Clustering
Part 1

Mohammed Brahimi & Sami Belkacem

Outline

- Overview of Clustering
- Major Clustering Approaches
    - K-means Clustering
    - Hierarchical Clustering
    - DBSCAN Clustering
- Cluster Evaluation

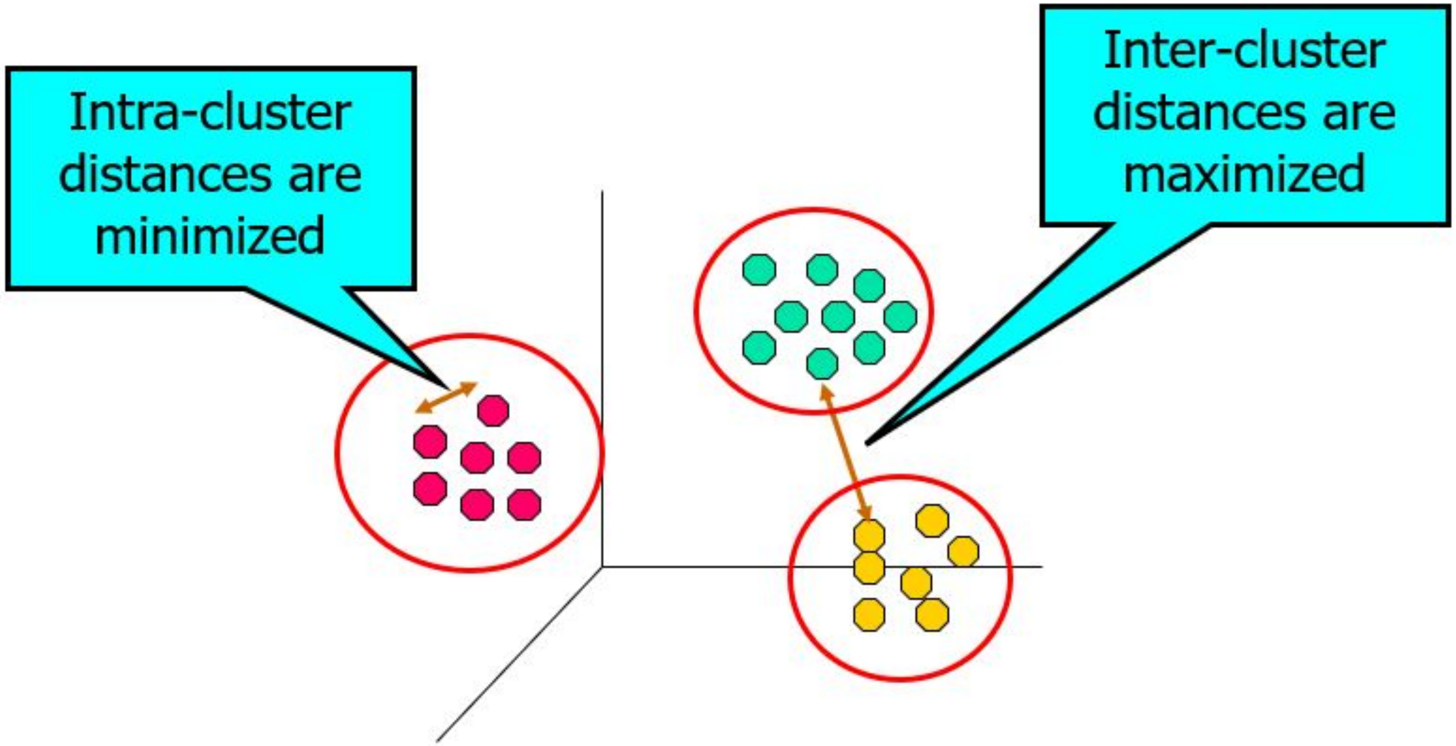
Outline

- **Overview of Clustering**
- **Major Clustering Approaches**
  - **K-means Clustering**
  - **Hierarchical Clustering**
  - **DBSCAN Clustering**
- **Cluster Evaluation**

What is Clustering?

