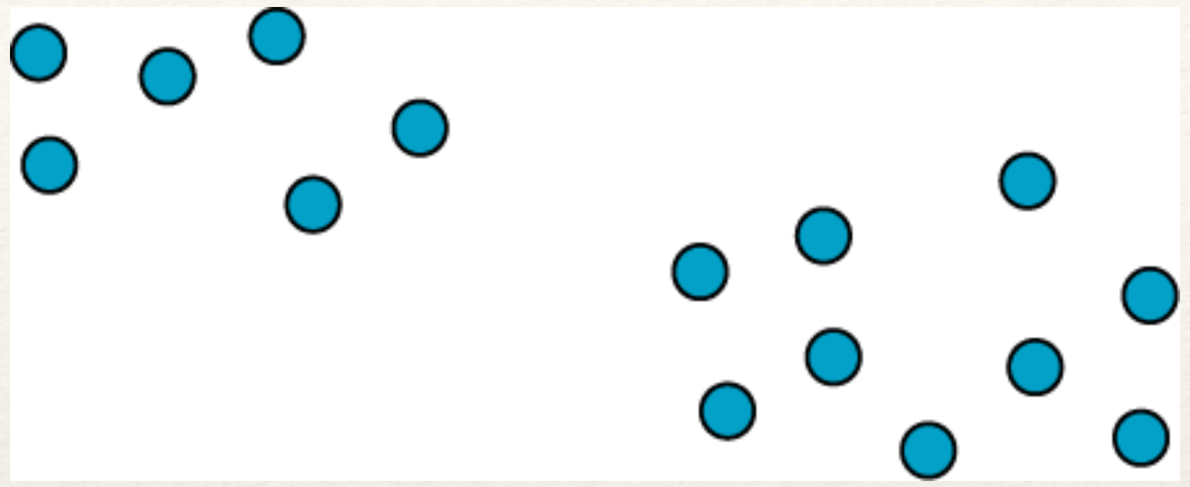
Clustering

Unsupervised Learning

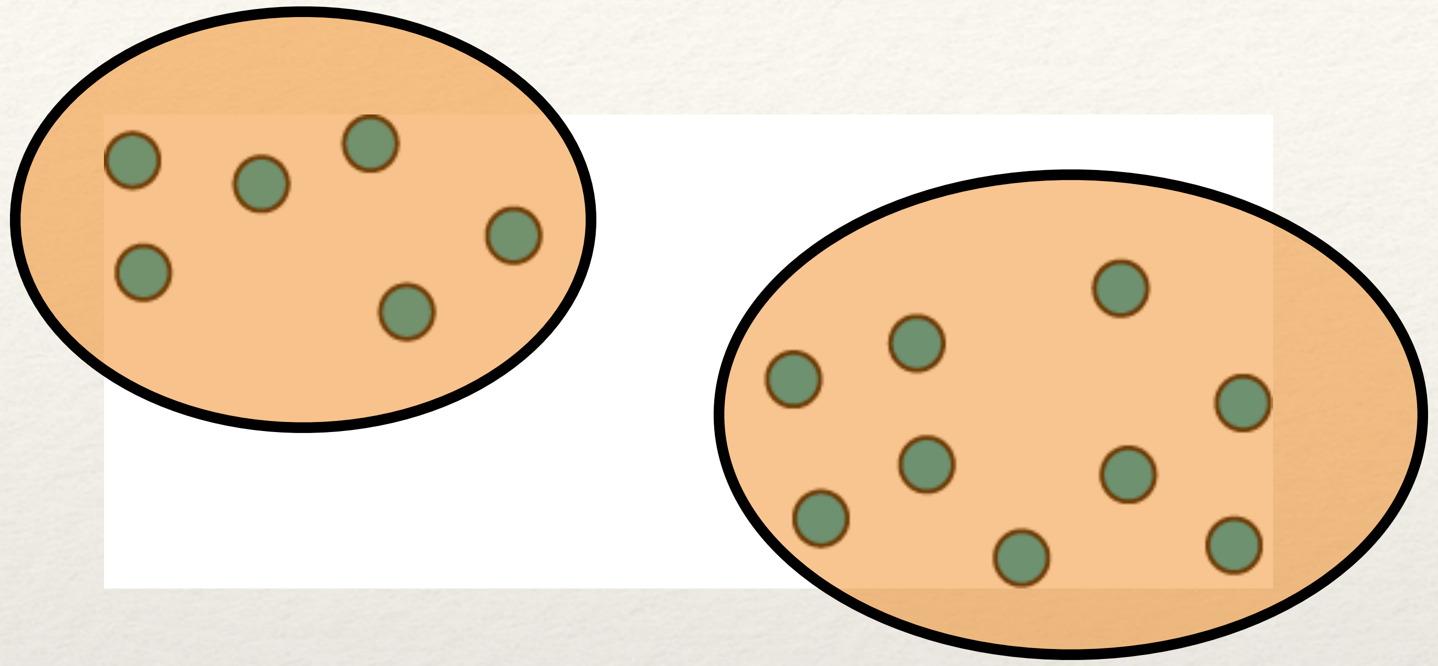
Clustering is a hard problem



Many possibilities; What is best clustering ?

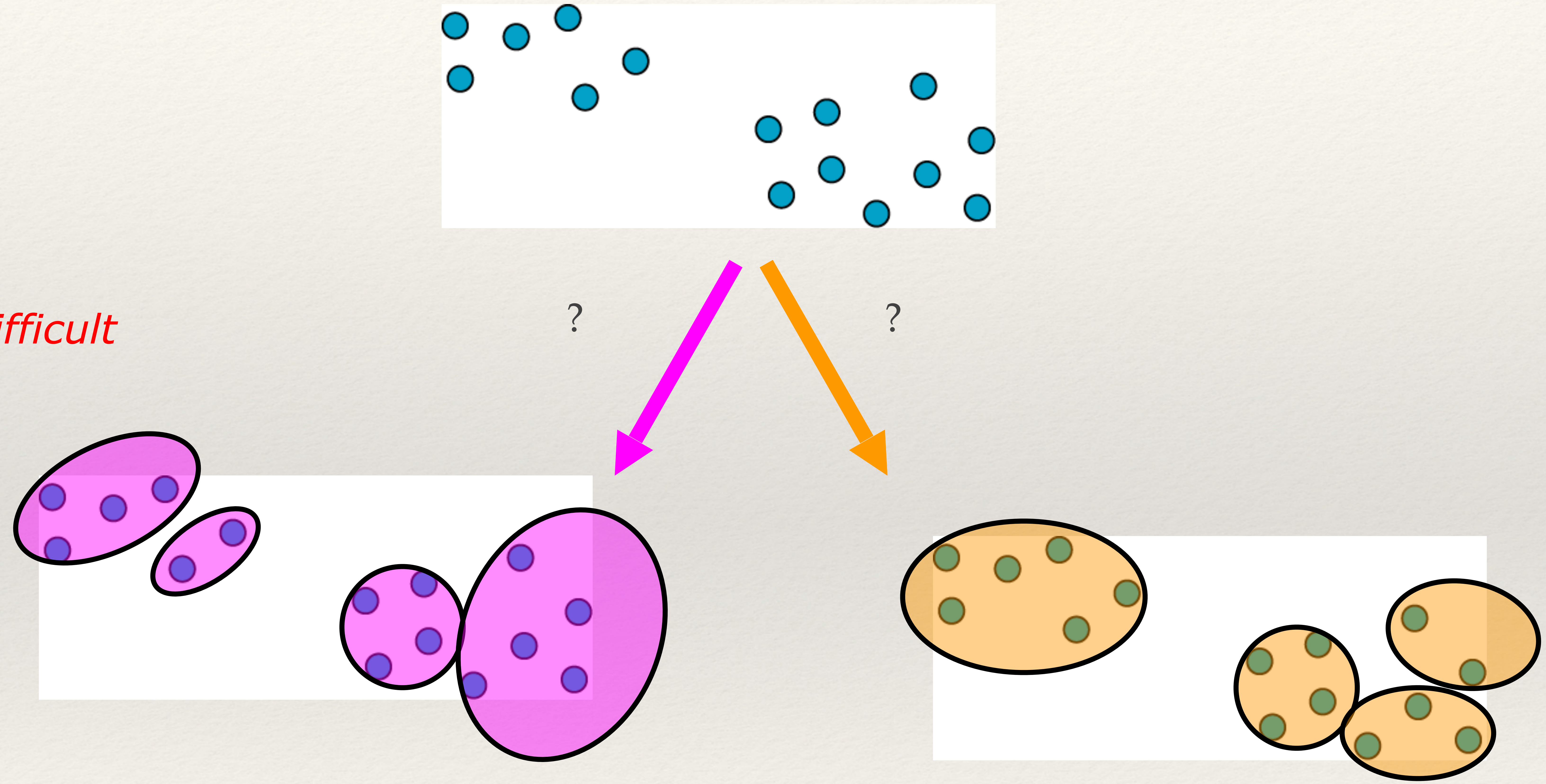
Clustering is a hard problem

2 clusters: easy



Clustering is a hard problem

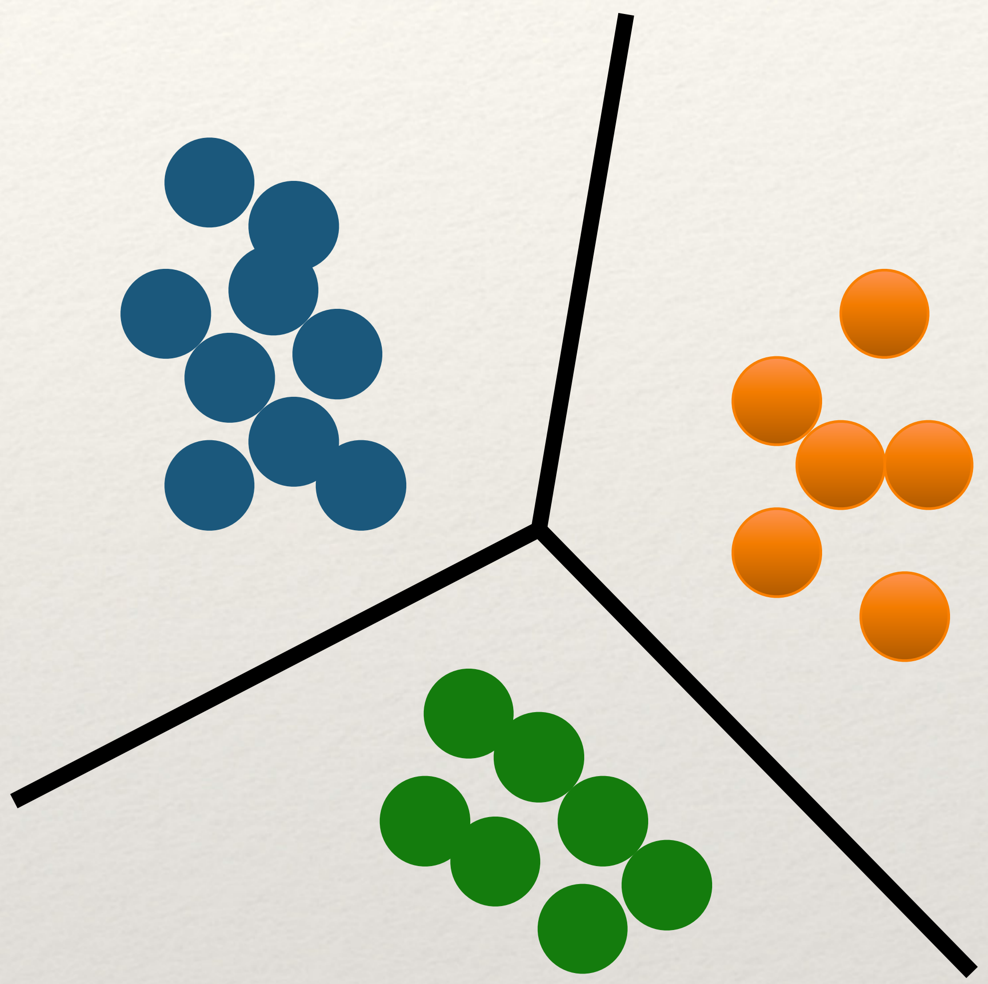
4 clusters: difficult



Many possibilities; What is best clustering ?

