



UNIVERSIDADE
CATÓLICA
PORTUGUESA

BRAGA

Machine Learning

Session 5 - T

Unsupervised Learning - Clustering

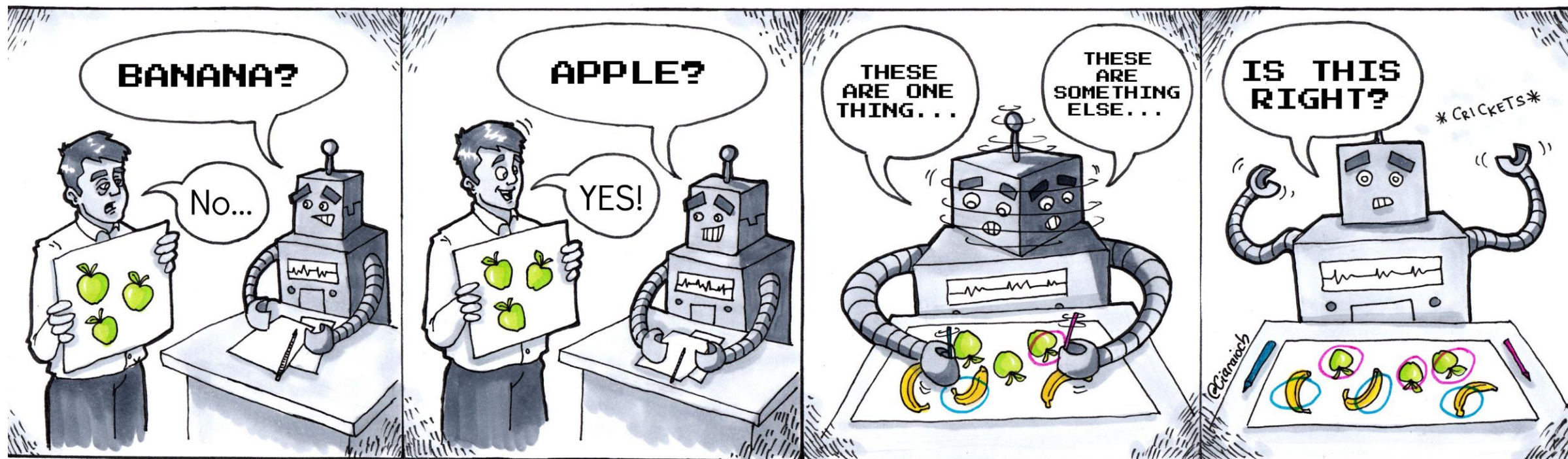
Degree in Applied Data Science

2024/2025

Unsupervised vs Supervised Learning

- **Unsupervised:** involves working with **unlabeled data**, where the algorithm explores the inherent **structure and patterns** within the input without explicit output guidance.
- **Supervised:** the algorithm is trained on a **labeled dataset**, where the input data is paired with corresponding output labels. The goal is to learn a **mapping from inputs to outputs**, allowing the algorithm to make predictions on new, unseen data.

Unsupervised vs Supervised Learning



Supervised Learning

Unsupervised Learning

Illustration by [@Ciaraioch](#)

Unsupervised Learning

- What can we do in the absence of target labels?
 - Group data based on similarity \Rightarrow **clustering**
 - Simplify/reduce data \Rightarrow **dimensionality reduction**
 - Visualize data \Rightarrow **data visualization**

Supervised

X_1	X_2	X_p	Y

Target

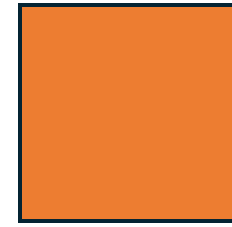
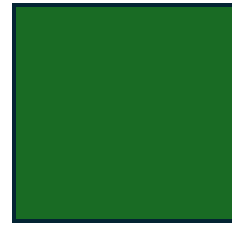
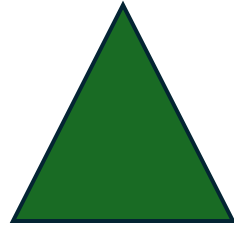
Unsupervised

X_1	X_2	X_p	Y

No
Target

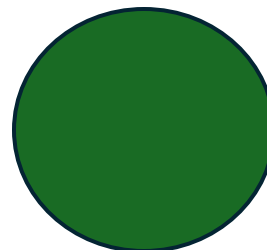
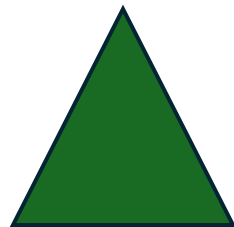
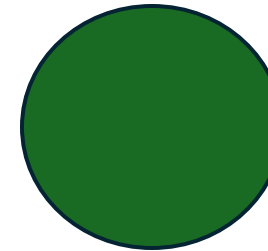
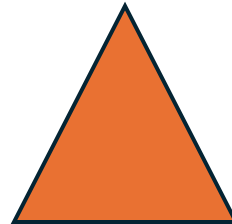
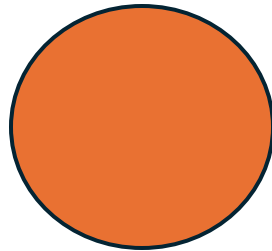
- Clustering is not a well-defined term with varying definitions in the literature:
 - Finding **groups in data**;
 - Dividing data into homogeneous groups;
 - Dividing data into groups where **points within each group are close or similar**;
 - Dividing data into groups where points within each group are close or similar, and **points of different groups are far or dissimilar**;
 - Dividing the feature space into regions with relatively **high density of points**, separated by regions with relatively **low density of points**.

Our First Clustering Task



How many clusters?

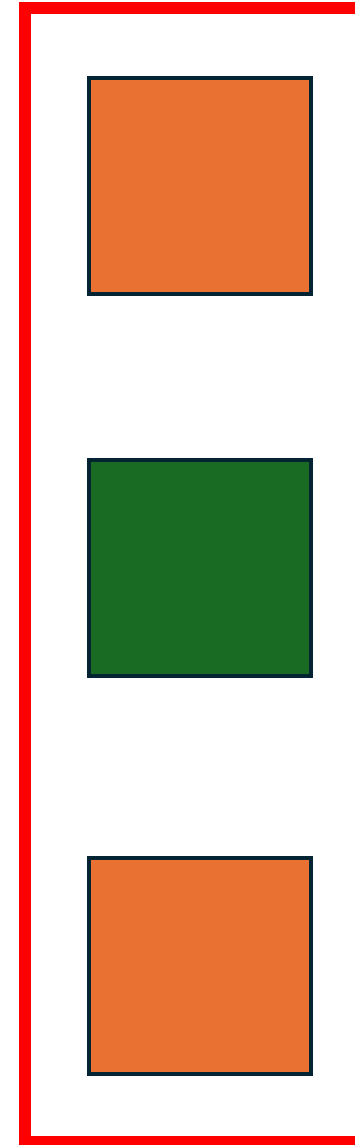
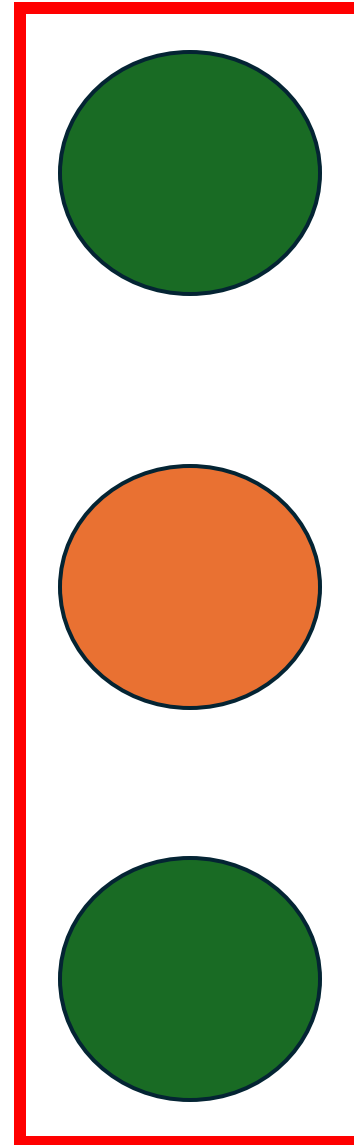
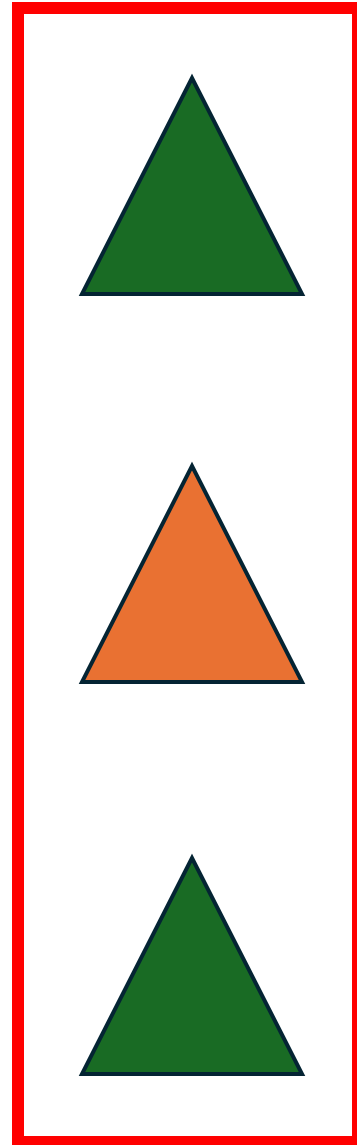
Which forms belong to
each cluster?



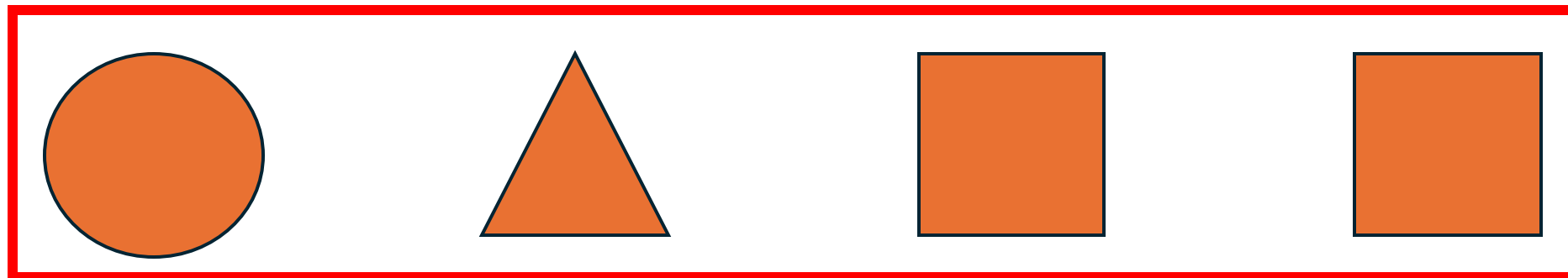
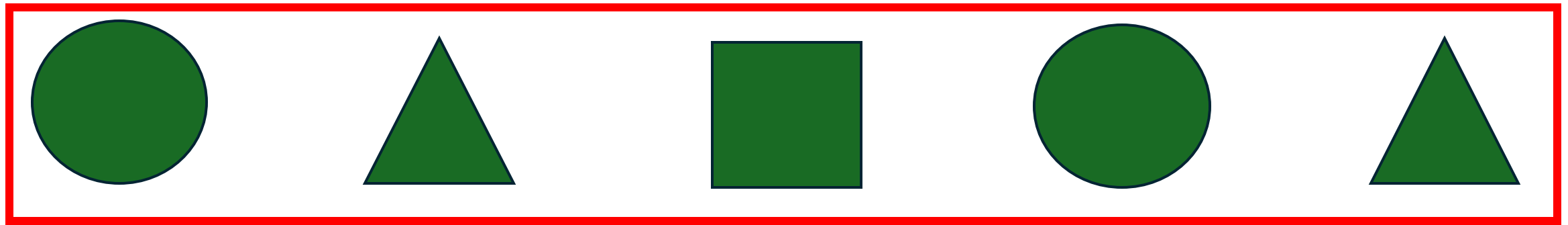
Our first clustering task



3 clusters
by shape



Our first clustering task



2 clusters

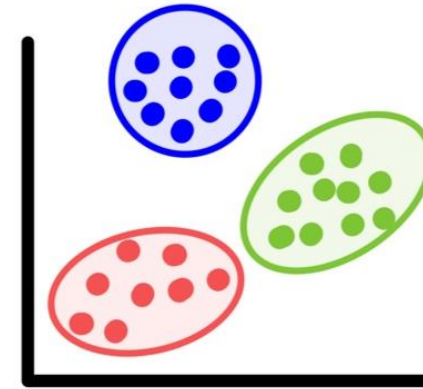
by color

Clustering

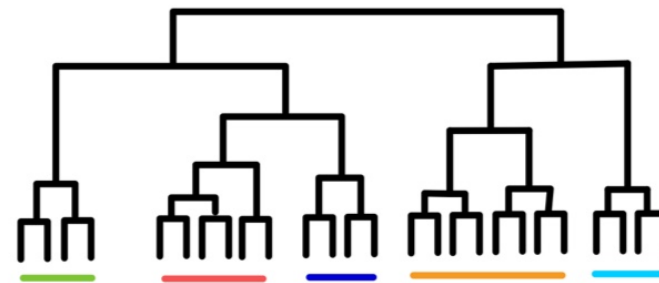
- Mathematically, clustering can be defined as a set of **optimization problems** with several variants.
- The clustering technique to use depends on:
 - **Data type** (e.g. numeric, nominal);
 - Desired **output format** (e.g. exclusive clusters, probabilities, hierarchies);
 - **Objective function** (e.g. homogeneity vs separation);
 - **Similarity/distance measure** (e.g. euclidean, manhattan distance).

Defining Clusters

- **Partitional clustering:**
 - Data is divided into groups at the same level.

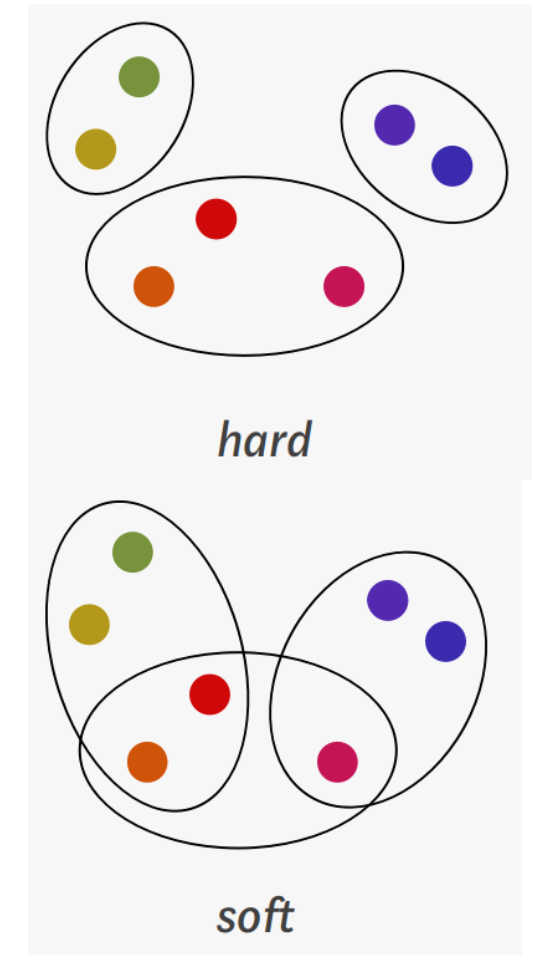


- **Hierarchical clustering:**
 - Clusters are nested within larger clusters, in a tree.