- Given a set of objects, place them in groups such that:

  The objects in a group are **similar** (or related) to one another and **different** from (or unrelated to) the objects in other groups



Think of some practical applications of clustering?

Applications of Clustering

- **Marketing:** Discover customer segments for targeted marketing
- **Information retrieval:** Document clustering
- **Land use:** Identifying similar land use areas in an Earth database
- **Biology:** Taxonomy levels (kingdom to species)
- **City planning:** Grouping houses by type, value, and location
- **Earthquake studies:** Clustering observed epicenters along fault lines
- **Climate:** Analyzing atmospheric and ocean patterns
- **Economic science:** Market research

Clustering as Preprocessing tool

- **Summarization:**
  - Preprocessing for classification, regression, PCA, and association analysis
- **Compression:**
  - Image processing using vector quantization
- **Finding K-nearest Neighbors**
  - Localizing search to one or a small number of clusters
- **Outlier detection**
  - Outliers are often viewed as those “far away” from any cluster

What is a **good** clustering and what are the **factors** that contribute to it?

What is a Good Clustering? (1)

- A **good clustering** method will produce high-quality clusters
  - high intra-class similarity: cohesive within clusters
  - low inter-class similarity: distinctive between clusters
- The **quality** of a clustering method depends on:
  - the similarity measure used by the method
  - the implementation of the clustering method
  - the ability to discover some or all of the hidden patterns

What is a Good Clustering? (2)

- **Dissimilarity/Similarity metric**
  - Similarity is expressed in terms of a distance function $d(i, j)$
  - The definitions of **distance functions** depend on the attribute type: boolean, categorical, interval-scaled, ordinal ratio, and vector variables
  - Weights should be associated with different attributes based on the domain application and data semantics
- **Quality of clustering**
  - There is a “quality” function that measures the “goodness” of a cluster
  - It is hard to define “similar enough” or “good enough” due to subjectivity

