Specialist Programme on Artificial Intelligence for IT & ITES Industry

Pattern Recognition Using Clustering

By *Dr Zhu Fangming*
fangming@nus.edu.sg

Singapore e-Government Leadership Centre
National University of Singapore

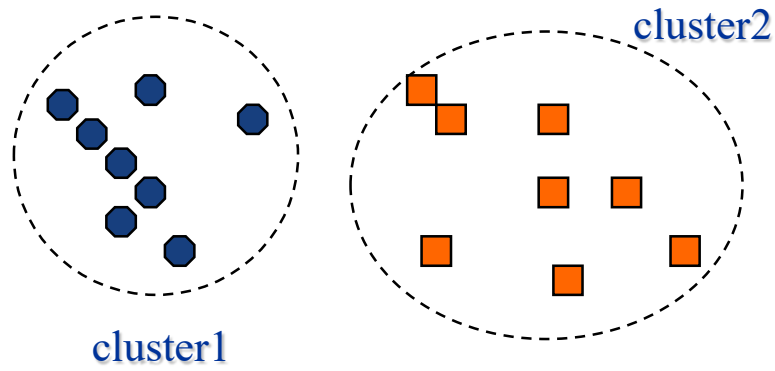
Transform
Lead
Inspire

Clustering

- Cluster Analysis (or simply "Clustering") is a multivariate data analysis technique.

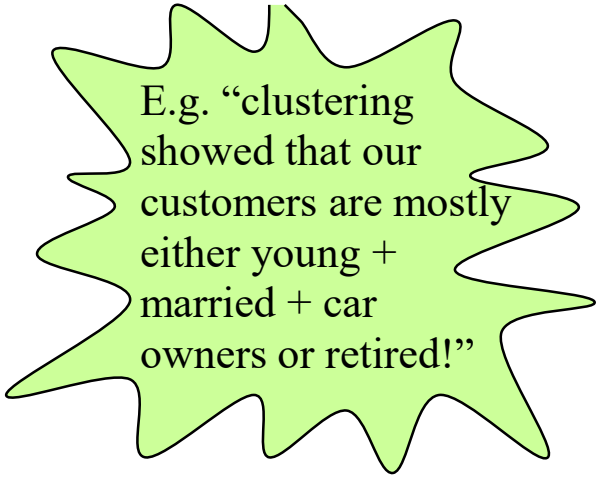
- It is an unsupervised learning method because there are no predefined classes or labels.

- Clustering usually uses algorithms to perform grouping based on a (potentially) large set of variables.



Why Do Clustering?

- Learn something new about the data
  - Understanding the natural structure in the data may lead to knowledge discovery



- Simplify the problem
  - Big databases often have too much complex structure for successful analysis. Analysis of smaller, homogenous clusters may yield better results

- Use the clusters as predictive models
  - E.g. cluster customer sales data to find groups of "typical" buyers. Predict new buyers by measuring their similarity to these clusters

Applications of Clustering

- Clustering is versatile and can be used in many problems across many domains:
  - Sales & Marketing: help marketers discover groups in their customer databases, and then use this insight to develop more targeted marketing campaigns
  - Fraud Detection: Identify groups of customers whose transaction behavior is uncharacteristic
  - Health & Bioinformatics: help physicians discover groups of patients with similar profiles and with a similar risk pattern, and use this insight to make predictions about diseases risks
  - Insurance: Help identify groups of policy holders with high average claim cost
  - ……

Clustering Algorithms

- Clustering algorithms generally calculate the distance between different records and try to group the ones that are closest together.
- Hierarchical Clustering
  - Furthest Neighbour
  - Nearest Neighbour
  - …..
- Non-hierarchical Clustering
  - K-Means Clustering
  - DBSCAN: Density-based clustering
  - …..

