K-means

Yasaman Amannejad,
Department of Mathematics and Computing,
Mount Royal University

Outline

- Unsupervised learning
- Clustering
- K-means
  - Algorithm
  - Similarity Measures
  - Elbow Method
  - Cluster Evaluation

Activities

- In this session we will:
  - Using K-means algorithm from Scikit-learn.
  - Implementing a K-means algorithm.

ML vs. Programming

> Field of study that gives computers the ability to learn **without being explicitly programmed**.
> - Arthur Samuel, 1959

Traditional Programming

Program

If (eyes==2) & (legs == 4) & (tail == 1) ... then print "Cat"

Output

Computer

"Cat"

Output

Machine Learning

Output

"Cat"

Data

Computer

Program

Cat Recognition

Supervised vs. Unsupervised

Machine Learning

Supervised	Reinforcement Learning	Unsupervised

Supervised

Output

"Cat"

Data

Computer

Program

Cat Recognition!

Unsupervised

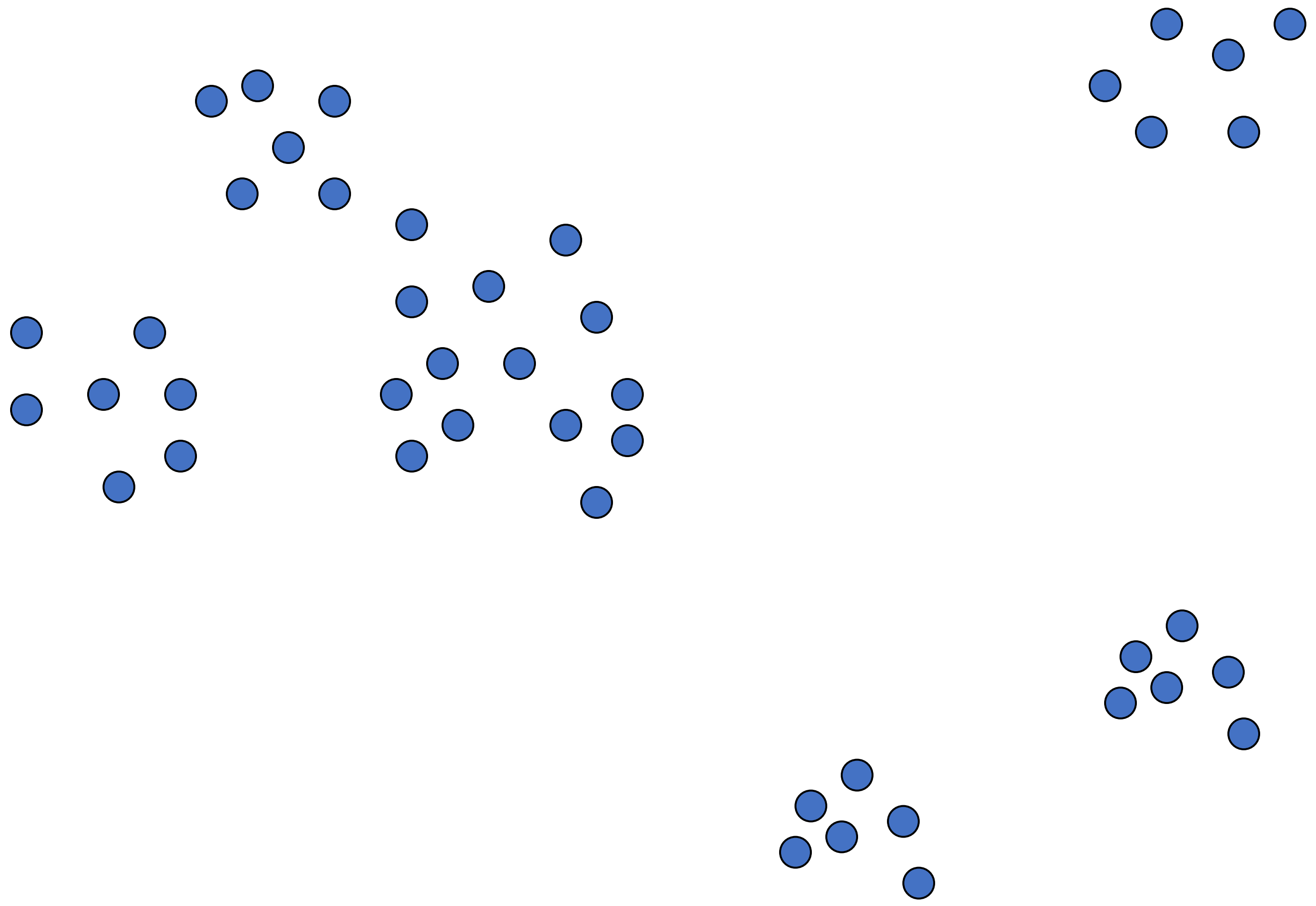
Data

Computer

Program

Grouping Cats!

Clustering



What is Clustering?

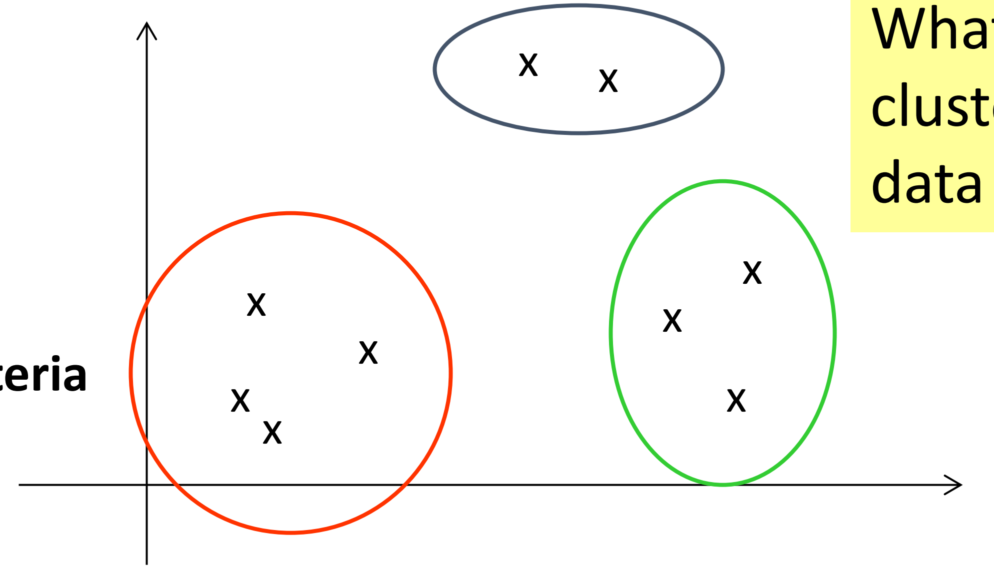
- Find K clusters so that the objects of one cluster are **similar** to each other whereas objects of different clusters are **dissimilar**.

