Master Degree in Artificial Intelligence for Science and Technology

Cluster Analysis: Clustering Validation

Fabio Stella

Department of Informatics, Systems and Communication

University of Milan-Bicocca

fabio.stella@unimib.it

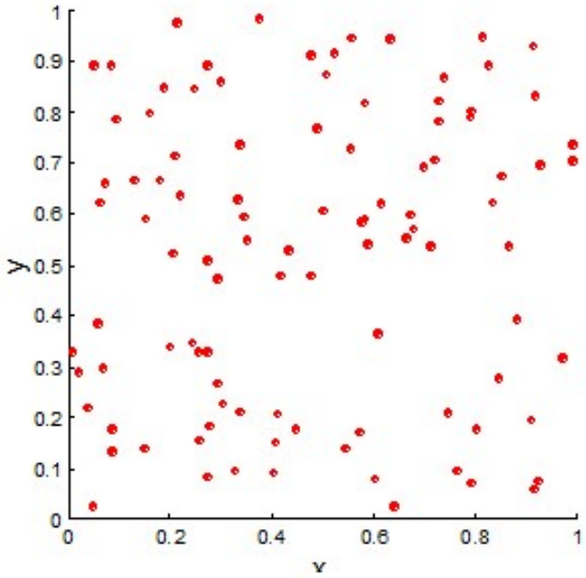
OUTLOOK

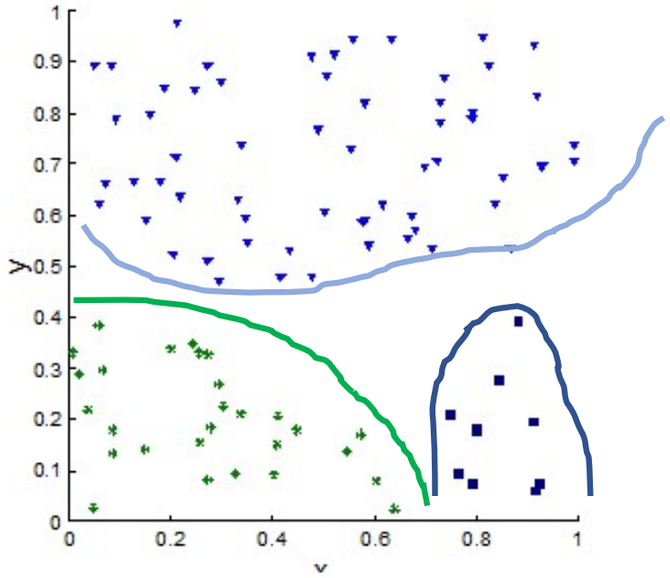
- Validation Measures
  - Supervised
  - Unsupervised
- Correlation and Visual Methods
- Optimal Number of Clusters
- Final Comment on Clustering Analysis

CLUSTERING VALIDATION

- For supervised classification we have a variety of measures to evaluate how good our model is
  - accuracy, precision, recall, AUC, …
- For cluster analysis, the analogous question is how to evaluate the "goodness" of the resulting clusters?
- But "clusters are in the eye of the beholder"!
  - in practice the clusters we find are defined by the clustering algorithm
- Then why do we want to evaluate them?
  - to avoid finding patterns in noise
  - to compare clustering algorithms
  - to compare two sets of clusters
  - to compare two clusters

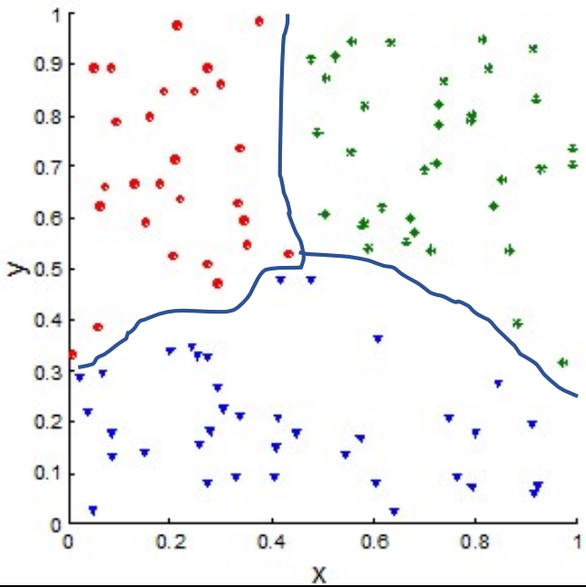
Random Points

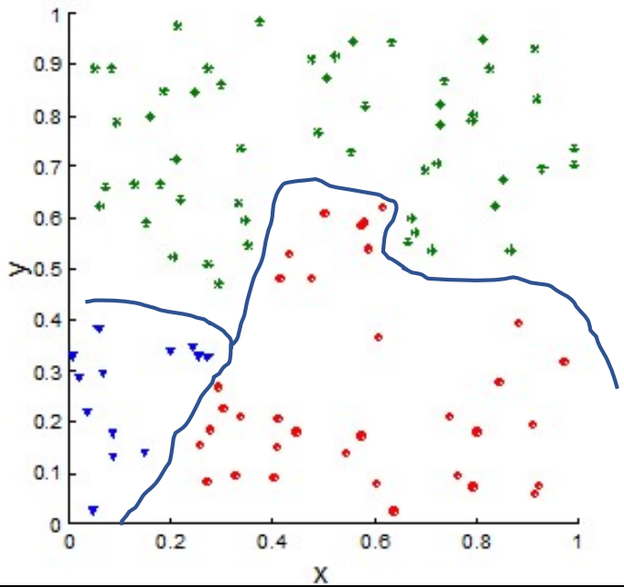




DBSCAN

K-means





Complete Link

CLUSTERS FOUND IN RANDOM DATA

MEASURES OF CLUSTER VALIDITY

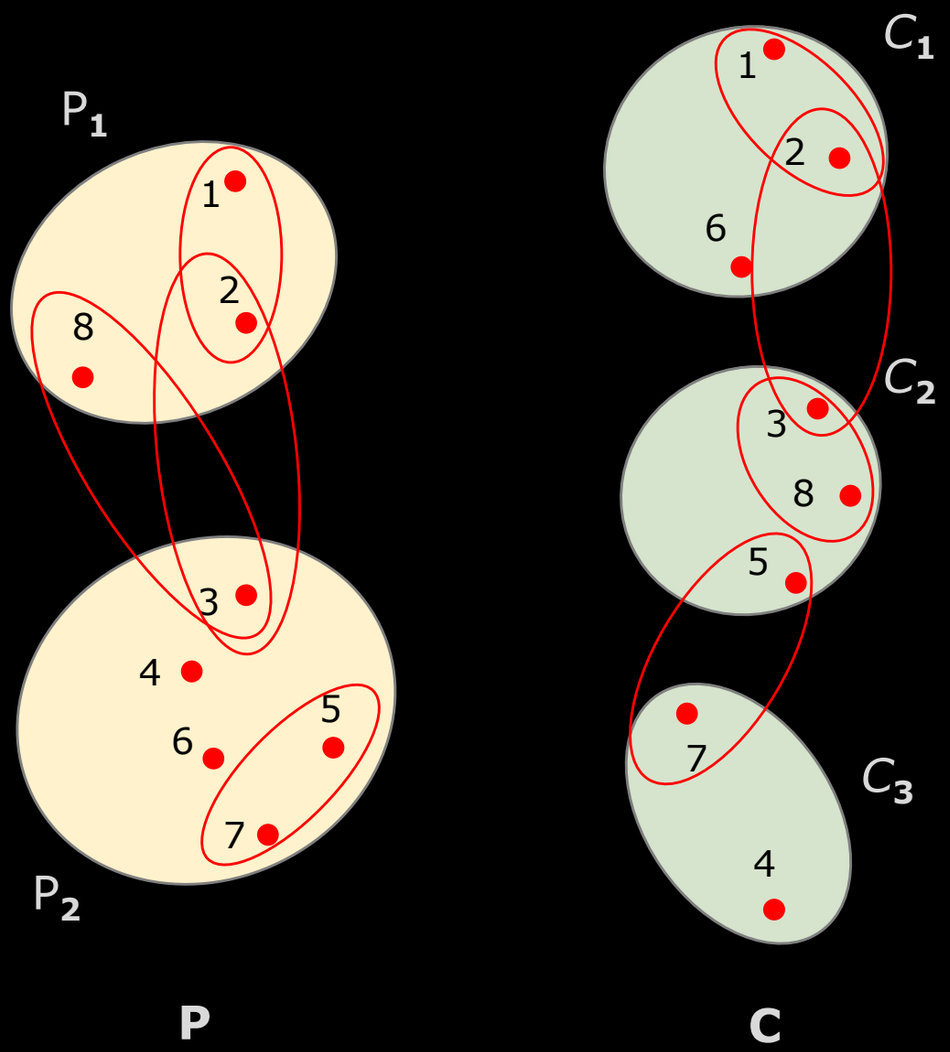
- Numerical measures that are applied to judge various aspects of cluster validity, are classified into the following two types.

  — **SUPERVISED**: used to measure the extent to which cluster labels match externally supplied class labels.
    - entropy
    - often called external indices because they use information external to the data

  — **UNSUPERVISED**: used to measure the goodness of a clustering structure without respect to external information.
    - sum of squared error (SSE)
    - often called internal indices because they only use information in the data

- You can use supervised or unsupervised measures to compare clusters or clusterings

SUPERVISED MEASURES



Let

$$P = \{P_1,...,P_R\}$$

be a partitioning, of a data set consisting of "*m*" objects (records), into "*R*" categories.

$$C = \{C_1,...,C_K\}$$

be the partition obtained with a clustering algorithm into "*K*" clusters.

SUPERVISED OR EXTERNAL INDICES compare **P** to **C**:

3 (**a**) **Case 1: x** and **y** belong to the same cluster of **C** and to the same category of **P**

4 (**b**) **Case 2: x** and **y** belong to the same cluster of **C** but to different categories of **P**

10 (**c**) **Case 3: x** and **y** belong to different clusters of **C** but to the same category of **P**

11 (**d**) **Case 4: x** and **y** belong to different clusters of **C** and to different categories of **P**

SUPERVISED MEASURES

The overall number of pairs amounts to

$$M = \frac{m \times (m-1)}{2} = a + b + c + d$$

RAND $\mathrm{R} = \frac{\mathrm{a} + \mathrm{d}}{\mathrm{M}}$ ***R*** **∈ [0,1]**

JACCARD $\mathrm{J} = \frac{\mathrm{a}}{\mathrm{a} + \mathrm{b} + \mathrm{c}}$ ***J*** **∈ [0,1]**

FOWLKES AND MALLOWS $\mathrm{FM} = \sqrt{\frac{\mathrm{a}}{\mathrm{a} + \mathrm{b}} \times \frac{\mathrm{a}}{\mathrm{a} + \mathrm{c}}}$ ***FM*** **∈ [0,1]**

Γ STATISTICS $\Gamma = \frac{M \times a - (a+b) \times (a+c)}{\sqrt{(a+b) \times (a+c)(M-a-b) \times (M-a-c)}}$ Γ **∈ [-1,1]**

The larger the values, the more similar are **C** and **P**.

Let

$$P = \{\mathrm{P}_1,\ldots,\mathrm{P}_\mathrm{R}\}$$

be a partitioning, of a data set consisting of "*m*" objects (records), into "*R*" categories.

$$C = \{\mathrm{C}_1,\ldots,\mathrm{C}_\mathrm{K}\}$$

be the partition obtained with a clustering algorithm into "*K*" clusters.

SUPERVISED OR EXTERNAL INDICES compare **P** to **C**:

3 (**a**) **Case 1: x** and **y** belong to the same cluster of **C** and to the same category of **P**

4 (**b**) **Case 2: x** and **y** belong to the same cluster of **C** but to different categories of **P**

10 (**c**) **Case 3: x** and **y** belong to different clusters of **C** but to the same category of **P**

11 (**d**) **Case 4: x** and **y** belong to different clusters of **C** and to different categories of **P**

UNSUPERVISED MEASURES

- Cluster COHESION: measures how closely related are objects x in a cluster.
  - Example: SSE
- Cluster SEPARATION: measures how distinct or well-separated a cluster is from other clusters
- Example: squared error
  - COHESION is measured by the within cluster sum of squares $$SSE = \sum_i \sum_{x \in C_i} (x - m_i)^2$$
  - SEPARATION is measured by the between cluster sum of squares $$SSB = \sum_i |C_i|(m - m_i)^2$$

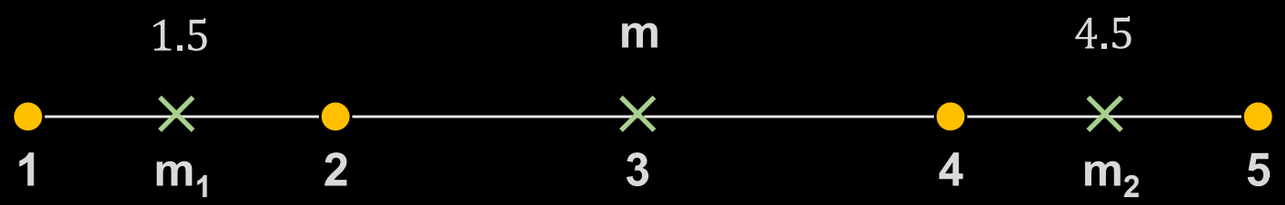
where $|C_i|$ is the size of cluster i, m_i is the centroid of cluster C_i, x is an object, and m is the overall centroid.