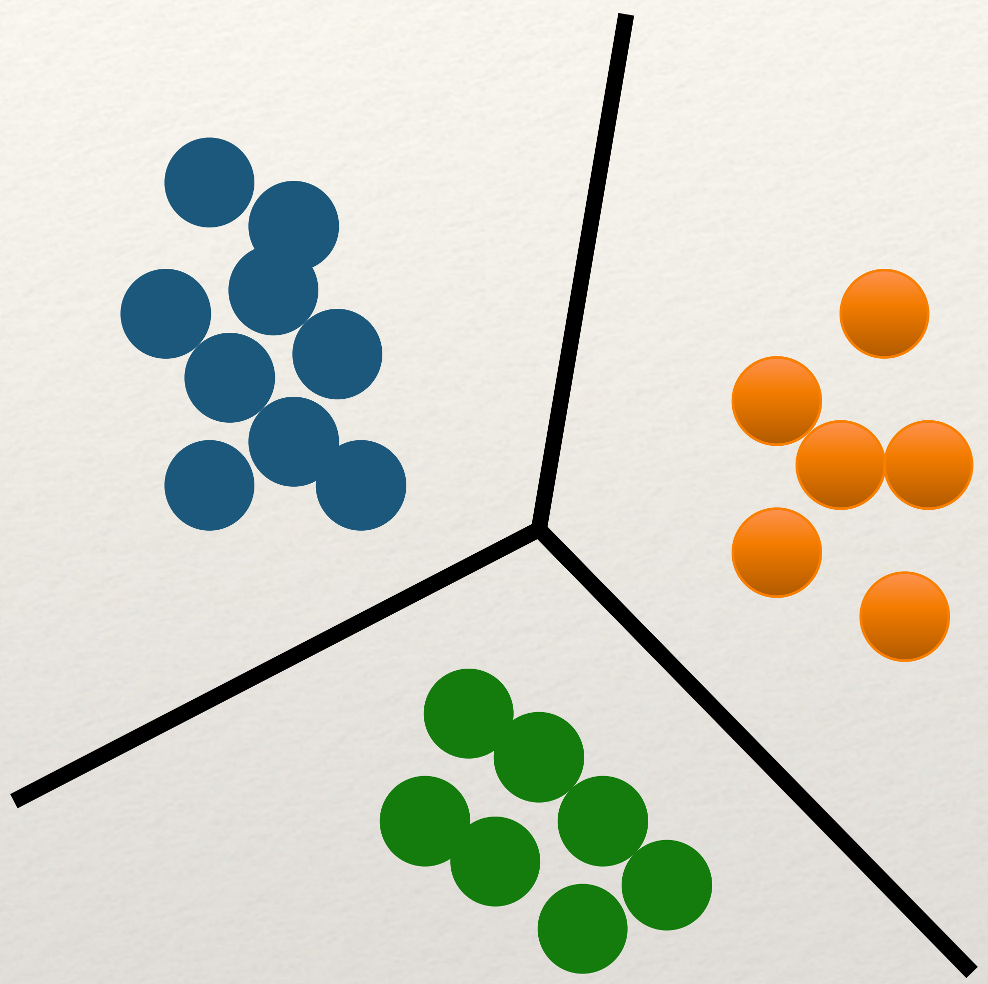
Clustering

- Hierarchical clustering
- K-means clustering
- How many clusters?



Clustering

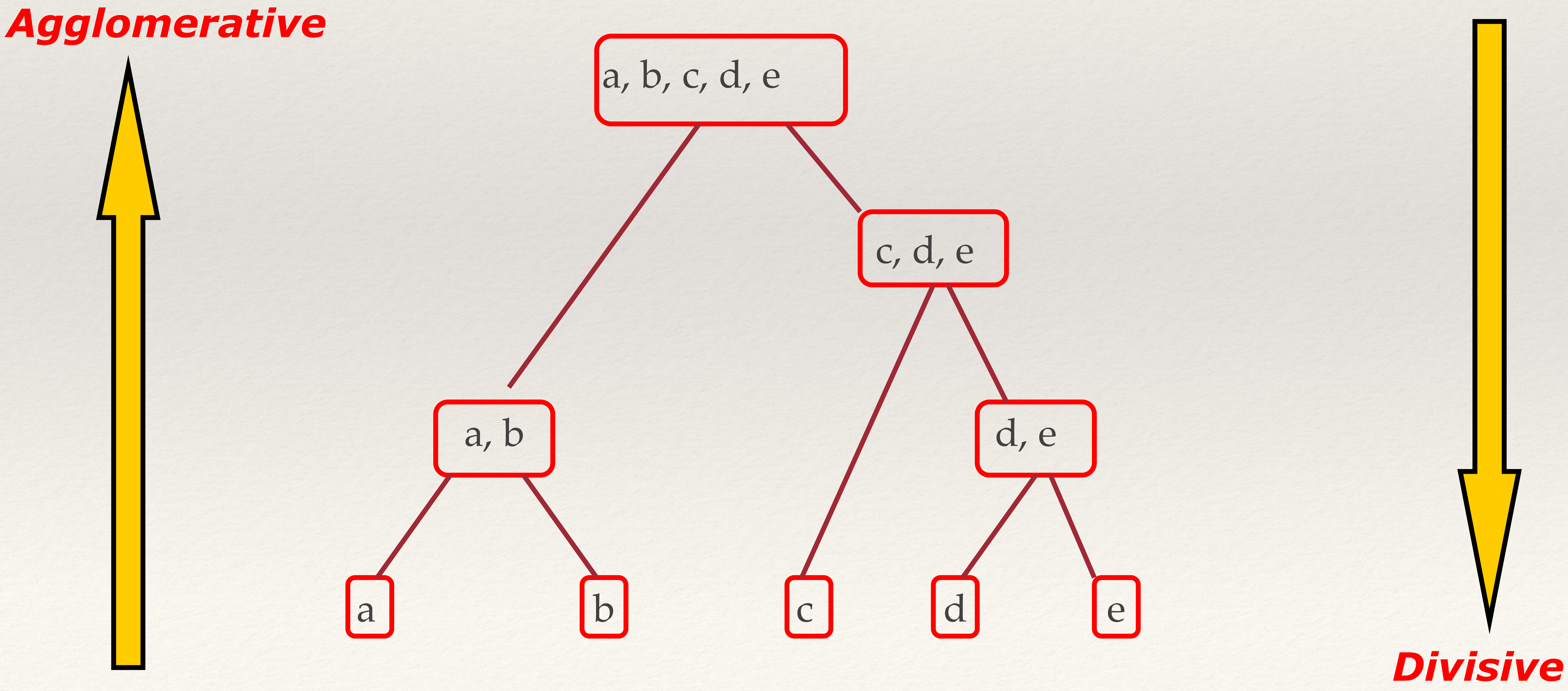
- Hierarchical clustering
- K-means clustering
- How many clusters?



Hierarchical Clustering

To cluster a set of data D={P_1, P_2, …,P_N}, hierarchical clustering proceeds through a series of partitions that runs from a single cluster containing all data points, to N clusters, each containing 1 data points.

Two forms of hierarchical clustering:



Hierarchical Clustering

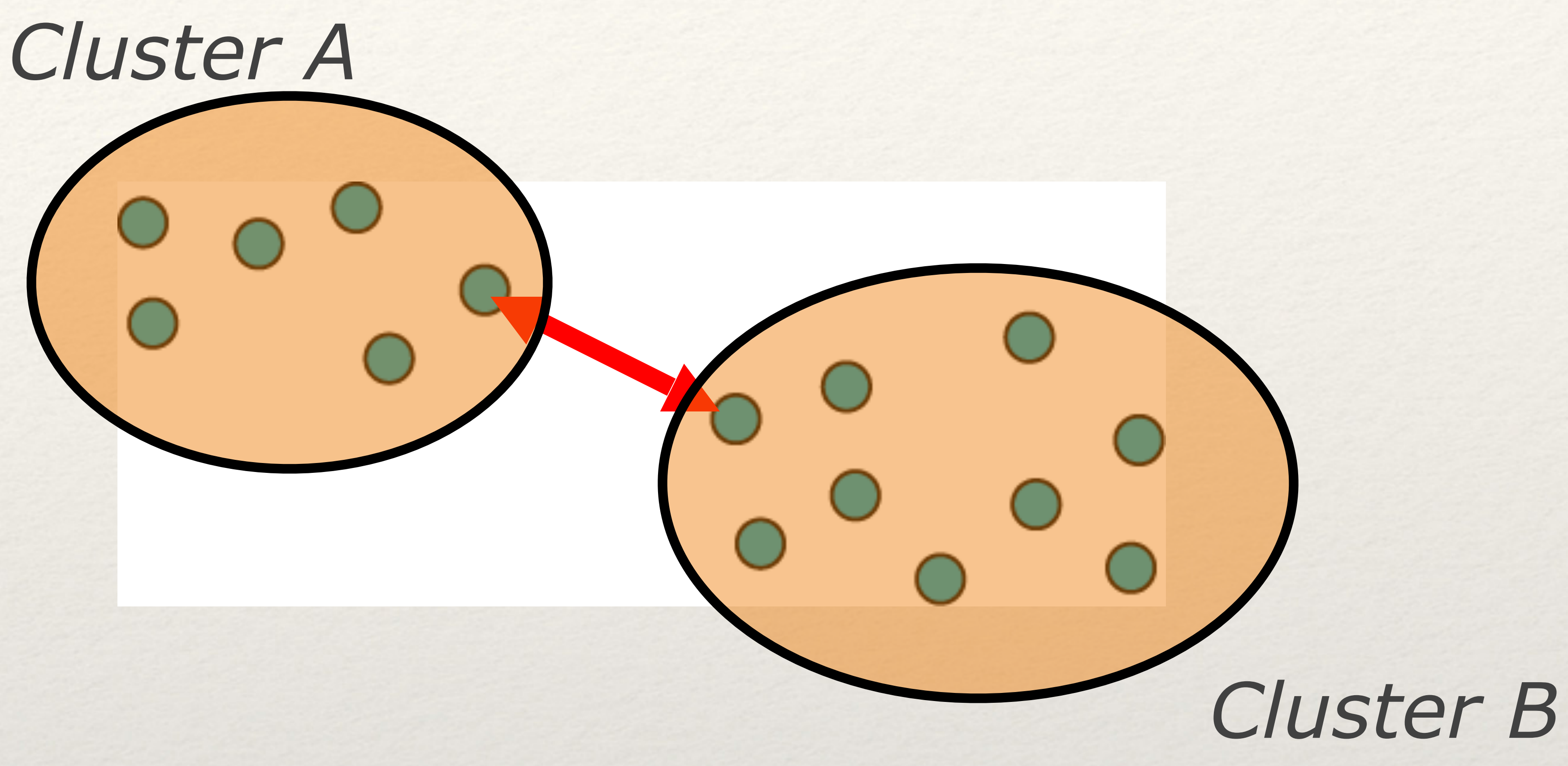
Methods differ in their definition of inter-cluster distance (or similarity)

Hierarchical Clustering

1) Single linkage clustering

Distance between closest pairs of points:

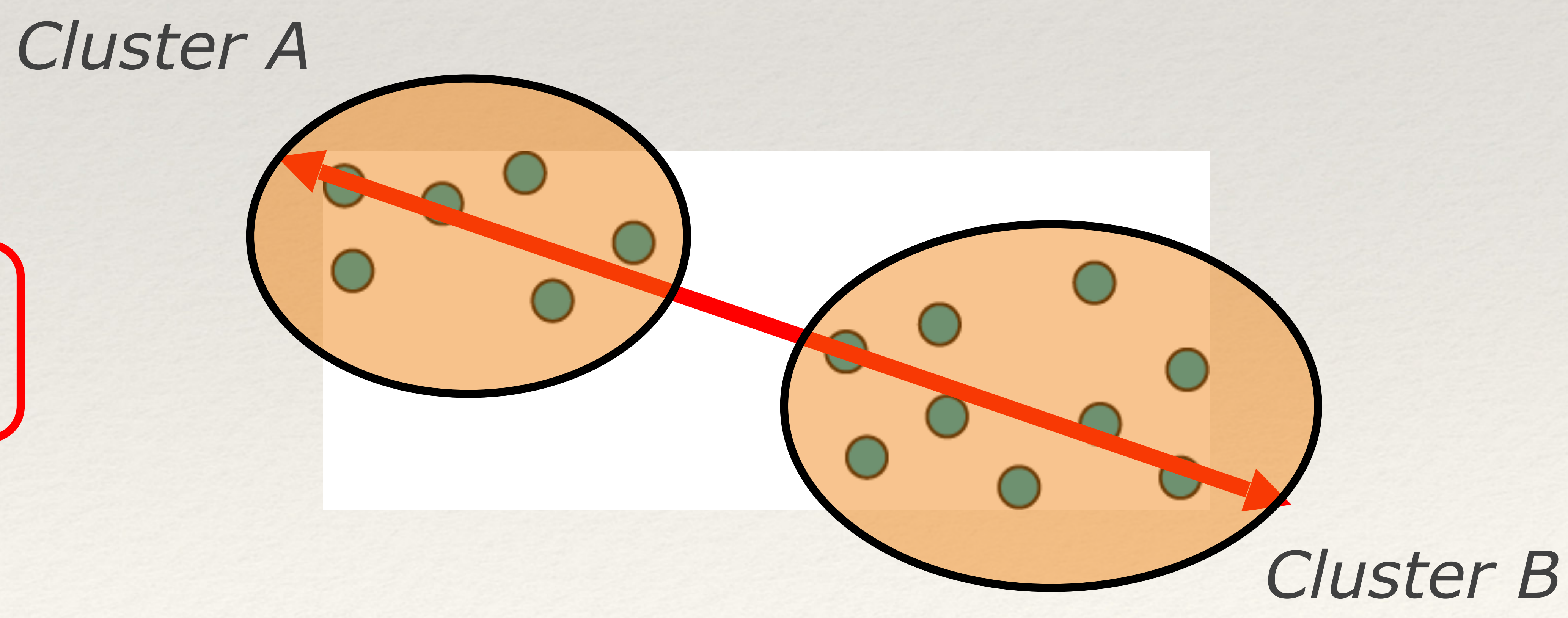
$$d(A, B) = \min\{d(P_i, P_j), P_i \in A, P_j \in B\}$$