- Identify such groupings (or clusters) in an ***unsupervised*** manner.

$$\sum_{i=0}^{n} \min_{\mu_j \in C}(||x_i - \mu_j||^2)$$

inertia

Within-cluster sum-of-square criteria

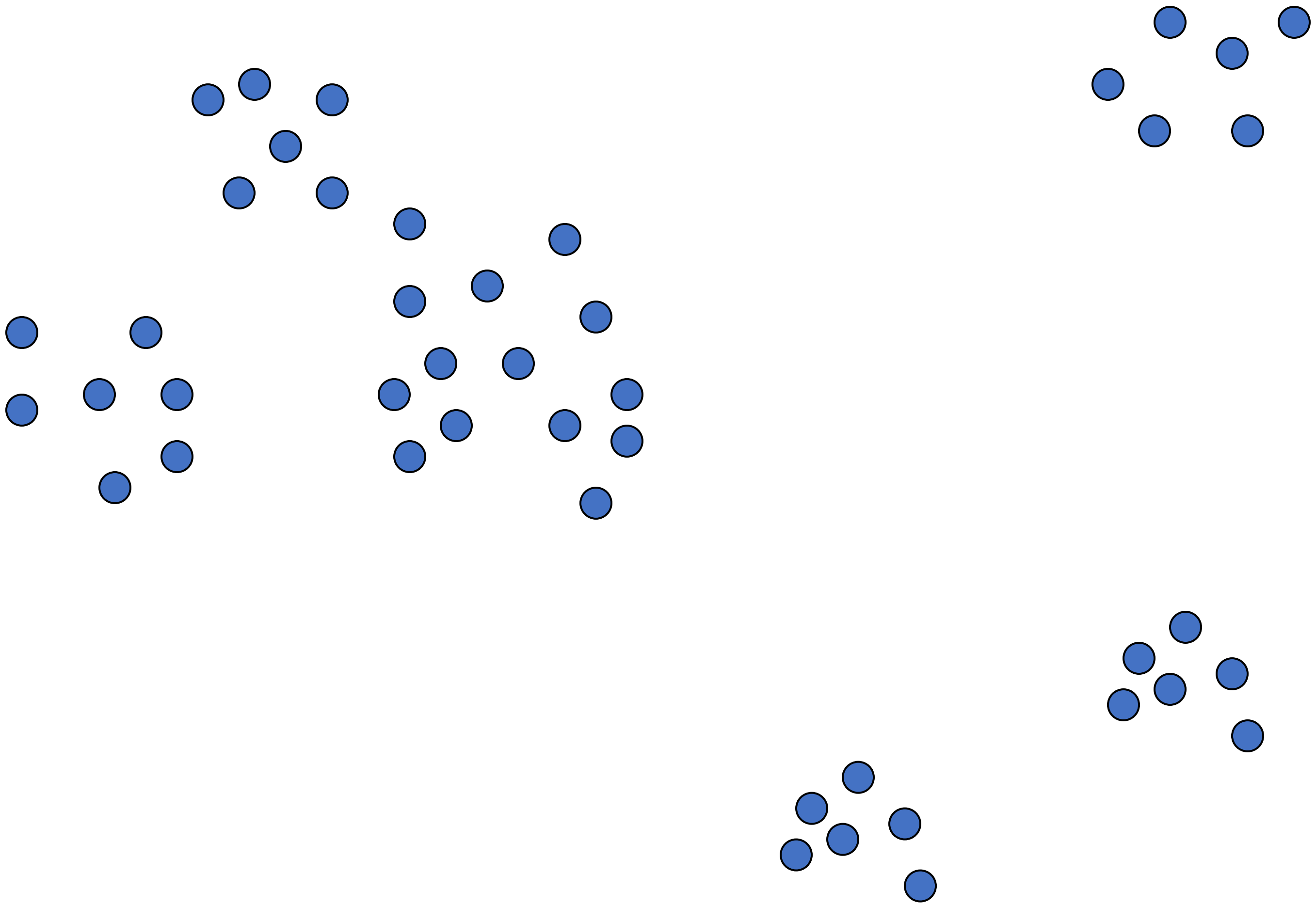


What should the clusters be for these data points?

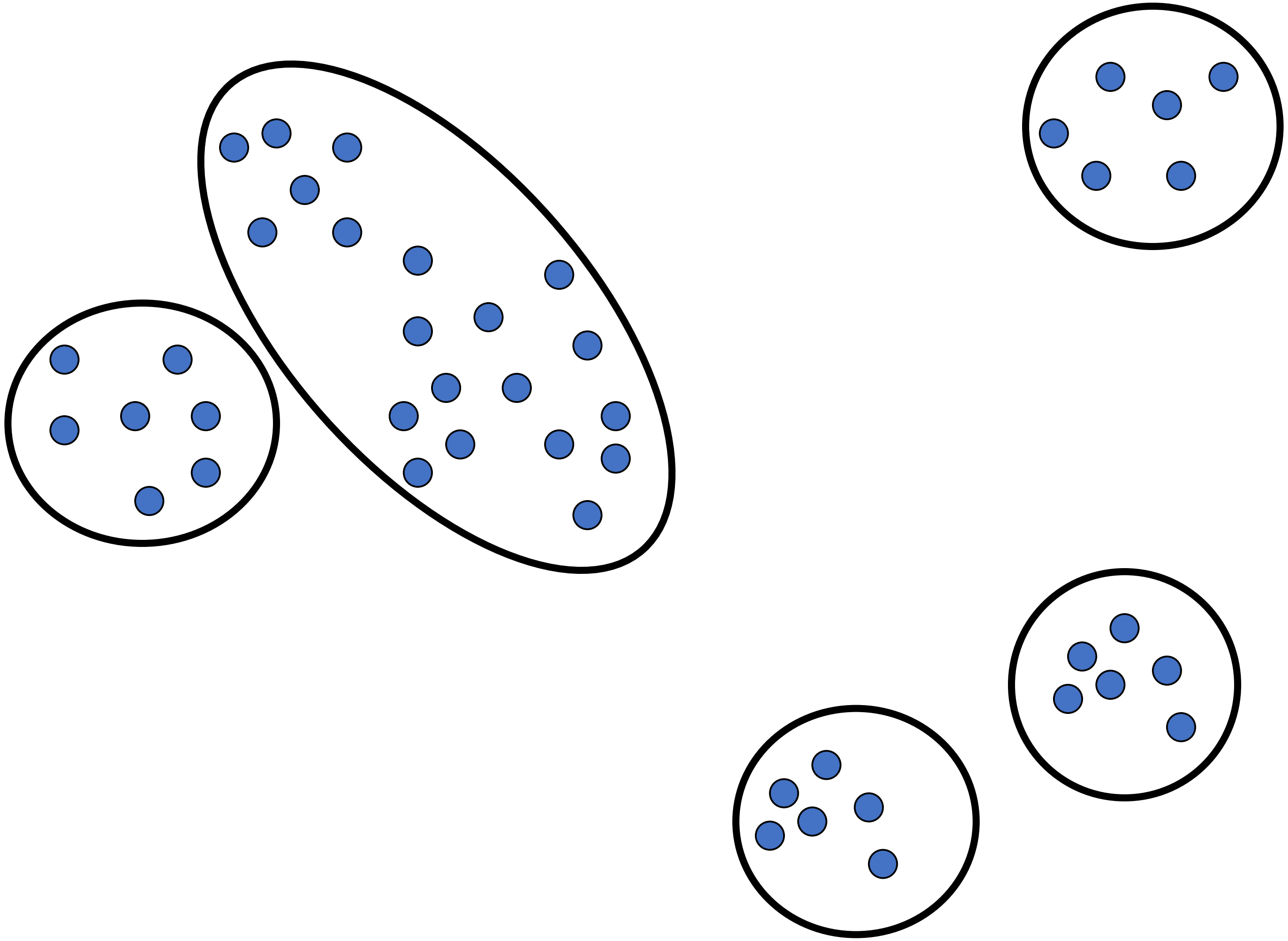
Yasaman Amannejad, Mount Royal University

