Unsupervised and Unstructured Machine Learning

BA820 – Mohannad Elhamod

BOSTON UNIVERSITY

Boston University Questrom School of Business

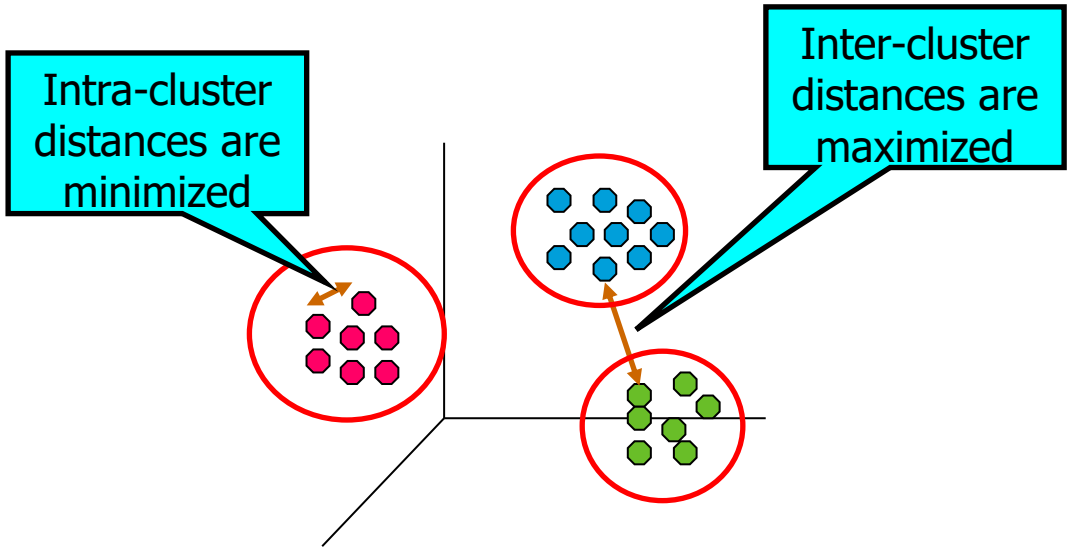
Intro to Clustering

BOSTON UNIVERSITY

Boston University Questrom School of Business

What is Cluster Analysis?

- Placing objects in groups such that:
  - the objects in a group are similar (or related) to one another.
  - They are different from (or unrelated to) the objects in other groups.
- We need a (metric/measure/objective function) to measure the (distance/similarity) of the (objects/clusters).



Clusters are in the eye of the beholder

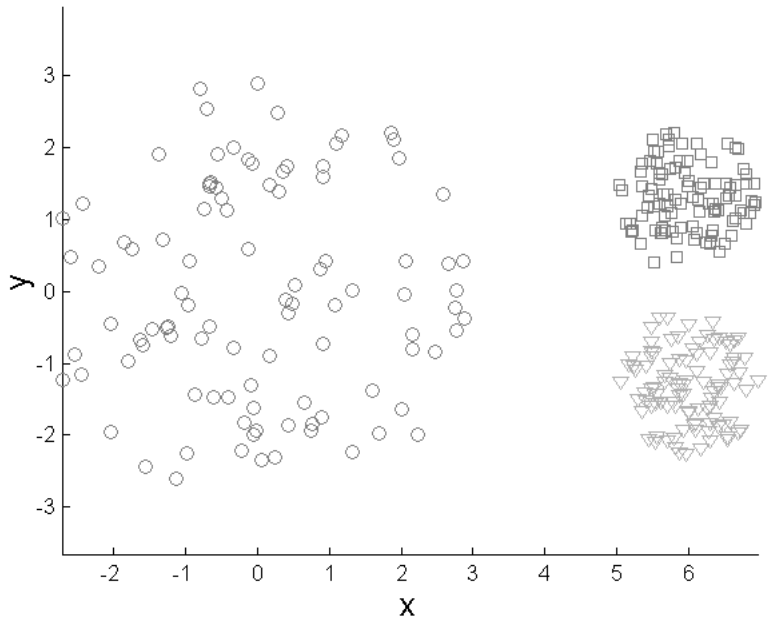
How many clusters?

