# Classical Data Analysis


**Master in Big Data Solutions 2017-2018**

Francisco Gutierres
francisco.gutierres@bts.tech

Sara Hajian
sara.hajian@bts.tech

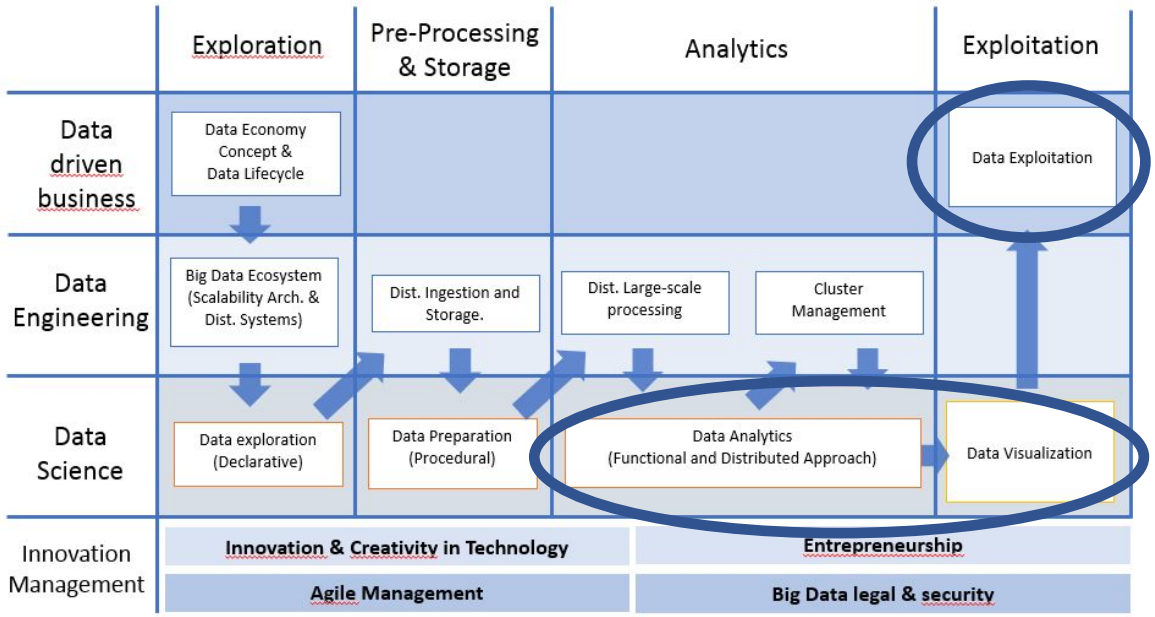
# Session 2.2.6 - Clustering
# K-means

Sara Hajian

# What we will learn

**Session1: Clustering**

K-means

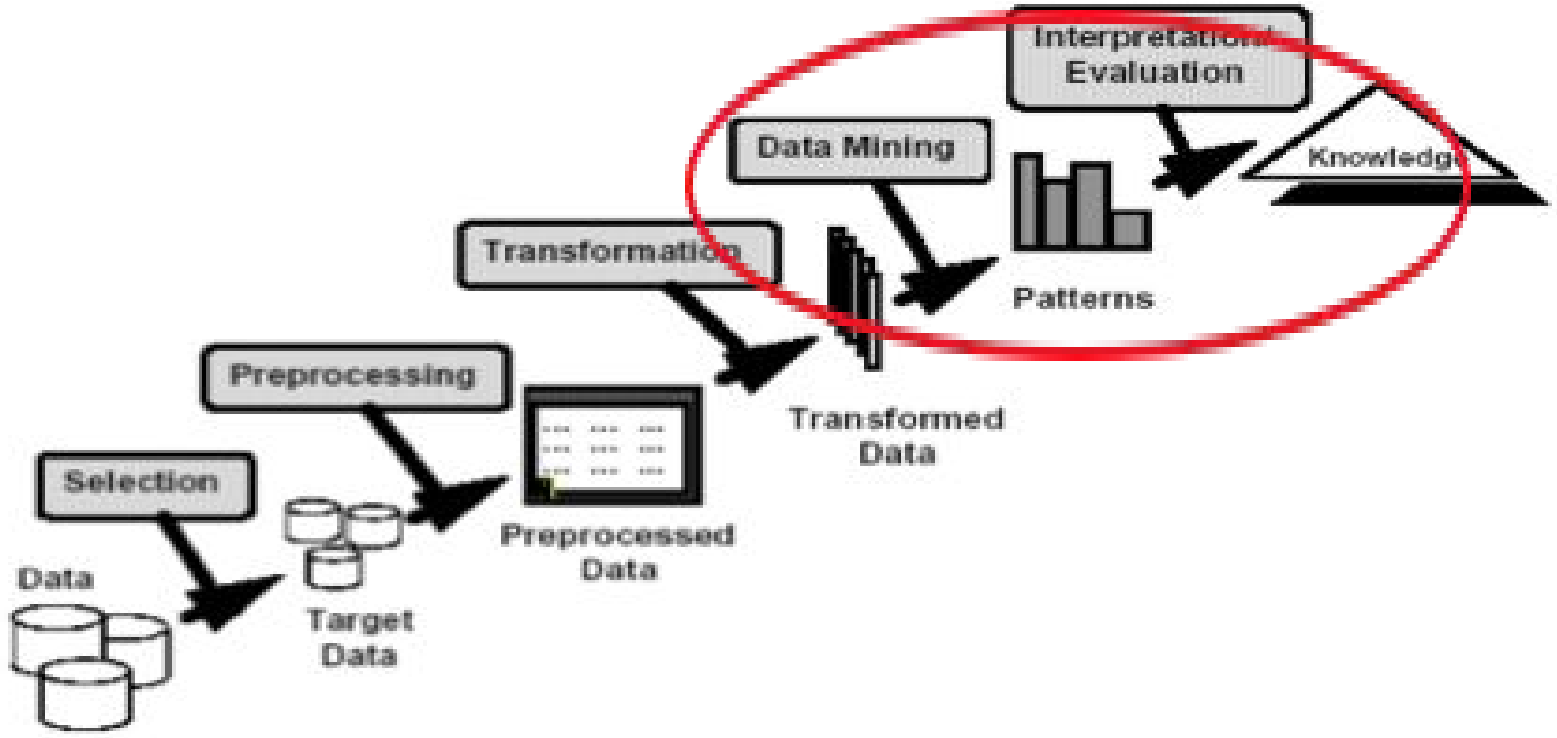



We will learn how to:

- What is cluster analysis, and know about its practical applications
- Different types of clustering and clusters
- K-means
- K-means with Python

# What is data mining

Non-trivial extraction of implicit, previously unknown and potentially useful information from data (i.e., **discovery of meaningful patterns**)