Email: yamannejad@mtroyal.ca

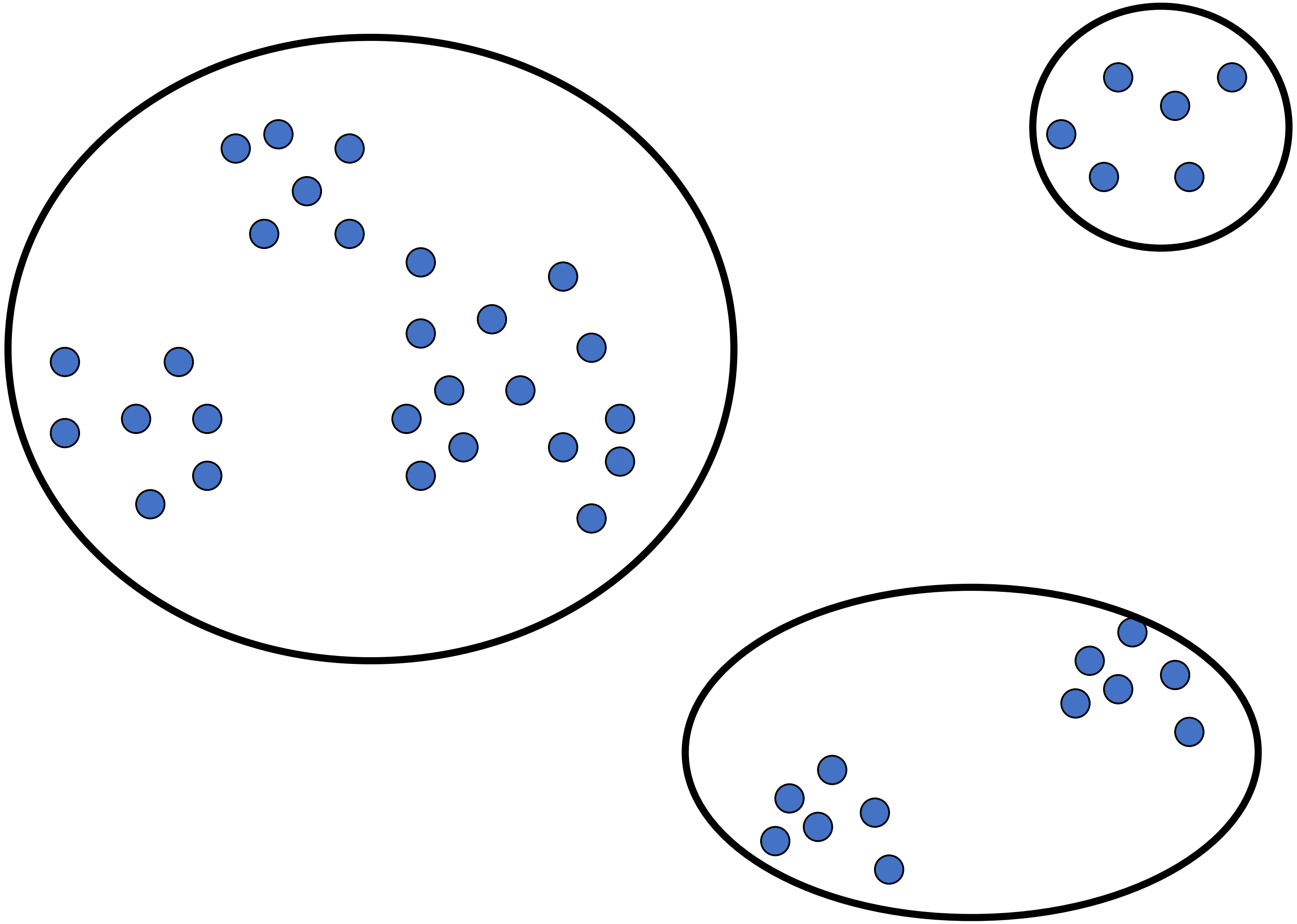
Clustering



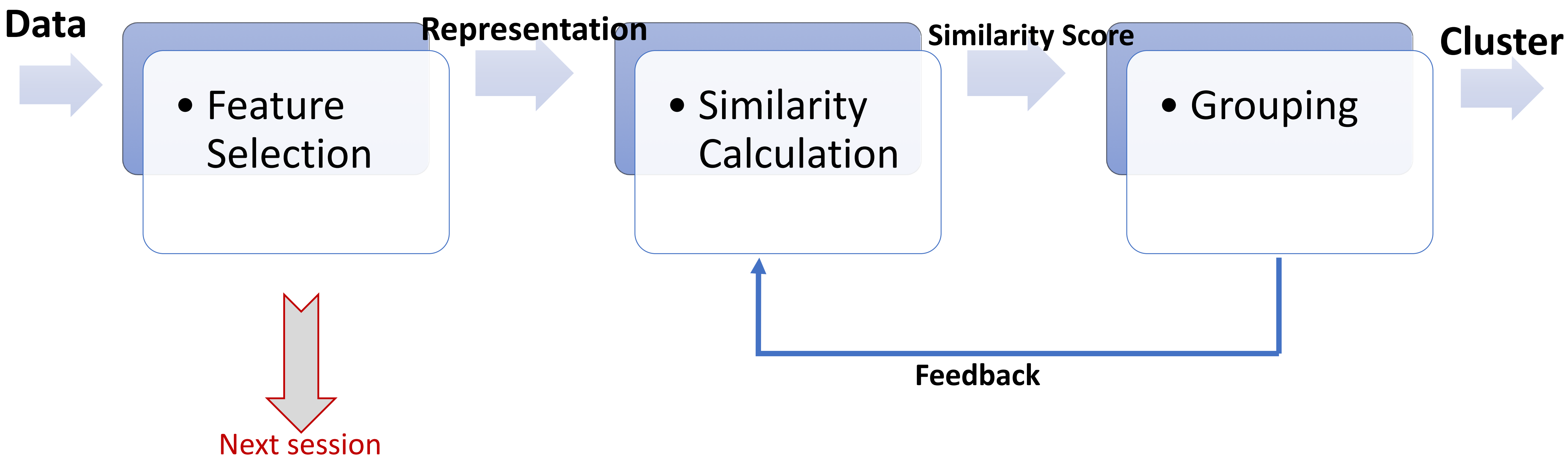
Clustering



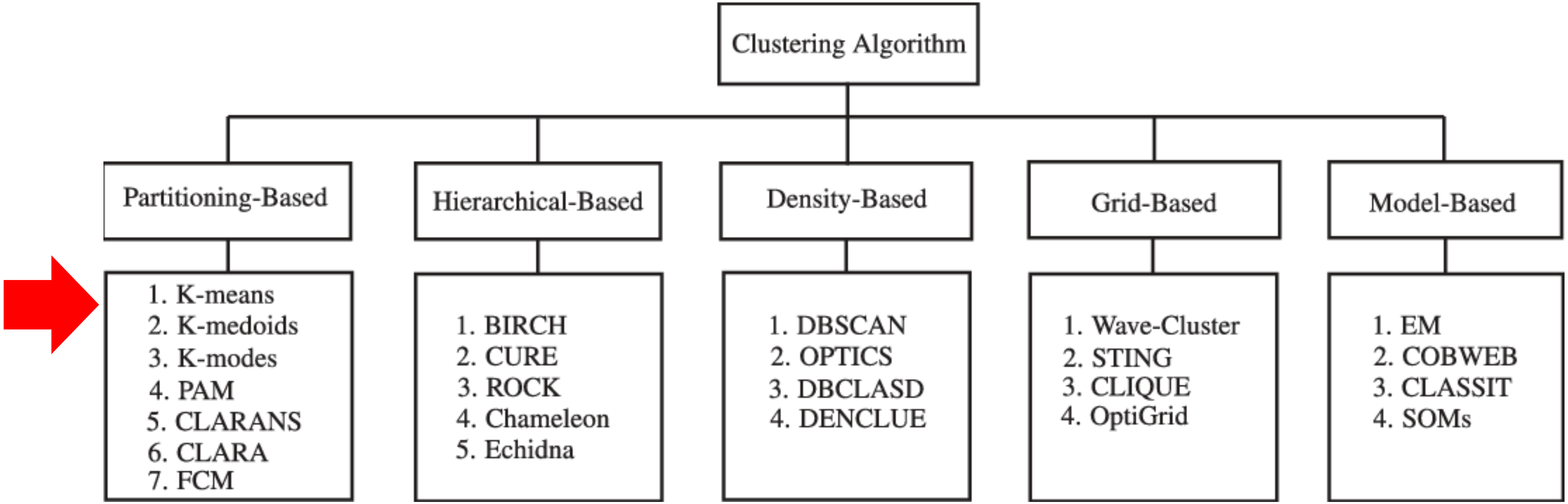
Clustering



Stages in clustering



Clustering Algorithms



K-Means

- **Step 1**: Start with a random points as cluster centers
- **Step 2**: Assigning each data to its closest cluster center
- **Step 3**: Compute new cluster centers as the centroids of the clusters.
- **Step 4**: Steps 2 and 3 are repeated until there is no change in the membership (also cluster centers remain the same)

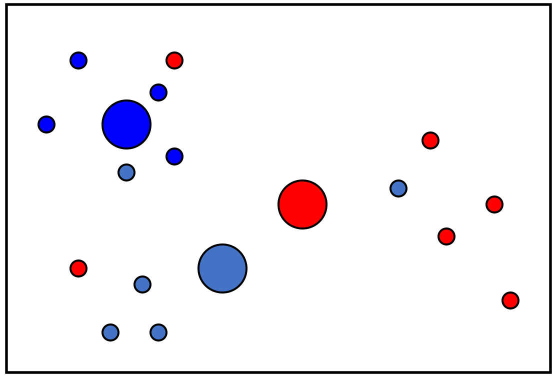
K-Means

- **Stopping criteria:**
