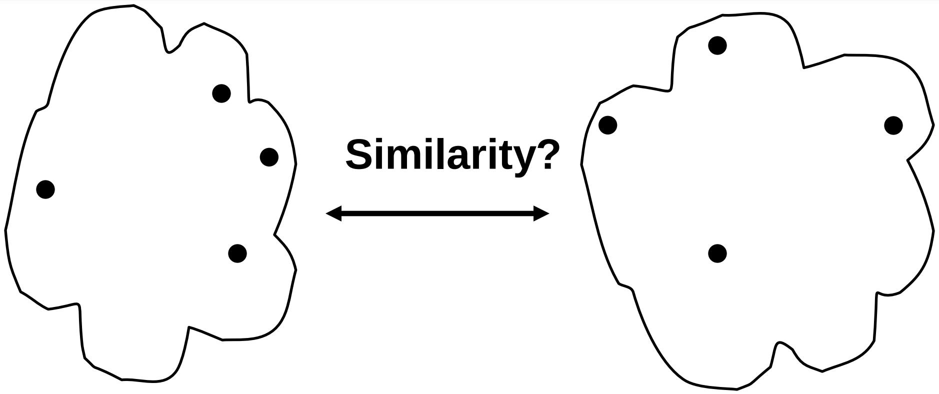
Quick Review

How to Define Inter-Cluster Similarity

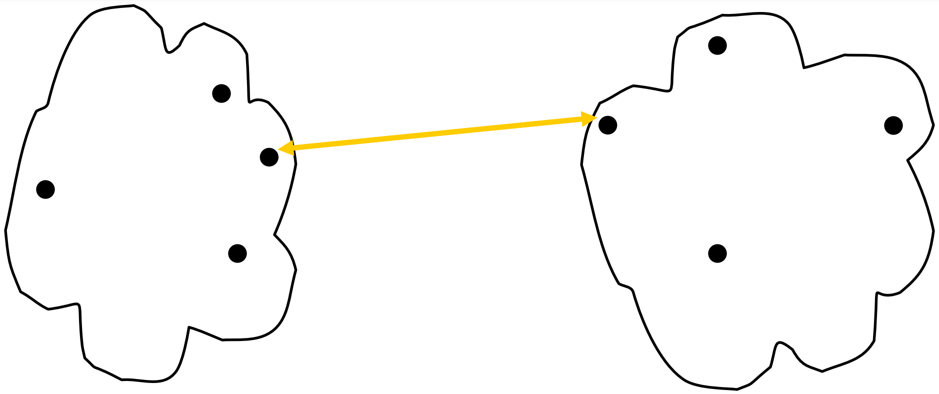


- MIN
- MAX
- Group Average

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

Proximity Matrix

How to Define Inter-Cluster Similarity

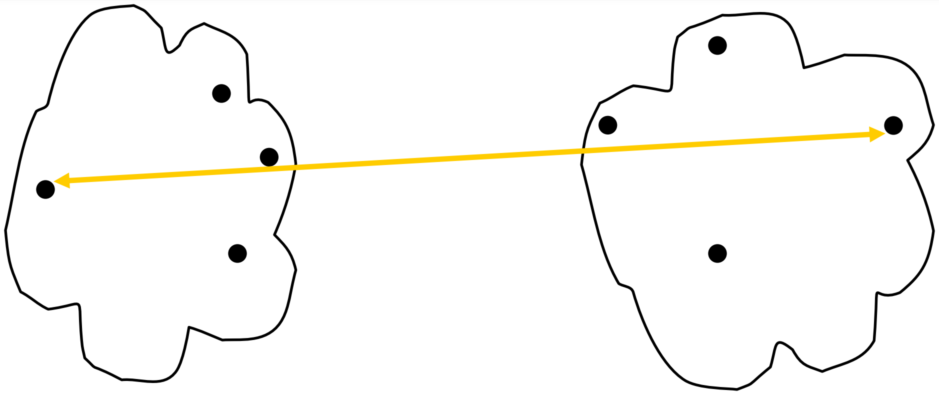


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity

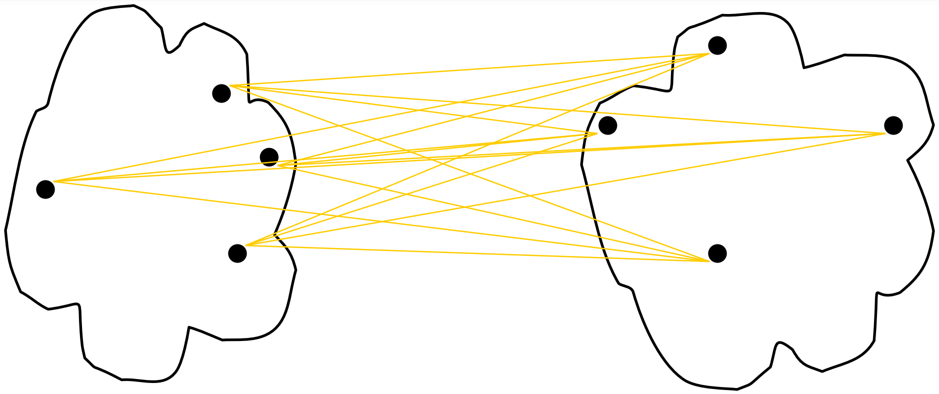


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

How to Define Inter-Cluster Similarity



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

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids
- Other methods driven by an objective function
  - Ward's Method uses squared error

Single Link

	a	b	c	d	e	f
a	0	12	6	3	25	4
b		0	19	8	14	15
c			0	12	5	18
d				0	11	9
e					0	7
f						0

	ad	b	c	e	f
ad	0	8	6	11	4
b		0	19	14	15
c			0	5	18
e				0	7
f					0

	adf	b	ce
adf	0	8	6
b		0	14
ce			0

	$adfce$	b
$adfce$	0	8
b	8	0

Complete Link min(max distances)

	a	b	c	d	e	f
a	0	12	6	3	25	4
b		0	19	8	14	15
c			0	12	5	18
d				0	11	9
e					0	7
f						0

	ad	b	c	e	f
ad	0	12	12	25	9
b		0	19	14	15
c			0	5	18
e				0	7
f					

	ad	b	ce	f
ad	0	12	25	9
b		0	19	15
ce				18
f				

Properties of intergroup similarity

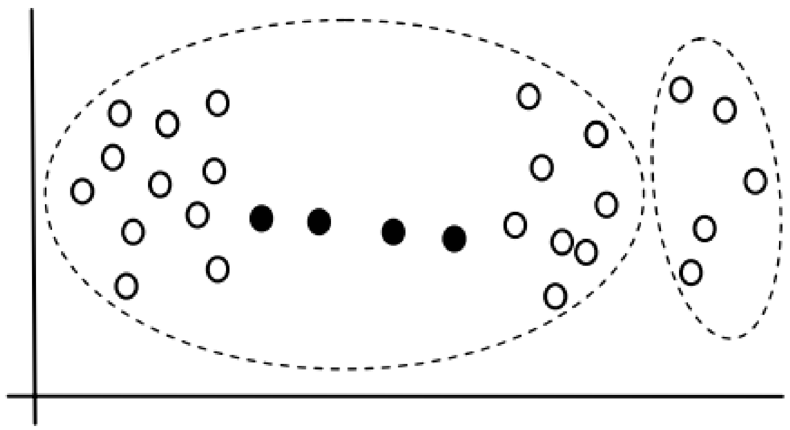
- Single linkage
  - can produce “chaining,” where a sequence of close observations in different groups cause early merges of those groups

- Complete linkage has the opposite problem.
  - It might not merge close groups because of outlier members that are far apart.