Measuring Similarity/Distance

- Euclidean Measure of Distance is commonest

$$d_{xy} = \sqrt{\sum_{k=1}^{p} (x_i - y_i)^2}$$

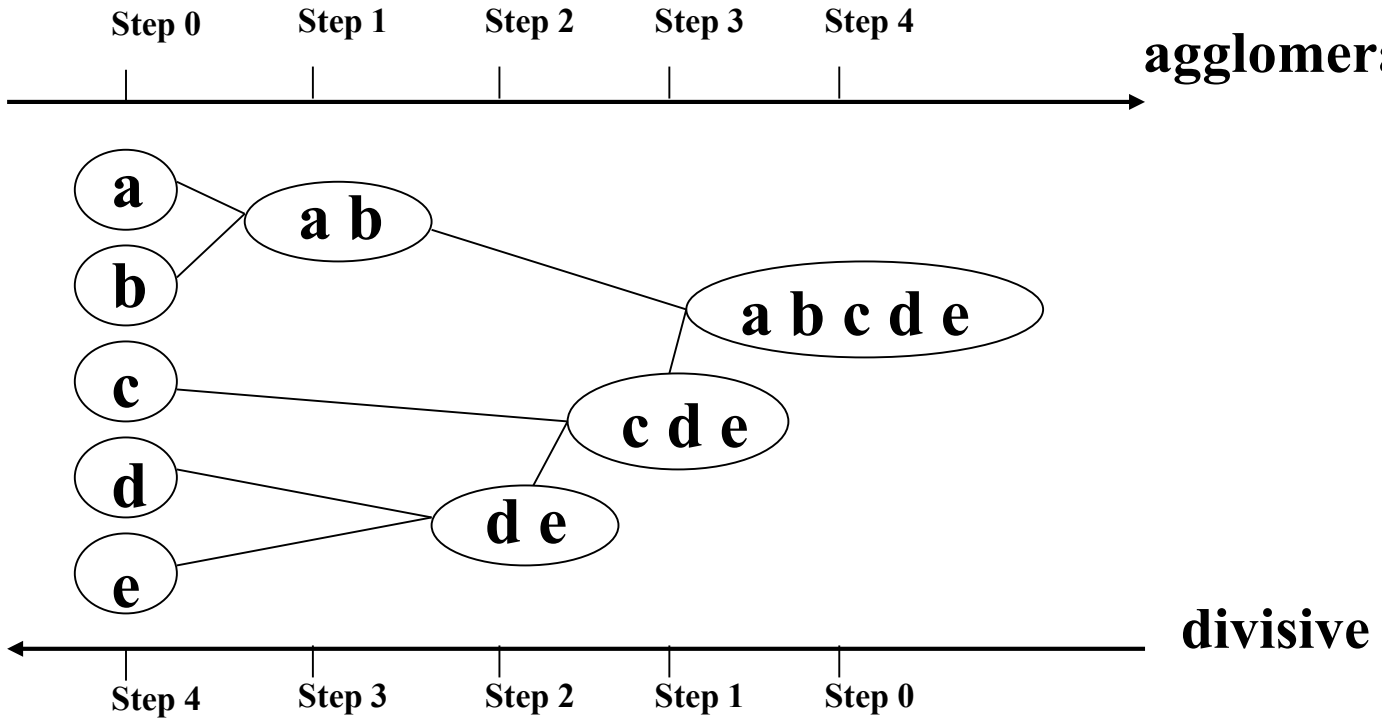
ID	Age	Income
S1234567D	21	5600
S3456782X	56	4600
B1725353Y	39	7000

$$\sqrt{(21-56)^2 + (5600-4600)^2}$$

First normalise each variable to the range 0-1 to eliminate bias of "big" numbers – usually done by the tool