  - No change in the members of all clusters
  - when the **squared error** is less than some small threshold value α
    - Squared error *se*

$$se = \sum_{i=1}^{k} \sum_{p \in c_i} \left\| p - m_i \right\|^2$$

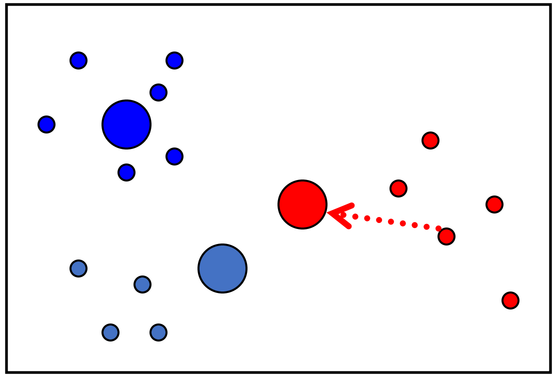
      - where m_i is the mean of all instances in cluster c_i
    - $se^{(j)} < \alpha$

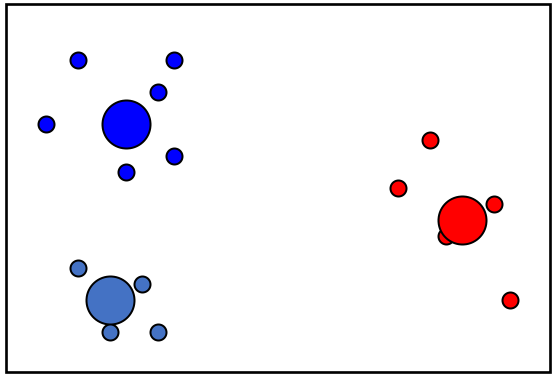
K-means: Example, k = 3



Step 1: Make random assignments and compute centroids (big dots)

Step 2: Assign points to nearest centroids

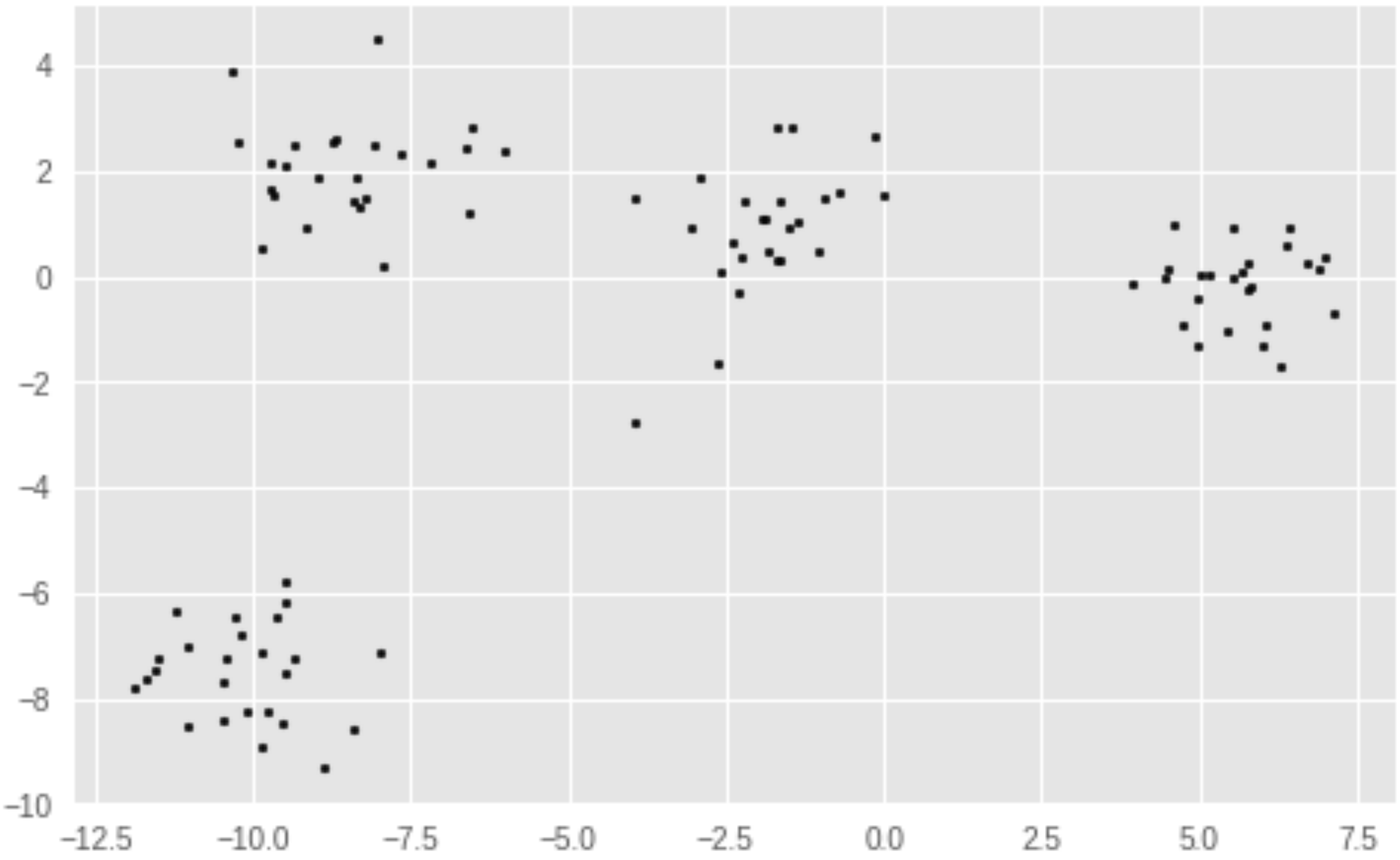




Step 3: Re-compute centroids (in this example, solution is now stable)