UNSUPERVISED MEASURES: COHESION AND SEPARATION

- Example: SSE
  - SSB + SSE = constant

$$SSE = \sum_i \sum_{x \in C_i} (x - m_i)^2$$

$$SSB = \sum_i |C_i| (m - m_i)^2$$



K=1 cluster:

$$SSE = (1-3)^2 + (2-3)^2 + (4-3)^2 + (5-3)^2 = 10$$

$$SSB = 4(3-3)^2 = 0$$

$$Total = 10 + 0 = 10$$

K=2 clusters:

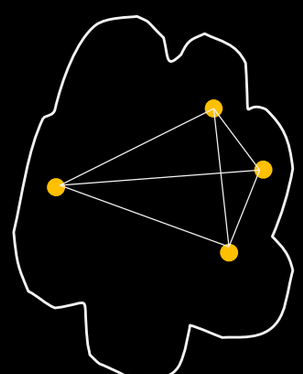
$$SSE = (1-1.5)^2 + (2-1.5)^2 + (4-4.5)^2 + (5-4.5)^2 = 1$$

$$SSB = 2\,(3-1.5)^2 + 2(4.5-3)^2 = 9$$

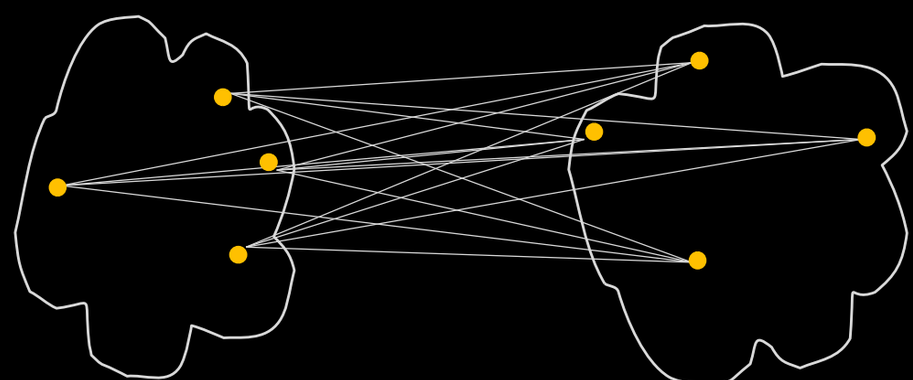
$$Total = 1 + 9 = 10$$

UNSUPERVISED MEASURES: COHESION AND SEPARATION

- A proximity graph-based approach can also be used for cohesion and separation.
  - cluster COHESION is the sum of the weight of all links within a cluster.
  - cluster SEPARATION is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



separation

UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- Silhouette coefficient combines ideas of both cohesion and separation, but for individual points, as well as clusters and clusterings

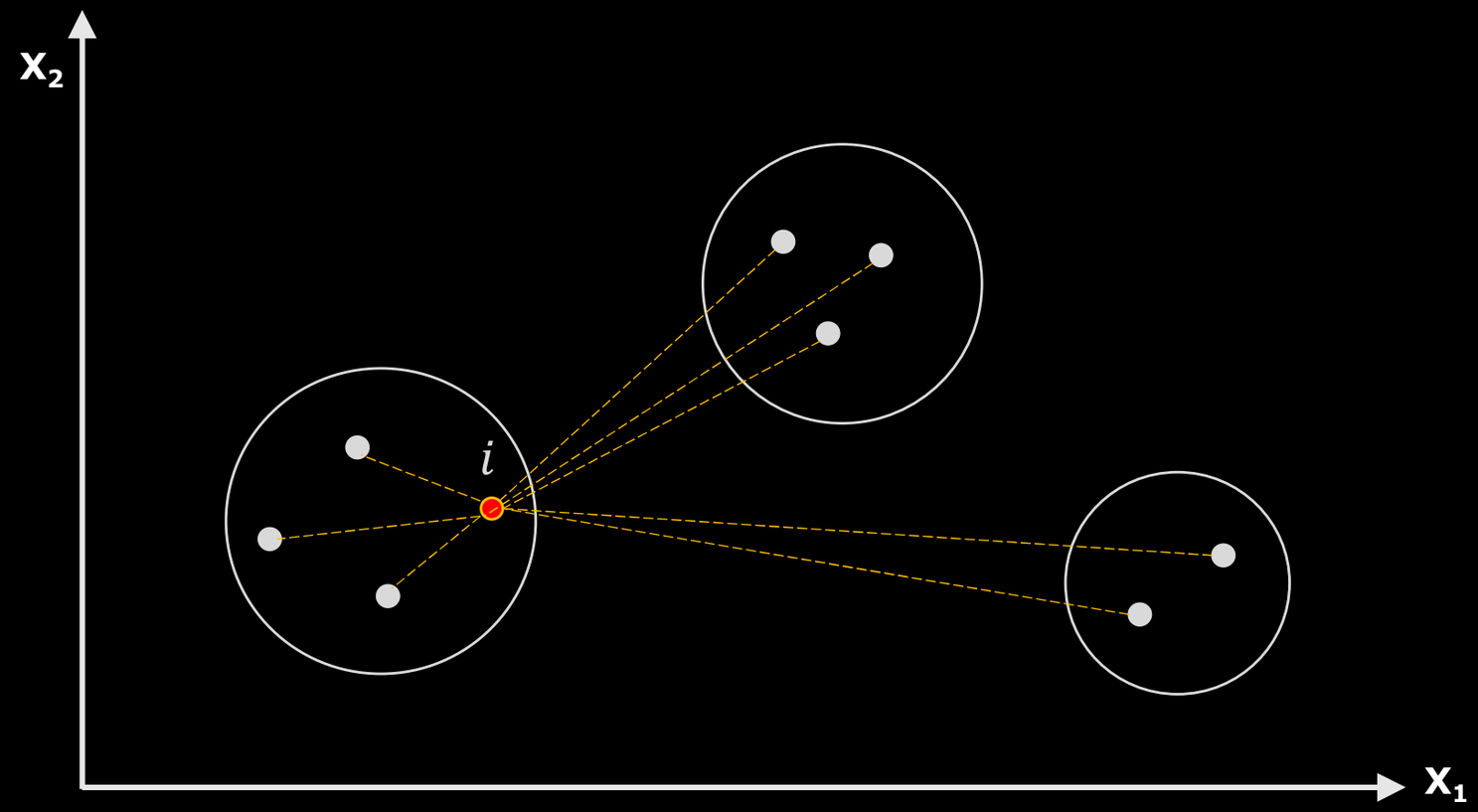
- For an individual point i
  - calculate a = average distance of point i to the points in its cluster
  - calculate b = minimum of the average distances of i to all points in any other cluster
  - the silhouette coefficient for a point i is then given by $s_i = \frac{b-a}{\max(a,b)}$
  - value can vary between -1 and 1
  - typically ranges between 0 and 1.
  - the closer to 1 the better.

- Can calculate the average silhouette coefficient for a cluster or a clustering

UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- calculate a = average distance of point i to the points in its cluster
- calculate b = minimum of the average distances of i to all points in any other cluster

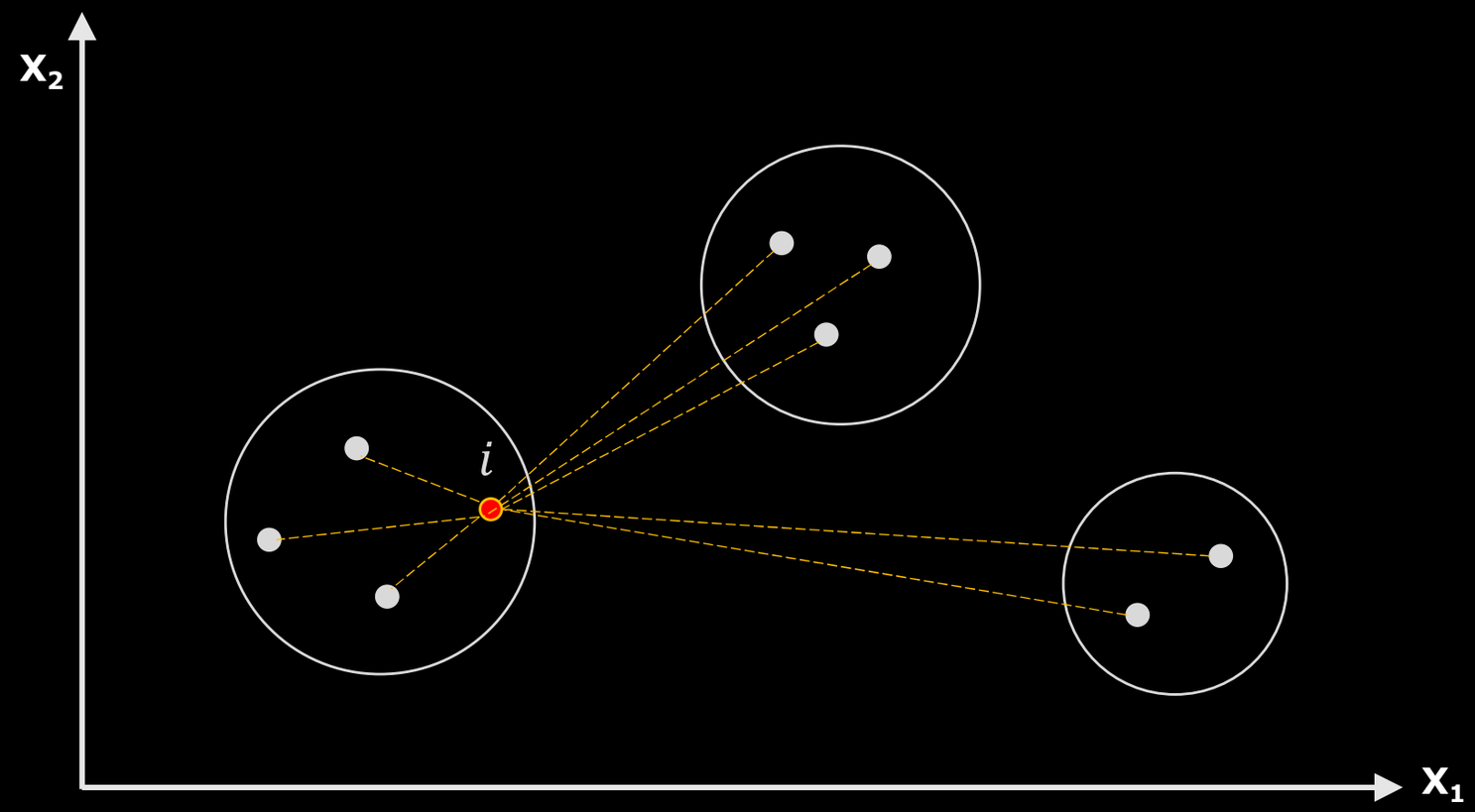
$$s_i = \frac{b - a}{\max(a, b)}$$



UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

– negative Silhouette Coefficient means that the average distance to points in its cluster (a) is greater than the minimum average distance to points in another cluster (b).