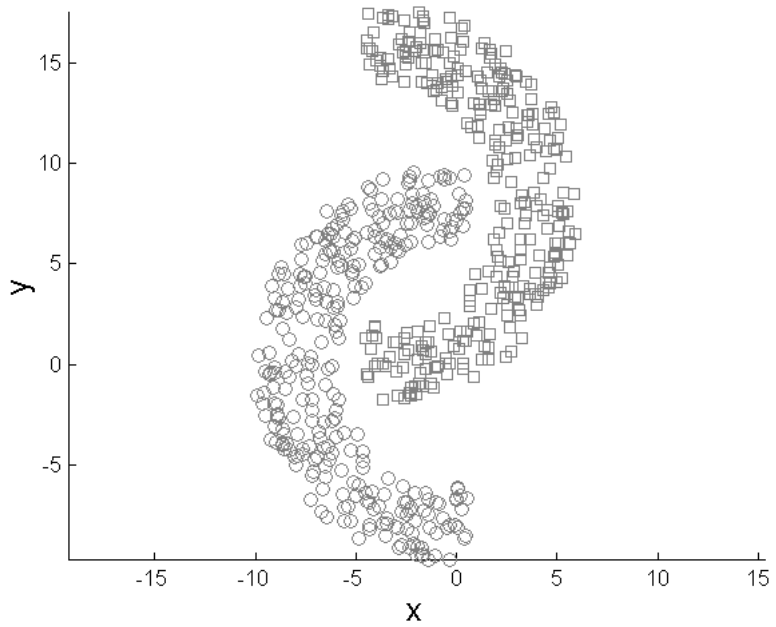
Six Clusters

Two Clusters

Four Clusters

Clusters come in all shapes and sizes



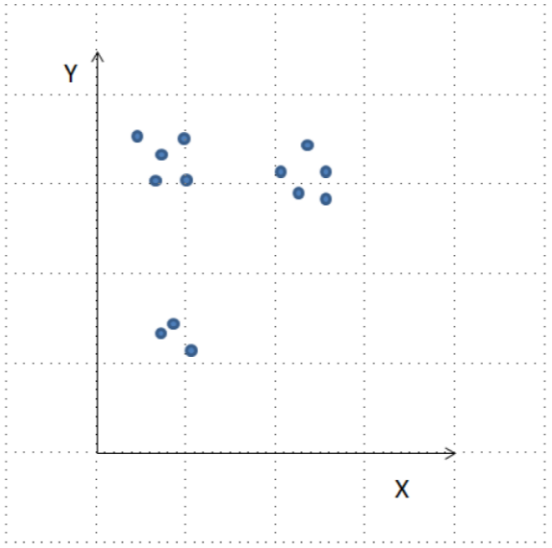


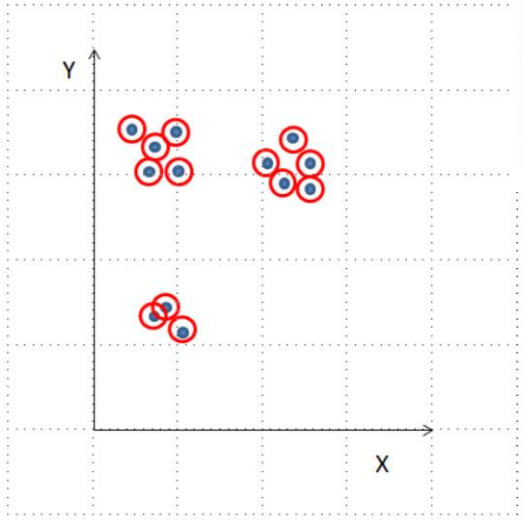
Hierarchical Clustering

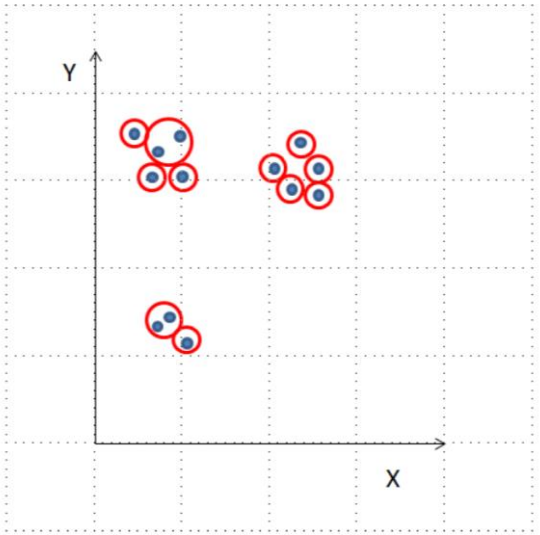
BOSTON UNIVERSITY

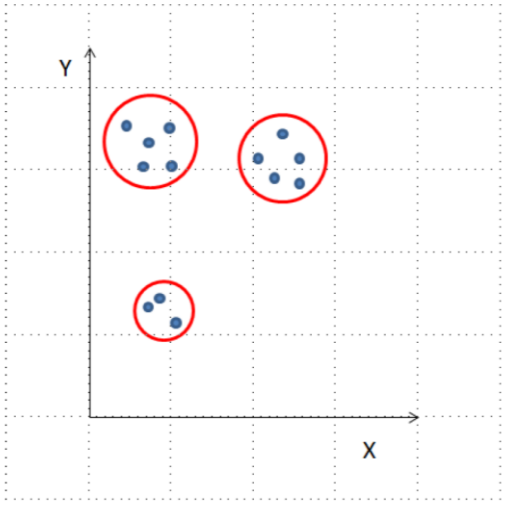
Boston University Questrom School of Business