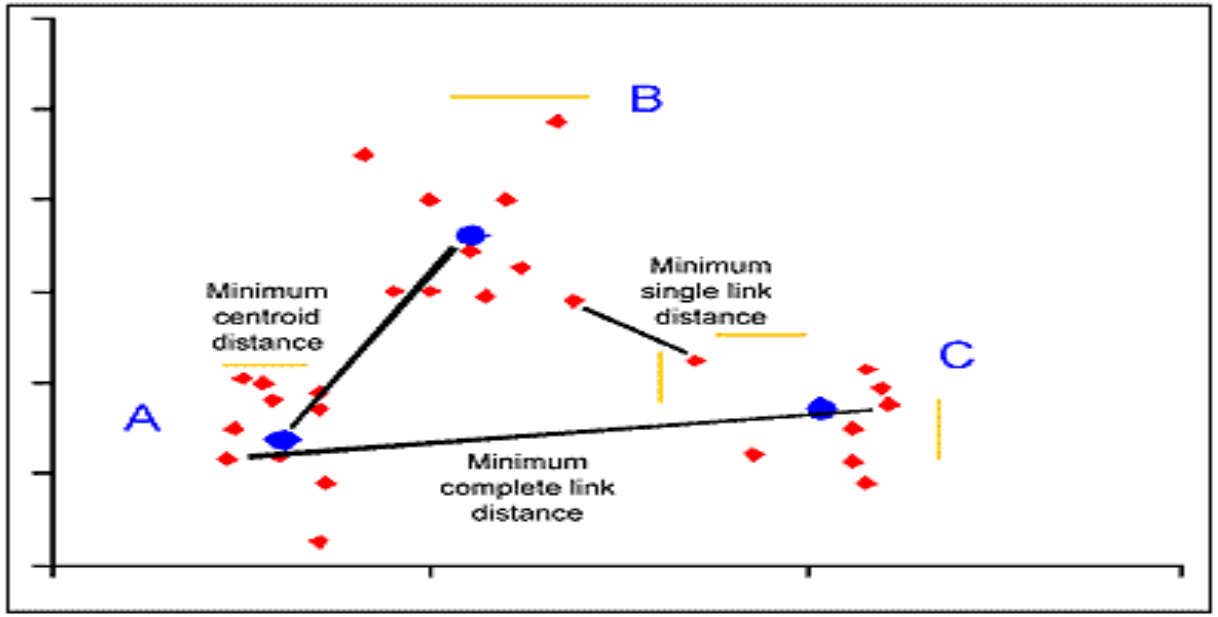
Hierarchical Clustering

- Use distance matrix as clustering criteria. This method does not require the number of clusters ***k*** as an input, but needs a termination condition.

