UNC CHARLOTTE

The WILLIAM STATES LEE COLLEGE *of* ENGINEERING

Introduction to ML
Lecture 11: Dimension Reduction

Hamed Tabkhi

Department of Electrical and Computer Engineering,
University of North Carolina Charlotte (UNCC)
htabkhiv@uncc.edu

UNC CHARLOTTE

Dimensionality Reduction

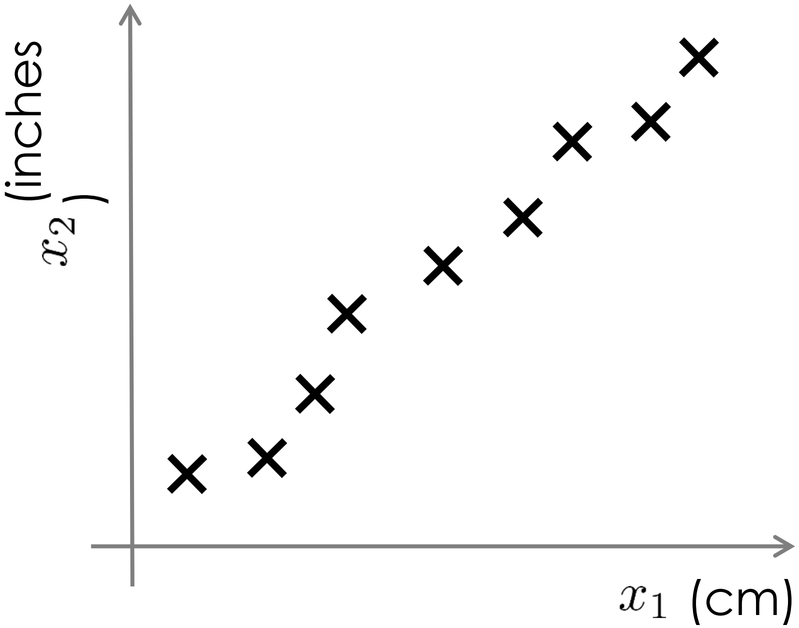
Why Dimensionality Reduction?

- It is so easy and convenient to collect data
- Data accumulates in an unprecedented speed
- Data preprocessing is an important part for *effective* machine learning and data mining
- Dimensionality reduction is an effective approach to downsizing data

Why Dimensionality Reduction?

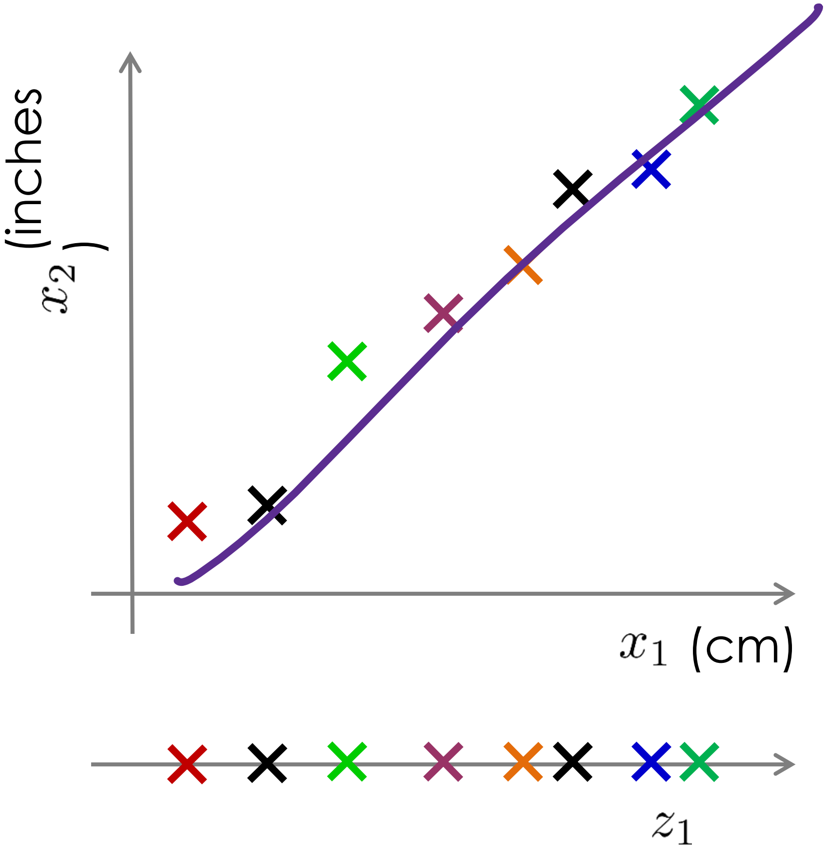
- Visualization: projection of high-dimensional data onto 2D or 3D.
- Data compression: efficient storage and retrieval.
- Noise removal: positive effect on query accuracy.