Hierarchical Clustering





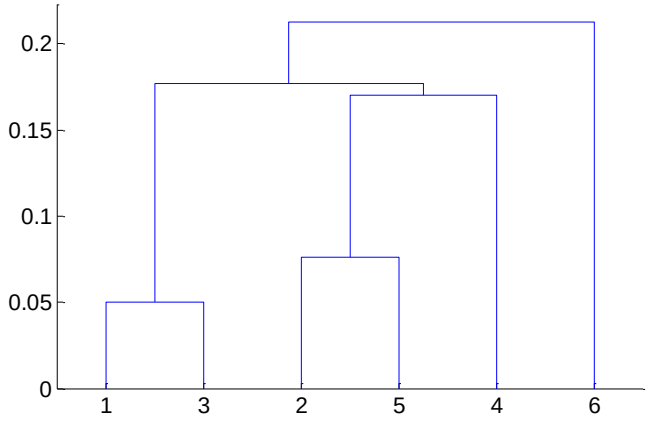


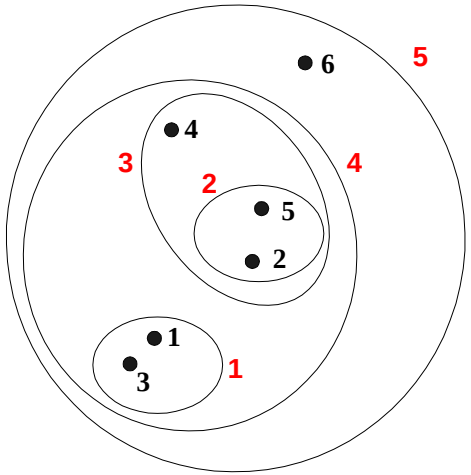


datacamp CLUSTER ANALYSIS IN PYTHON

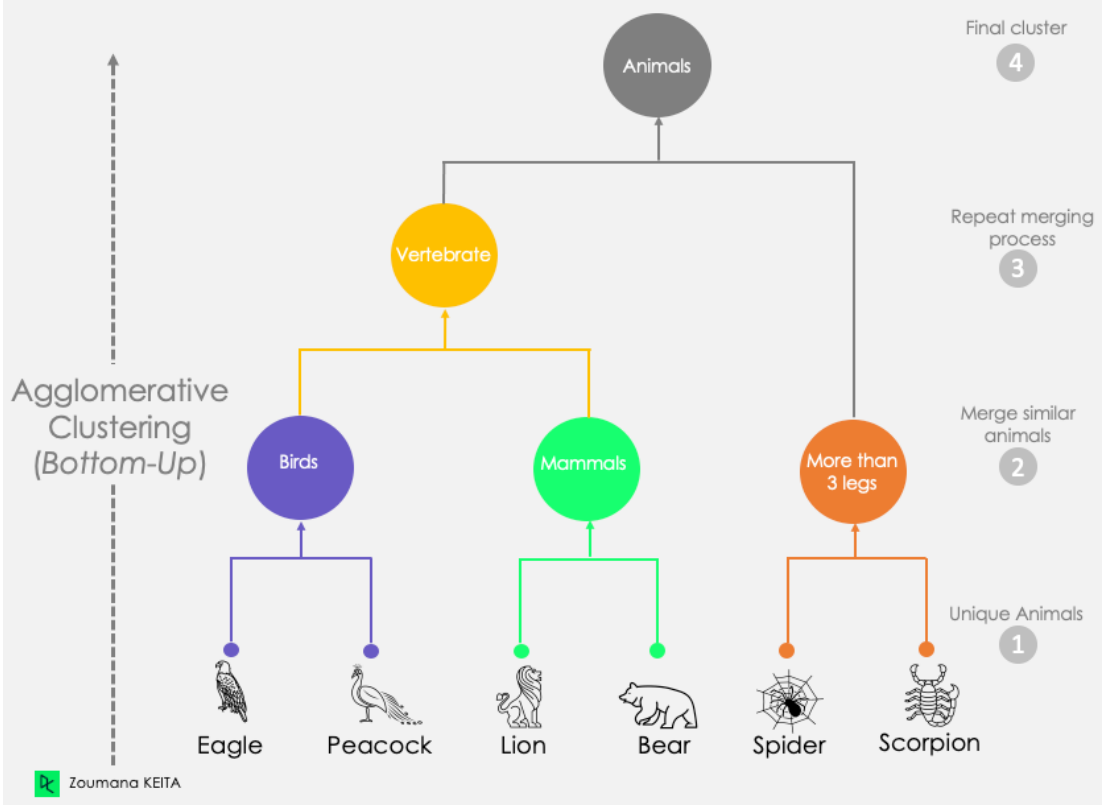
Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree
- Can be visualized as a dendrogram
  - A tree like diagram that records the sequences of merges or splits.





