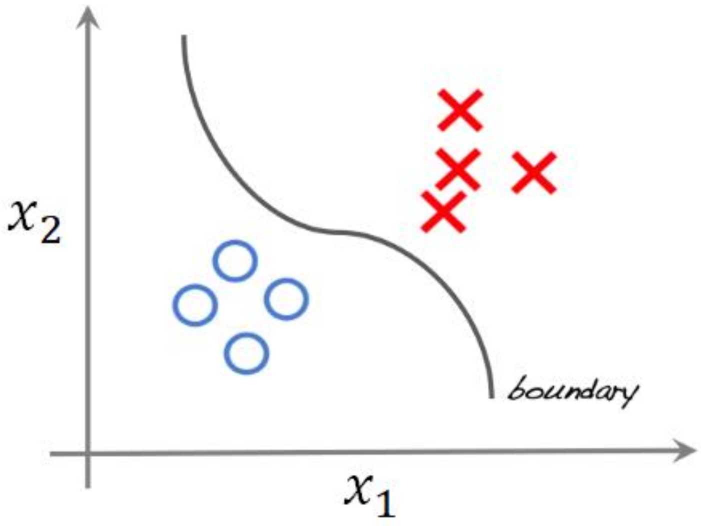
Unsupervised learning
Clustering

Abdessalam Bouchekif
abdessalam.bouchekif@epita.fr

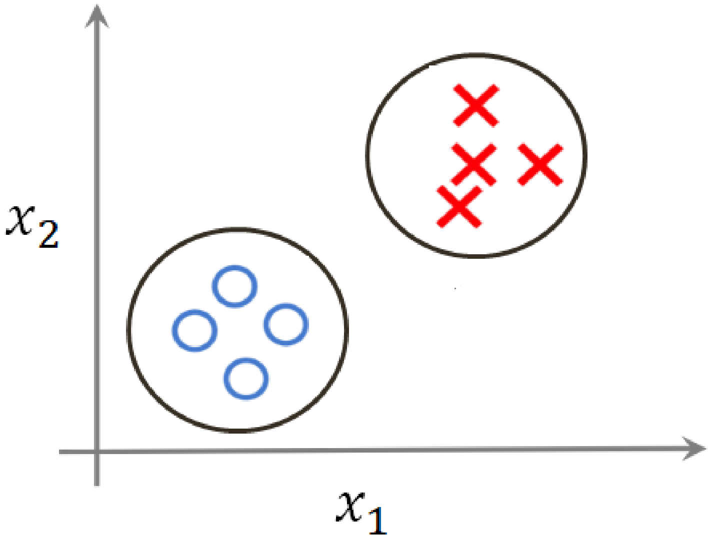
2017/2018

Supervised learning



Training set: $\{(x^{(1)}, x^{(1)}), (x^{(2)}, x^{(2)}), \dots, (x^{(m)}, x^{(m)})\}$

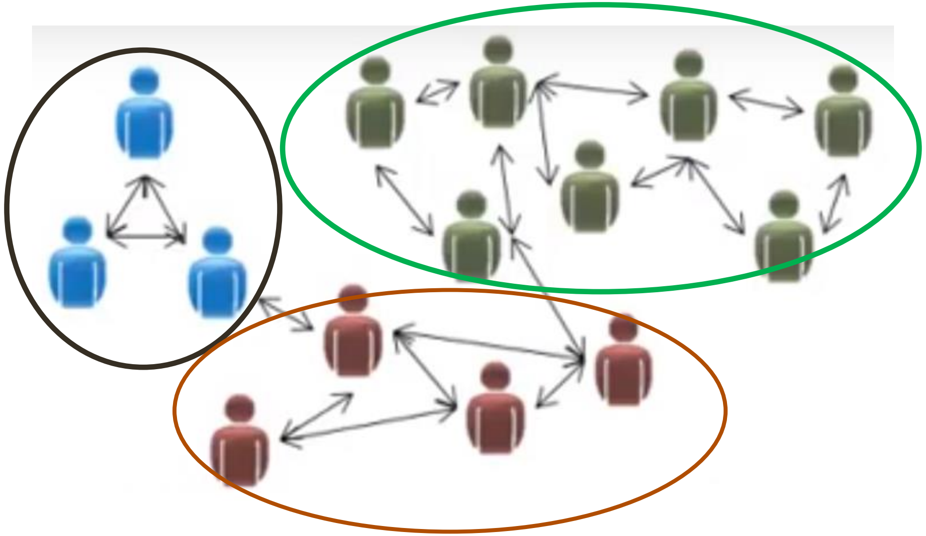
Unsupervised learning



Training set: $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$

K-Means Clustering

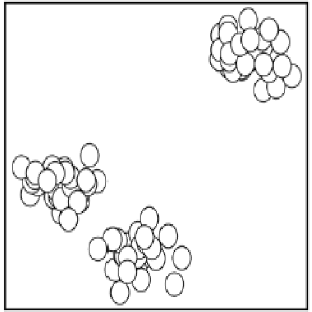
- Organizing data into classes such that there is
  - High intra-class similarity
  - Low extra-class similarity



