# Dimensionality Reduction

Lesson

AI Academy

# Data Preprocessing

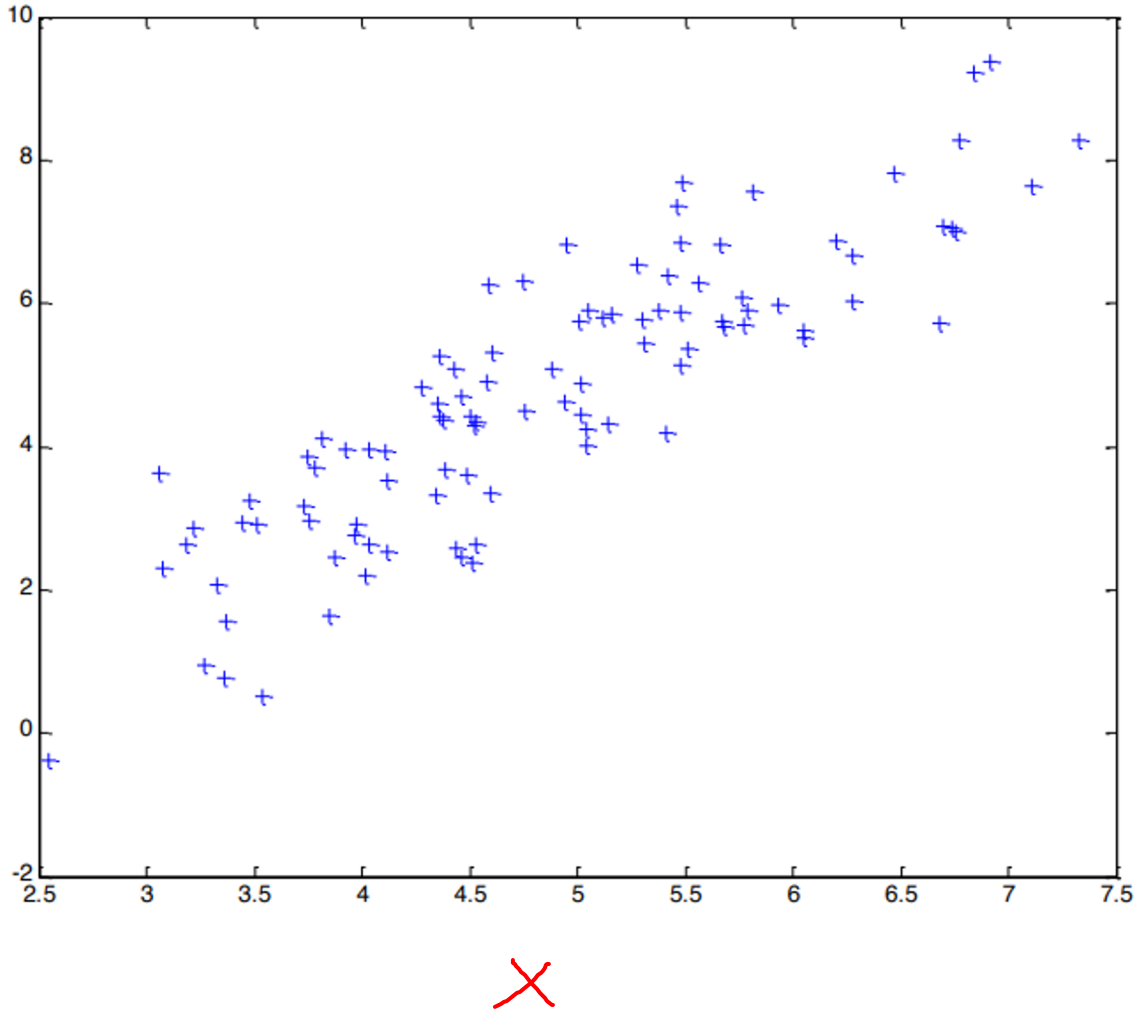
- Sampling
- Feature subset selection
- **Dimensionality Reduction**
- Feature creation
- Discretization and binarization
- Attribute Transformation

# Dimensionality

- **Dimensionality:** The number of attributes in a dataset.