Hierarchical Clustering

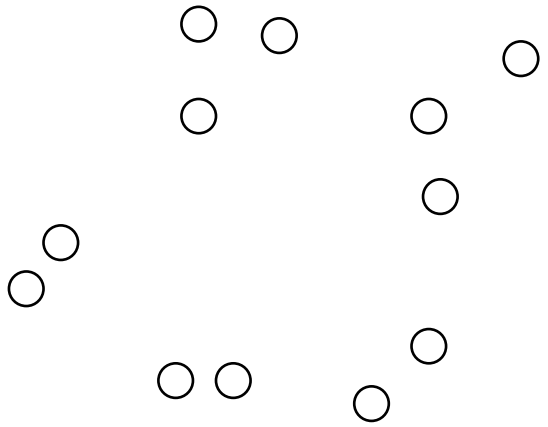


Hierarchical Clustering

- **Key Idea: Successively merge closest clusters**
- Basic algorithm
  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 to defining the distance between clusters distinguish the different algorithms

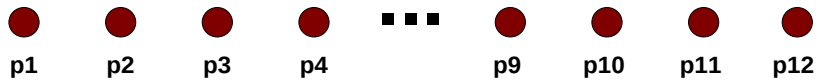
Steps 1 and 2

- Start with clusters of individual points and a proximity matrix



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

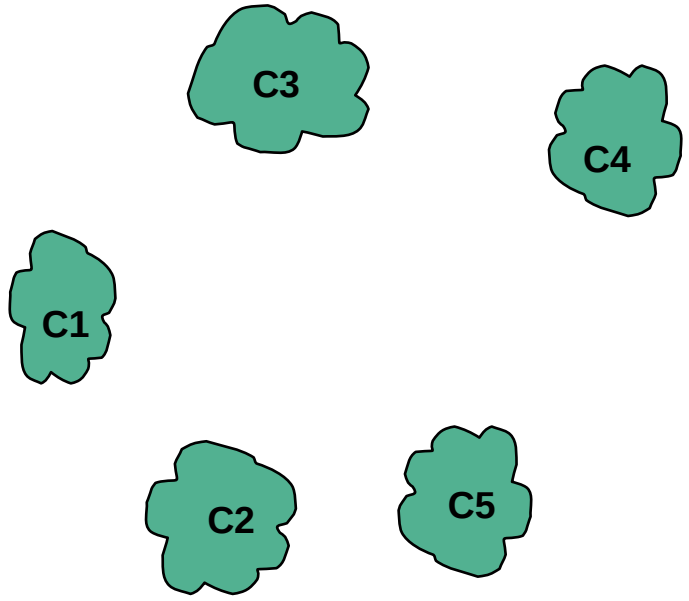
Proximity Matrix

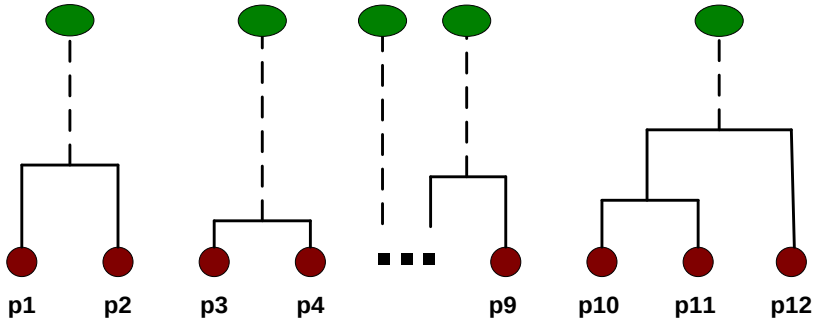


Intermediate Situation

After some merging steps, we have some clusters

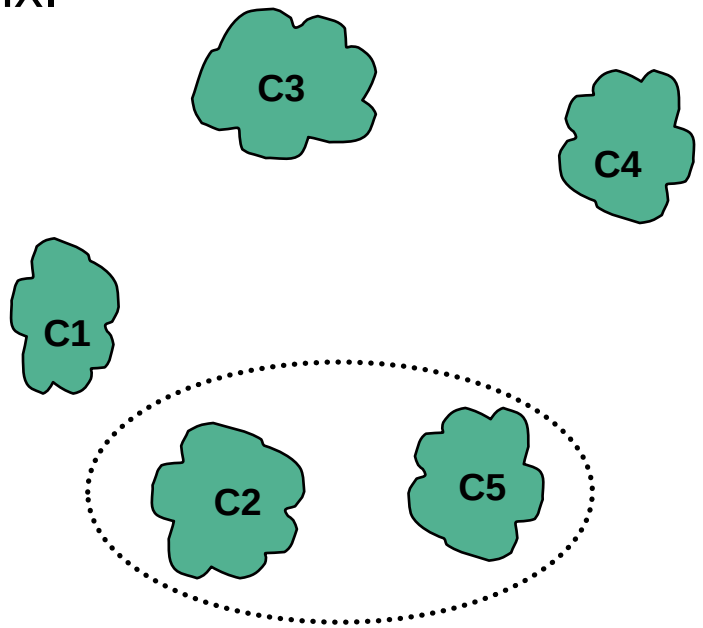
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