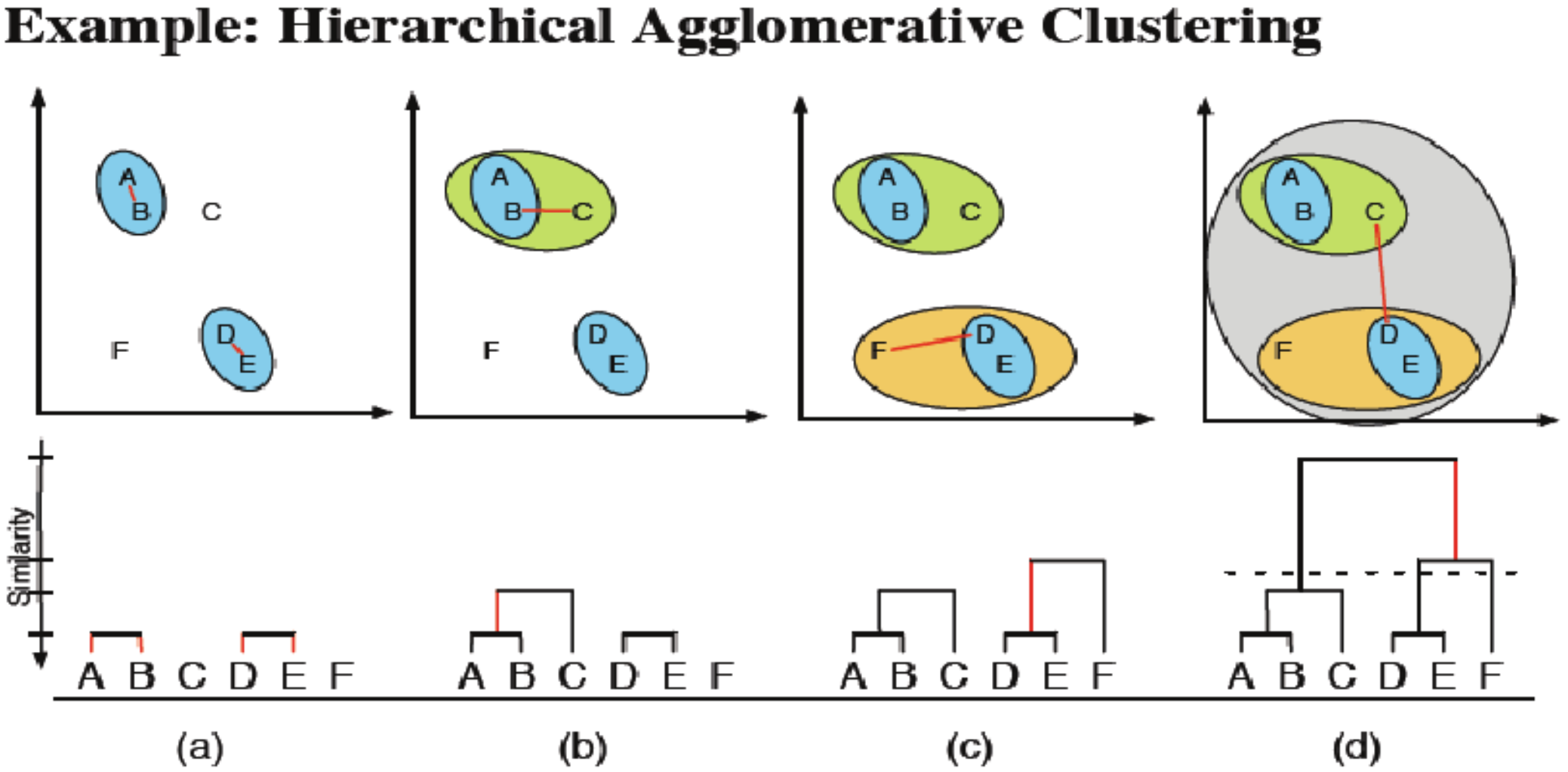


Distance Measures between Clusters

- ***Single link method***- The distance between two clusters is equal to the distance between the two closest records in them, aka *nearest neighbor method*.
- ***Complete link method***- The distance between two clusters is equal to the distance between the two most distant records in them, aka *furthest neighbor method*.
- ***Centroid method-*** The distance between two clusters is equal to the distance between their centroids.



Hierarchical Clustering - Dendrogram



Source: https://towardsdatascience.com/

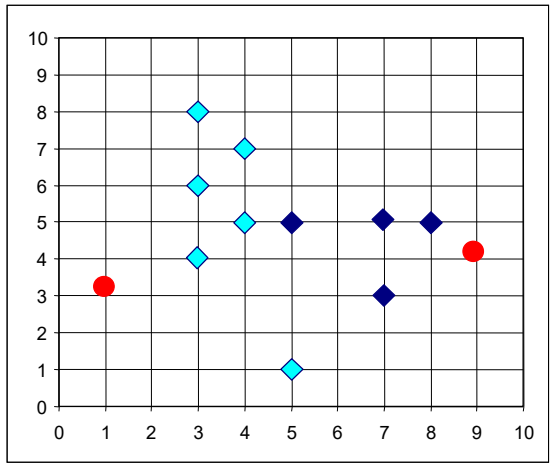
Non-hierarchical Clustering: K-Means

- Algorithm

  1. Randomly partition dataset into a specified number (K) of clusters
  2. Calculate centroids or average values of clusters
  3. Assign each data point to nearest cluster centroid
  4. Compute new centroids or averages of clusters and update clusters after complete pass of data
  5. Execute 2 & 3 above till there is no change in clusters by data points

