8 Machine Learning

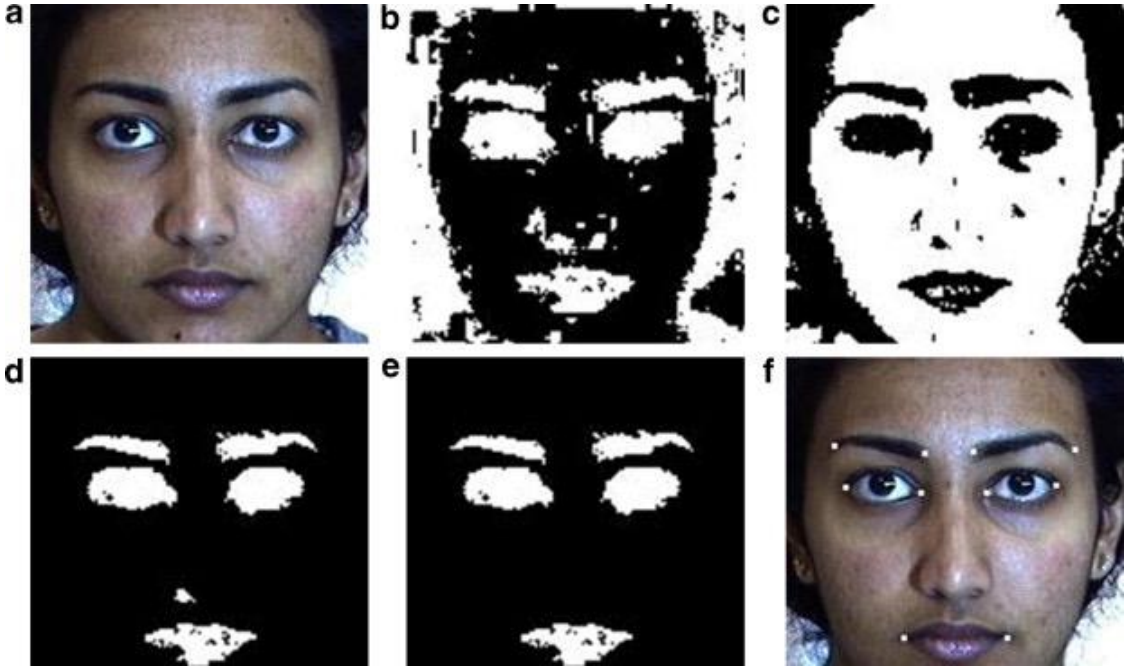
- Unsupervised Learning
  - Dimensionality Reduction
    - Principal Component Analysis (PCA)

Prof. Umut ORHAN

Some Problems in Unsupervised Learning

In the ever-growing field of data science, while this learning approach saves humans from the difficult task of labeling datasets that have hundreds or even thousands of features, it brings some challenges:

1. Curse of Dimensionality
2. Computational Efficiency
3. Noise Reduction
4. Improved Visualization



PCA

vs

Feature Selection & Feature Extraction

PCA vs.
Feature Selection & Feature Extraction

- Independence from Expertise
- Versatility
- Efficiency
- Variance Maximization
- Noise Reduction

University of California Machine Learning Repository			
Name	Instances	Attributes	Year
URL Reputation	2.396.130	3.231.961	2009
Gas sensor arrays in open sampling settings	18.000	1.950.000	2013
YouTube Multiview Video Games Dataset	120.000	1.000.000	2013
Twin gas sensor arrays	640	480.000	2016
Gas sensor array exposed to turbulent gas mixtures	180	150.000	2014
ElectricityLoadDiagrams	370	140.256	2015
PEMS-SF	440	138.672	2011
Gas sensor array under flow modulation	58	120.432	2014
Bag of Words	8.000.000	100.000	2008
Dorothea	1.950	100.000	2008
Farm Ads	4.143	54.877	2011
Dexter	2.600	20.000	2008

How Does PCA Work?

- Standardization
- Covariance Matrix Calculation
- Eigenvalue and Eigenvector
- Sorting and Selection
- Projection