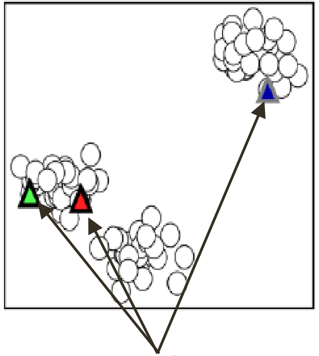
- k is number of class

K-Means

Data input $\boldsymbol{k = 3}$



Initialization

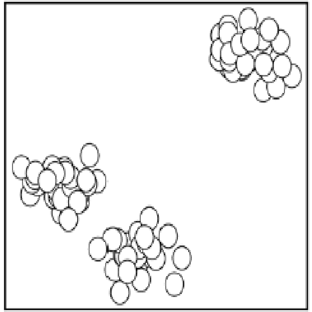


cluster
centroids

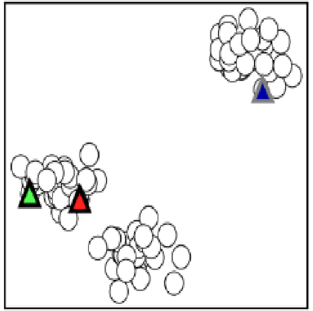
Randomly initialize k cluster centroids

K-Means

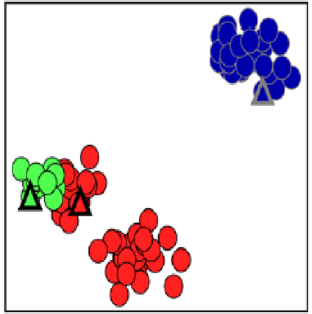
Data input



Initialization



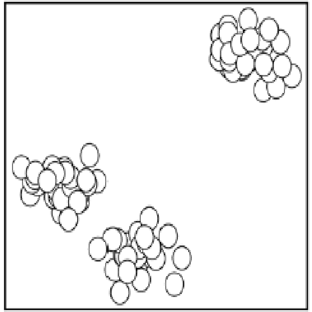
Assign points



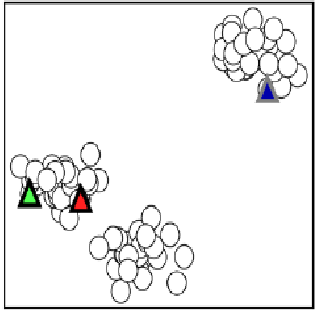
Each data point is assigned to
the cluster center it is closest to