- Issues
  - Resulting clusters may not be the best and highly dependent upon initial partitioning or division of the data set
  - Difficult to determine the best clustering

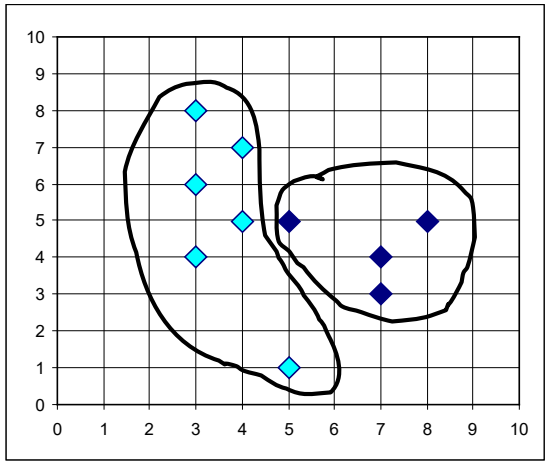
The *K-Means* Clustering Method



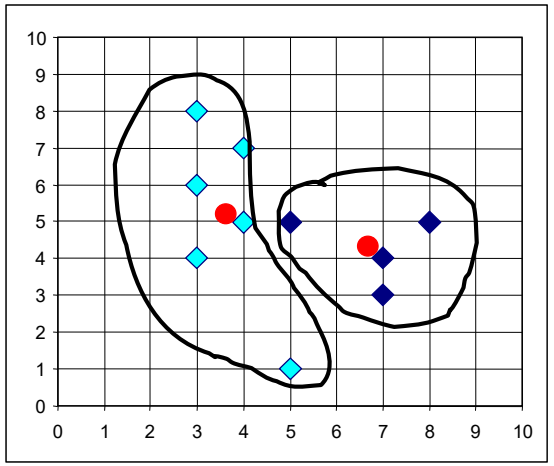
K=2

Arbitrarily choose K objects as initial cluster center

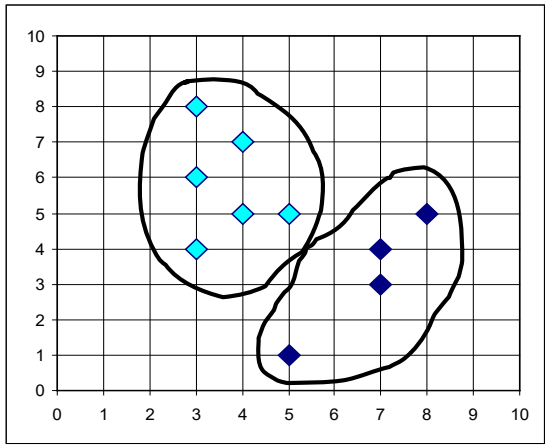
Assign each object to the nearest center