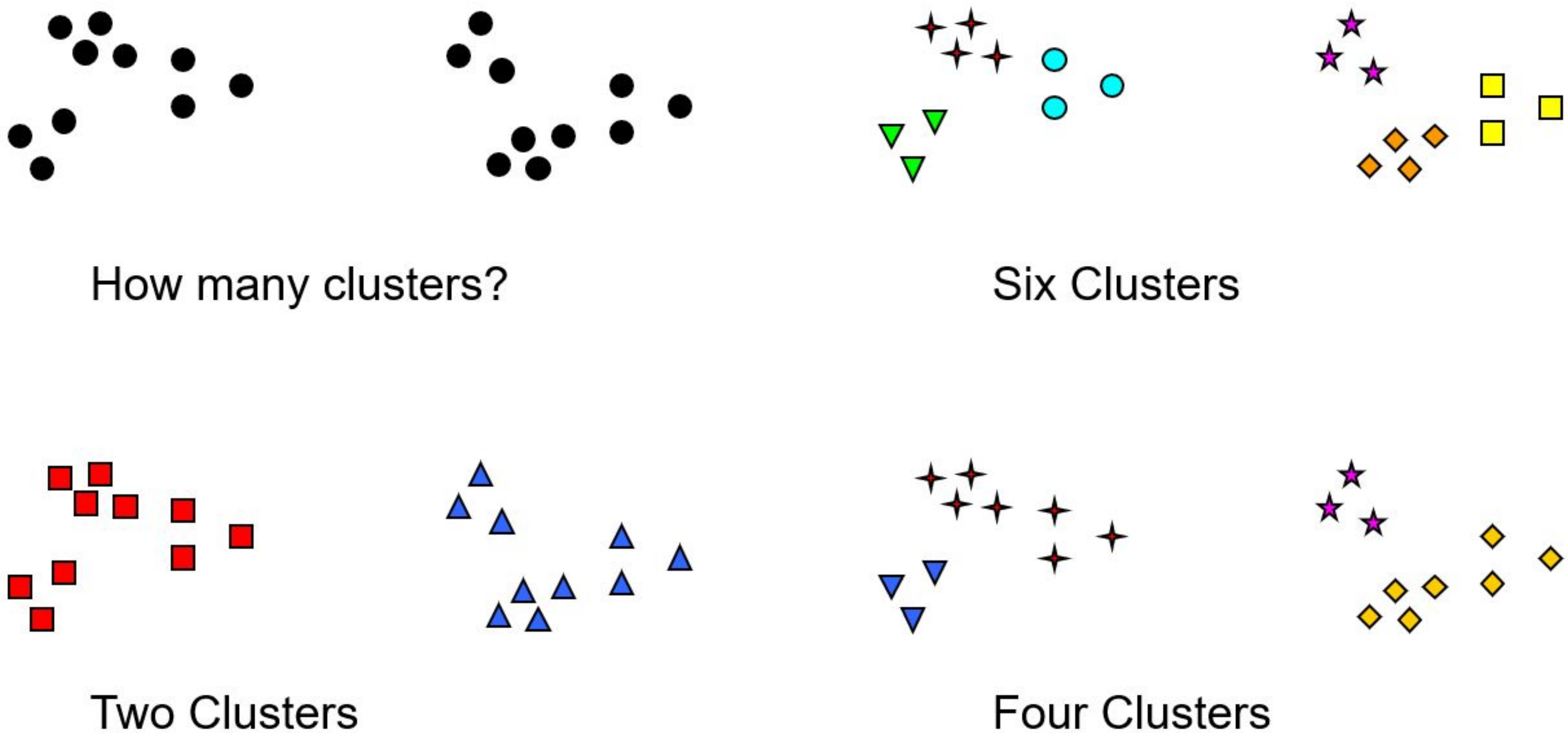
Considerations for Clustering

- **Partitioning criteria**
  - Single-level
  - Hierarchical partitioning (often, multi-level partitioning is desirable)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region)
  - Non-exclusive (e.g., one document may belong to more than one class)
- **Similarity measure**
  - Distance-based (e.g., Euclidean, road network, vector)
  - Connectivity-based (e.g., density or contiguity)
- **Clustering space**
  - Full space (often when low dimensional)
  - Subspaces (often in high-dimensional clustering)

Notion of a Cluster can be Ambiguous

How many clusters?