**Example**

2-dimensional (2D) data

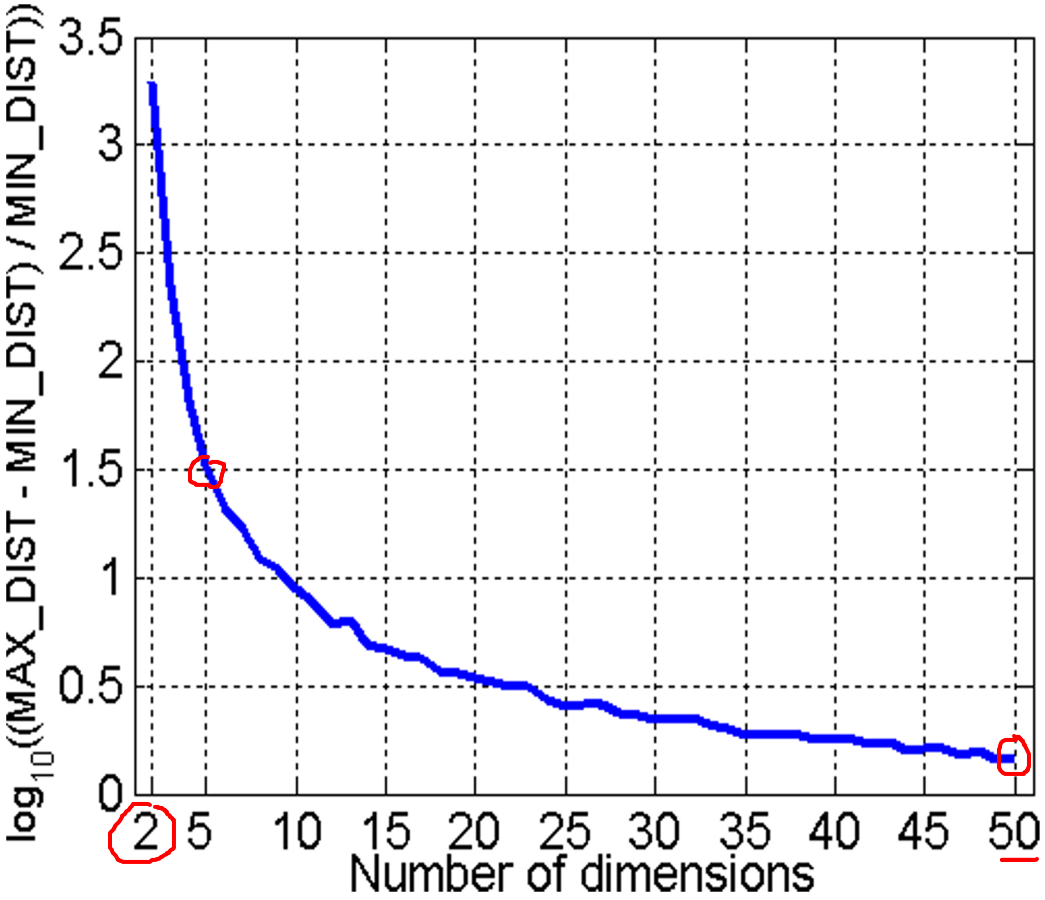


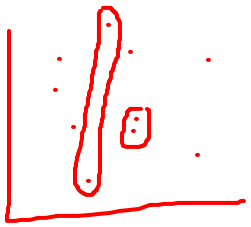
# Curse of Dimensionality

- As **dimensionality increases**, data becomes **more sparse**
- Sparse data is more **difficult to cluster and meaningfully measure** distance

**Example**

- Randomly generate 500 points
- Computer difference between max and min distance between any pair of points