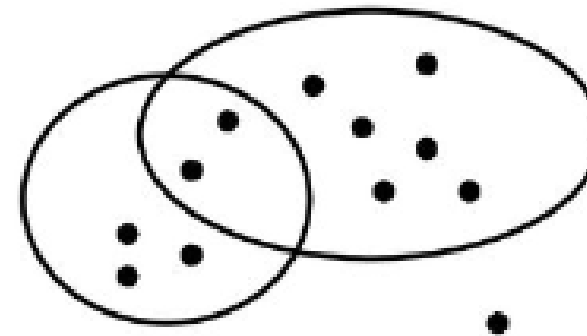
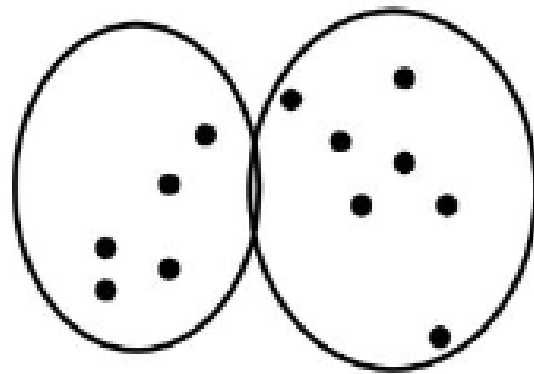
Clustering Membership

- **Hard clustering:**
 - Each example belongs only to one cluster.
- **Soft clustering:**
 - Examples may belong to more than one cluster.
- **Fuzzy clustering:**
 - Each example belongs to clusters with probabilities.



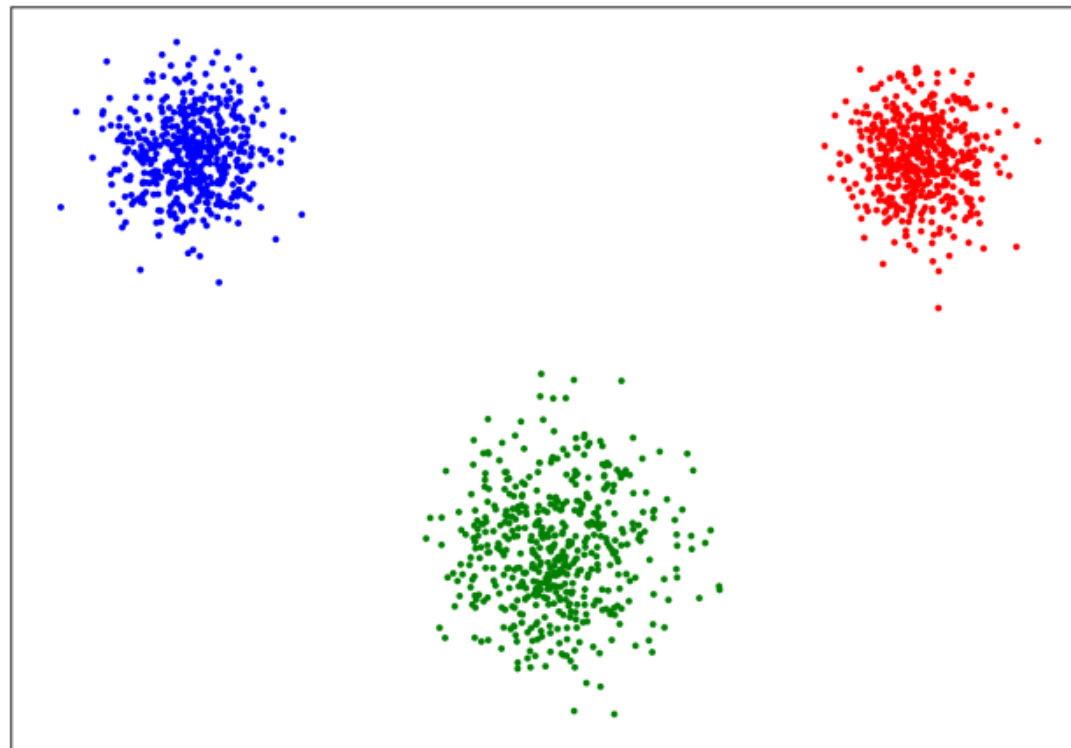
Clustering Coverage

- **Complete clustering:**
 - All examples are assigned to cluster (or clusters);
- **Partial clustering:**
 - Some examples unassigned (e.g. noise, irrelevant data)



Types of Clustering

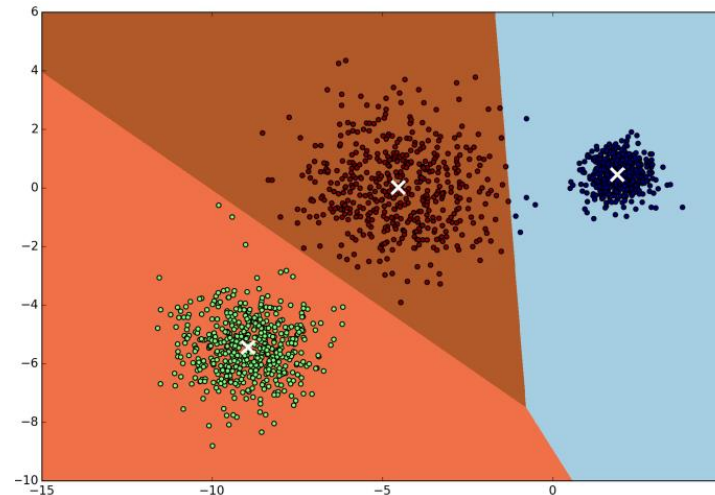
- **Well separated clusters:**
 - Distance between any two points in different clusters is larger than the distance between any two points in the same group.



Types of Clustering

- **Prototype-based clustering:**

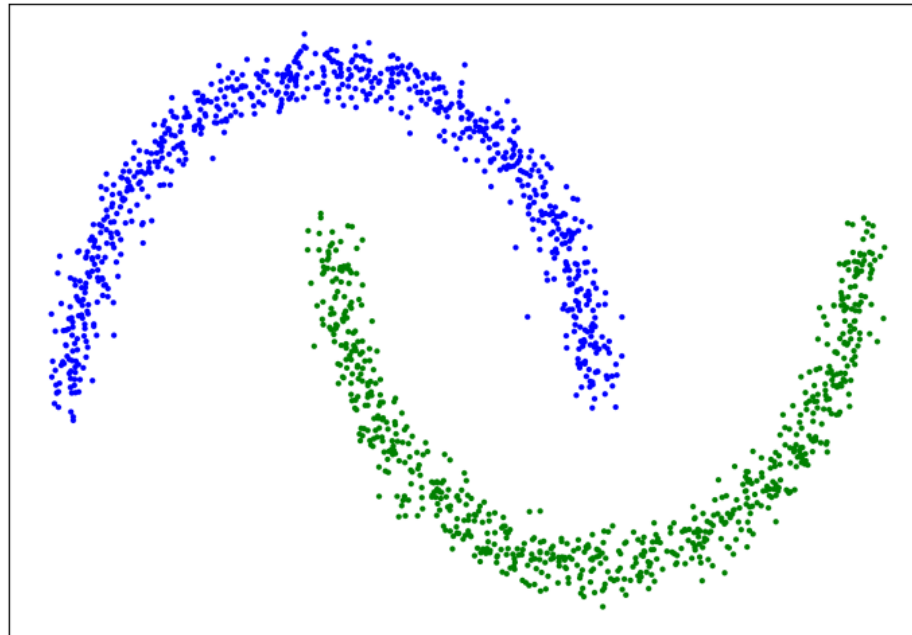
- Examples in a cluster are closer to the prototype of the cluster than to the prototype of any other cluster.
- If the data is **numerical**, the prototype of the cluster is often a **centroid** i.e., the average of all the points in the cluster.
- If the data has **categorical** attributes, the prototype of the cluster is often a **mode** i.e., the most representative point of the cluster.



Types of Clustering

- **Contiguity-based clustering:**

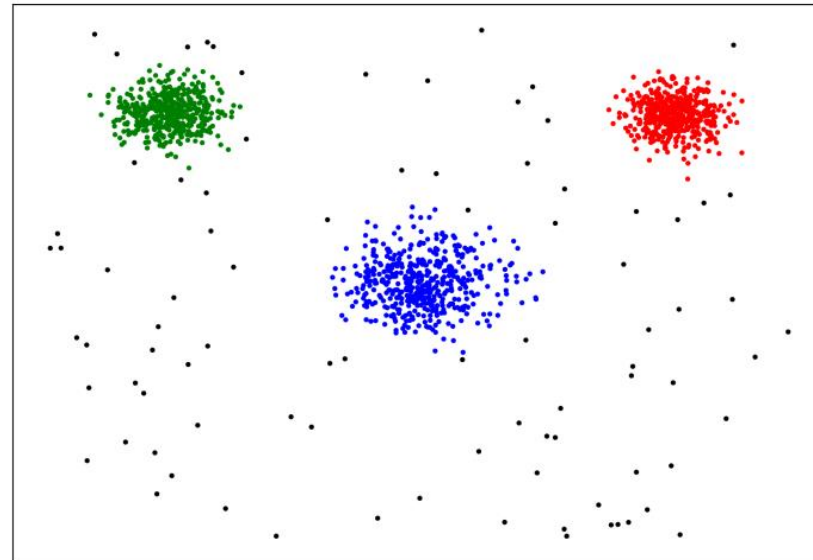
- Each example in a cluster is closer to at least one example in the same cluster than to any example in a different cluster.
- Useful when clusters are irregular and intertwined.
- Does not work well when there is noise in the data.



Types of Clustering

- **Density-based clustering:**

- Cluster is a dense region of examples that is surrounded by a region of low density.
- Used when the clusters are irregular, intertwined and when noise and outliers are present.
- Examples in low density region are classified as noise and omitted.



Similarity

- Sometimes is difficult to determine what is similar or not!
- Distance measures:

- **Euclidean distance:**

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$

- **Manhattan distance:**

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$

Similarity

- Sometimes is difficult to determine what is similar or not!
- Similarity measures:
 - **Jaccard similarity:**

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

- **Pearson correlation:**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Homogeneity and Separation

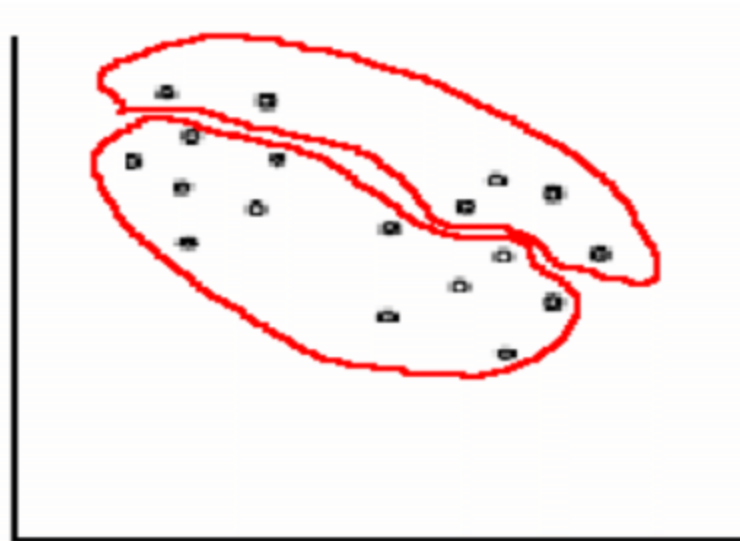
- **Homogeneity:** elements within a cluster must be close to each other (low distances) – intra-cluster
- **Separation:** elements in different clusters should be quite separate from each other (high distances) - inter-cluster

Which clustering would you choose (and why)?

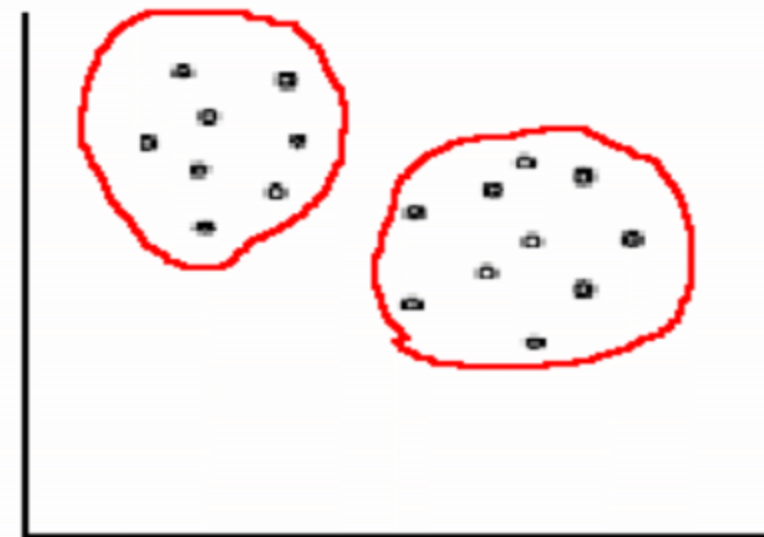


UNIVERSIDADE
CATOLICA
PORTUGUESA

BRAGA

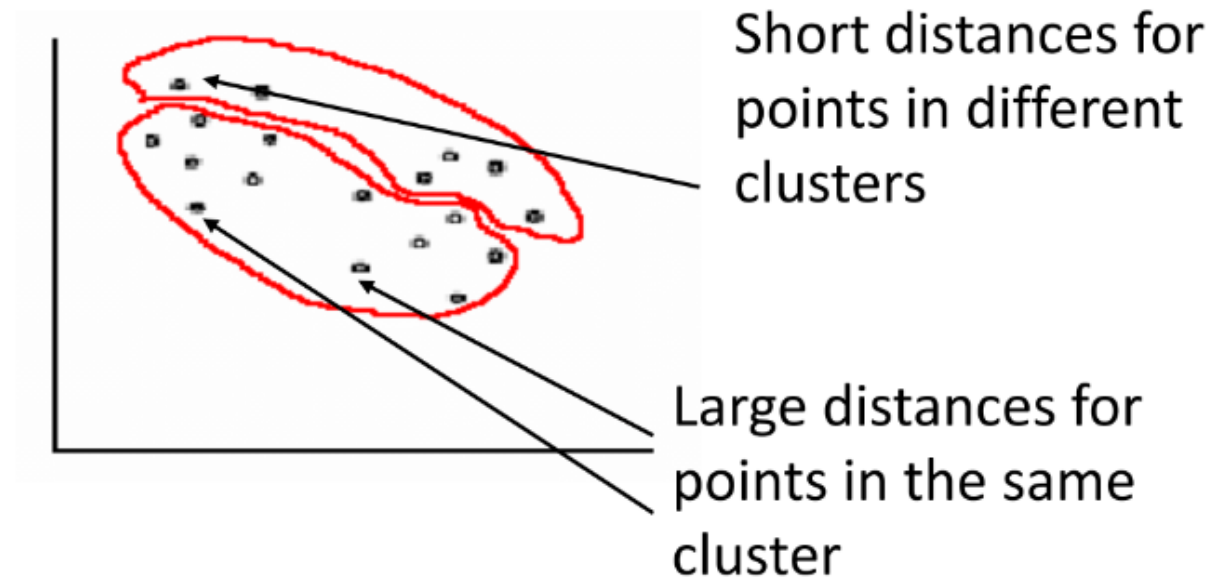


VS



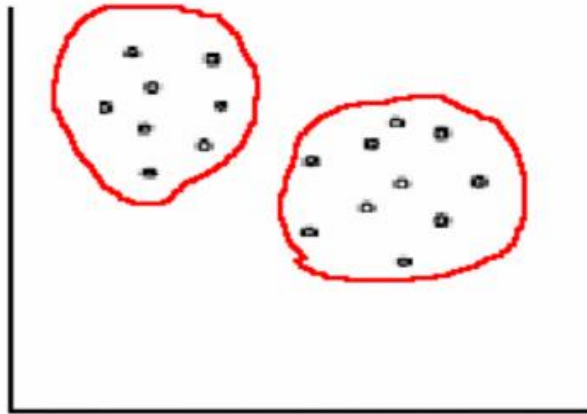
Clustering: "bad" solution

Violates homogeneity and separation



Clustering: "good" solution

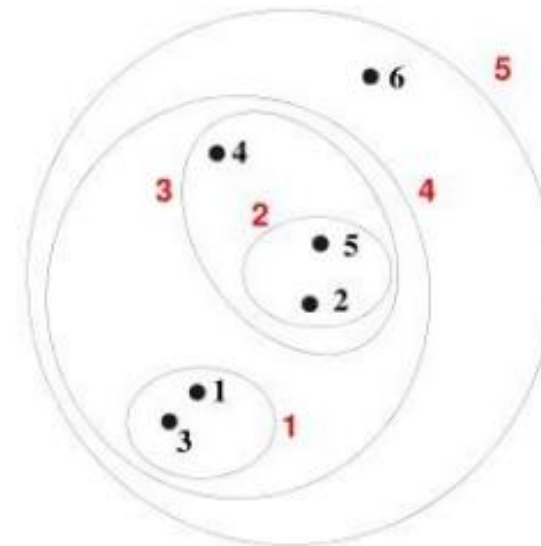
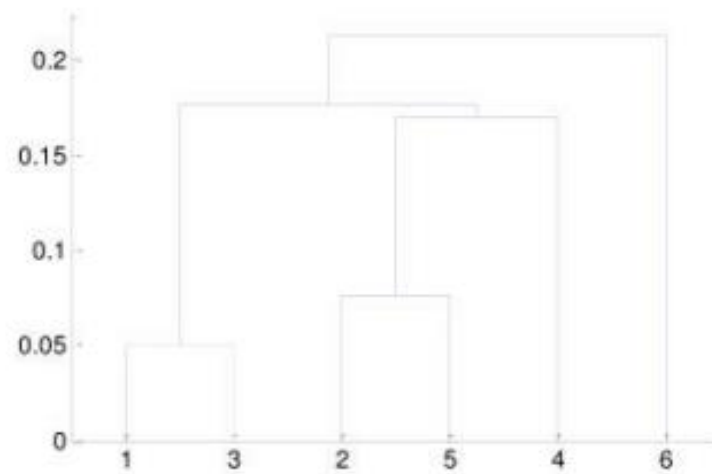
Solution with good homogeneity and separation



