巨量資料與統計分析

Fall 2024

授課教師：統計系余清祥

日期：2024年10月22日

第六週：群聚與分類

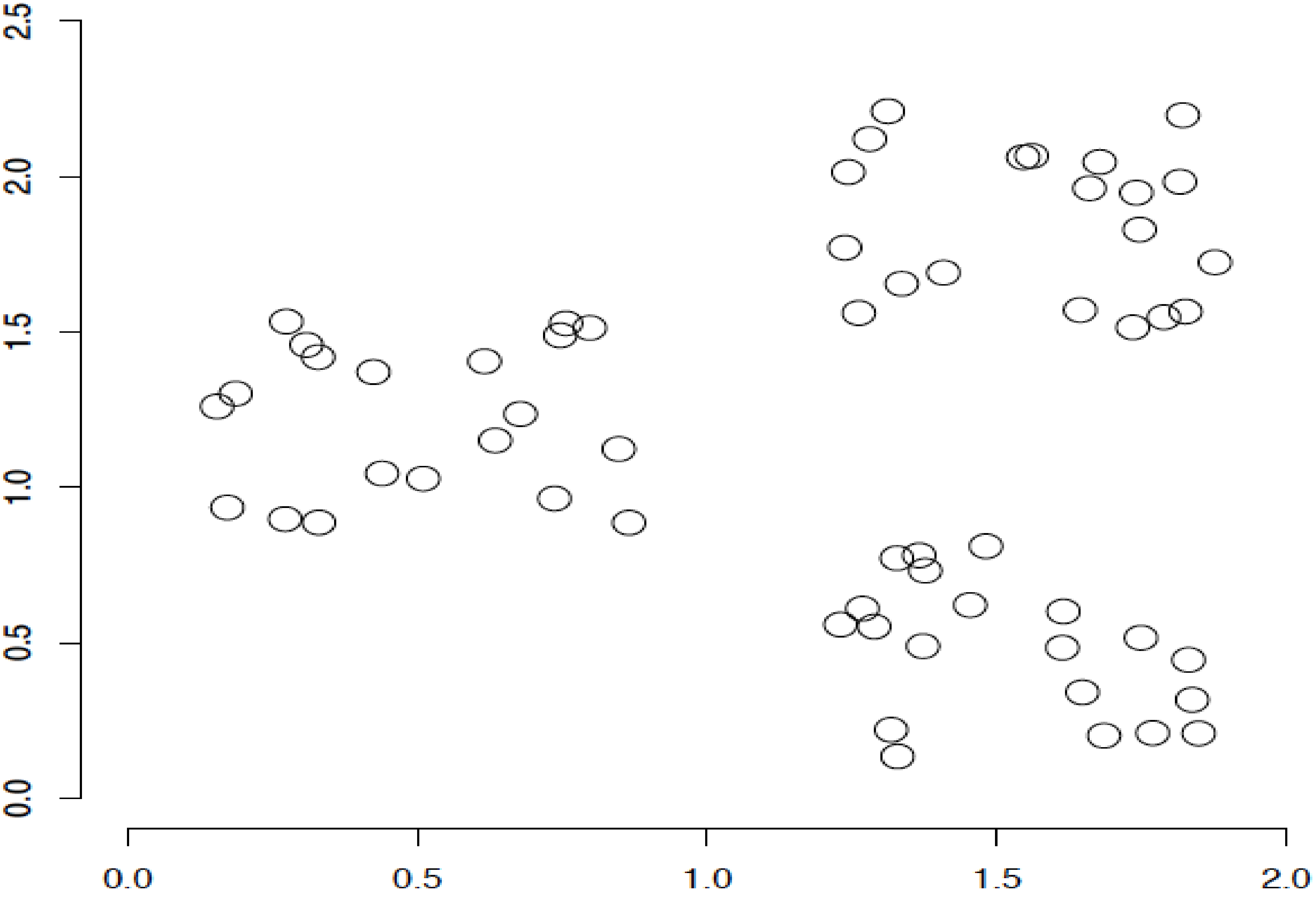
什麼是群集(Clustering)？

- **Clustering**: the process of grouping a set of objects into classes of similar objects

→到同一組文件有類似特性，不同組別的文件特性大不相同。如紅樓夢前八十回、後四十回作者不同，風格應該略有差異。

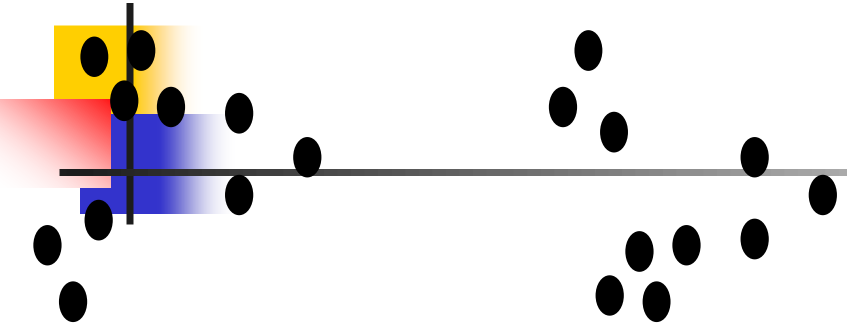
→公司、客戶、產品也可如此區隔。

- 問題：如何定義相似性(Similarity)？如何劃分不同類別的界線？

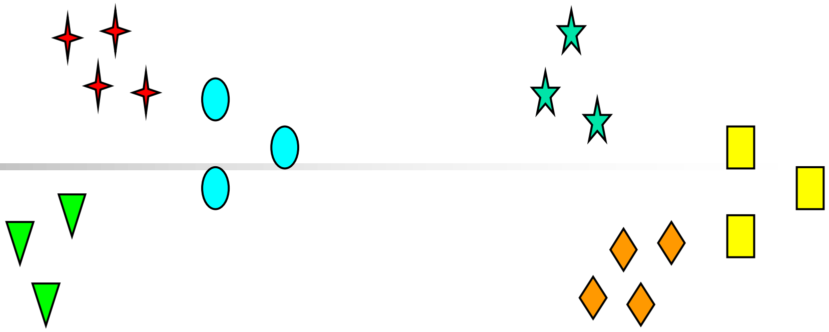


如何區隔群集、總共有幾個群集？

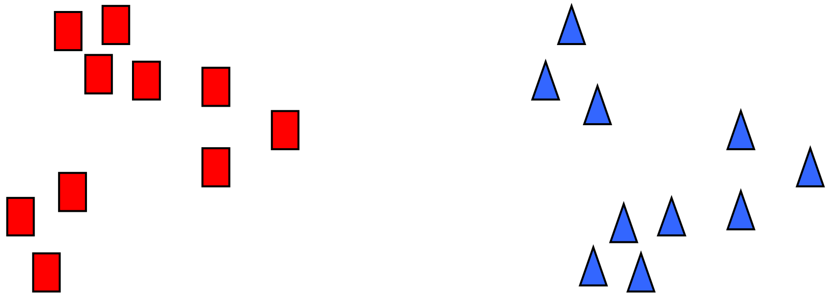
Notion of Cluster can be Ambiguous!!



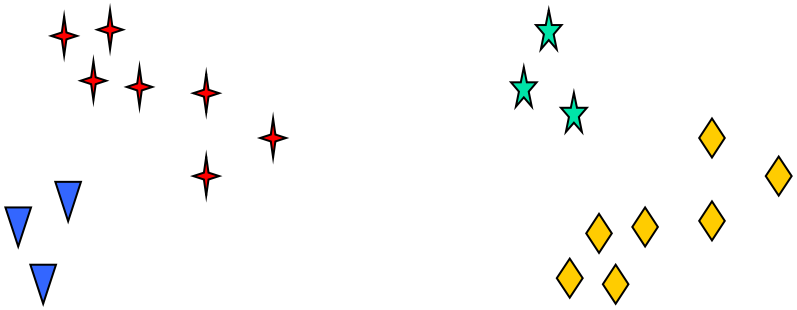
How many clusters?



Six Clusters

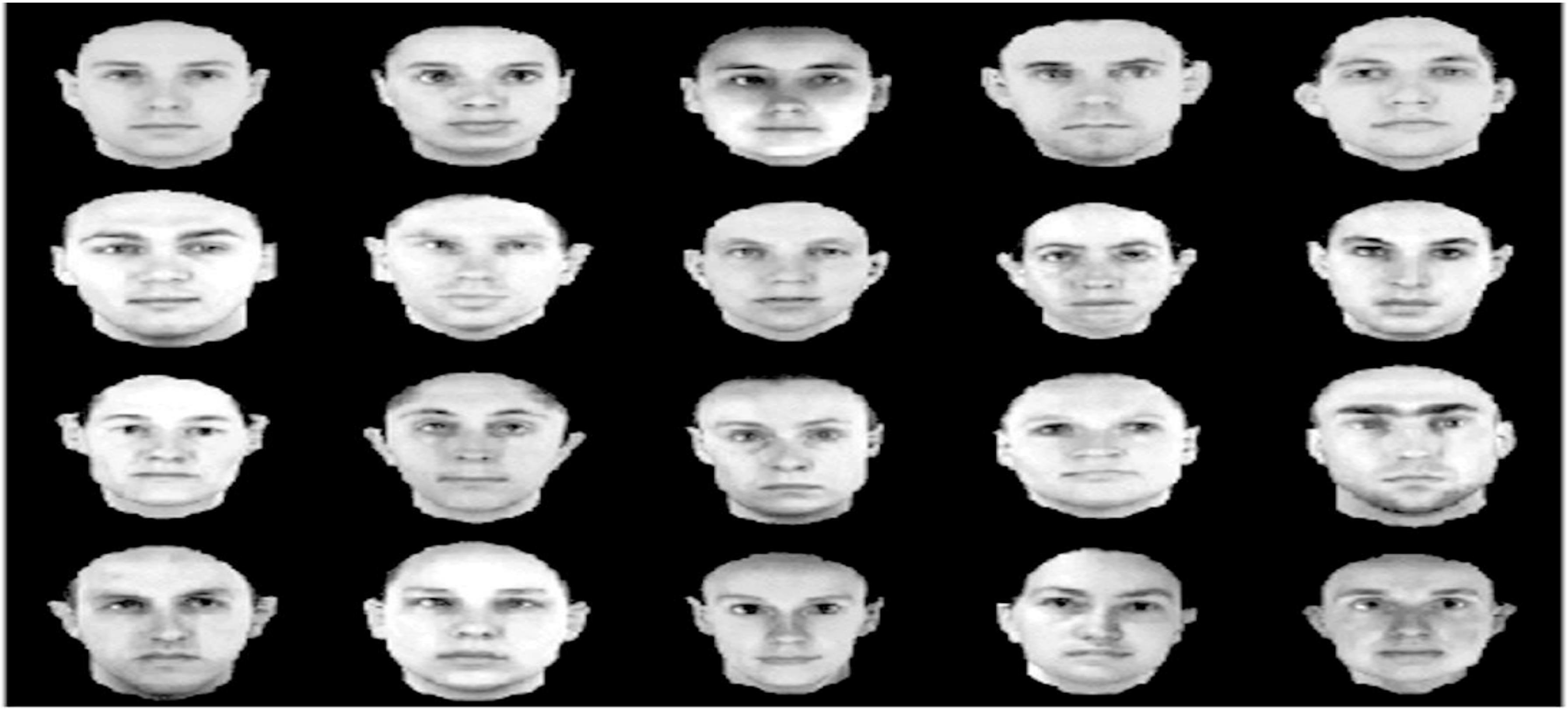


Two Clusters

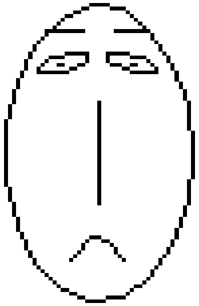


Four Clusters

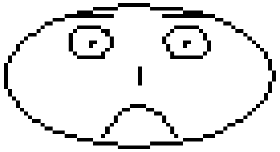
Q：How many groups are there in the following 20 faces?



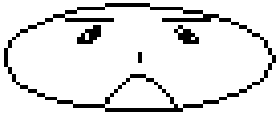
Converting them into Chernoff faces …
→ Which two faces are the most similar?