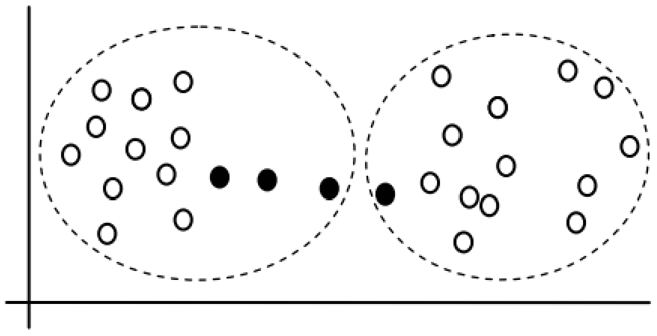
- Group average represents a natural compromise,
  - but depends on the scale of the similarities. Applying a monotone transformation to the similarities can change the results.

Properties of intergroup similarity

- Single-link is suitable for non-elliptical shape clusters
- But, it is sensitive to noise and may cause a chain effect and produce straggly clusters



Chain effect of the single link method



Clustering using the complete link method

Hierarchical Clustering: Time and Space

- SPACE
  - $O(N^2)$ space since it uses the proximity matrix.
    - N is the number of points.

- TIME
  - $O(N^3)$ time in many cases
    - There are N steps, and at each step, the size, N^2, proximity matrix must be updated and searched
    - Complexity can be reduced to $O(N^2 \log(N))$ time if we use a special structure like a heap or sorted lists

Hierarchical Clustering: Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- Do not scale well: time complexity of at least $O(N^2 \log N)$, where n is the number of total objects
- Different schemes have problems with one or more of the following:
  - Sensitivity to noise and outliers
  - Difficulty handling different sized clusters and convex shapes
  - Breaking large clusters

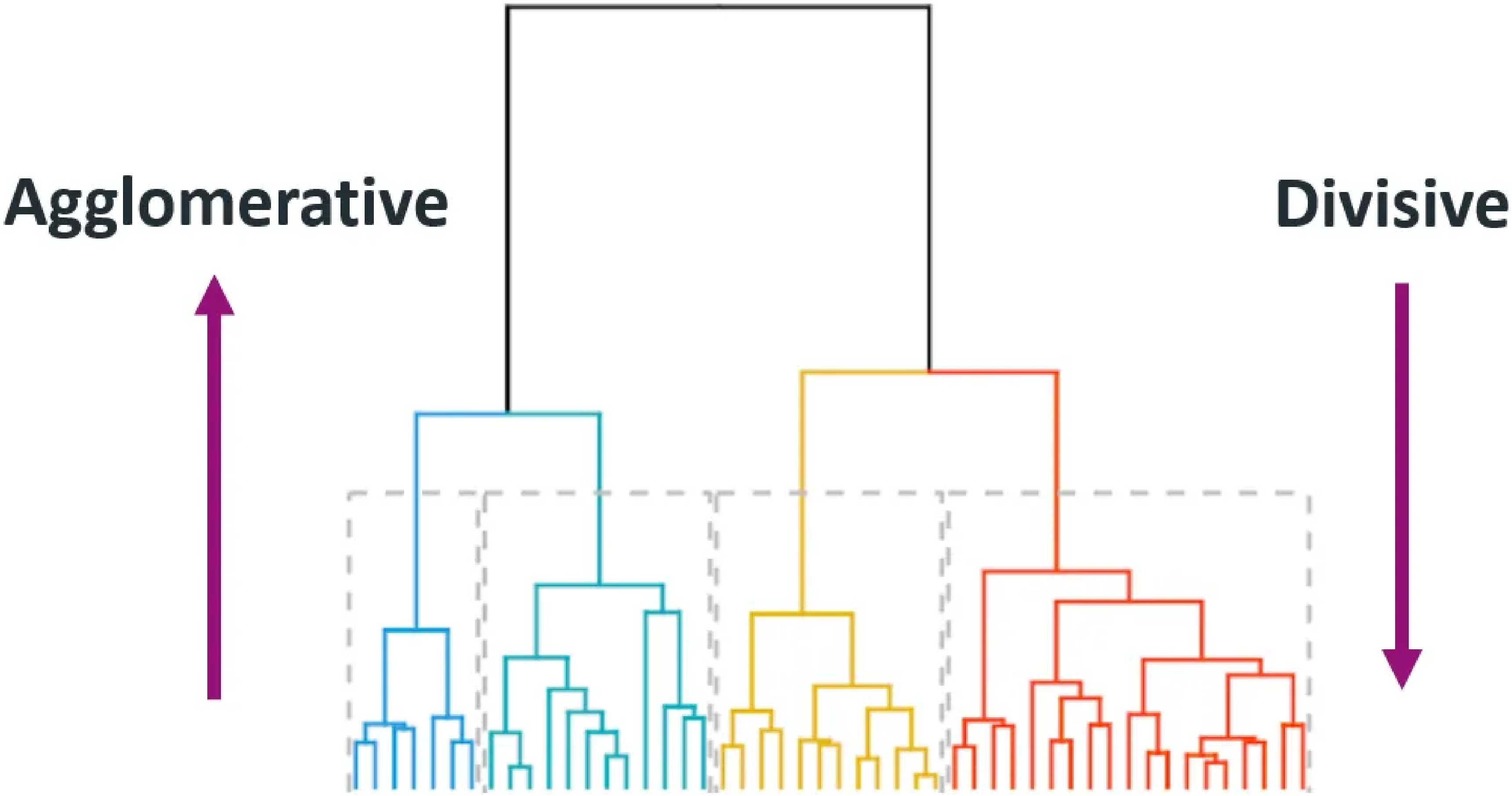
Hierarchical Clustering

Two main types of hierarchical clustering

Agglomerative	Divisive
• Start with the points as individual clusters • At each step, merge the closest pair of clusters until only one cluster (or k clusters) left	• Start with one, all-inclusive cluster • At each step, split a cluster until each cluster contains a point (or there are k clusters)

Traditional hierarchical algorithms use a similarity or distance matrix to Merge or split one cluster at a time

Hierarchical Clustering



https://harshsharma1091996.medium.com/hierarchical-clustering-996745fe656b

Agglomerative Clustering Algorithm

More popular hierarchical clustering technique

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

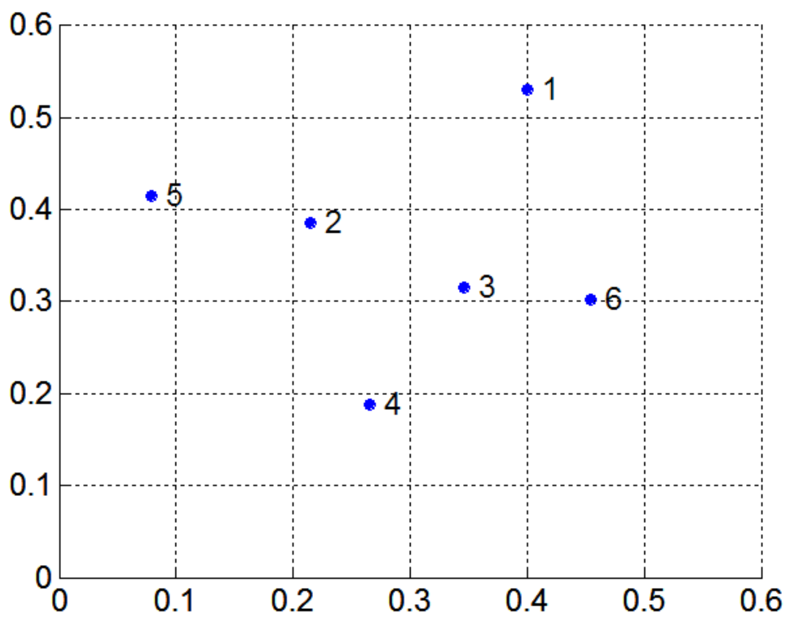
Key operation is the computation of the proximity of two clusters