# Dimensionality Reduction

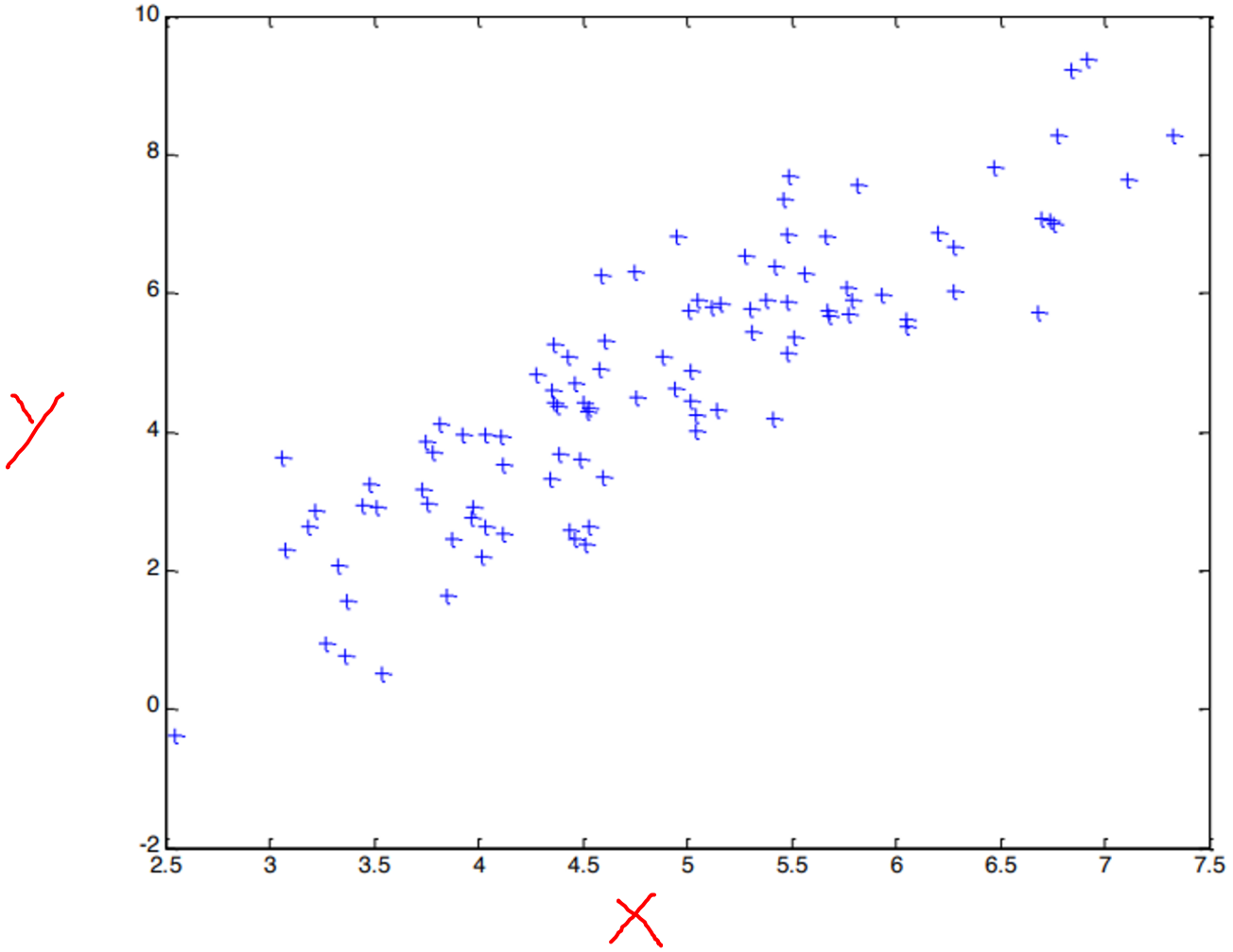
**Purposes:**

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Allow data to be more easily visualized
- Maybe help eliminate irrelevant features or reduce noise

**Techniques:**

- **Principle Component Analysis** PCA
- Singular value Decomposition
- Others: supervised and non-linear techniques