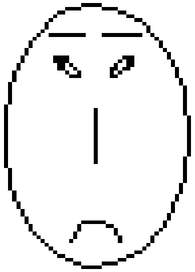
1



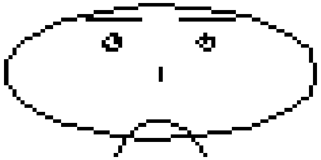
2



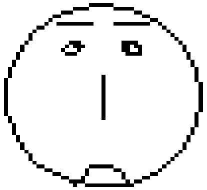
3



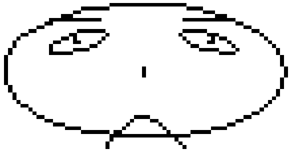
4



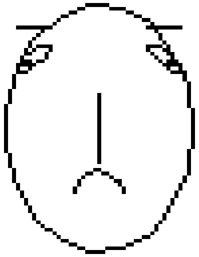
5



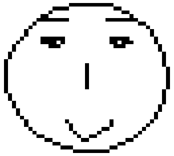
6



7



8



9

Applications of clustering

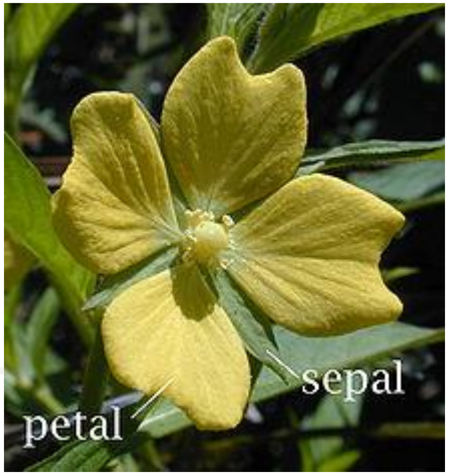
- Pattern Recognition
- Spatial Data Analysis
  - Create thematic maps in GIS by clustering feature spaces
  - Detect spatial clusters or for other spatial mining tasks
- Image Processing
- Economic Science (especially market research)
- WWW
  - Document classification
  - Cluster Weblog data to discover groups of similar access patterns

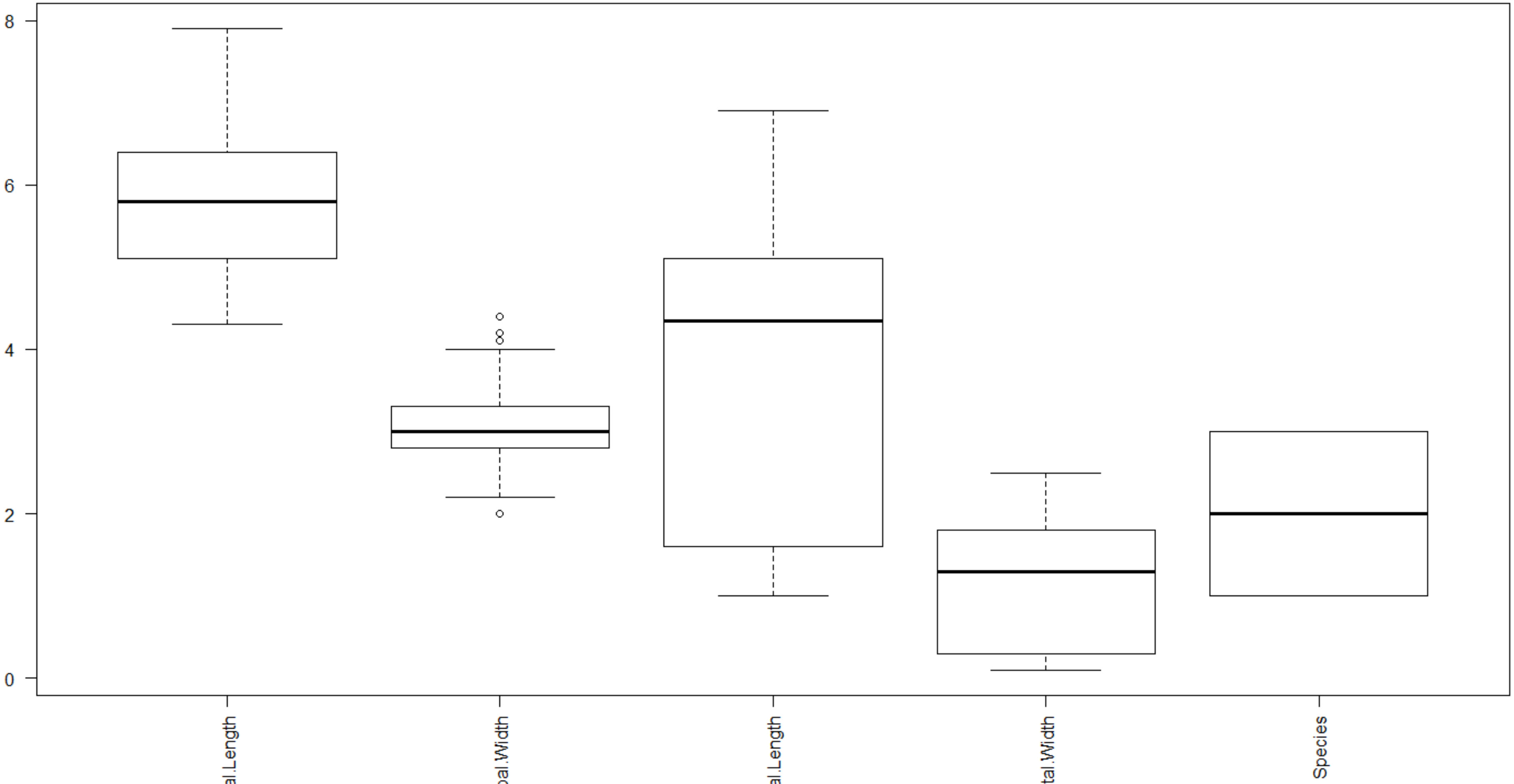
http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.htmlt