Data Compression



Reduce data from 2D to 1D

Data Compression



Reduce data from 2D to 1D

$$x^{(1)} \rightarrow z^{(1)}$$

$$x^{(2)} \rightarrow z^{(2)}$$

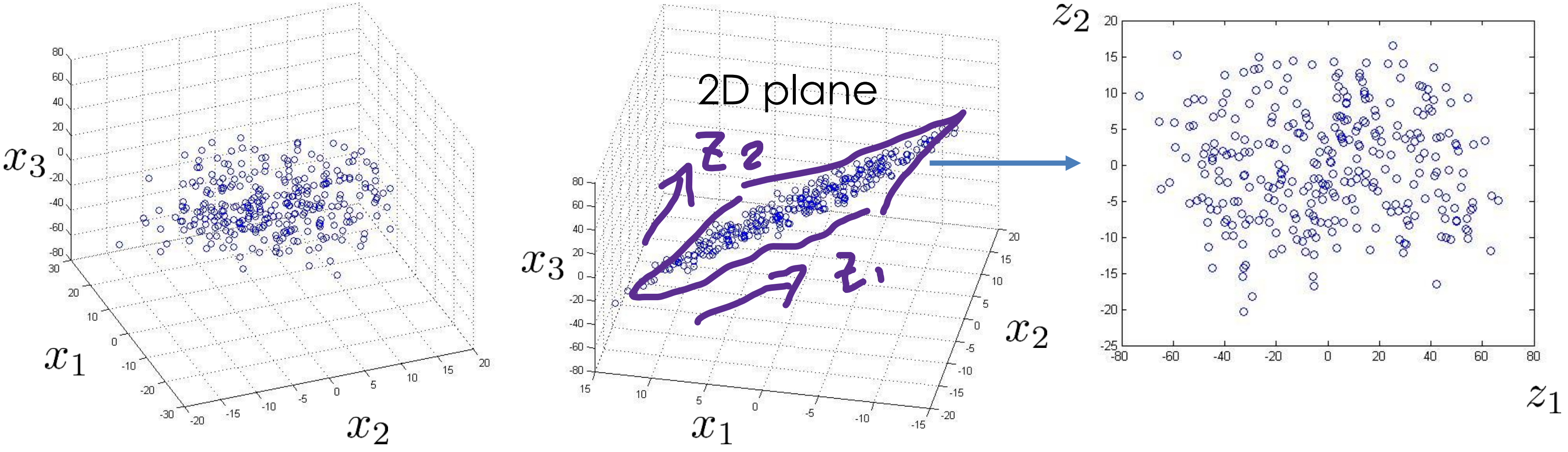
$$\vdots$$

$$x^{(m)} \rightarrow z^{(m)}$$

Andrew N

UNC CHARLOTTE

Reduce data from 3D to 2D



Easy to visualize

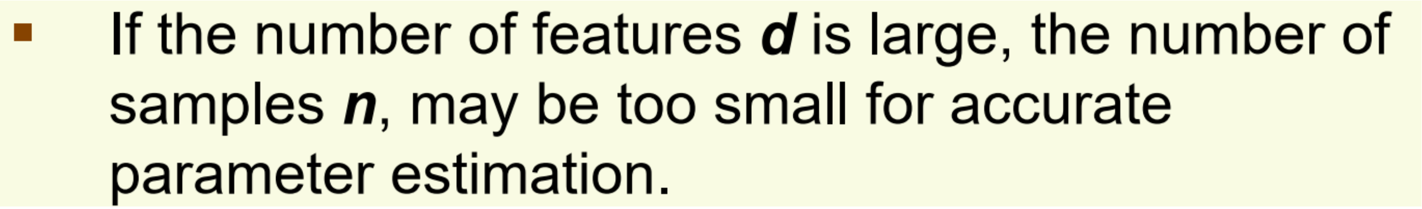
Andrew N

Why Dimensionality Reduction?

- Most machine learning and data mining techniques may not be effective for high-dimensional data
  - Curse of Dimensionality
  - Accuracy and efficiency degrade rapidly as the dimension increases.

- The intrinsic dimension may be small.
  - For example, the number of genes responsible for a certain type of disease may be small.

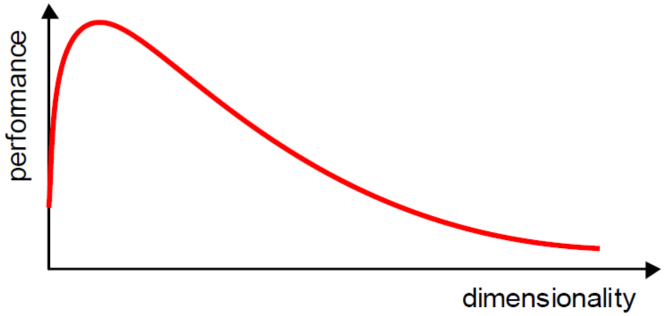
Curse of Dimensionality



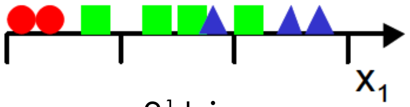
- If the number of features ***d*** is large, the number of samples ***n***, may be too small for accurate parameter estimation.

Curse of Dimensionality

- Increasing the number of features will not always improve classification accuracy.

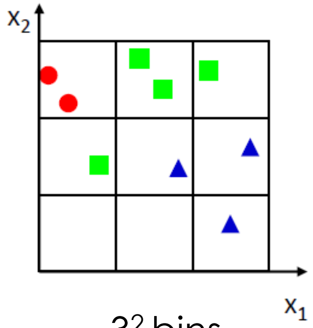


- In practice, the inclusion of more features might actually lead to worse performance.

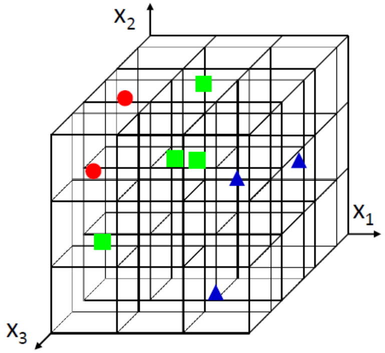


3^1 bins

- The number of training examples required increases exponentially with dimensionality **d** (i.e., k^d).