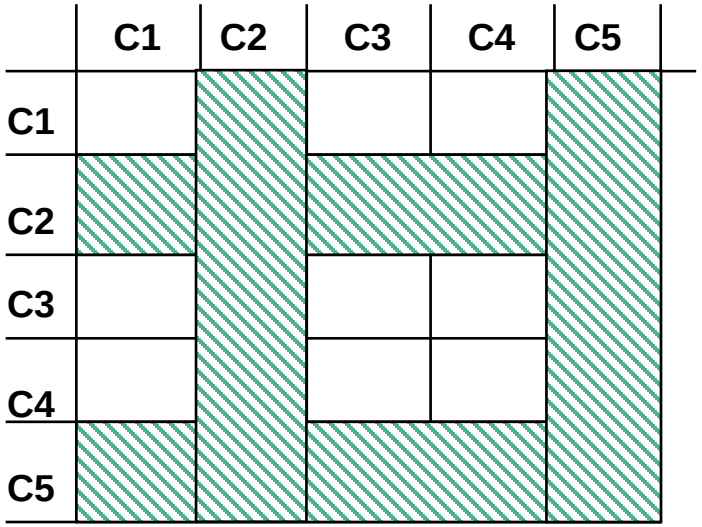




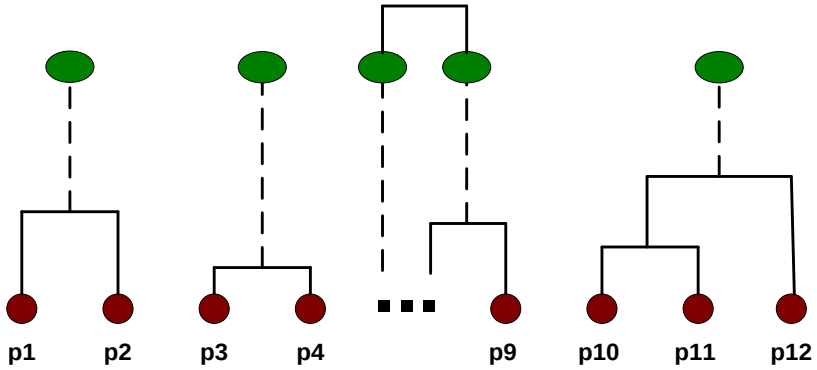
Step 4

We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



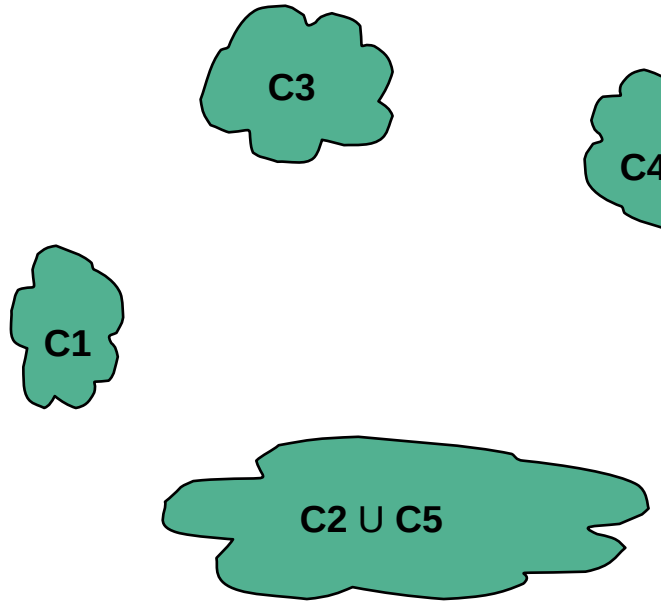


Proximity Matrix



