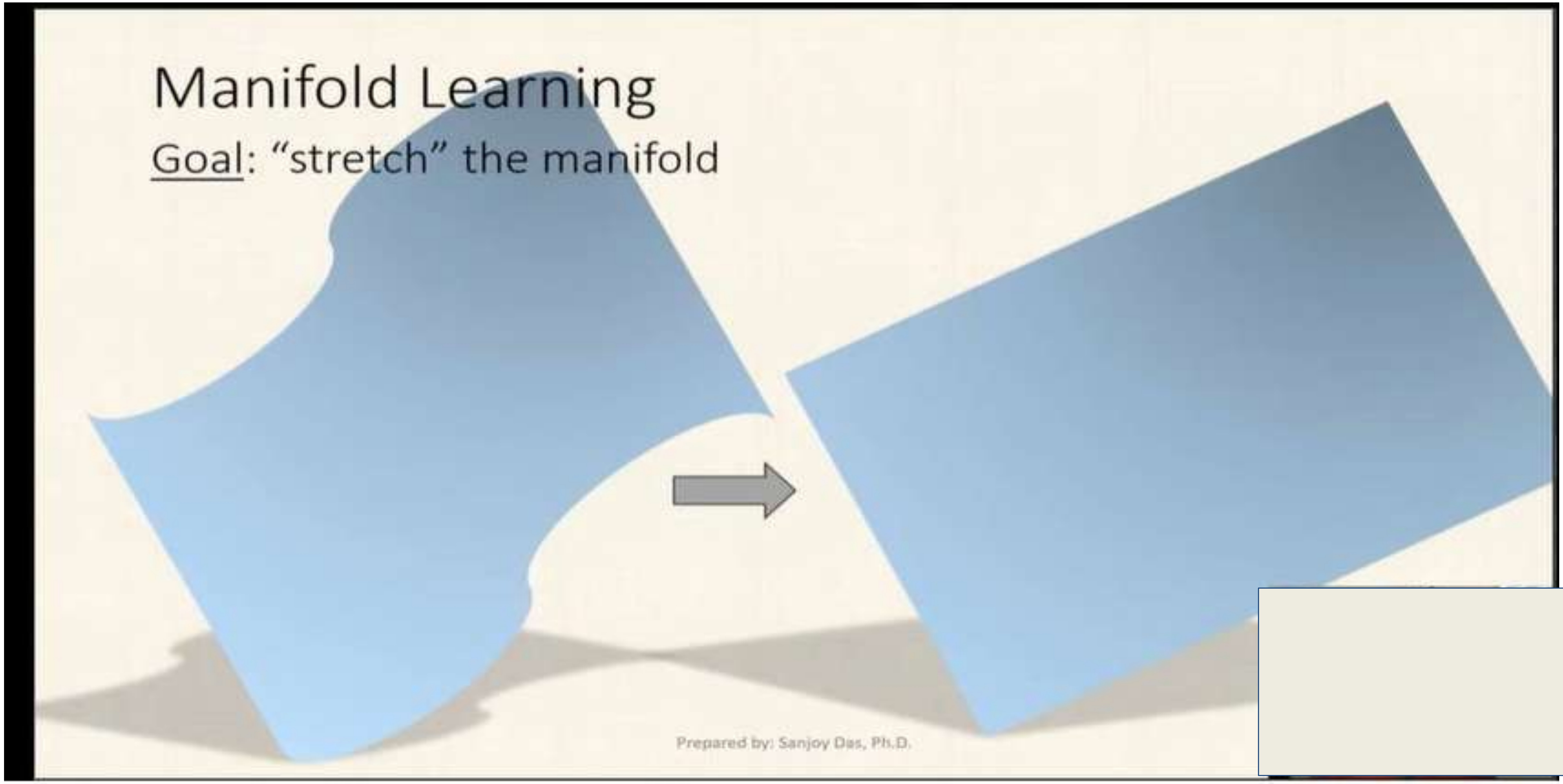
- High dimensional data is very difficult to handle
- Difficulty increases rapidly with number of dimensions
- Need to transform high dimensional data into low dimensional data
    - Dimension reduction is needed to make data more tractable
- Linear methods (classical):
    - PCA, LDA, MDS
- Nonlinear methods (manifold learning):
    - LLE, ISOMAP, Laplacian Eigenmaps, MVU, LTSA, etc.

Linear (PCA, LDA) and manifolds




Manifold Learning
Manifold assumption:
data points in high dim space
appear in lower dim manifold
Prepared by: Sanjoy Das, Ph.D.

Linear (PCA, LDA) and manifolds



Manifold Learning

Real example

Metric learning is a machine learning technique that can be used in deep learning to establish the similarity or dissimilarity between objects. It can be used to perform tasks like clustering, information retrieval, and k-NN classification.

Metric learning aims to:

- Reduce the distance between similar objects
- Increase the distance between dissimilar objects
- Learn a representation function that maps objects into an embedded space

In metric learning, a distance metric is learned over objects, which means that a model can be trained to provide a number for any pair of objects. This number represents the degree of similarity between the objects.

Activities:

1. **PCA**: Take a dataset (e.g., Iris), reduce dimensions to 2 or 3, and visualize clusters.
2. **LDA**: Train an LDA classifier using a labeled dataset (e.g., Iris with target labels) and test accuracy on unseen data.
3. **Manifold Learning**: Apply t-SNE or Isomap to MNIST digits, then visualize the results in 2D to identify clusters.

THANK YOU