2) Complete linkage clustering

Distance between farthest pairs of points:

$$d(A, B) = \max\{d(P_i, P_j), P_i \in A, P_j \in B\}$$

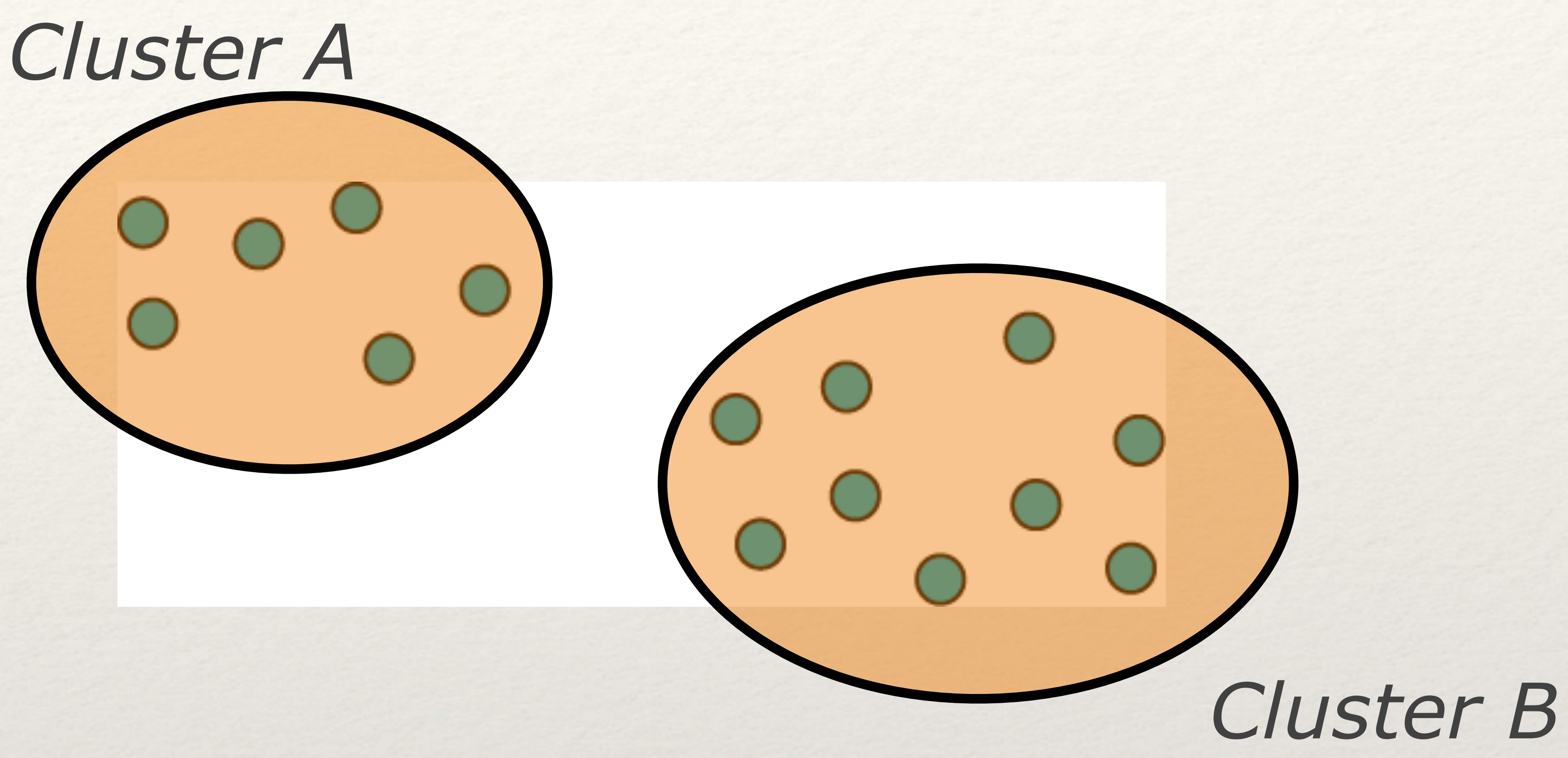


Hierarchical Clustering

3) Average linkage clustering

Mean distance of all mixed pairs of points:

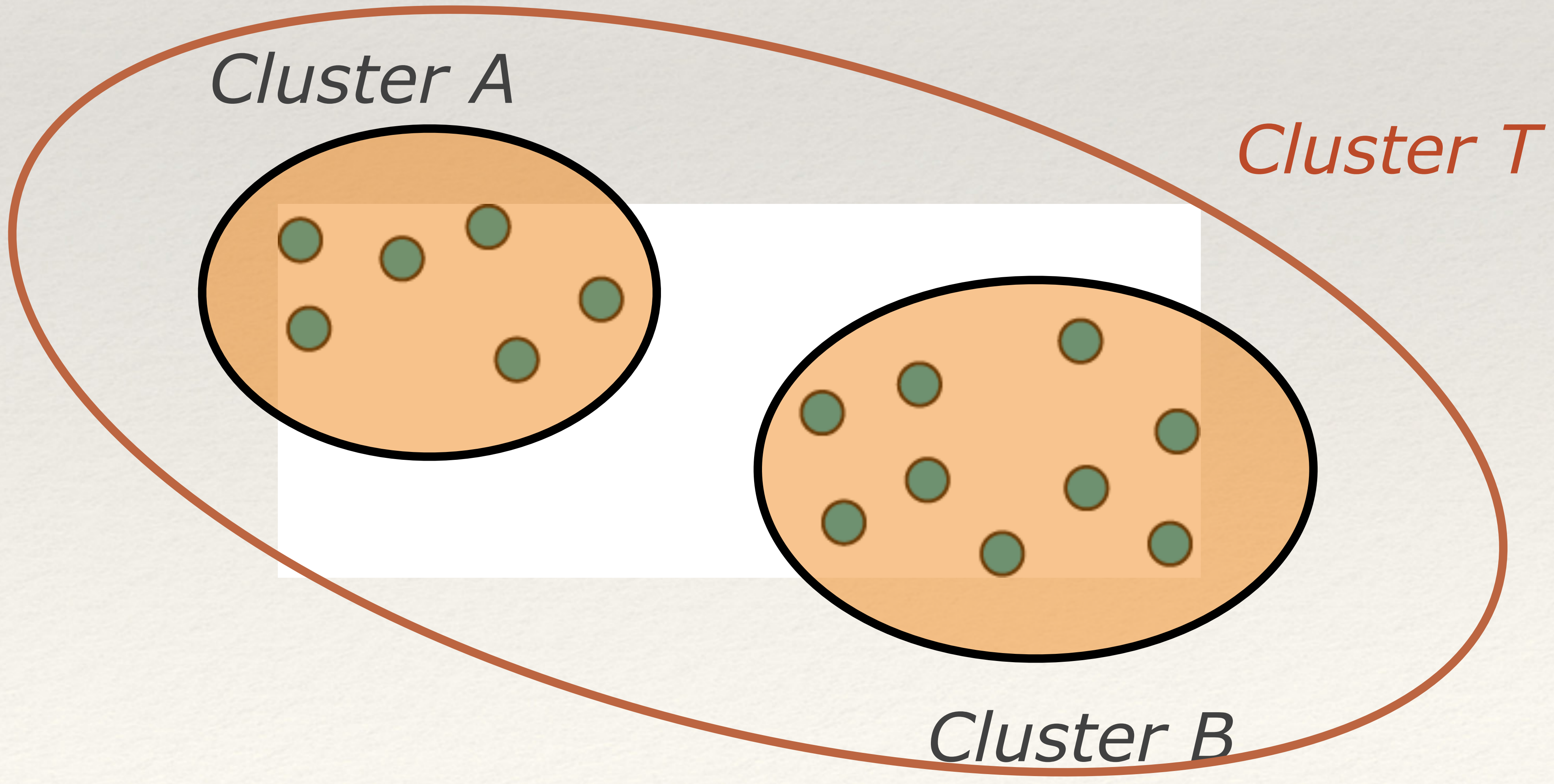
$$\boxed{d(A,B) = \frac{\sum_{i=1}^{N_A} \sum_{j=1}^{N_B} d(P_i, P_j)}{N_A N_B}}$$



4) Average group linkage clustering

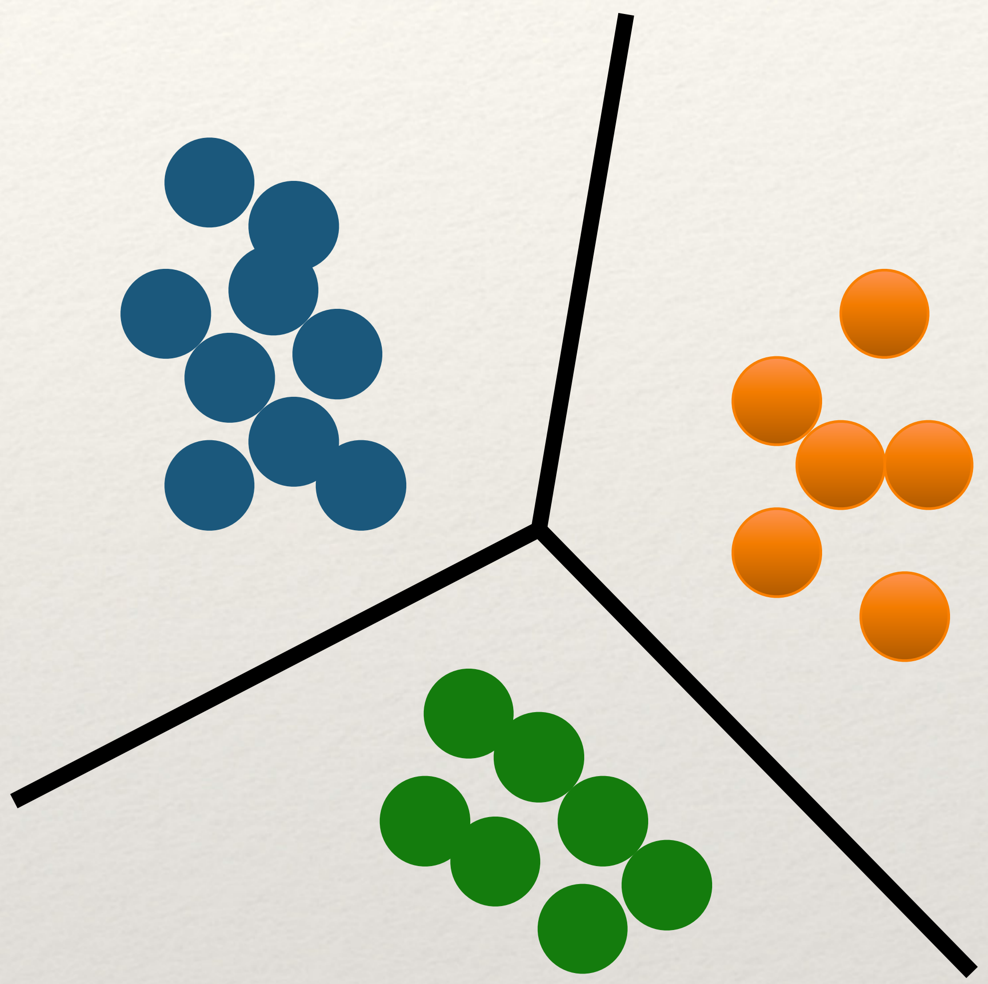
Mean distance of all pairs of points:

$$\boxed{d(A,B) = \frac{\sum_{i=1}^{N_T} \sum_{j=1}^{N_T} d(P_i, P_j)}{N_T^2}}$$

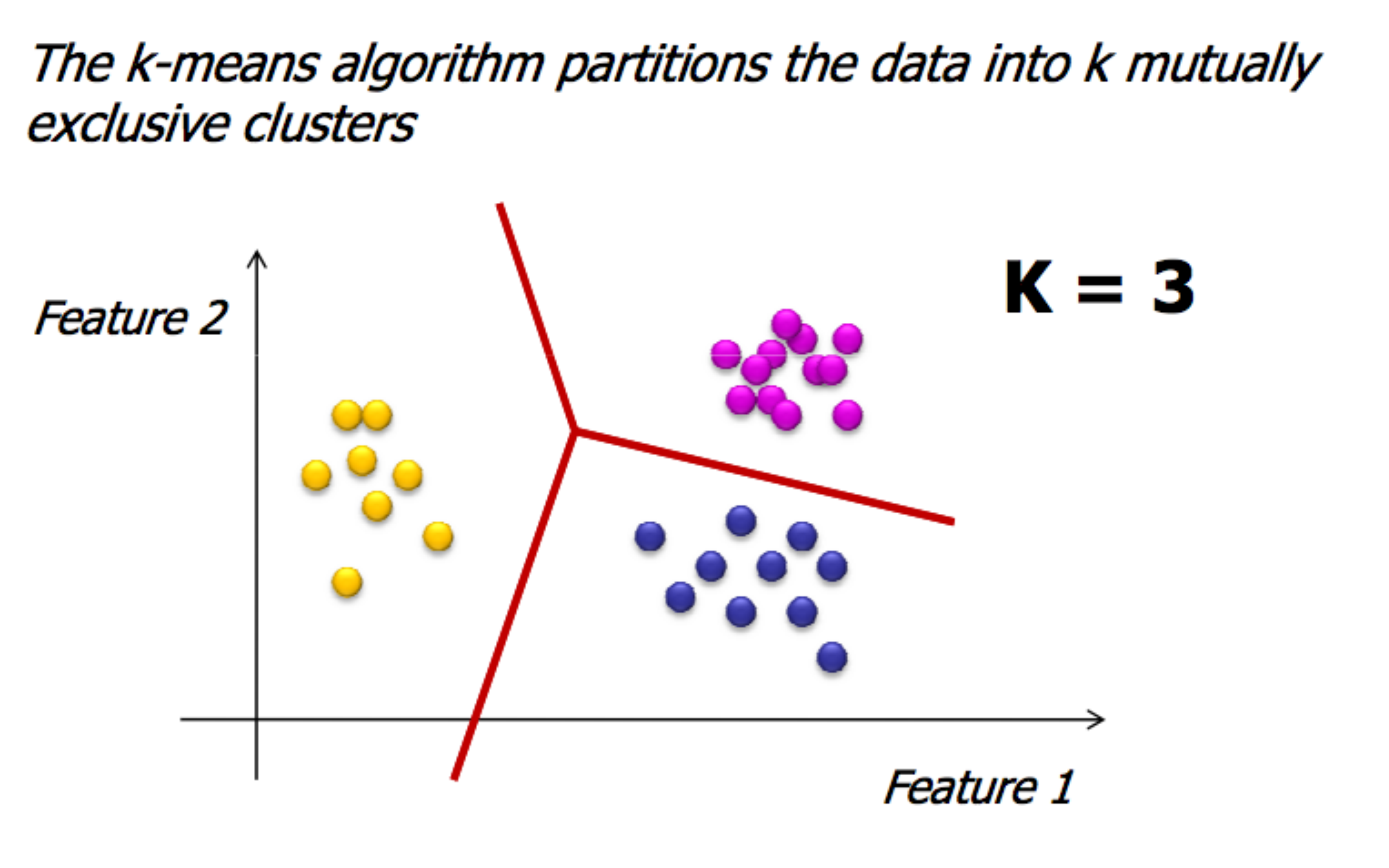


Clustering

- Hierarchical clustering
- K-means clustering
- How many clusters?



K-means clustering

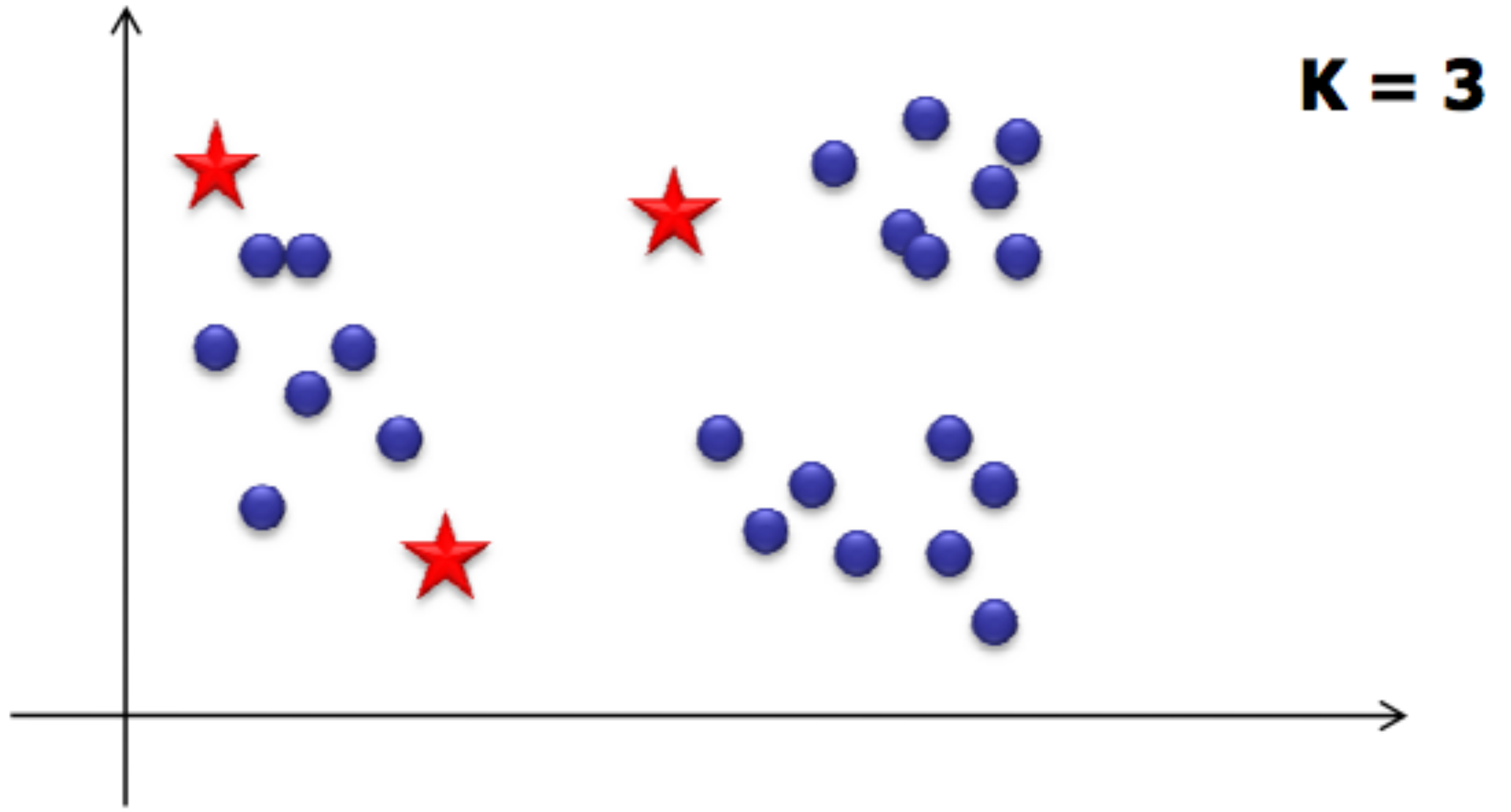


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters, K
- Randomly choose initial positions of K centroids

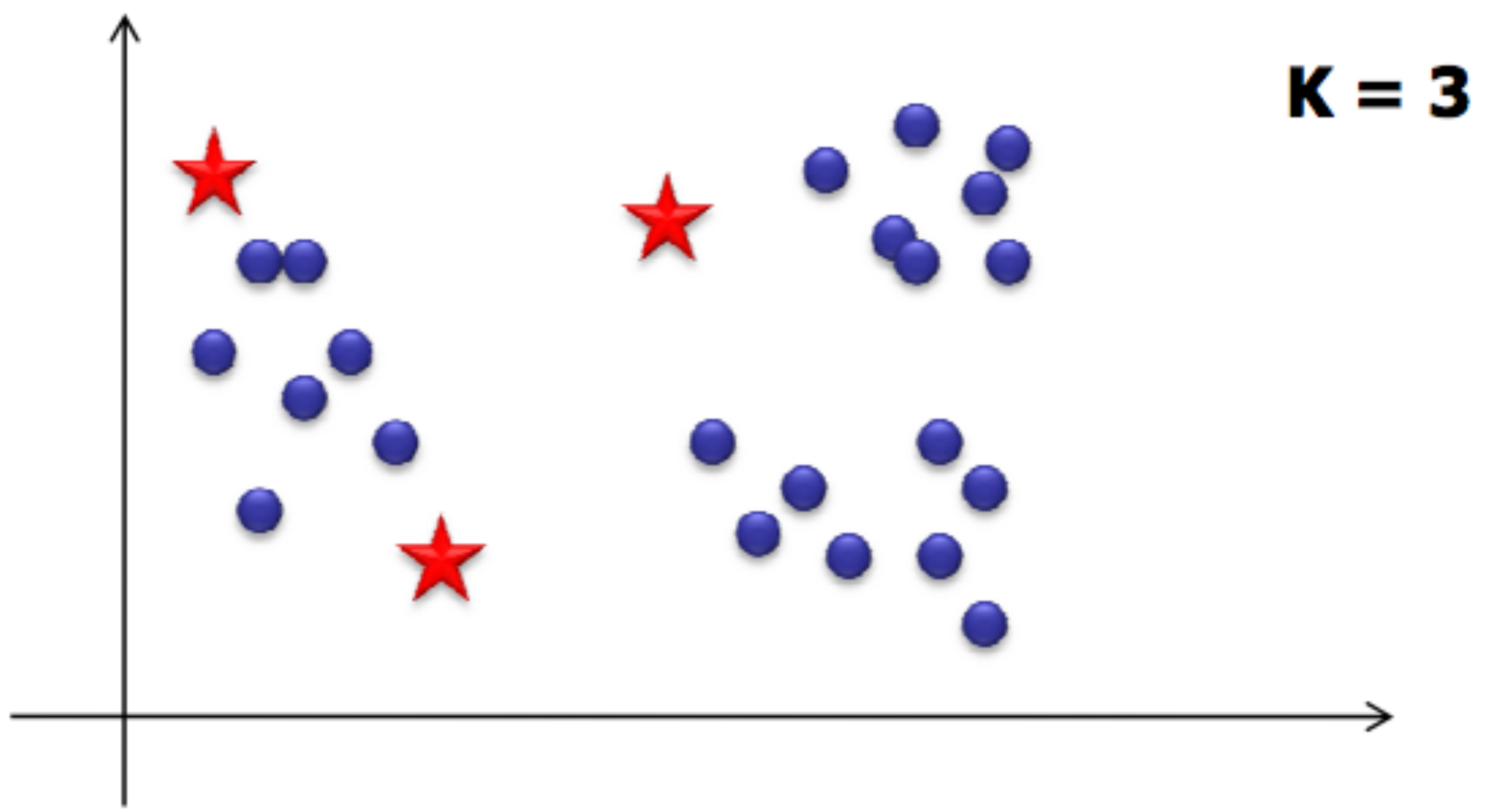


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters, K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)

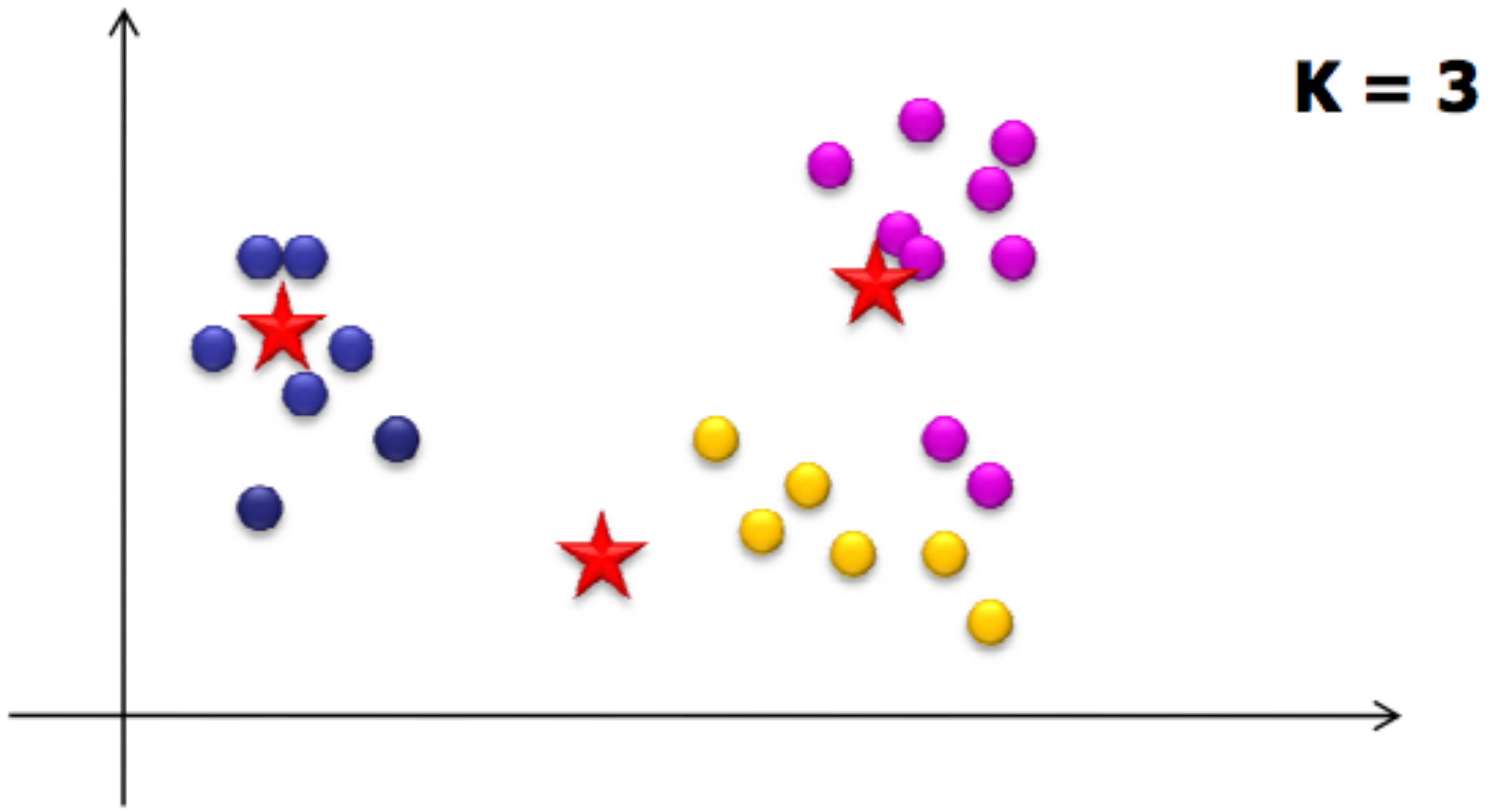


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)
- Re-compute centroid positions
- If solution converges → Stop!