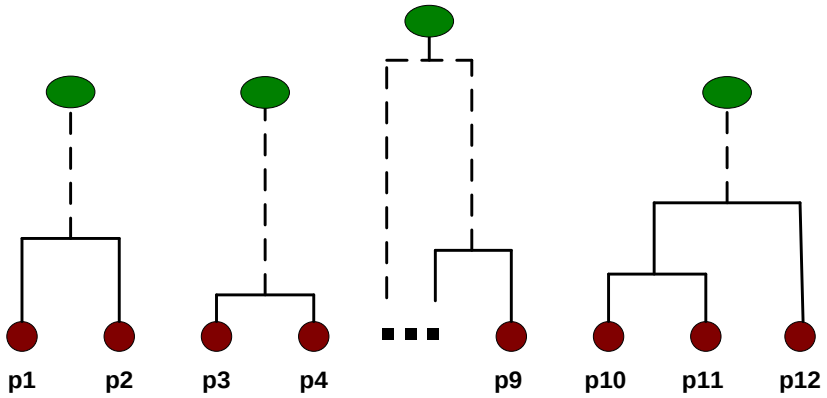
Step 5

The question is “How do we update the proximity matrix?”



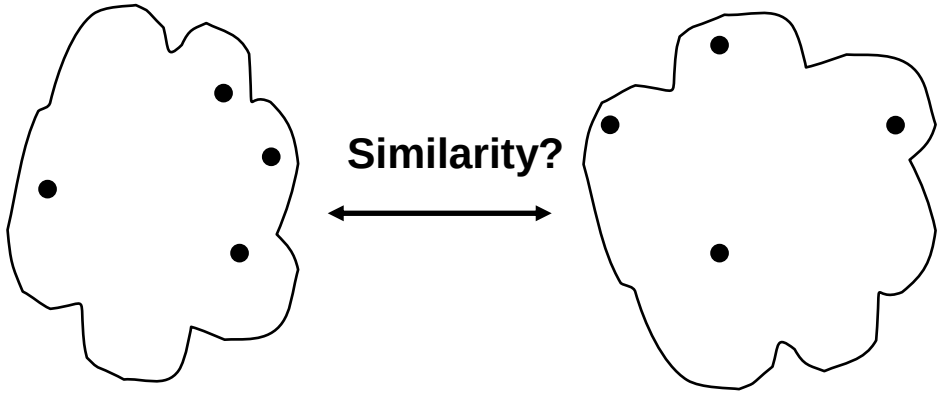
	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