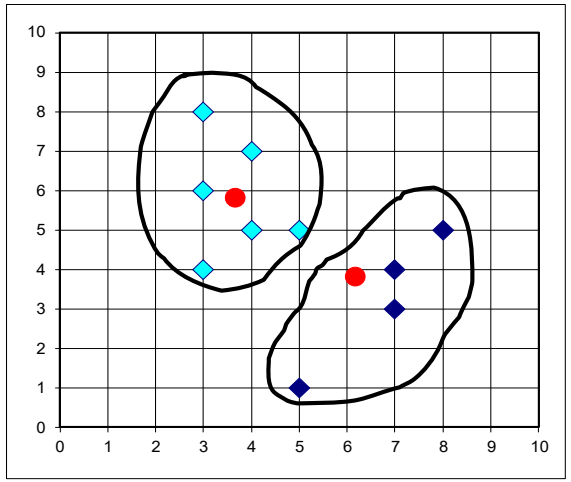
Update the cluster means



reassign

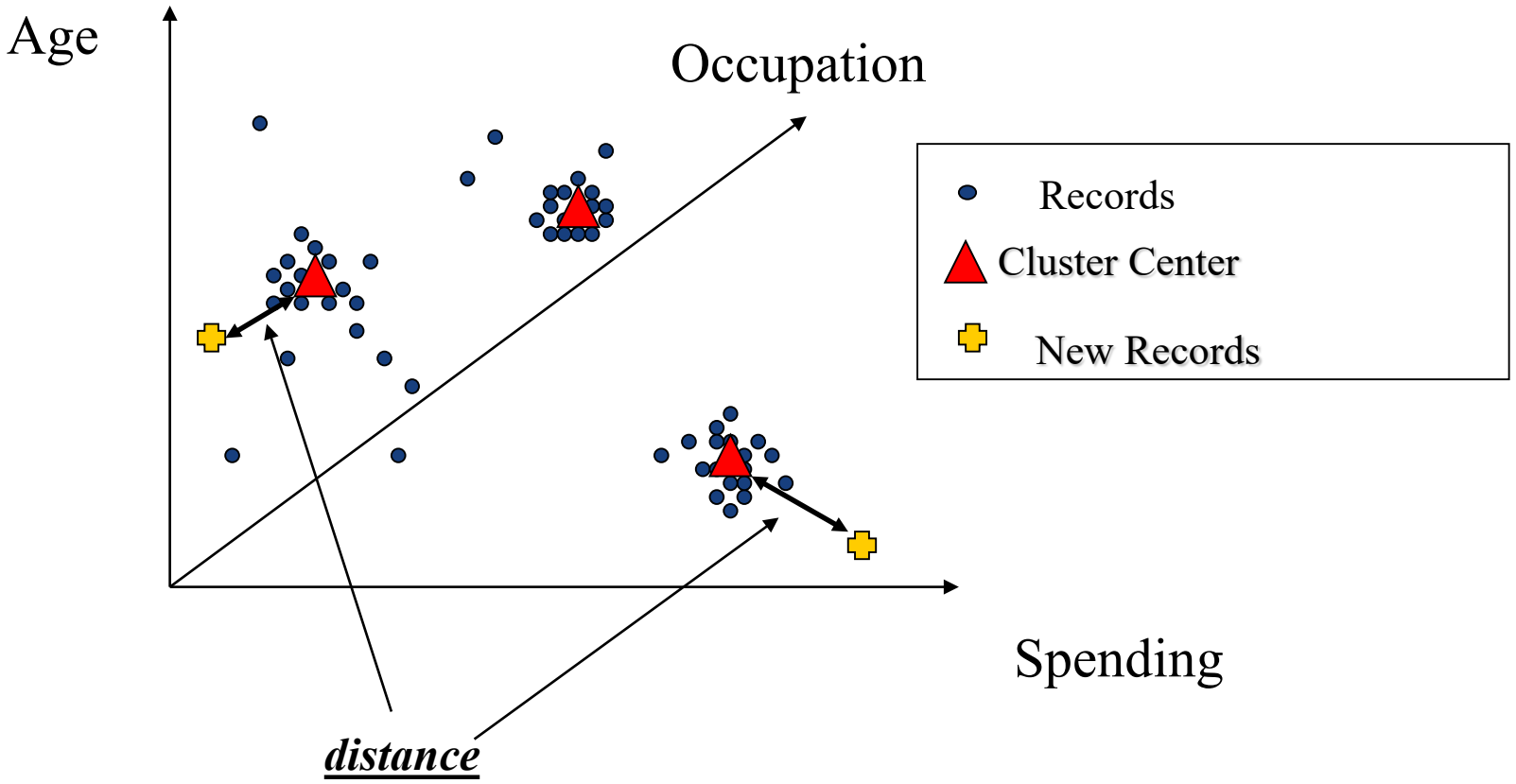


Update the cluster means



reassign

K-Means



Clustering Issues

- Clustering can be the most difficult data analytics /pattern recognition activity.