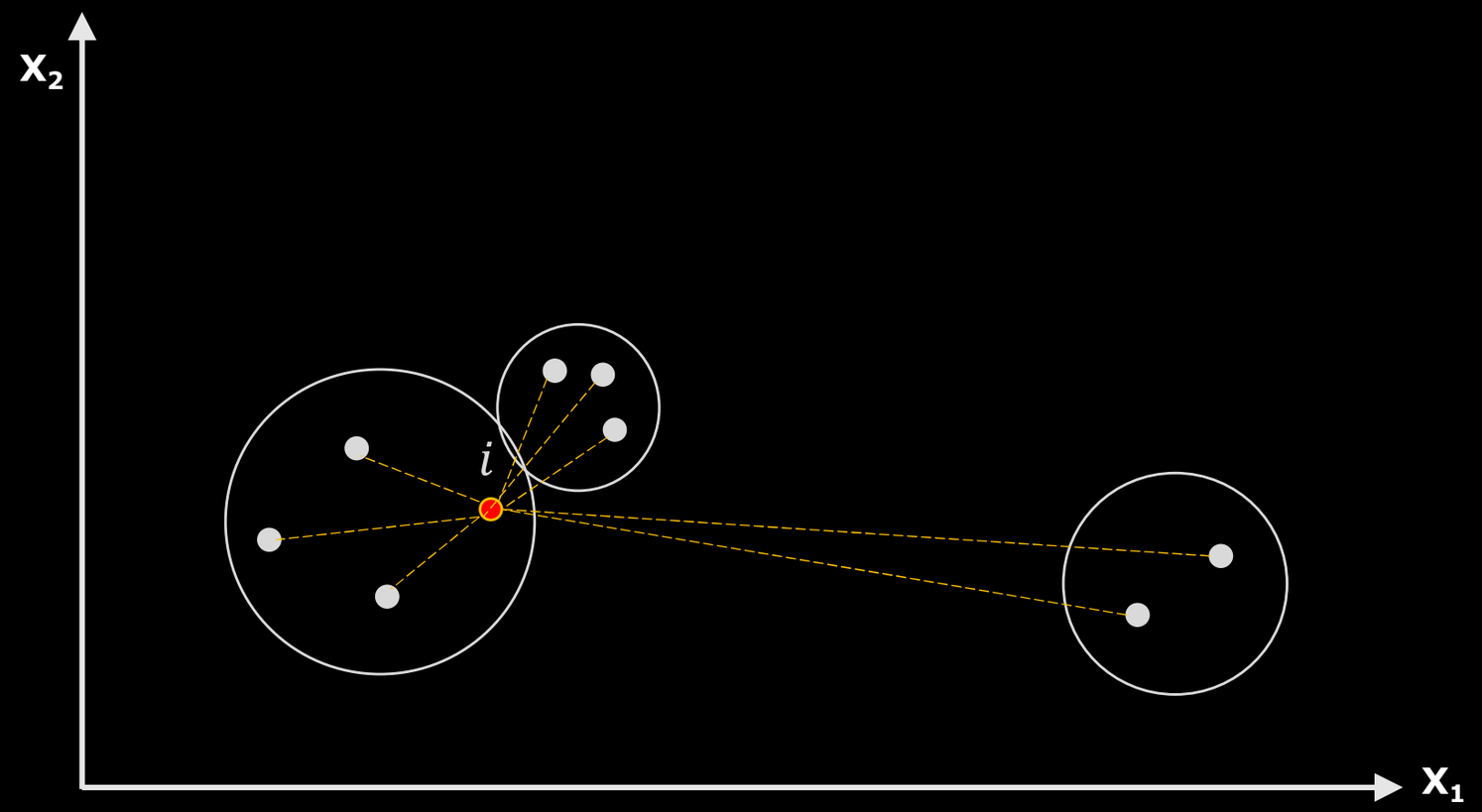
$$s_i = \frac{b - a}{\max(a, b)}$$



UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

– negative Silhouette Coefficient means that the average distance to points in its cluster (a) is greater than the minimum average distance to points in another cluster (b).

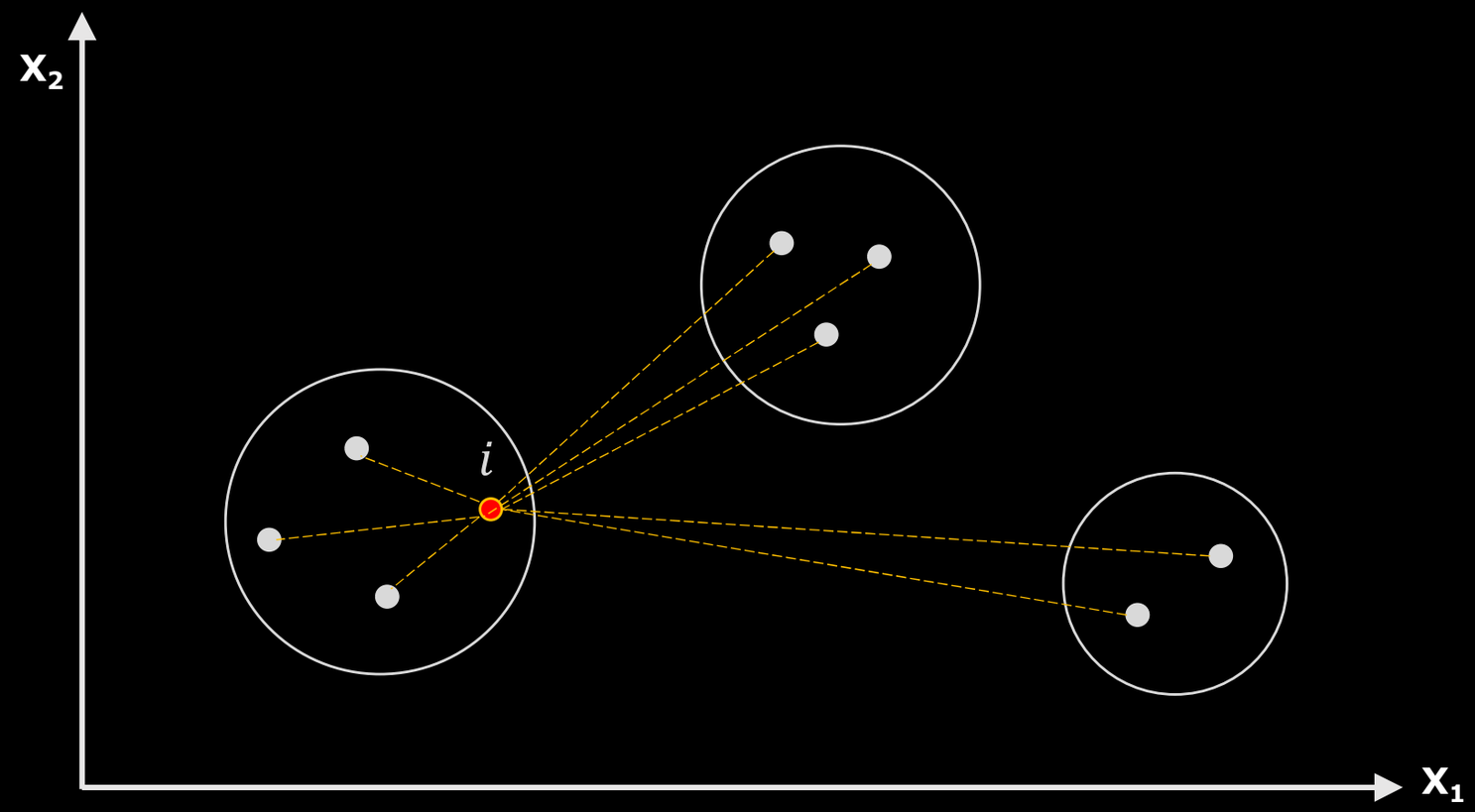
$$s_i = \frac{b - a}{\max(a, b)}$$



UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- negative Silhouette Coefficient means that the average distance to points in its cluster (a) is greater than the minimum average distance to points in another cluster (b).
- we want that the Silhouette Coefficient is positive ($a < b$), and for a to be as close to 0 as possible, since the Silhouette Coefficient assumes its maximum value of 1 when $a = 0$.

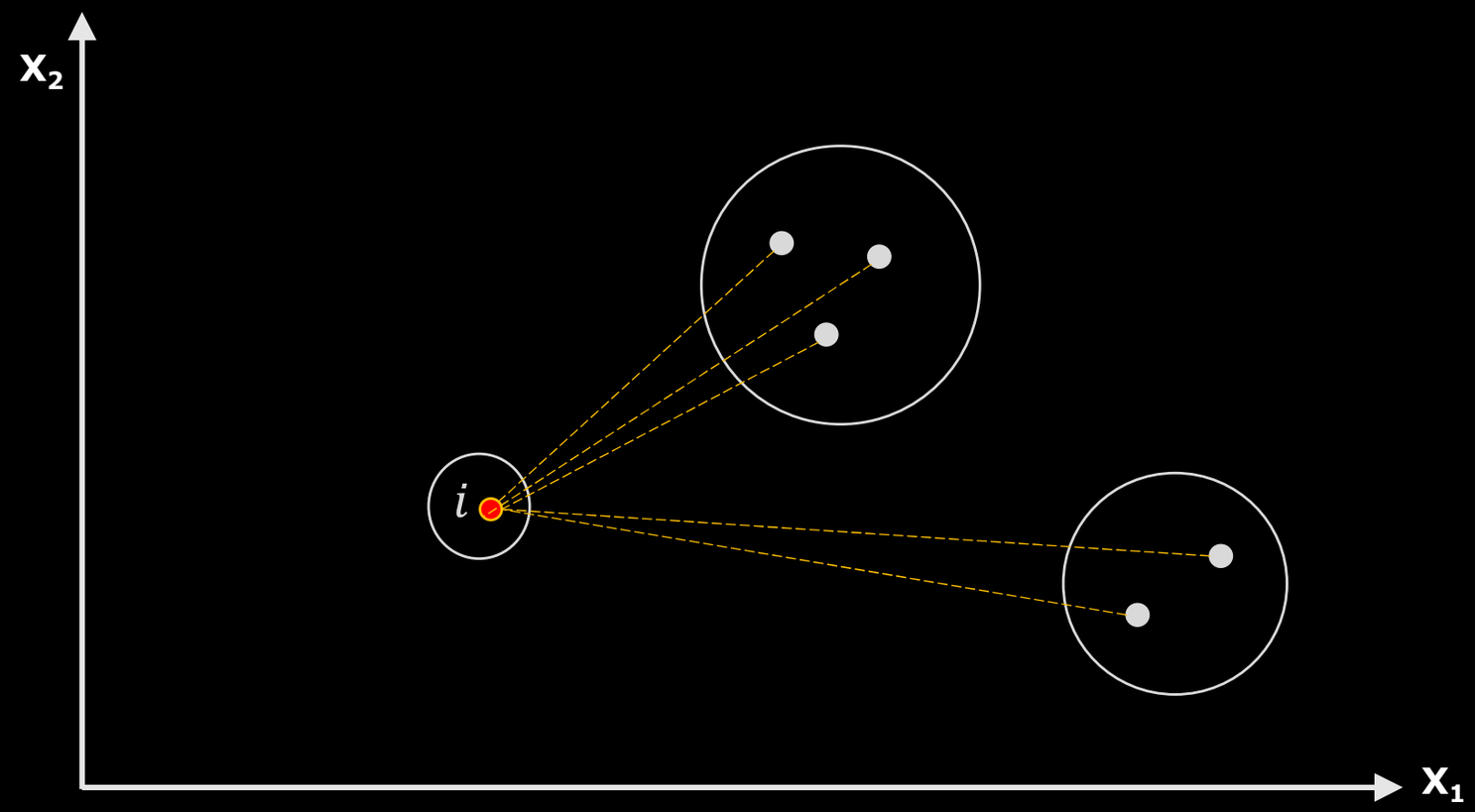
$$s_i = \frac{b - a}{\max(a, b)}$$



UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- negative Silhouette Coefficient means that the average distance to points in its cluster (a) is greater than the minimum average distance to points in another cluster (b).
- we want that the Silhouette Coefficient is positive ($a < b$), and for a to be as close to 0 as possible, since the Silhouette Coefficient assumes its maximum value of 1 when $a = 0$.