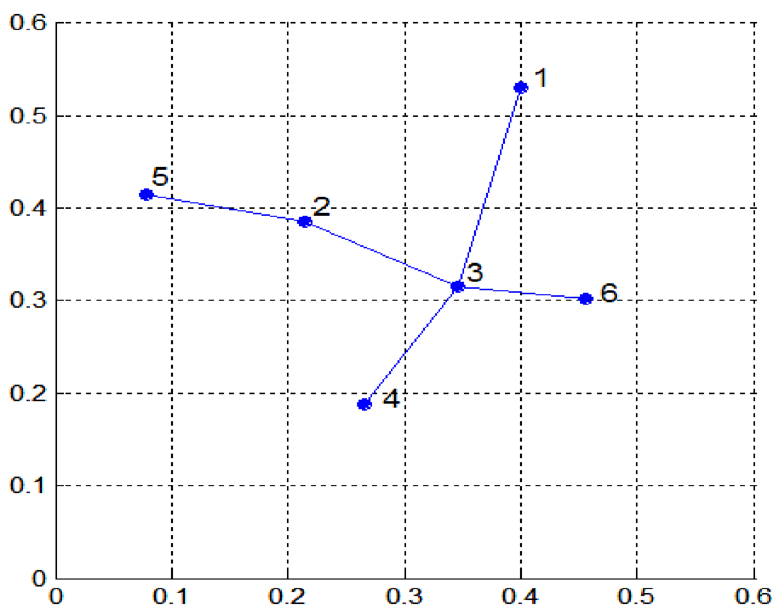
Different approaches exists to define the distance between clusters

MST: Divisive Hierarchical Clustering

Build MST (Minimum Spanning Tree)

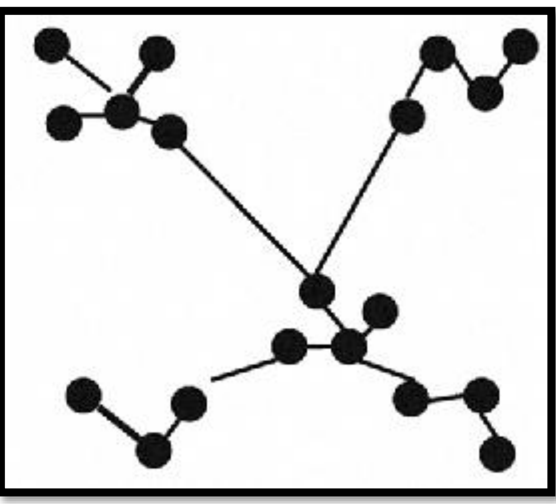
- Start with an arbitrary vertex (consider it a tree with one vertex)
- In successive steps, look for the closest pair of points (p, q) such that one point (p) is in the current tree but the other (q) is not
- Add q to the tree and put an edge between p and q



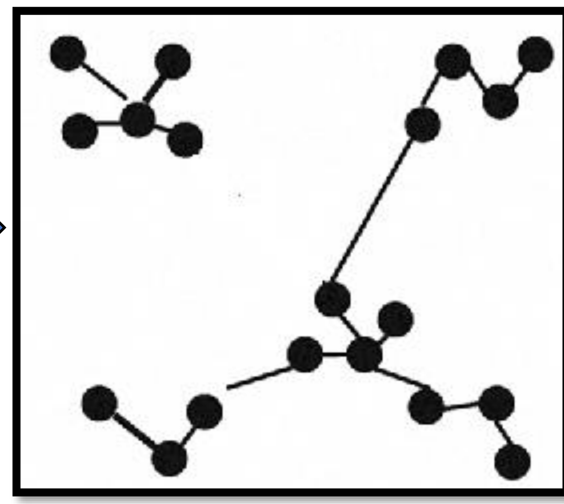


Algorithm MST Divisive Hierarchical Clustering

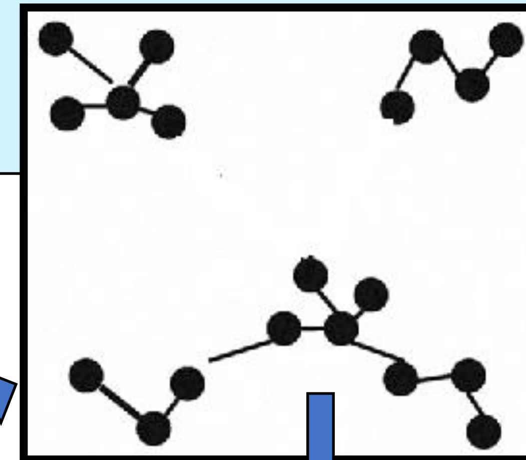
- Compute MST for the proximity graph
- **Repeat**
  - ◆ Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity)
- **Until only singleton clusters remain**

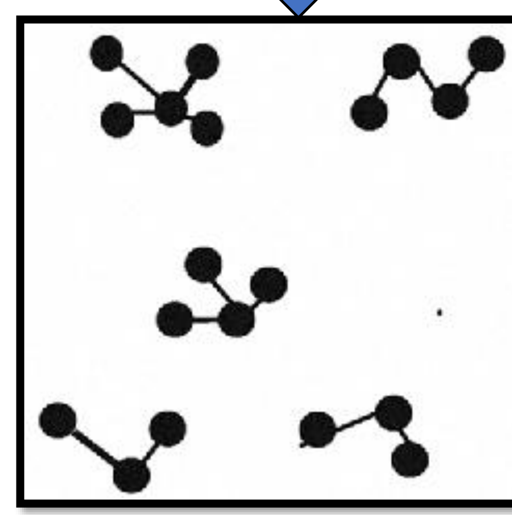


MST



Clusters after removing longest edges

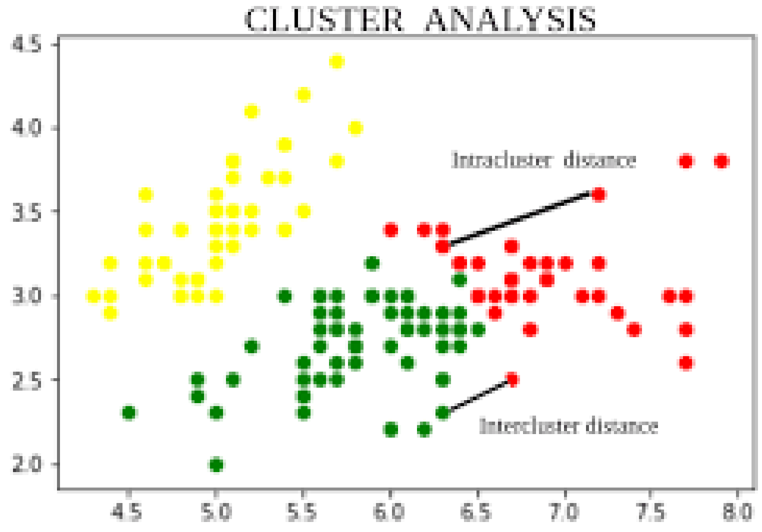




Remarks

	Partitioning Clustering	Hierarchical Clustering
Time Complexity	$O(n)$	$O(n^2 \log n)$
Pros	Easy to use and Relatively efficient	Outputs a dendrogram that is desired in many applications.
Cons	Sensitive to initialization; bad initialization might lead to bad results. Need to store all data in memory.	higher time complexity; Need to store all data in memory.