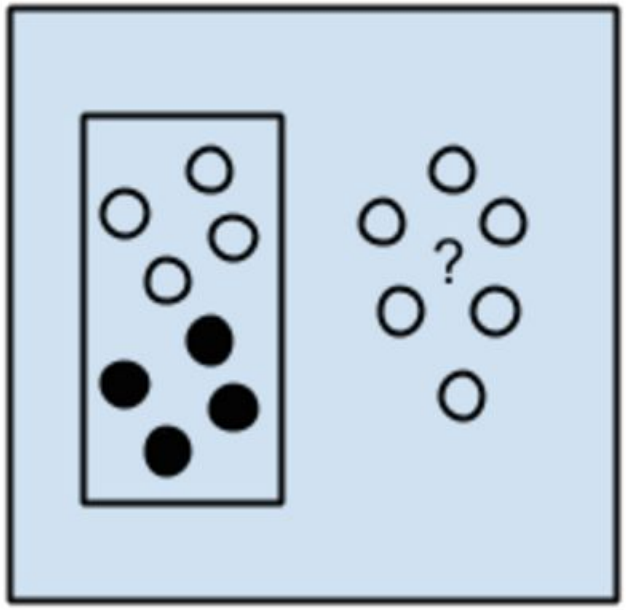

Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

# Data mining algorithms

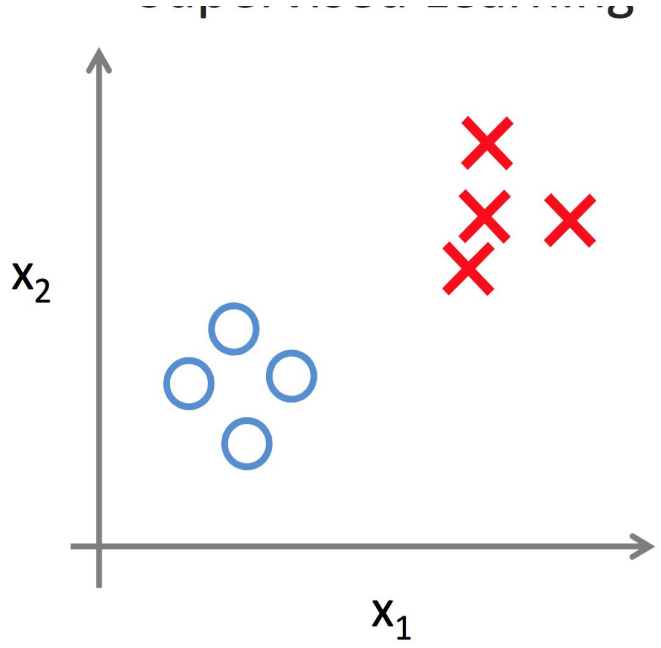
**Supervised**: Classification and regression

**Unsupervised**: Clustering

# Data mining algorithms

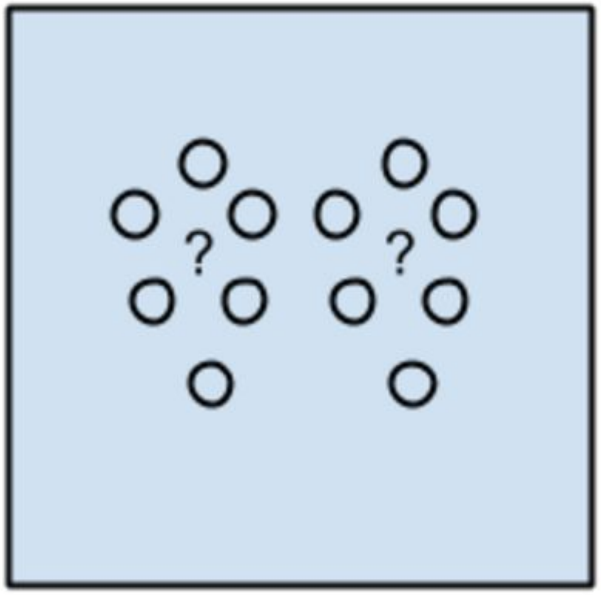



Supervised Learning Algorithms

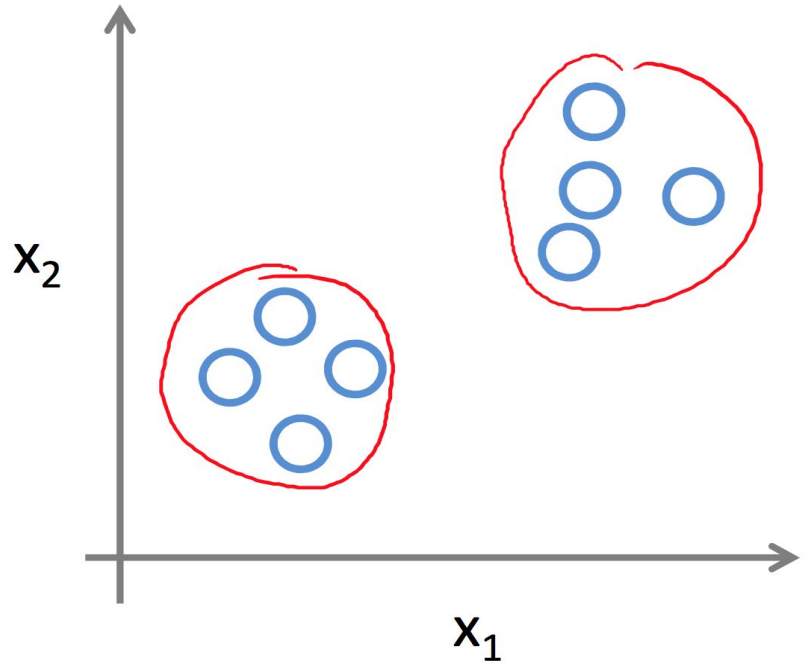



Examples: regression and classification

# Data mining algorithms




Unsupervised Learning Algorithms




Examples: clustering

# Classical Data analysis: Curriculum

**Part I: Regression**

Linear regression

Logistic regression

**Part II: Classification**

Support vector machines (SVM)

Decision trees

Ensemble methods

**Clustering**

K-means

Hierarchical clustering

Supervised learning algorithms

# Classical Data analysis: Curriculum

**Part I: Regression**

Linear regression

Logistic regression

**Part II: Classification**

Support vector machines (SVM)

Decision trees

Ensemble methods

**Clustering**

K-means

Hierarchical clustering

Supervised learning algorithms

# Classical Data analysis: Curriculum

**Part I: Regression**

Linear regression

Logistic regression

**Part II: Classification**

Support vector machines (SVM)