3^2 bins



3^3 bins

k: number of bins per feature

Slide Credit: George Bebis

Gene Expression Microarray Analysis

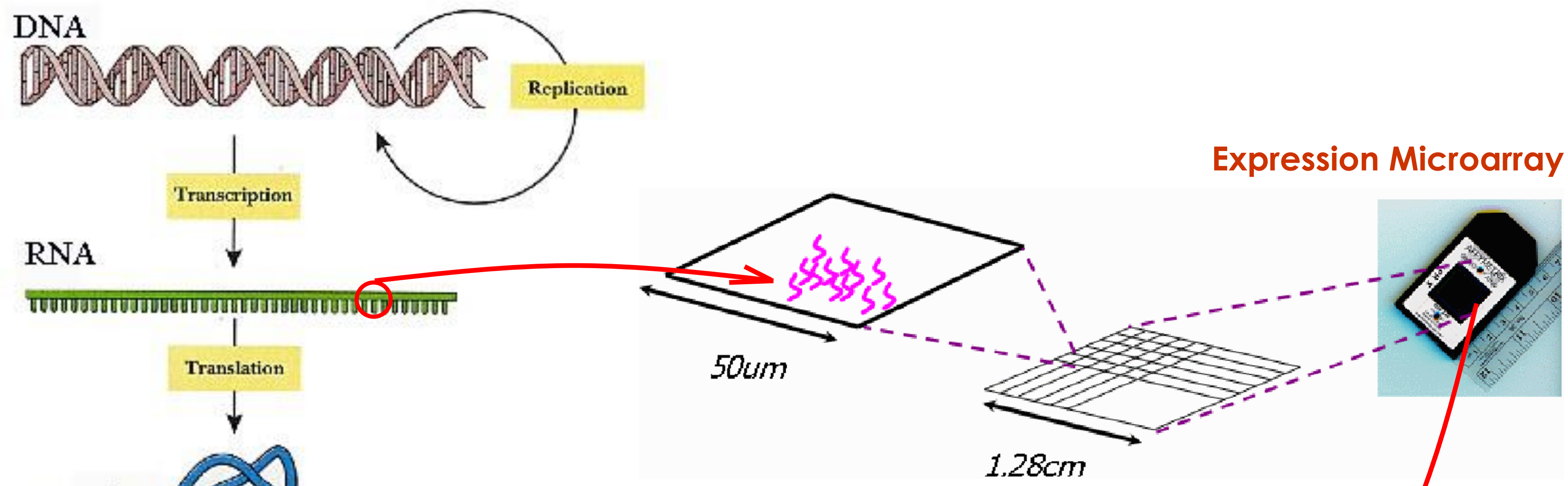
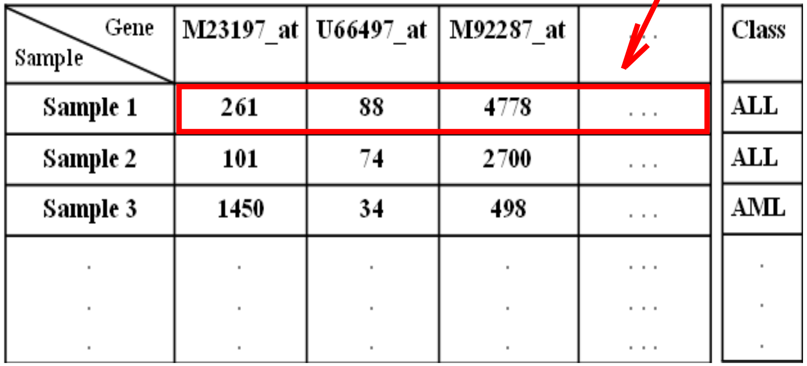


Image Courtesy of Affymetrix

- **Task:** To classify novel samples into known disease types (disease diagnosis)
- **Challenge:** thousands of genes, few samples
- **Solution:** to apply dimensionality reduction

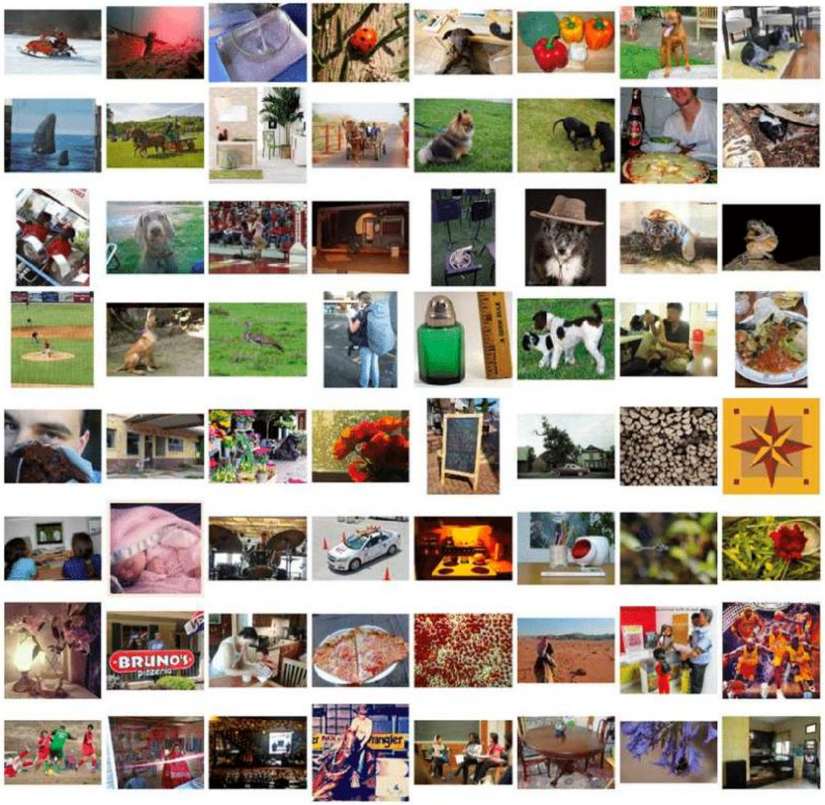


Gene / Sample	M23197_at	U66497_at	M92287_at	. . .	Class
Sample 1	261	88	4778	. . .	ALL
Sample 2	101	74	2700	. . .	ALL
Sample 3	1450	34	498	. . .	AML
.	.	.	.	. . .	.
.	.	.	.	. . .	.
.	.	.	.	. . .	.