Still, some pairs of points would be better grouped together in the previous solution

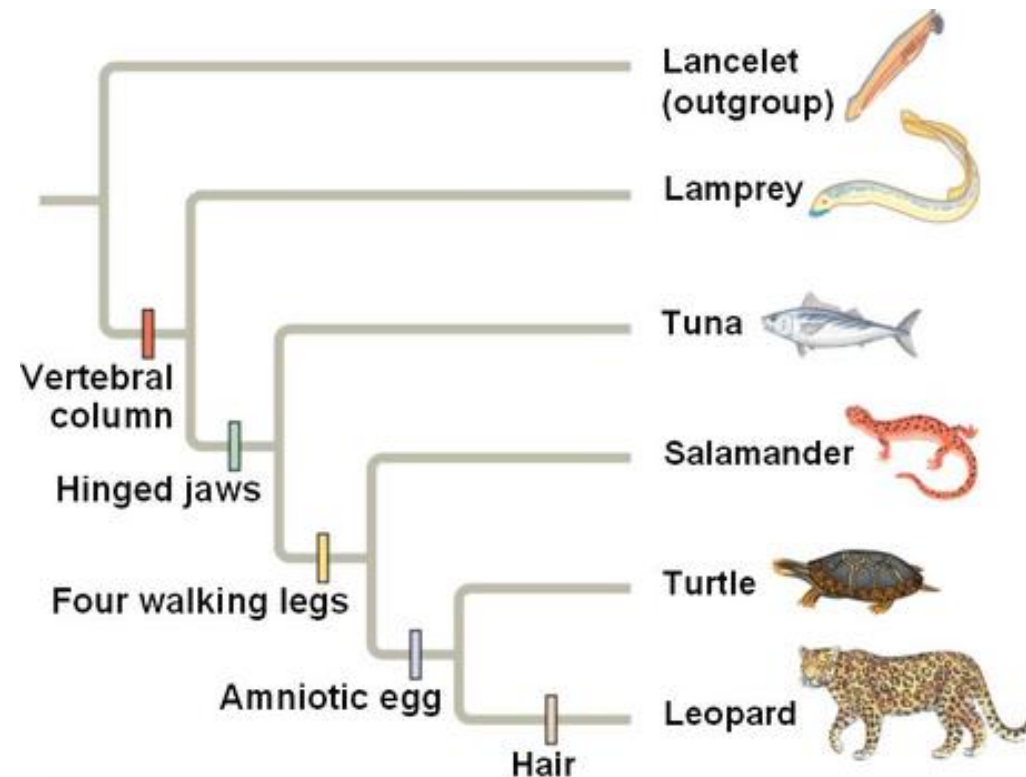
Hierarchical Clustering

- Generates a set of **nested clusters** organized as a **hierarchical tree**;
- Visual representation often depicted as a **dendrogram**:
 - A tree like diagram representing a hierarchy of nested clusters
 - Clustering obtained by cutting at desired level



Hierarchical Clustering - Advantages

- Do not have to assume any particular number of clusters;
- May correspond to meaningful taxonomies.



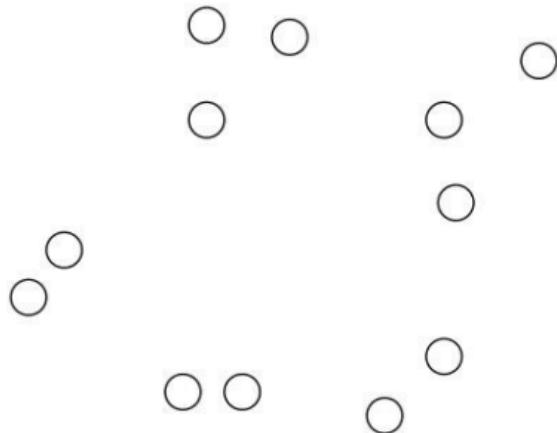


Hierarchical Clustering Types

- **Agglomerative:**
 - Start with the examples as individual clusters;
 - At each step, merge the closest pair of clusters until only one cluster (or k clusters) left.
- **Divisive:**
 - Start with one cluster with all examples;
 - At each step, split a cluster until each cluster contains a example (or there are k clusters).

Agglomerative Hierarchical Clustering

1. Compute the proximity matrix;
2. Let each example be a cluster;
3. Merge the two closest clusters;
4. Update the proximity matrix;
5. Repeat 3 and 4 until a single cluster remains (or k clusters).



	p1	p2	p3	p4	p5	...
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

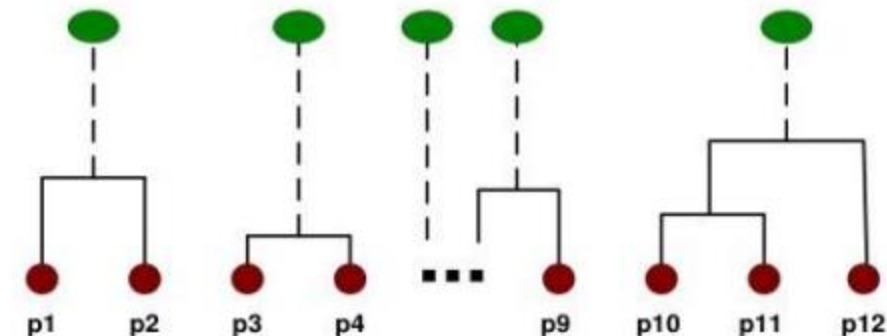
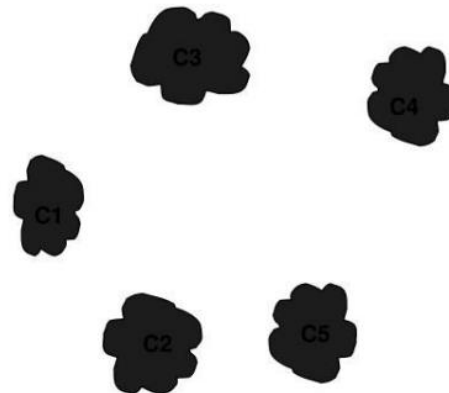


Agglomerative Hierarchical Clustering

1. Compute the proximity matrix;
2. Let each example be a cluster;
3. **Merge the two closest clusters;**
4. **Update the proximity matrix;**
5. Repeat 3 and 4 until a single cluster remains (or k clusters).

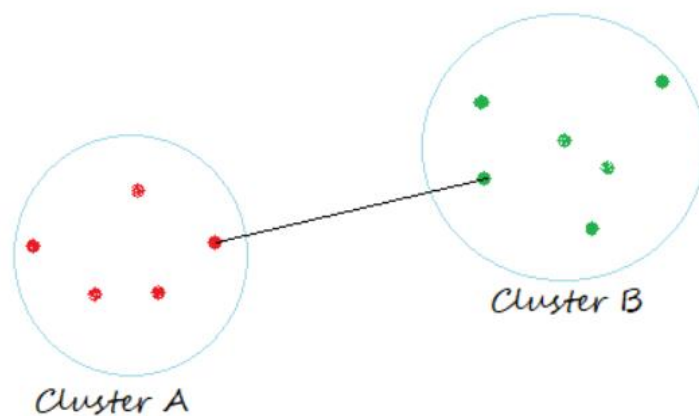
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

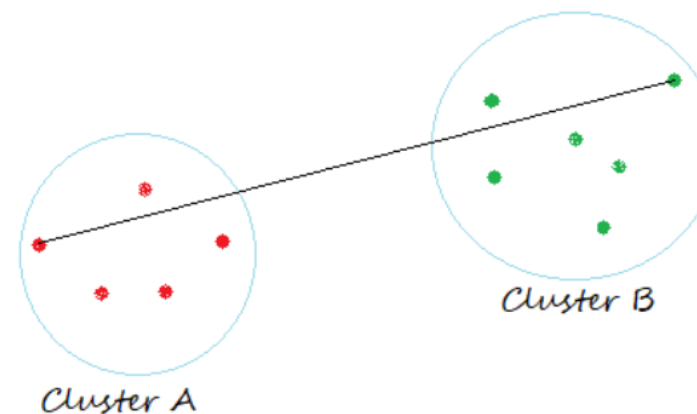


Hierarchical Clustering – Cluster Similarity

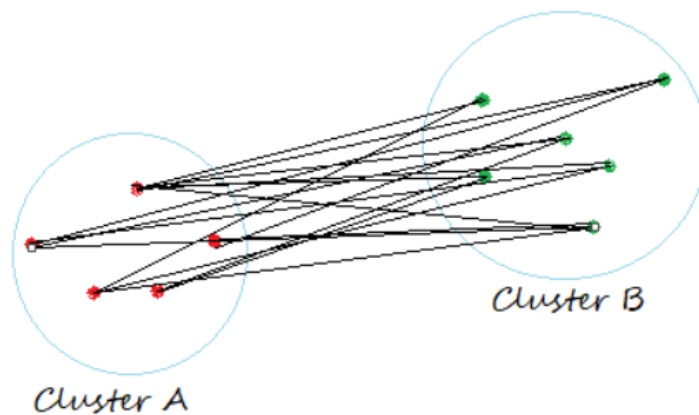
Single Linkage



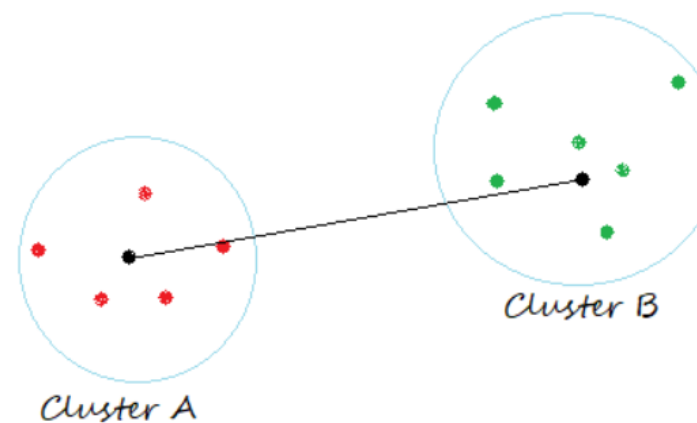
Complete Linkage



Average Linkage



Centroid Linkage

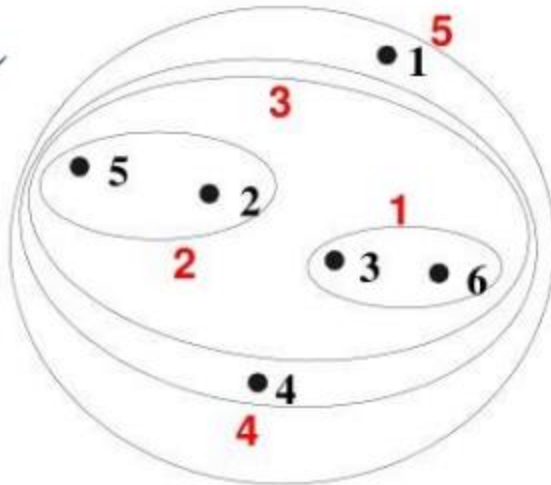


Hierarchical Clustering – Single vs Complete vs Average

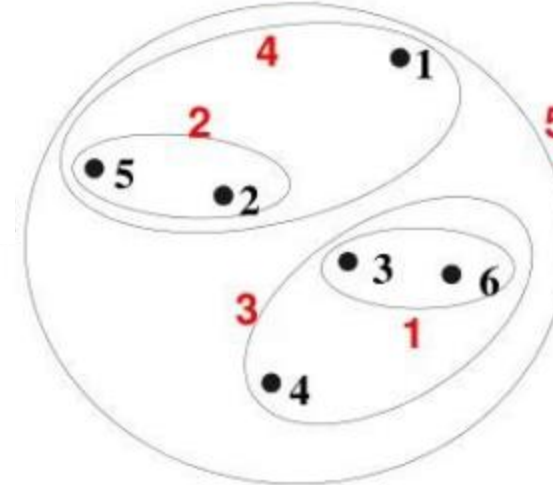


UNIVERSIDADE
CATOLICA
PORTUGUESA
BRAGA

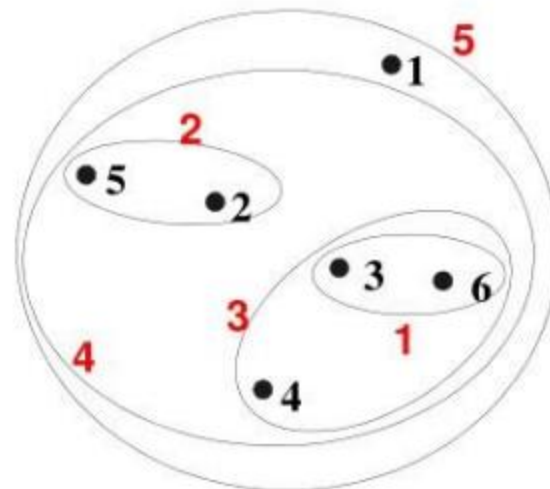
Single Linkage



Complete Linkage



Average Linkage



Hierarchical Clustering – Single vs Complete vs Average

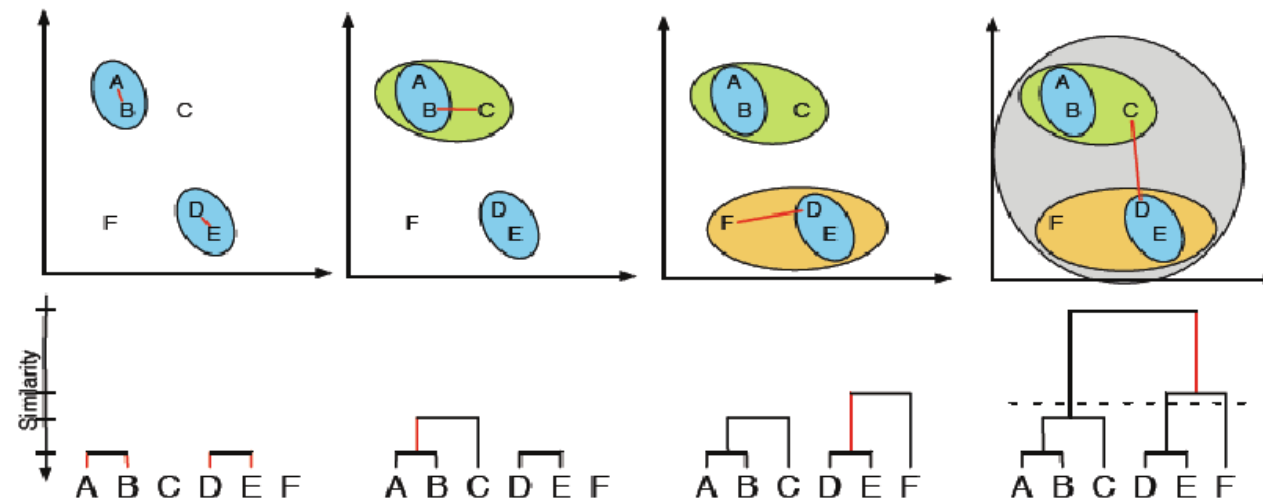
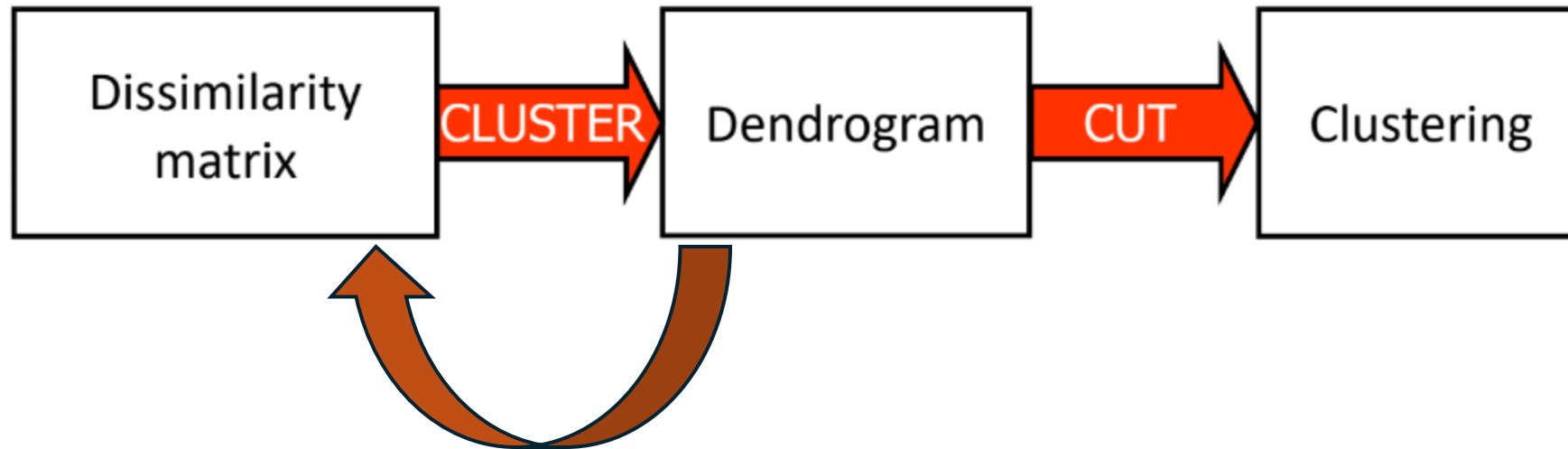