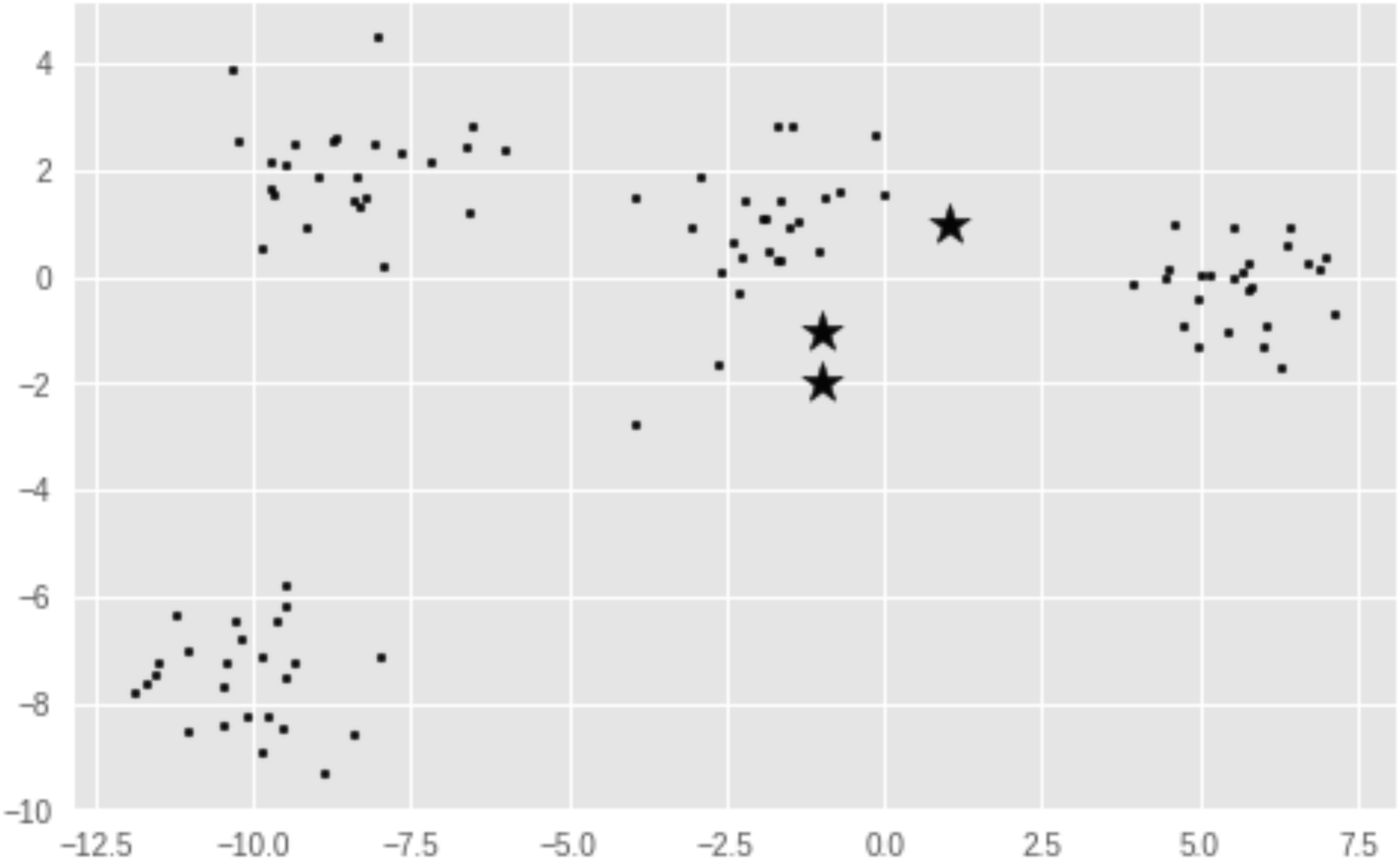
Steps of K-means, k = 3

Round 0



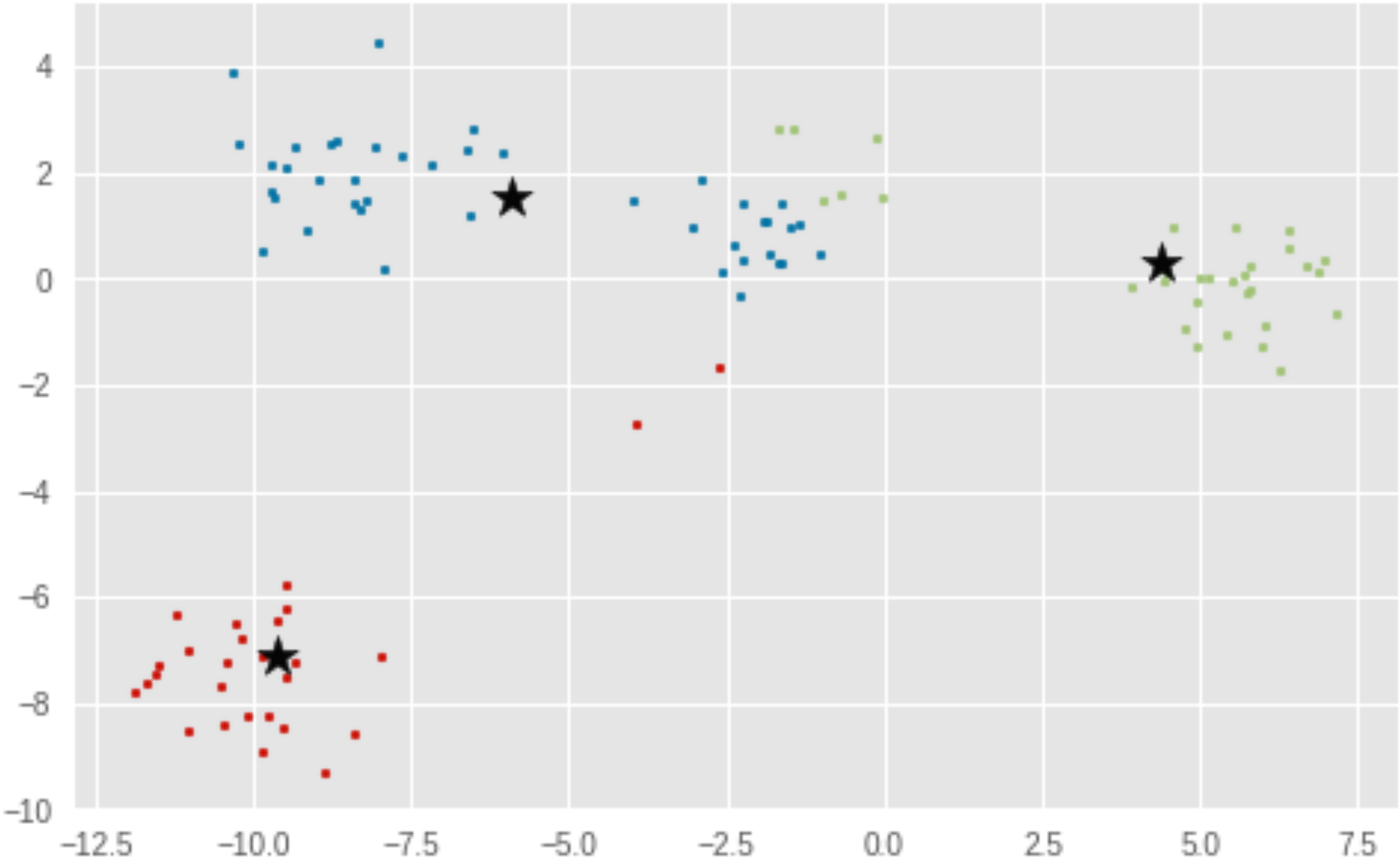
Steps of K-means

Round 1



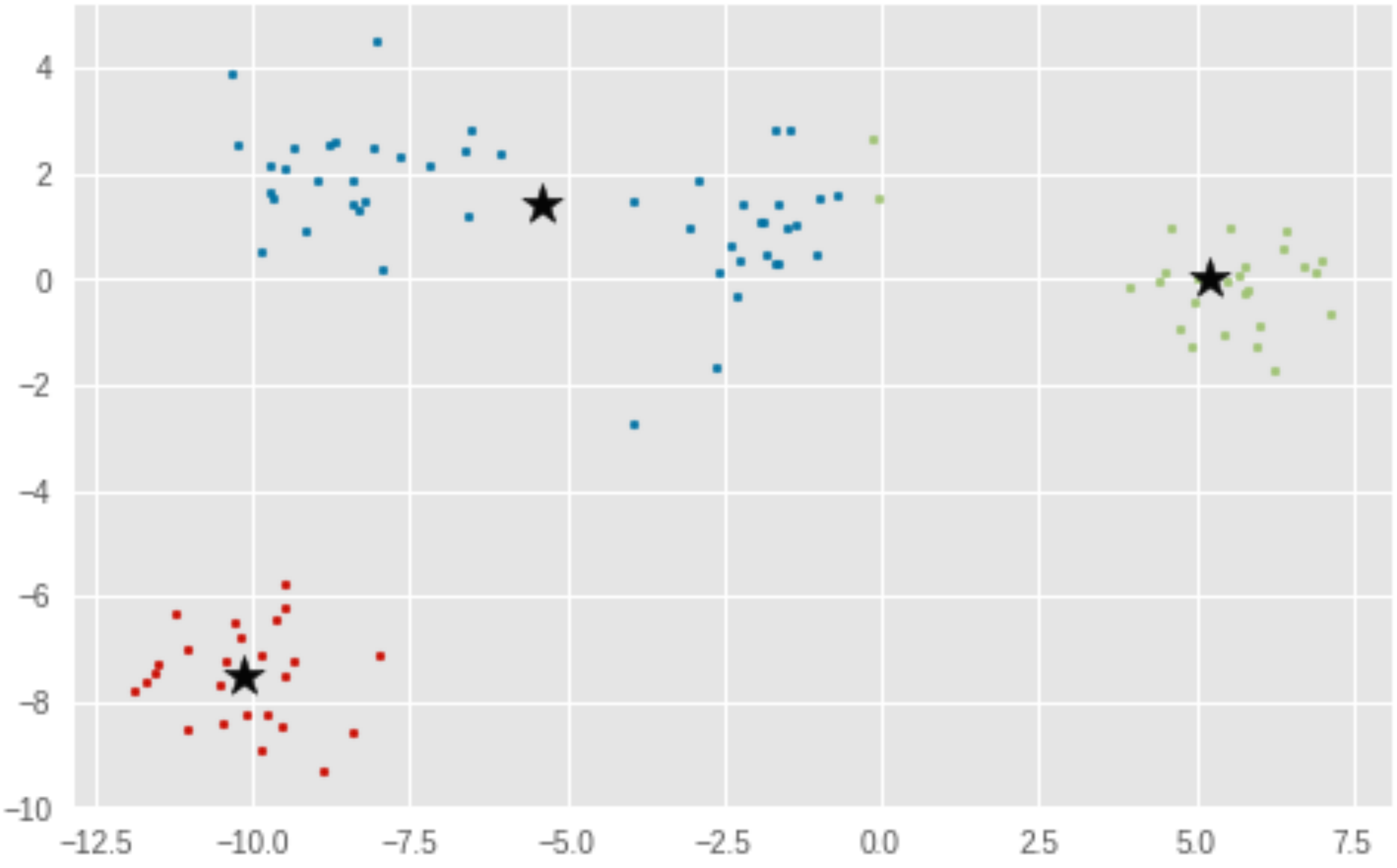
Steps of K-means

Round 2



Steps of K-means

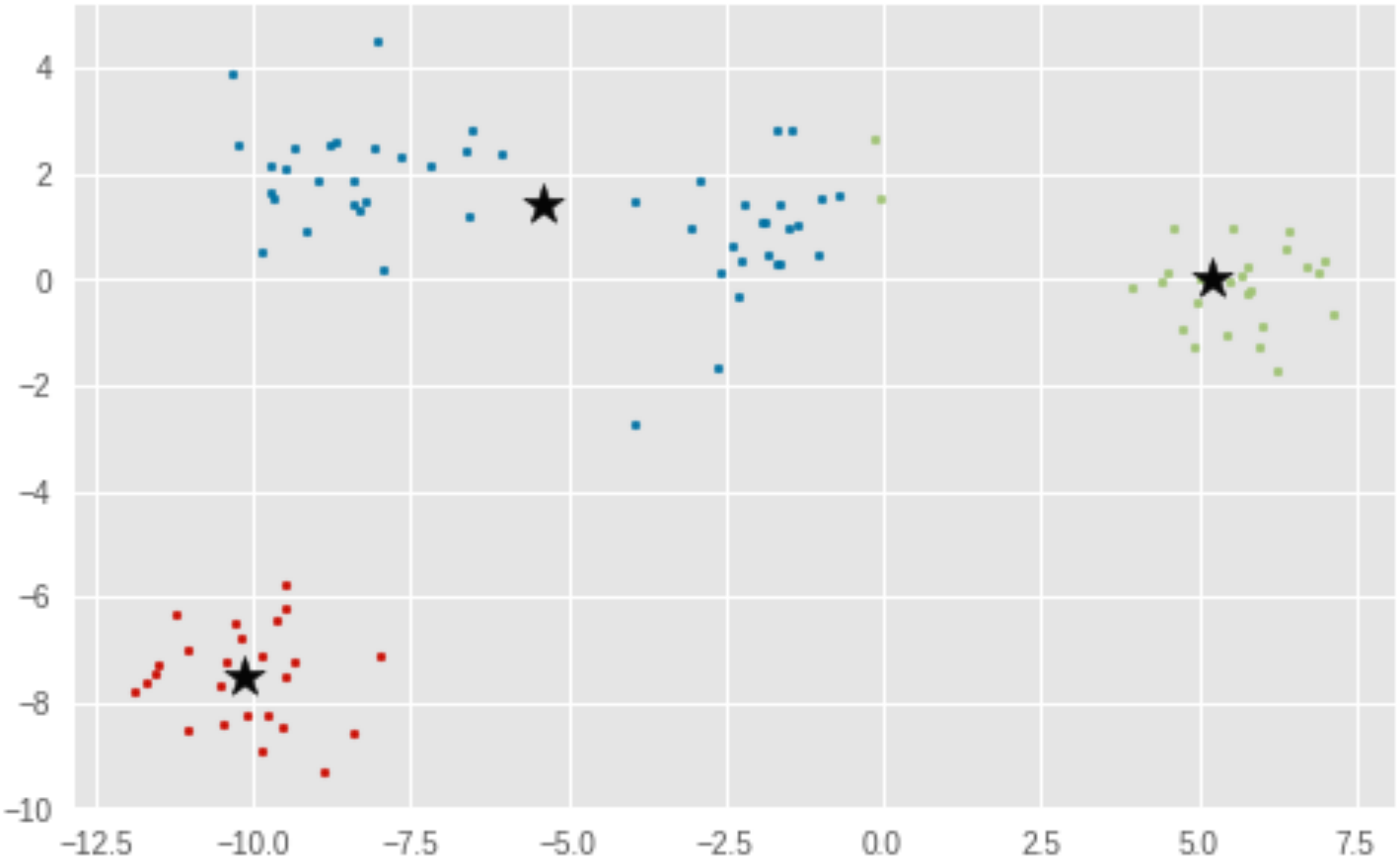
Round 3



Steps of K-means

Round 4

STOP!



(Dis)similarity Measure

How to determine similarity between data points? using various distance metrics!

Let **x** = (x_1,…,x_n) and **y** = (y_1,…y_n) be n-dimensional vectors of data points of objects g_1 and g_2

- Euclidean distance

$$d(g_1, g_2) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Manhattan distance

$$d(g_1, g_2) = \sum_{i=1}^{n} \left|(x_i - y_i)\right|$$

- Minkowski distance

$$d(g_1, g_2) = \sqrt[m]{\sum_{i=1}^{n} (x_i - y_i)^m}$$