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

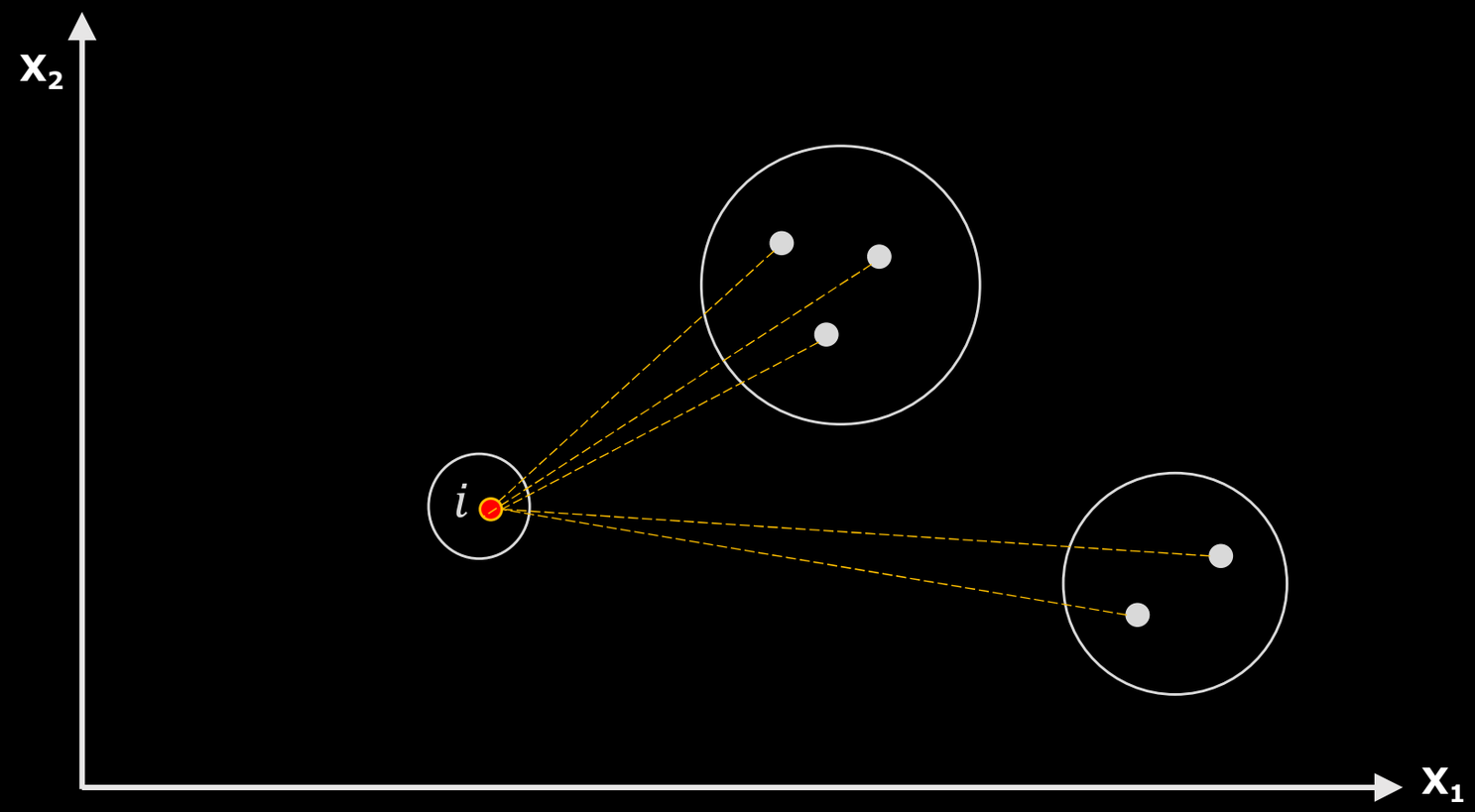


UNSUPERVISED MEASURES: SILHOUETTE COEFFICIENT

- we can compute the Average Silhouette Coefficient of a cluster by simply taking the average of the Silhouette Coefficients of data points (records) belonging to the considered cluster.

- an overall measure of goodness of a clustering can be obtained by computing the Average Silhouette Coefficient of all points.

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

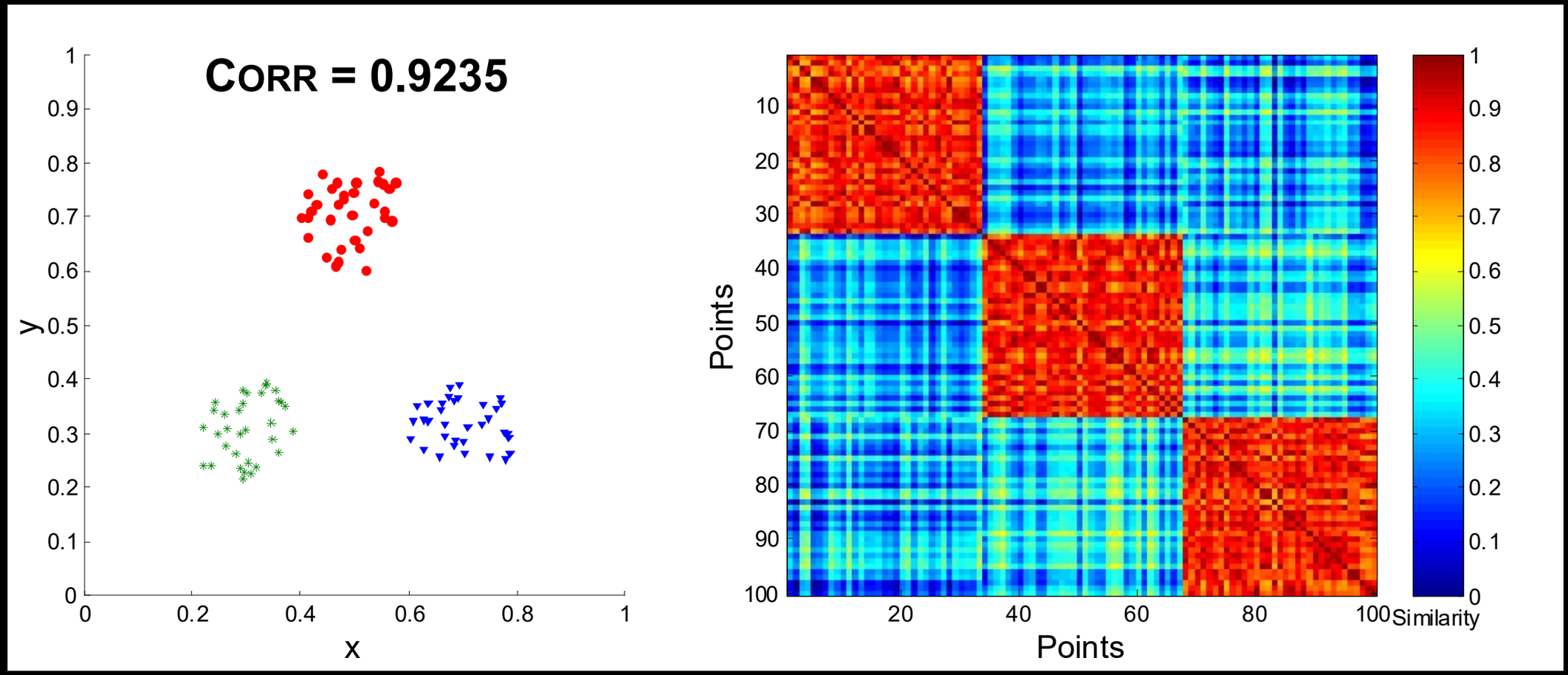


UNSUPERVISED MEASURES: CORRELATION

- Two matrices
  - proximity matrix
  - ideal similarity matrix
    - one row and one column for each data point
    - an entry is 1 if the associated pair of points belong to the same cluster
    - an entry is 0 if the associated pair of points belongs to different clusters
- Compute the correlation between the two matrices
  - since the matrices are symmetric, only the correlation between $\frac{n(n-1)}{2}$ entries needs to be calculated.
- High magnitude of correlation indicates that points that belong to the same cluster are close to each other.
  - correlation may be positive or negative depending on whether the similarity matrix is a similarity or dissimilarity matrix
- Not a good measure for some density or contiguity based clusters.

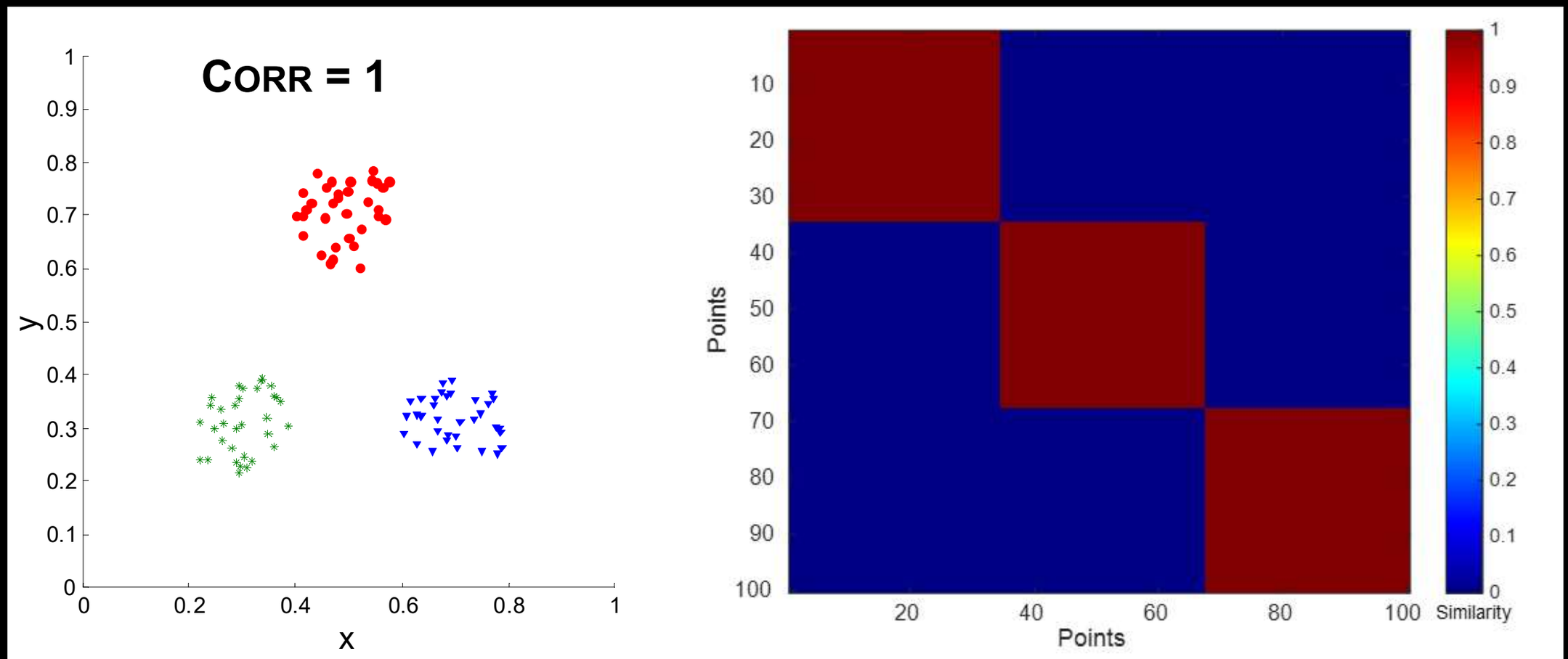
UNSUPERVISED MEASURES: CORRELATION

- correlation of ideal similarity and proximity matrices for the K-means clusterings of the following well-clustered data set.



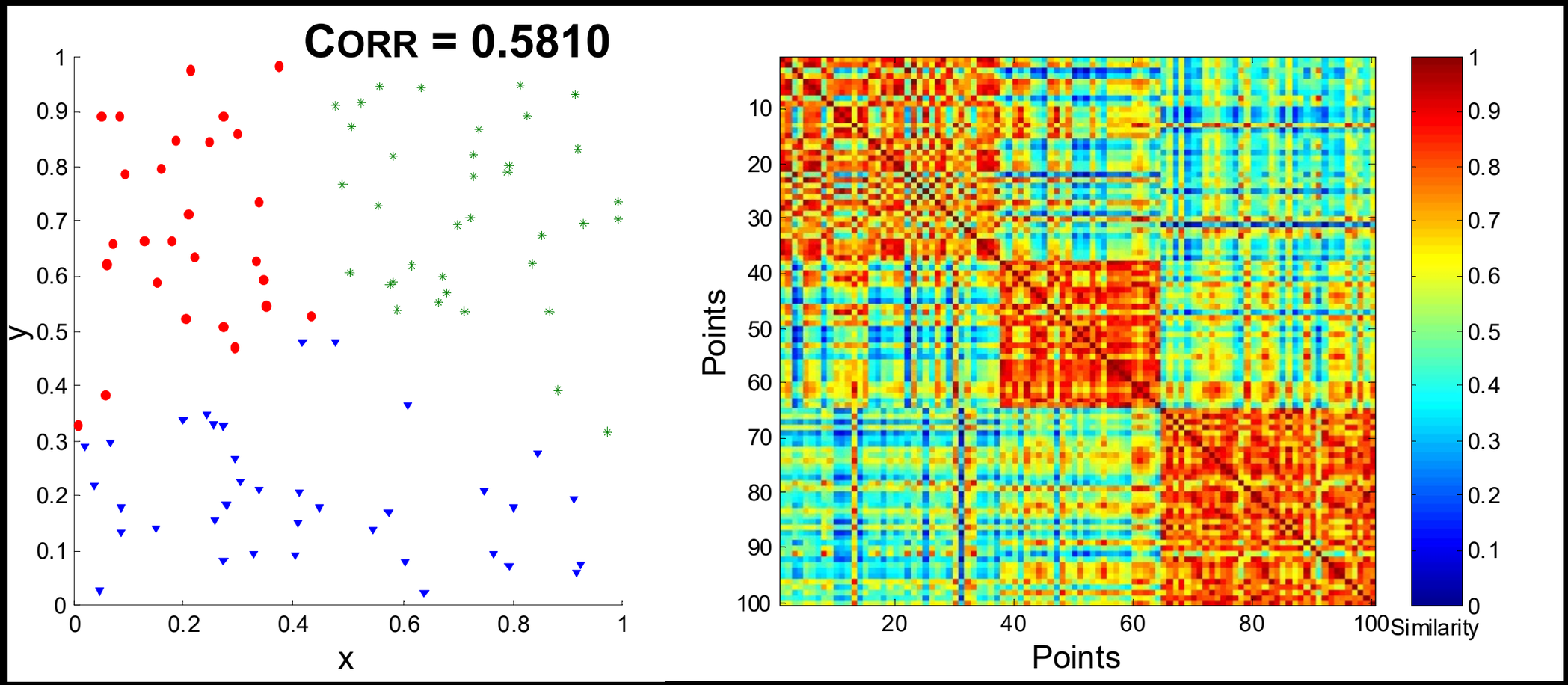
UNSUPERVISED MEASURES: CORRELATION

- correlation of ideal similarity and proximity matrices for the K-means clusterings of the following well-clustered data set.