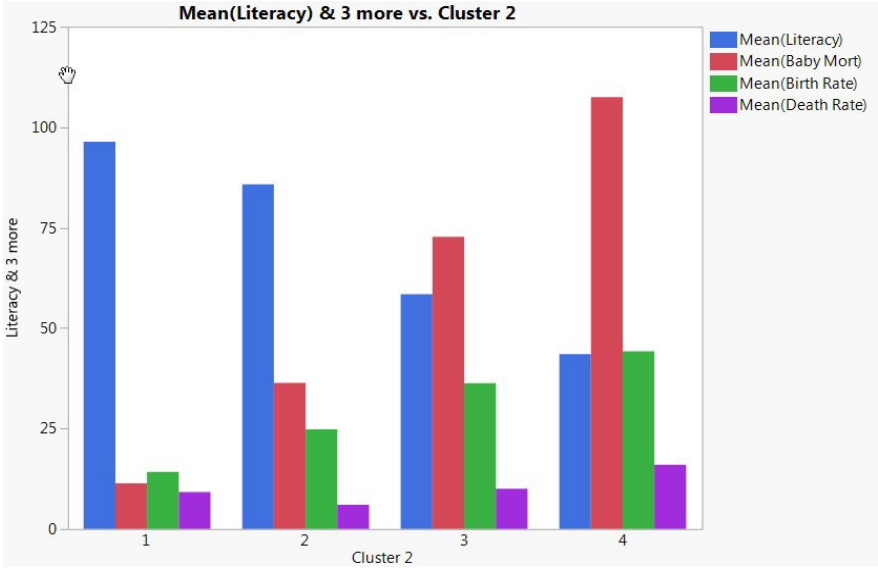
- Problems and issues include:
  - Variable selection
  - Understanding the resulting clusters
  - Assessing the quality of the clusters
  - Utilising the clusters

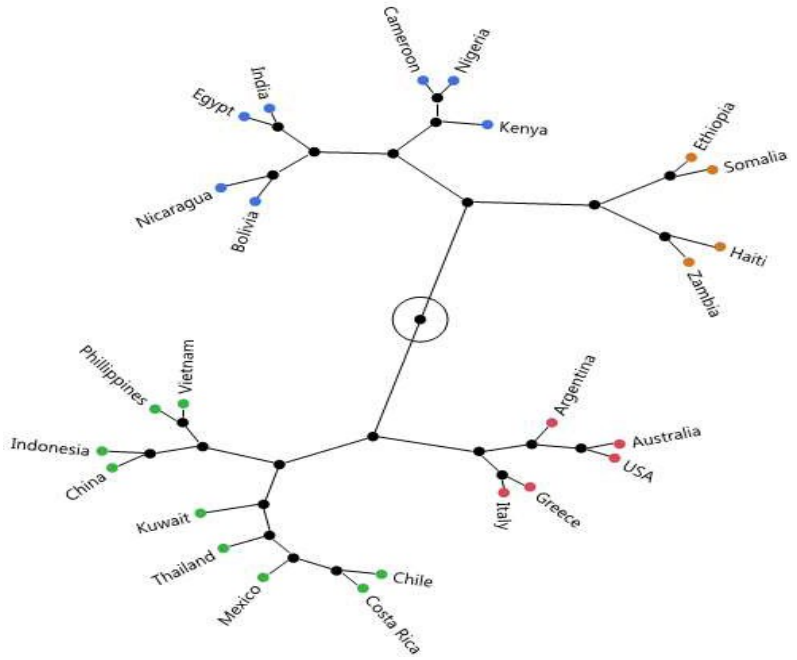
Clustering Issues

- Variable selection
  - Clustering with too many variables produces poor clusters that are not homogeneous within clusters and heterogeneous between clusters.
  - Clustering is “unsupervised learning” - there is no target variable to guide the selection of relevant versus non-relevant variables.