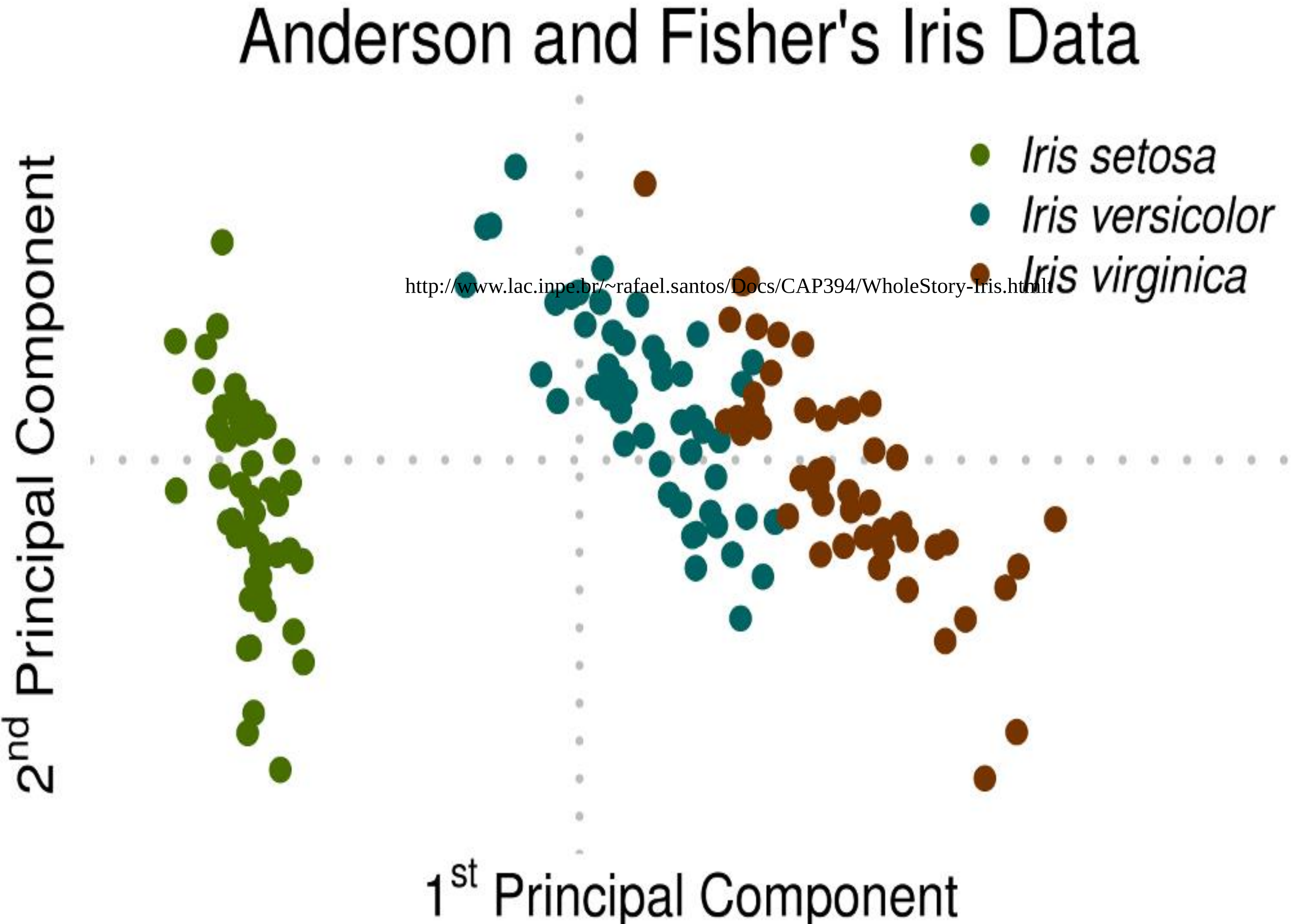
Iris setosa

Iris versicolor

Iris virginica

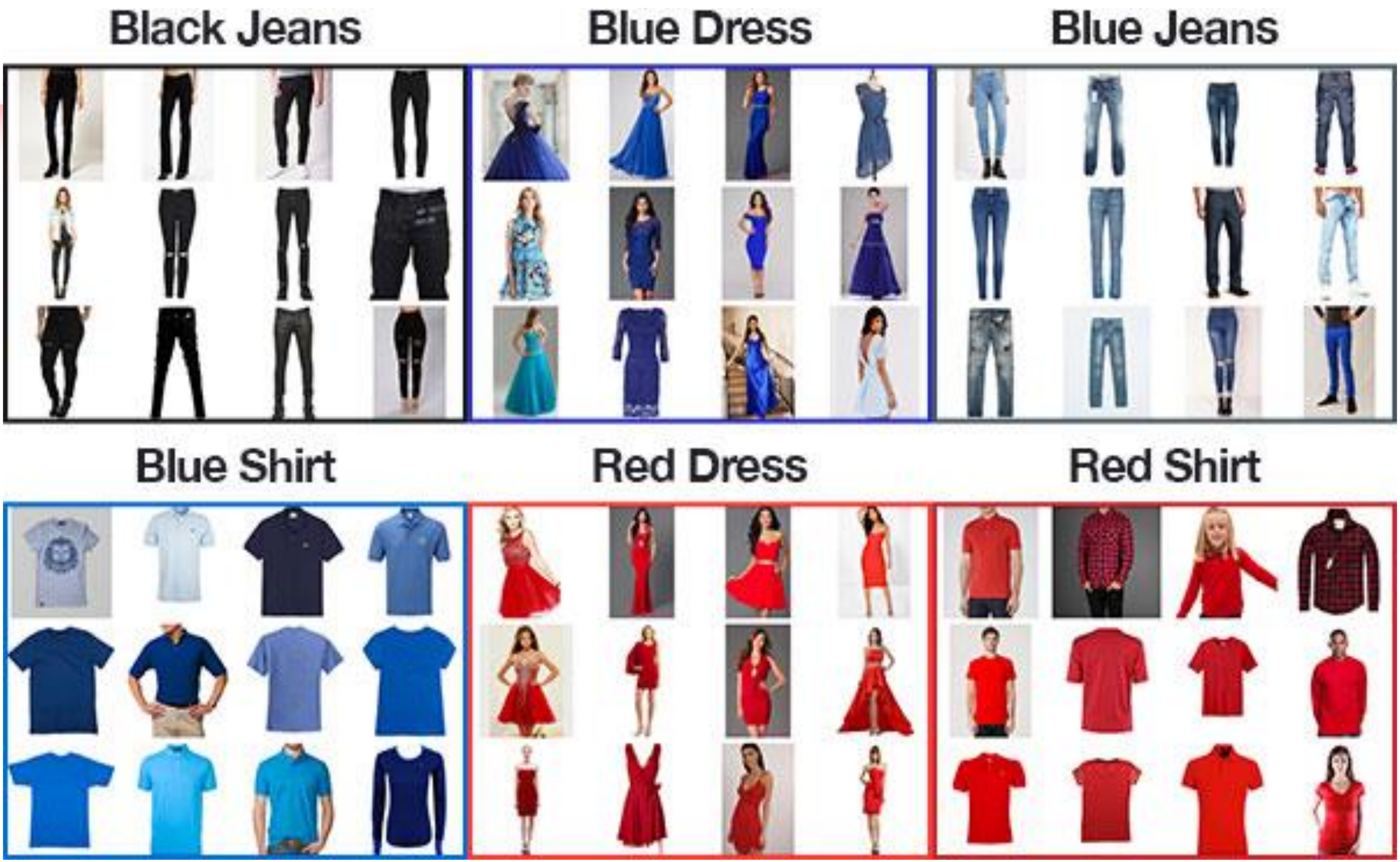






http://www.lac.inpe.br/~rafael.santos/Docs/CAP394/WholeStory-Iris.htmlt

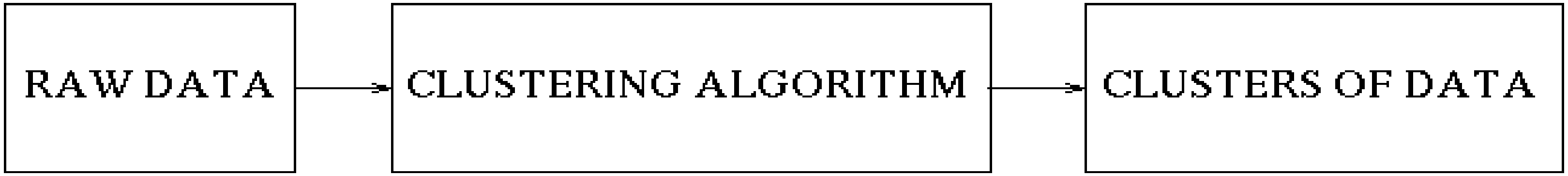
Multi-label classification with Keras



https://pyimagesearch.com/wp-content/uploads/2018/04/keras_multi_label_dataset.jpg

Clustering Algorithms

- A clustering algorithm tries to find natural groups of components based on **similarity** & the **centroid** of a group of data sets. Most algorithms evaluate the **distance** between a point and the cluster centroids. The output from a clustering algorithm is basically a statistical description of the cluster centroids with the number of components in each cluster.



RAW DATA → CLUSTERING ALGORITHM → CLUSTERS OF DATA

Partitioning Clustering Approach

- A typical approach via iteratively partitioning training data set to learn a partition of the given data
- Learning a partition on a data set to produce several non-empty clusters (given the number of clusters)
- In principle, optimal partition achieved via minimising the sum of squared distance to its "representative object" in each cluster

$$E = \Sigma_{k=1}^{K} \Sigma_{\mathbf{x} \in C_k} d^2(\mathbf{x}, \mathbf{m}_k)$$

e.g., Euclidean distance $d^2(\mathbf{x}, \mathbf{m}_k) = \sum_{n=1}^{N} (x_n - m_{kn})^2$

What is K-Means?

- Given a *K*, find a partition of *K clusters* to optimise the chosen partitioning criterion
  - global optimum: exhaustively search all partitions
- The *K-means* algorithm: a heuristic method
  - K-means algorithm (MacQueen'67): each cluster is represented by the centre of the.
  - K-means algorithm is the simplest partitioning method for clustering.

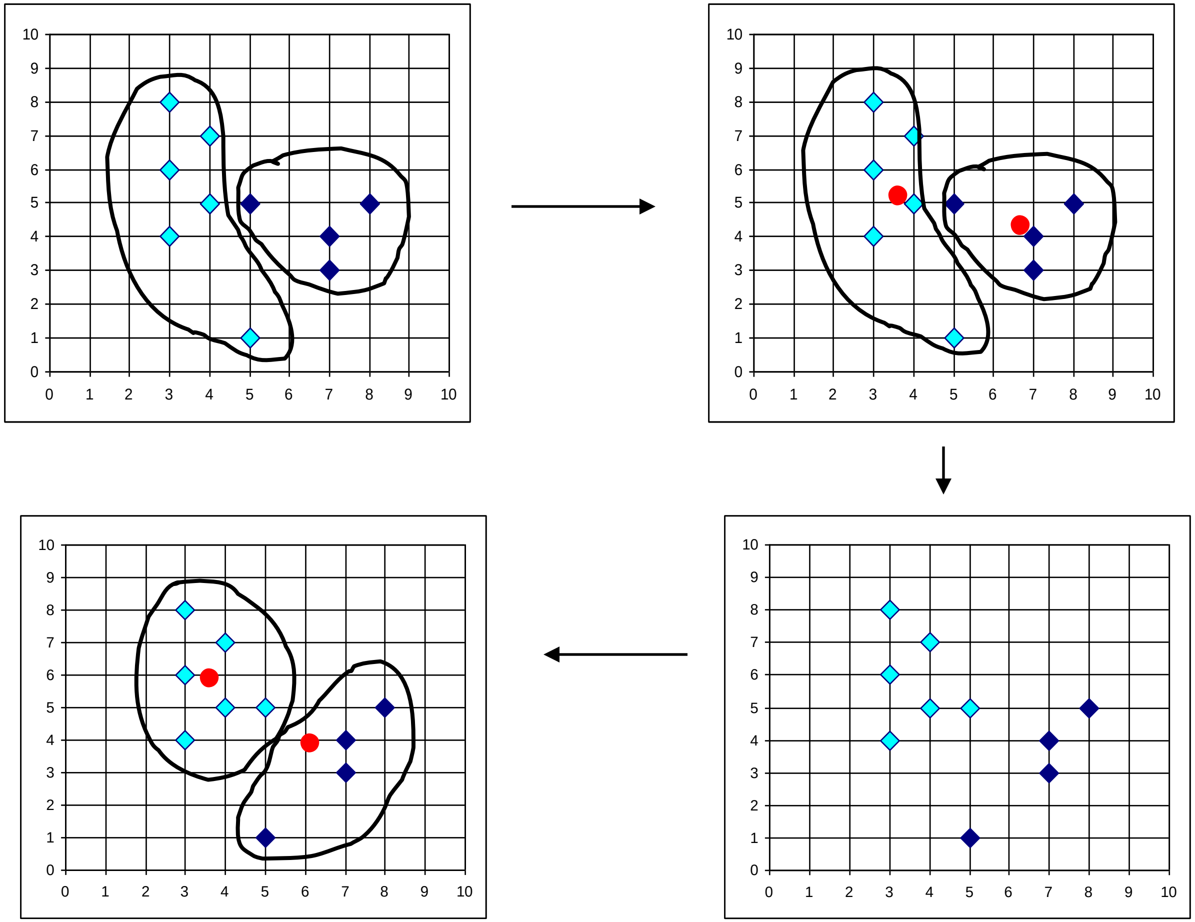
K-means Algorithm

- Given the number *K*, the *K-means* algorithm is carried out in three steps after initialisation:

Initialisation: set seed points (randomly)

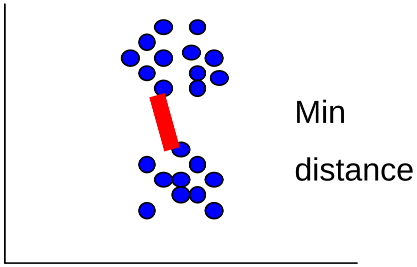
1) Assign each object to the cluster of the nearest seed point measured with a specific distance metric
2) Compute new seed points as the centroids of the clusters of the current partition (the centroid is the centre, i.e., *mean point*, of the cluster)
3) Go back to Step 1), stop when no more new assignment (i.e., membership in each cluster no longer changes)

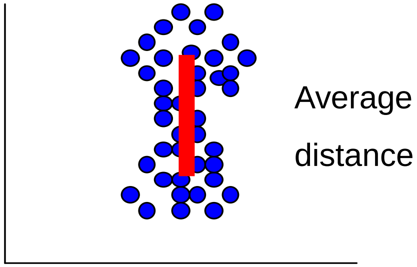
The *K-Means* Clustering Method

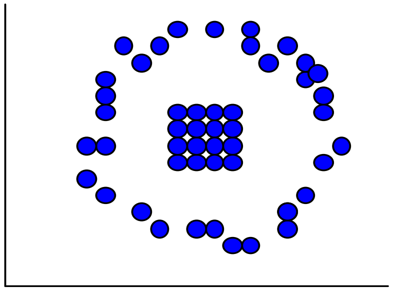


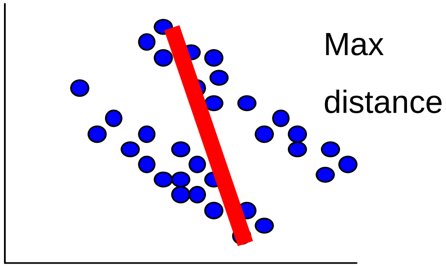
Distance Between Two Clusters

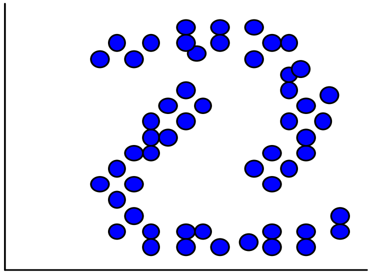
- ❑Single-Link Method / Nearest Neighbor
- ❑Complete-Link / Furthest Neighbor
- ❑Their Centroids.
- ❑Average of all cross-cluster pairs.

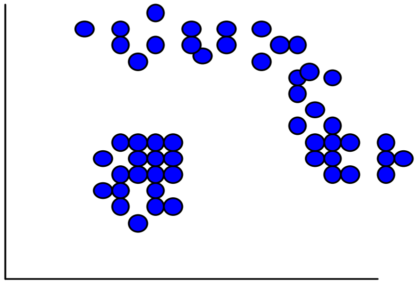






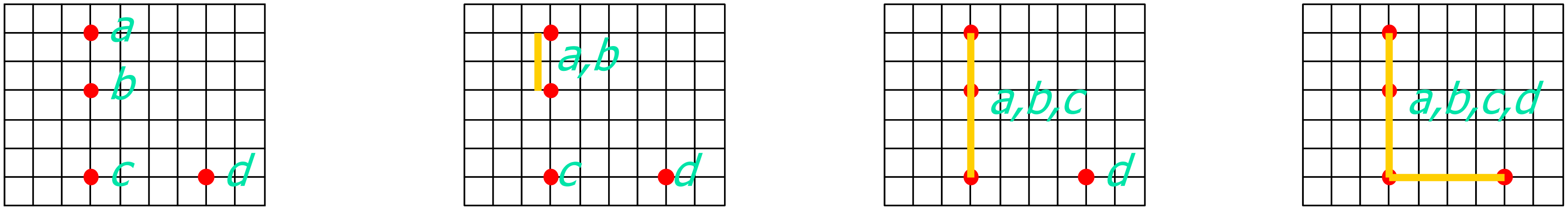






Single-Link Method

Euclidean Distance



(1)

(2)