# 2D Data

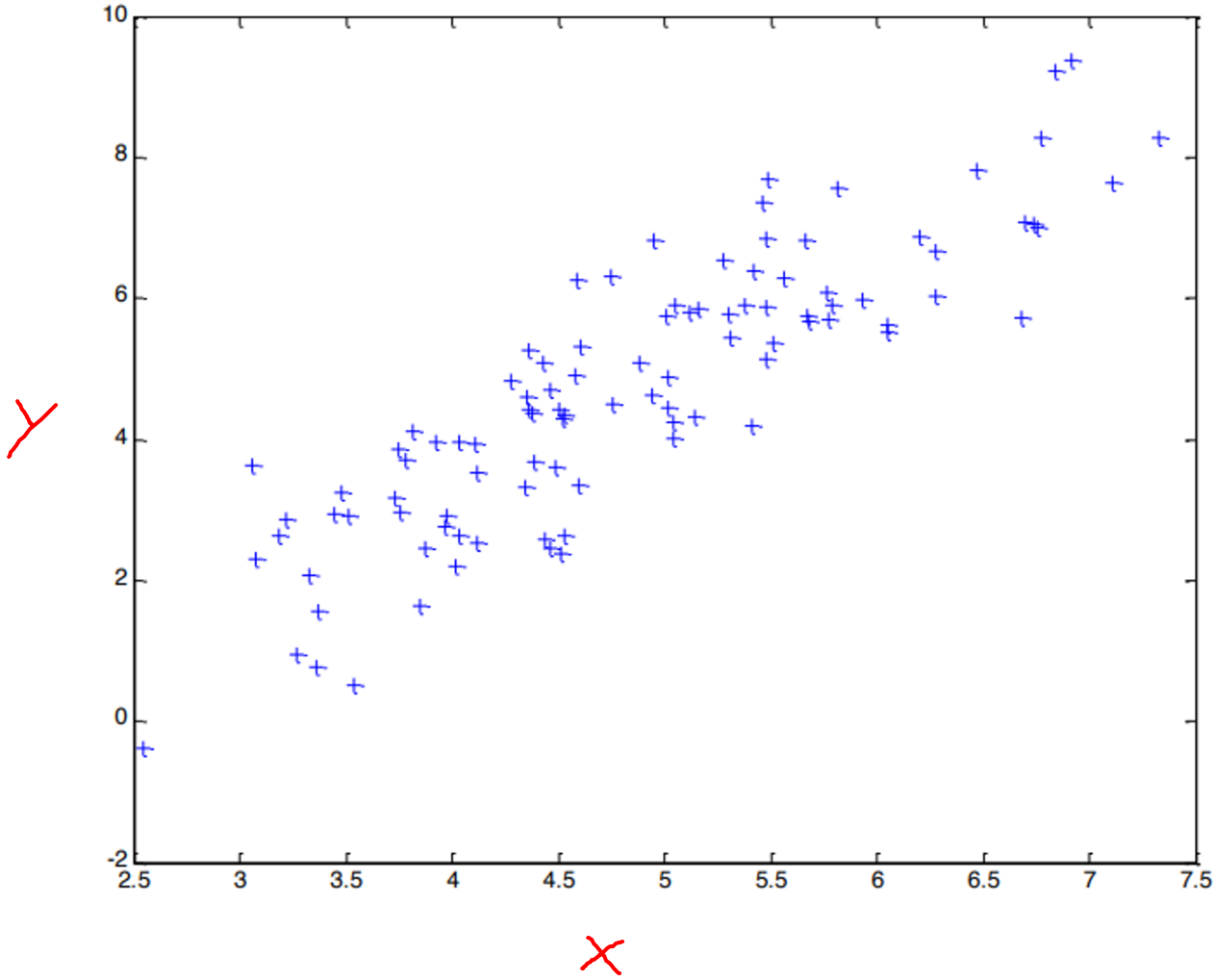


# Dimensionality Reduction: PCA

- Principle Component Analysis
- Goal is to find a projection that captures the largest amount of variation in the data

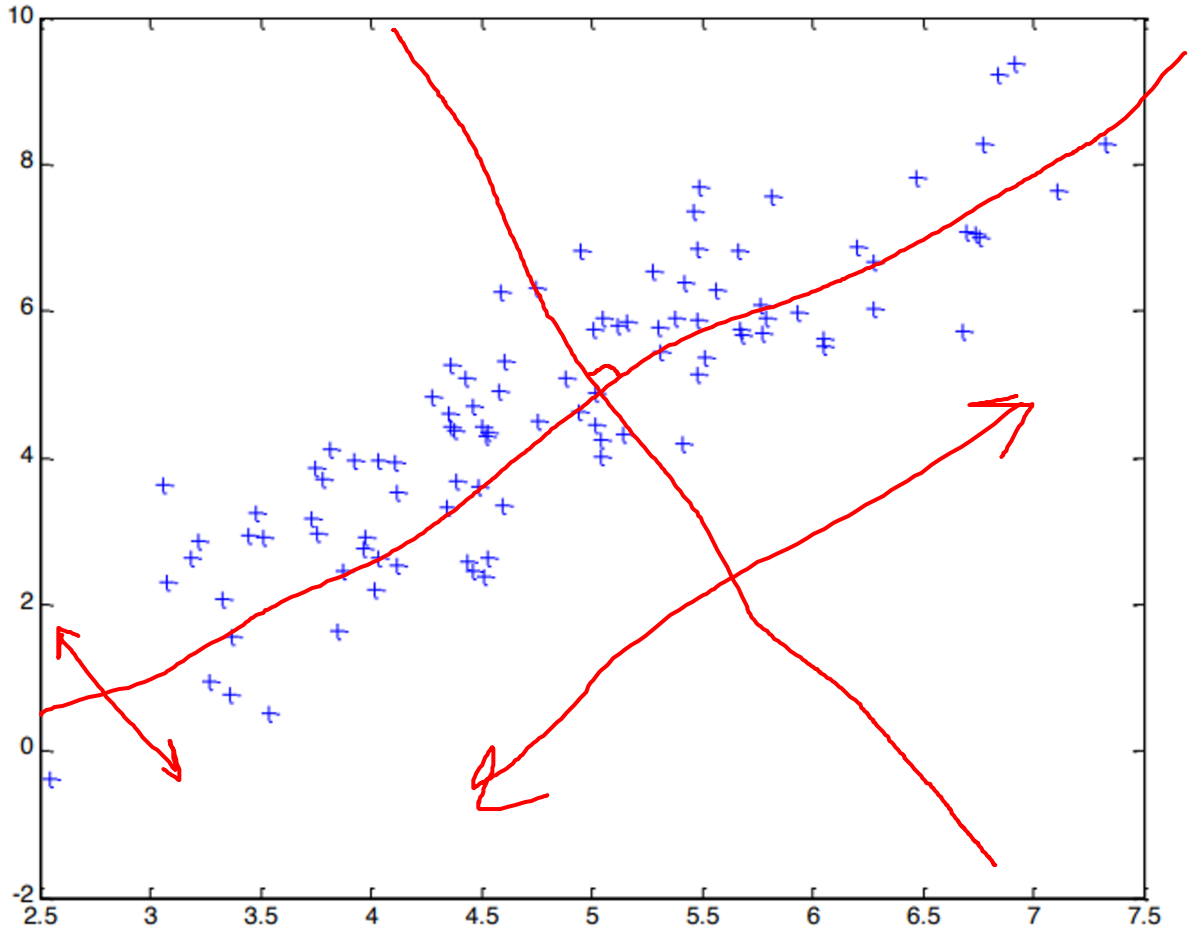
- How can we capture the most variance in the data with just one feature?
  - What about two?