Notion of a Cluster can be Ambiguous



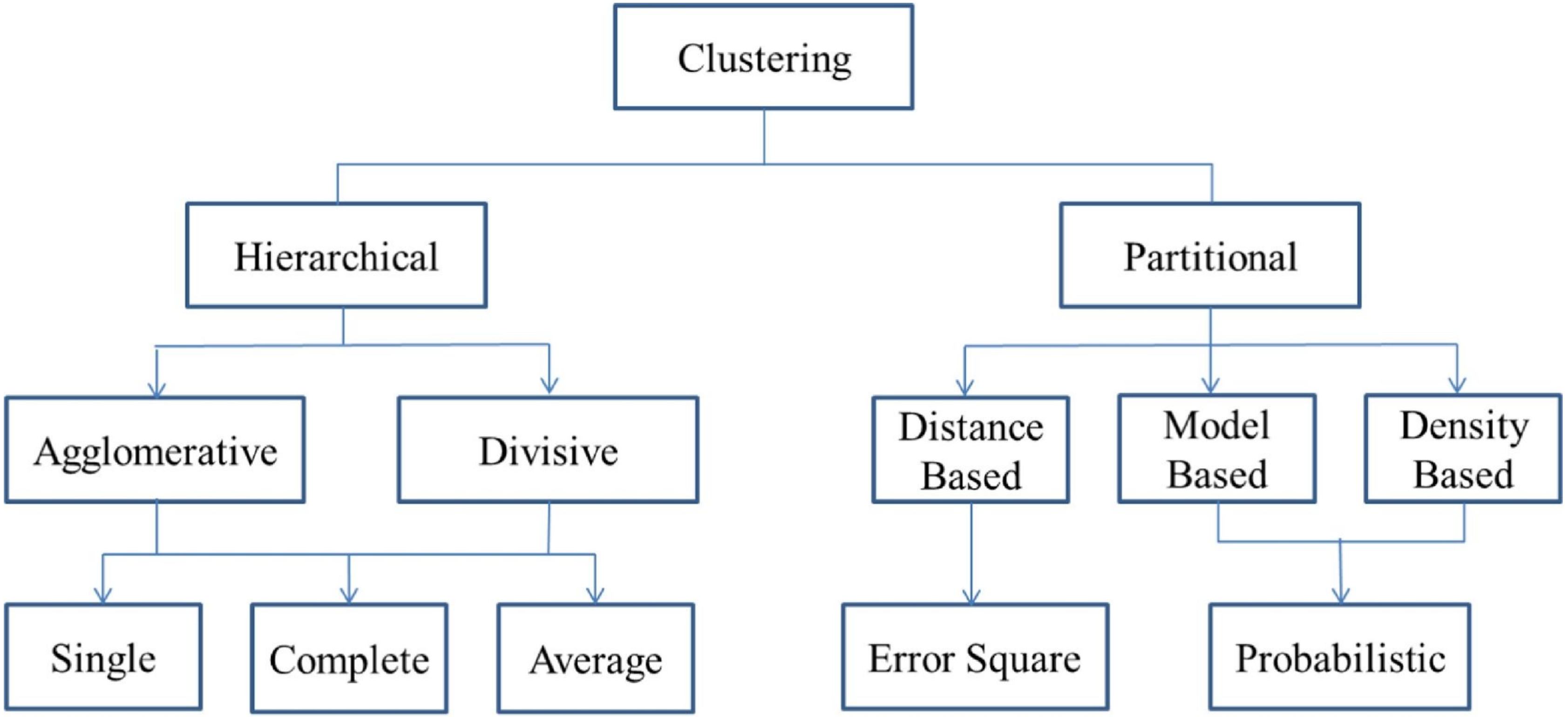
Requirements and Challenges

- **Interpretability**
  - Explain and use the different clusters
- **Scalability**
  - Clustering all the data instead of samples
- **Deal with different types of attributes**
  - Numerical, binary, categorical, ordinal, linked, and a mixture of these
- **Constraint-based clustering**