K-Means

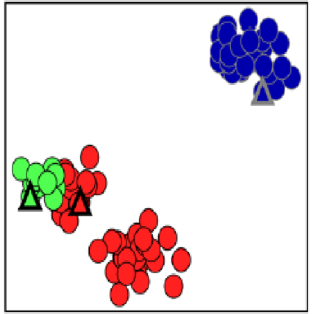
Data input



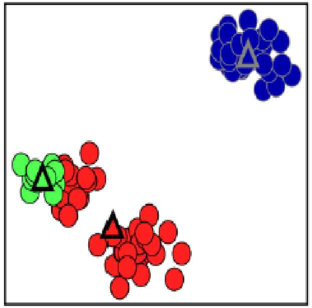
Initialization



Assign points

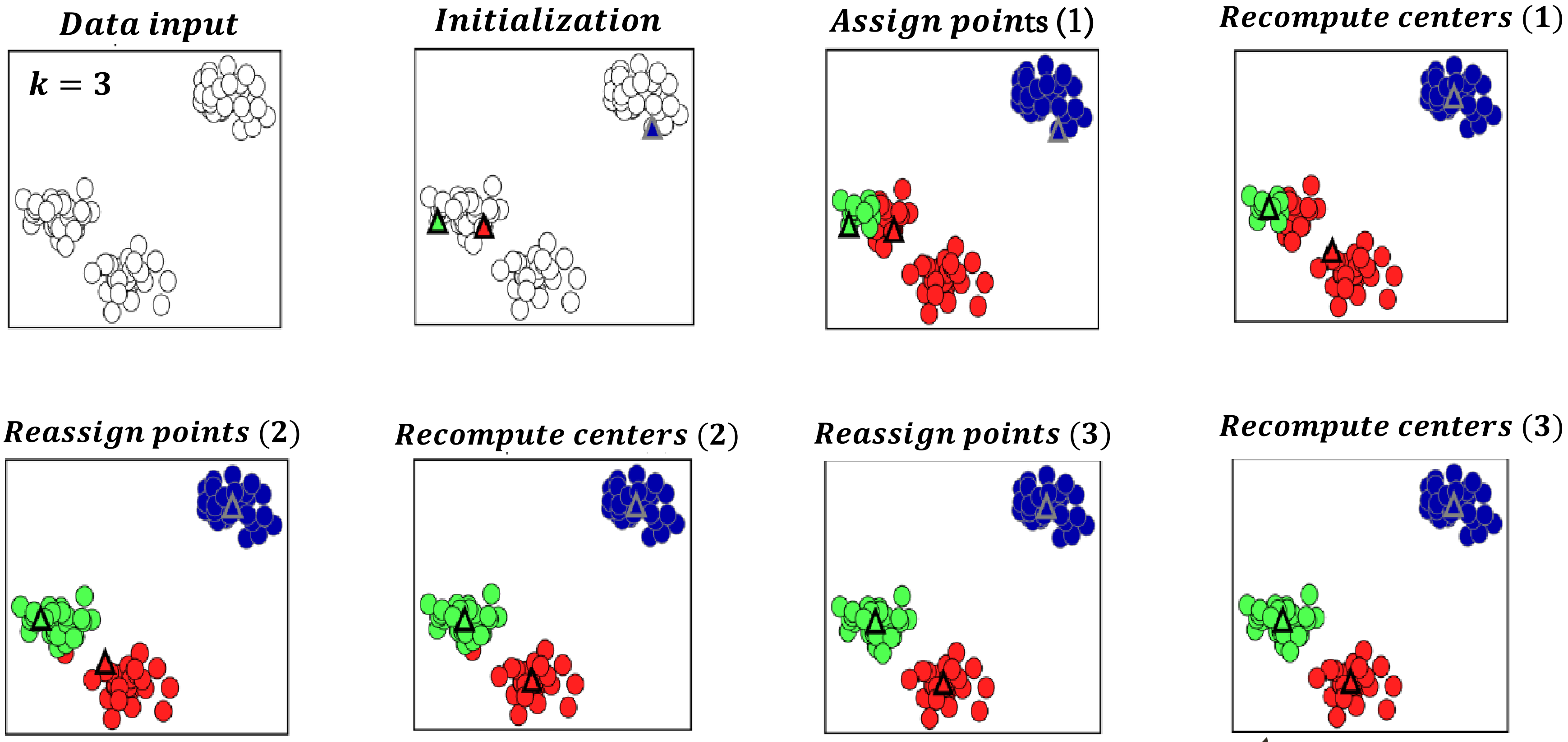


***Recompute centers* (1)**



The cluster centers are updated to be the mean of the assigned points

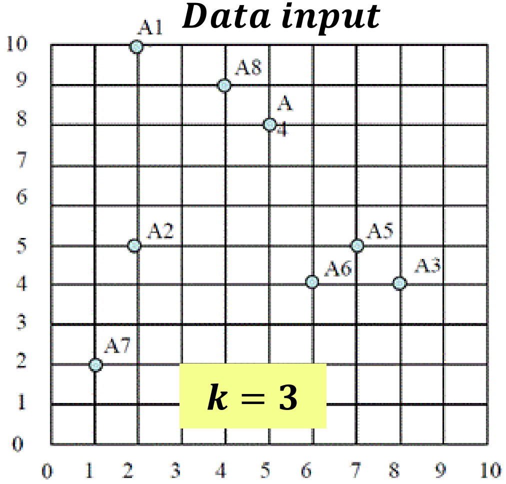
K-Means



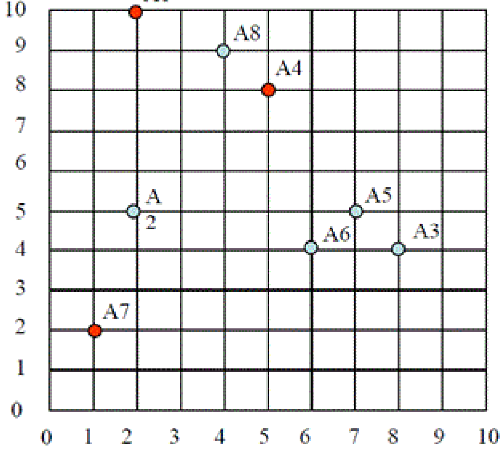
cluster centers remained unchanged ➡ stop

K-Means: Example

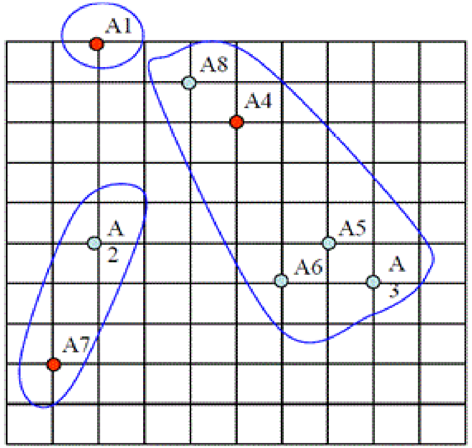
Data input



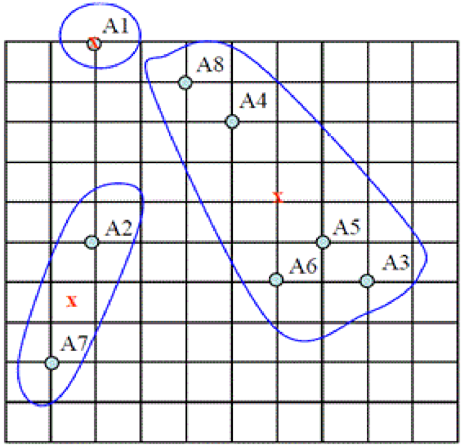
Initialization



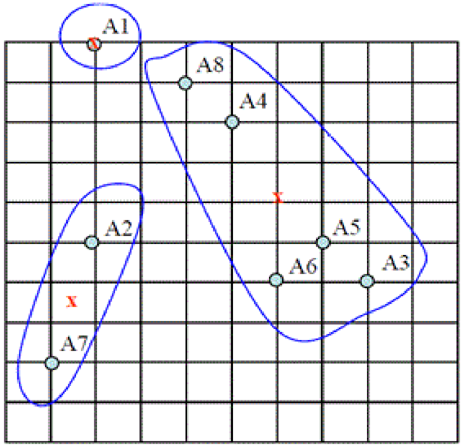
A_1	$d(A_1,\mu_1) = 0$; $d(A_1,\mu_2) = \sqrt{13}$; $d(A_1,\mu_1) = \sqrt{65}$
A_2	$d(A_2,\mu_1) = 5$; $d(A_2,\mu_2) = \sqrt{18}$; $d(A_2,\mu_3) = \sqrt{10}$
A_3	$d(A_3,\mu_1) = 6$; $d(A_3,\mu_2) = 5$; $d(A_3,\mu_3) = \sqrt{53}$
A_4	$d(A_4,\mu_1) = \sqrt{13}$; $d(A_4,\mu_2) = 0$; $d(A_4,\mu_3) = \sqrt{52}$
A_5	$d(A_5,\mu_1) = \sqrt{50}$; $d(A_5,c_2) = \sqrt{13}$, $d(A_5,\mu_2) = \sqrt{45}$
A_6	$d(A_6,\mu_1) = \sqrt{52}$; $d(A_6,\mu_2) = \sqrt{17}$, $d(A_6,\mu_3) = \sqrt{29}$
A_7	$d(A_7,\mu_1) = \sqrt{65}$; $d(A_7,\mu_2) = \sqrt{52}$, $d(A_7,\mu_3) = 0$
A_8	$d(A_8,\mu_1) = \sqrt{5}$; $d(A_8,\mu_2) = \sqrt{2}$, $d(A_8,\mu_3) = \sqrt{52}$