Expression Microarray Data Set

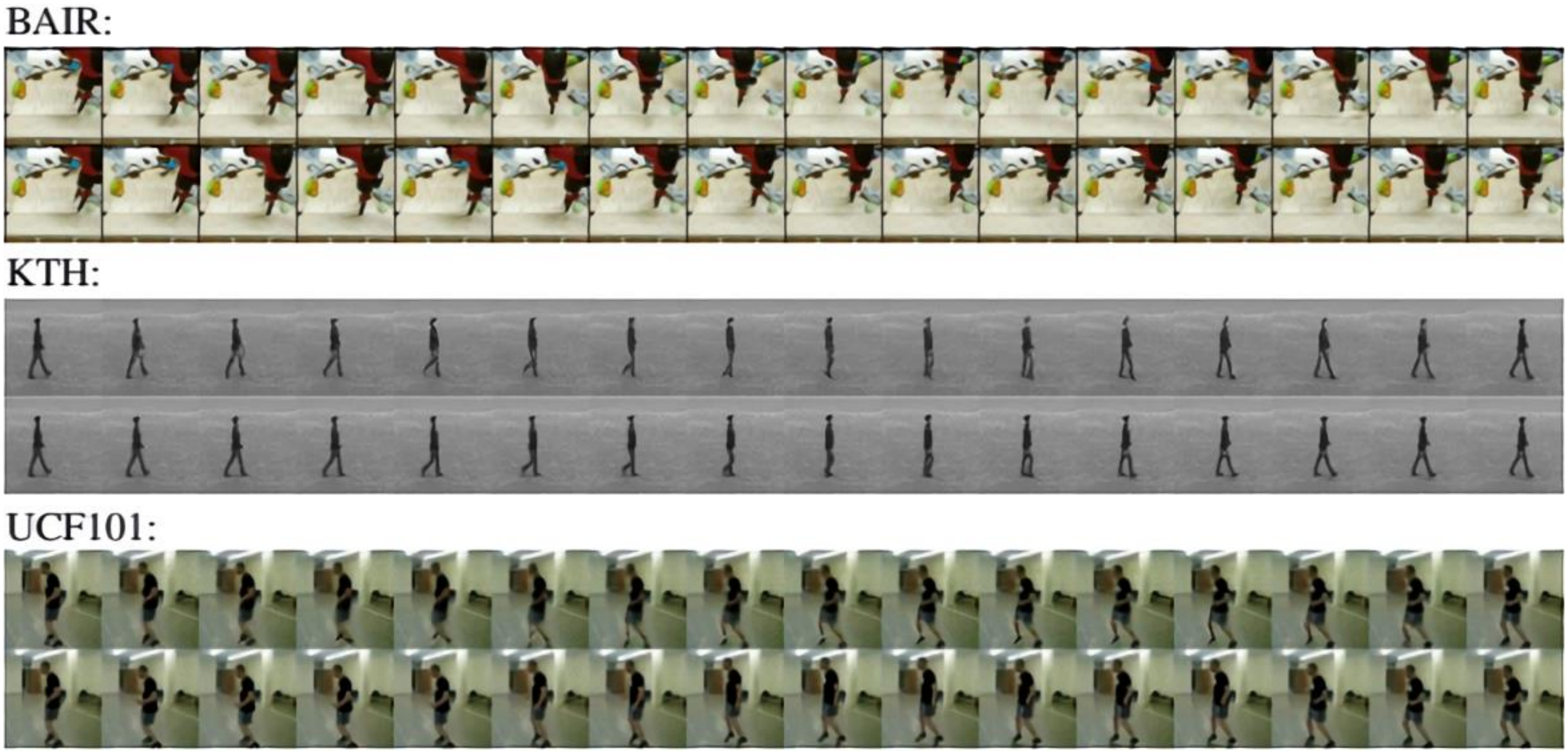
Other Types of High-Dimensional Data

Face images



Natural images

Other Types of High-Dimensional Data



Videos (action recognition)

Major Techniques of Dimensionality Reduction

- Feature selection
- Feature extraction (reduction)

Feature Selection
(a very brief overview)

Feature Selection

- Definition
  - A process that chooses an optimal subset of features according to an objective function

- Objectives
  - To reduce dimensionality and remove noise
  - To improve mining performance
    - Speed of learning
    - Predictive accuracy
    - Simplicity and comprehensibility of mined results

Feature selection

Horse vs. Zebra

Features:

4-leg

Shape

Color

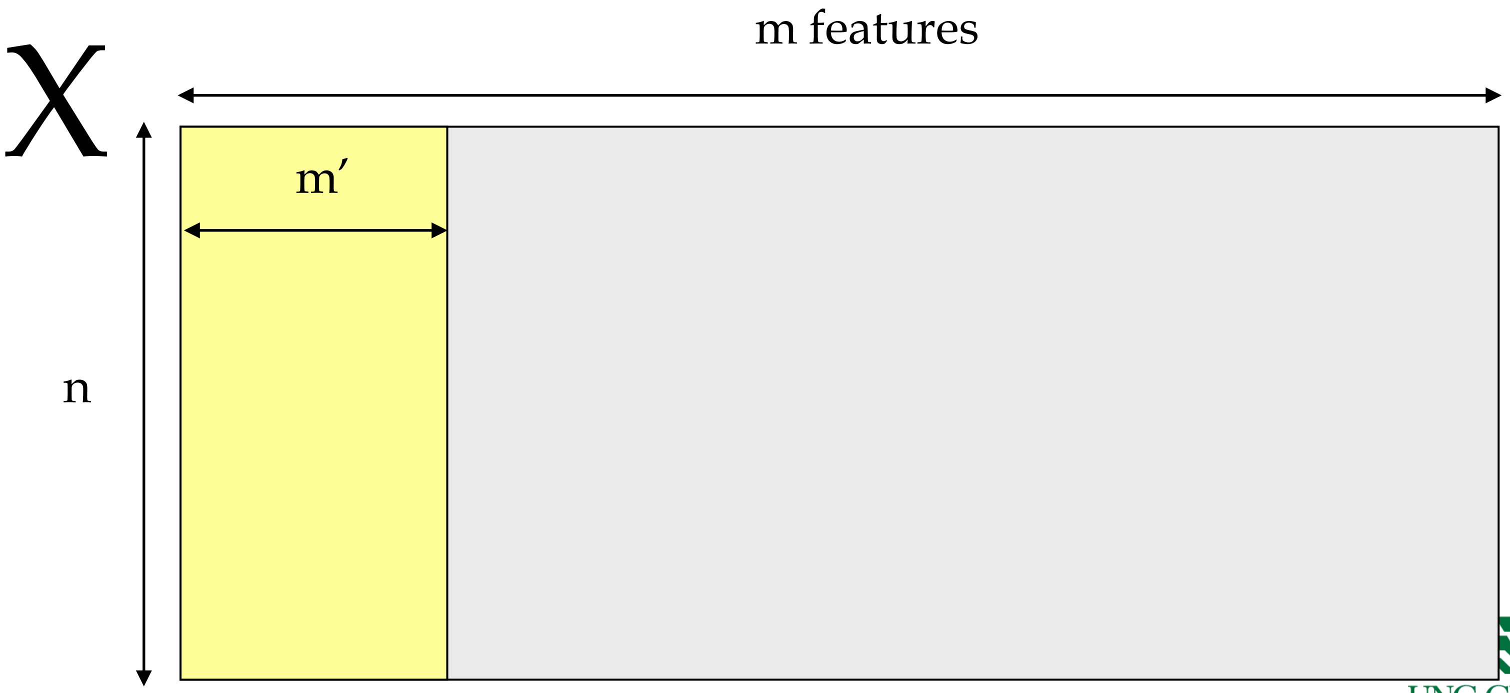
⋮

Most discriminative feature

Feature selection

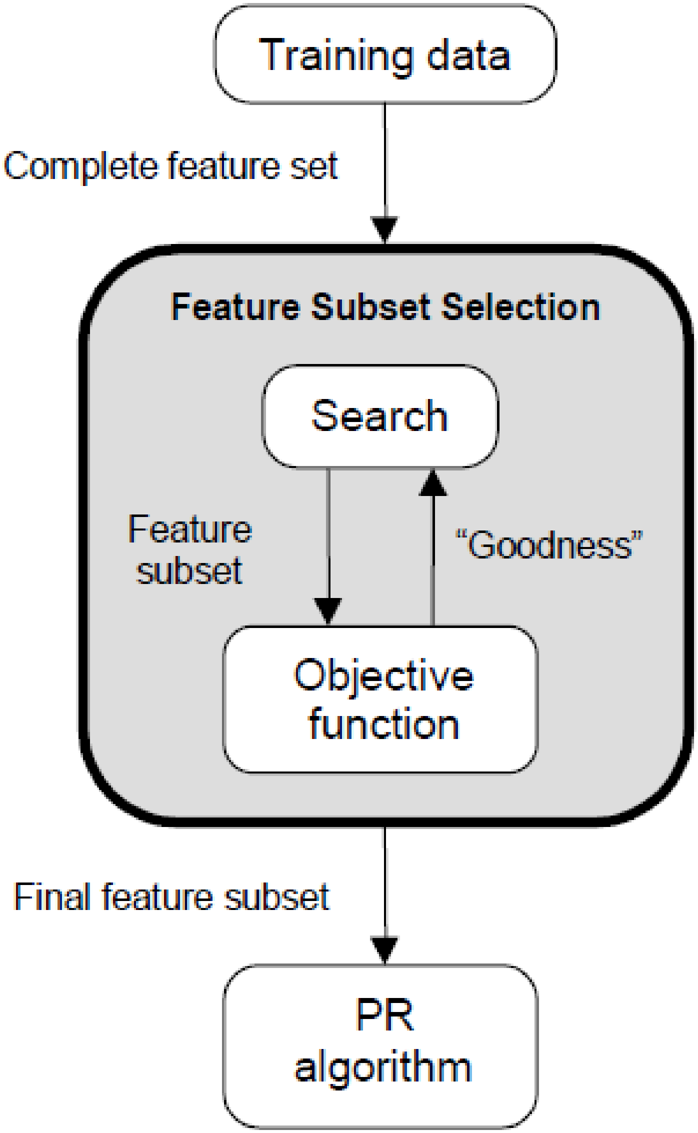
In the presence of **millions of features/attributes/inputs/variables**, select the most relevant ones.

Advantages: build better, faster, and easier to understand learning machines.



Feature selection

- Feature selection is an **optimization** problem.
  - Step 1: Search the space of possible feature subsets.
  - Step 2: Pick the subset that is optimal or near-optimal with respect to some objective function.