Decision trees

Ensemble methods

**Clustering**

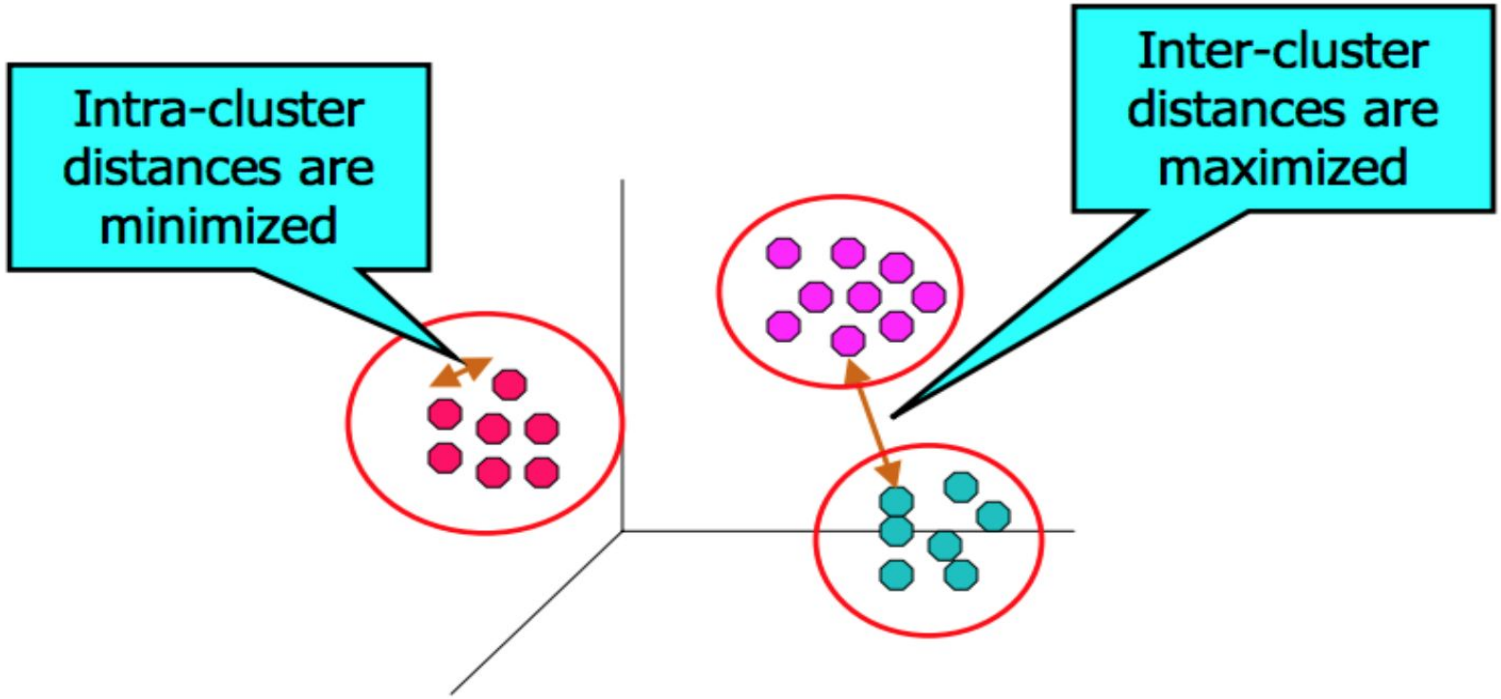
K-means

Hierarchical clustering

Unsupervised learning algorithms

# Clustering analysis: definition

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



# Clustering analysis: applications

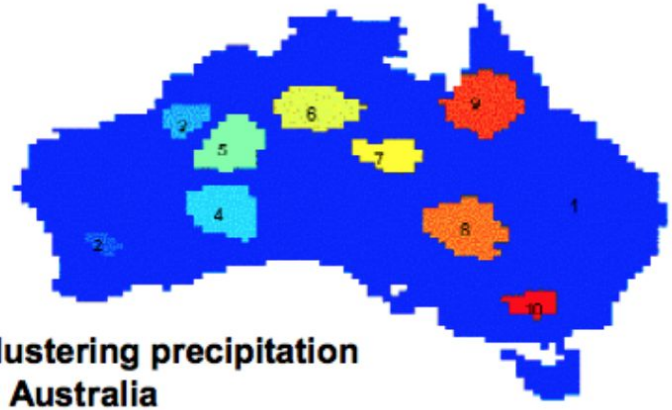
## Understanding

- Group related documents for browsing, group genes and proteins that have similar functionality, or group stocks with similar price fluctuations

| | *Discovered Clusters* | *Industry Group* |
|---|---|---|
| 1 | Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN | Technology1-DOWN |
| 2 | Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN | Technology2-DOWN |
| 3 | Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN | Financial-DOWN |
| 4 | Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP | Oil-UP |

## Summarization

- Reduce the size of large data sets



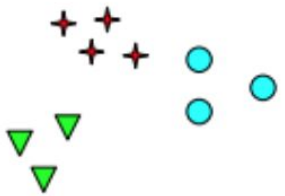

**Clustering precipitation in Australia**

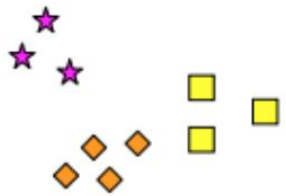
# Clustering analysis: applications

- A key intermediate step for other data mining tasks
  - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
  - Outlier detection: Outliers—those "far away" from any cluster
- Data summarization, compression, and reduction
  - Ex. Image processing: Vector quantization
- Collaborative filtering, recommendation systems, or customer segmentation
  - Find like-minded users or similar products
- Dynamic trend detection
  - Clustering stream data and detecting trends and patterns
- Multimedia data analysis, biological data analysis and social network analysis
  - Ex. Clustering images or video/audio clips, gene/protein sequences, etc.

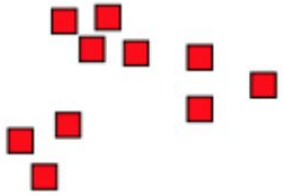
# What is a cluster?

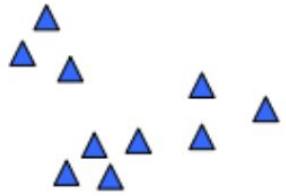
How many clusters?



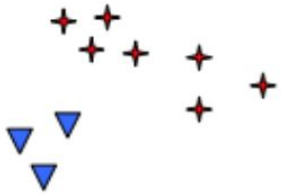


Six Clusters





Two Clusters



Four Clusters

# Types of clusterings

A clustering is a set of clusters

Important distinction between hierarchical and partitional sets of clusters

## Partitional Clustering

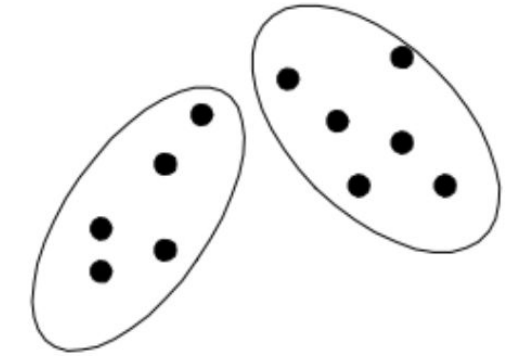
- A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

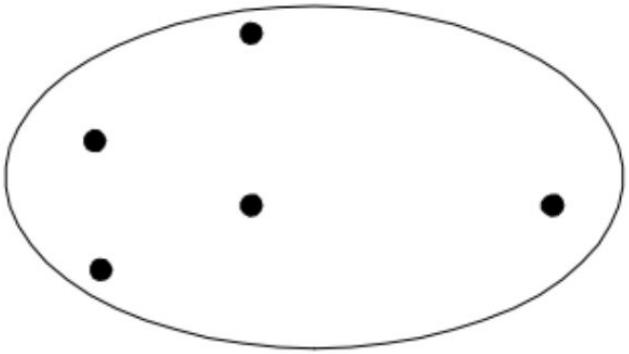
## Hierarchical clustering

- A set of nested clusters organized as a hierarchical tree

# Partitional clustering

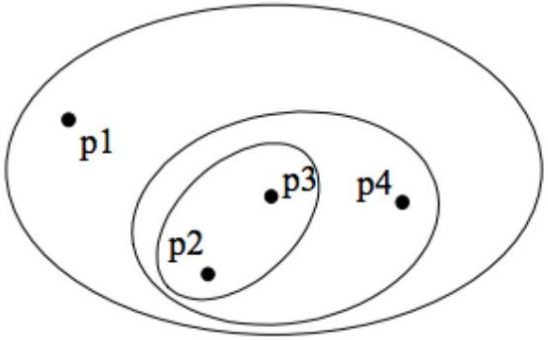
Original Points



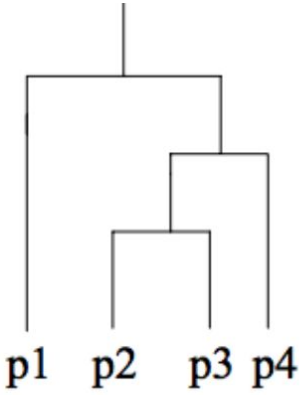


A Partitional Clustering

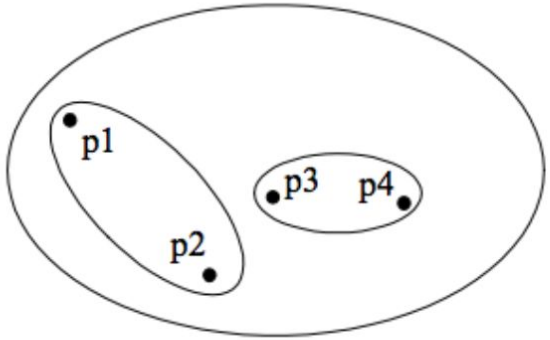
# Hierarchical clustering



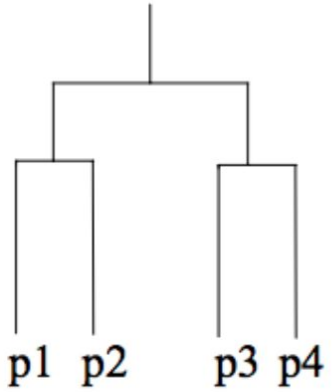

**Traditional Hierarchical Clustering**




**Traditional Dendrogram**




**Non-traditional Hierarchical Clustering**




**Non-traditional Dendrogram**

# Other distinctions between clusterings

## Exclusive versus non-exclusive

- In non-exclusive clusterings, points may belong to multiple clusters.
- Can represent multiple classes or 'border' points