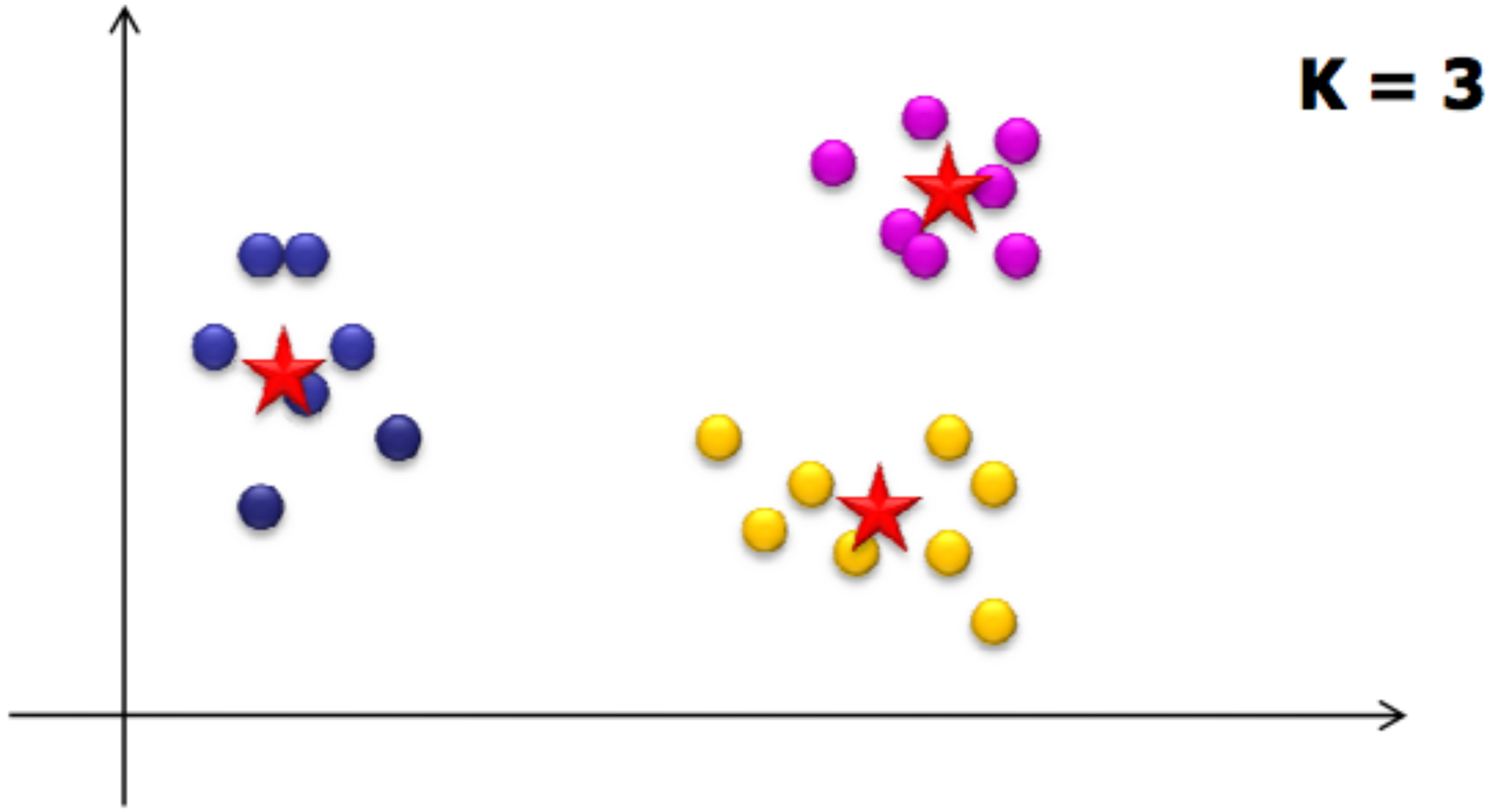


(http://www.weizmann.ac.il/midrasha/courses/)

K-means clustering

Algorithm description

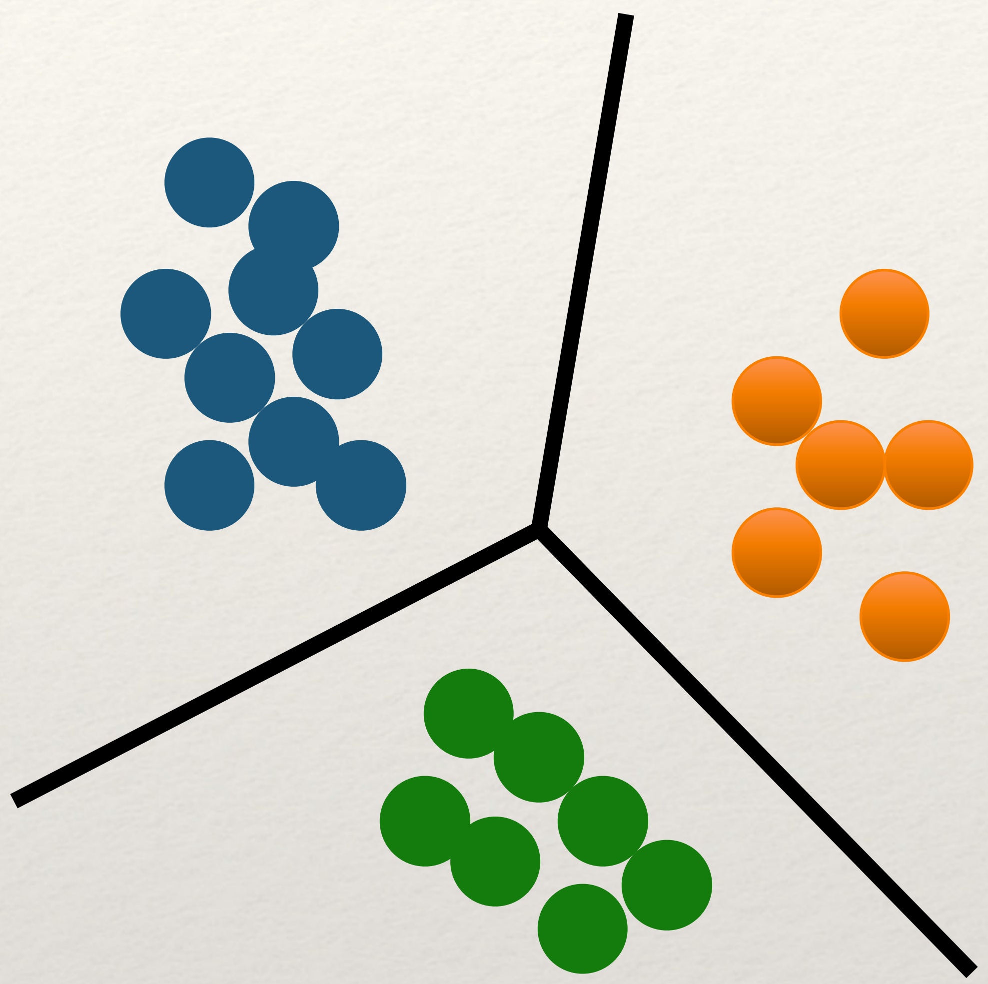
- Choose the number of clusters - K
- Randomly choose initial positions of K centroids
- Assign each of the points to the "nearest centroid" (depends on distance measure)
- Re-compute centroid positions
- **If solution converges → Stop!**



(http://www.weizmann.ac.il/midrasha/courses/)

Clustering

- Hierarchical clustering
- K-means clustering
- How many clusters?



Cluster Validation

Clustering is hard: it is an unsupervised learning technique. Once a Clustering has been obtained, it is important to assess its validity!

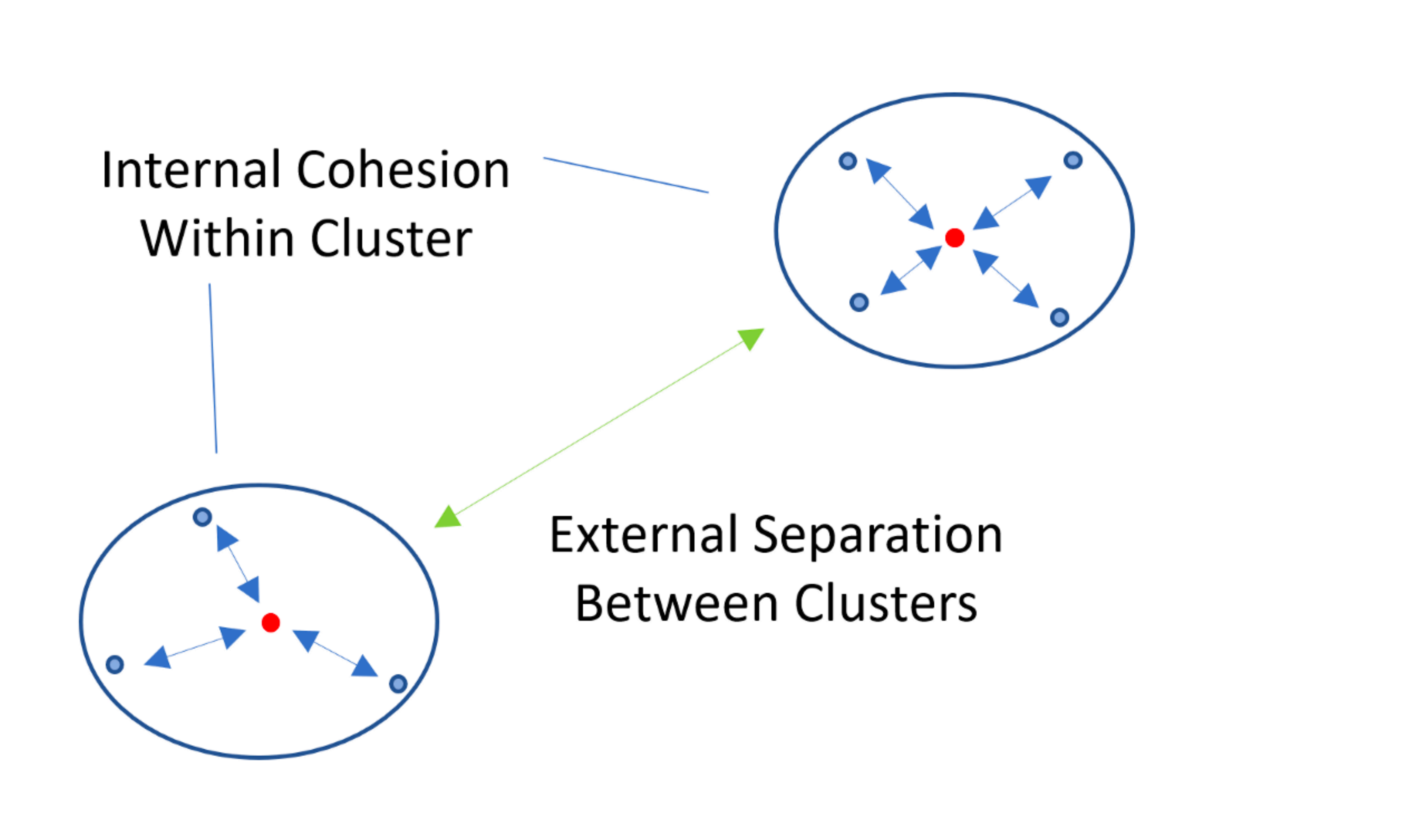
The questions to answer:

- ➢Did we choose the right number of clusters?
- ➢Are the clusters compact?
- ➢Are the clusters well separated?