(Dis)similarity Measure

- Correlation

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{(Var(X) \cdot Var(Y)}}$$

- maximum value of 1 if X and Y are perfectly correlated
- minimum value of 1 if X and Y are exactly opposite
- $d(X,Y) = 1 - r_{xy}$

- Cov(X,Y) stands for covariance of X and Y
  - degree to which two different variables are related
- Var(X) stands for variance of X
  - measurement of a sample differ from their mean

(Dis)similarity Measure

- Example:
  - Euclidean Distance: number of inserts and deletes to change one stringinto another.

Cluster Evaluation: the Silhouette Score

- Measures of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

$$b(i) = \min_{i \neq j} \frac{1}{|C_j|} \sum_{j \in C_j} d(i, j)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

Cluster Distortion

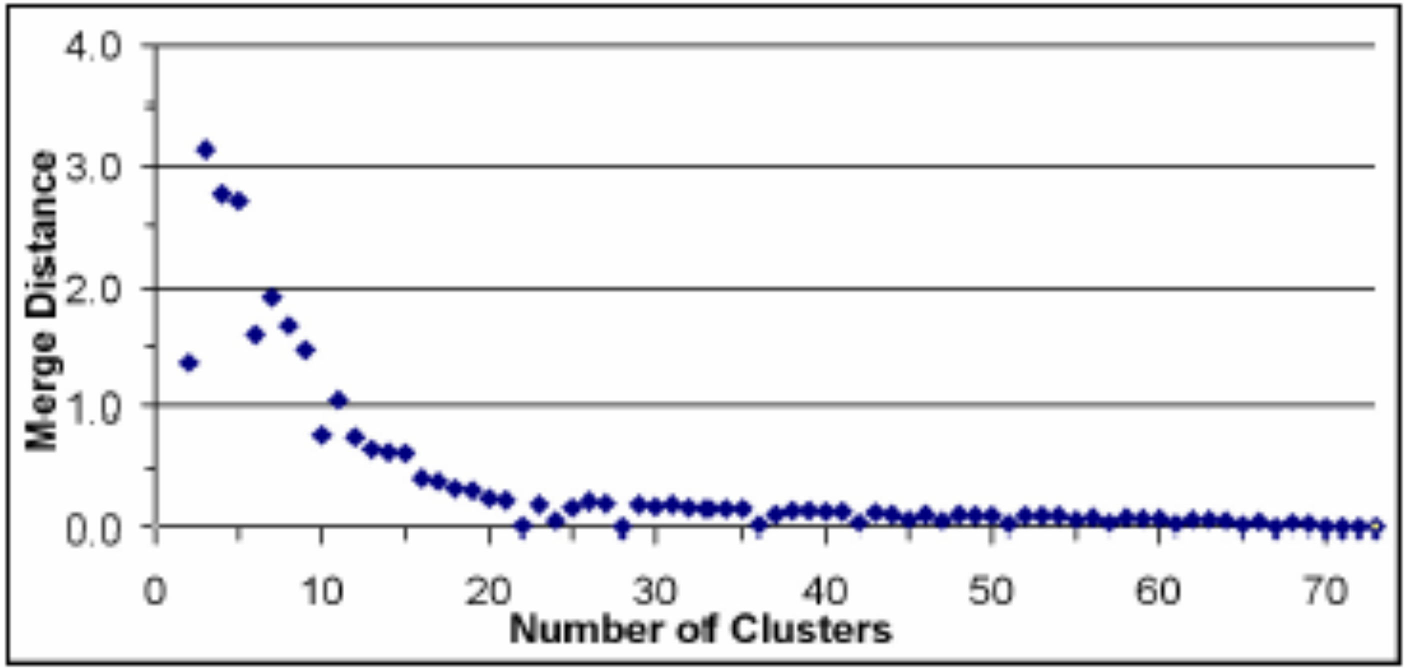
$$J(c,\mu) = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_j^{(i)} - \mu_{c^{(i)},j})^2$$

Sum of squared distances of samples to their closest cluster center.

How Many Clusters?

- Number of clusters *K* is given
  - Partition *n* docs into predetermined number of clusters
- Finding the "right" number of clusters is part of the problem
  - Given data, partition into an "appropriate" number of subsets.
  - E.g., for query results - ideal value of *K* not known up front - though UI may impose limits.
- Can usually take an algorithm for one flavor and convert to the other.

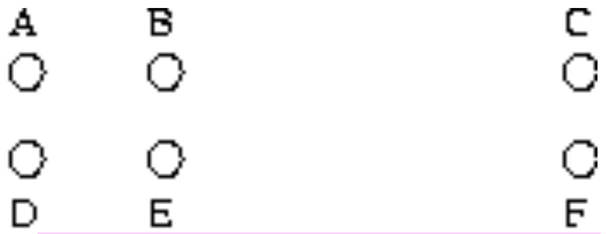
How Many Clusters? Elbow Method



The knee of a curve is defined as the point of maximum curvature.