# 2D Data



# 1st Principal Component (PC)

- The 1st PC is a new **axis** that captures the **most variance**.
- The next PCs are **orthogonal**.

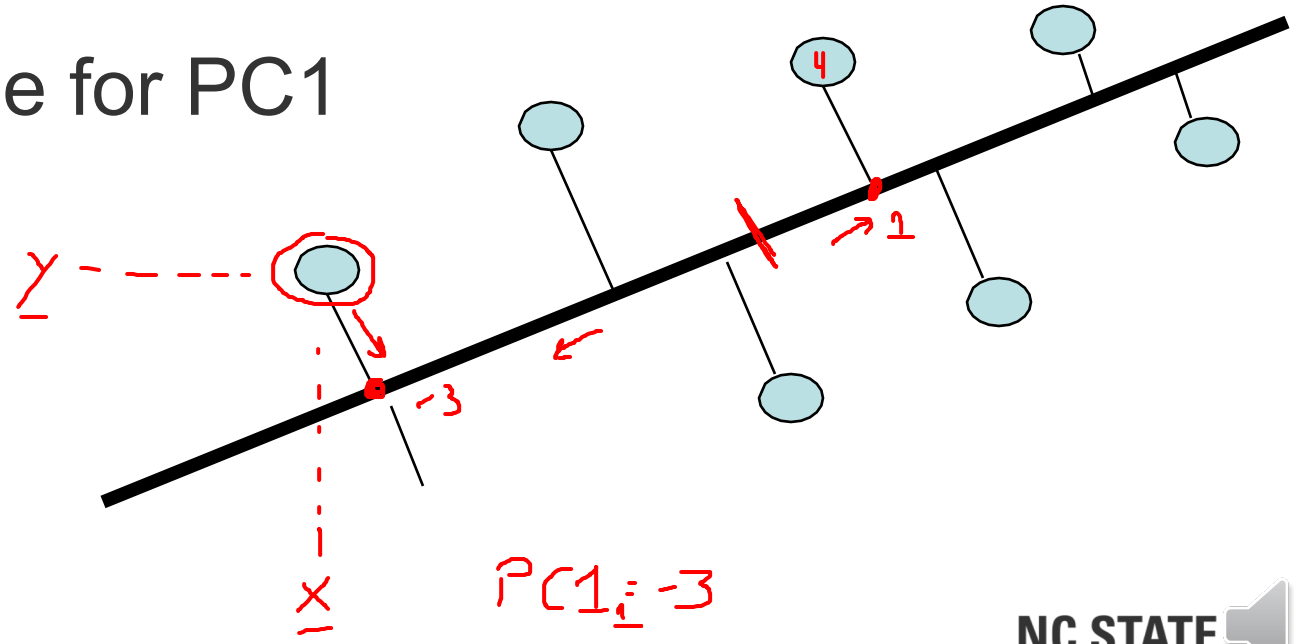