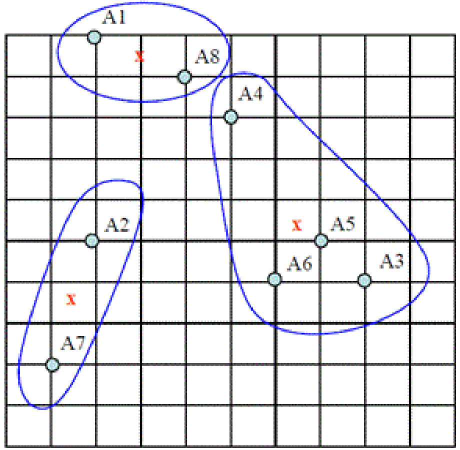


$$\mu_1 = (2,10)$$
$$\mu_2 = (6,6)$$
$$\mu_3 = (1.5,3,5)$$



K-Means: Example

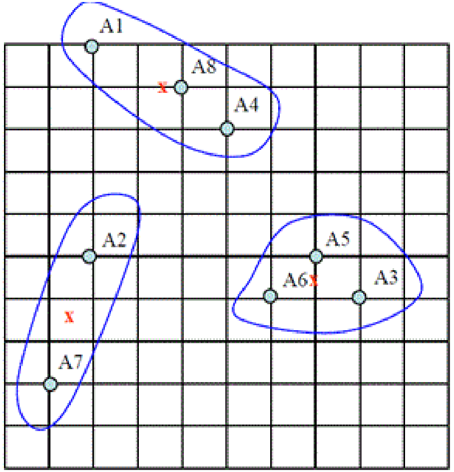




$\{A1, A8\}$ $\mu_1 = (3, 9.5)$

$\{A3, A4, A5, A6\}$ $\mu_2 = (6.5, 5.25)$

$\{A2, A7\}$ $\mu_3 = (1.5, 3.5)$



$\{A1, A4, A8\}$ $\mu_1 = (3.66, 9)$

$\{A3, A5, A6\}$ $\mu_2 = (7, 4.33)$

$\{A2, A7\}$ $\mu_3 = (1.5, 3.5)$

The assignment of points to cluster centers remained Unchanged ➡ stop

K-Means optimization objective

- $c^{(i)}$ index of cluster $(1, 2, \dots, k)$ to which example $x^{(i)}$ is currently assigned
- μ_k cluster centroid k ($\mu_k \in \mathbb{R}^n$)
- $\mu_{c^{(i)}}$ cluster centroid of cluster to which example $x^{(i)}$ has been assigned

Optimization objective

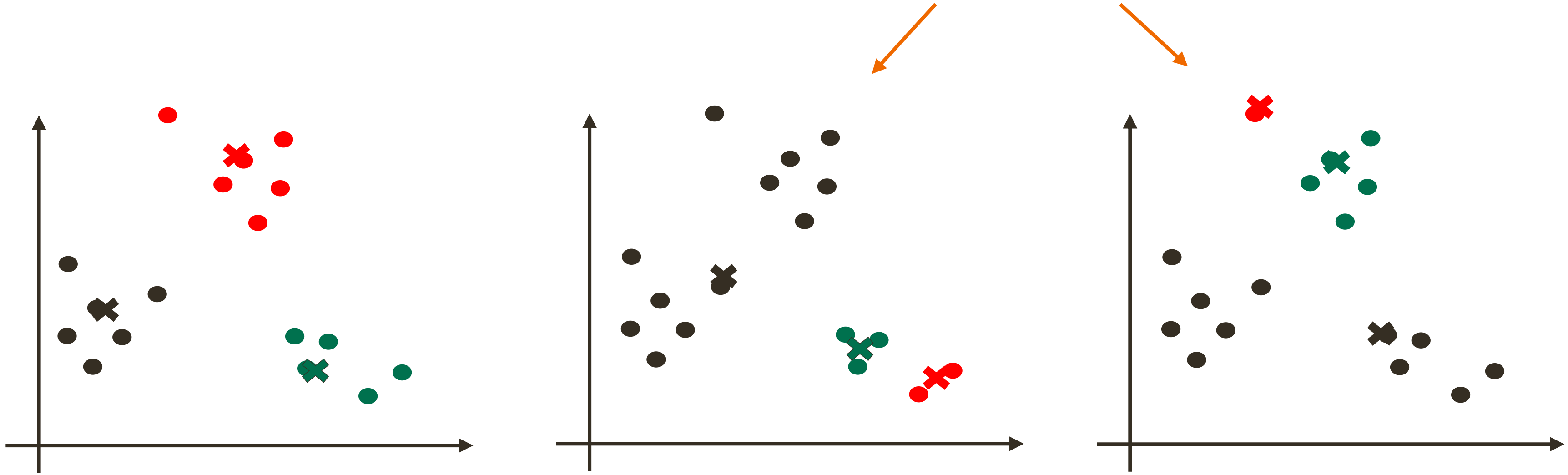
Objective function $\longleftarrow J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k) = \frac{1}{m}\sum_{i=1}^{m}\left\|x^{(i)} - \mu_{c^{(i)}}\right\|^2$

$$\min_{\substack{c^{(1)}, \dots, c^{(m)} \\ \mu_1, \dots, \mu_k}} J(c^{(1)}, \dots, c^{(m)}, \mu_1, \dots, \mu_k)$$