$$Z = \frac{X - \mu}{\sigma}$$

Standardization

$$C = \frac{1}{n-1} Z^T Z$$

Covariance Matrix Calculation

$$C\ v = \lambda\ v$$

Eigenvalue and Eigenvector

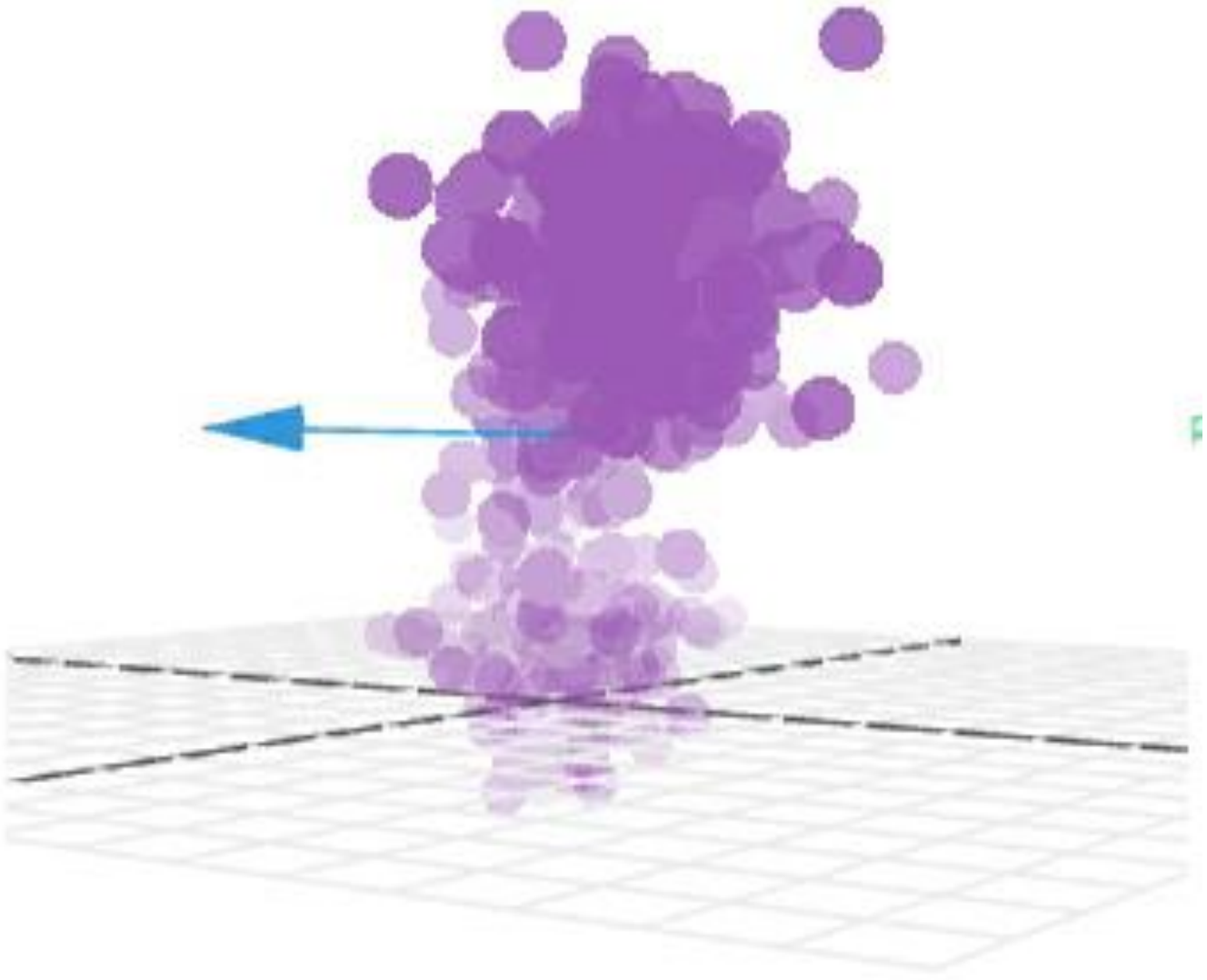
$$V = \underset{v_i : 1 \le i \le k}{arg\max}(\lambda_i)$$