- **Single Linkage:**
 - Can handle non-elliptical shapes;
 - Sensitive to noise and outliers.
- **Complete Linkage:**
 - Less susceptible to noise and outliers;
 - Tends to break large clusters;
 - Biased towards globular clusters.
- **Average Linkage:**
 - Compromise between single and complete linkage;
 - Less susceptible to noise and outliers;
 - Biased towards globular clusters.

Hierarchical Clustering – Limitations

- **Do not scale well:**
 - Space complexity: $O(N^2)$
 - Time complexity: $O(N^3)$
 - $O(N^2 \log(N))$ for some approaches.
- **Cannot undo what was previously done;**
- **Quality varies** a lot in terms of distance measure used.

Hierarchical Clustering - Summary

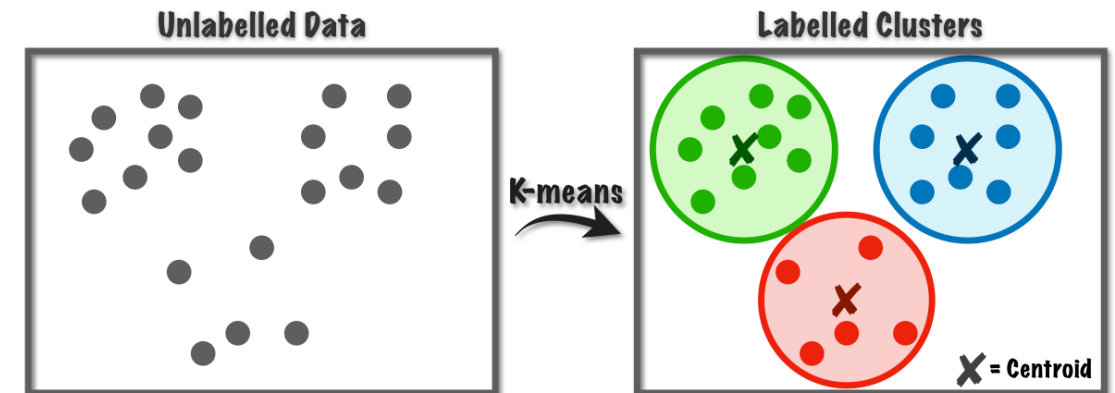


Hierarchical Clustering – Applications

- **Biology:** Used in genomics for grouping genes with similar expression patterns, clustering protein sequences, and phylogenetic analysis.
- **Marketing:** Segmentation of customers based on purchasing behavior or demographic characteristics to tailor marketing strategies.
- **Image Analysis:** Grouping similar images together for tasks such as image retrieval, object recognition, and image compression.
- **Document Clustering:** Organizing documents by topic for information retrieval, text mining, and document summarization.
- **Anomaly Detection:** Identifying outliers or unusual patterns in data, such as detecting fraudulent transactions or network intrusions.
- **Finance:** Segmenting financial data for portfolio optimization, risk assessment, and fraud detection.

K-Means Clustering

- Find best clustering of **k clusters**
 - Partitional, exclusive, complete and prototype-based;
 - Define clusters by proximity to the mean of the cluster (centroid);
 - The number of clusters is predefined (k).
- **Objective:** find the centroids that **minimize the distance between the examples and the centroids.**

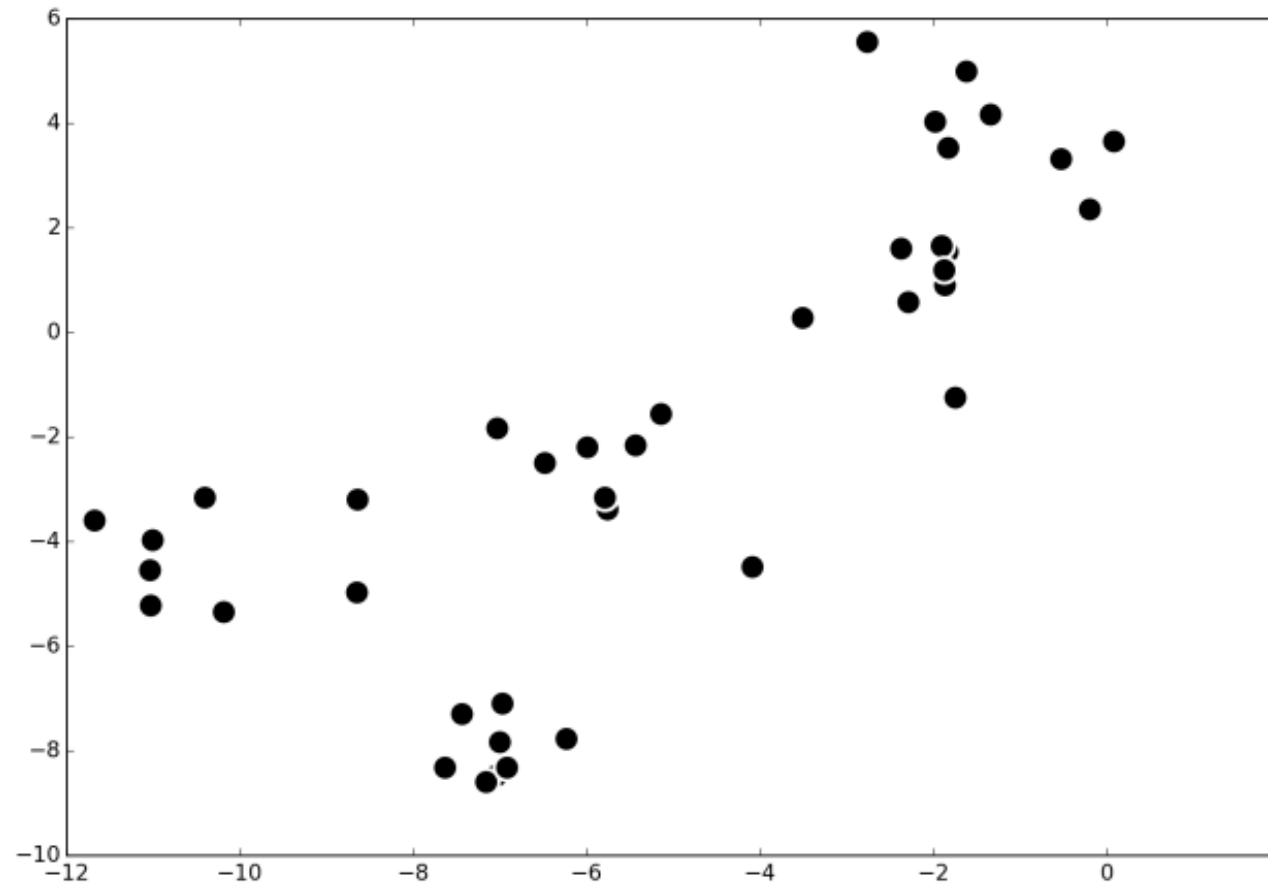


K-Means Clustering: Lloyd Algorithm

- Because the optimal solution for this problem is NP-hard, practical useful solutions can be obtained with simple heuristic algorithms such as the **Lloyd algorithm**.
- **Lloyd algorithm:**
 - Start with random centroids;
 - Assign each example to the closest centroid;
 - Update centroids to mean of respective cluster;
 - Recompute clusters and repeat until convergence.
- **Does not guarantee optimal solutions** as in practical implementations a maximum number of iterations is commonly defined.