## Fuzzy versus non-fuzzy

- In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1
- Weights must sum to 1
- Probabilistic clustering has similar characteristics

## Partial versus complete

- In some cases, we only want to cluster some of the data

## Heterogeneous versus homogeneous

- Cluster of widely different sizes, shapes, and densities

# Types of clusters

Well-separated clusters

Center-based clusters

Contiguous clusters

Density-based clusters

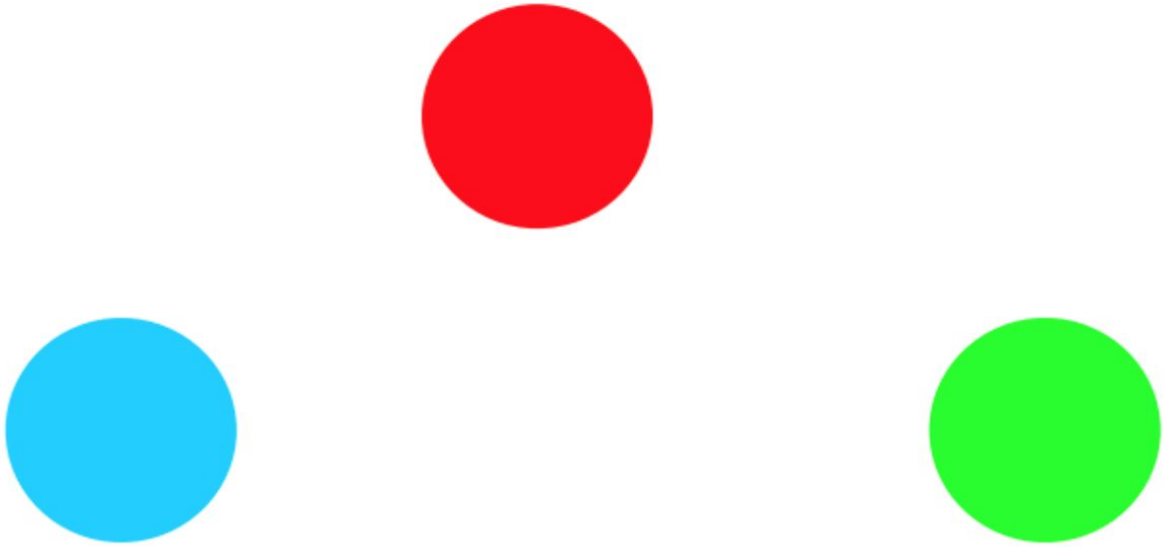
Property or Conceptual

Described by an Objective Function

# Types of clusters: well-separated

## Well-Separated Clusters:

- A cluster is a set of points such that any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



**3 well-separated clusters**

# Types of clusters: Center-Based

## Center-based

- A cluster is a set of objects such that an object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster
- The center of a cluster is often a centroid, the average of all the points in the cluster, or a medoid, the most "representative" point of a cluster

**4 center-based clusters**

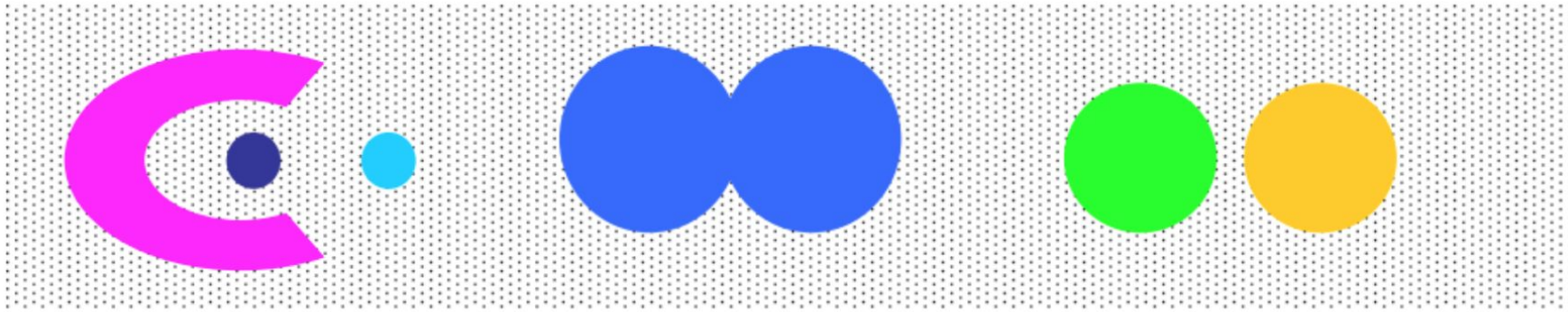
# Types of clusters: Contiguity-Based

- Contiguous Cluster (Nearest neighbor or Transitive)
  - A cluster is a set of points such that a point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.

**8 contiguous clusters**

# Types of clusters: Density-Based

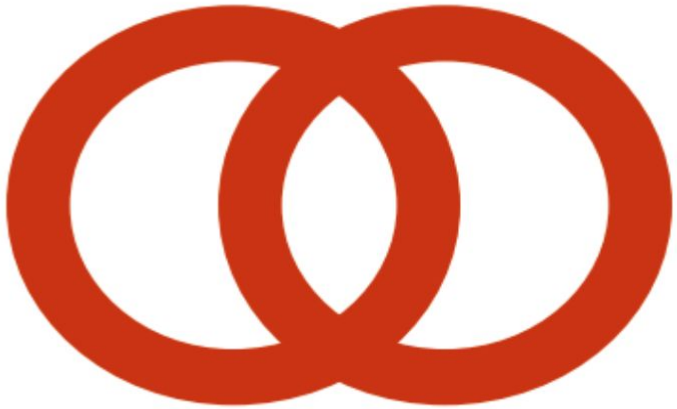
- Density-based
  - A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
  - Used when the clusters are irregular or intertwined, and when noise and outliers are present.



**6 density-based clusters**

# Types of clusters: Conceptual clusters

- Shared Property or Conceptual Clusters
  - Finds clusters that share some common property or represent a particular concept.



**2 Overlapping Circles**

# Types of clusters: Objective function

## Clusters Defined by an Objective Function

- Finds clusters that minimize or maximize an objective function.
- Enumerate all possible ways of dividing the points into clusters and evaluate the `goodness' of each potential set of clusters by using the given objective function. (NP Hard)
- Can have global or local objectives.
  - Hierarchical clustering algorithms typically have local objectives
  - Partitional algorithms typically have global objectives

# Types of clusters: Objective function

Map the clustering problem to a different domain and solve a related problem in that domain

- Proximity matrix defines a weighted graph, where the nodes are the points being clustered, and the weighted edges represent the proximities between points
- Clustering is equivalent to breaking the graph into connected components, one for each cluster.
- Want to minimize the edge weight between clusters and maximize the edge weight within clusters

# Characteristics of in the input data are important

Type of proximity or density measure
- This is a derived measure, but central to clustering

Sparseness
- Dictates type of similarity
- Adds to efficiency

Attribute type
- Dictates type of similarity

Type of Data
- Dictates type of similarity
- Other characteristics, e.g., autocorrelation

Dimensionality

Noise and Outliers

Type of Distribution

# Clustering algorithms

K-means and its variants

Hierarchical clustering

Density-based clustering

# Clustering algorithms

K-means and its variants

Hierarchical clustering

Density-based clustering

# K-means clustering

- Partitional clustering approach
- Each cluster is associated with a centroid (center point)
- Each point is assigned to the cluster with the closest centroid
- Number of clusters, K, must be specified
- The basic algorithm is very simple

---

1: Select $K$ points as the initial centroids.
2: **repeat**
3: Form $K$ clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

---

# K-means clustering

Initial centroids are often chosen randomly.