Cluster Validity



Cluster Validity

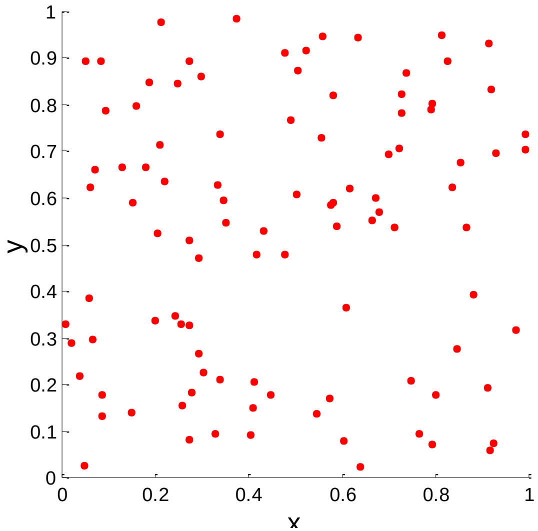
- For cluster analysis, we want to evaluate the "goodness" of the resulting clusters?

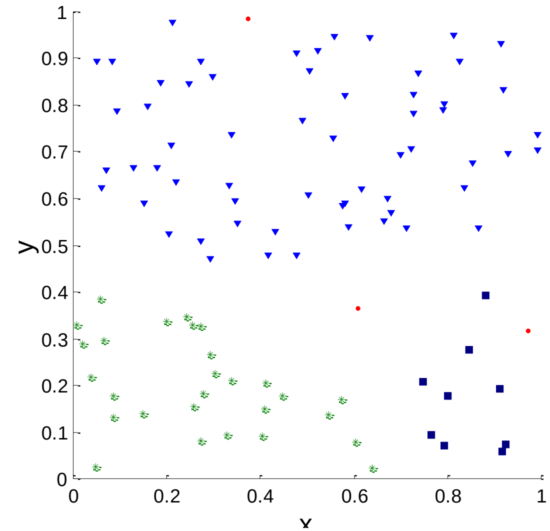
- But "clusters are in the eye of the beholder"!

- **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

Clusters found in Random Data

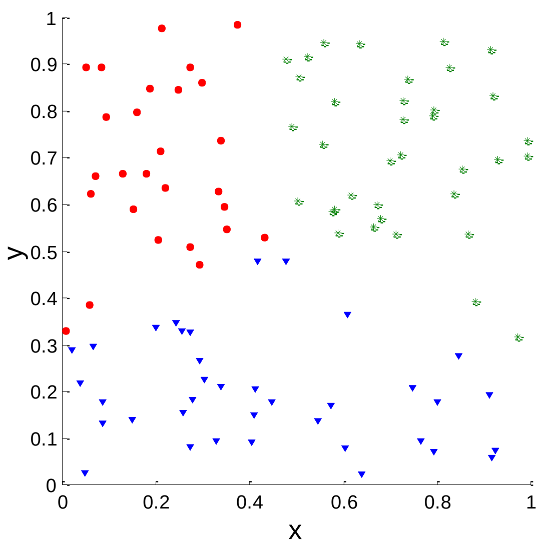
Random Points

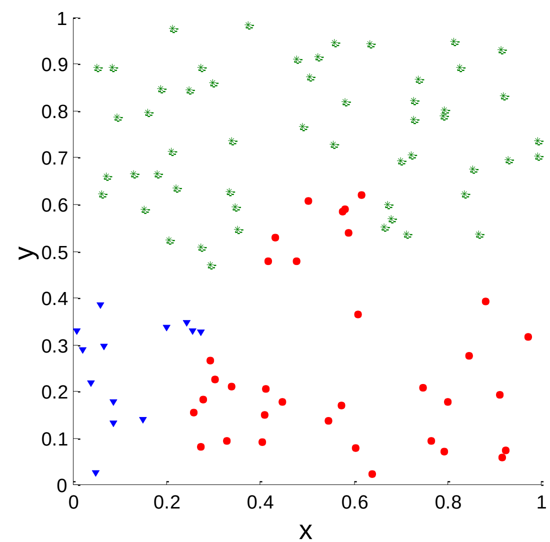




DBSCAN

K-means





Complete Link

Measures of Cluster Validity

Numerical measures to judge cluster validity:

- External Index: measure the extent to which cluster labels match externally supplied class labels.
  - Entropy
- Internal Index: measure the goodness of a clustering structure *without* respect to external information.
  - Sum of Squared Error (SSE)
- Relative Index: Used to compare two different clusterings or clusters.
  - Often, an external or internal index is used for this function, e.g., SSE or entropy

Unsupervised Cluster Evaluation

- Consider two unsupervised approaches for cluster evaluation using the Proximity Matrix
  - **Correlation of actual and Ideal Proximity matrices**
  - **Visualization**

- Two matrices
  - **Similarity Matrix for the data set**
  - **Ideal Similarity Matrix** (cluster label from cluster analysis)
    - 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

Measuring Cluster Validity Via Correlation

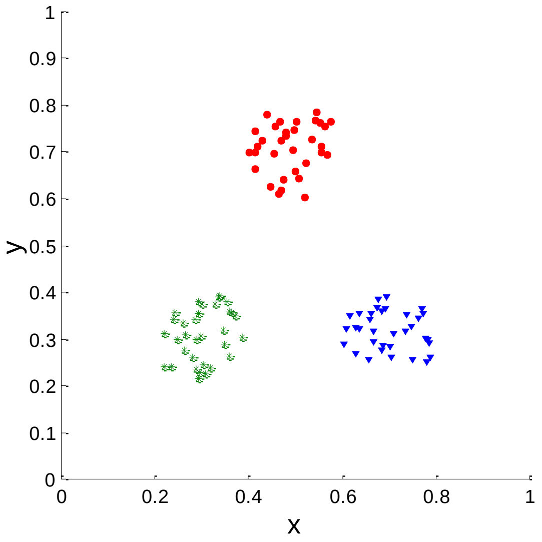
- Compute the correlation between the two matrices
  - Since the matrices are symmetric, only the correlation between **n(n-1) / 2** entries needs to be calculated.

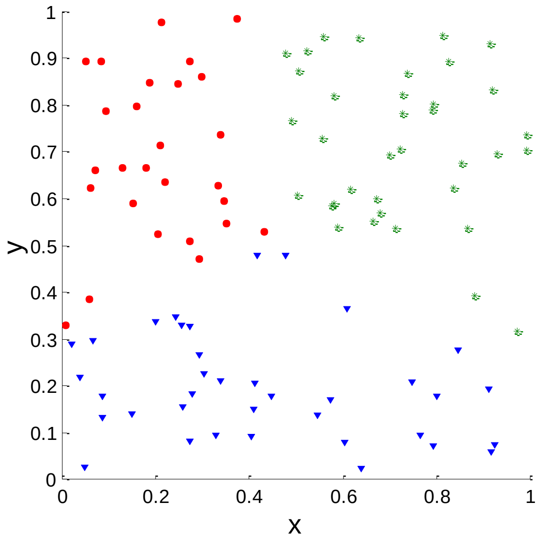
- High correlation indicates that points that belong to the same cluster are close to each other.