  - User may give inputs on constraints
  - Use domain knowledge to determine input parameters
- **Others**
  - Ability to deal with noisy data and outliers
  - Ability to detect clusters of any shape

Outline

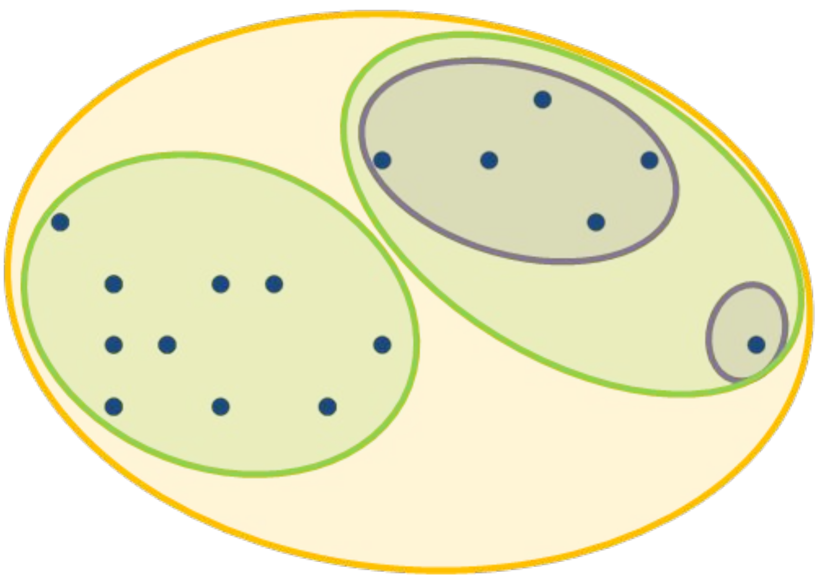
- Overview of Clustering
- Major Clustering Approaches
  - K-means Clustering
  - Hierarchical Clustering
  - DBSCAN Clustering
- Cluster Evaluation