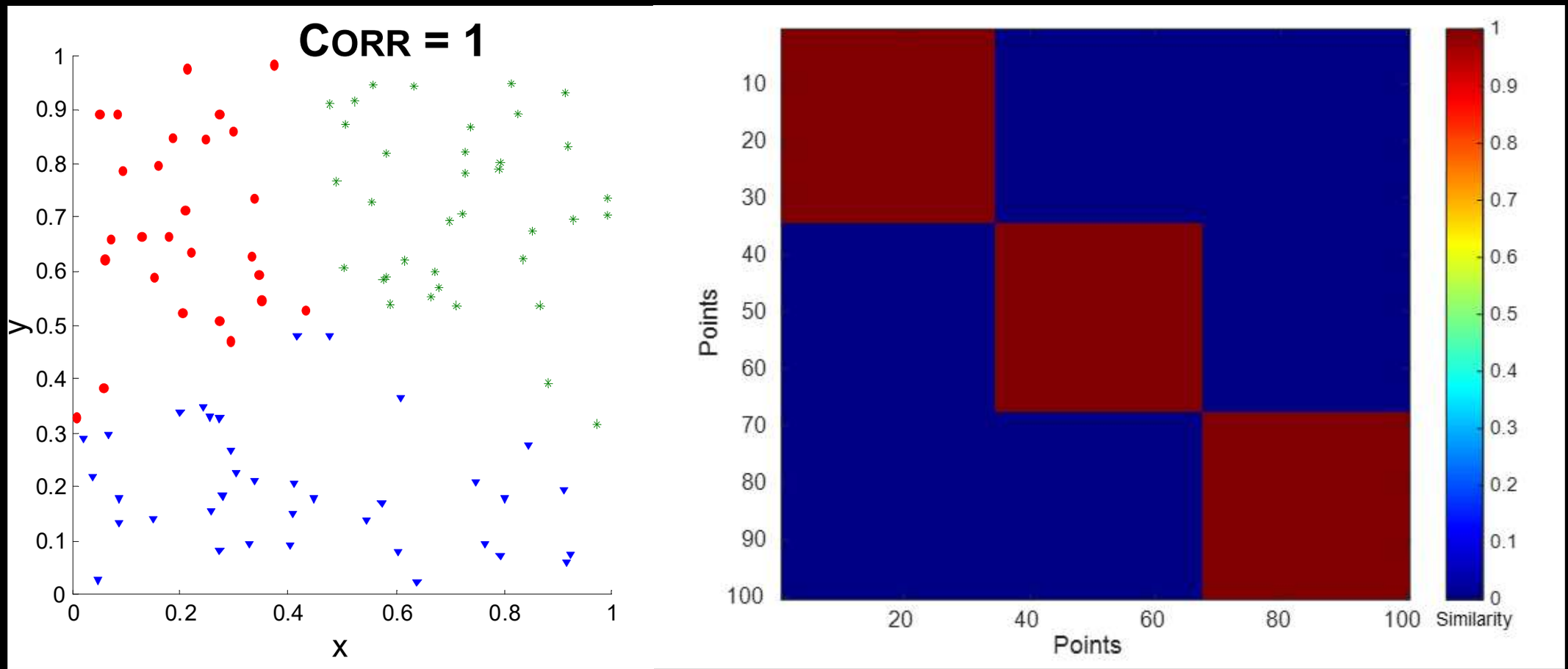
UNSUPERVISED MEASURES: CORRELATION

- correlation of ideal similarity and proximity matrices for the K-means clusterings of the following random data set.



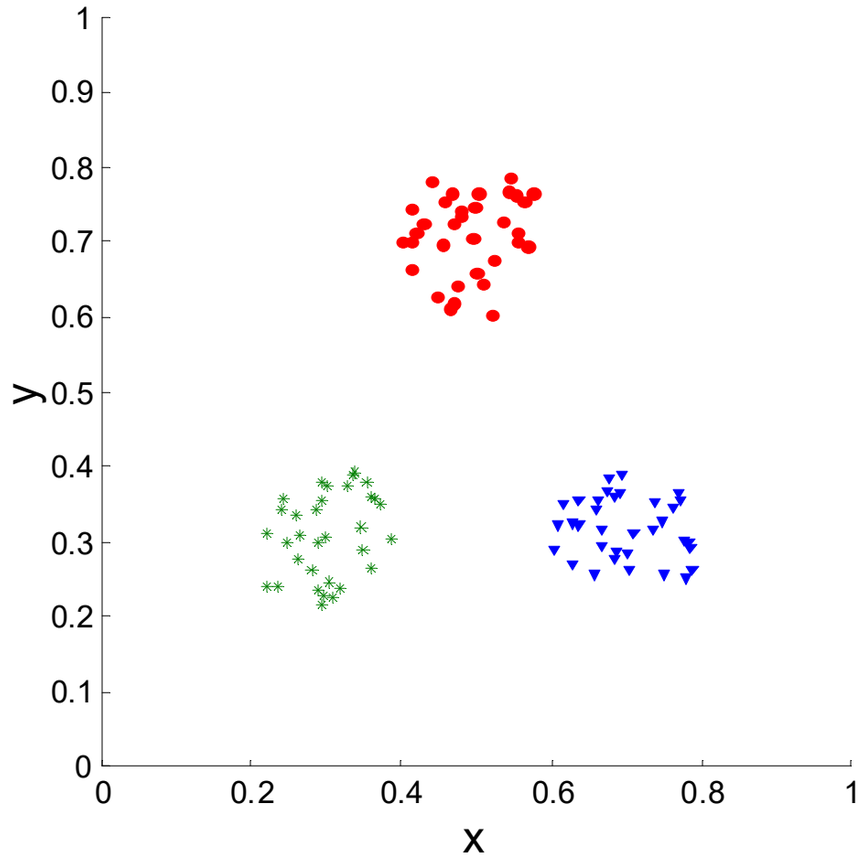
UNSUPERVISED MEASURES: CORRELATION

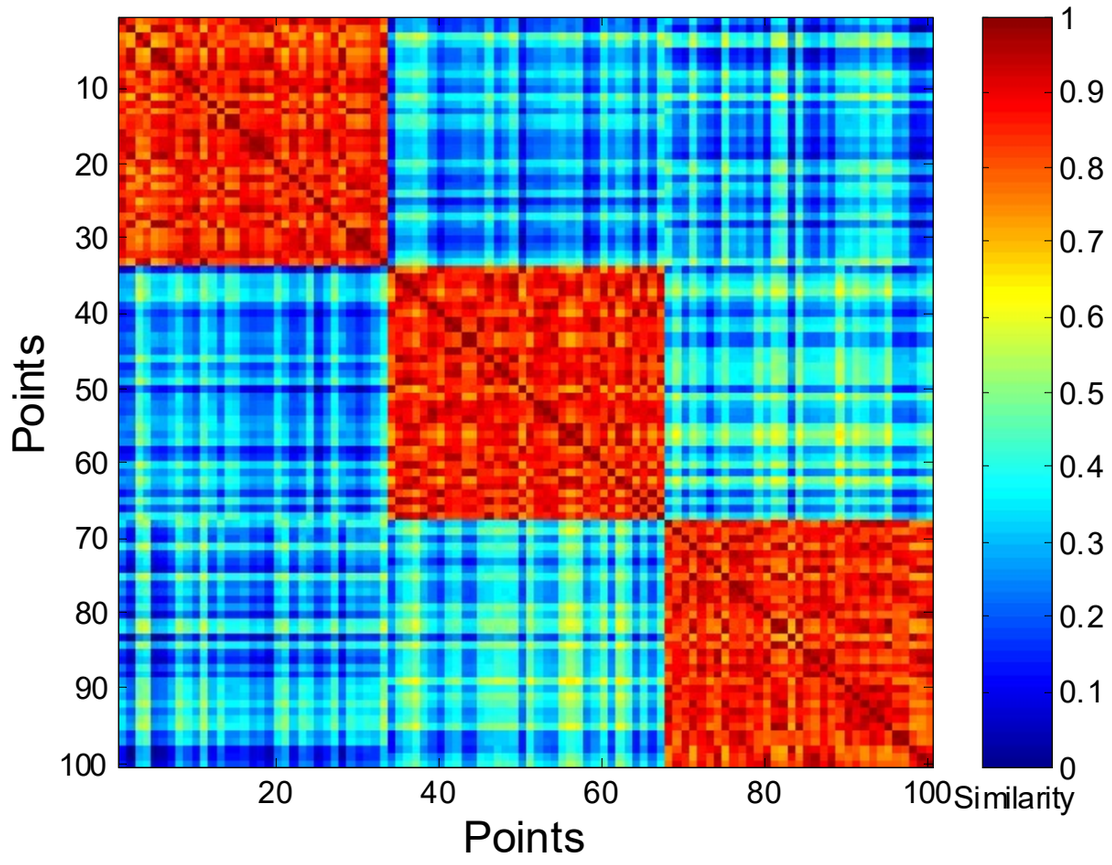
- correlation of ideal similarity and proximity matrices for the K-means clusterings of the following random data set.



UNSUPERVISED MEASURES: VISUAL JUDGEMENT BY SIMILARITY MATRIX

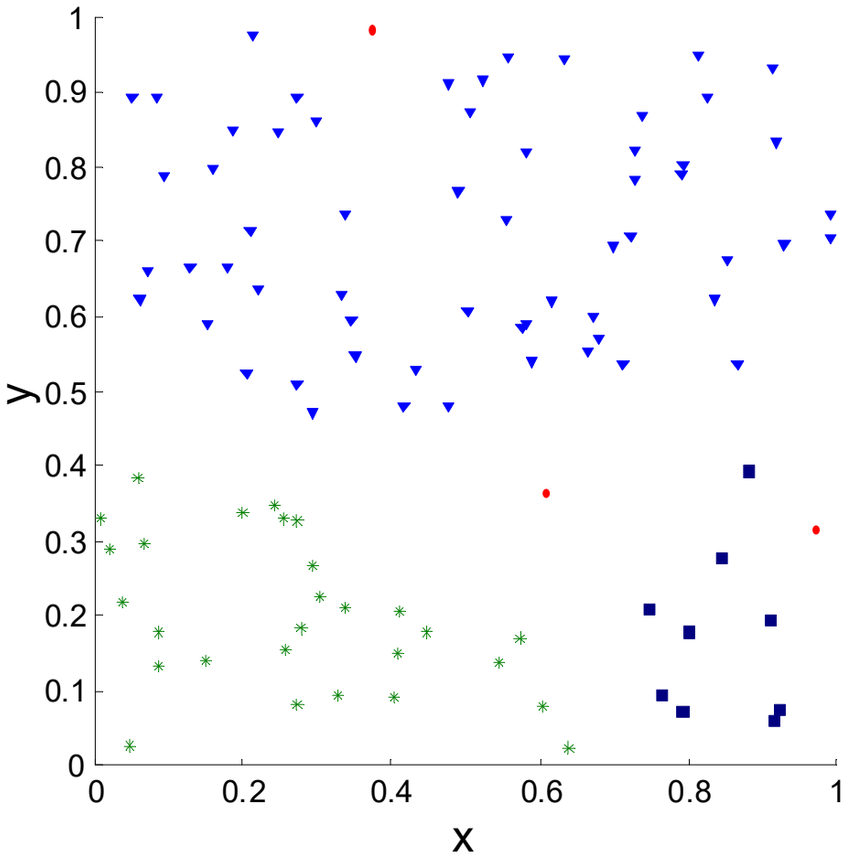
- order the similarity matrix with respect to cluster labels and inspect visually.