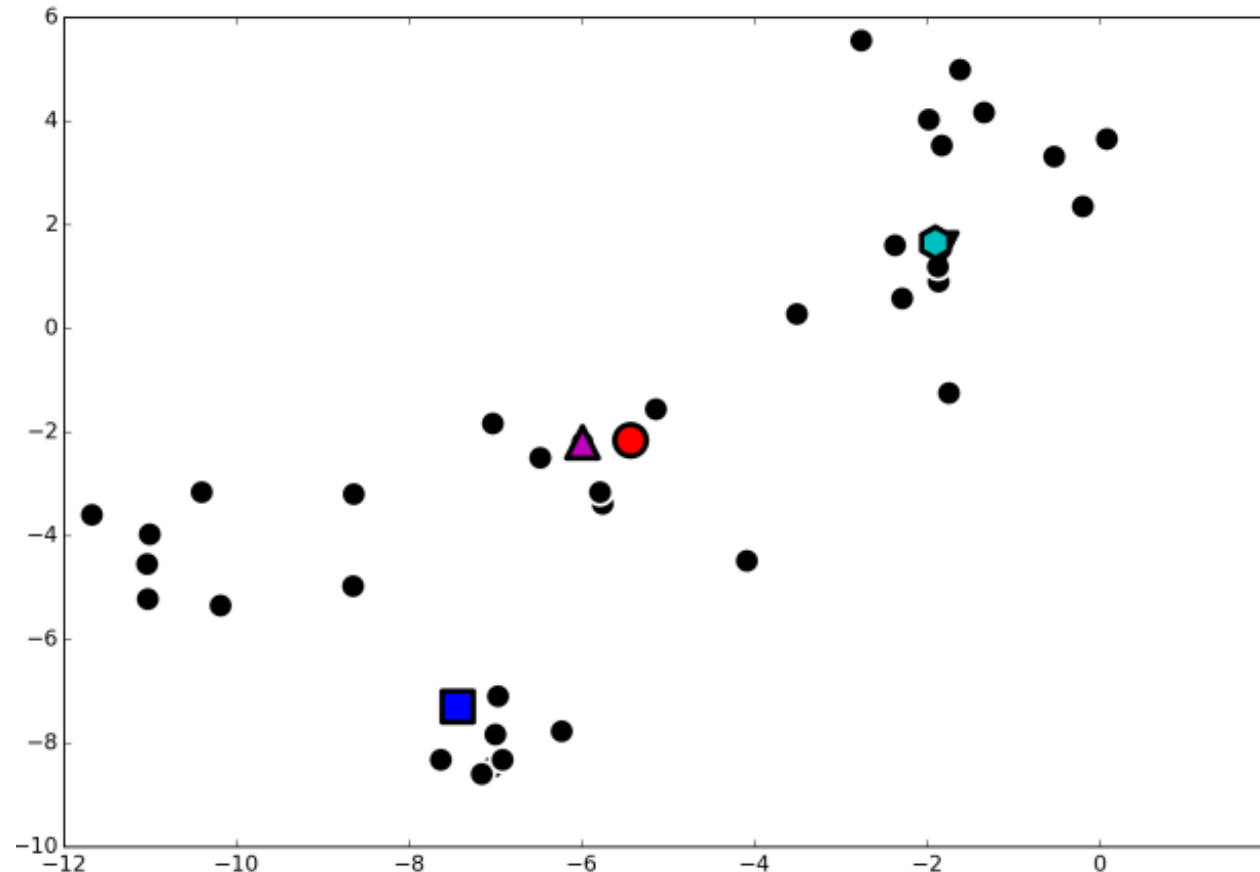
K-Means Clustering: Initialization

- **Forgy:** start with coordinates of a random set of k examples



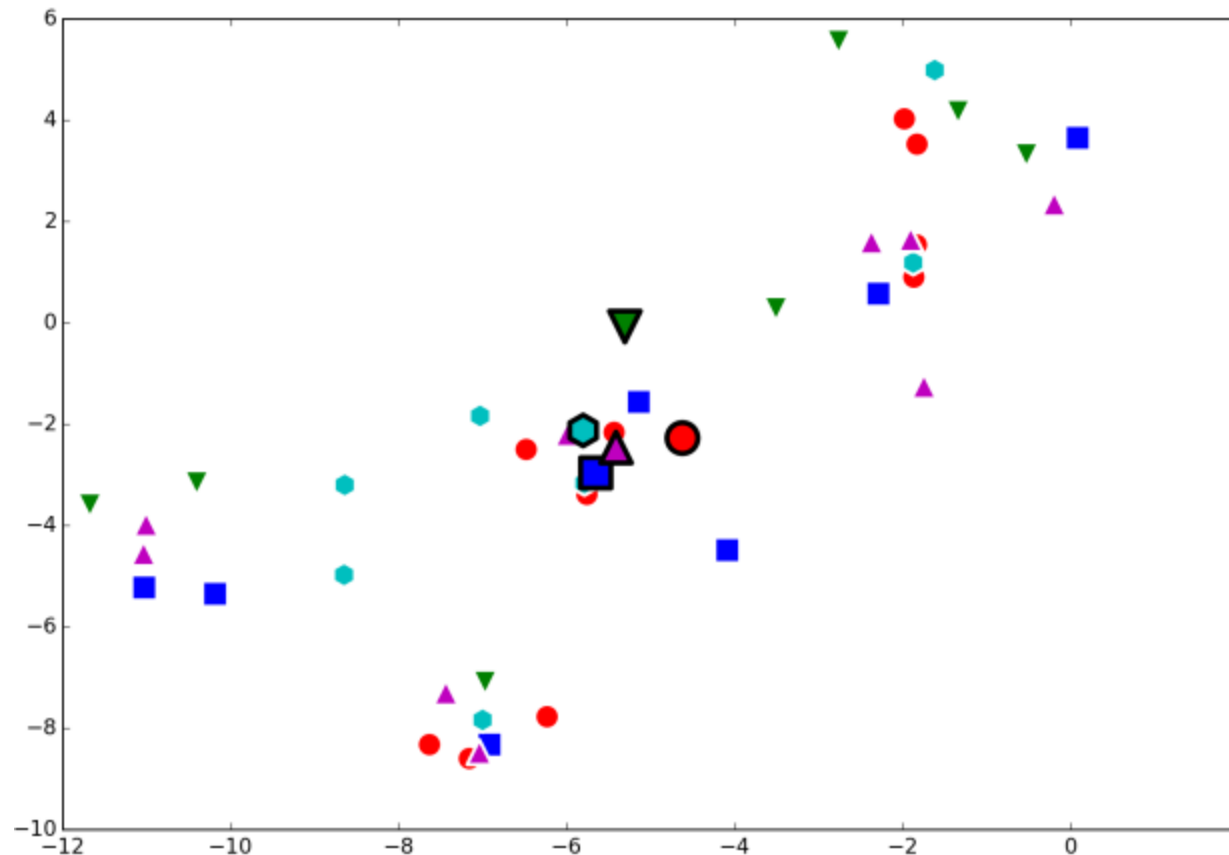
K-Means Clustering: Initialization

- **Forgy:** start with coordinates of a random set of k examples



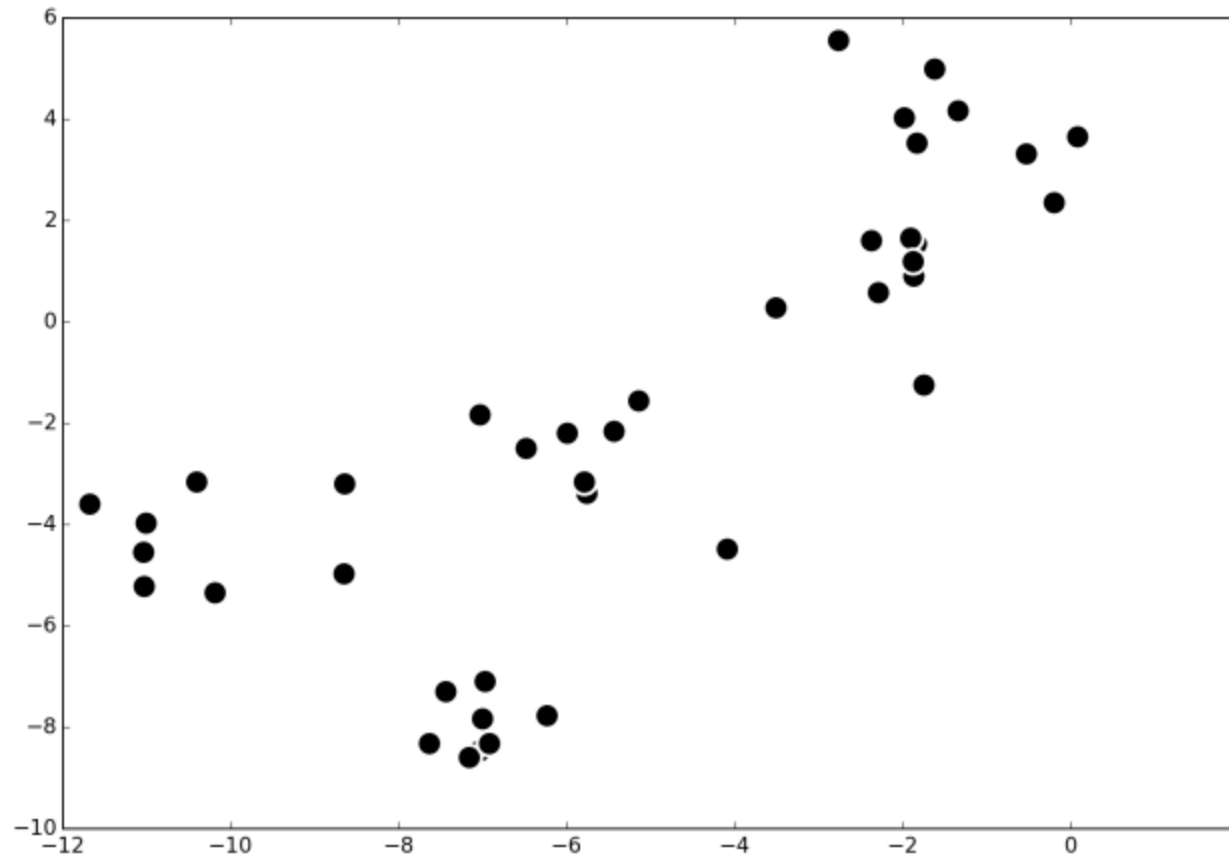
K-Means Clustering: Initialization

- **Random:** random assignment, compute means



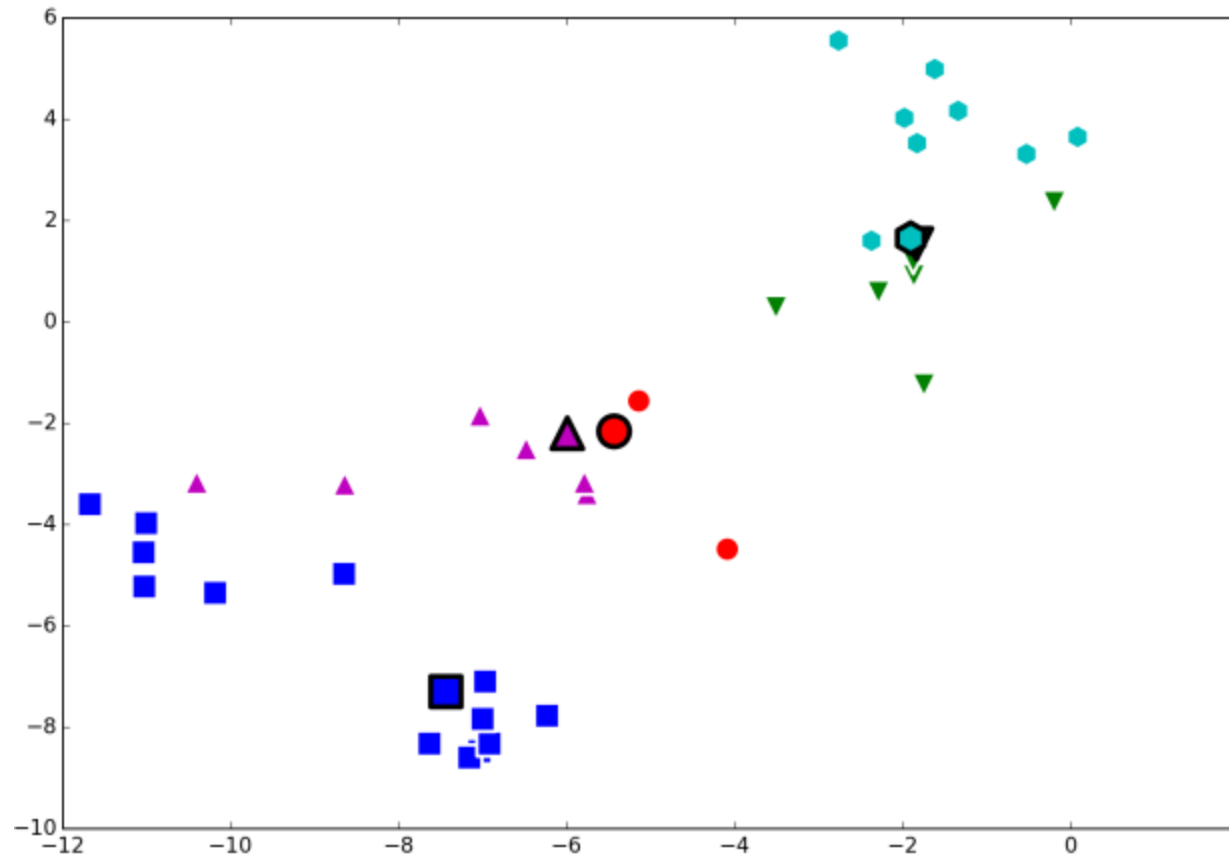
K-Means Clustering: Lloyd Algorithm

- Original data



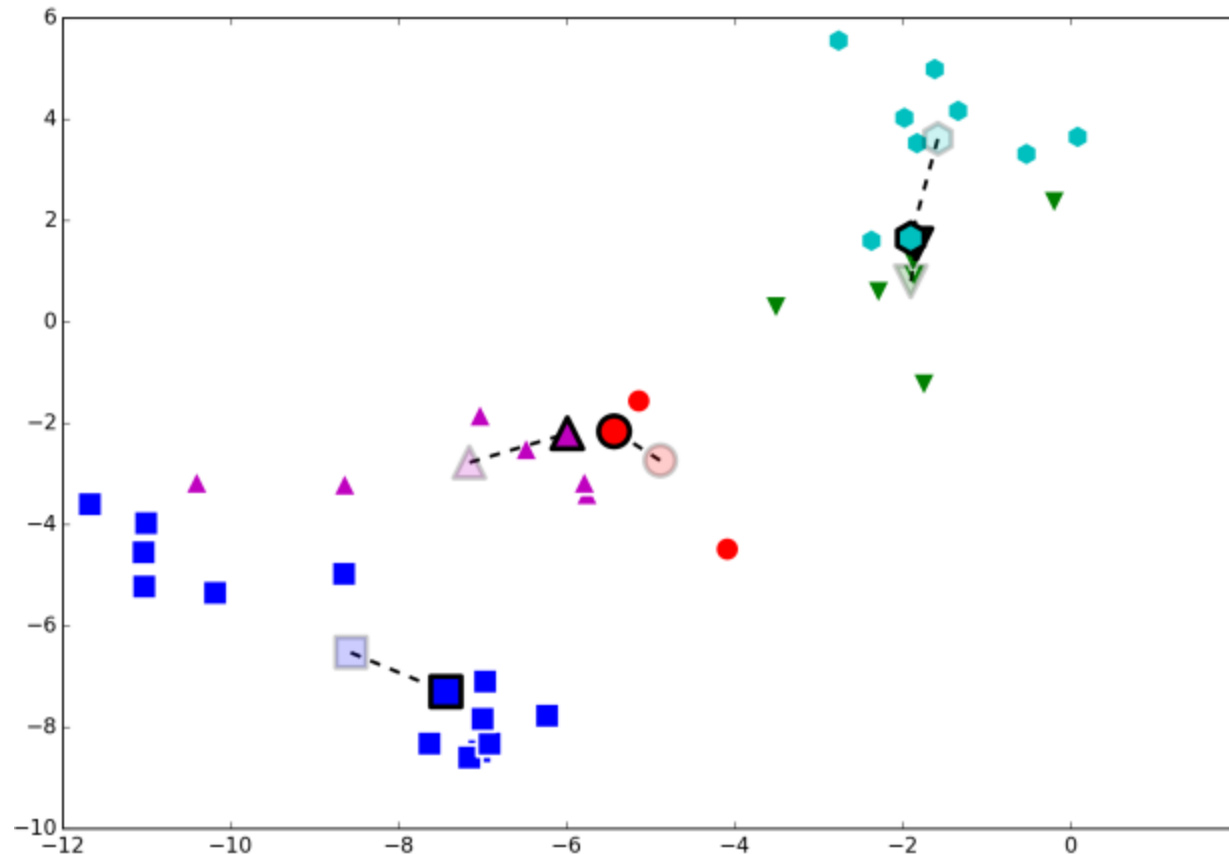
K-Means Clustering: Lloyd Algorithm

- Initialize (Forgy), compute clusters



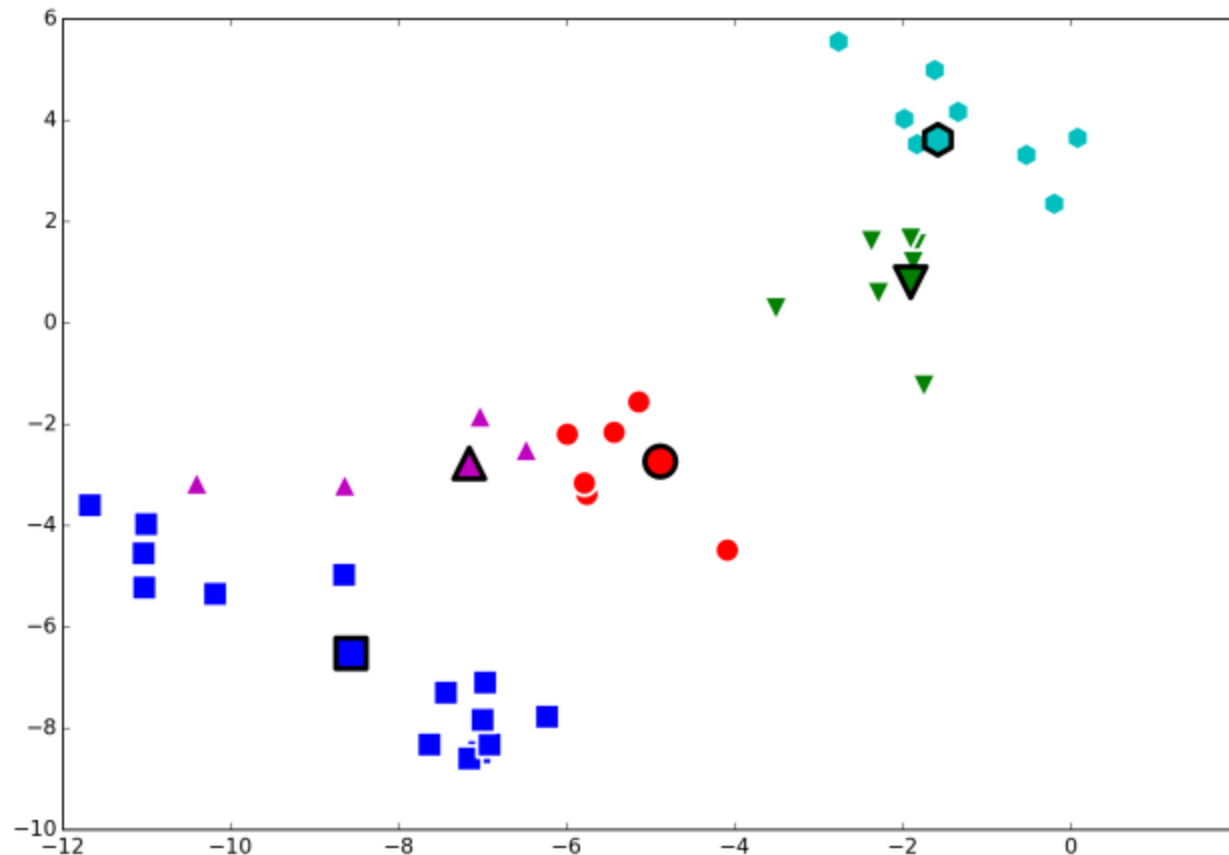
K-Means Clustering: Lloyd Algorithm

- Compute new means, update centroids



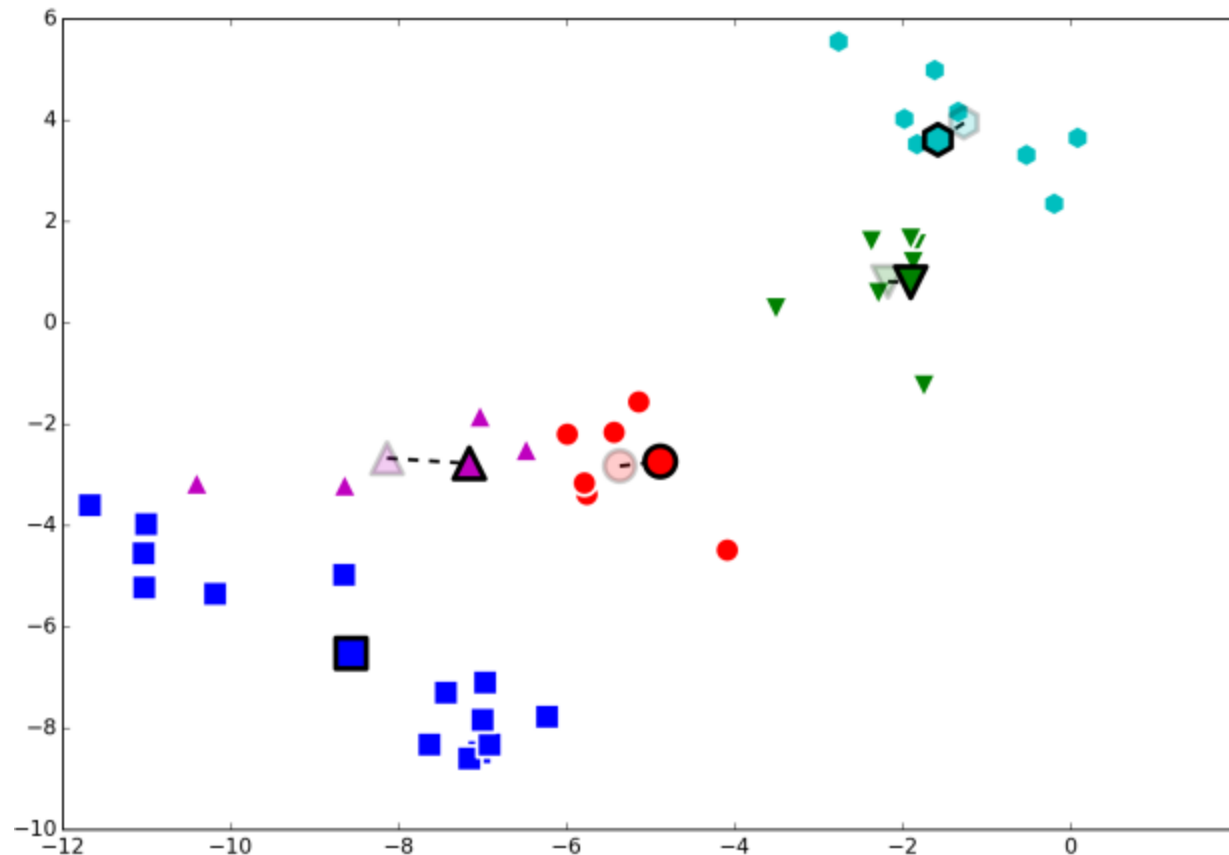
K-Means Clustering: Lloyd Algorithm

- Recompute clusters