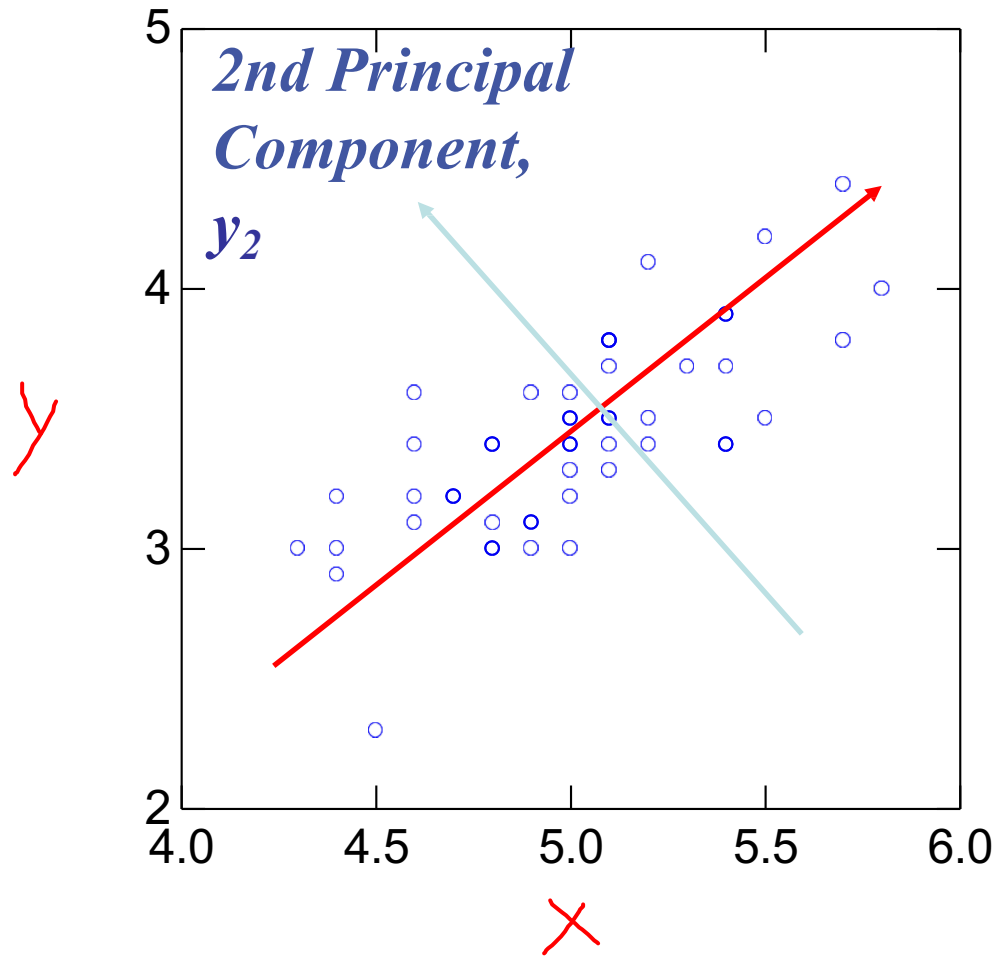
# PC Projection

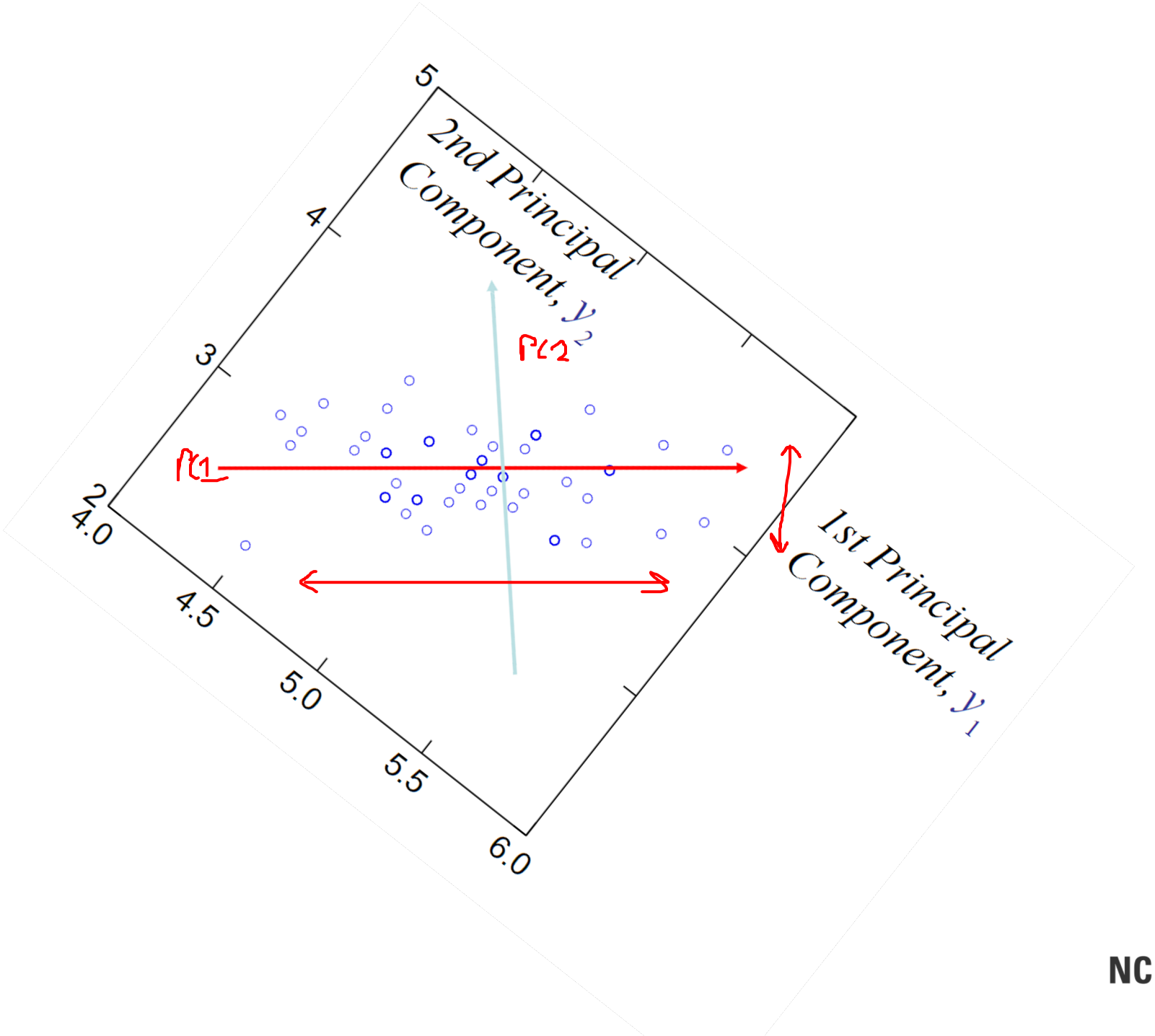
- We then project each data point onto the PC vector to calculate the PC value