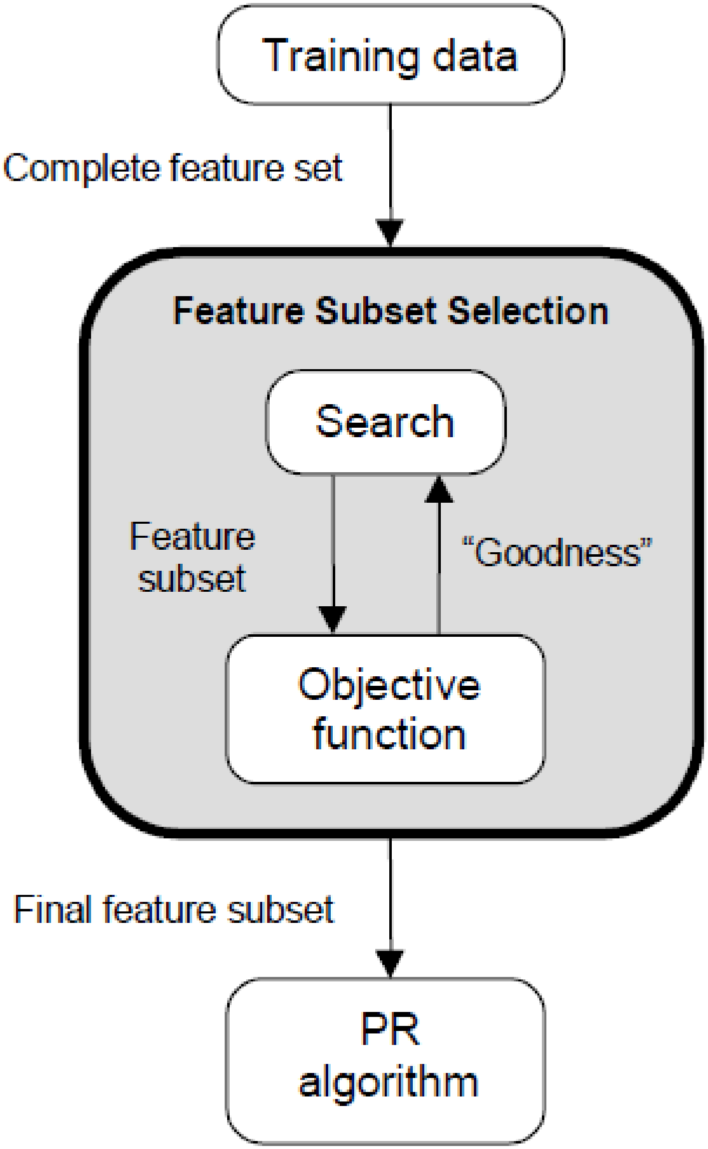
Feature selection

Search strategies

- Optimum
- Heuristic
- Randomized

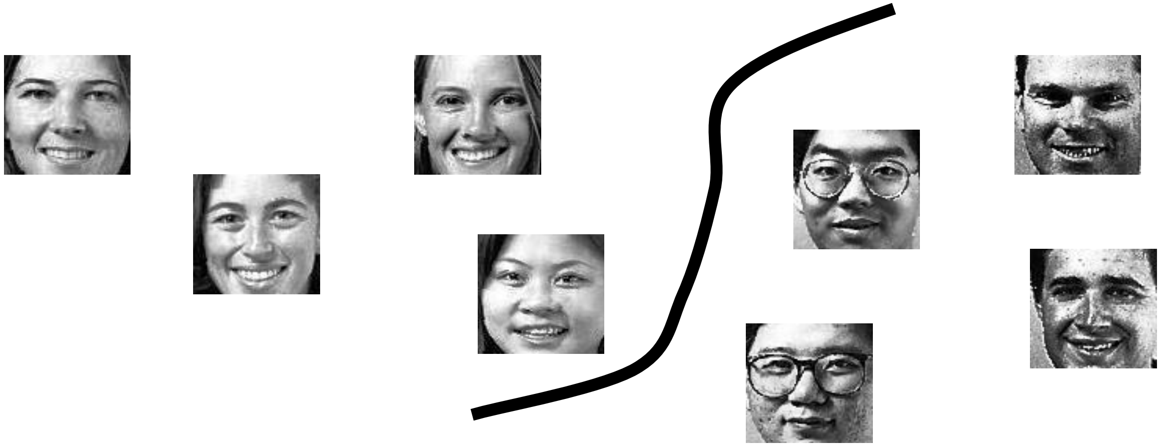
Evaluation strategies

- Filter methods
- Wrapper methods



Case Study: Gender Classification

- Determine the gender of a subject from facial images.
  - Challenges: race, age, facial expression, hair style, etc.

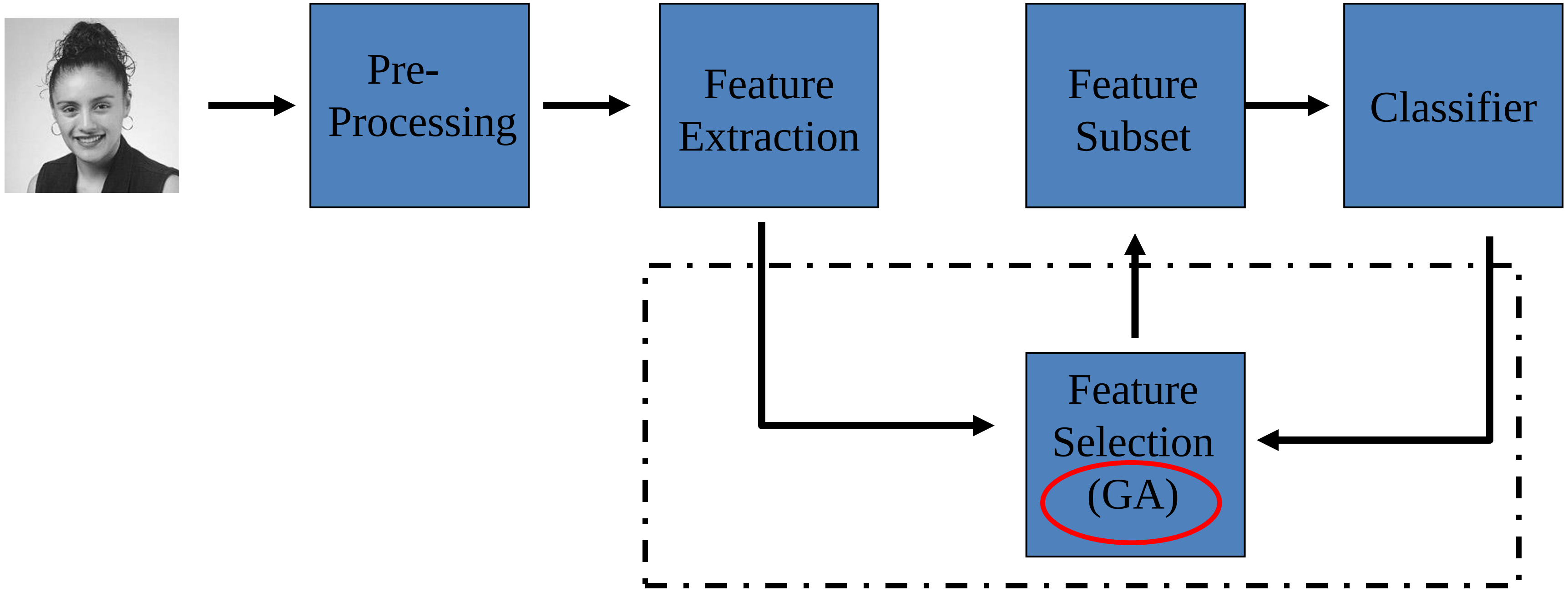


Z. Sun, G. Bebis, X. Yuan, and S. Louis, "Genetic Feature Subset Selection for Gender Classification: A Comparison Study", **IEEE Workshop on Applications of Computer Vision**, pp. 165-170, Orlando, December 2002.

UNC CHARLOTTE

Feature Selection using Genetic Algorithms

- GAs provide a simple, general, and powerful framework for feature selection.