- Clusters produced vary from one run to another.

The centroid is (typically) the mean of the points in the cluster.

'Closeness' is measured by Euclidean distance, cosine similarity, correlation, etc.

K-means will converge for common similarity measures mentioned above.

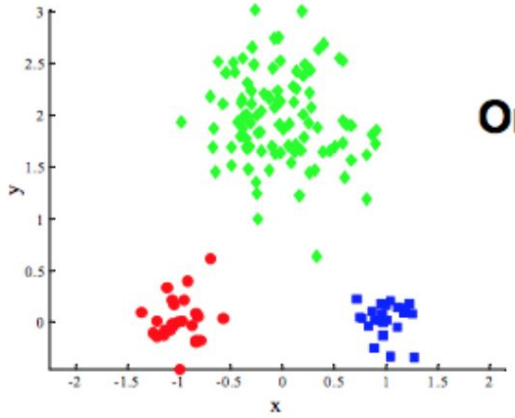
Most of the convergence happens in the first few iterations.

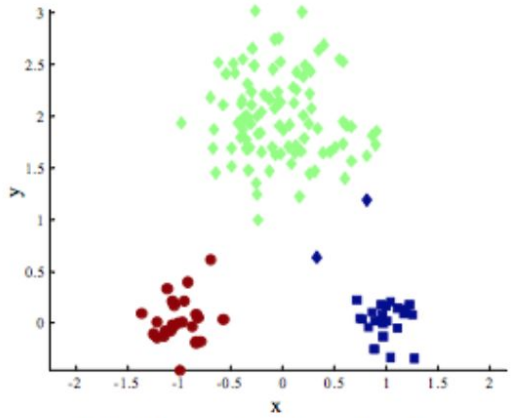
- Often the stopping condition is changed to 'Until relatively few points change clusters'

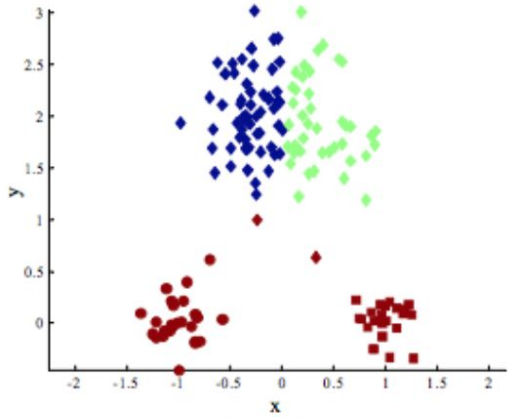
Complexity is O( n * K * I * d )

- n = number of points, K = number of clusters, I = number of iterations, d = number of attributes