What's the problems

- **Sensitive to initialization**

Local optima



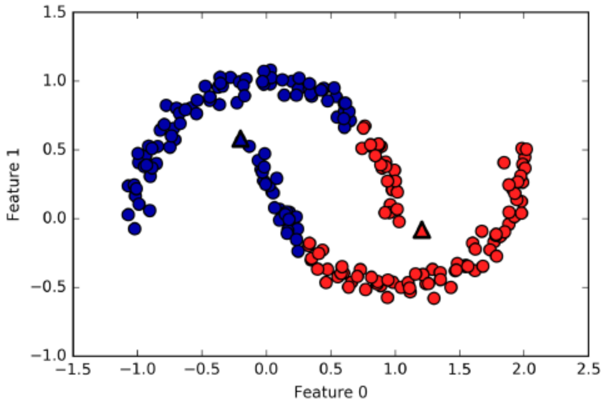
True clustering *Random seed 1* *Random seed 2*

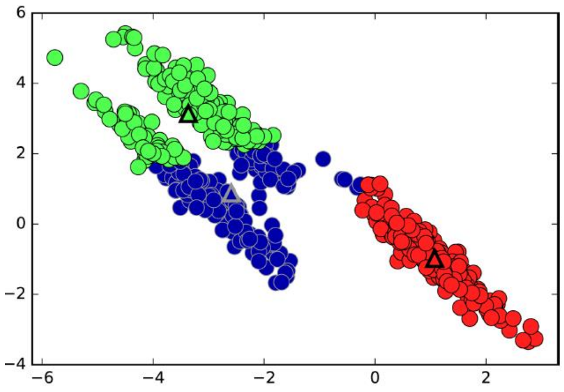
- Solution: repeat many times and take the best solution

What's the problems

- Clusters of non-globular shape

Each cluster is defined by its center ➡ convex shape ➡ *k*-means can only capture relatively simple shapes

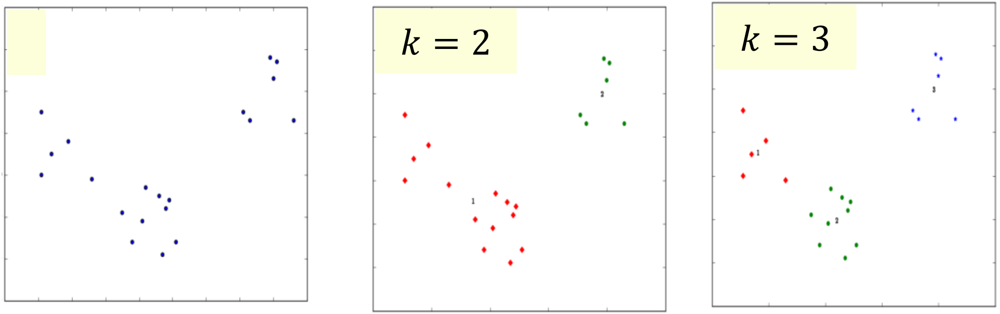




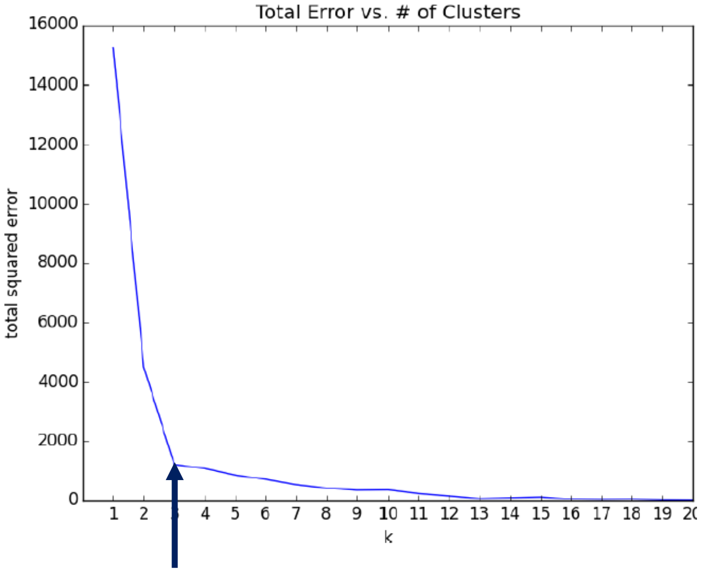
What's the problems

- **Select ion of k**

What is the right value of k?