Sorting and Selection

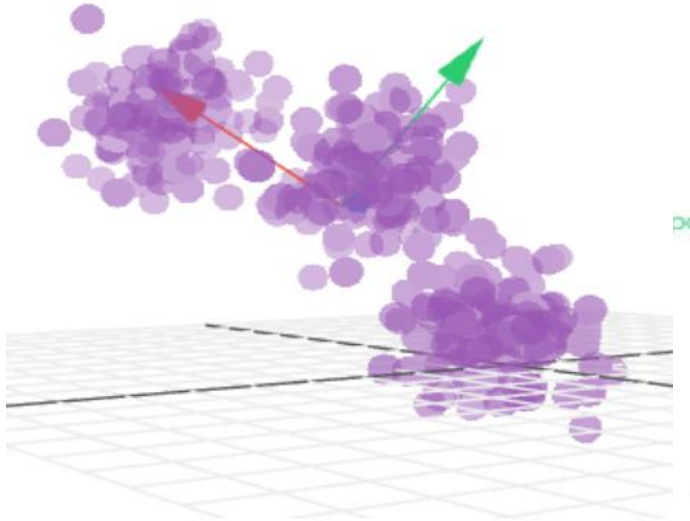
$$Y = Z\ V$$

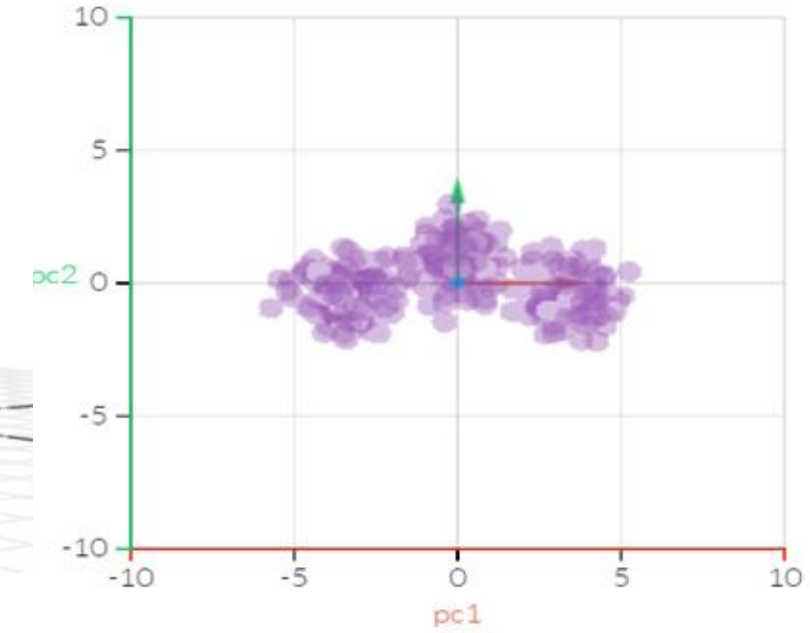
Projection

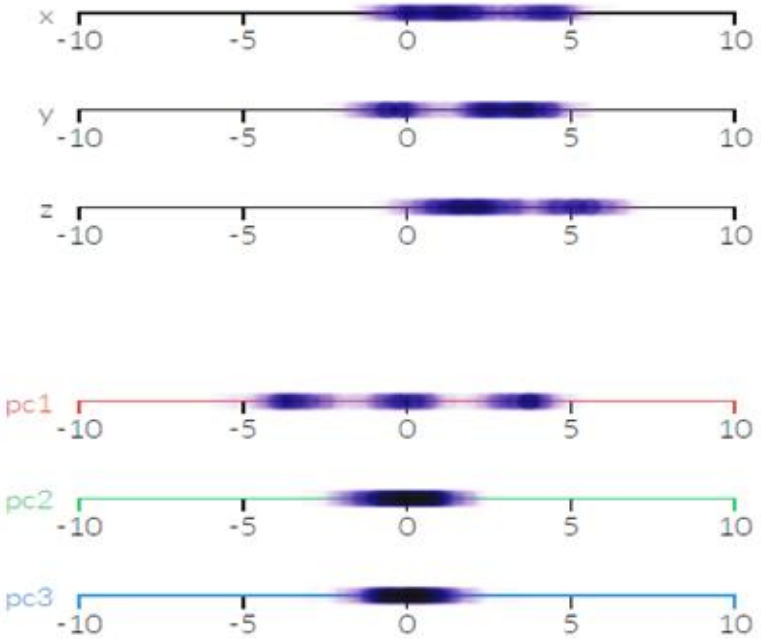
What is Principal Components?