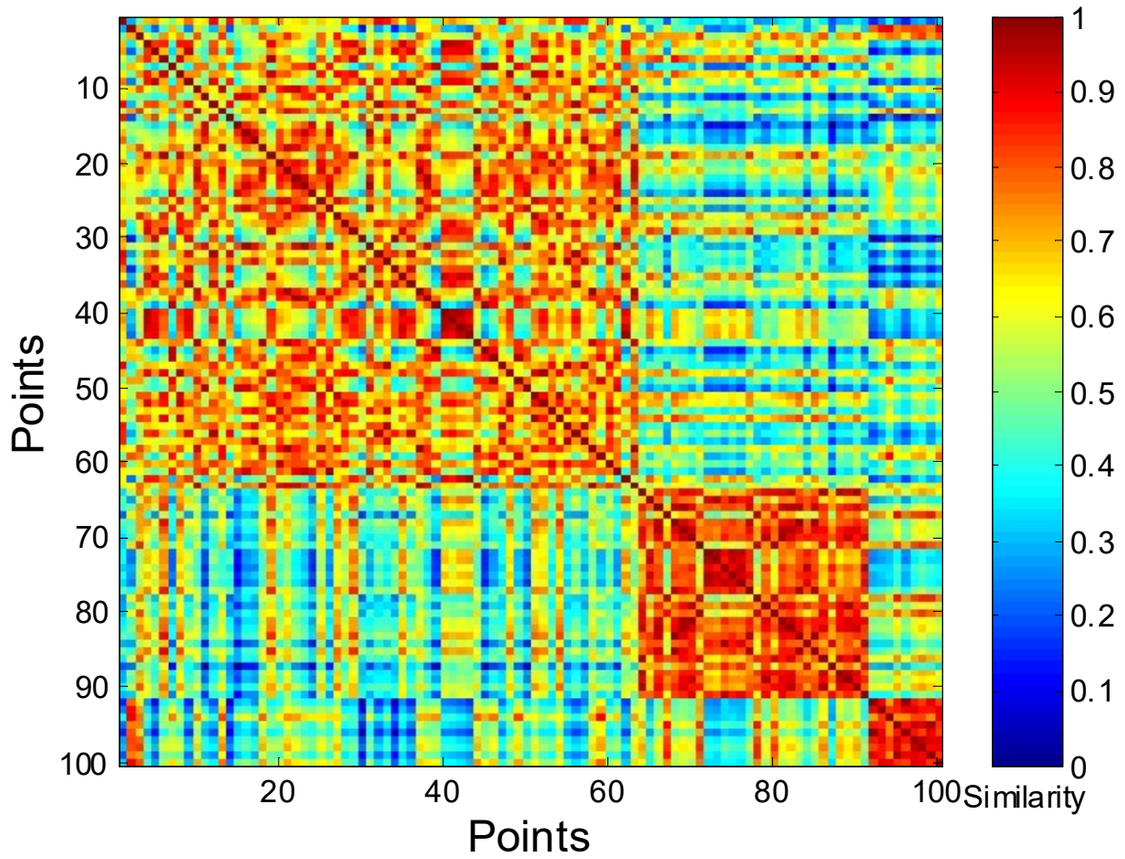




UNSUPERVISED MEASURES: VISUAL JUDGEMENT BY SIMILARITY MATRIX

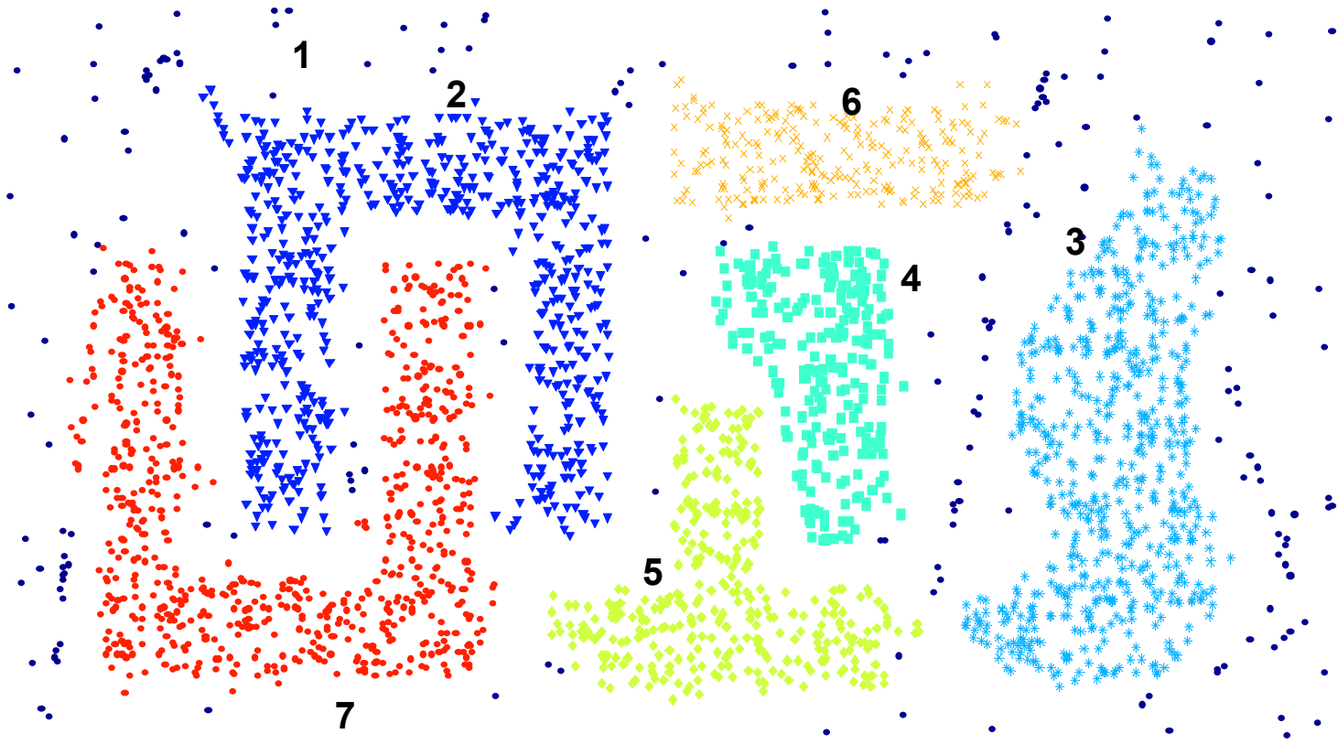
- clusters in random data are not so crisp

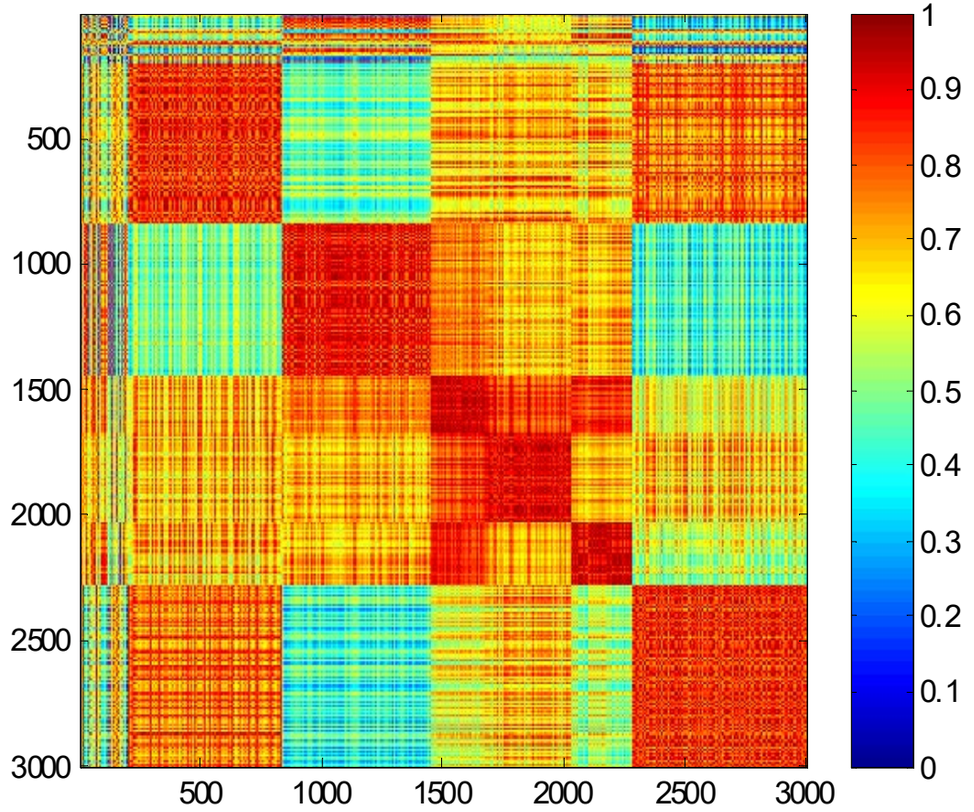




UNSUPERVISED MEASURES: VISUAL JUDGEMENT BY SIMILARITY MATRIX

- DBSCAN





UNSUPERVISED MEASURES: ADDITIONAL INDICES

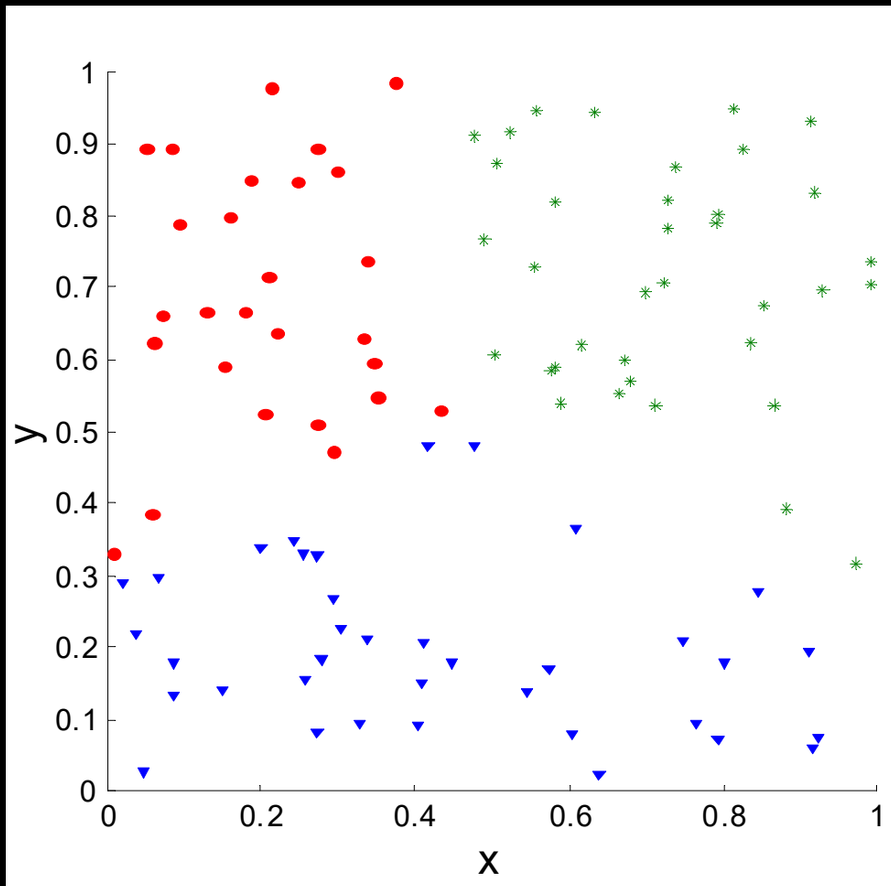
- Calinski and Harabasz
  - the value of K corresponding to the maximum is taken to be the optimal number of clusters
- Dunn
  - the value of K corresponding to the maximum is taken to be the optimal number of clusters (a large value suggests the presence of compact and well-separated clusters)
- Davies-Bouldin
  - the value of K corresponding to the minimum is taken to be the optimal number of clusters