Seed Choice

Example showing sensitivity to seeds



In the above, if you start with B and E as centroids you converge to {A,B,C} and {D,E,F}
If you start with D and F you converge to {A,B,D,E} {C,F}

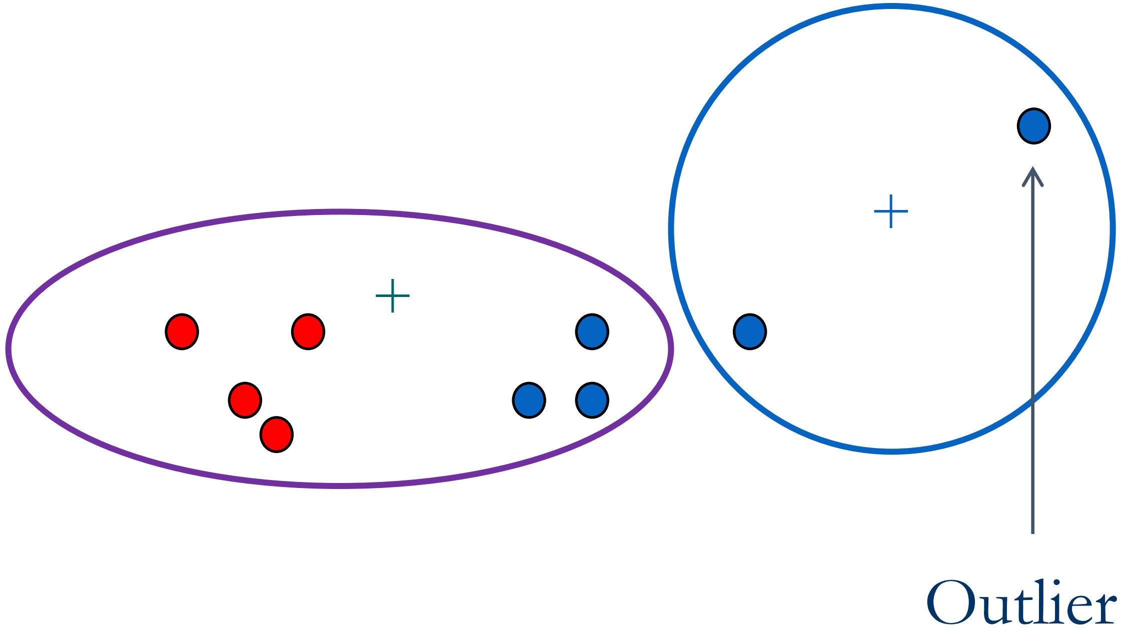
- Results can vary based on random seed selection.
- Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings.
  - Select good seeds using a heuristic (e.g., doc least similar to any existing mean)
  - Try out multiple starting points
  - Initialize with the results of another method.

K-means

- **Pros**
  - Low complexity
    - *complexity is O(nkt), where t = #iterations*
- **Cons**
  - Necessity of specifying k
  - Sensitive to noise and outlier data points
    - Outliers: a small number of such data can substantially influence the mean value)
  - Clusters are sensitive to initial assignment of centroids
    - K-means is not a deterministic algorithm
    - Clusters can be inconsistent from one run to another

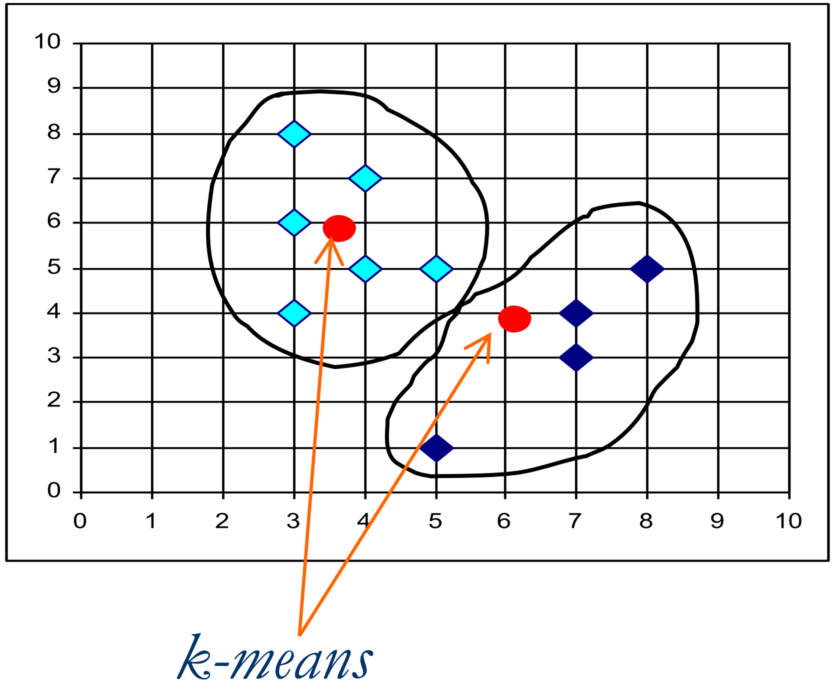
A Problem of K-means

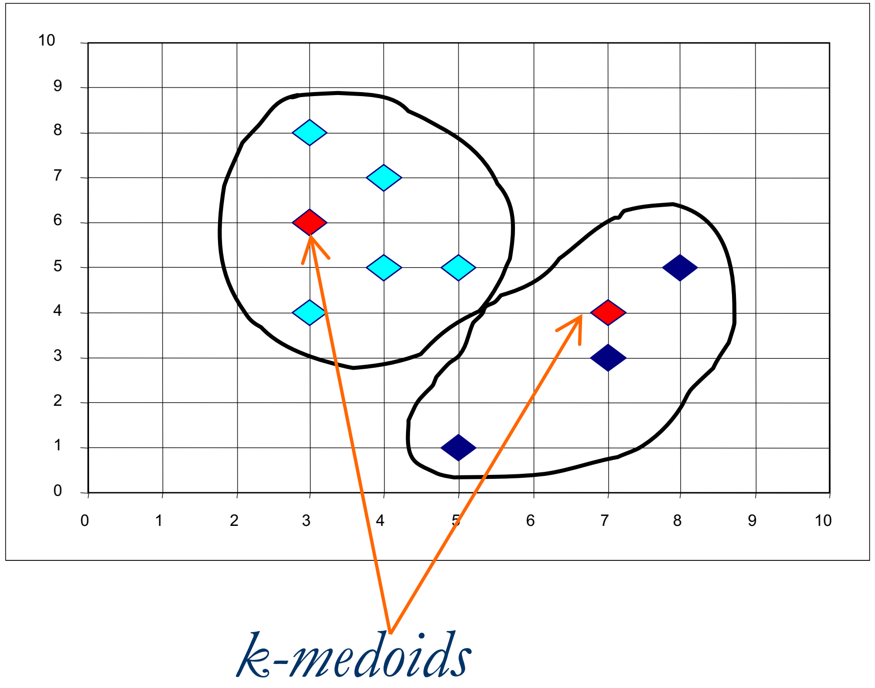
- **Sensitive to outliers**
  - Outlier: objects with extremely large (or small) values
    - May substantially distort the distribution of the data



k-Medoids Clustering Method

- *k*-medoids: Find *k representative objects*, called *medoids*





Yasaman Amannejad, PhD
Assistant Professor, Mount Royal University

Email: yamannejad@mtroyal.ca
Website: mru.ca/amannejad