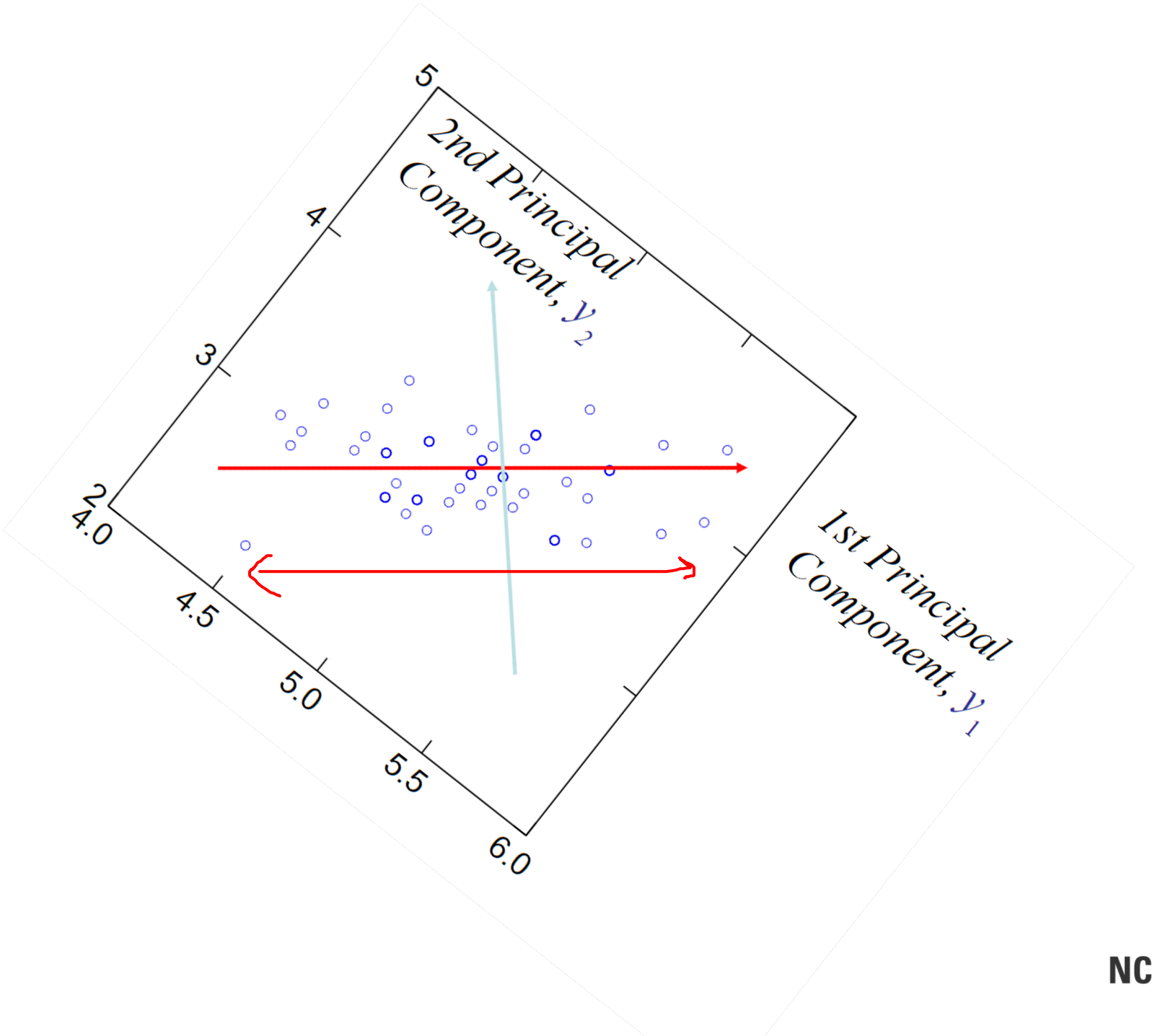
- This is a *new feature* (e.g. PC1)