- Not a good measure for some density or contiguity based clusters.

Measuring Cluster Validity Via Correlation

- Correlation of Ideal and proximity matrices for the K-means clusterings of the following two data sets.



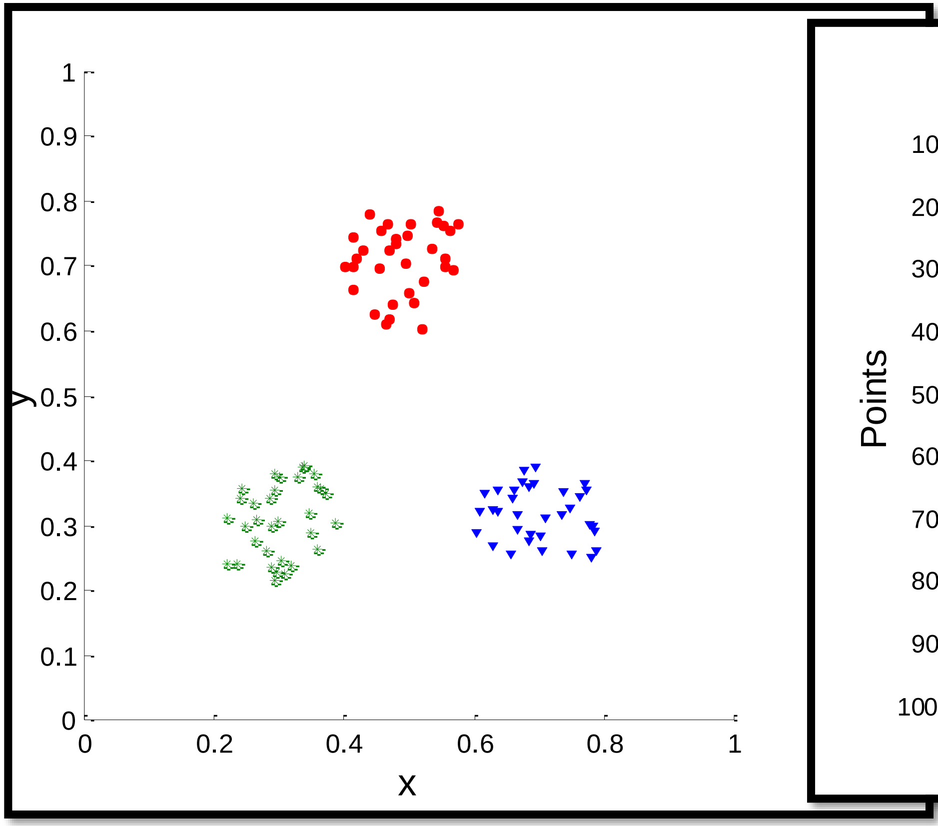


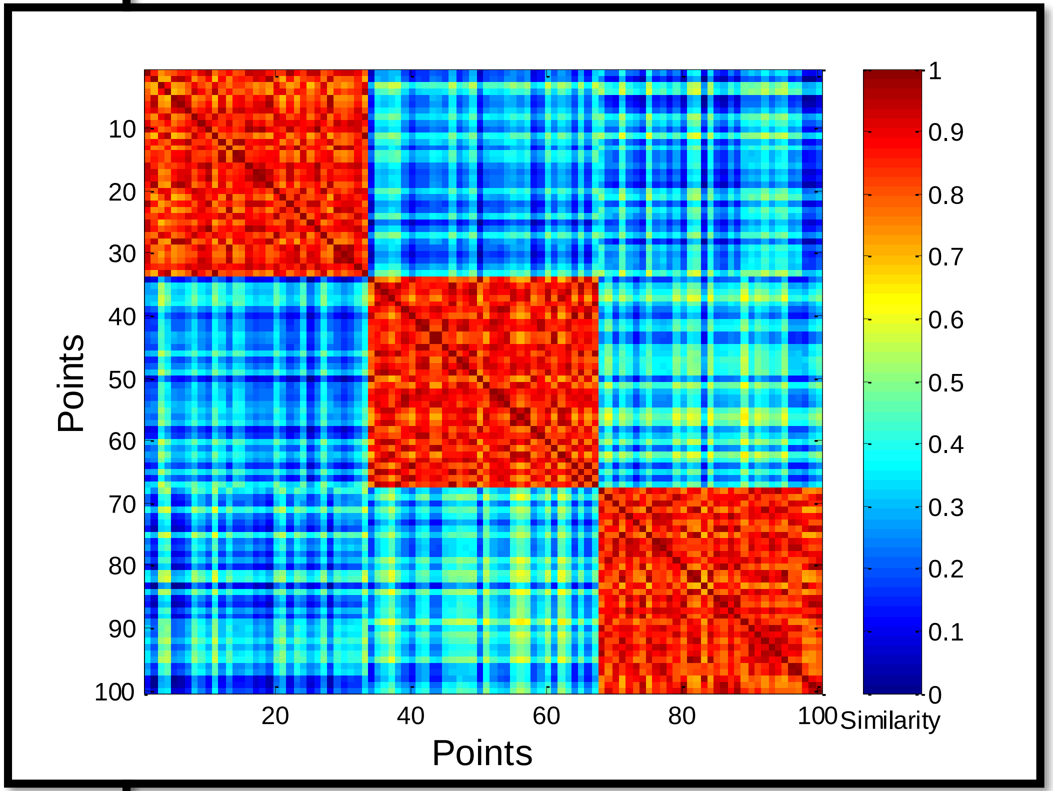
Corr = 0.9235

Corr = 0.5810

Using Similarity Matrix for Cluster Validation

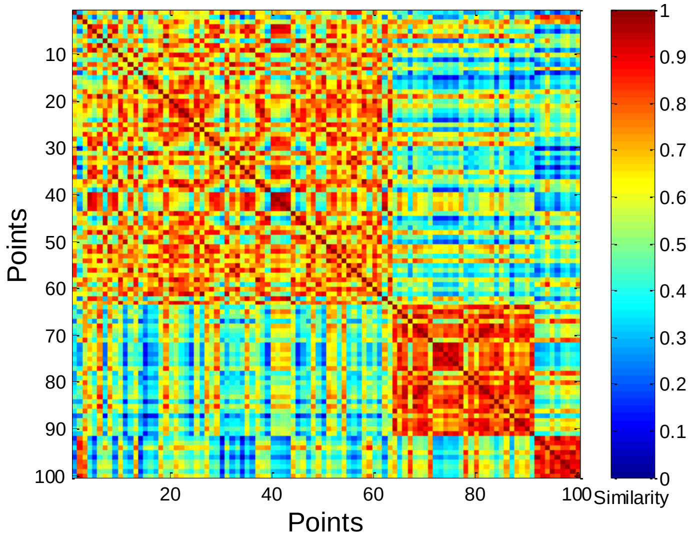
- Order the similarity matrix with respect to cluster labels and inspect visually.