(3)

	b	**c**	**d**
a	2	5	6
b		3	5
c			4

	b	**c**	**d**
a	2	5	6
b		3	5
c			4

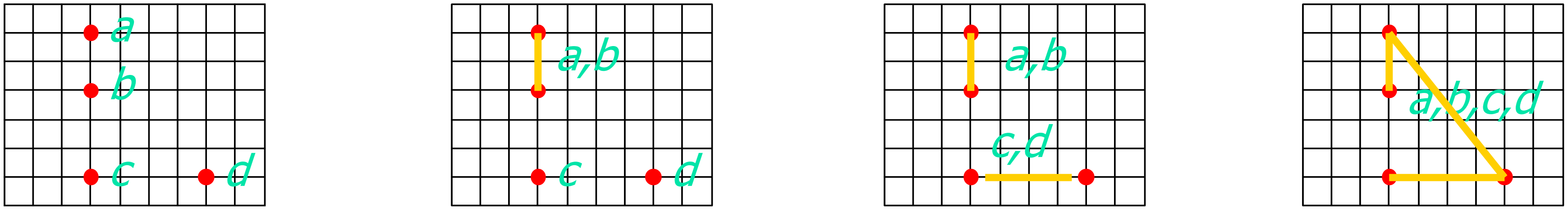
	c	**d**
a,b	3	5
c		4

	d
a,b,c	4

Distance Matrix

Complete-Link Method

Euclidean Distance



(1)

(2)

(3)