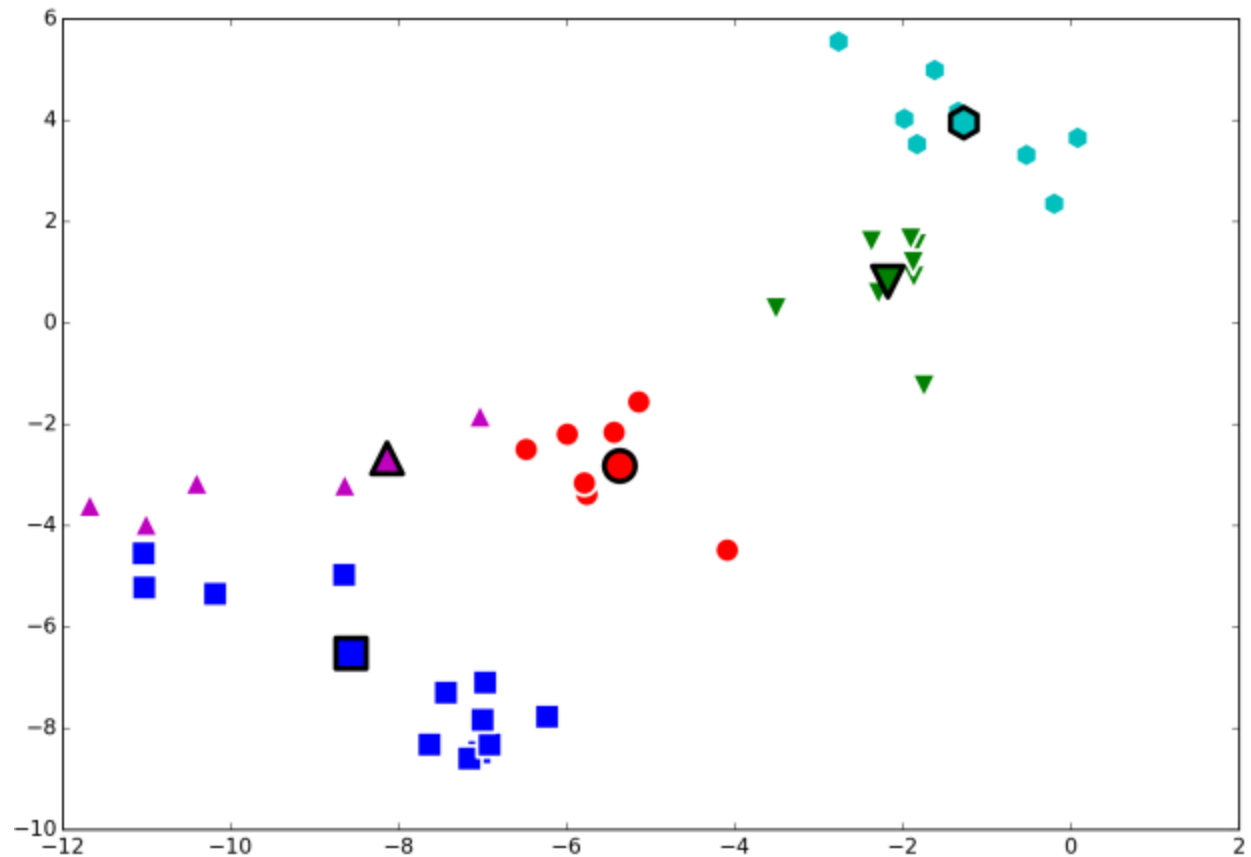
K-Means Clustering: Lloyd Algorithm

- Compute new means, update centroids



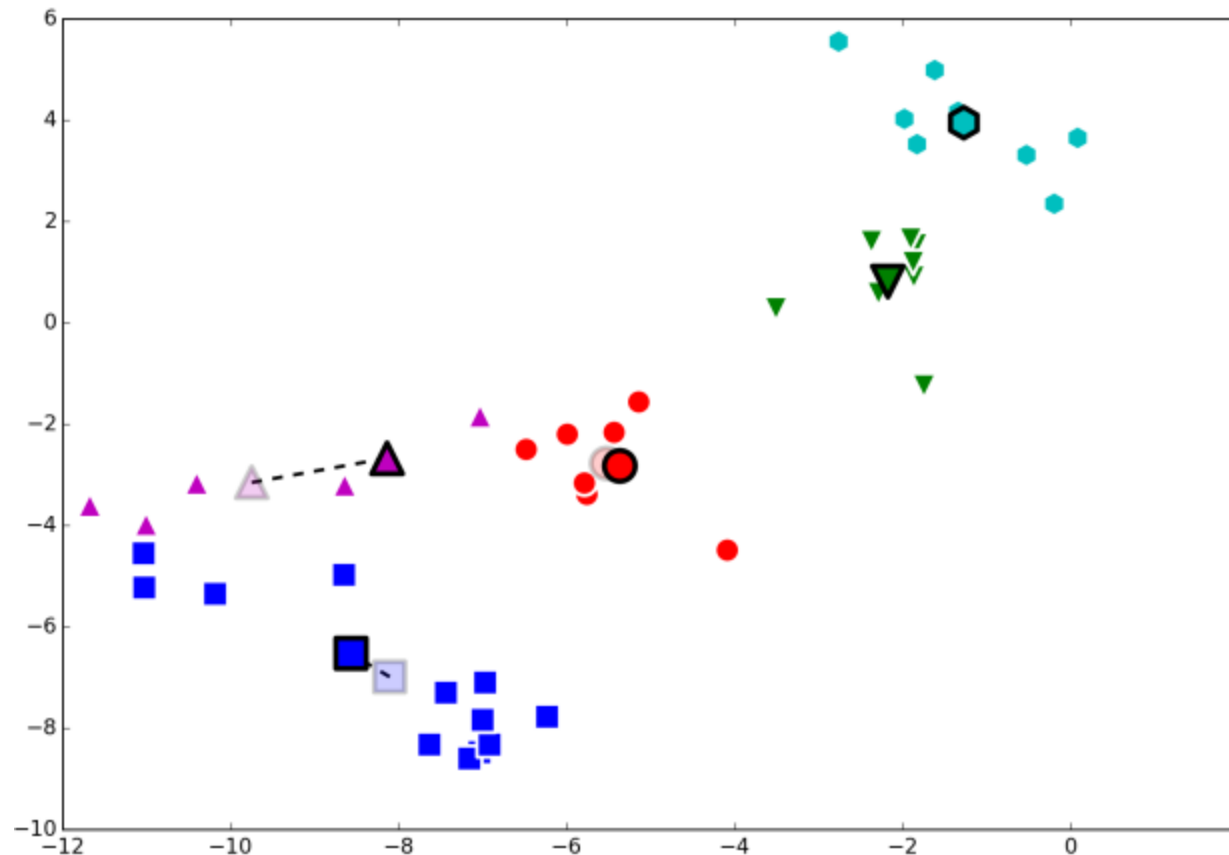
K-Means Clustering: Lloyd Algorithm

- Recompute clusters



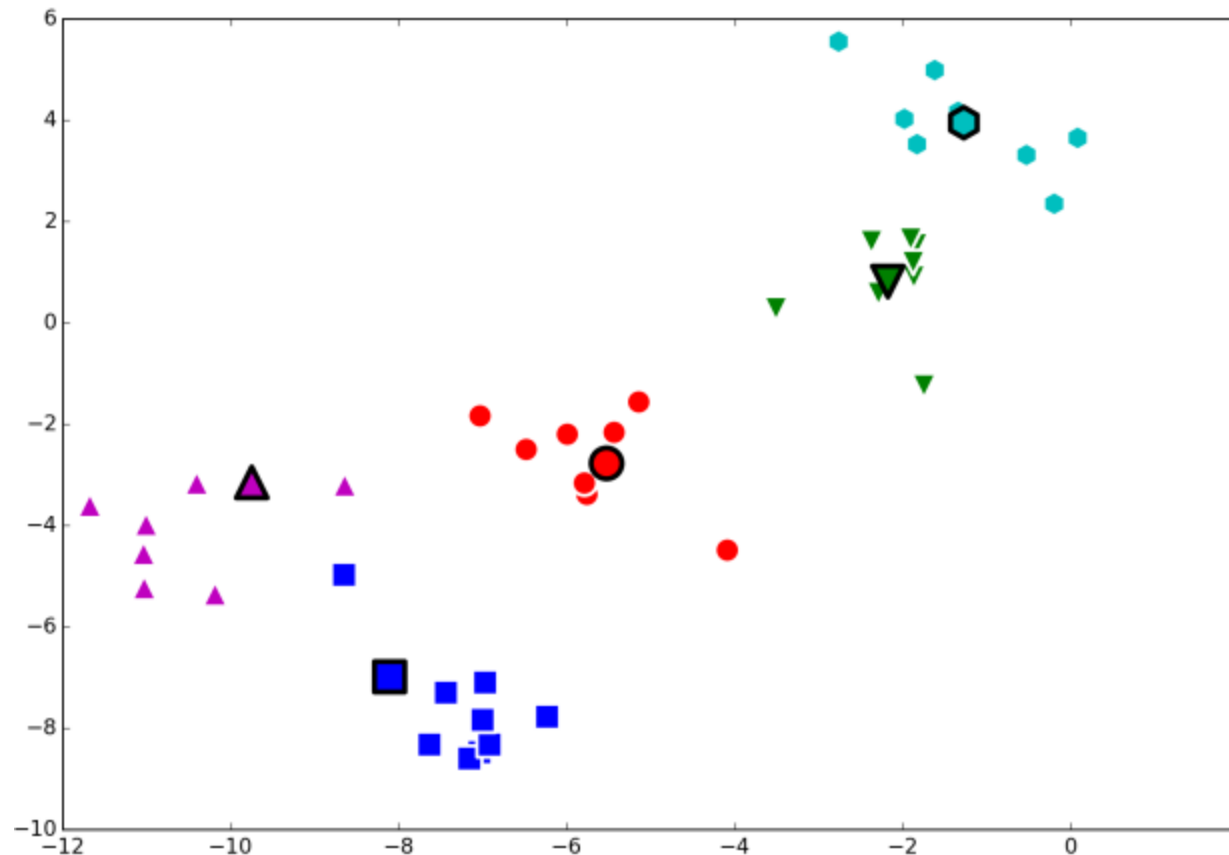
K-Means Clustering: Lloyd Algorithm

- Compute new means, update centroids



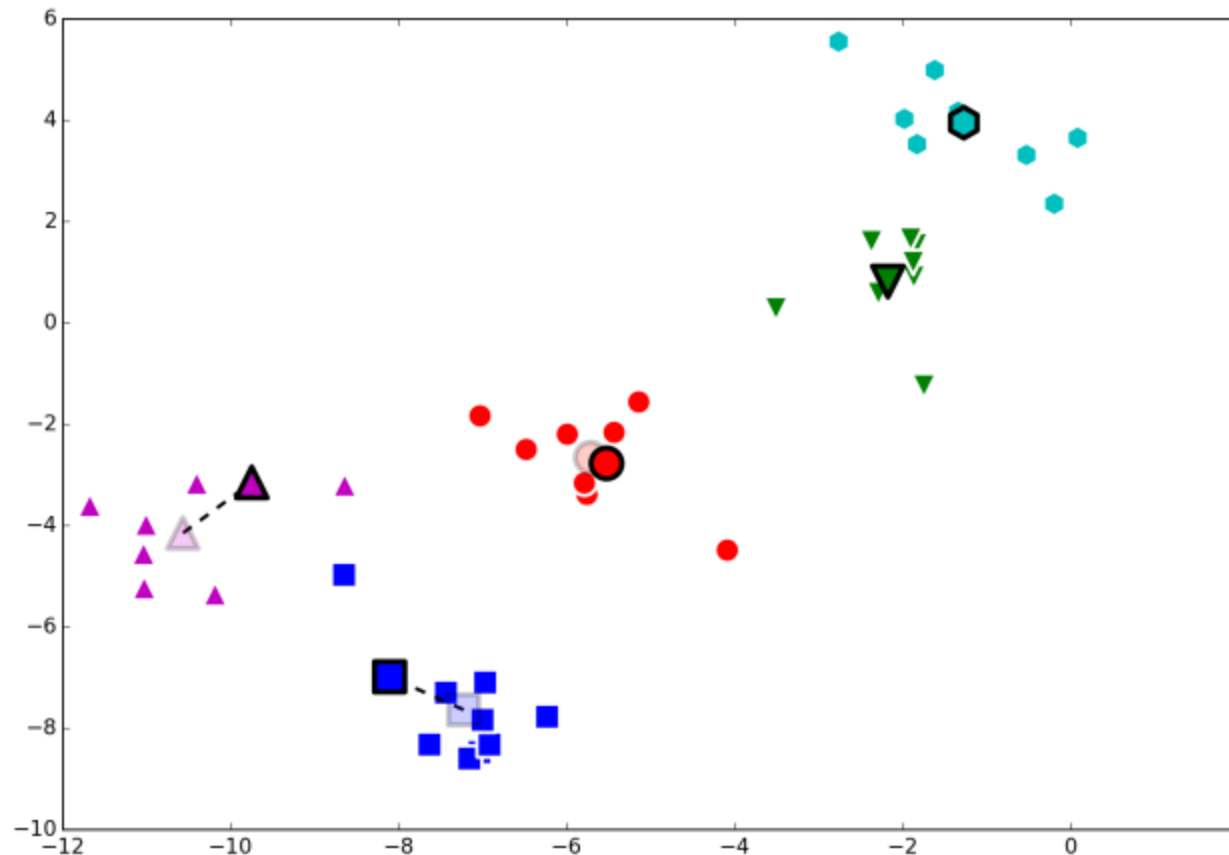
K-Means Clustering: Lloyd Algorithm

- Recompute clusters



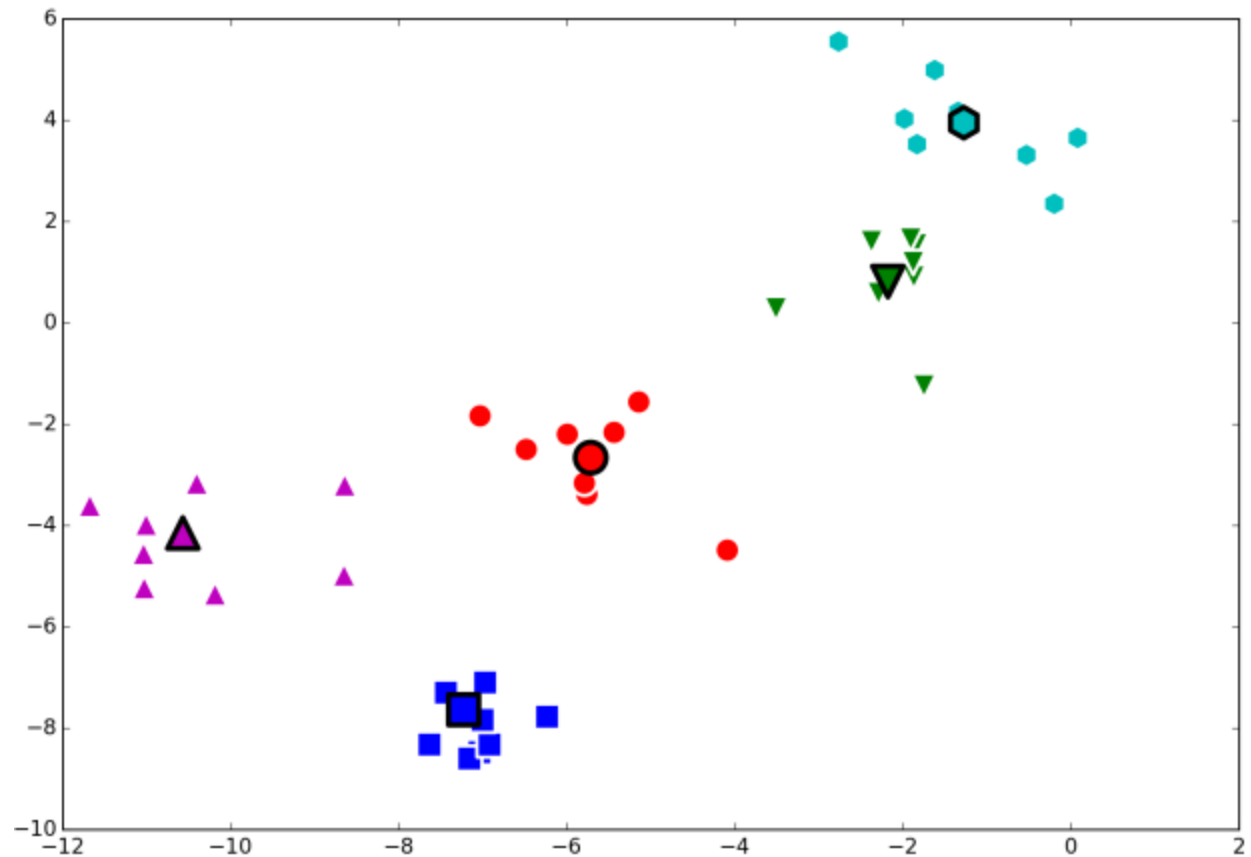
K-Means Clustering: Lloyd Algorithm

- Compute new means, update centroids



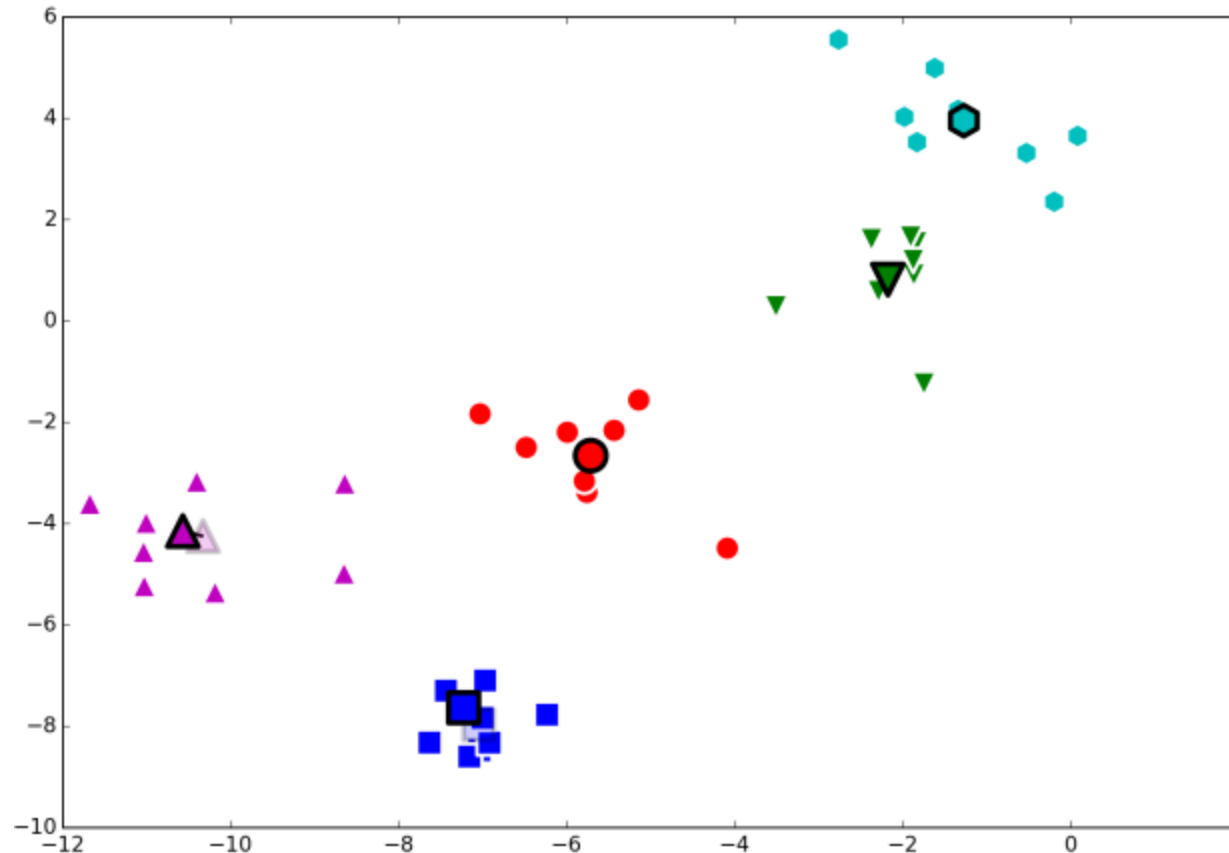
K-Means Clustering: Lloyd Algorithm

- Recompute clusters



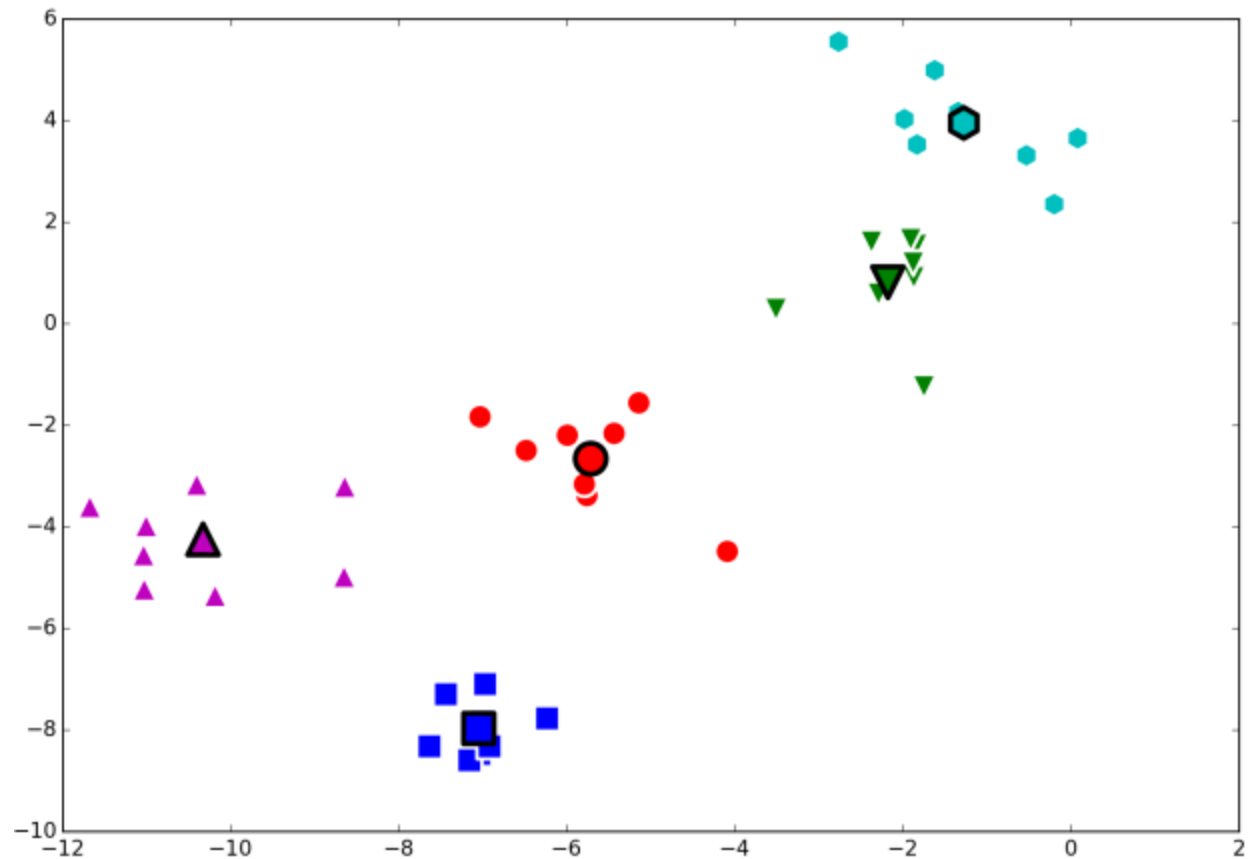
K-Means Clustering: Lloyd Algorithm

- Compute new means, update centroids



K-Means Clustering: Lloyd Algorithm

- Until convergence



K-Means Clustering: Strengths and Limitations