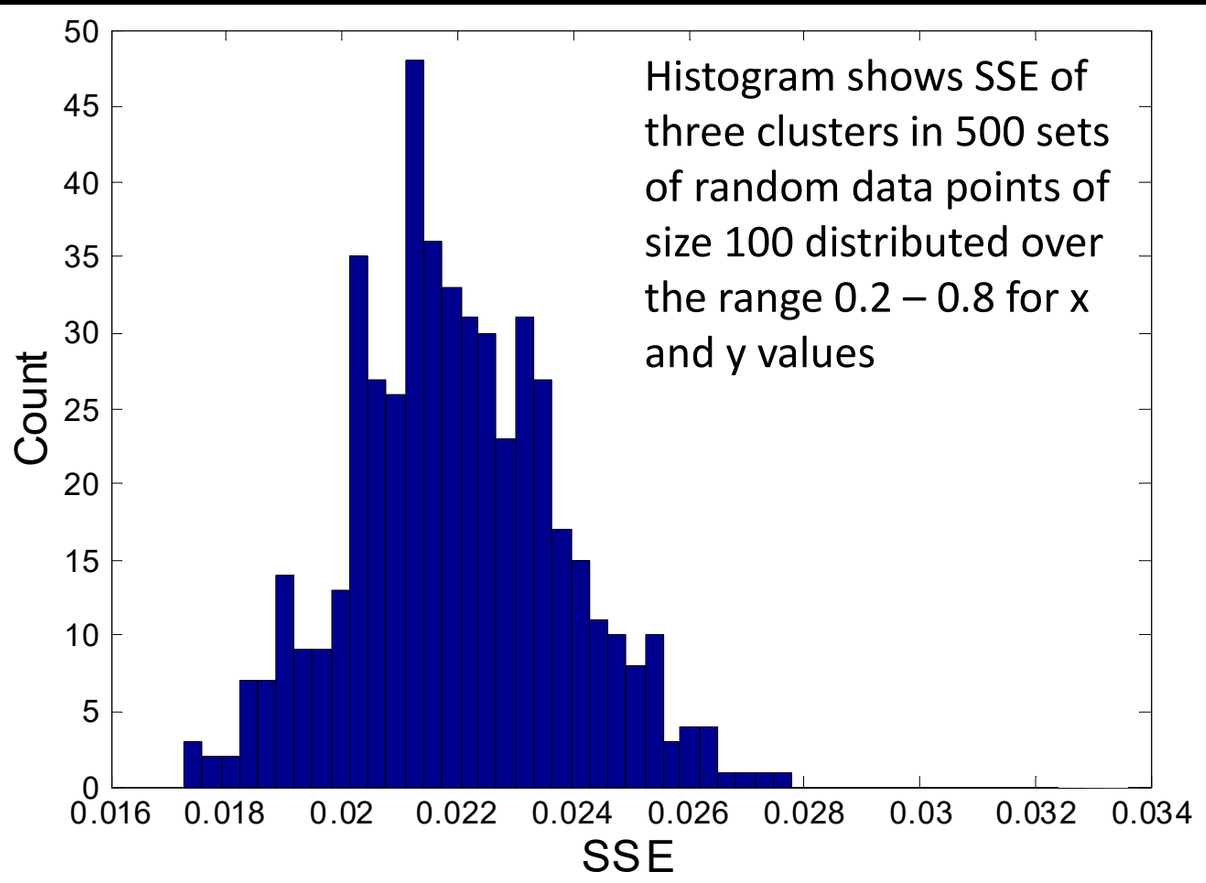
CLUSTERING VALIDATION

- The most important issues for cluster validation are:
  - determine the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data
  - determine the “correct number of clusters” (whatever it means!!!)
- Need a framework to interpret any measure
  - for example, if our measure of evaluation has the value, 10, is that good, fair, or poor?
- Statistics provide a framework for cluster validity
  - clustering result is, the more likely it represents valid structure in the data
  - compare the value of an index obtained from the given data with those resulting from random data.
  - if the value of the index is unlikely, then the cluster results are valid

CLUSTERING VALIDATION: THROUGH SSE

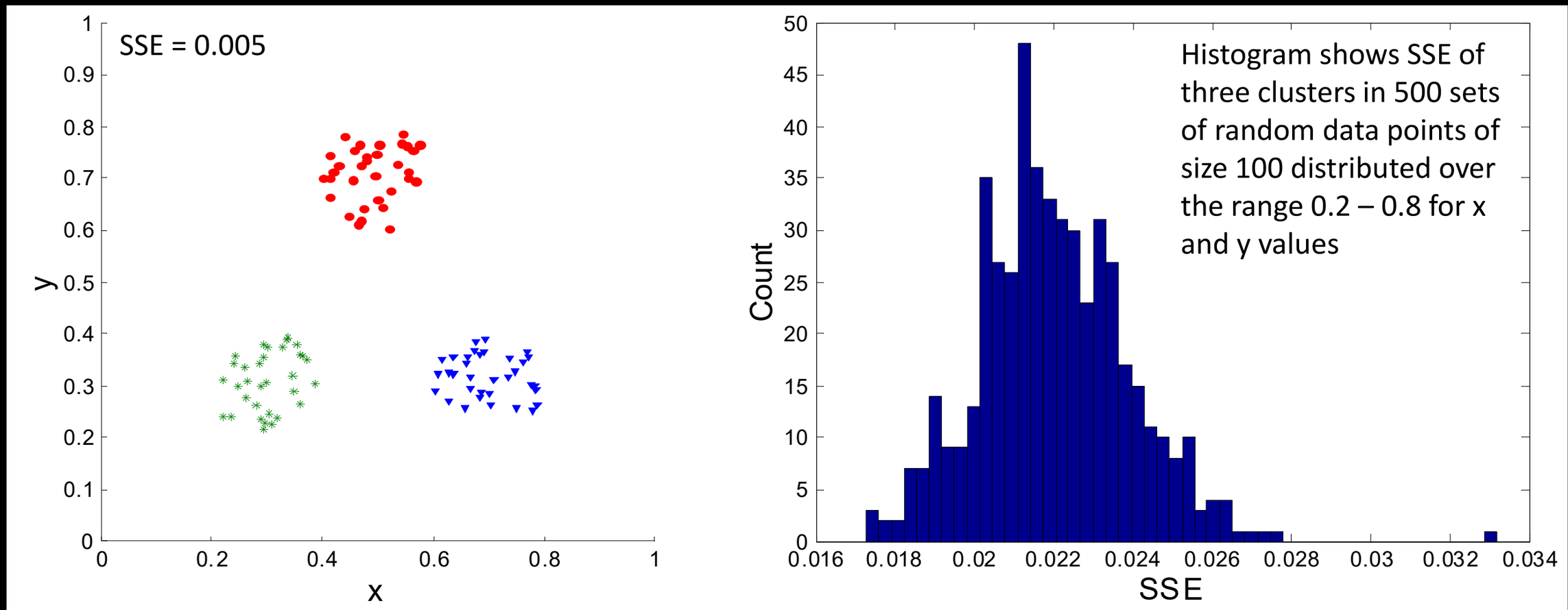
- Example: compare SSE of three cohesive clusters against three clusters in random data





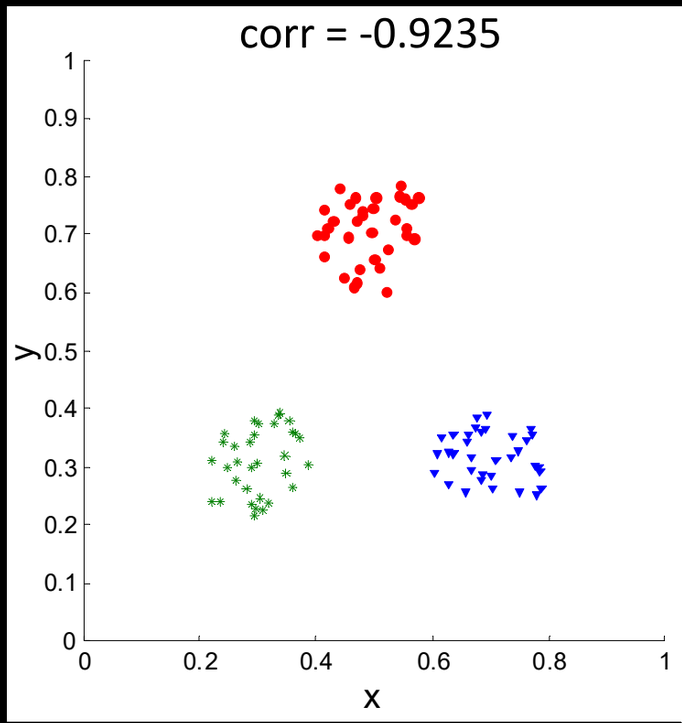
CLUSTERING VALIDATION: THROUGH SSE

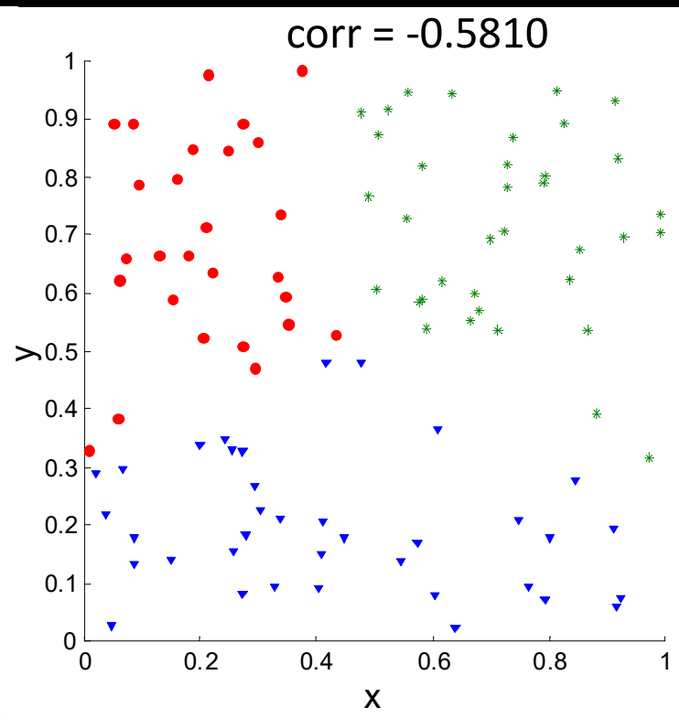
- Example: compare SSE of three cohesive clusters against three clusters in random data

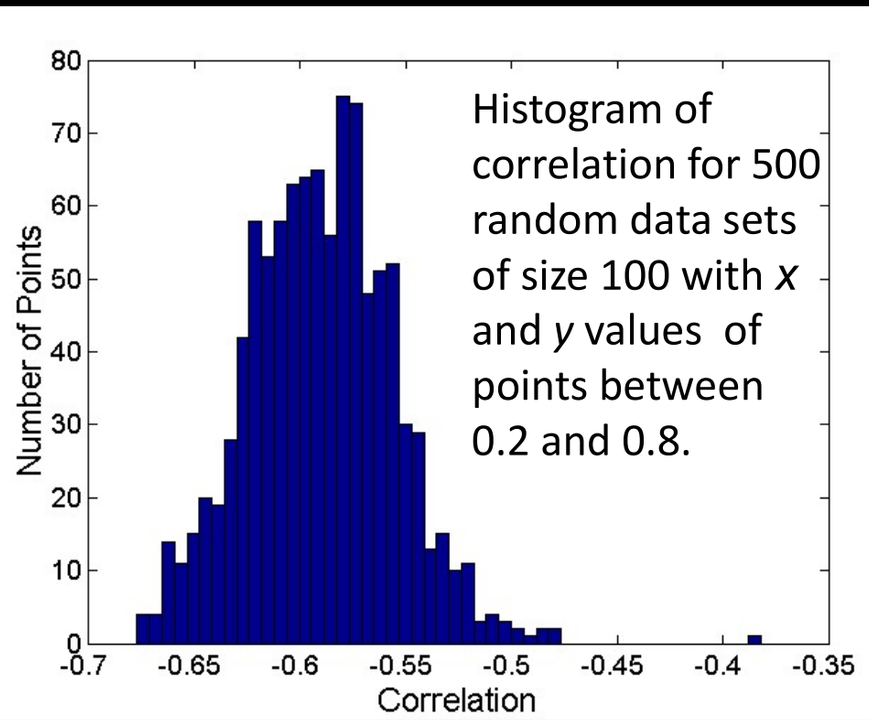


CLUSTERING VALIDATION: THROUGH CORRELATION

- correlation of ideal similarity and proximity matrices for the K-means clusterings of the following two data sets.