Solution (Elbow method)



Best solution = where the graph « bends»

Hierarchical clustering

- Produces a set of nested clusters organized as a hierarchical tree

- Two types of hierachical clustering algorithms

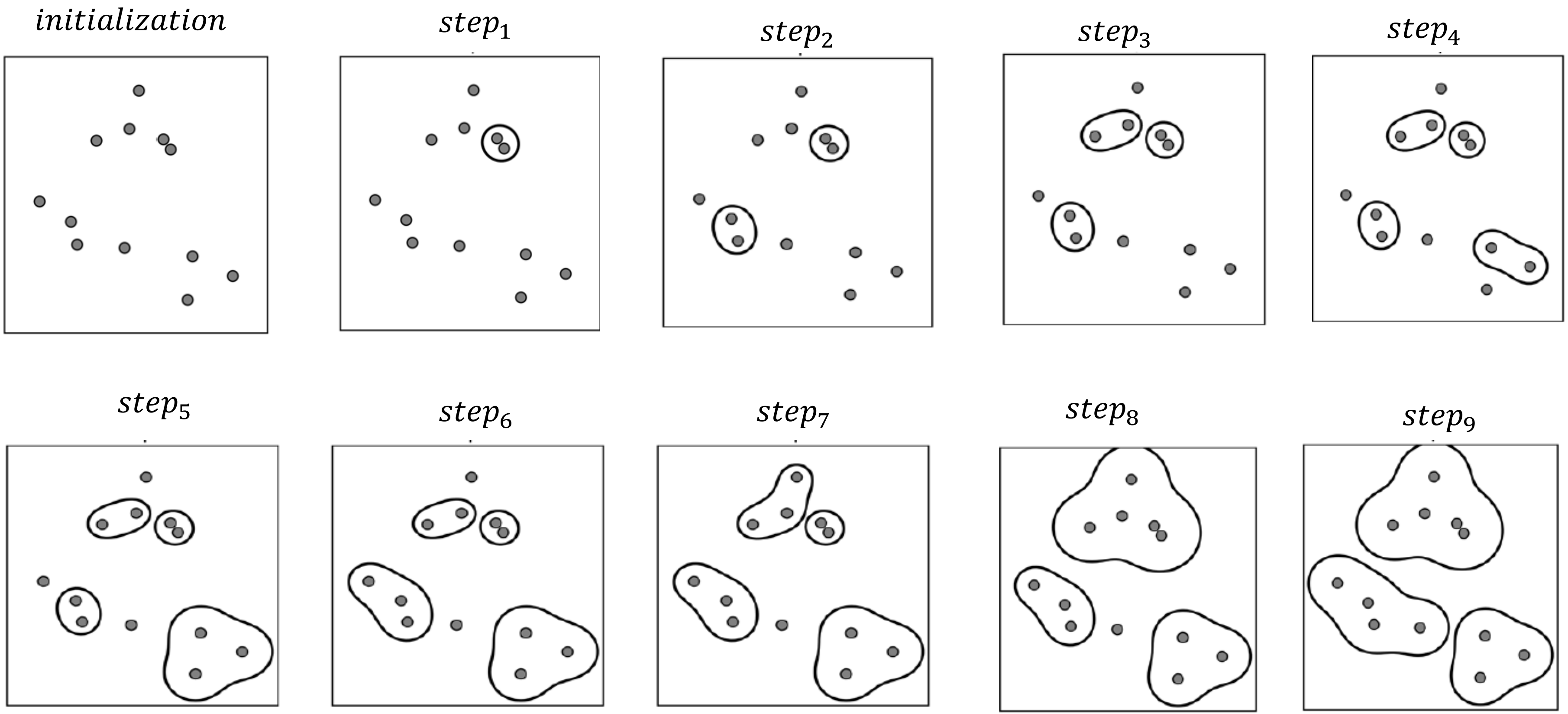
  Agglomerative ($i.e$ bottom-up)

  Start with each sample as an individual cluster and merge the closest pairs of clusters until only one cluster remains.

  Divisive ($i.e$ top-down)

  Start with one cluster that encompasses all our samples, and we iteratively split the cluster into smaller cluster until each cluster only contains one sample

Agglomerative



Distance between two clusters

- How do we define distance between two sets of points?

  ✓ Single-link distance

  ✓ Complete-link

  ✓ Average-link

	1	2	3	4	5
1	0				
2	9	0			
3	3	7	0		
4	6	5	9	0	
5	11	10	2	8	0

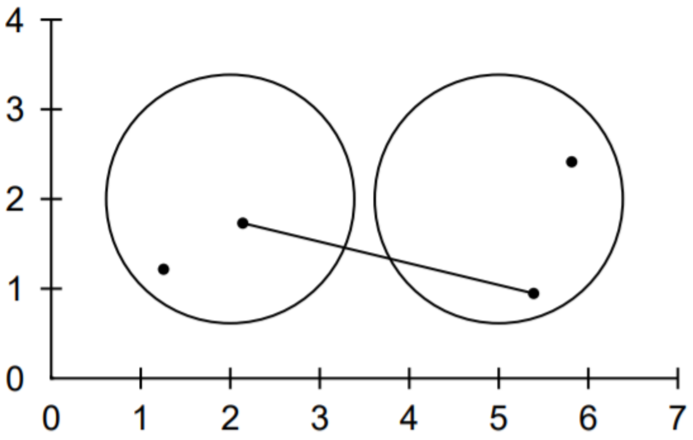
	1	2	35	4
1	0		?	
2	9	0	?	
35	?	?	?	?
4	6	5	?	0