Feature Extraction
(a very brief overview)

Feature Extraction (or Reduction)

- Feature extraction refers to the mapping of the original high-dimensional data onto a lower-dimensional space
- Given a set of data points of p variables $\{x_1, x_2, \cdots, x_n\}$

  Compute their low-dimensional representation:

$$x_i \in \Re^d \rightarrow y_i \in \Re^p \ (p << d)$$

- Criterion for feature reduction can be different based on different problem settings.
  - Unsupervised setting: minimize the information loss
  - Supervised setting: maximize the class discrimination

Given class labels

Feature Reduction vs. Feature Selection

Feature extraction (or reduction): finds a set of new features (i.e., through some mapping f()) from the existing features.

The mapping f() could be linear or non-linear

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_N \end{bmatrix} \xrightarrow{f(\mathbf{x})} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_K \end{bmatrix}$$

K<<N

Feature selection: chooses a subset of the original features.

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ . \\ . \\ x_N \end{bmatrix} \rightarrow \mathbf{y} = \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ . \\ . \\ x_{i_K} \end{bmatrix}$$

K<<N

Slide Credit: George Bebis

Feature Extraction

- Linear combinations are particularly attractive because they are simpler to compute and analytically tractable.

- Given x ϵ R^N, find an N x K matrix U such that:

$$y = U^T x \; \epsilon \; R^K \text{ where } K<N$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ x_N \end{bmatrix} \overset{U^T}{\xrightarrow{f(\mathbf{x})}} \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_K \end{bmatrix}$$

This is a projection from the N-dimensional space to a K-dimensional space.

Slide Credit: George Bebis

Feature Extraction (cont'd)

- From a mathematical point of view, finding an optimum mapping **y**=f(**x**) is equivalent to optimizing an **objective** function.
- Different methods use different objective functions, e.g.,
  - Information Loss: The goal is to represent the data as accurately as possible (i.e., no loss of information) in the lower-dimensional space.
  - Discriminatory Information: The goal is to enhance the class-discriminatory information in the lower-dimensional space.

Slide Credit: George Bebis

Feature Extraction (cont'd)

- Commonly used linear feature extraction methods:
  - Principal Components Analysis (PCA): Seeks a projection that **preserves** as much **information** in the data as possible.
  - Linear Discriminant Analysis (LDA): Seeks a projection that **best discriminates** the data.

- Some other interesting methods:
  - Retaining interesting directions (Projection Pursuit),
  - Making features as independent as possible (Independent Component Analysis or ICA),
  - Embedding to lower dimensional manifolds (Isomap, Locally Linear Embedding or LLE).

Principal Component Analysis (PCA)

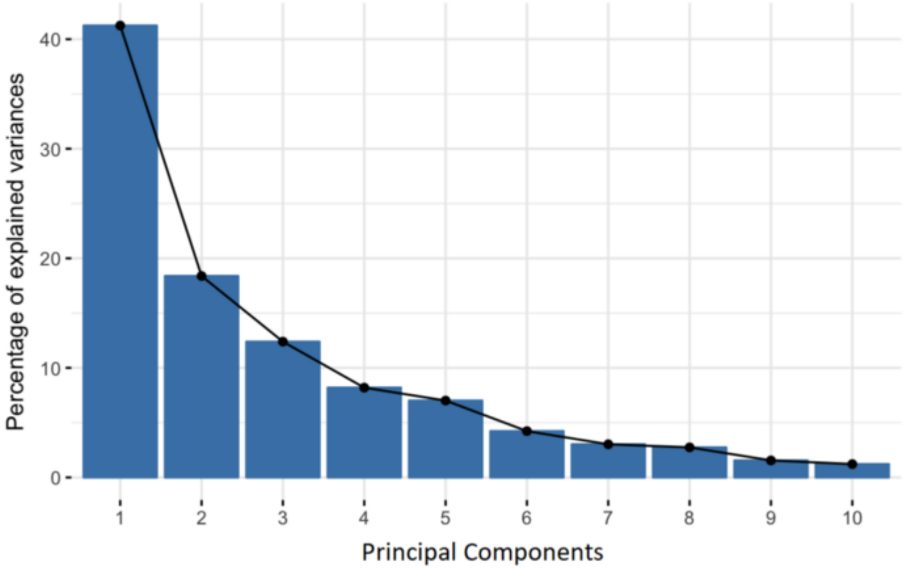
What is Principal Component Analysis (PCA)

- Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets.
- It transforms a large set of variables into a smaller one that still contains most of the information in the large set.
- The trick in dimensionality reduction is to trade a little accuracy for simplicity.
- The idea of PCA is simple — reduce the number of variables of a data set, while preserving as much information as possible.

What are Principal Components?

“Principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. These combinations are done in such a way that **the new variables (which we call principal components) are uncorrelated and most of the information within the initial variables is squeezed or compressed into the first components.”**

PCA tries to put maximum possible information in the first component, then maximum remaining information in the second and so on, until having something like shown in the scree plot below.



What are Principal Components

- Note: the principal components are less interpretable and don’t have any real meaning since they are constructed as linear combinations of the initial variables.
- Organizing information in principal components this way, will reduce dimensionality without losing much information
- The larger the variance carried by a line, the more the information it has.
- As there are as many principal components as there are variables in the data, principal components are constructed in such a manner that the first principal component accounts for the **largest possible variance** in the data set, and so on.