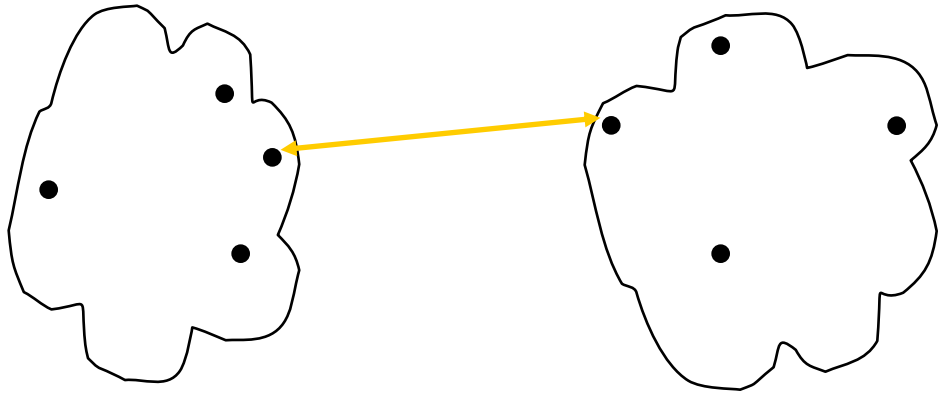
How to Define Inter-Cluster Distance

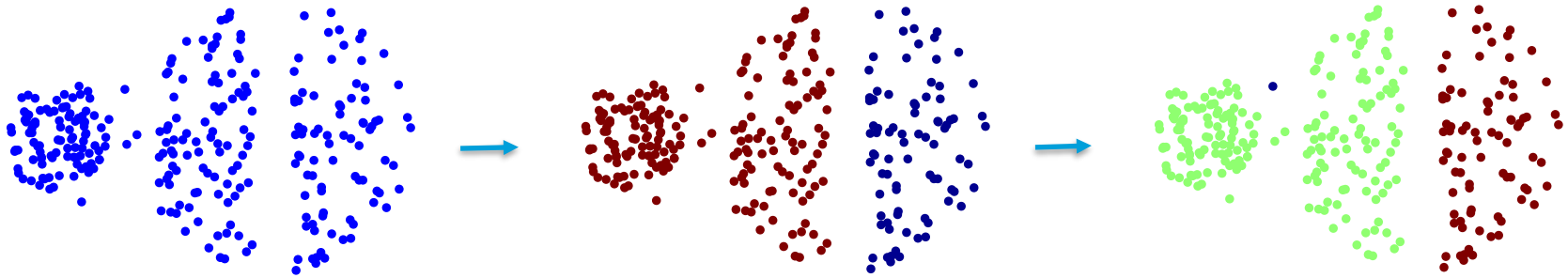
- MIN (Single Link)
- MAX (Complete Linkage)
- Group Average
  - Ward's Method uses squared error
- Distance Between Centroids



How to Define Inter-Cluster Similarity

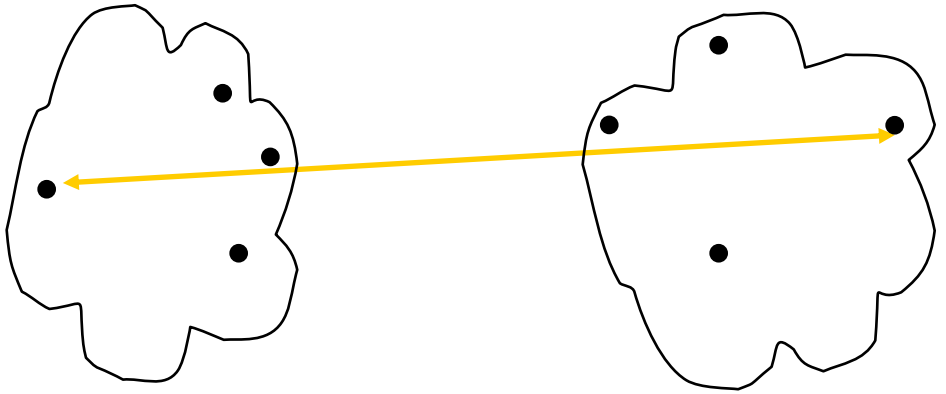
- MIN (Single Link)
  - Sensitive to noise.

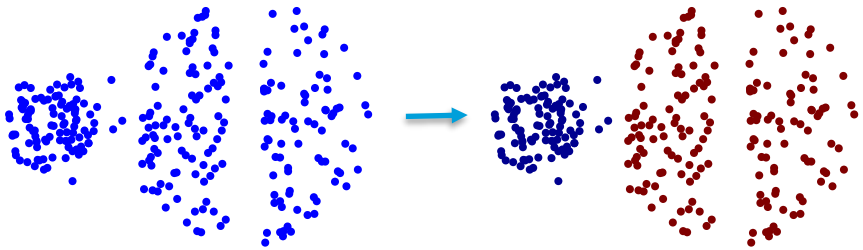


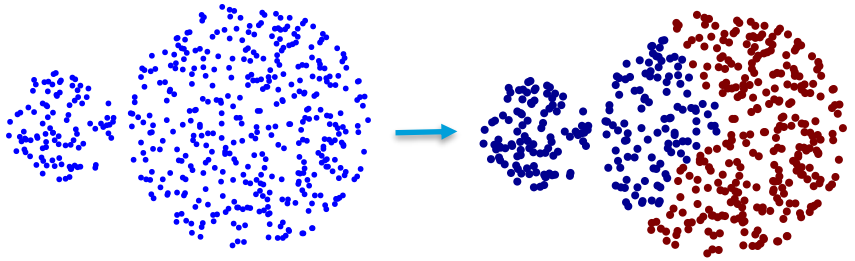


How to Define Inter-Cluster Similarity

- MAX (Complete Linkage)
  - Less susceptible to noise.
  - Breaks larger clusters.