- Each data point has a value for PC1




2nd Principal Component, $y_2$
1st Principal Component, $y_1$
y
x
5
4
3
2
4.0
4.5
5.0
5.5
6.0


5
4
3
2
4.0
4.5
5.0
5.5
6.0
2nd Principal Component, $y_2$
1st Principal Component, $y_1$
PC2
PC1

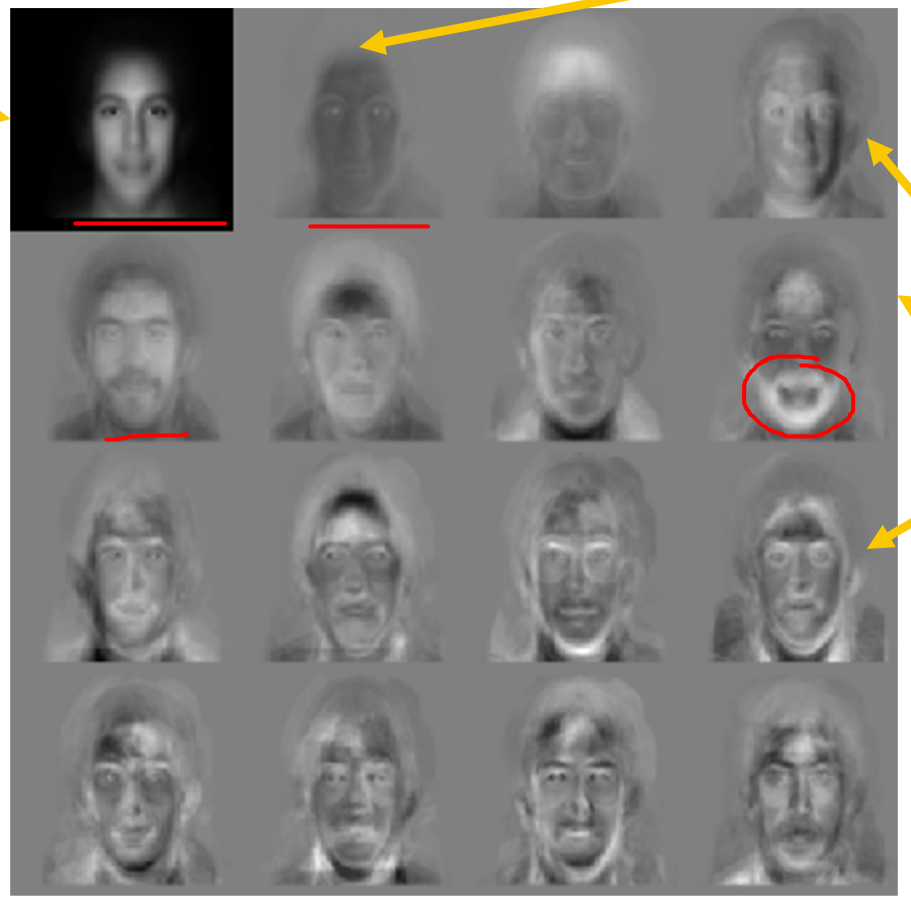

5
4
3
2
4.0
4.5
5.0
5.5
6.0
2nd Principal Component, $y_2$
1st Principal Component, $y_1$

# PCA on Faces: “Eigenfaces”



Average Face

First principal component



Other components

# PCA for Relighting

Images under different illumination

[Matusik & McMillan]

# PCA for Relighting

- Images under different illumination
- Most variation capture by first 5 principal components
- Can re-illuminate by combining only a few images


[Matusik & McMillan]

# Learning Objectives: Dimensionality Reduction

## You now should be able to:

- Explain the challenges of high-dimensional data
- Identify when dimensionality reduction is useful

AI Academy
NC STATE