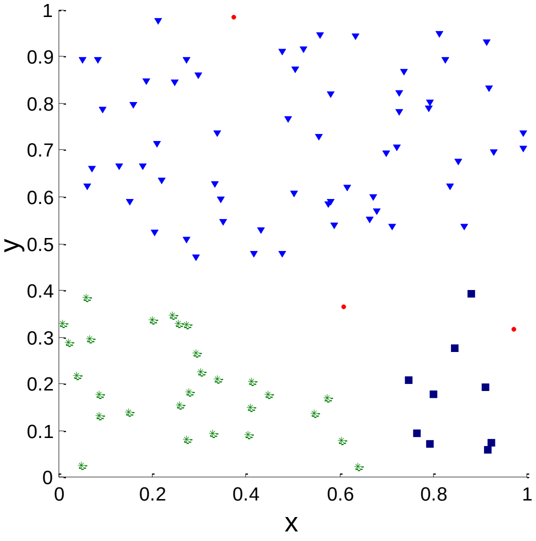




Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

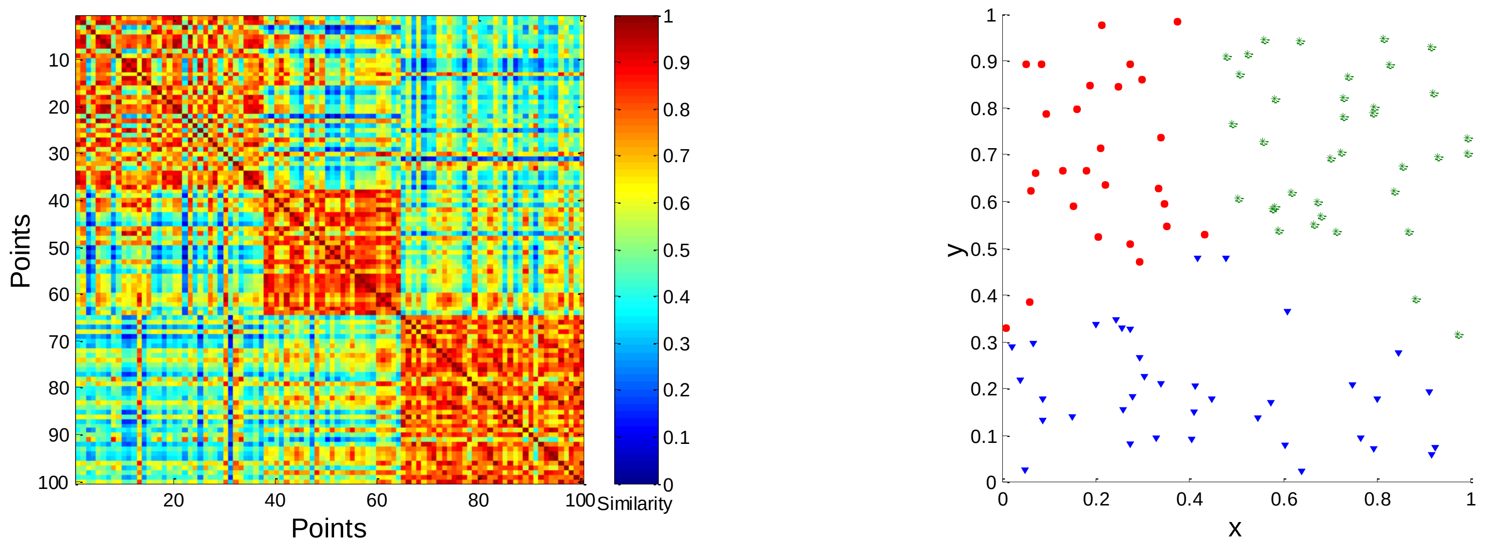




DBSCAN

Using Similarity Matrix for Cluster Validation

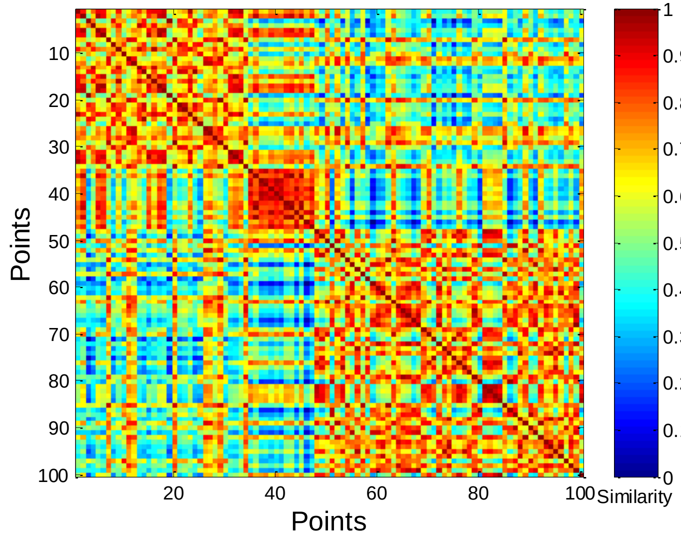
- Clusters in random data are not so crisp

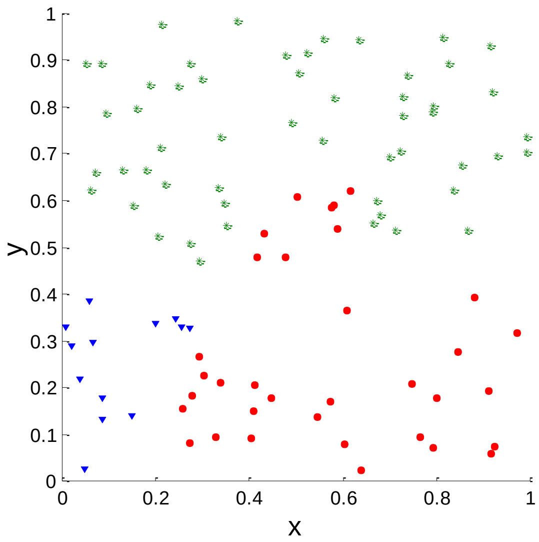


K-means

Using Similarity Matrix for Cluster Validation

- Clusters in random data are not so crisp

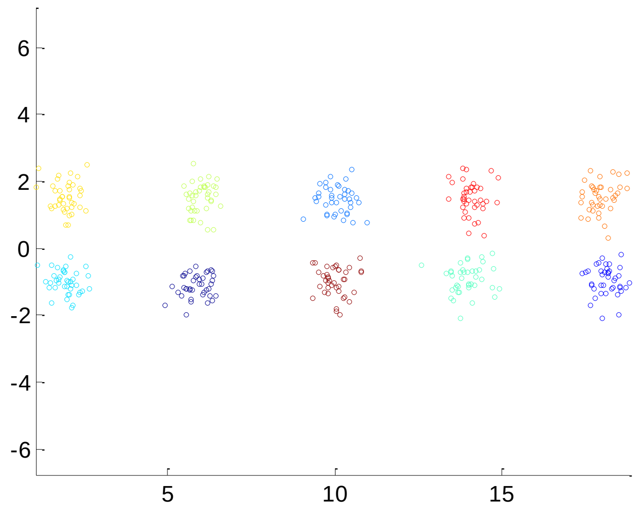


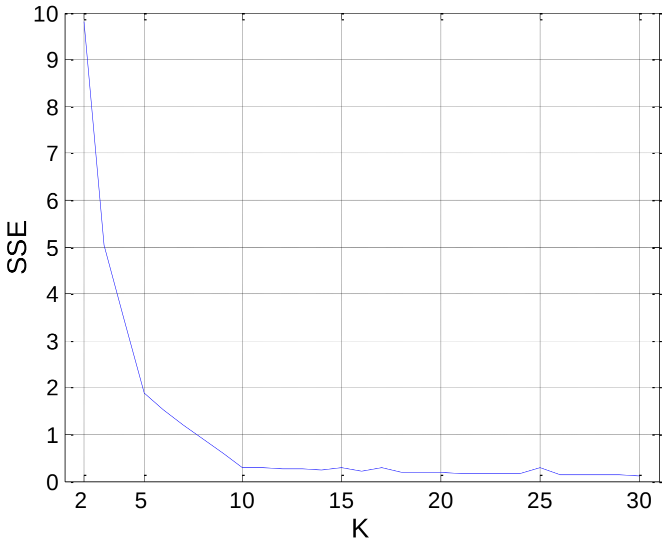


Complete Link

Internal Measures: SSE

- Clusters in more complicated data aren’t well separated
- **Internal Index:** Used to measure the goodness of a clustering structure without respect to external information
  – SSE
- SSE is good for comparing two clusterings or two clusters (average SSE).
- ***Can also be used to estimate the number of clusters***