- **Strengths:**

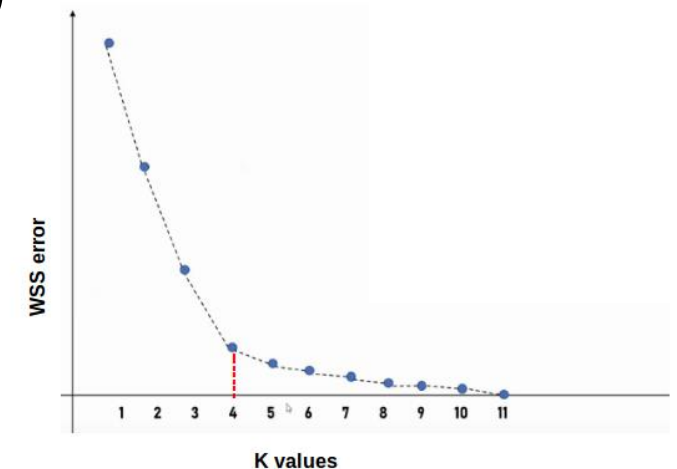
- Simple and works well for disjoint clusters;
- Relatively efficient and scalable (Lloyd algorithm);

- **Limitations:**

- k needs to be defined in advance;
- Highly dependent on the initialization;
- Unable to handle well noisy data and outliers;
- Not suitable for clusters of different sizes and non-convex shapes.

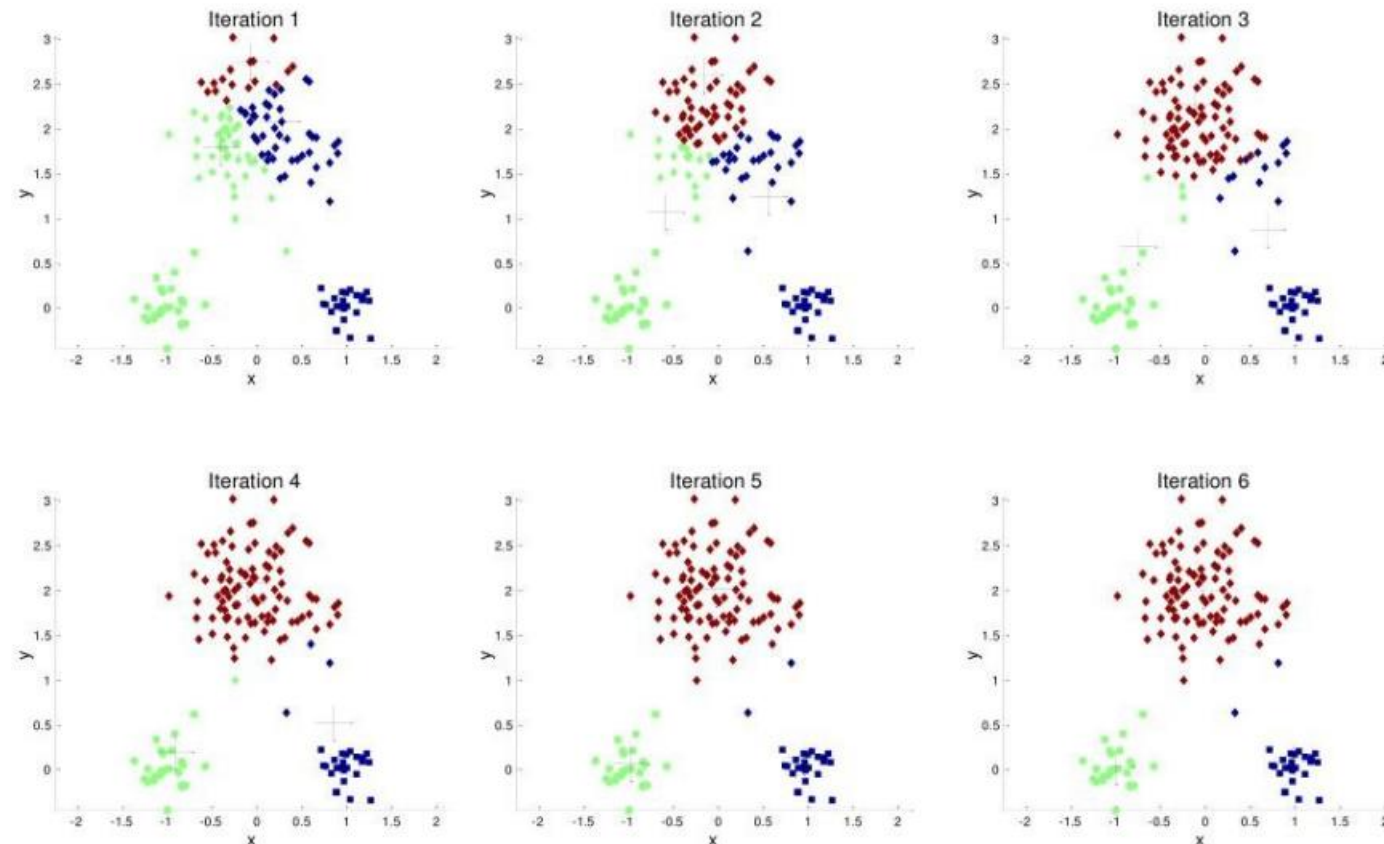
K-Means Clustering: Choosing k

- Choosing the value of k is not a trivial task that heavily affects the outcome of the algorithm.
- **Elbow method:** determine the optimal k
 - Iteratively applies K-means clustering with an increasing number of clusters.
 - Calculates the within-cluster sum of squares (WCSS) for each iteration.
 - WCSS measures the compactness of clusters; lower WCSS indicates better clustering.