Let’s remember: Covariance

- Variance (one dimension):
  - Measure of the deviation from the mean for points in one dimension

- Covariance (two-dimensional):
  - Measure of how much each of the dimensions vary from the mean with **respect to each other**

- **Covariance is measured between two dimensions**
- **Covariance sees if there is a relation between two dimensions**
- **Covariance between one dimension is the variance**

UNC CHARLOTTE

Math basics

- (Sample) Mean

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- (Sample) Variance

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

- (Sample) Covariance

$$cov(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

PDA Process: (1) Standardization

- The aim of this step is to standardize the range of the continuous initial variables so that each one of them contributes equally to the analysis.
- More specifically, the reason why it is critical to perform standardization prior to PCA, is that the latter is quite sensitive regarding the variances of the initial variables.
- That is, if there are large differences between the ranges of initial variables, those variables with larger ranges will dominate over those with small ranges, which will lead to biased results.
- Mathematically, this can be done by subtracting the mean and dividing by the standard deviation for each value of each variable.

$$z = \frac{value - mean}{standard\ deviation}$$

PDA Process: (2) COVARIANCE MATRIX COMPUTATION

- The aim of this step is to understand how the variables of the input data set are varying from the mean with respect to each other, or in other words, to see if there is any relationship between them.
- In order to identify these correlations, we compute the covariance matrix.
- The covariance matrix is a *p* × *p* symmetric matrix (where *p* is the number of dimensions)
- For example, for a 3-dimensional data set with 3 variables *x*, *y*, and *z*, the covariance matrix is a 3×3 matrix.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

PDA Process:(2) Convergence Matrix

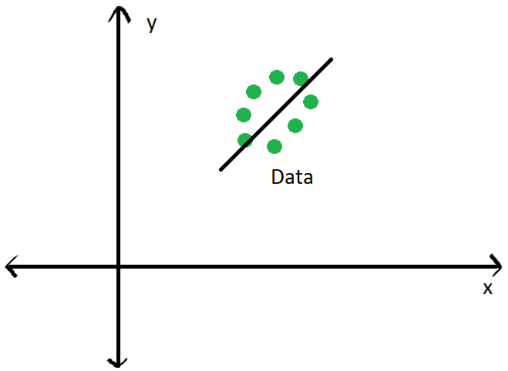
- The covariance of a variable with itself is its variance Cov(a,a)=Var(a)
- The covariance is commutative Cov(a,b)=Cov(b,a)
- The entries of the covariance matrix are symmetric with respect to the main diagonal, which means that the upper and the lower triangular portions are equal.
- if Cov value is positive : the two variables increase or decrease together (correlated)
- if Cov value negative : One increases when the other decreases (Inversely correlated)

PDA Process:(3) Eigenvectors and Eigenvalues

- Eigenvectors and eigenvalues are the linear algebra concepts that we need to compute from the covariance matrix in order to determine the *principal components* of the data.

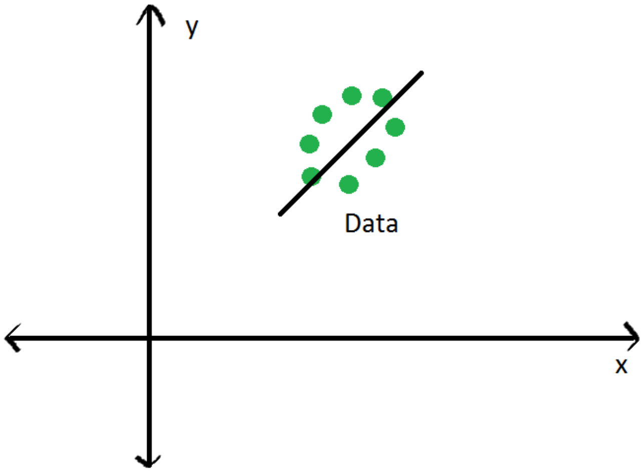
PDA Process:(3) Eigenvectors and Eigenvalues

- Suppose we have plotted a *scatter plot* of random variables, and a line of best fit is drawn between these points..
- **Eigenvector** is the direction of that line.
- **Eigenvalue** is a number that tells us how the data set is spread out on the line which is an Eigenvector.



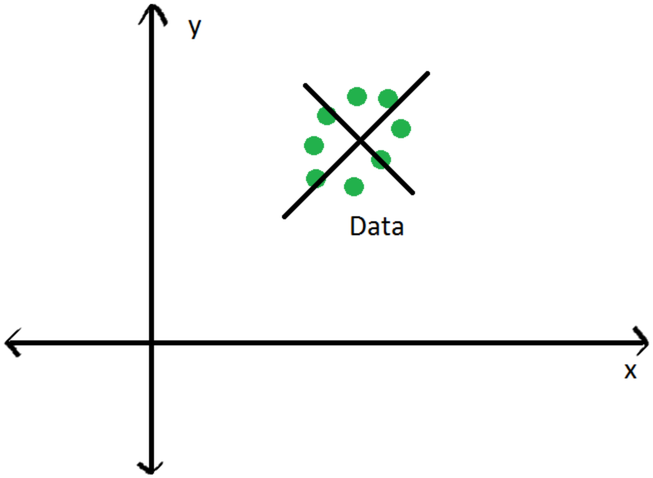
Main Principal Component

- This *line of best fit*, shows the direction of maximum variance in the dataset
- The main principal component, depicted by the black line, is the first Eigenvector.



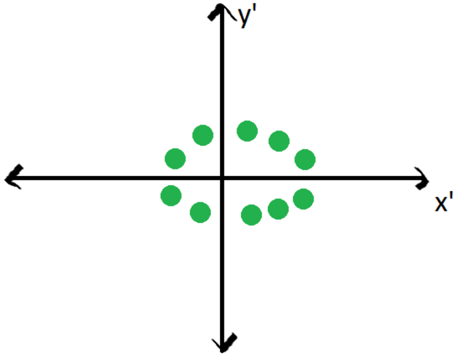
Second Principal Component

- The second Eigenvector will be p**erpendicular or orthogonal** to the first one.
- The reason the two Eigenvectors are orthogonal to each other is because the Eigenvectors should be able to span the whole x-y area.
- Naturally, a line perpendicular to the black line will be our new Y axis, the other principal component.



Generalizing Principle Components:

- Since the Eigenvectors indicate the direction of the principal componen (new axes), we will multiply the original data by the eigenvectors to **orient** our data onto the new axes..



- Geometrically speaking, principal components represent the directions of the data that explain a **maximal amount of variance**, that is to say, the lines that capture most information of the data.

- The larger the variance carried by a line, the larger the dispersion of the data points along it, and the larger the dispersion along a line, the more the information it has.