	b	c	d
a	2	5	6
b		3	5
c			4

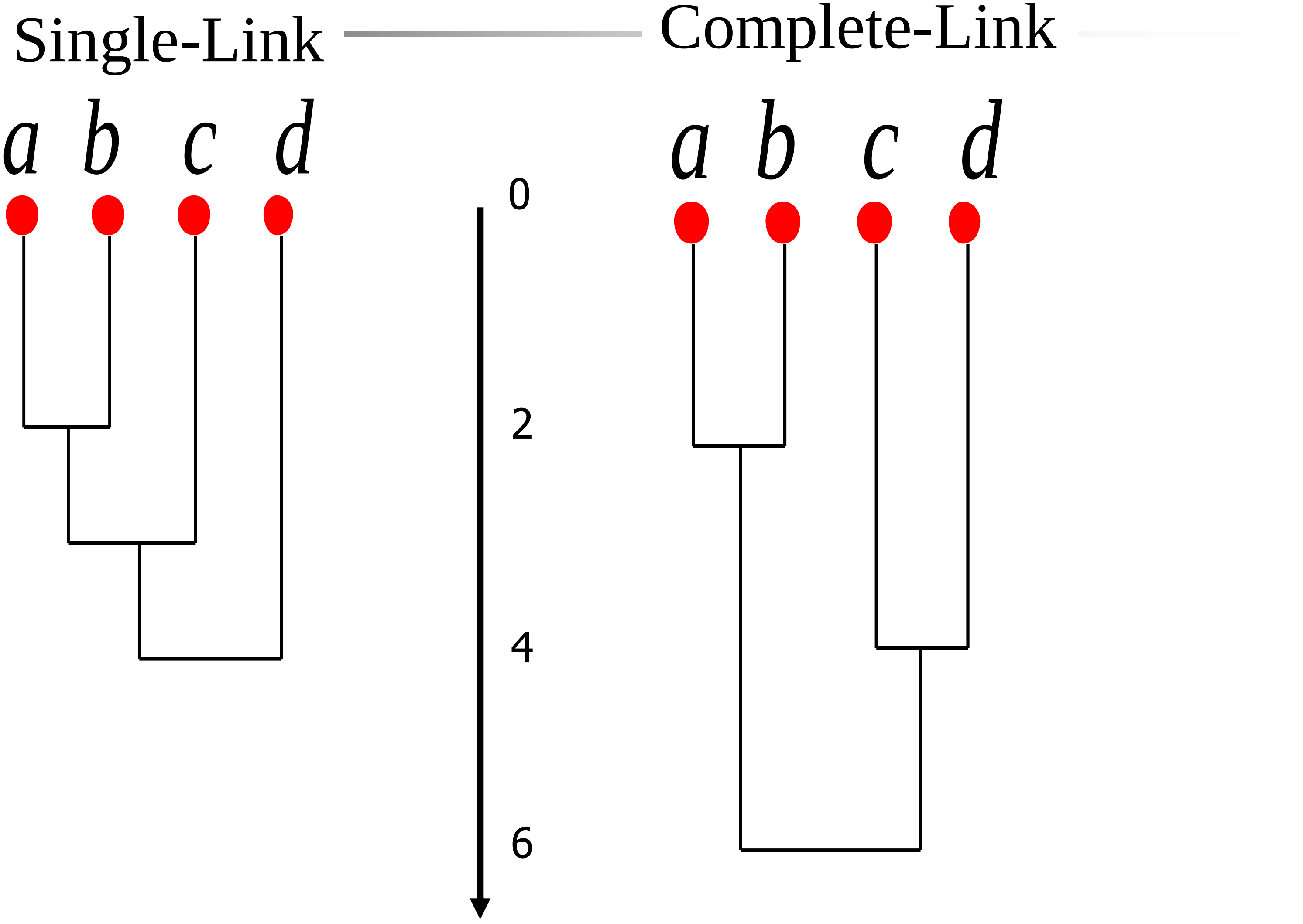
	b	c	d
a	2	5	6
b		3	5
c			4

	c	d
a,b	5	6
c		4

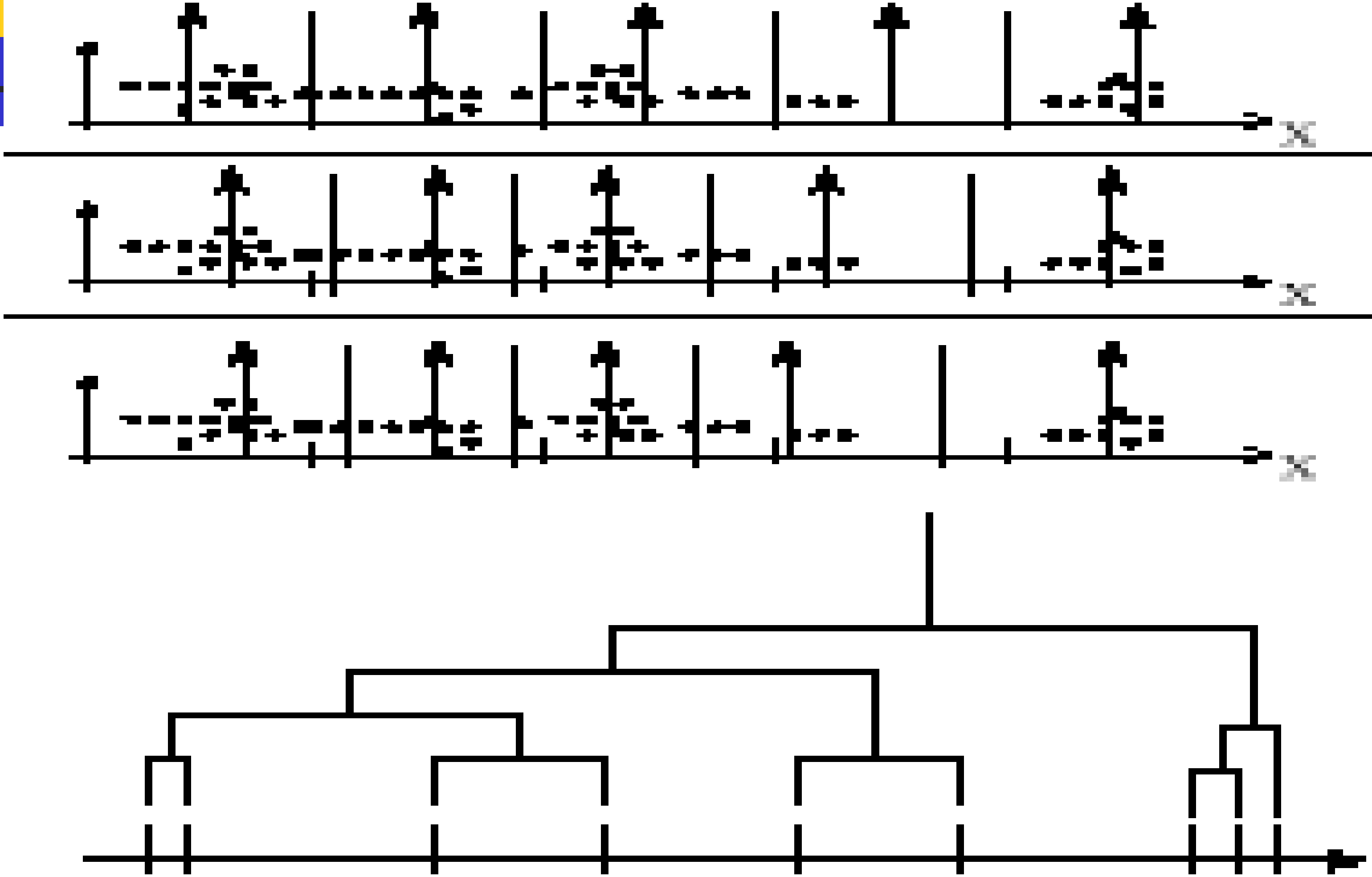
	c,d
a,b	6

Distance Matrix

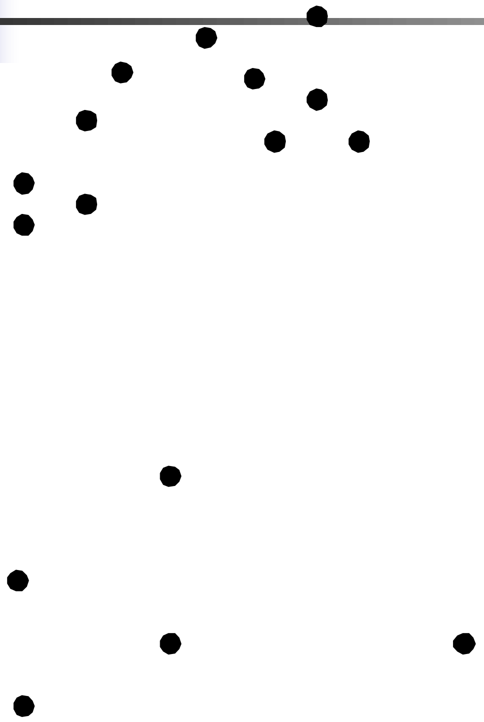
Compare Dendrograms



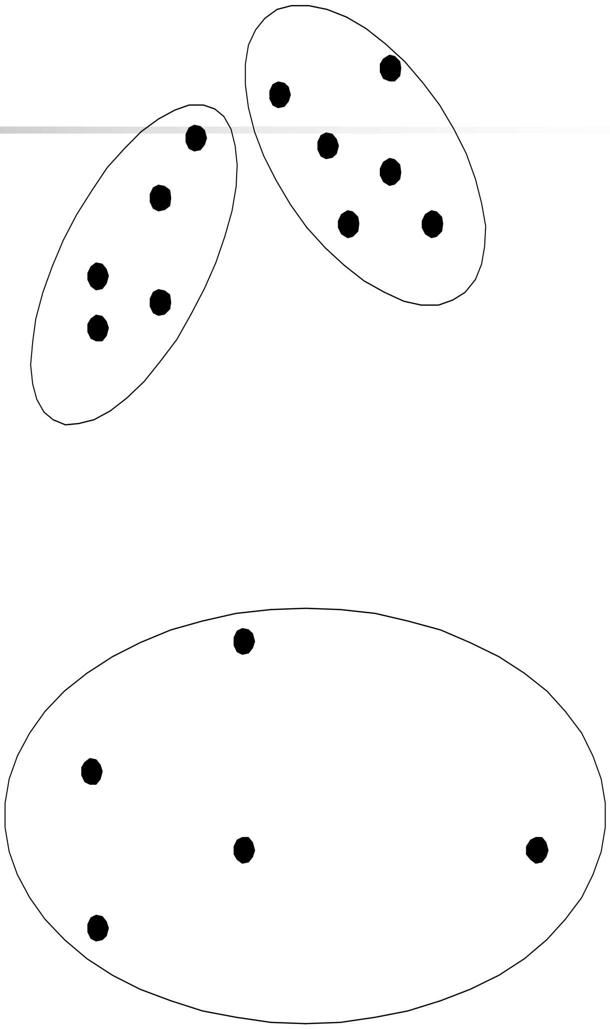
K-Means vs. Hierarchical Clustering



Partitional Clustering

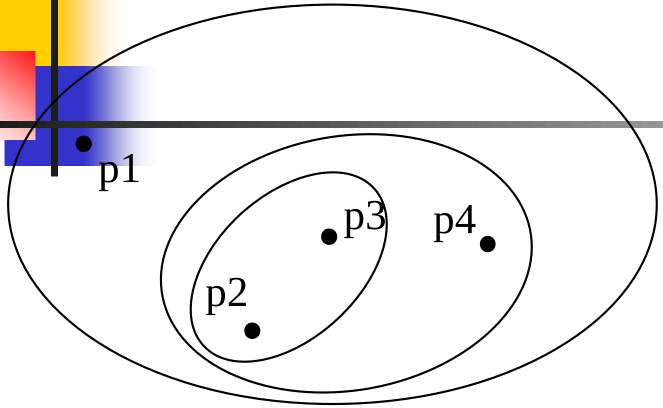


Original Points

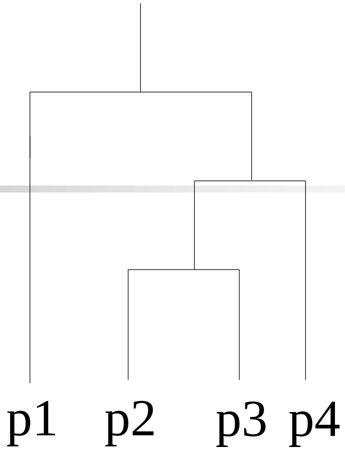


A Partitional Clustering

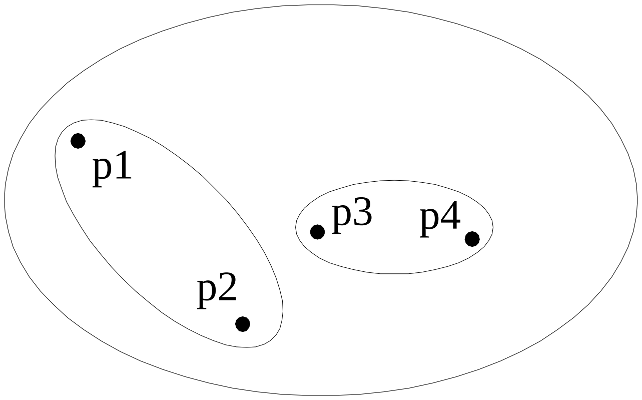
Hierarchical Clustering



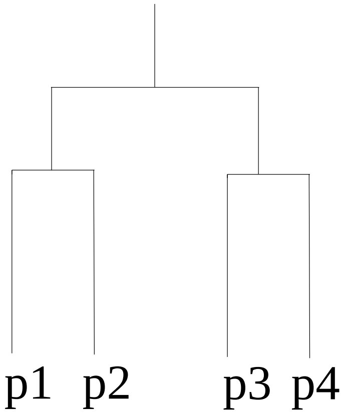
Traditional Hierarchical Clustering



Traditional Dendrogram



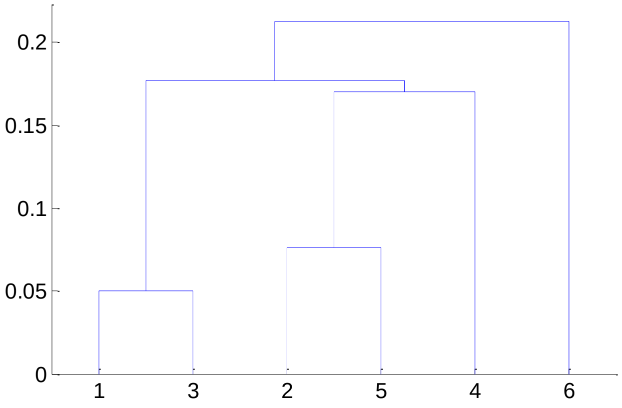
Non-traditional Hierarchical Clustering

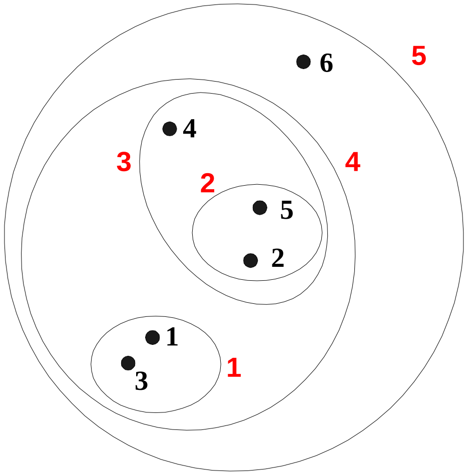


Non-traditional Dendrogram

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

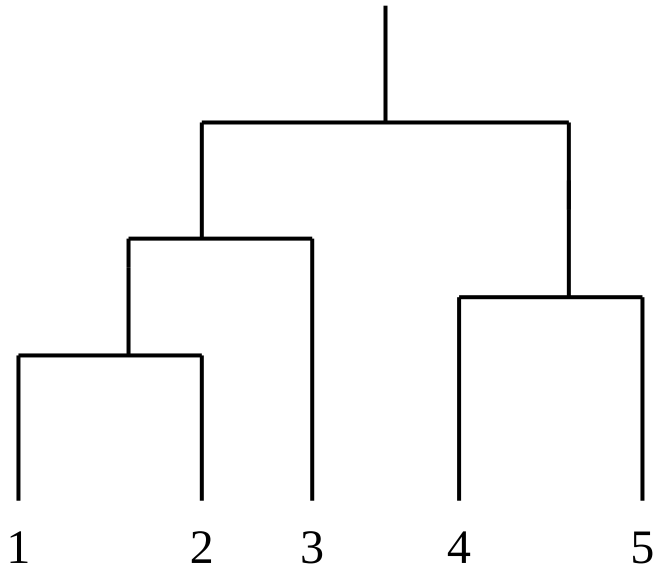




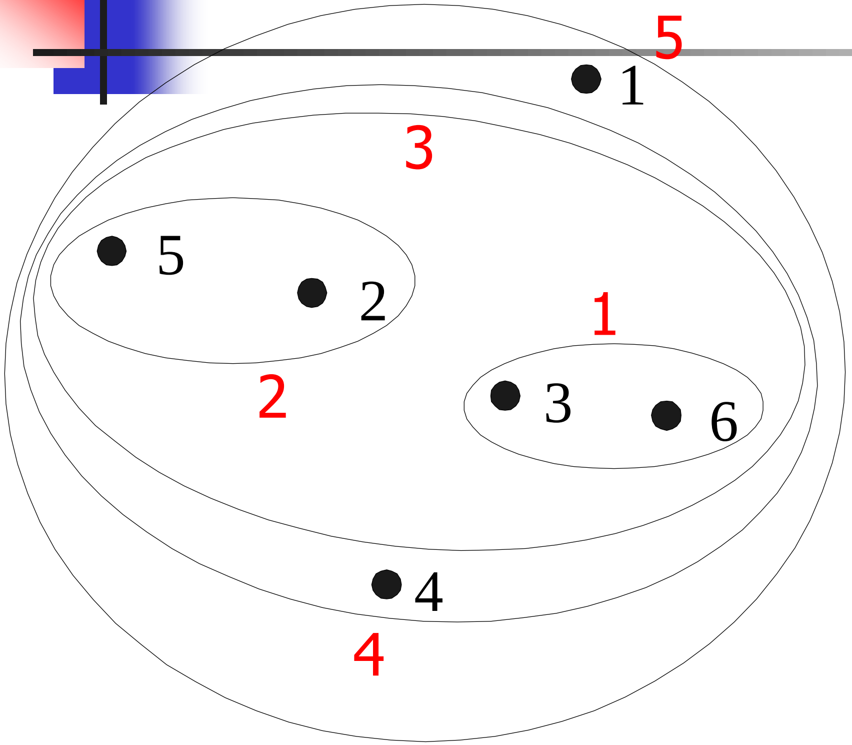
Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two most similar (closest) points in the different clusters
  - Determined by one pair of points, i.e., by one link in the proximity graph.

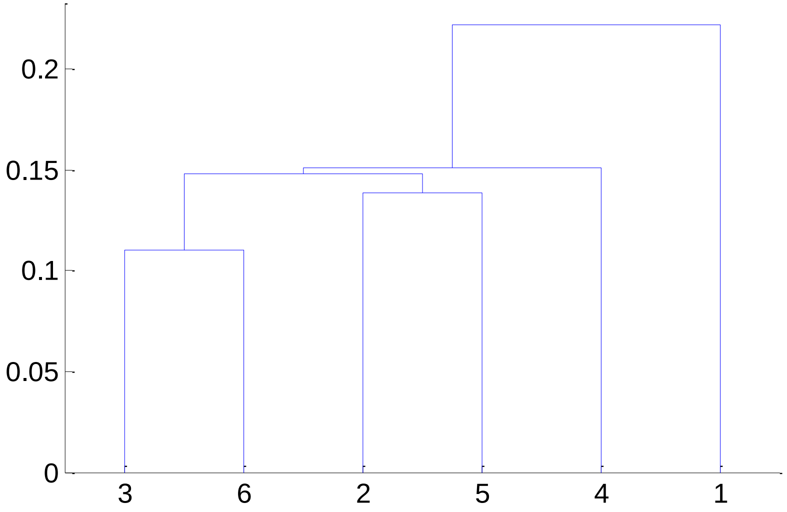
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: MIN