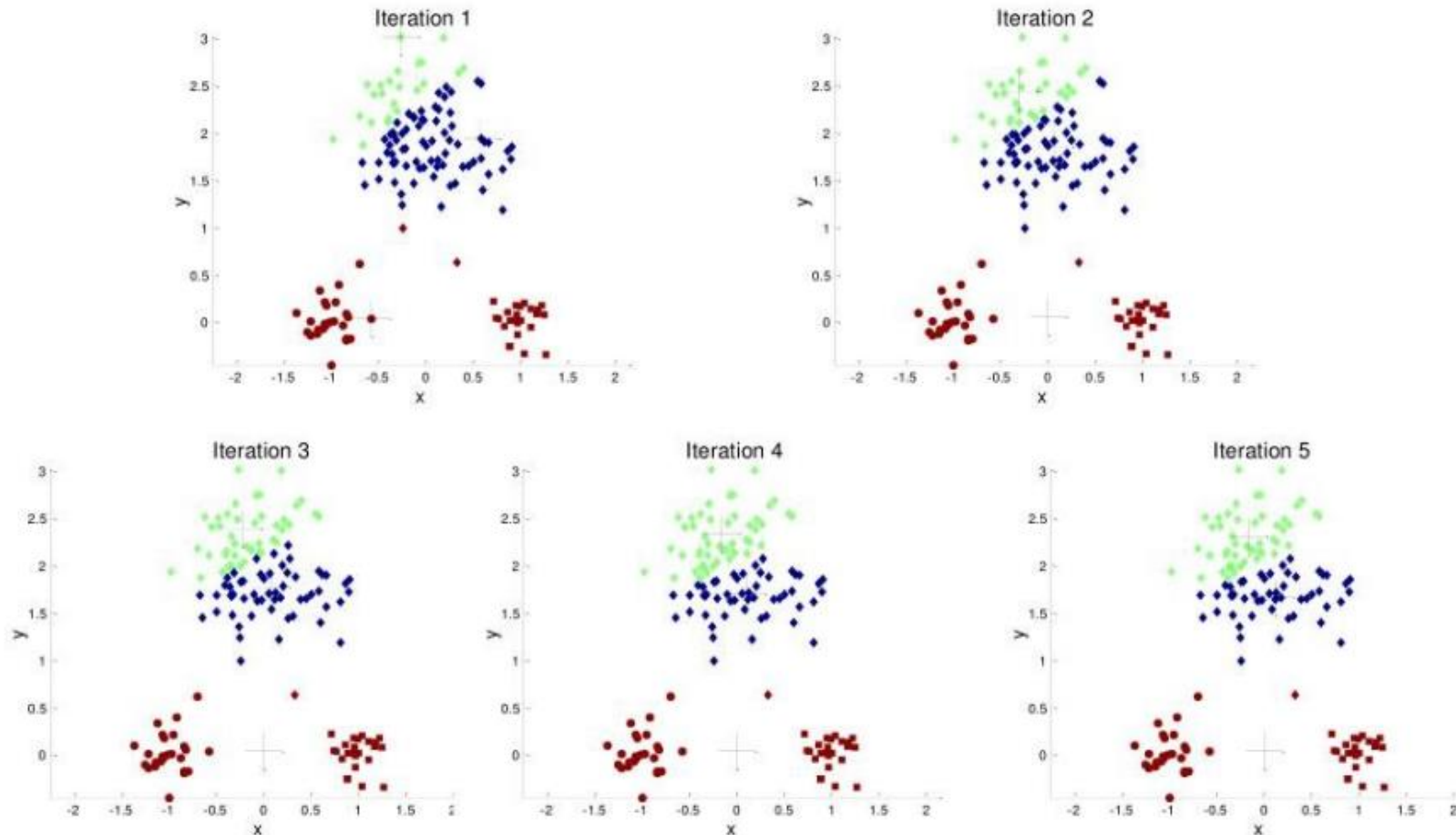
K-Means Clustering: Initialization

- Importance of initialization: case 1

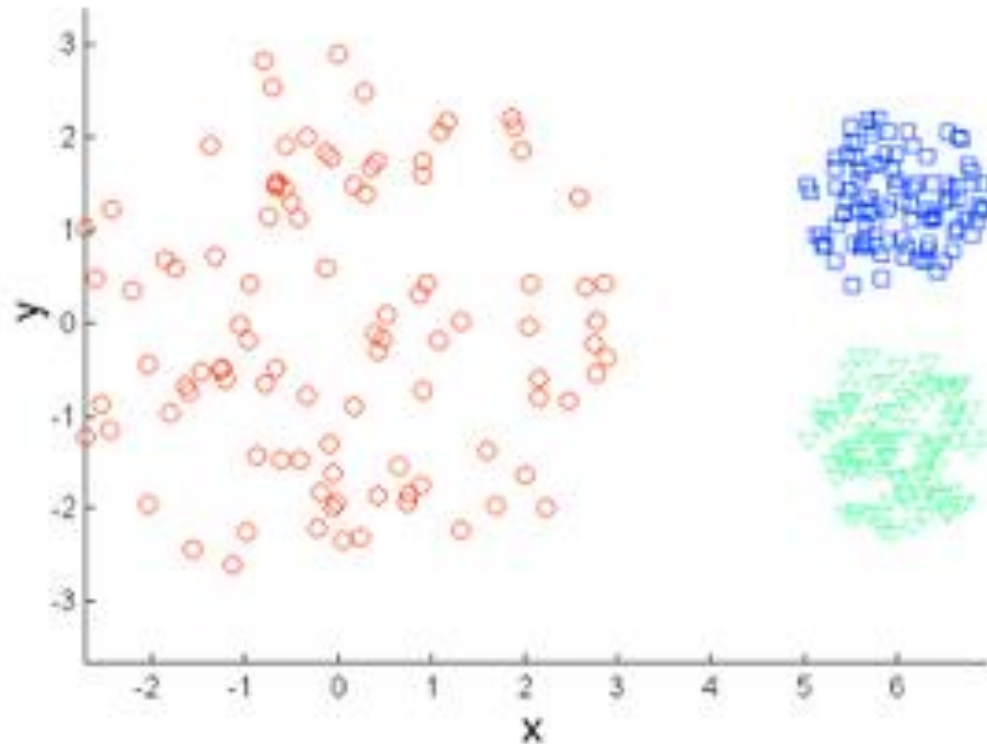


K-Means Clustering: Initialization

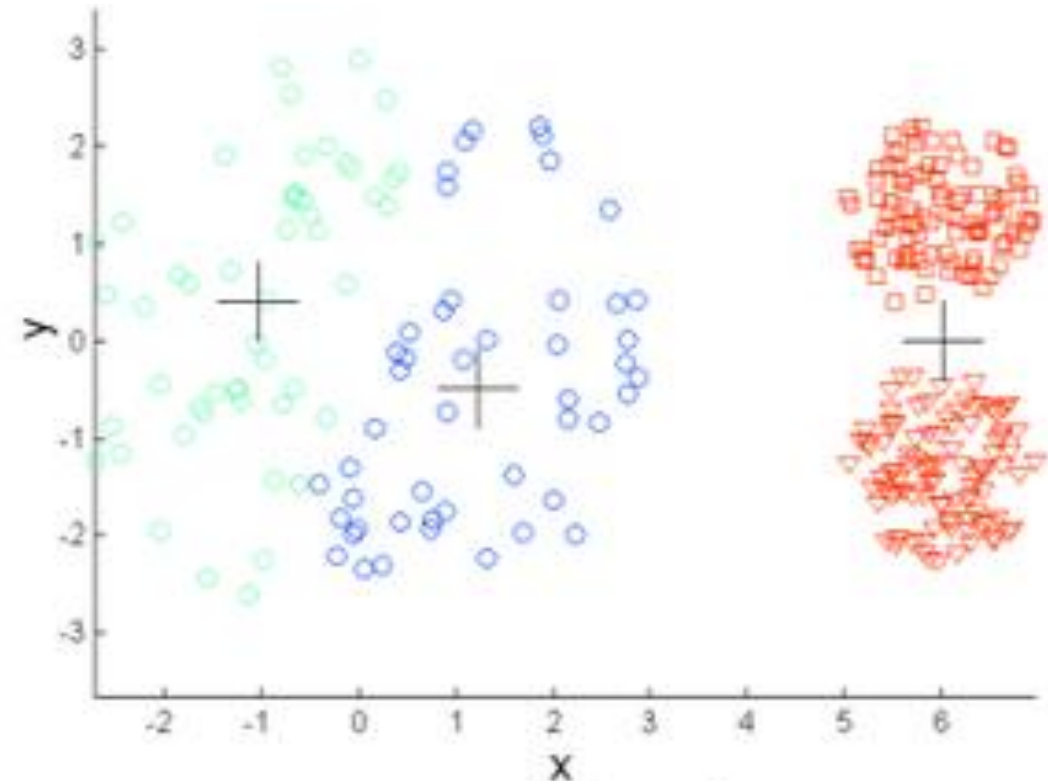
- Importance of initialization: case 2



K-Means Clustering: Different Sizes

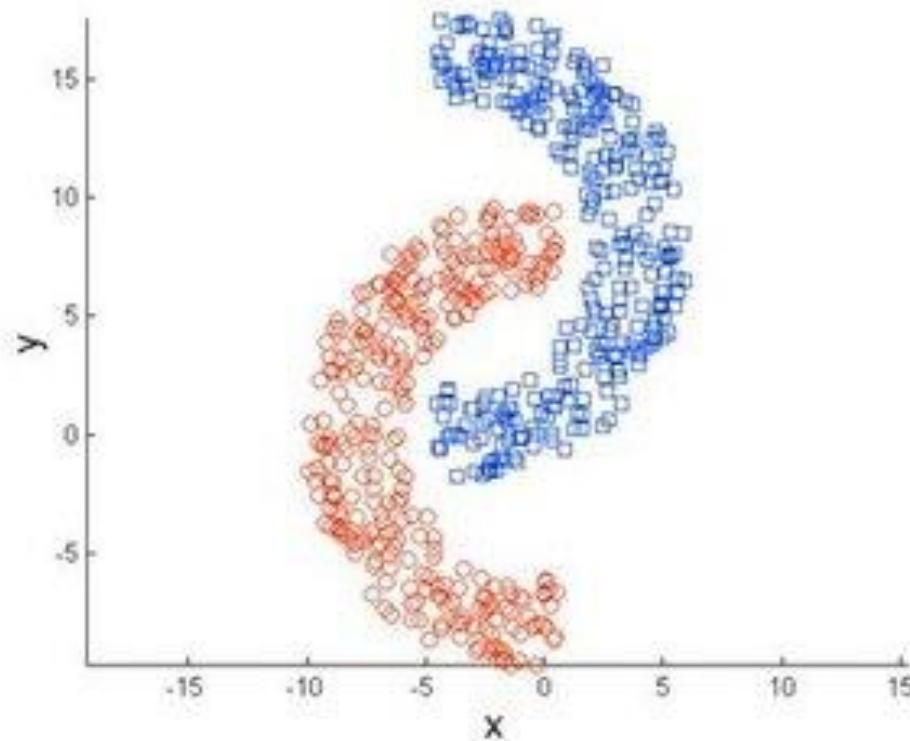


Original Points

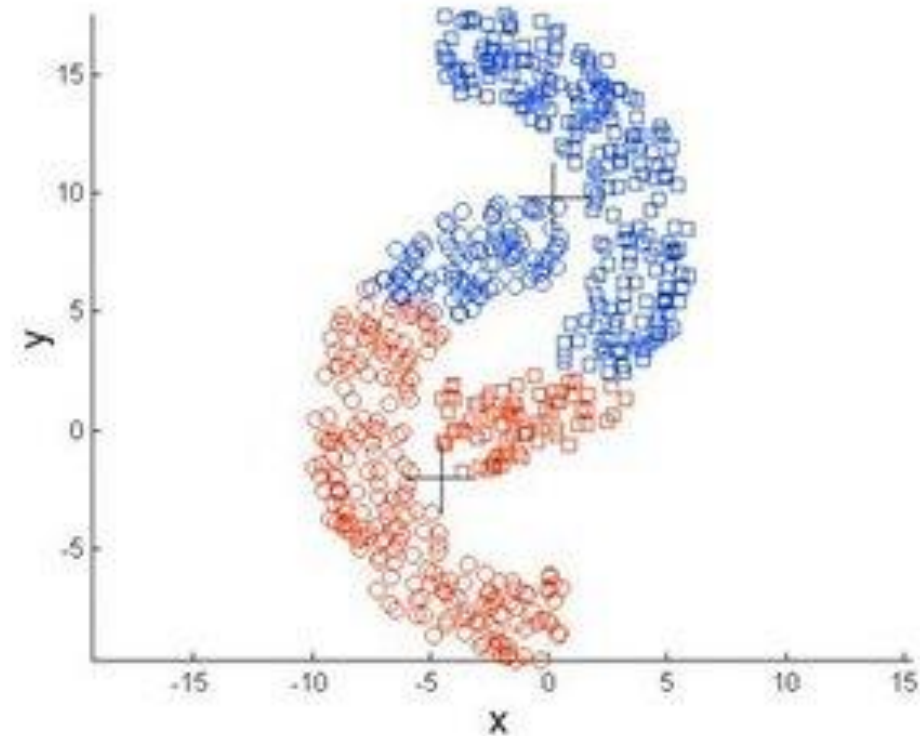


K-means ($k = 3$)

K-Means Clustering: Non-Convex Shapes



Original Points



K-means (2 Clusters)



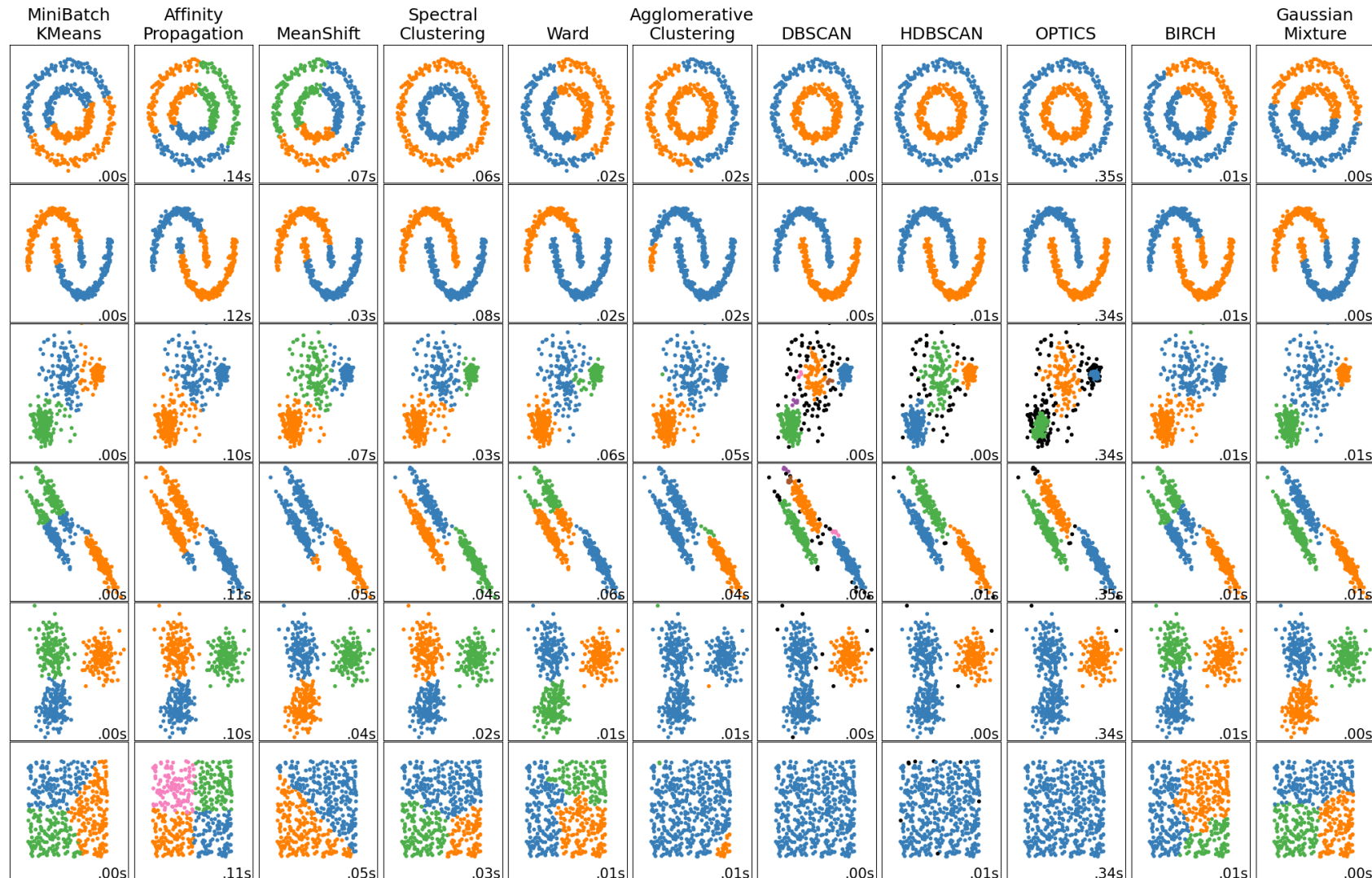
K-Means Variants

- **K-Medoids**: prototypes are data points (medoids);
- **K-Modes**: for categorical data. Utilizes mode-based distance measures (e.g., Hamming distance)

K-Means - Applications

- **Customer Segmentation:** Identifying groups of customers with similar traits for targeted marketing and personalized services.
- **Image Compression:** Simplifying images by reducing the number of colors, preserving visual quality while saving storage space.
- **Anomaly Detection:** Spotting outliers in datasets, useful for fraud detection and quality control.
- **Document Clustering:** Organizing documents by content similarity for efficient retrieval and topic modeling.
- **Retail Inventory Management:** Optimizing inventory levels and product placement based on sales patterns.
- **Healthcare Data Mining:** Grouping patients with similar medical profiles for diagnosis and treatment planning.
- **Climate Pattern Analysis:** Identifying weather patterns and trends in climate data for forecasting and resource management.

Other Clustering Algorithms



<https://scikit-learn.org/stable/modules/clustering.html>

Resources

- Berry, M. W., Mohamed, A., & Yap, B. W. (2019). Supervised and unsupervised learning for Data Science. Cham, Switzerland: Springer Nature.
- Patel, A. A. (2019). Hands-on unsupervised learning using python. Sebastopol, CA: O'Reilly Media.