Single-link clustering

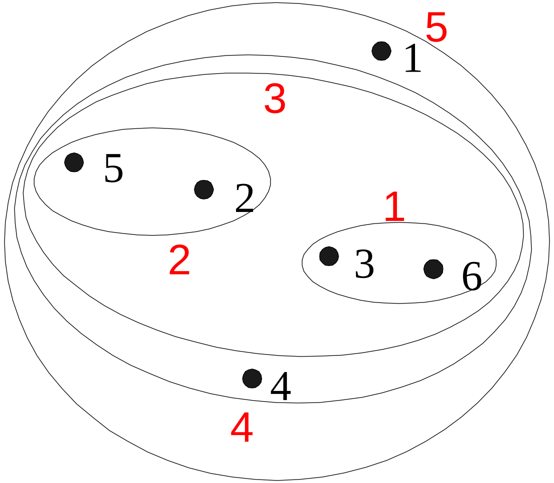
- **Single-link distance** between clusters $\boldsymbol{C_i}$ and is $\boldsymbol{C_j}$ the **minimum distance** between any object in $\boldsymbol{C_i}$ and any object in $\boldsymbol{C_j}$

- The distance is defined by the two most similar objects

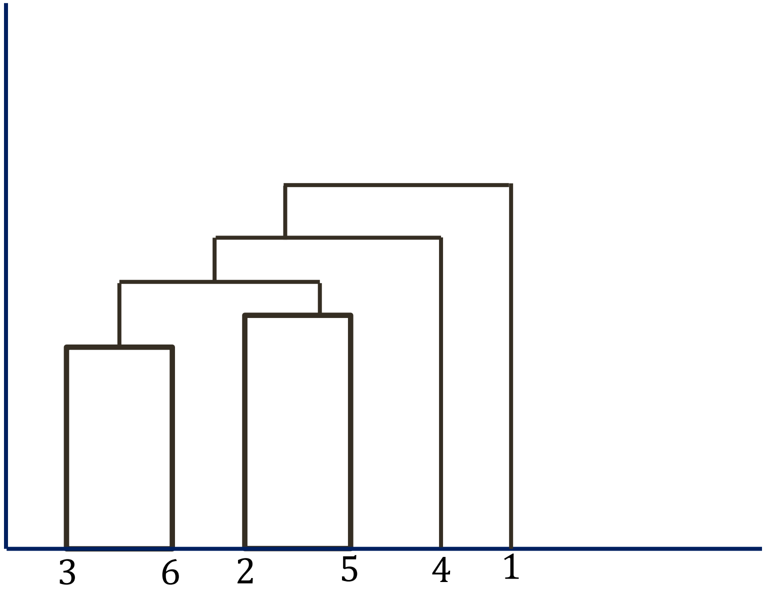
$$D_{single} = \min\{d(x,y) \mid x \in C_i, y \in C_j\}$$



Single-link clustering: example



Nested clusters



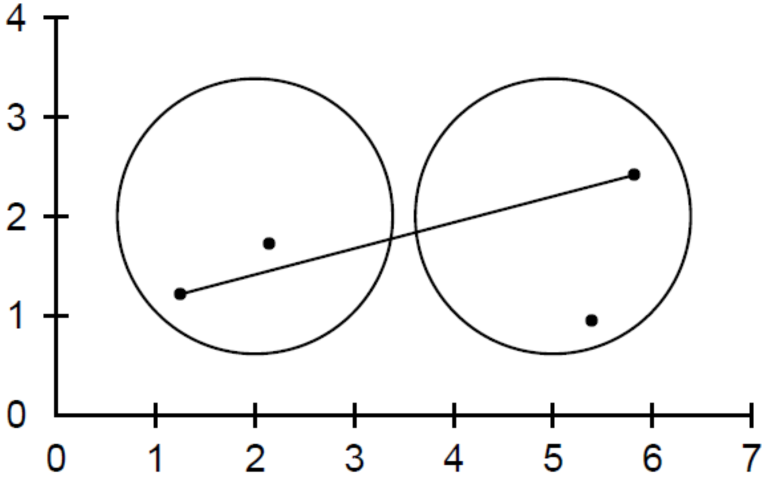
Dendrogram

Complete-link clustering

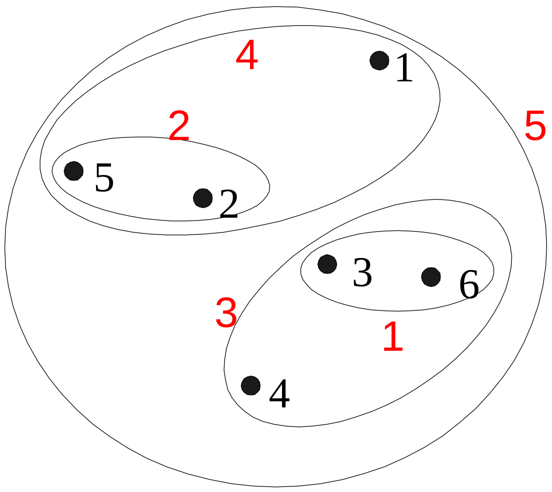
- Complete-link distance between clusters $\boldsymbol{C_i}$ and $\boldsymbol{C_j}$ is the **maximum distance** between any object in $\boldsymbol{C_i}$ and any object in $\boldsymbol{C_j}$