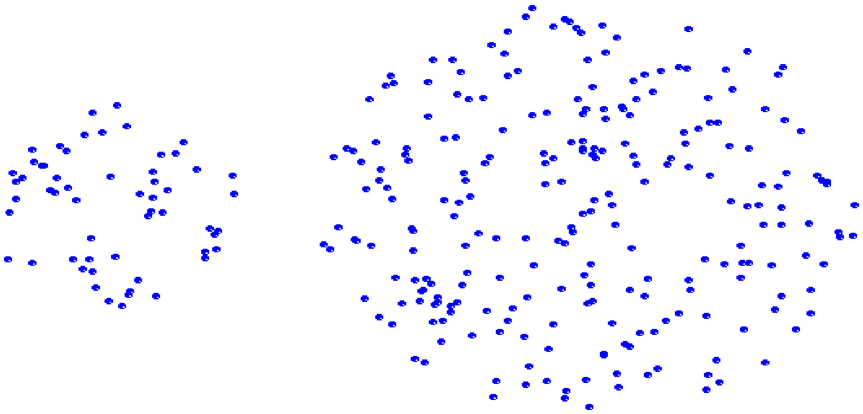


Nested Clusters

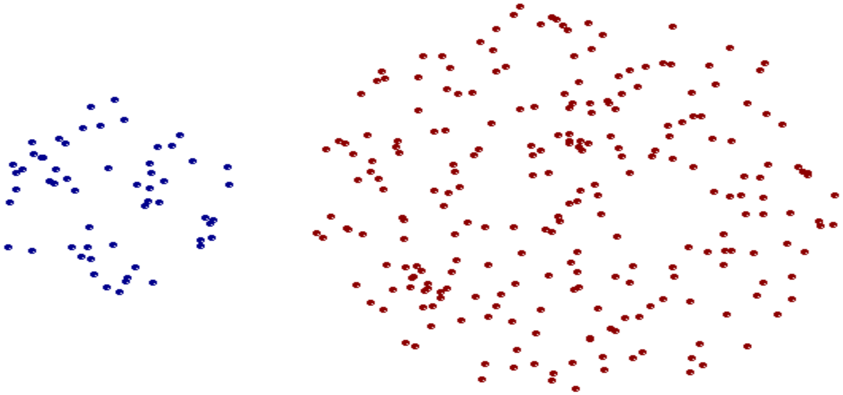


Dendrogram

Strength of MIN



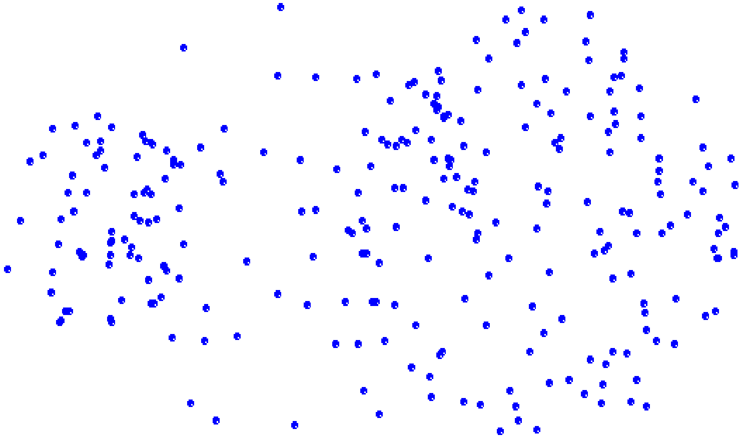
Original Points



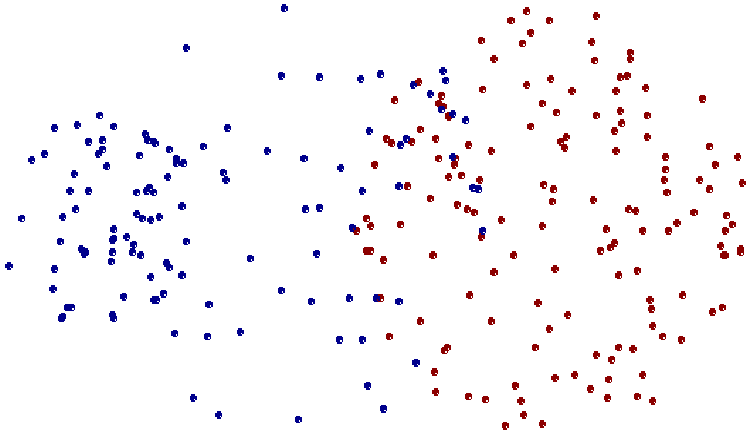
Two Clusters

- Can handle non-elliptical shapes

Limitations of MIN



Original Points



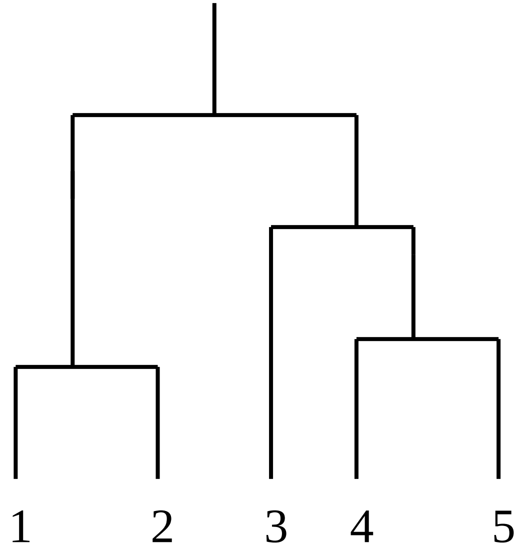
Two Clusters

- Sensitive to noise and outliers

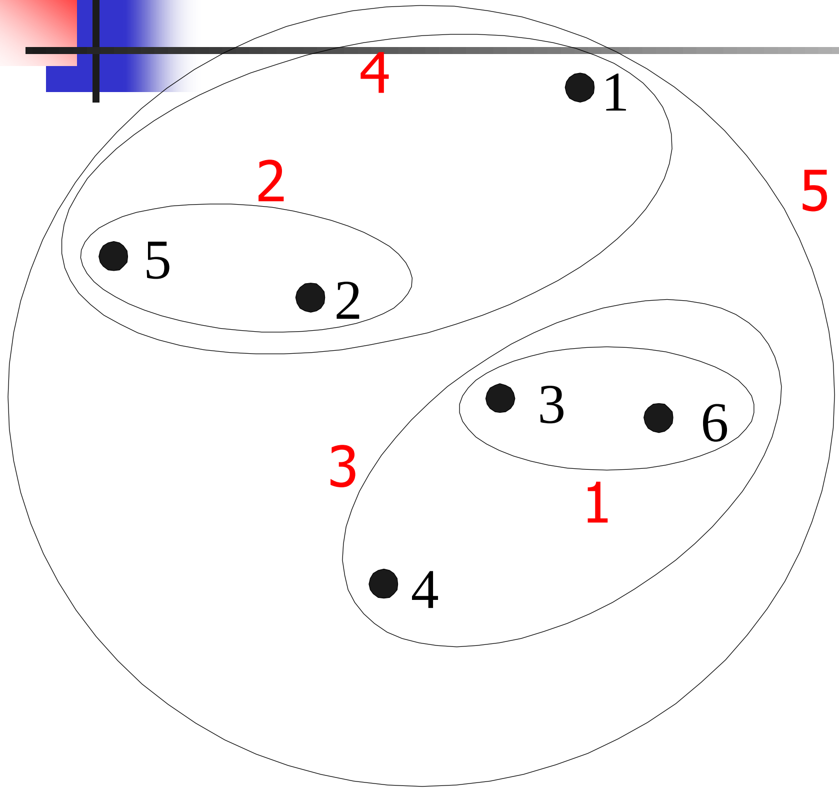
Cluster Similarity: MAX or Complete Linkage

- Similarity of two clusters is based on the two least similar (most distant) points in the different clusters
  - Determined by all pairs of points in the two clusters

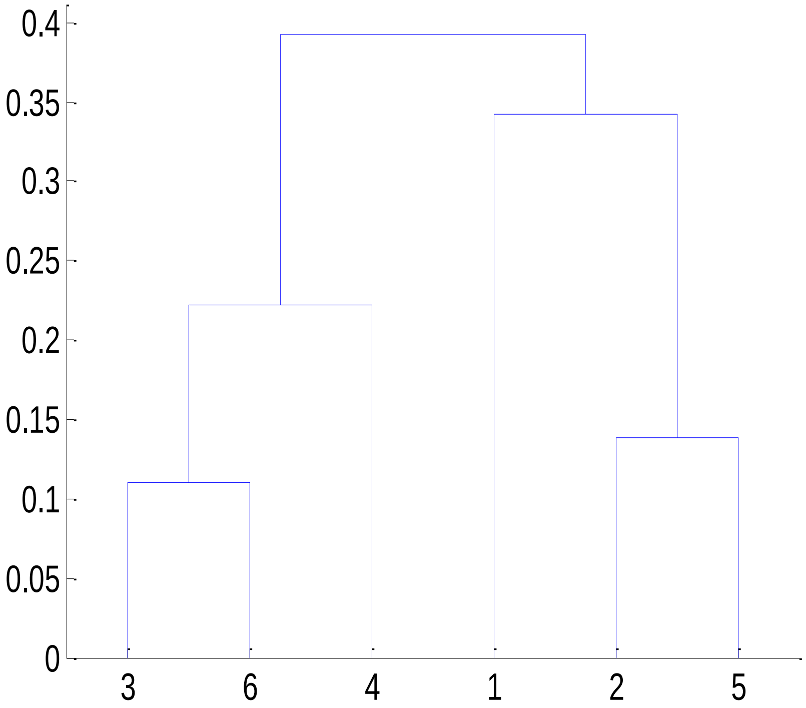
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: MAX

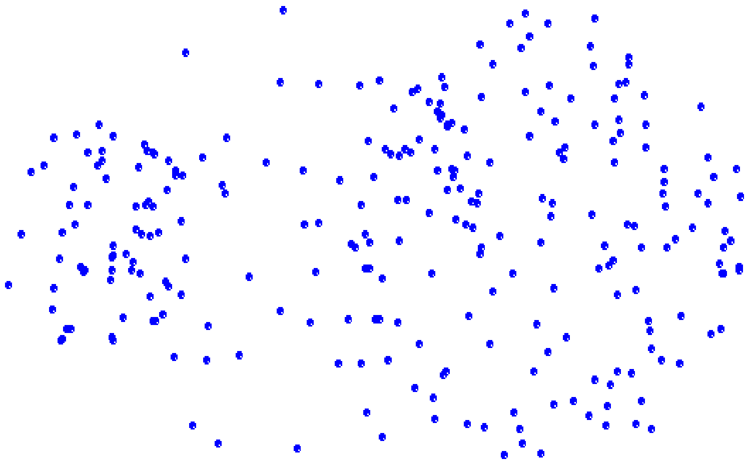


Nested Clusters

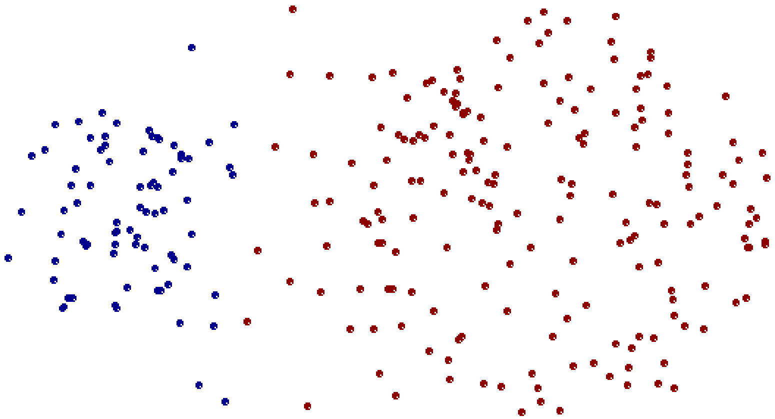


Dendrogram

Strength of MAX



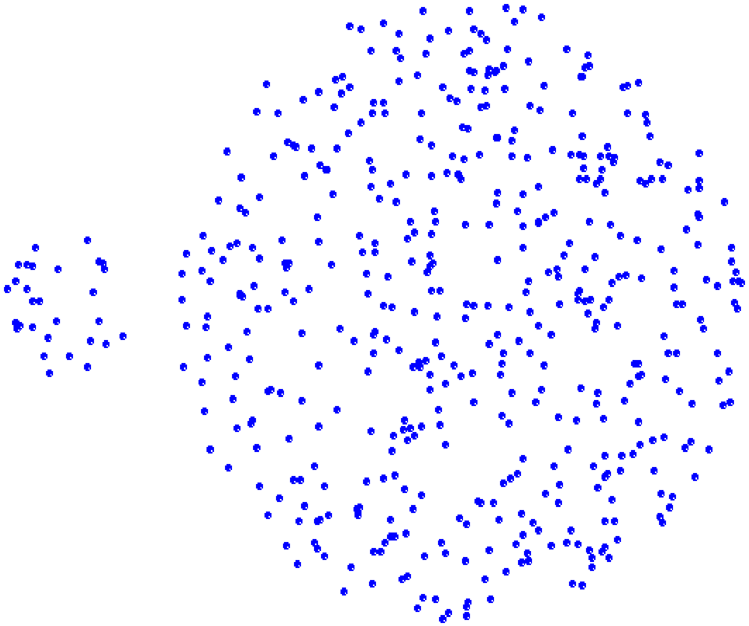
Original Points



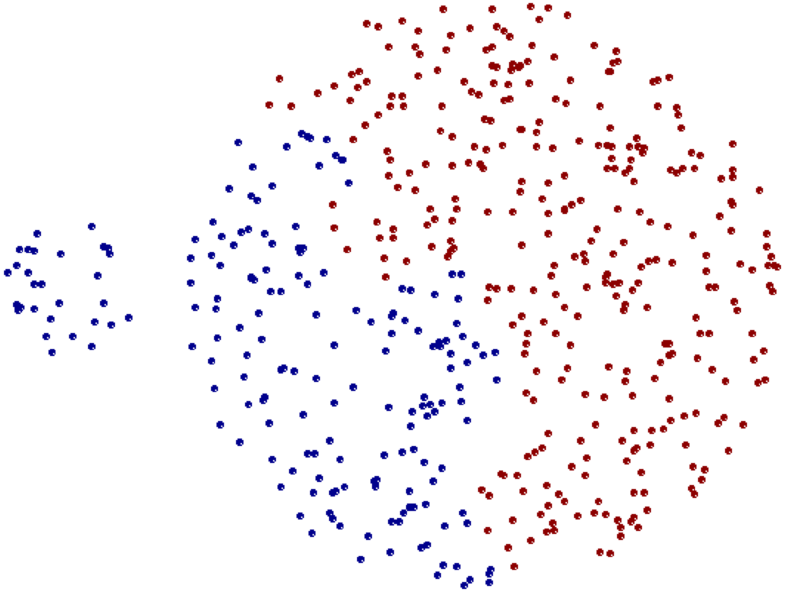
Two Clusters

- Less susceptible to noise and outliers

Limitations of MAX



Original Points



Two Clusters

- Tends to break large clusters
- Biased towards globular clusters

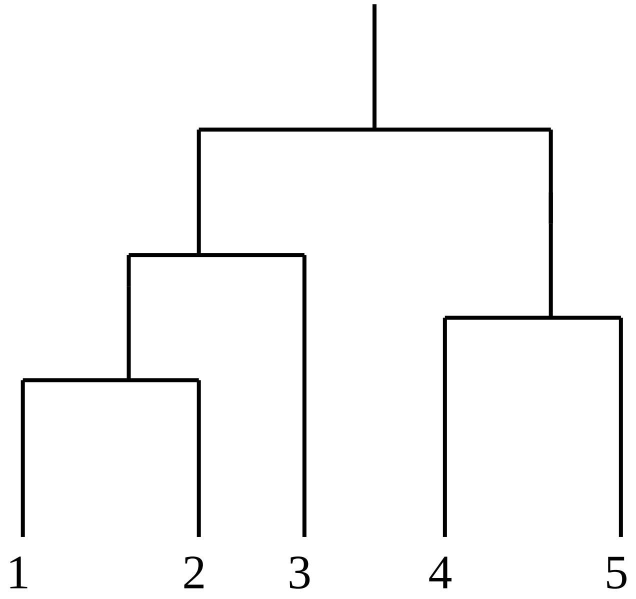
Cluster Similarity: Group Average

- Proximity of two clusters is the average of pairwise proximity between points in the two clusters.