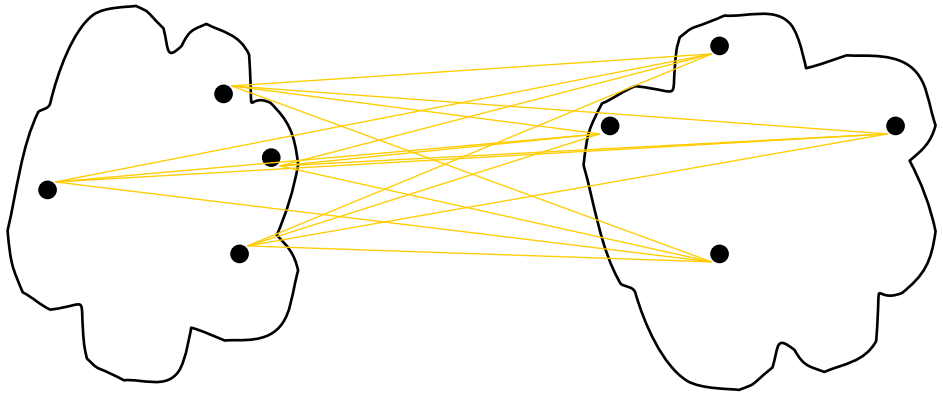






How to Define Inter-Cluster Similarity

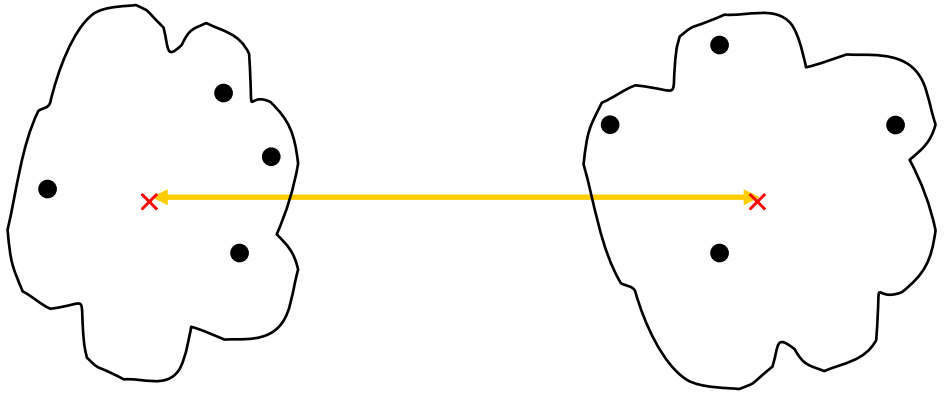
- Group Average
  - Middle ground between MIN and MAX.
  - If square distance is used, it is called Ward method.



$$\text{proximity}(\text{Cluster}_i, \text{Cluster}_j) = \frac{\sum_{\substack{p_i \in \text{Cluster}_i \\ p_j \in \text{Cluster}_j}} \text{proximity}(p_i, p_j)}{|\text{Cluster}_i| \times |\text{Cluster}_j|}$$

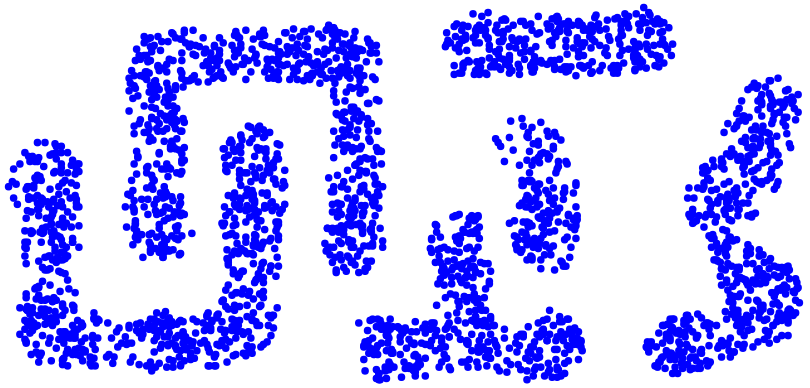
How to Define Inter-Cluster Similarity

- Distance Between Centroids

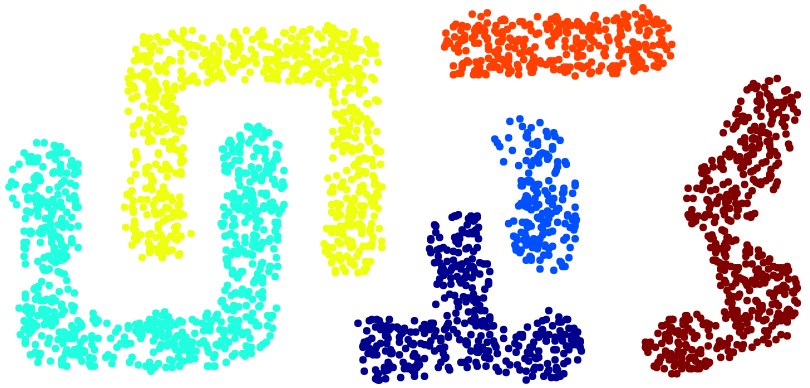


Hierarchical Clustering

Can handle non-elliptical shapes



Original Points



Six Clusters

Hierarchical Clustering

- Visualization + dendograms (easy to find answer when changing number of clusters)
- Hierarchical has a high time complexity (polynomial $O(n^3)$).

Demo time!