To answer these questions, we need a quantitative measure of the cluster sizes:

- ➢intra-cluster size
- ➢Inter-cluster distances

Cluster Validation

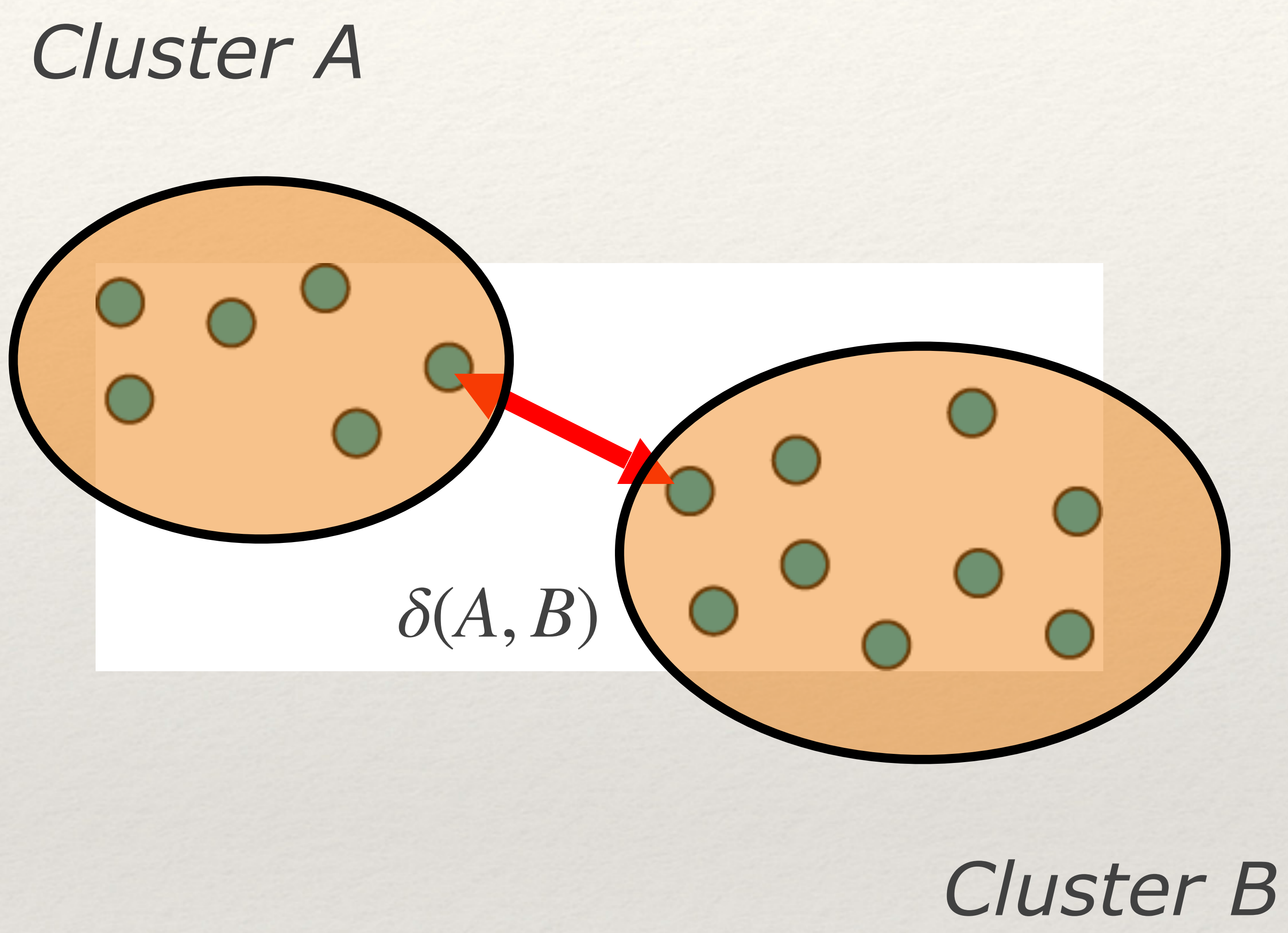


Inter cluster size

Computing $\delta(A, B)$:

Several options:

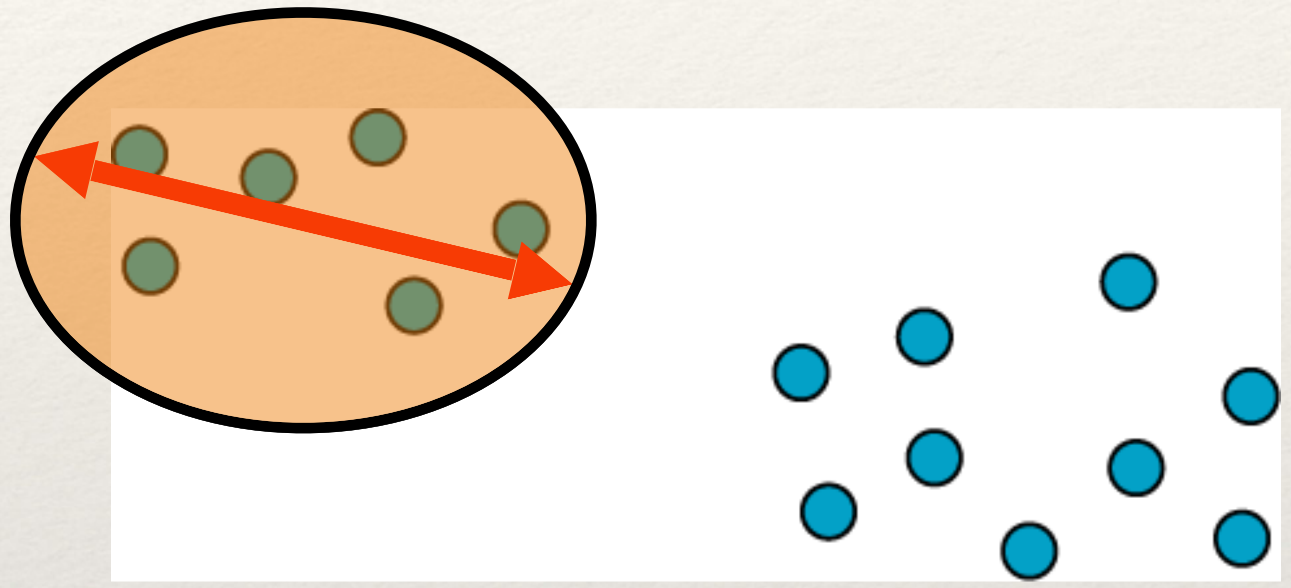
- ➢ Single linkage
- ➢ Complete linkage
- ➢ Average linkage
- ➢ Average group linkage



Intra cluster size

Several options:

For a cluster S, with N members and center C:



- Complete diameter:

$$\Delta(S) = \max_{(x,y)\in S^2}\big(d(x,y)\big)$$

- Average diameter:

$$\Delta(S) = \frac{1}{N(N-1)}\sum_{\substack{(x,y)\in S\\ x\neq y}} d(x,y)$$

- Centroid diameter:

$$\Delta(S) = \frac{2}{N}\sum_{x\in S} d(x,C)$$

Clustering Quality

For a clustering with K clusters:

1) Dunn's index

$$D = \min_{1 \le i \le K} \left(\min_{\substack{1 \le j \le K \\ j \ne i}} \left\{ \frac{\delta(S_i, S_j)}{\max\limits_{1 \le k \le K} \left(\Delta(S_k) \right)} \right\} \right)$$

-> Large values of D correspond to good clusters

2) Davies-Bouldin's index

$$DB = \frac{1}{K} \max_{i \ne j} \left(\frac{\Delta(S_i) + \Delta(S_i)}{\delta(S_i, S_j)} \right)$$

-> Low values of DB correspond to good clusters

Cluster Quality: Silhouette index

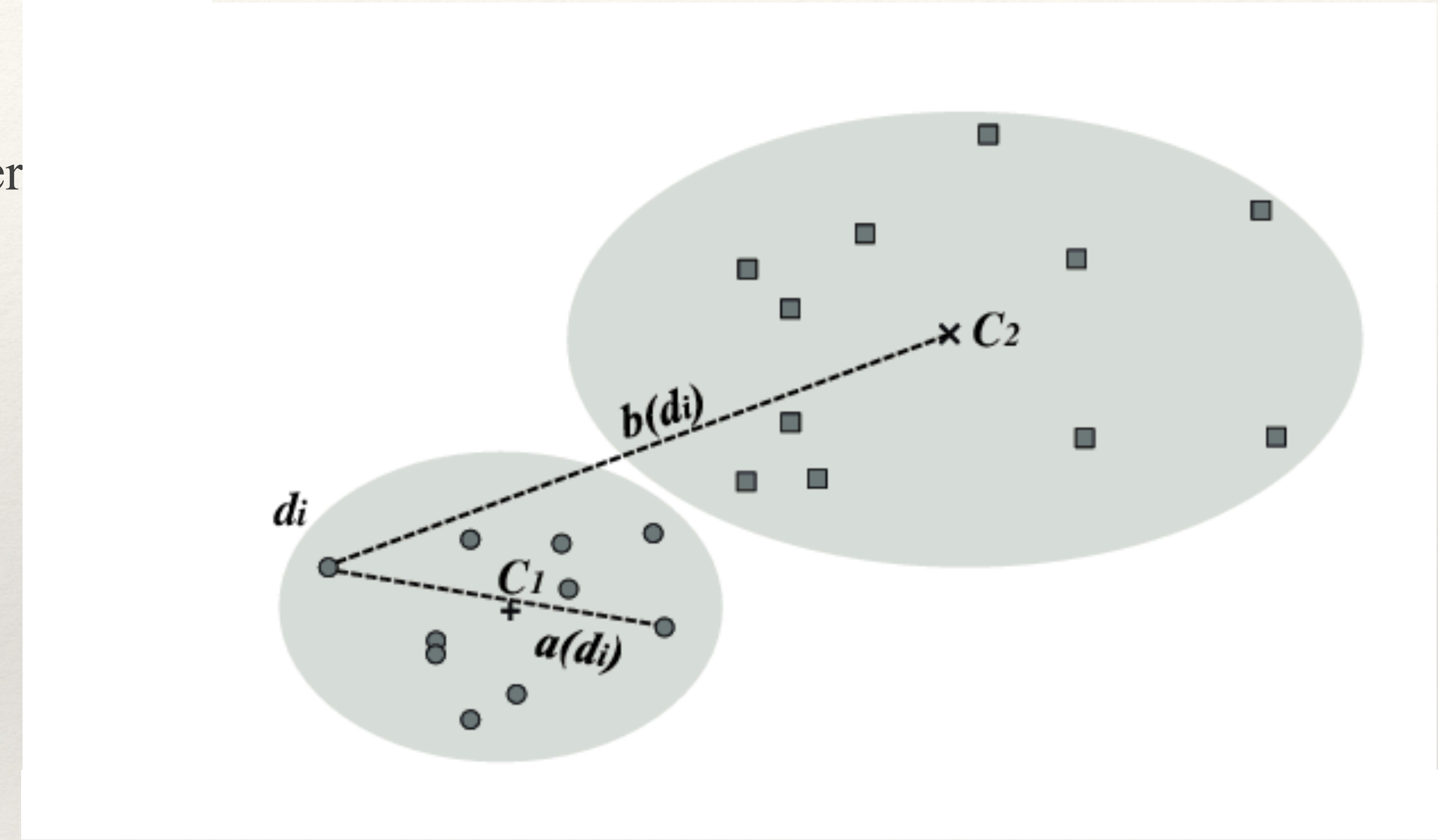
Define a quality index for each point in the original dataset:

➢For the ith object di, calculate its average distance to all other objects in its cluster. Call this value a(di).

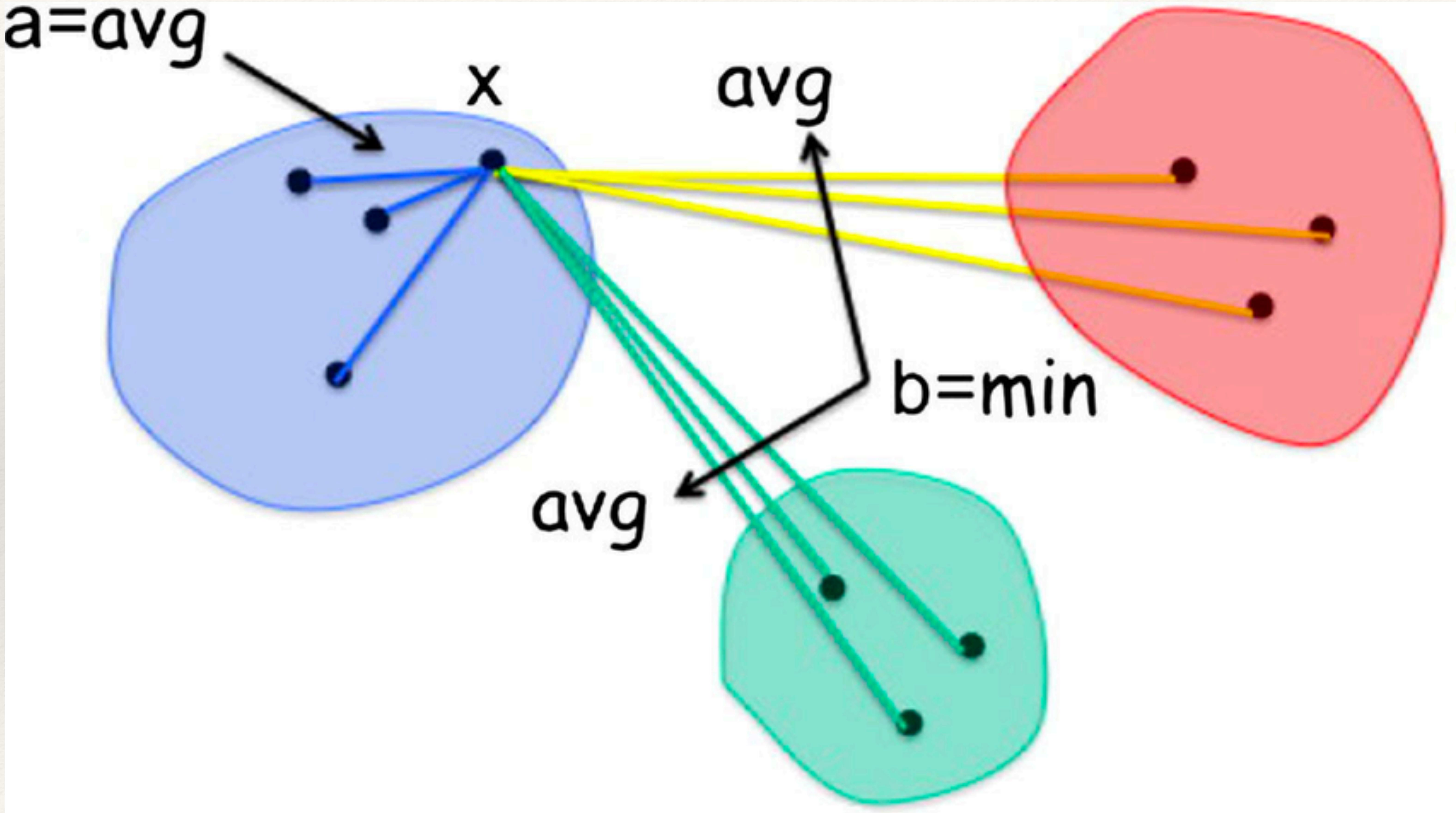
➢For the ith object and any cluster not containing the object, calculate the object's average distance to all the objects in the given cluster.
Find the minimum such value with respect to all clusters; call this value b(di).