Major Clustering Approaches



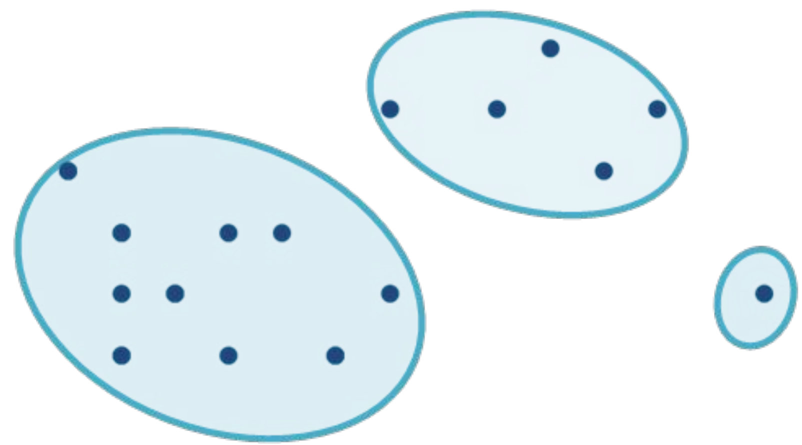
Partitional vs Hierarchical Clustering

Hierarchical Clustering



Nested clusters

Partitional Clustering



Non-nested clusters

Major Clustering Approaches

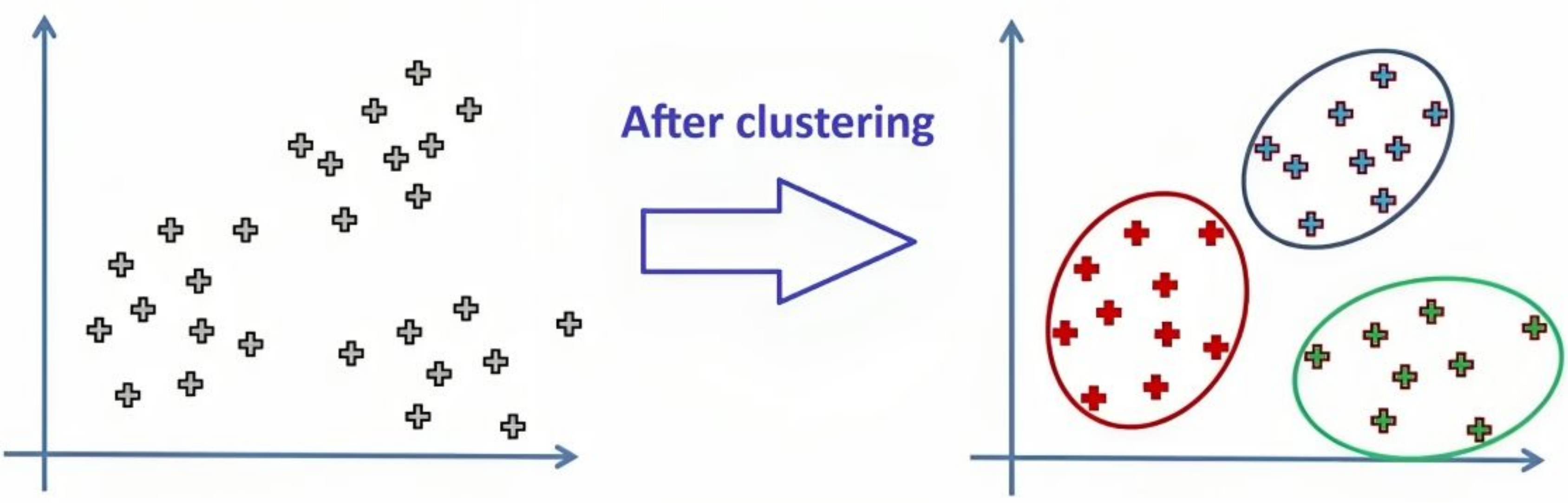
- **<u>Partitioning approach</u>**:
  - Construct various partitions and then evaluate them by some criterion
  - <u>Typical methods</u>: K-means, K-medoids, CLARANS
- **<u>Hierarchical approach</u>**:
  - Create a hierarchical decomposition of the set of data using some criterion
  - <u>Typical methods</u>: Diana, Agnes, BIRCH, CAMELEON
- **<u>Density-based approach</u>**:
  - Based on connectivity and density functions (detect regions where points are concentrated)
  - <u>Typical methods</u>: DBSCAN, OPTICS, DenClue
- **<u>Model-based</u>**:
  - A model is hypothesized for each of the clusters and tries to find the best fit
  - <u>Typical methods:</u> EM, SOM, COBWEB