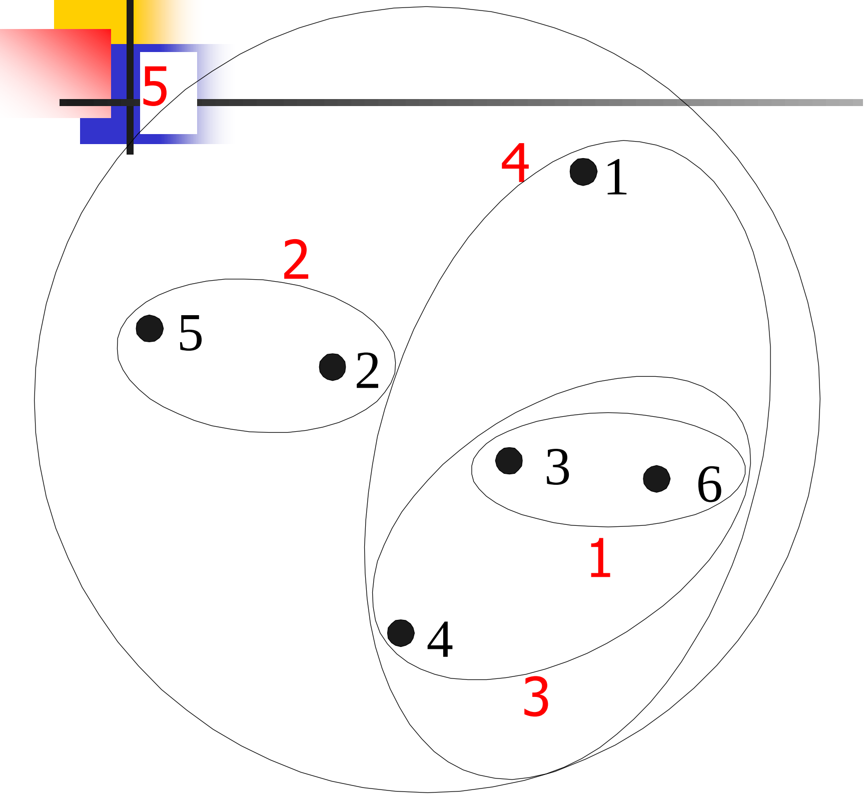
$$\textbf{proximity}(\textbf{Cluster}_i, \textbf{Cluster}_j) = \frac{\sum\limits_{\substack{p_i \in \textbf{Cluster}_i \\ p_j \in \textbf{Cluster}_j}} \textbf{proximity}(p_i, p_j)}{|\textbf{Cluster}_i| * |\textbf{Cluster}_j|}$$

- Need to use average connectivity for scalability since total proximity favors large clusters

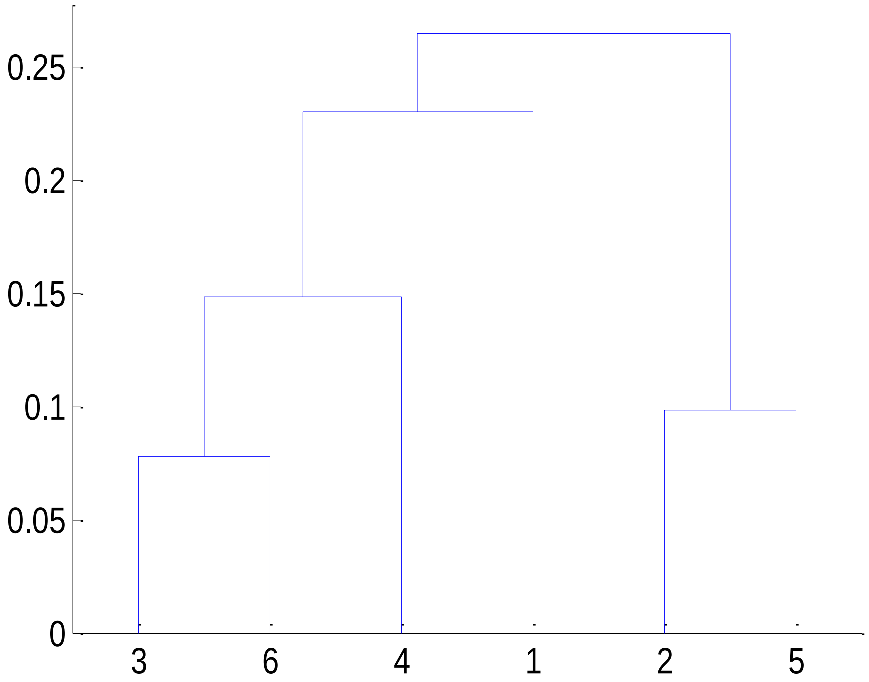
	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: Group Average



Nested Clusters



Dendrogram

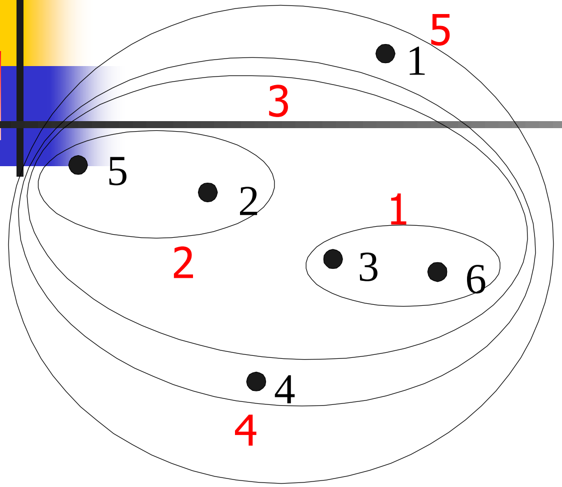
Hierarchical Clustering: Group Average

- Compromise between Single and Complete Link
- Strengths
  - Less susceptible to noise and outliers
- Limitations
  - Biased towards globular clusters

Cluster Similarity: Ward’s Method

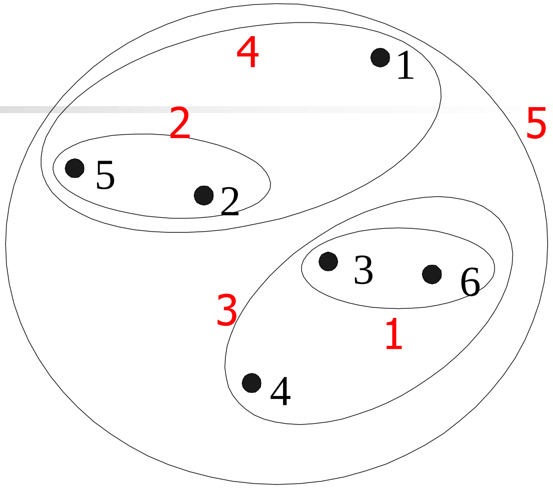
- Similarity of two clusters is based on the increase in squared error when two clusters are merged
  - Similar to group average if distance between points is distance squared
- Less susceptible to noise and outliers
- Biased towards globular clusters
- Hierarchical analogue of K-means
  - Can be used to initialize K-means

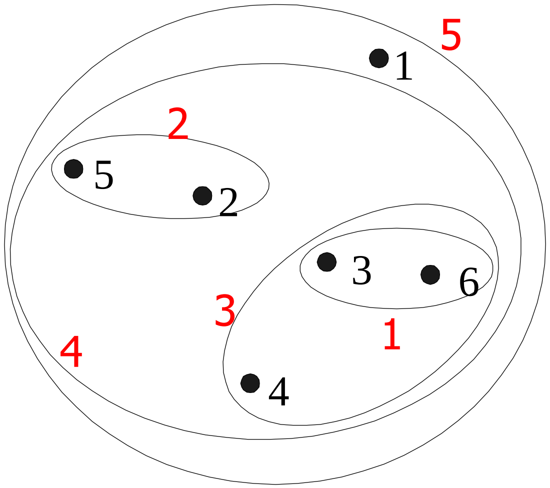
Hierarchical Clustering: Comparison



MIN

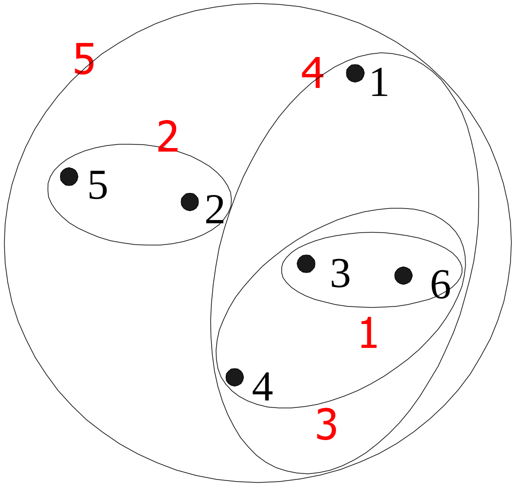
MAX





Group Average

Ward's Method



Application

- <u>Colour-Based Image Segmentation Using *K*-means</u>

Step 1: Loading a colour image of tissue stained with hemotoxylin and eosin (H&E)