- The distance is defined by the **two most dissimilar objects**

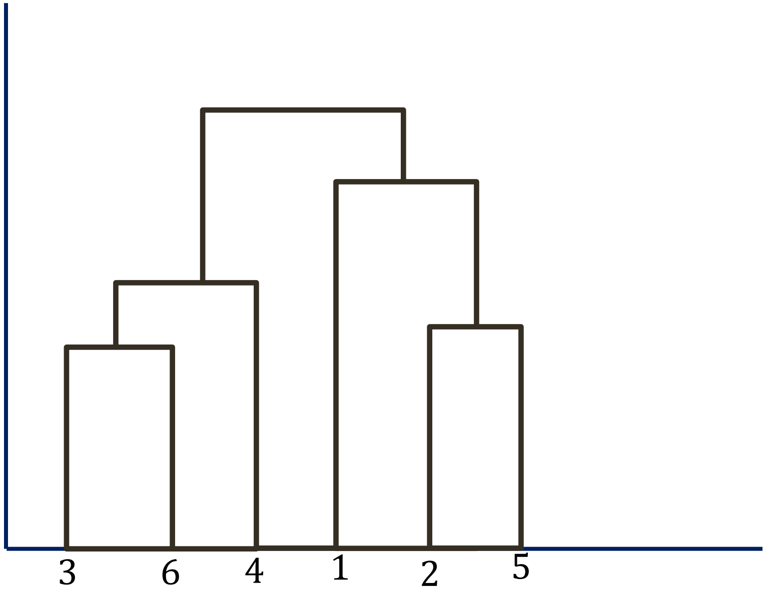
$$D_{complete} = \max\{d(x, y) \mid x \in C_i, y \in C_j\}$$



Complete-link clustering: example



Nested clusters

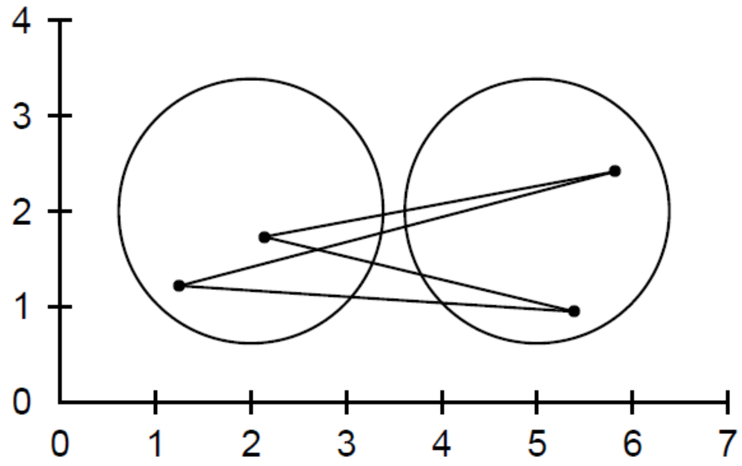


Dendrogram

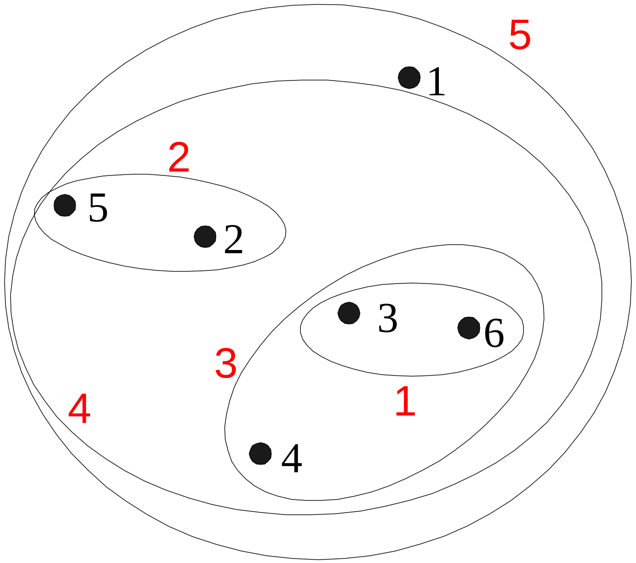
Average-link clustering

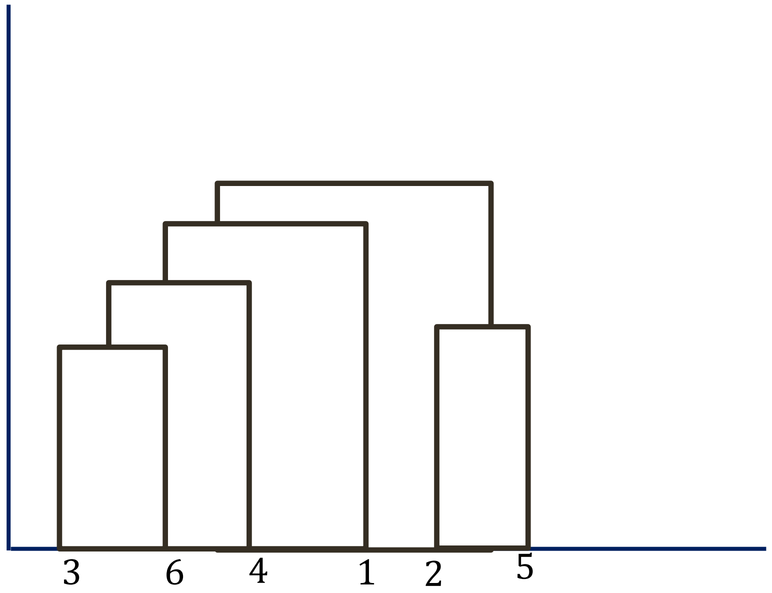
The distance between two clusters is represented by the the average distance of all pairs of data objects belonging to different clusters

$$D_{average} = \text{avg}\,\{d(x, y) \mid x \in C_i, y \in C_j\}$$



Average-link clustering: example





Centroid clustering

- **Centroid distance** between two clusters C_i et C_j is the distance between the centroid r_i of C_i and the centroid r_i of C_j

$$D_{centroids}(C_i, C_j) = d(r_i, r_j)$$