Outline

- Overview of Clustering
- Major Clustering Approaches
  - **K-means Clustering**
  - Hierarchical Clustering
  - DBSCAN Clustering
- Cluster Evaluation

Partitional Algorithms

- **Objective:** Partitioning a database ***D*** of ***n*** objects into a set of ***K*** clusters.
- K-Means algorithm is an example of a partitional clustering algorithm.
- Example of clustering data points with ***K***=3:



Which objective function should be used?

Objective Function

- A common objective function is minimize the Sum of Squared Distances (SSE)
- SSE is used with the Euclidean distance measure

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

  - ***x*** is a data point in cluster $\boldsymbol{C}_\mathrm{i}$ and $\boldsymbol{m_i}$ is the centroid or medoid for cluster $\boldsymbol{C}_\mathrm{i}$
  - For each point ***x***, the error is the distance to the nearest cluster center $\boldsymbol{m_i}$
  - To get SSE, we square these errors and sum them.
  - SSE improves in each iteration until it reaches a local or global minima.

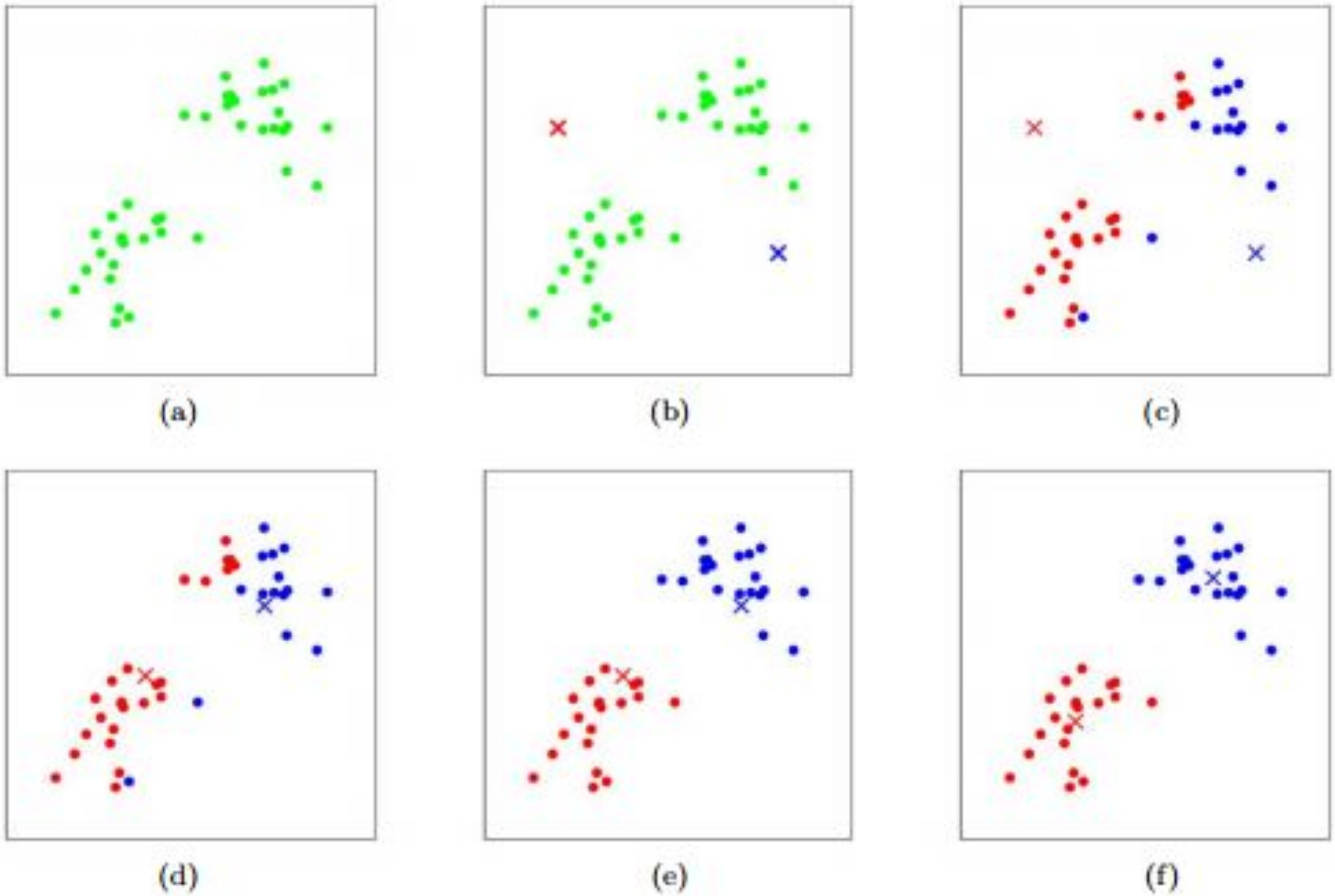
K-Means Algorithm

- The number of clusters ***K*** must be specified as input
- Each cluster is represented with a centroid (i.e. center point, e.g. the mean)
- In the first iteration, the ***K*** centroids are ***K*** random points (objects) from the data
- Each point in the data is assigned to the cluster with the closest centroid
- The centroid of each cluster is updated at each iteration
- The algorithm keeps iterating until the centroid don't change