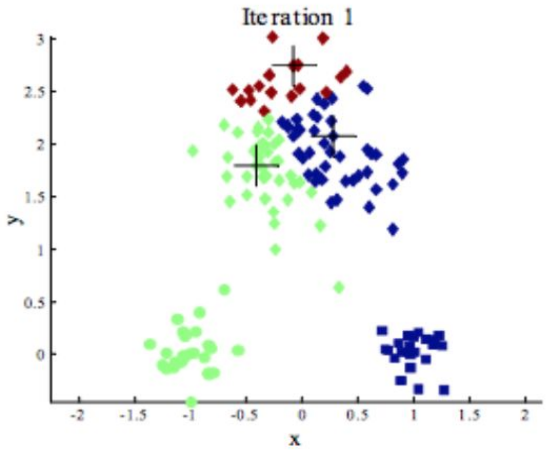
# Two different K-means clusterings

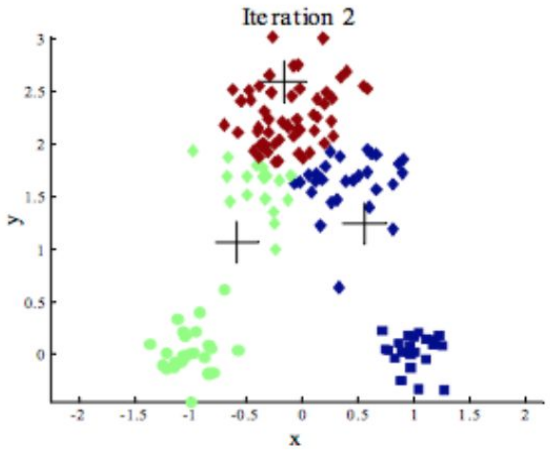


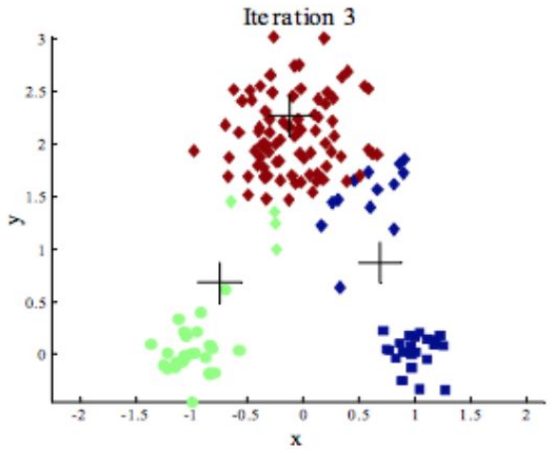



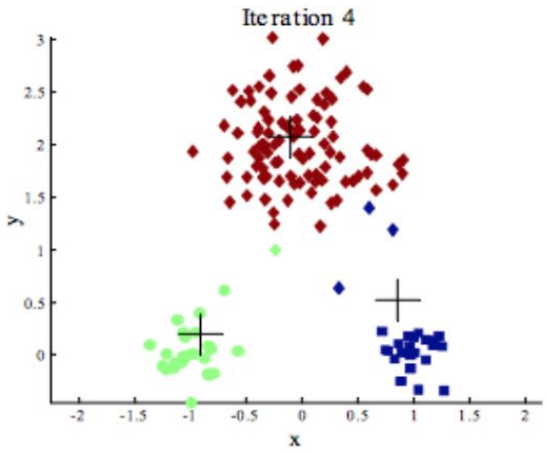



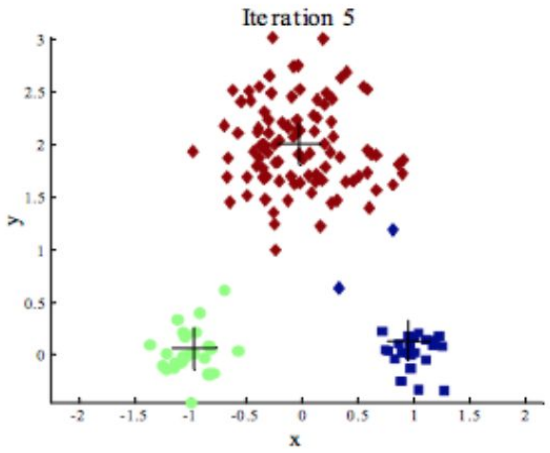
# K-means: example

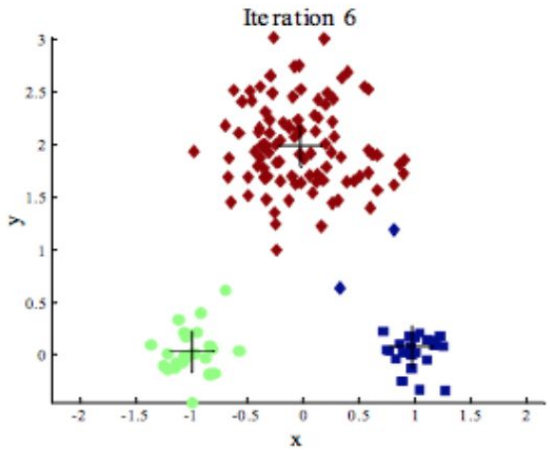


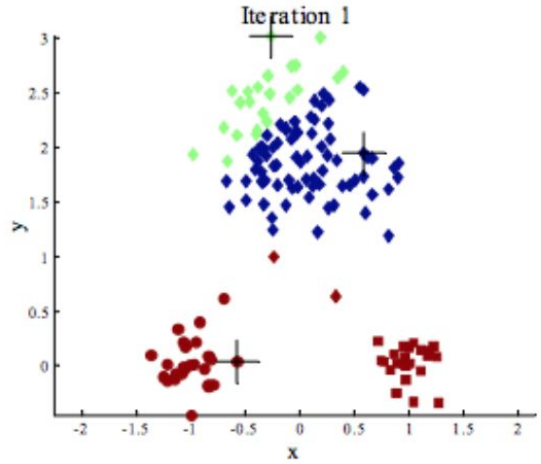



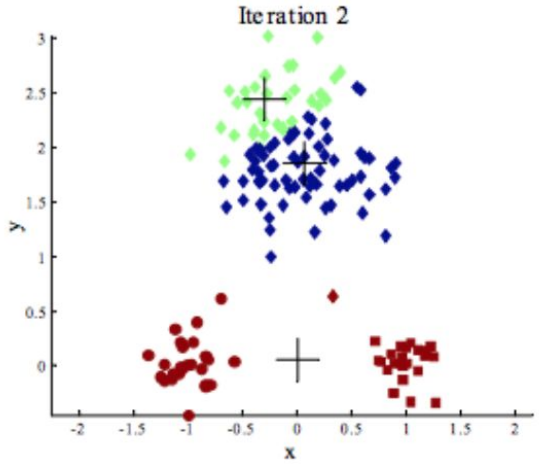



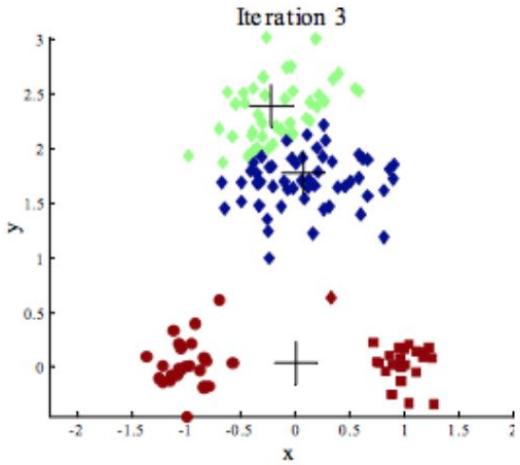



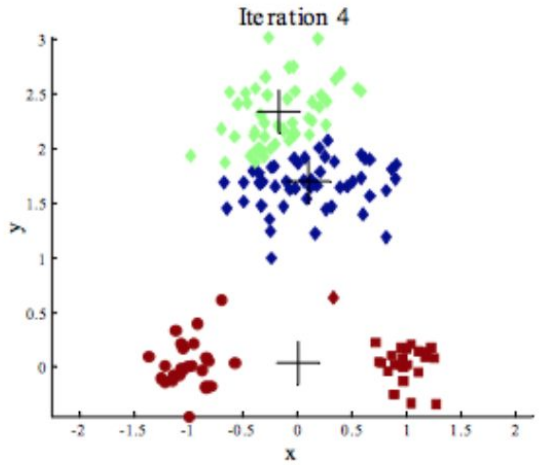



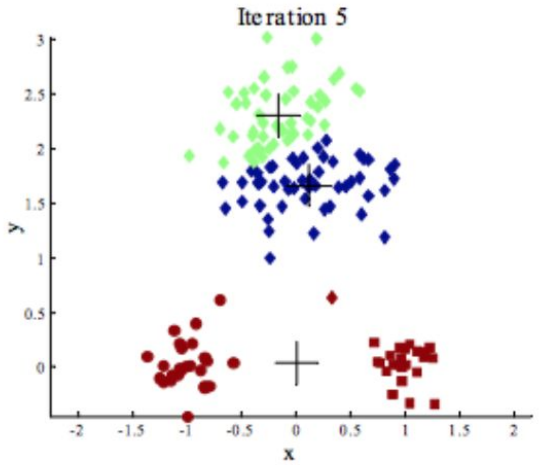
# Importance of choosing initial centroids



# Evaluating K-means Clusters

Most common measure is Sum of Squared Error (SSE)

- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x)$$

- $x$ is a data point in cluster $C_i$ and $m_i$ is the representative point for cluster $C_i$
  - can show that $m_i$ corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase K, the number of clusters
  - A good clustering with smaller K can have a lower SSE than a poor clustering with higher K

# Solutions to Initial centroids problem

- Multiple runs
  - Helps, but probability is not on your side
- Sample and use hierarchical clustering to determine initial centroids
- Select more than k initial centroids and then select among these initial centroids
  - Select most widely separated
- Postprocessing
- Bisecting K-means
  - Not as susceptible to initialization issues

# Pre-processing and post-processing

- Pre-processing
  - Normalize the data
  - Eliminate outliers
- Post-processing
  - Eliminate small clusters that may represent outliers
  - Split 'loose' clusters, i.e., clusters with relatively high SSE
  - Merge clusters that are 'close' and that have relatively low SSE
  - Can use these steps during the clustering process
    - ISODATA

# Bisecting K-means

The **bisecting k-means** algorithm is a variant of **k-means** algorithm. Considering the whole dataset to be one cluster C, in each iteration, we select one cluster and divide it into two partitions using the KM algorithm, until the desired number of clusters **K** is reached.