H&E image

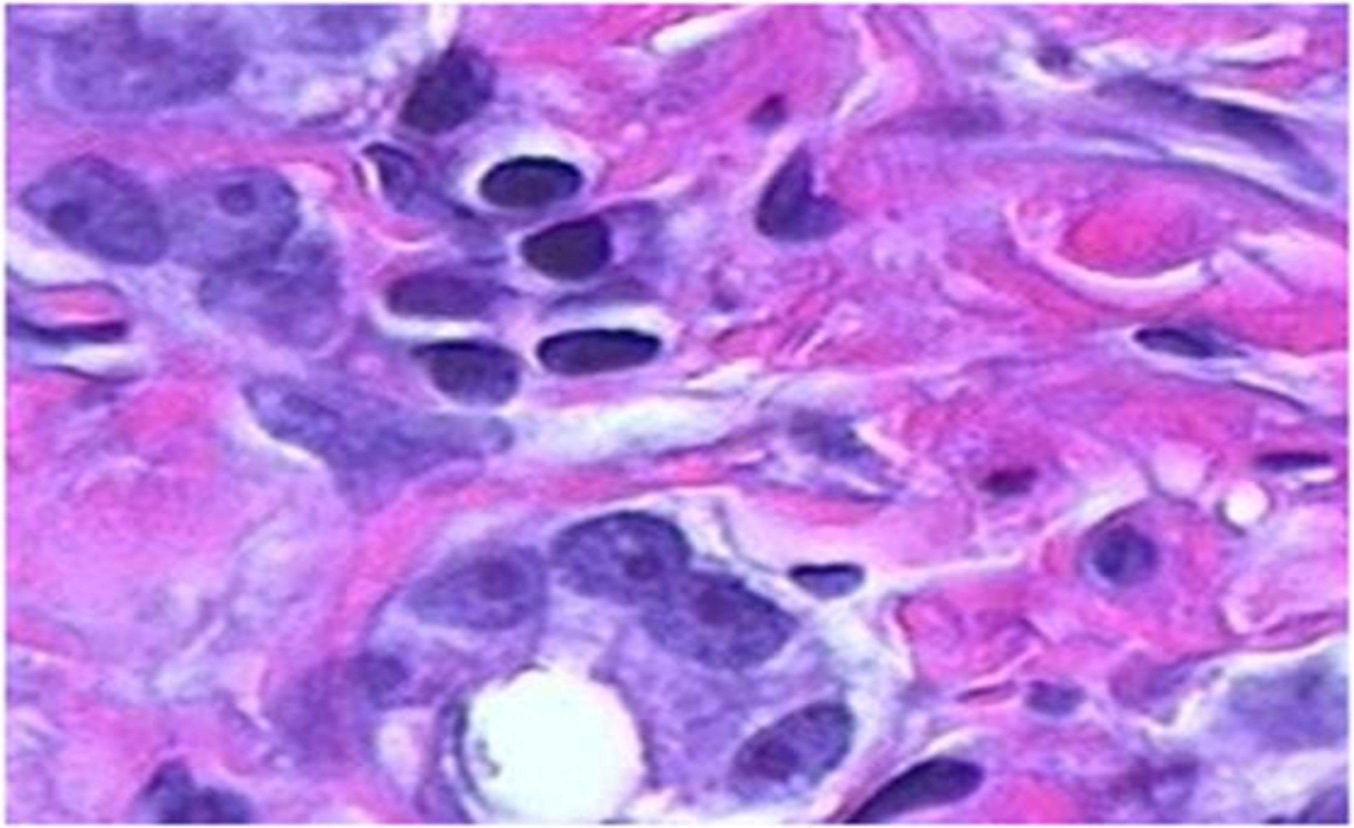


Image courtesy of Alan Partin, Johns Hopkins University

Application

- Colour-Based Image Segmentation Using *K*-means

Step 2: Convert the image from RGB colour space to L*a*b* colour space

- Unlike the RGB colour model, L*a*b* colour is designed to approximate human vision.
- There is a complicated transformation between RGB and L*a*b*.

$$(L^*, a^*, b^*) = T(R, G, B).$$

$$(R, G, B) = T'(L^*, a^*, b^*).$$

Application

- Colour-Based Image Segmentation Using *K*-means

Step 3: Undertake clustering analysis in the (a*, b*) colour space with the *K*-means algorithm

- In the L*a*b* colour space, each pixel has a properties or feature vector: (L*, a*, b*).
- Like feature selection, L* feature is discarded. As a result, each pixel has a feature vector (a*, b*).
- Applying the *K*-means algorithm to the image in the a*b* feature space where *K* = 3 by applying the domain knowledge.

Application

- Colour-Based Image Segmentation Using *K*-means

Step 4: Label every pixel in the image using the results from *K*-means clustering (indicated by three different grey levels)

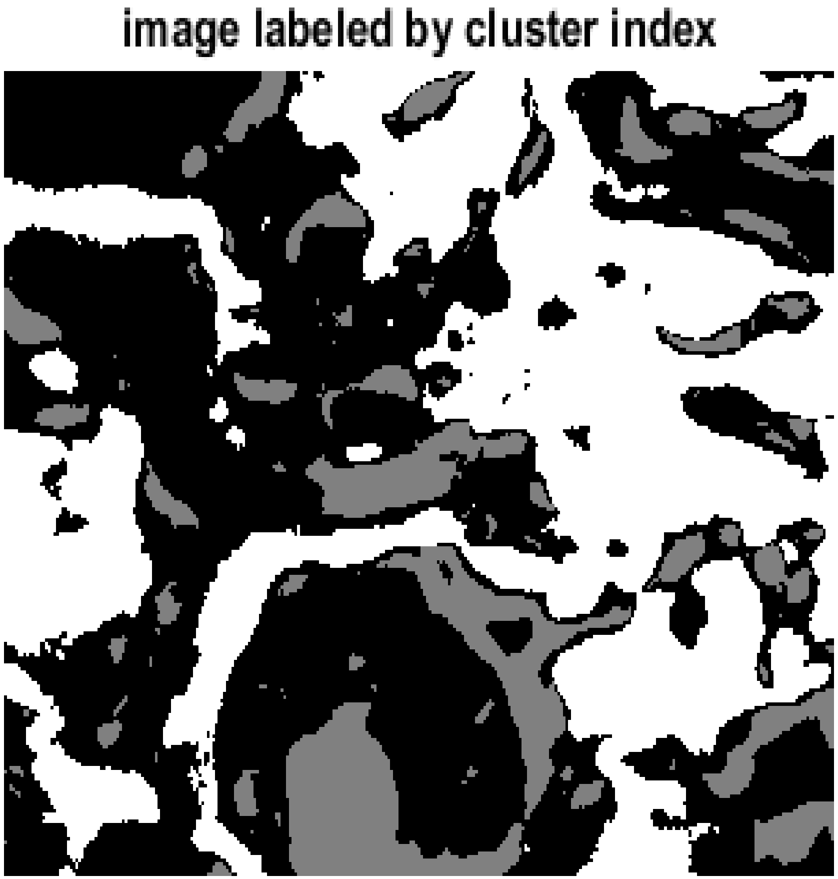


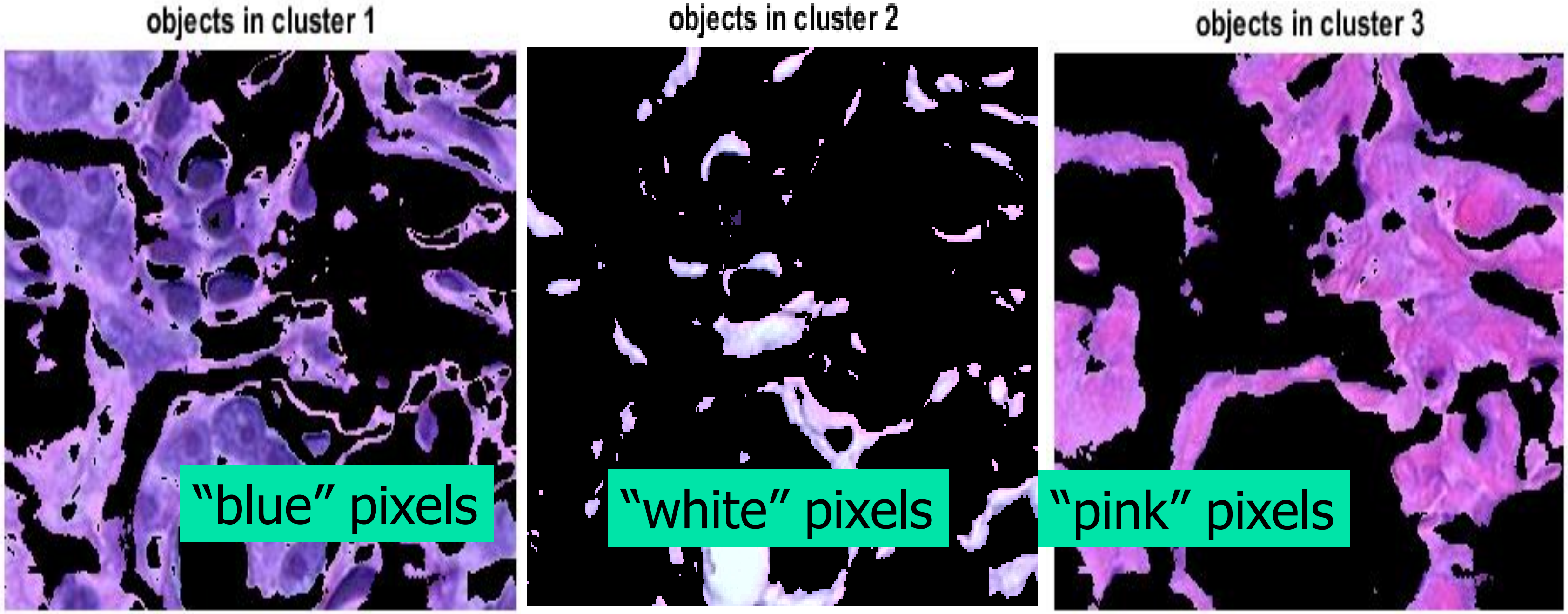
image labeled by cluster index

Application

- Colour-Based Image Segmentation Using *K*-means

Step 5: Create Images that Segment the H&E Image by Colour

- Apply the label and the colour information of each pixel to achieve separate colour images corresponding to three clusters.

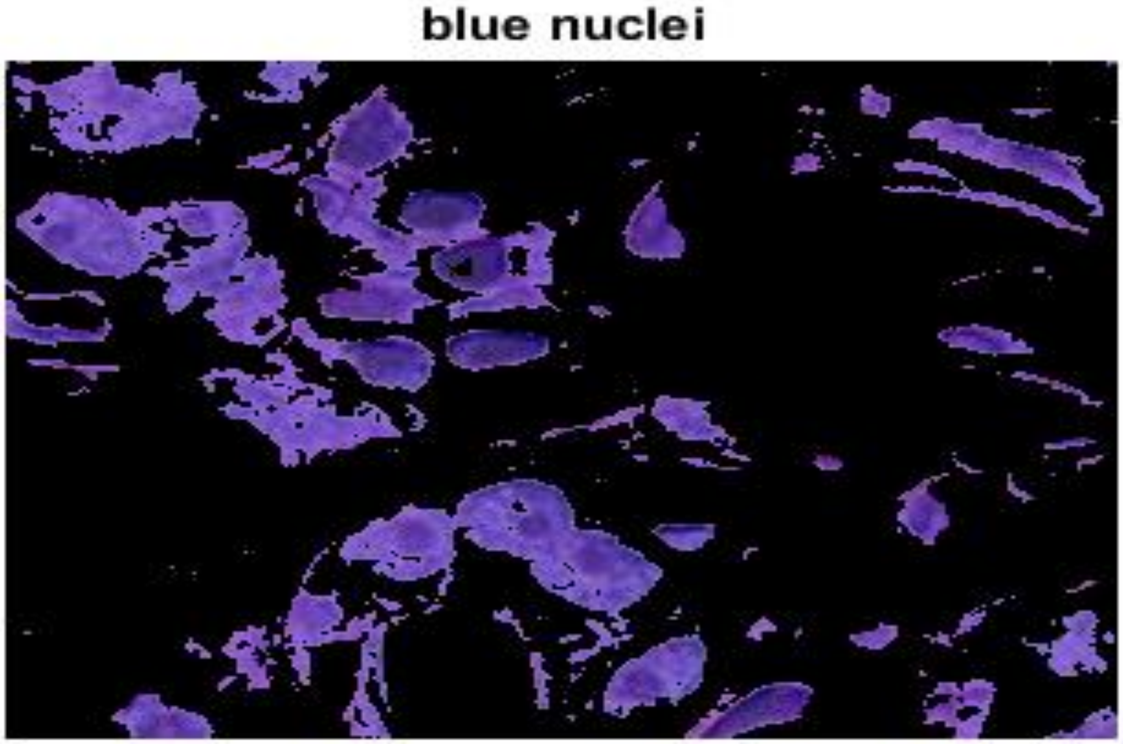


Application

- Colour-Based Image Segmentation Using *K*-means

Step 6: Segment the nuclei into a separate image with the L* feature

- In cluster 1, there are dark and light blue objects (pixels). The dark blue objects (pixels) correspond to nuclei (with the domain knowledge).
- L* feature specifies the brightness values of each colour.
- With a threshold for L*, we achieve an image containing the nuclei only.



blue nuclei

Summary

- *K*-means algorithm is a simple yet popular method for clustering analysis
- Its performance is determined by initialisation and appropriate distance measure
- There are several variants of *K*-means to overcome its weaknesses
  - *K*-Medoids: resistance to noise and/or outliers
  - *K*-Modes: extension to categorical data clustering analysis
  - CLARA: extension to deal with large data sets
  - Mixture models (EM): handling uncertainty of clusters