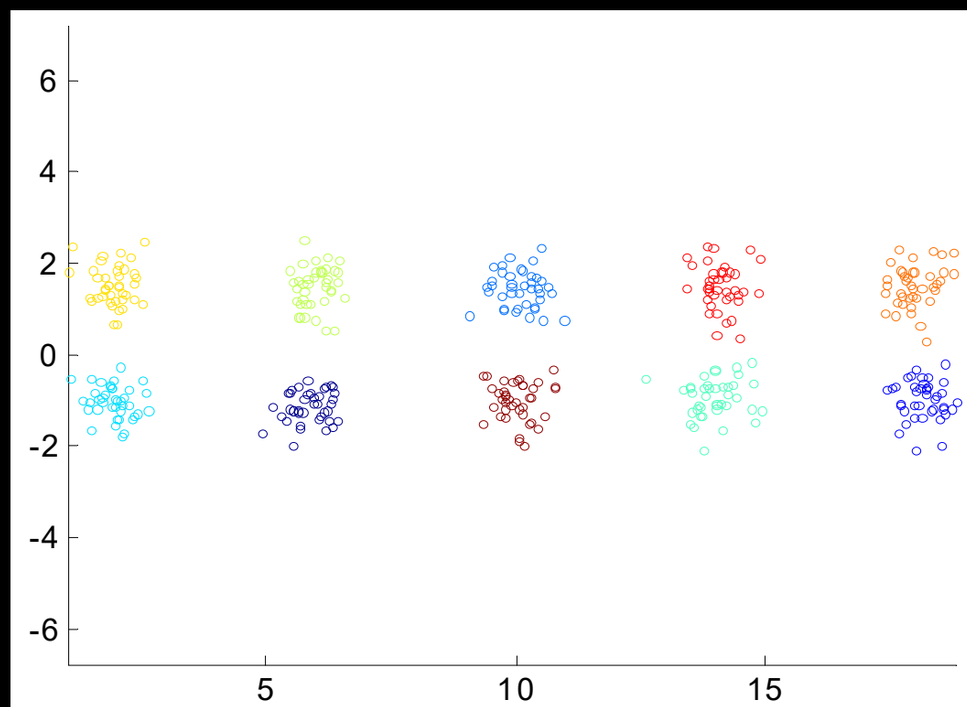


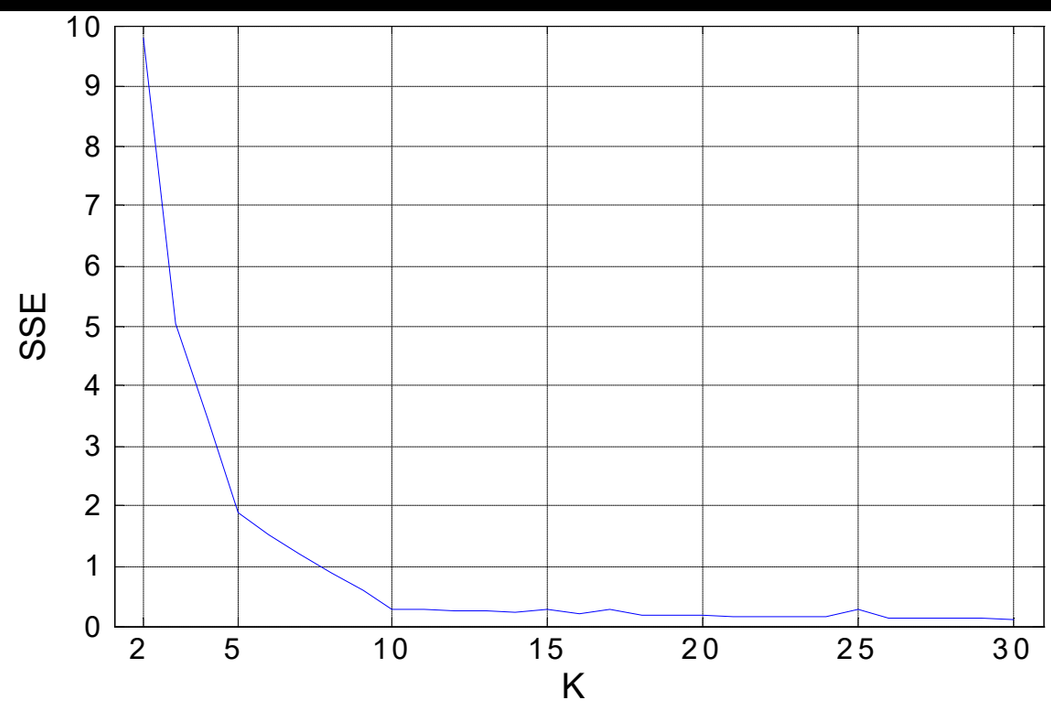




CLUSTERING VALIDATION: OPTIMAL NUMBER OF CLUSTERS?

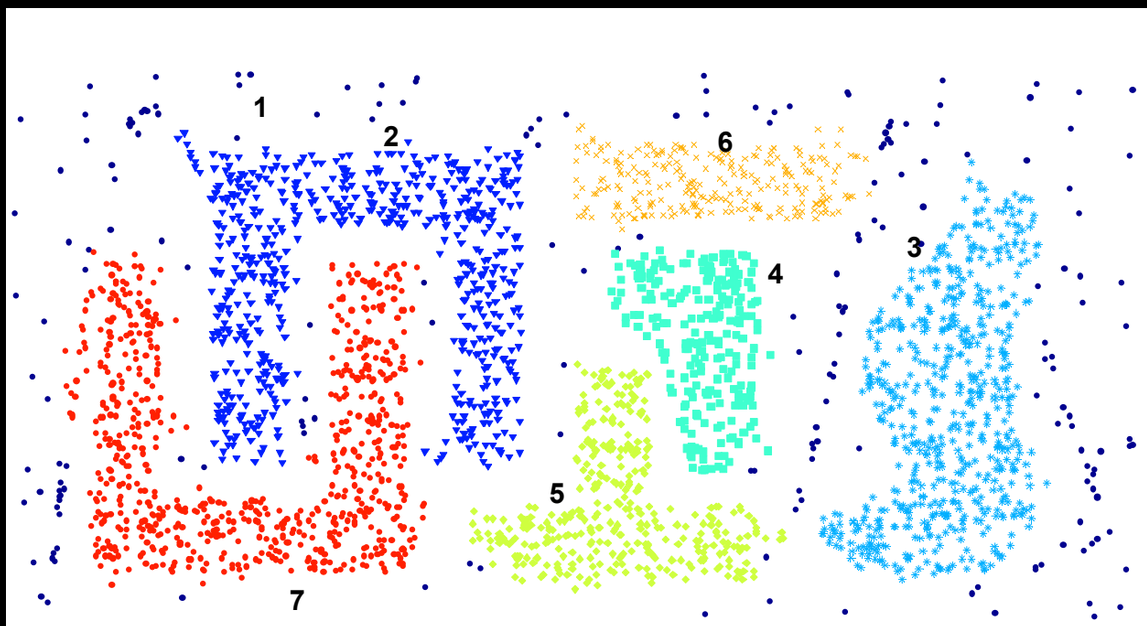
- An important issue for cluster validation is:
  - determine the "correct number of clusters" (whatever it means!!!)
    - SSE is good for comparing two clusterings or two clusters
    - SSE can also be used to estimate the number of clusters

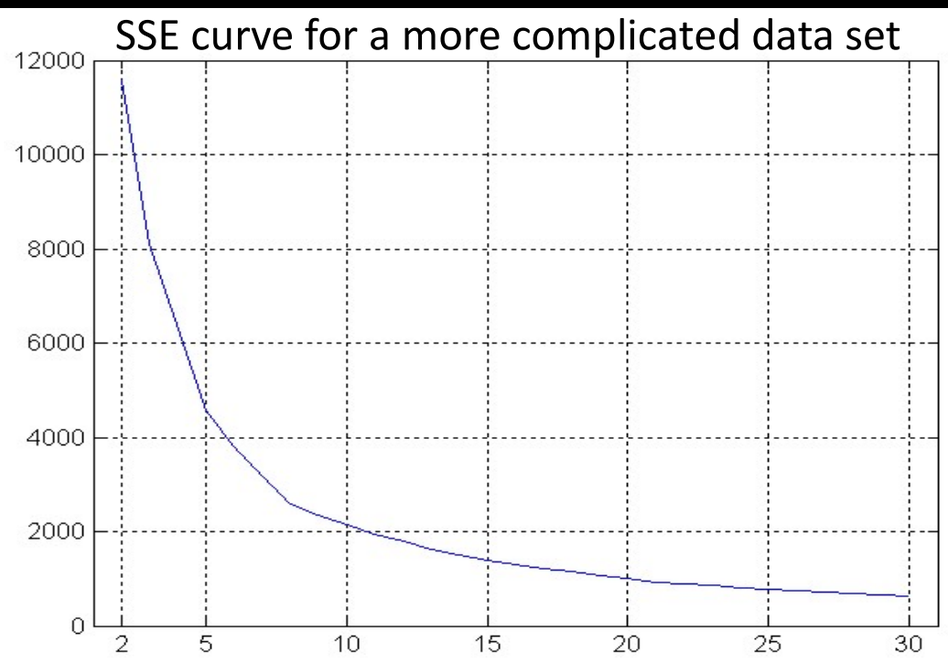




CLUSTERING VALIDATION: OPTIMAL NUMBER OF CLUSTERS?

- An important issue for cluster validation is:
  - determine the "correct number of clusters" (whatever it means!!!)
    - SSE is good for comparing two clusterings or two clusters
    - SSE can also be used to estimate the number of clusters





CLUSTERING VALIDATION: OPTIMAL NUMBER OF CLUSTERS?

- Supervised and Unsupervised indices can be used
  - Calinski and Harabasz
  - Dunn
  - Davies-Bouldin
  - Silhouette coefficient
  - Rand
  - Fowlkes and Mallows
  - AIC
  - BIC
  - MDL
  - …

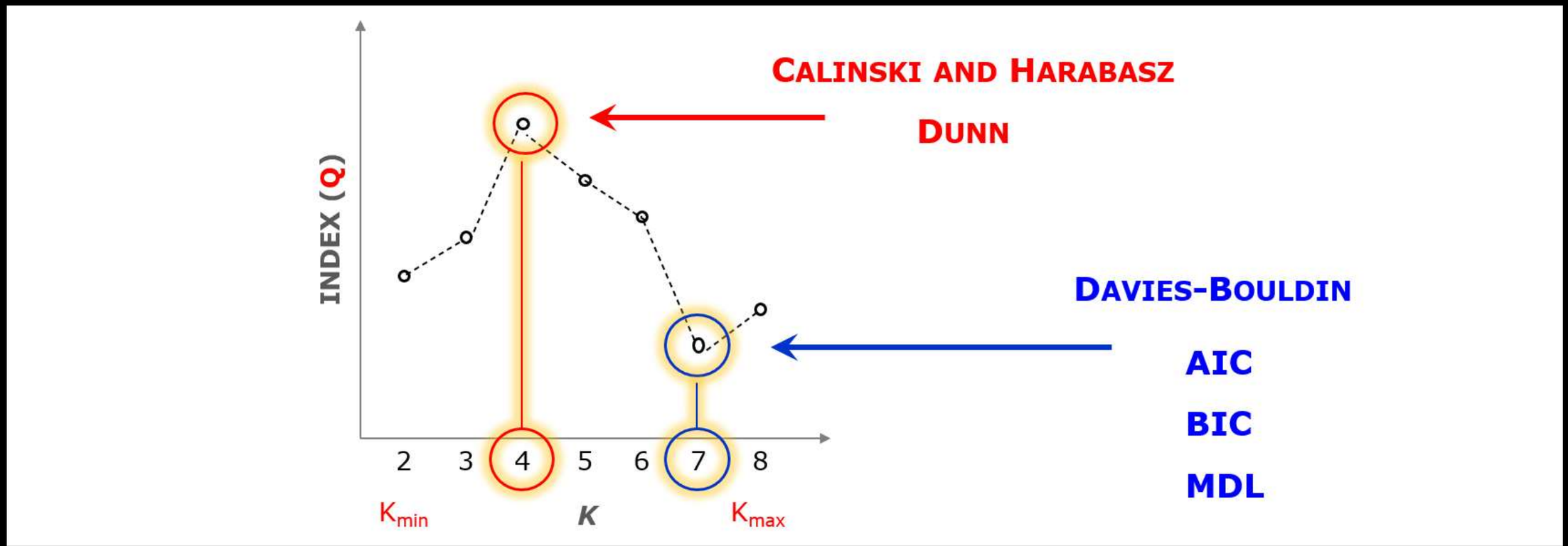
- For a clustering algorithm that requires the input of the number of clusters K from users, a sequence of clustering structures is obtained by running a clustering algorithm several times (r) where K ranges from K_{min} to K_{max}.

GENERATE SEQUENCE OF CLUSTERING STRUCTURES

1. Choose a clustering algorithm and a validity index
2. **FOR** K=K_{min} to K_{max} **DO**
3. **FOR** i=1 to r **DO**
4. run the clustering algorithm with K and use parameter values different from the previous running
1. compute the value q of the validity index and set $q(i) = q$
2. **END FOR**
3. Choose the best value $\mathbf{q}^*$ of the validity index $\{q_1, \ldots, q_r\}$
4. set $Q(K) = \mathbf{q}^*$
5. **END FOR**

CLUSTERING VALIDATION: OPTIMAL NUMBER OF CLUSTERS?

- The clustering structures are then evaluated based on the computed index, and the "optimal clustering solution" is determined by choosing the one with the best value of the index.



In the case of hierarchical clustering structures, the indices are also known as stopping rules, which tell where the best level is in order to cut the dendrogram.

FINAL COMMENT ON CLUSTERING

"The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage."

Algorithms for Clustering Data, Jain and Dubes

H. Xiong and Z. Li. Clustering Validation Measures. In C. C. Aggarwal and C. K. Reddy, editors, Data Clustering: Algorithms and Applications, pages 571–605. Chapman & Hall/CRC, 2013.

RECAP

- Validation Measures
  - Supervised
  - Unsupervised
- Correlation and Visual Methods
- Optimal Number of Clusters
- Final Comment on Clustering Analysis