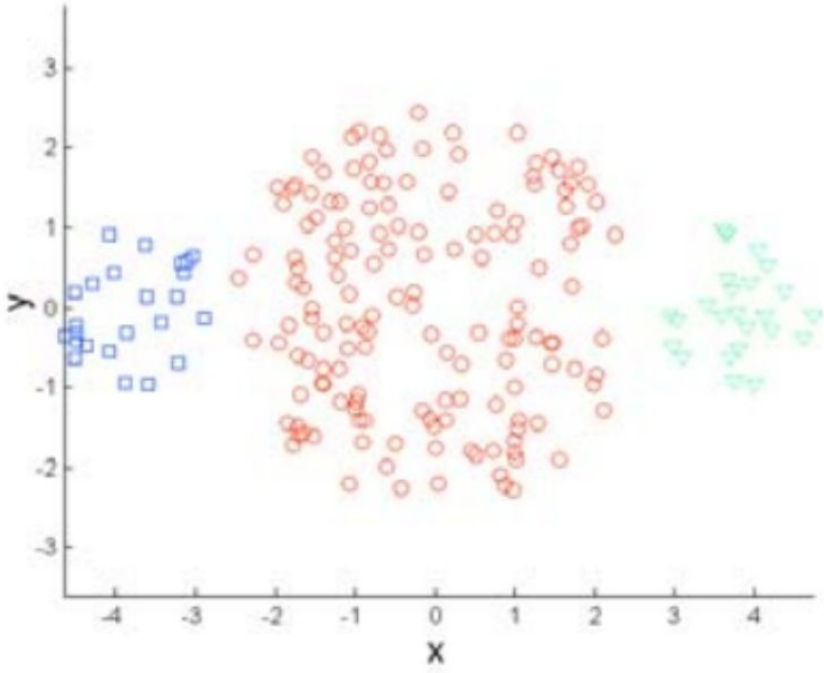
# Limitations of K-means

K-means has problems when clusters are of differing

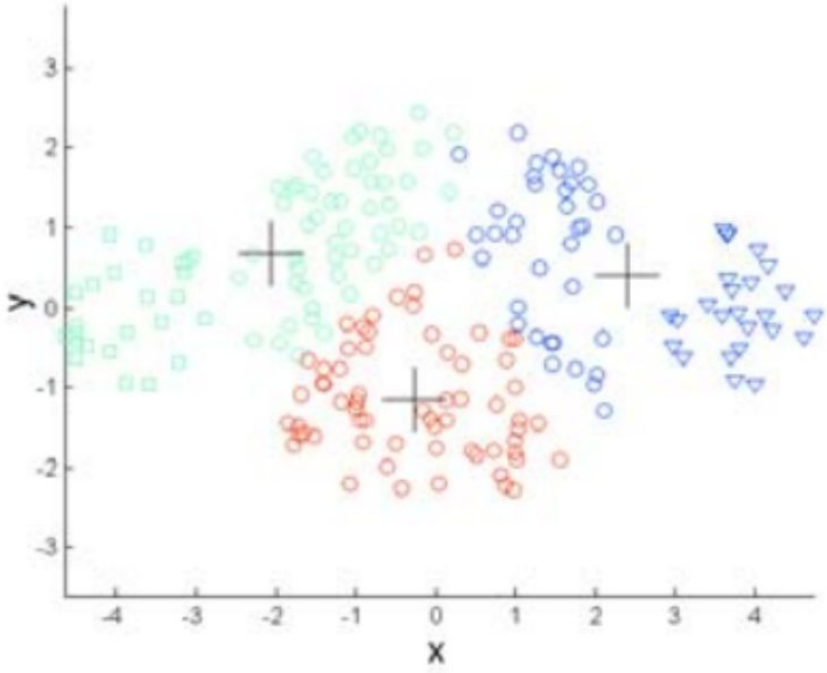
- Sizes
- Densities
- Non-globular shapes

K-means has problems when the data contains outliers.