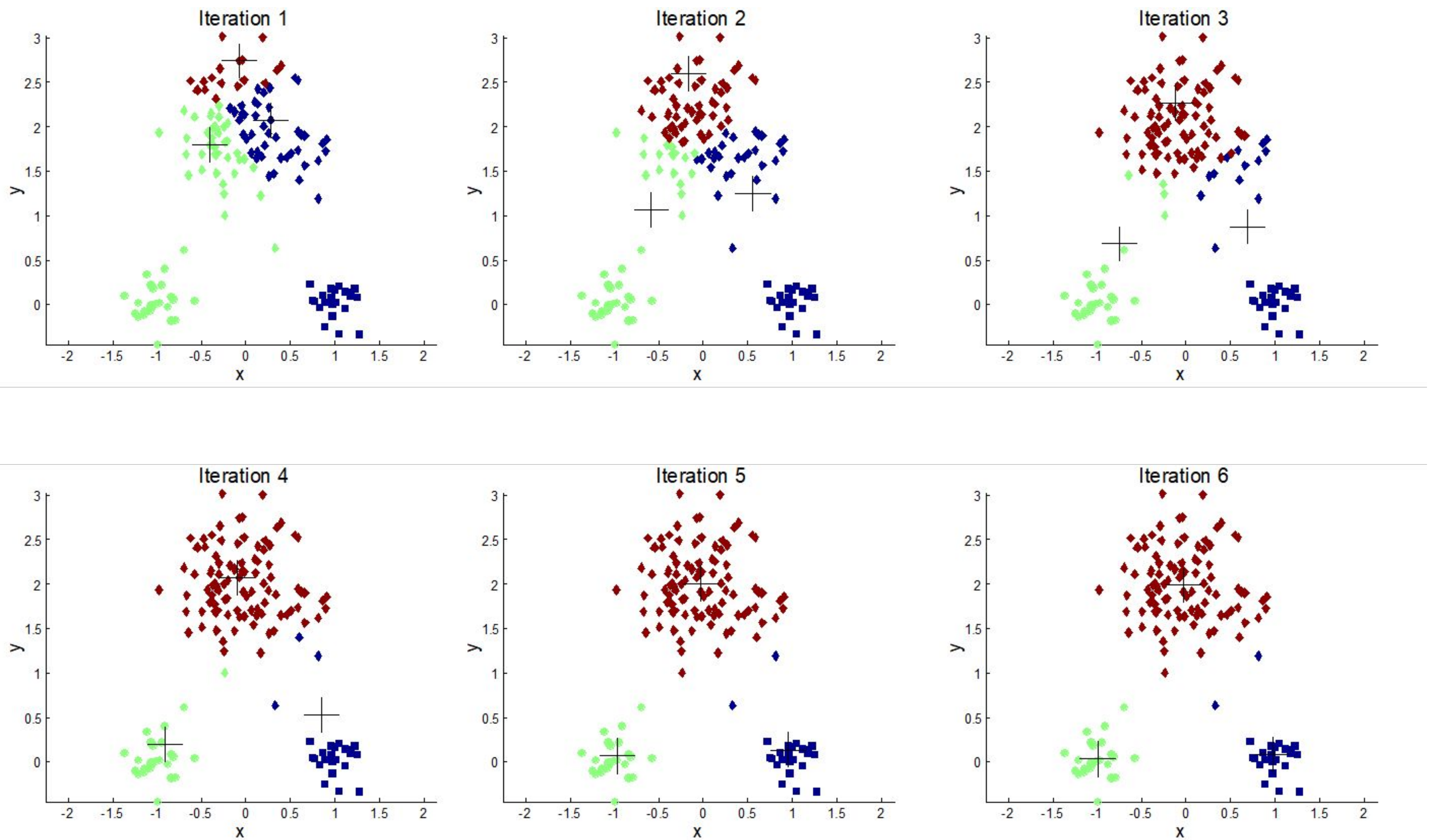
---

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

Example of *K-Means* (K=2)



Example of *K-Means* (K=3)



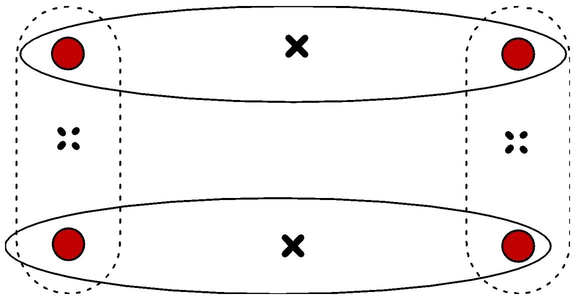
Strength and Weakness of K-Means

- **Strength:** *Fast*: $O(tkn)$
  - ***n***: number of objects, ***k***: number of clusters, ***t:*** number of iterations
  - Normally: ***k***, ***t*** << ***n***
- **Weakness**
  - Need to specify ***k***, the *number* of clusters, in advance
  - The random choice of the first ***k*** centroids may result in different clustering
  - Applicable only to objects in a continuous n-dimensional space
  - Often terminates at a local optimal
  - Sensitive to noisy data and outliers
  - Not suitable to discover clusters with non-convex shapes

How to improve the K-Means algorithm?

Variations of the *K-Means* Algorithm

- Most of the variants of the *k-means* differ in:
  - Selection of the initial ***k*** means
  - Dissimilarity calculations
  - Strategies to calculate cluster means



- Handling categorical data with *k-modes:*
  - Replacing means of clusters with <u>modes</u>
  - Using new dissimilarity measures to deal with categorical objects
  - Using a <u>frequency</u>-based method to update modes of clusters
  - A mixture of categorical and numerical data: *k-prototype* method

Outline

- Overview of Clustering
- Major Clustering Approaches
  - K-means Clustering
  - Hierarchical Clustering
  - DBSCAN Clustering
- Cluster Evaluation