➢For the ith object, the silhouette coefficient is

$$S(di) = \frac{b(di) - a(di)}{\max(a(di), b(di))}$$



Cluster Quality: Silhouette index



Cluster Quality: Silhouette Index

Note that:

$$-1 \leq s(di) \leq 1$$

- ➢s(i) = 1, i is likely to be well classified
- ➢s(i) = -1, i is likely to be incorrectly classified
- ➢s(i) = 0, indifferent

Cluster Quality: Silhouette Index

Cluster silhouette index:

$$S(X_i) = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

Global silhouette index:

$$GS = \frac{1}{K} \sum_{i=1}^{K} S(X_i)$$

Large values of GS correspond to good clusters

Comparing two clustering

Given a set of n elements $S = \{a_1, a_2, \ldots, a_n\}$ and two partitions of S to compare, $X = \{X_1, \ldots, X_r\}$, a partition of S into r subsets, and , $Y = \{Y_1, \ldots, Y_s\}$ a partition of S into s subsets, define the following:

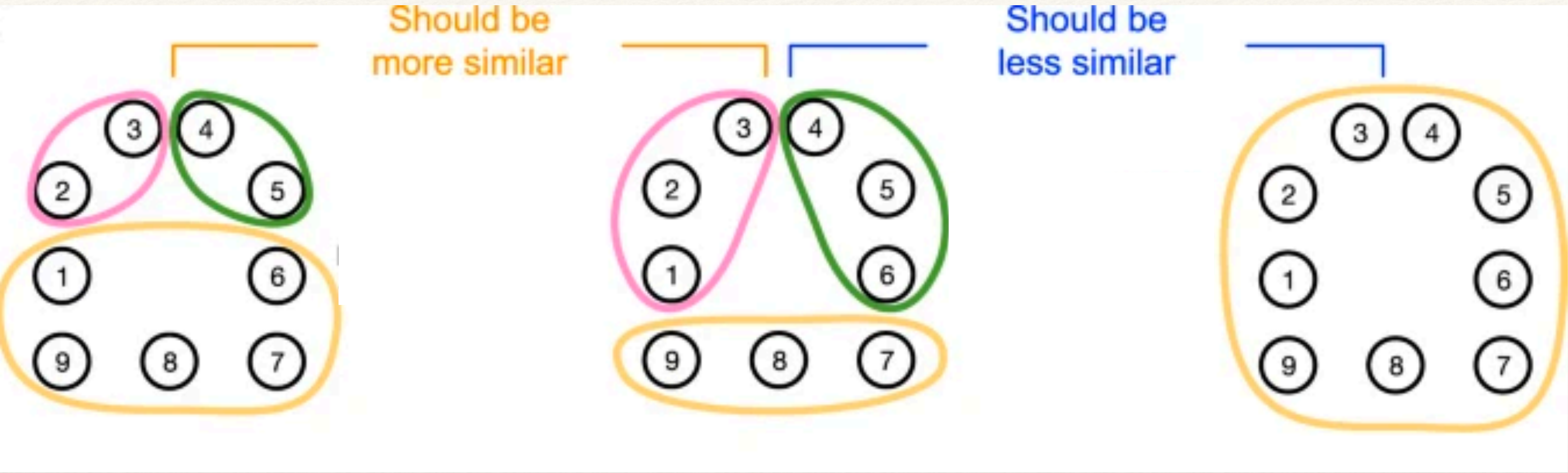
- ***a***, the number of pairs of elements in that are in the **same** subset in and in the **same** subset in S
- ***b***, the number of pairs of elements in that are in **different** subsets in and in **different** subsets in S
- ***c***, the number of pairs of elements in that are in the **same** subset in and in **different** subsets in S
- ***d***, the number of pairs of elements in that are in **different** subsets in and in the **same** subset in S

The Rand index, ***R***, is:

$$R = \frac{a+b}{a+b+c+d} = 2\frac{a+b}{n(n-1)}$$

Comparing two clustering

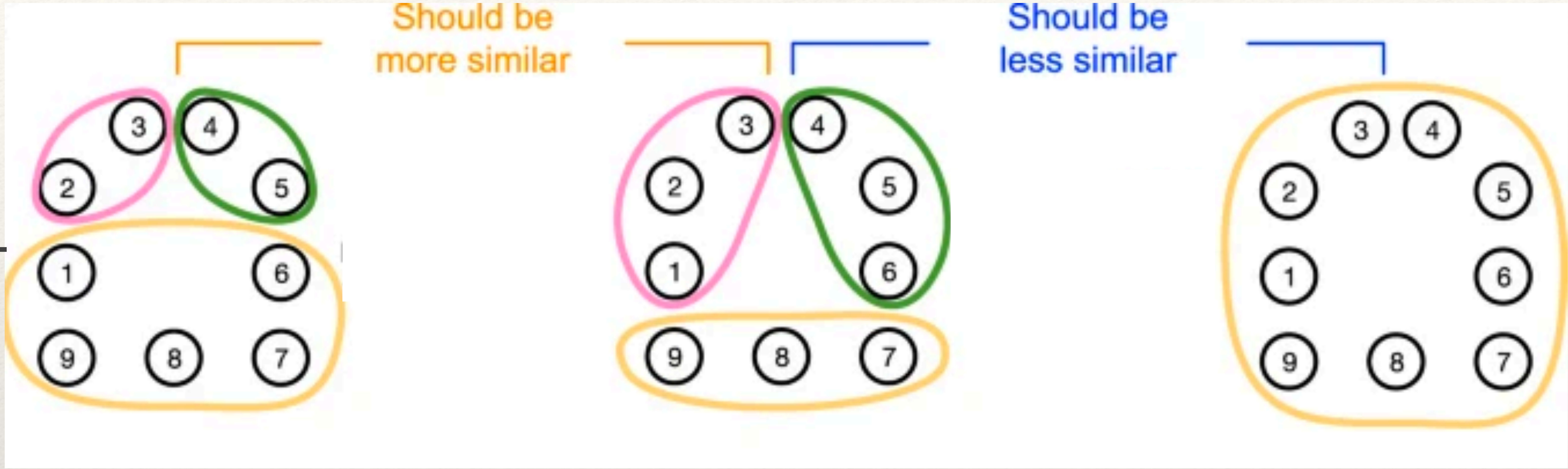
$$R = 2\frac{a+b}{n(n-1)}$$



n=9

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$

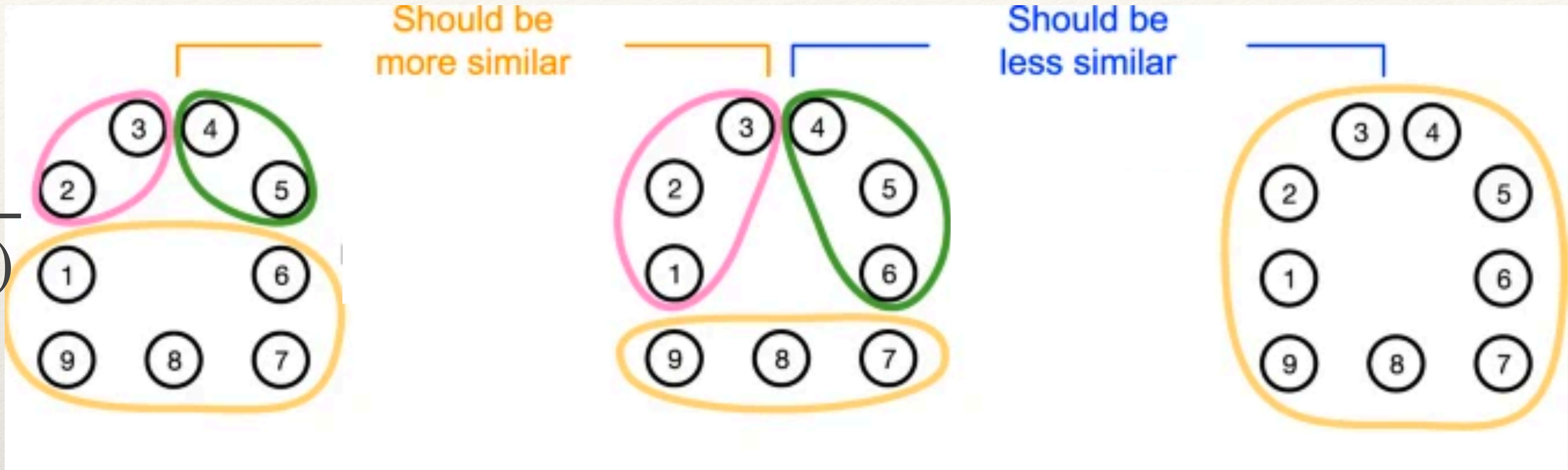


n=9

Pairs in same cluster is A:
(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



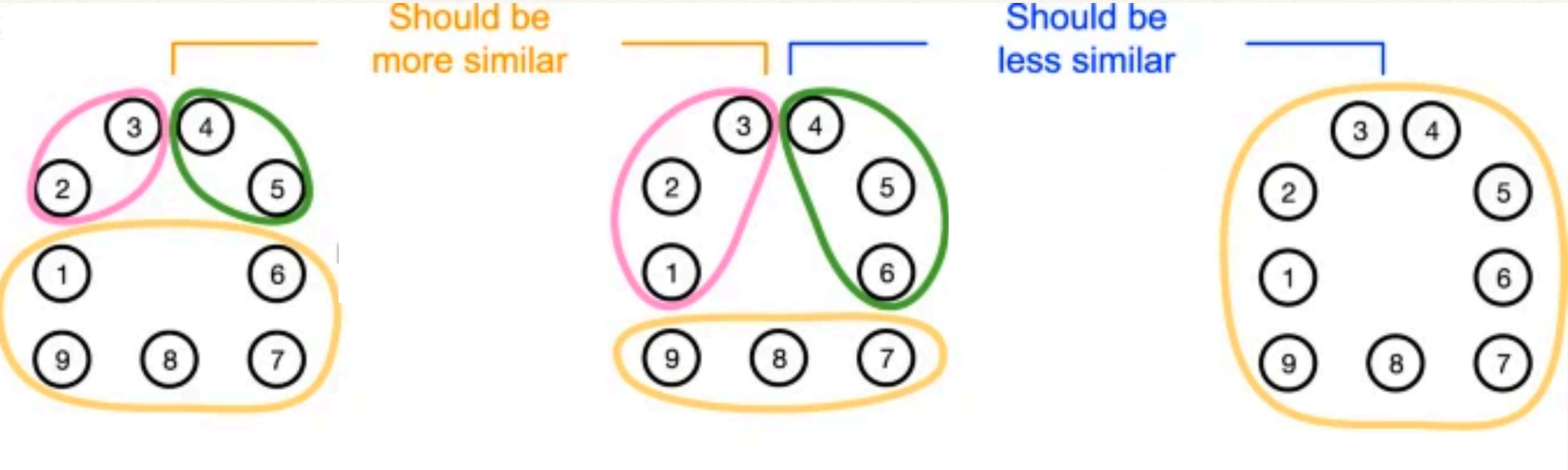
n=9

Pairs in same cluster is A and B:

(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



n=9

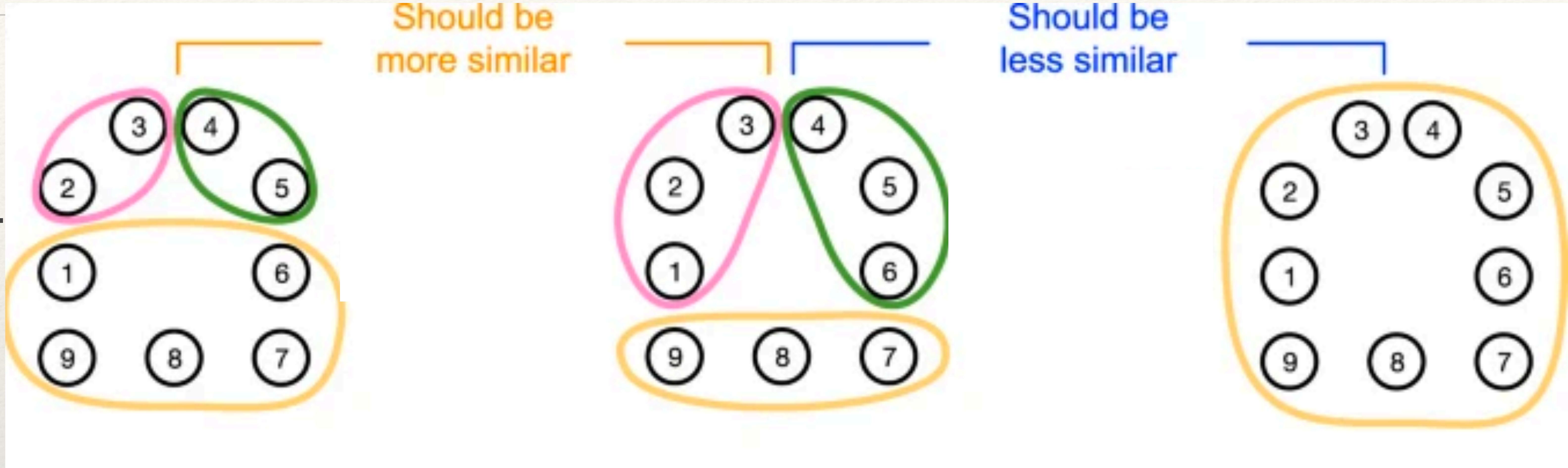
Pairs in same cluster is A and B:

(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

a =5

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



n=9

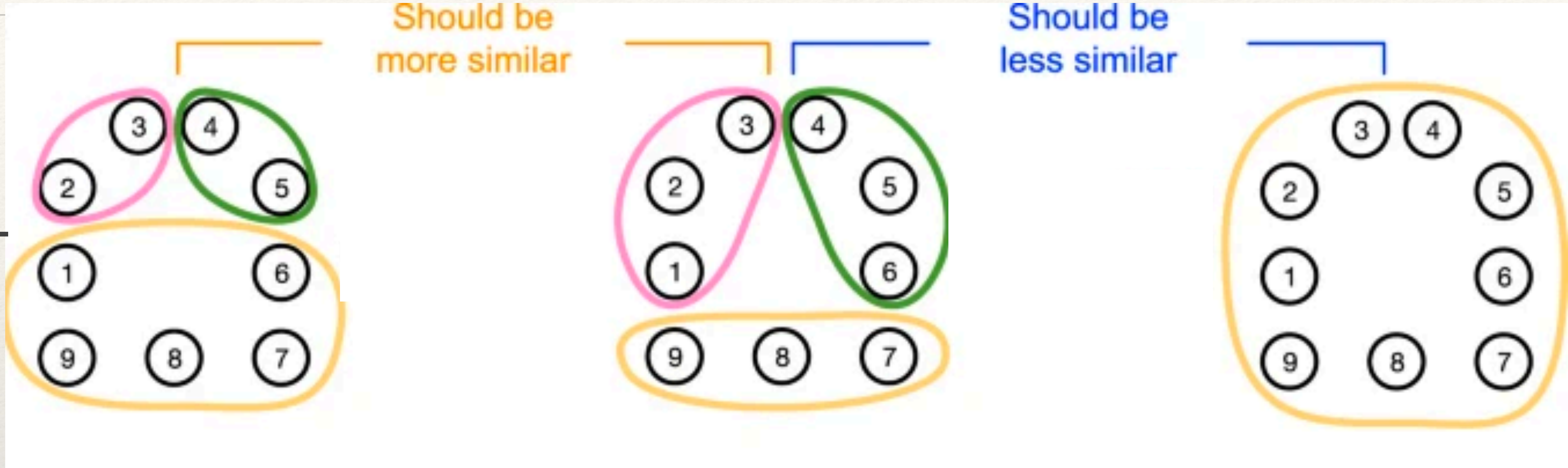
Pairs in same cluster is A and B:
(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

a =5

Pairs in different clusters in A:
(1,2) (1,3) (1,4) (1,5) (2,4) (2,5)
(2,6) (2,7) (2,8) (2,9) (3,4) (3,5)
(3,6) (3,7) (3,8) (3,9) (4,6) (4,7)
(4,8) (4,9) (5,6) (5,7) (5,8) (5,9)

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



n=9

Pairs in same cluster is A and B:

(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

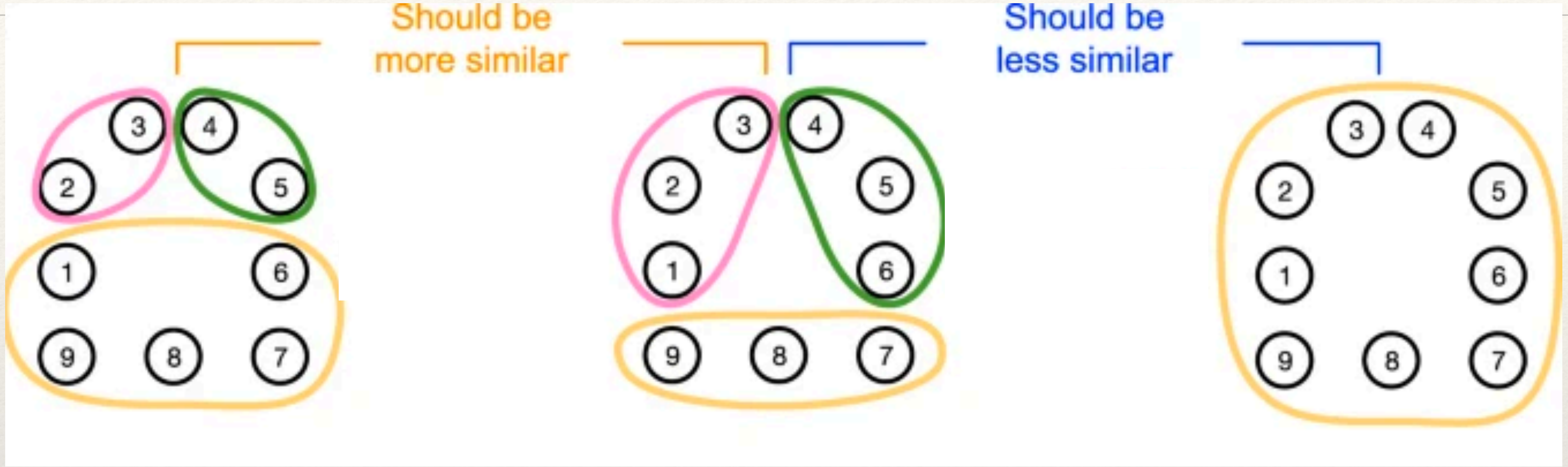
$a = 5$

Pairs in different clusters in A and B:

(1,2) (1,3) (1,4) (1,5) (2,4) (2,5)
(2,6) (2,7) (2,8) (2,9) (3,4) (3,5)
(3,6) (3,7) (3,8) (3,9) (4,6) (4,7)
(4,8) (4,9) (5,6) (5,7) (5,8) (5,9)

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



n=9

Pairs in same cluster is A and B:

(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)
(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

a =5

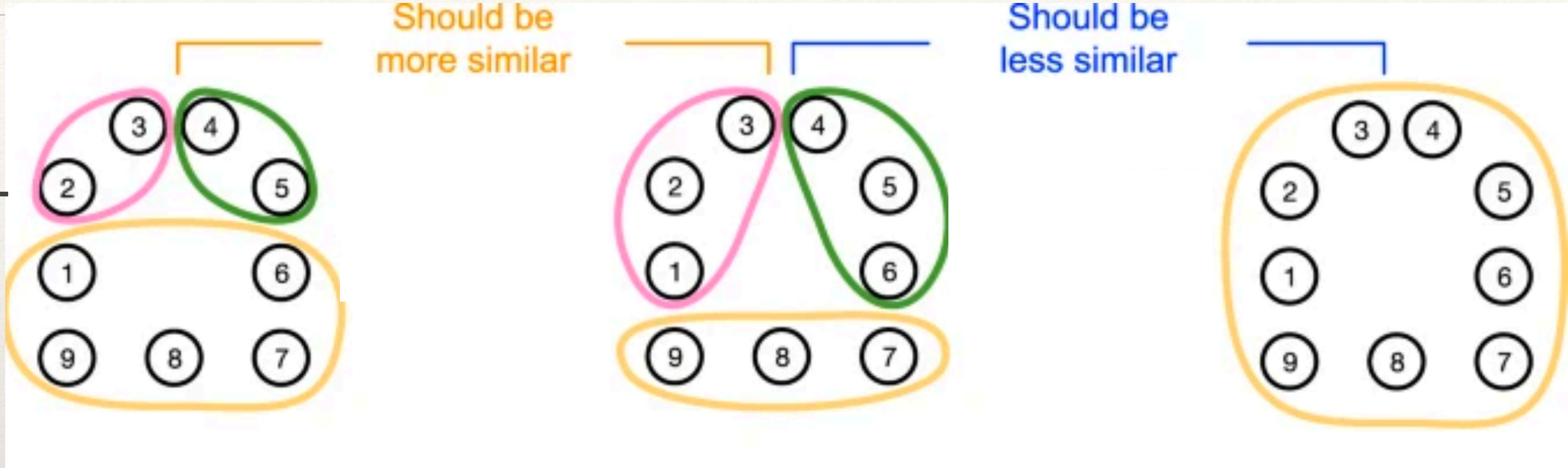
Pairs in different clusters in A and B:

(1,2) (1,3) (1,4) (1,5) (2,4) (2,5)
(2,6) (2,7) (2,8) (2,9) (3,4) (3,5)
(3,6) (3,7) (3,8) (3,9) (4,6) (4,7)
(4,8) (4,9) (5,6) (5,7) (5,8) (5,9)

b =20

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



n=9

Pairs in same cluster is A and B:

(2,3) (4,5) (1,6) (1,7) (1,8) (1,9)

(6,7) (6,8) (6,9) (7,8) (7,9) (8,9)

***a* =5**

$$\boxed{R = 2\frac{20+5}{9\times 8} = \frac{25}{36} = 0.69}$$

Pairs in different clusters in A and B:

(1,2) (1,3) (1,4) (1,5) (2,4) (2,5)

(2,6) (2,7) (2,8) (2,9) (3,4) (3,5)

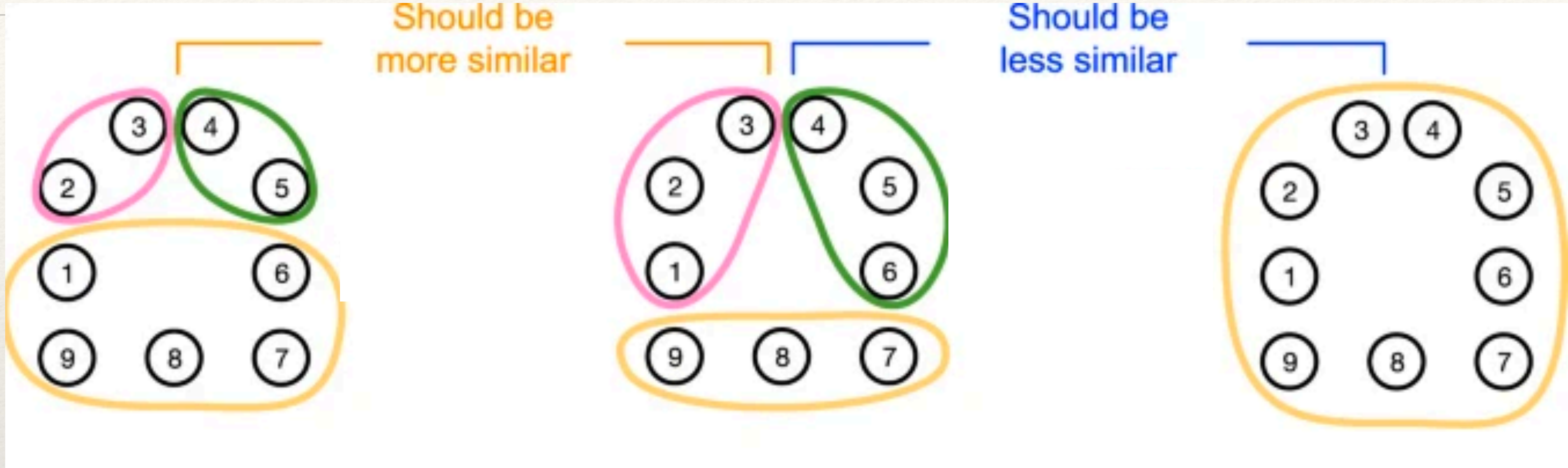
(3,6) (3,7) (3,8) (3,9) (4,6) (4,7)

(4,8) (4,9) (5,6) (5,7) (5,8) (5,9)

***b* =20**

Comparing two clustering

$$R = 2\frac{a+b}{n(n-1)}$$



$$R = 2\frac{20+5}{9\times 8} = \frac{25}{36} = 0.69$$

n=9

Pairs in same cluster is A and B:
(1,2) (1,3) (4,5) (4,6) (7,8) (7,9)
(8,9)

a =7

$$R = 2\frac{7+0}{9\times 8} = 0.19$$

Pairs in different clusters in A and B:
(1,4) (1,5) (1,6) (1,7) (1,8) (1,9) (2,4)
(2,5) (2,6) (2,7) (2,8) (2,9) (3,4) (3,5)
(3,6) (3,7) (3,8) (3,9) (4,7) (4,8) (4,9)
(5,7) (5,8) (5,9) (6,7) (5,8) (6,9)

b=0