# Limitations of K-means: Differing sizes

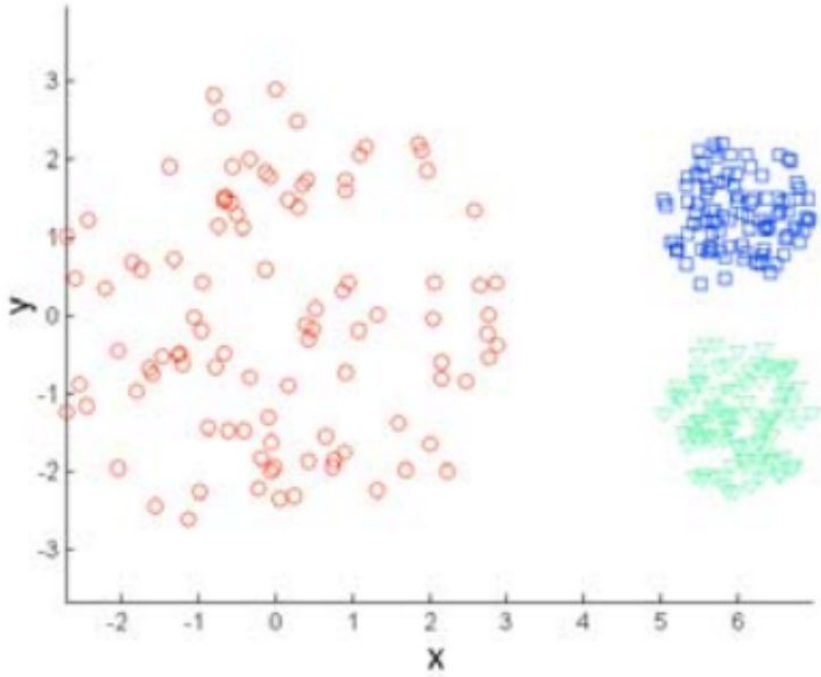



**Original Points**

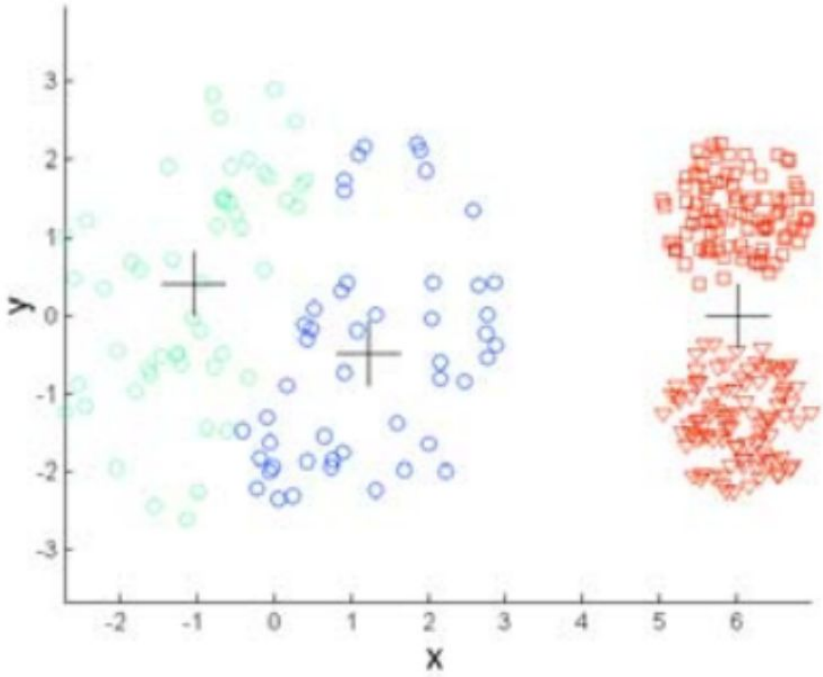



**K-means (3 Clusters)**

# Limitations of K-means: Differing density

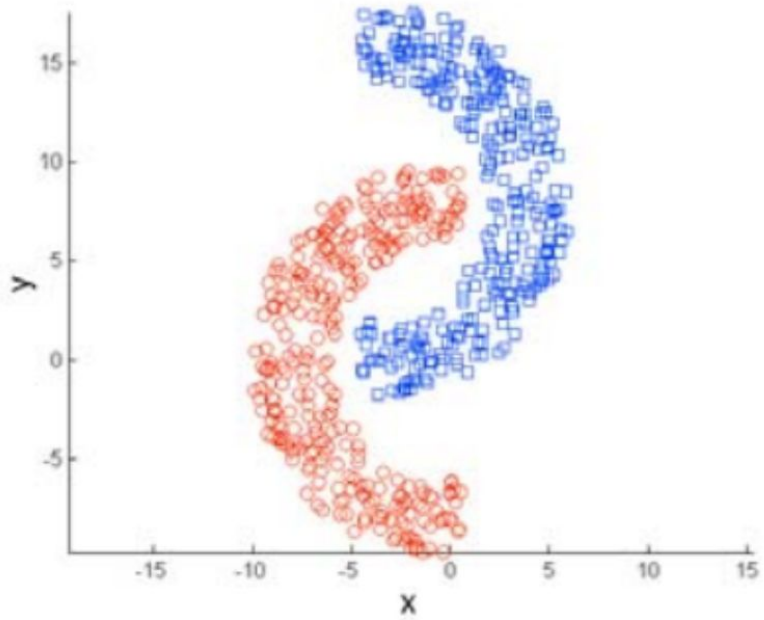



**Original Points**

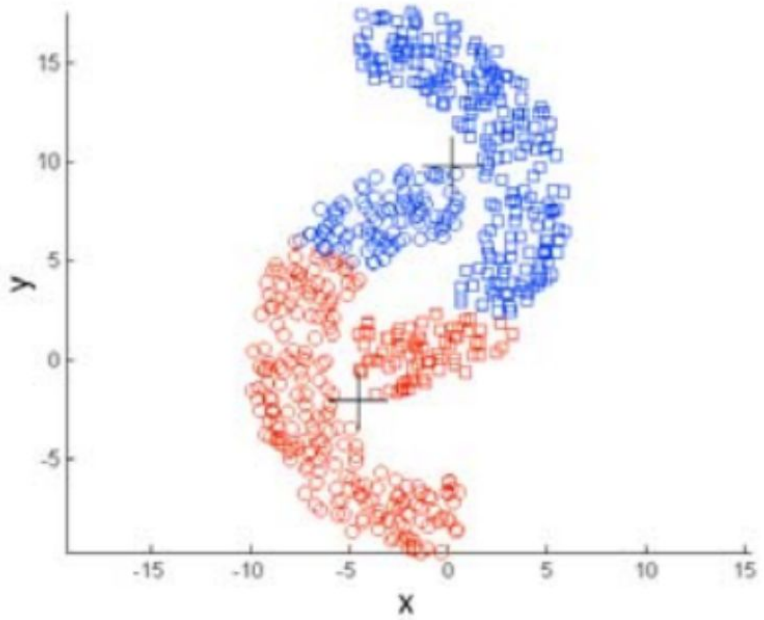



**K-means (3 Clusters)**

# Limitations of K-means: Non-globular shapes

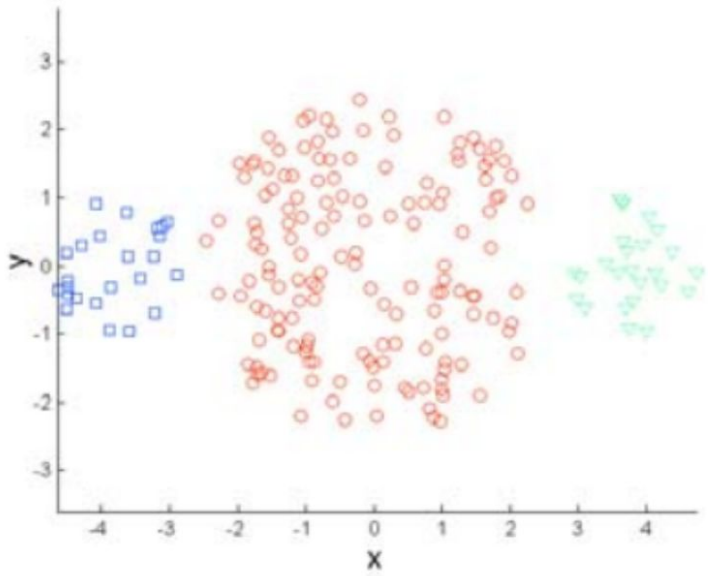



**Original Points**

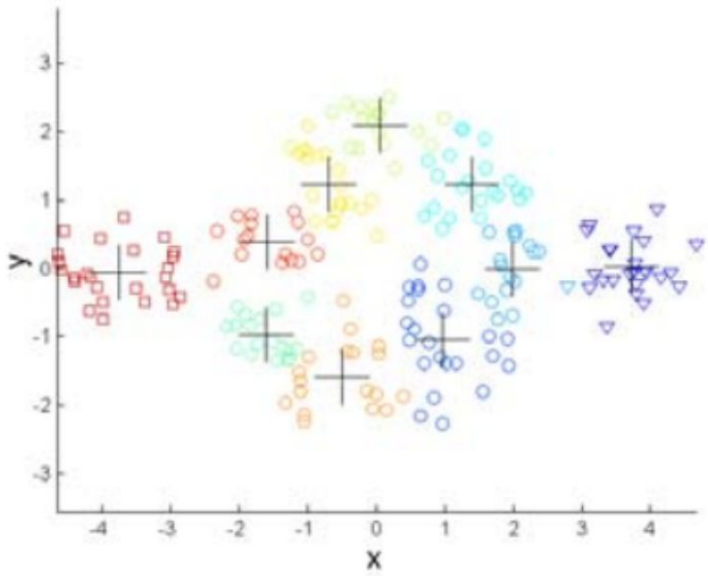



**K-means (2 Clusters)**

# Overcoming K-means limitations




**Original Points**




**K-means Clusters**

One solution is to use many clusters.
Find parts of clusters, but need to put together.

# Individual assignment

Install Weka from (https://www.cs.waikato.ac.nz/ml/weka/), load Titanic dataset in Weka, try different features of the tool for data preprocessing, regression, classification and visualization.

BTS | Barcelona Technology School

eurecat Centre Tecnològic de Catalunya

# Next class

**Clustering with Python**

**Hierarchical clustering:**

- description of hierarchical clustering
- different types of hierarchical clustering
- agglomerative hierarchical clustering algorithm
- measures for cluster proximity
- limitations and strengths

**Cluster evaluation**

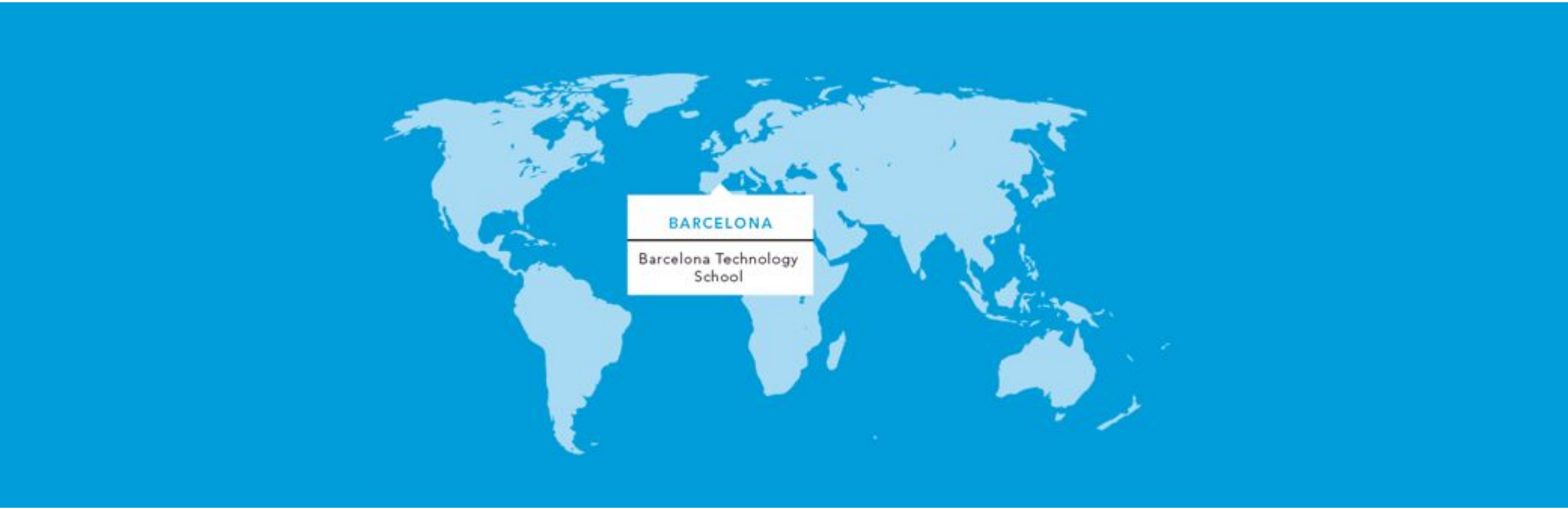
# Resources


- Tan, Pang-Ning. *Introduction to data mining*. Pearson Education India, 2006.

- Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. New York: Springer series in statistics, 2001.

BTS | Barcelona Technology School

eurecat Centre Tecnològic de Catalunya




# Thank you

Barcelona, 2017