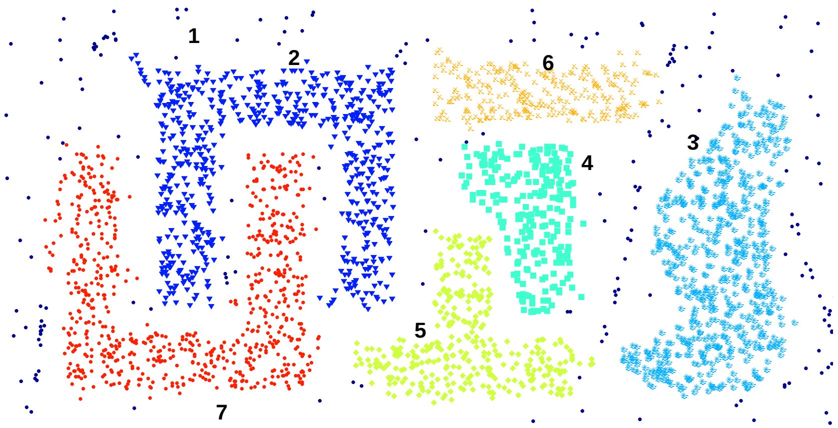


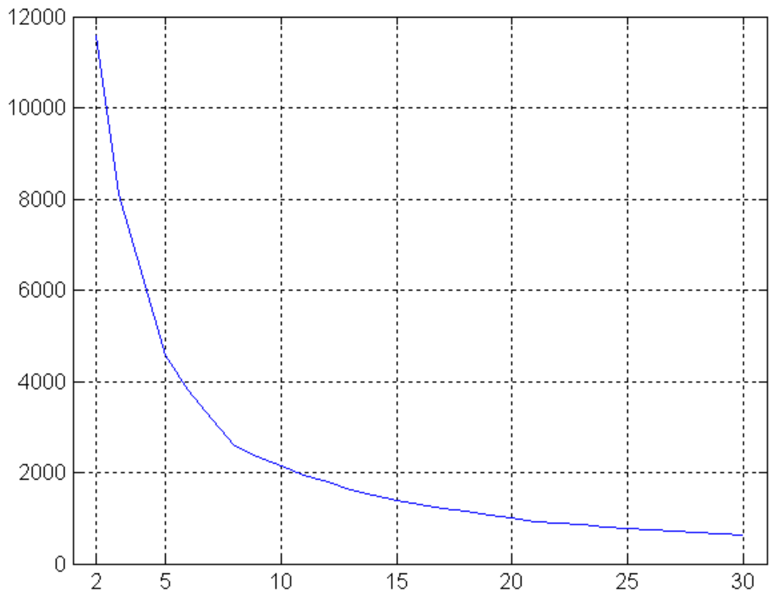


Shows a plot of the SSE vs the no of clusters for a (bisecting) k-means clustering of the data

Internal Measures: SSE

- SSE curve for a more complicated data set

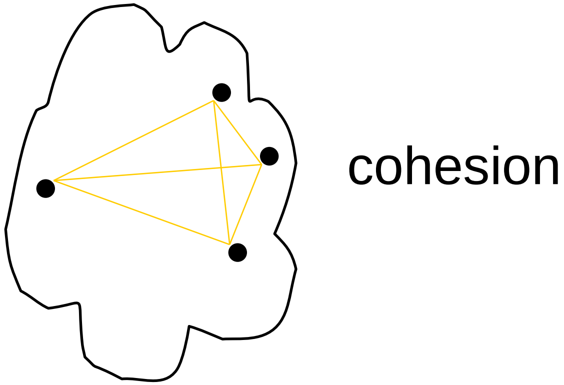




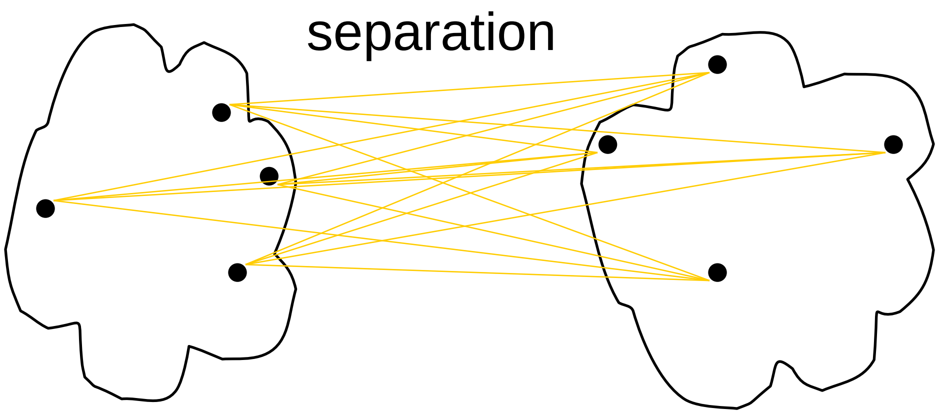
SSE of clusters found using K-means

Internal Measures: Cohesion and Separation

Cluster Cohesion: Measures how closely related are objects in a cluster

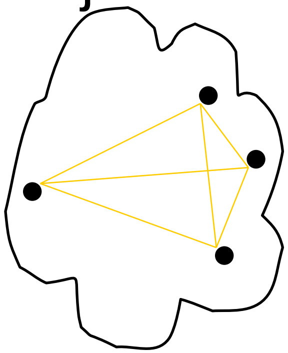


Cluster Separation: Measure how distinct or well-separated a cluster is from other clusters



Internal Measures: Cluster Cohesion

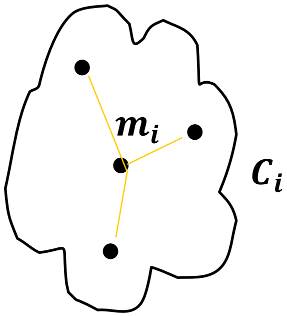
- Cohesion: Measures how closely related are objects in a cluster



Cohesion (C_i)$= \sum_{x,y \in C_i} \boldsymbol{proximity}(\boldsymbol{x}, \boldsymbol{y})$

The proximity function can be a similarity or a dissimilarity.