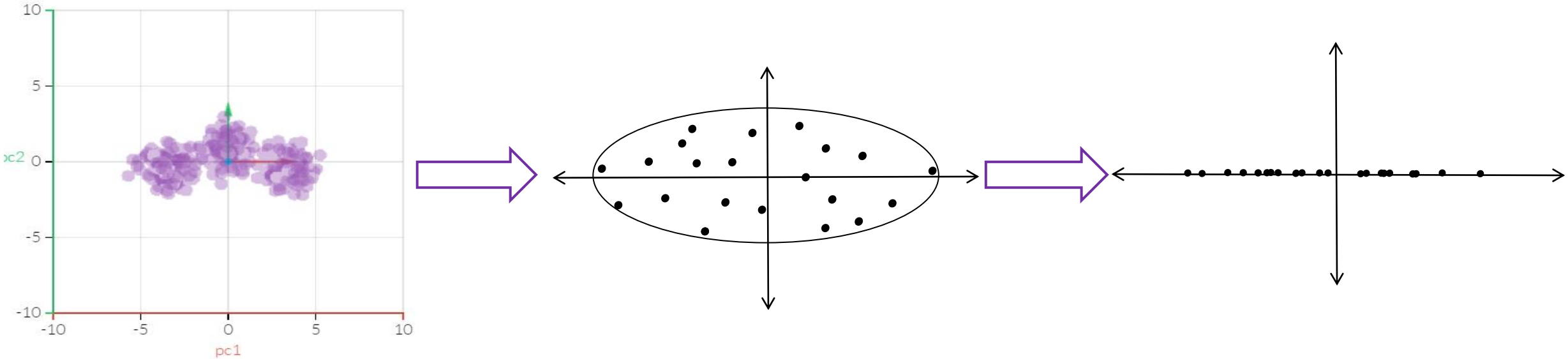


Choosing Principal Component





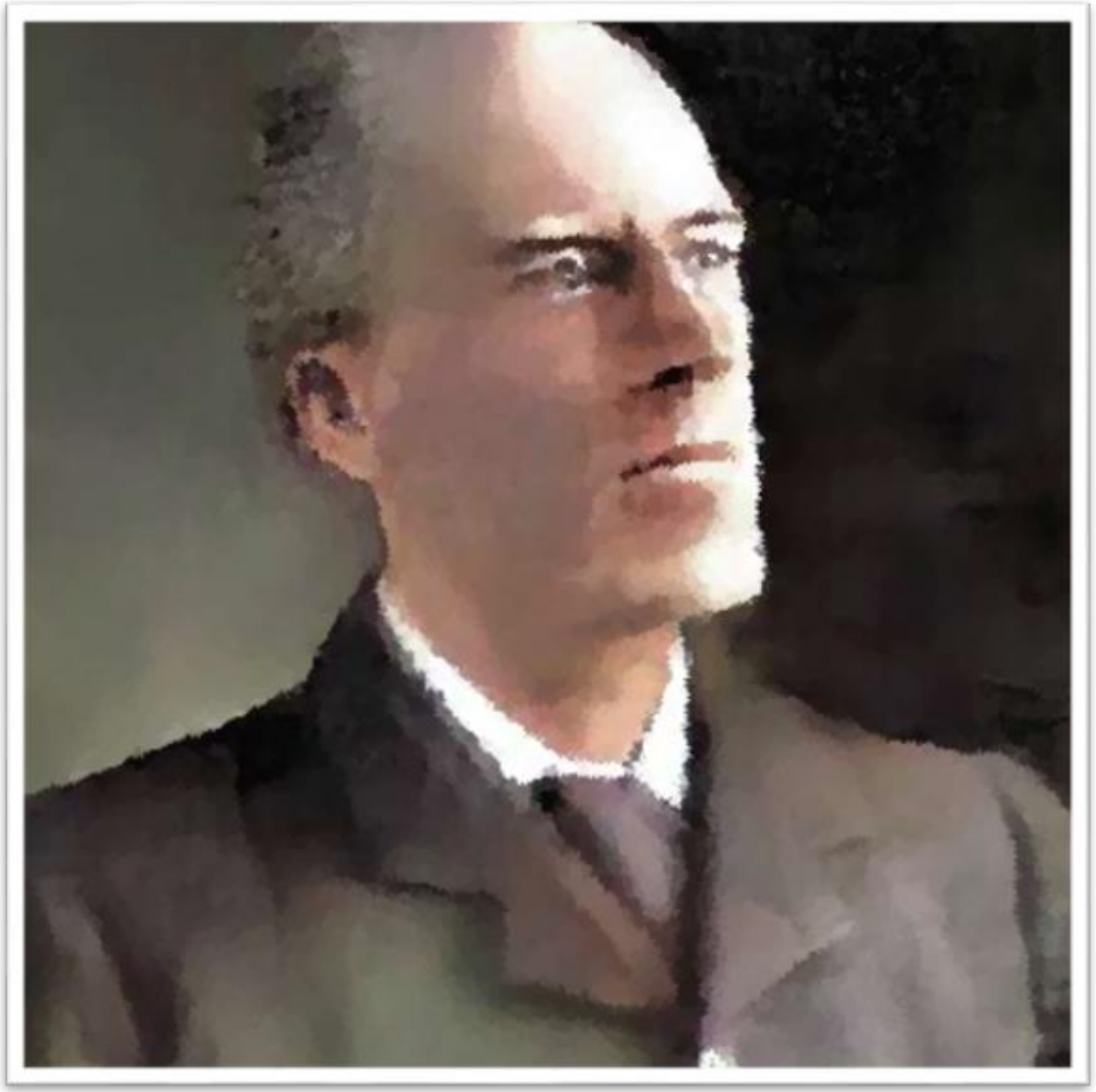




Dimensionality Reduction

Karl Pearson

1857-1936



Benefits of PCA

REDUCES COMPLEXITY

ENHANCES PERFORMANCE

REVEALS PATTERNS

Disadvantages of PCA

1. Linear Assumption
2. Interpretability of Features
3. Data Scaling Requirement
4. Information Loss
5. Sensitivity

Machine Learning

8. week

Thanks for watching