- Cohesion can be centroid based



Cohesion (C_i)$= \sum_{x \in C_i} \boldsymbol{proximity}(\boldsymbol{x}, \boldsymbol{m_i})$

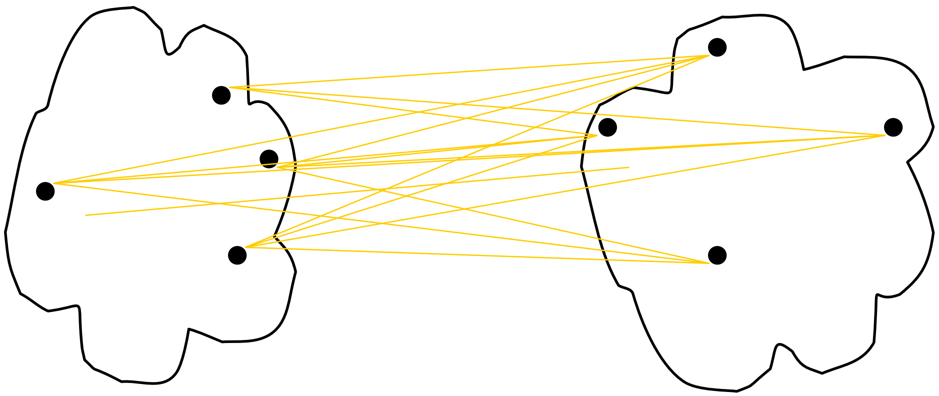
Cohesion is within cluster sum of squares (SSE) if we let proximity to be squared Euclidean distance

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

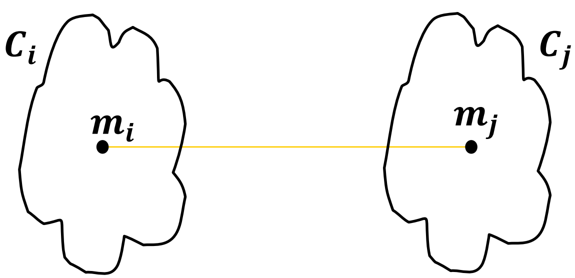
where $\mathbf{C_i}$ *is the cluster i ,* $\mathbf{m_i}$ *is the mean of cluster i*

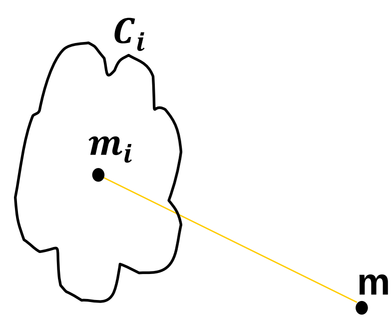
Internal Measures: Separation

Separation: Measure how distinct or well-separated a cluster is from other clusters



$$\text{Separation } (C_i, C_j) = \sum_{x \in C_i, y \in C_j} proximity(x, y)$$





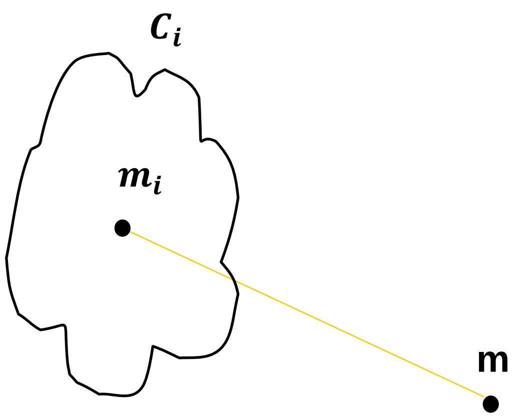
$$\text{Separation } (C_i, C_j) = proximity(m_i, m_j)$$

$$\text{Separation } (C_i) = proximity(m_i, m)$$

where **|C_i|** *is the size of cluster i ,* **m_i** *is the mean of cluster i and* **m** *is the overall mean of all data points*

Internal Measures: Separation

Separation: Measure how distinct or well-separated a cluster is from other clusters



Separation (C_i)= $proximity(C_i, C_j)$

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

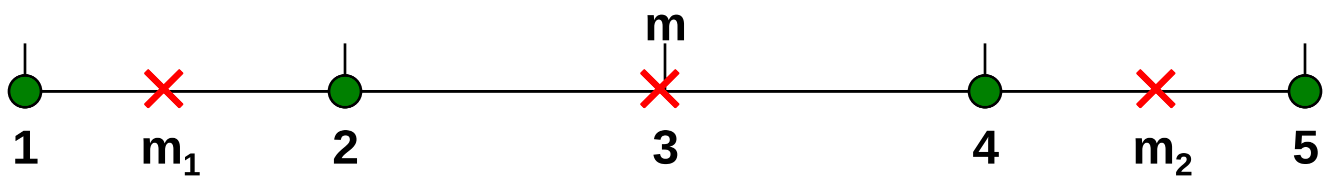
Separation is measured by the between cluster sum of squares

where $|C_i|$ *is the size of cluster i ,* m_i *is the mean of cluster i and* m *is the overall mean of all data points*

Internal Measures: Cohesion and Separation

- Example: SSE = WSS(Cohesion) +BSS(separation) = constant

$$\boxed{WSS = \sum_i \sum_{x \in C_i} (x - m_i)^2} \quad \boxed{BSS = \sum_i |C_i| (m - m_i)^2}$$



K=1 cluster:

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

$$BSS = 4 \times (3-3)^2 = 0$$

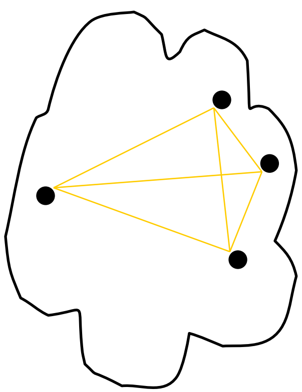
K=2 clusters:

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

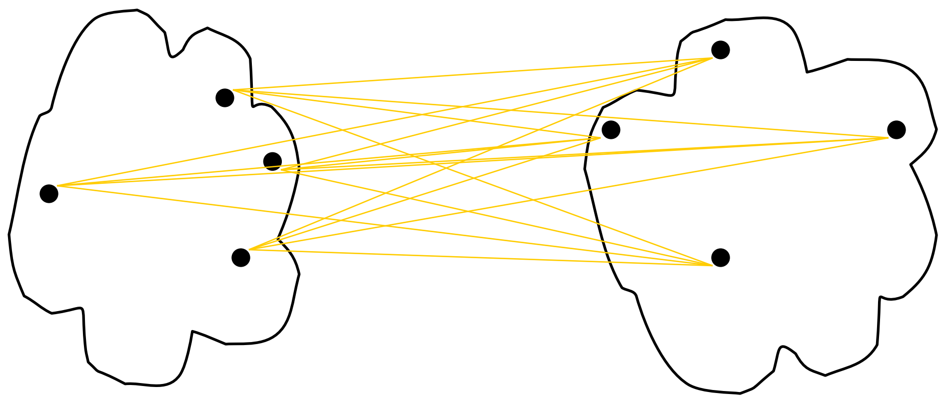
$$BSS = 2 \times (3-1.5)^2 + 2 \times (3-4.5)^2 = 9$$

Internal Measures: Combine Cohesion & Separation

- We can combine the idea of 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

Internal Measures: Silhouette Coefficient

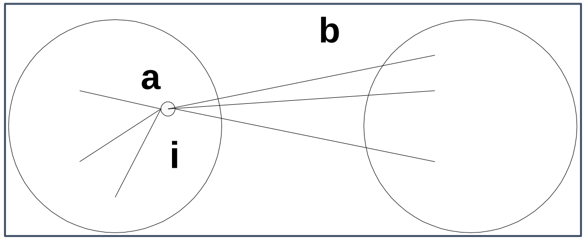
- **Silhouette Coefficient combines ideas of cohesion & separation**
- Silhouette Coefficient for an individual point, *i*
  - Calculate ***a*** = average distance of *i* to the points in its cluster
  - Calculate ***b*** = min (average distance of *i* to points in another cluster)

The silhouette coefficient for a point is

$$s = 1 - a/b \quad \text{if } a < b$$

(or $s = b/a - 1$ if $a \geq b$, not the usual case)

Typically between 0 and 1.
The closer to 1 the better.



Average Silhouette width for a cluster is the average of silhouette coefficients of points in the cluster