Clustering Issues

- Cluster Understanding
  - Clustering algorithms assign a cluster label (typically a number) to each record. How do we interpret what this means?
  - Clustering tools provide various aids to help cluster understanding – visualisation aids are particularly useful.



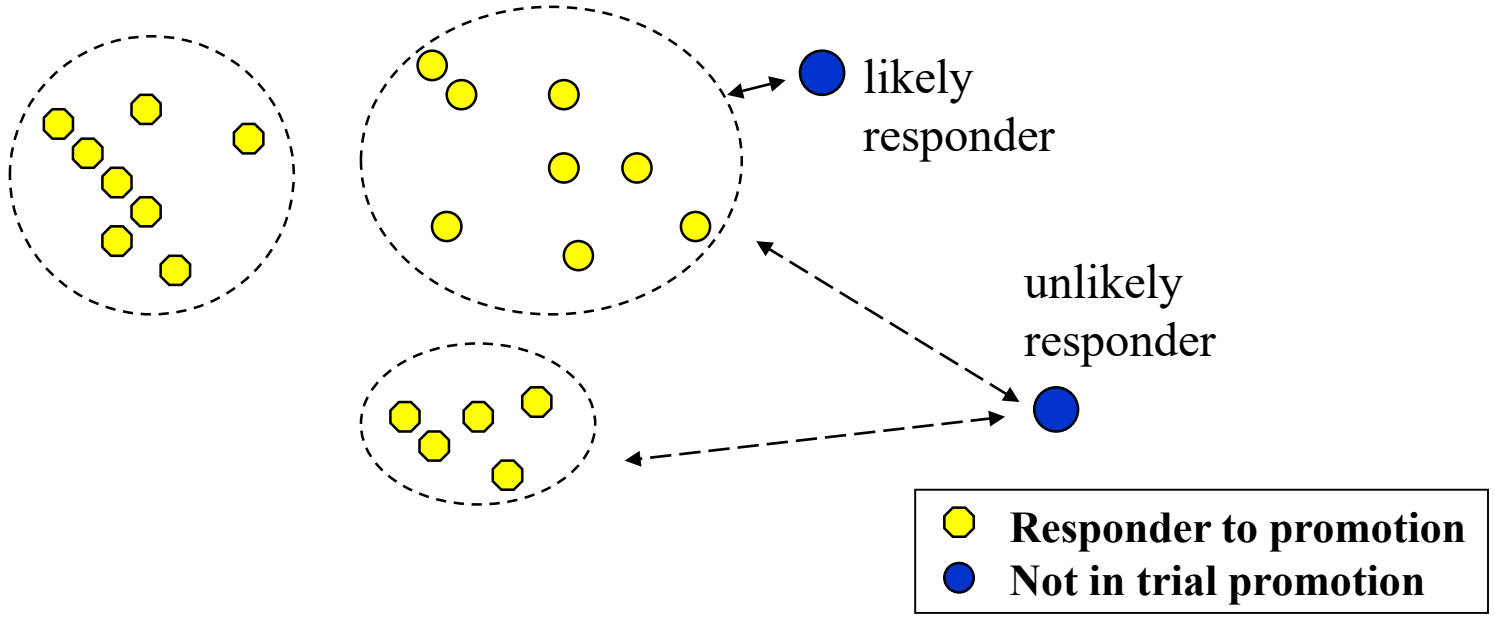


Clustering Issues

- Assessing the quality of the clusters
  - Assess the **number** of clusters and their relative **sizes**
    - If there are very small clusters, it could mean that they represent outliers. This should warrant further investigation.
  - Assess the cluster **cohesion** and **separation**
    - The **Sum of Square error** (SSE) is produced by clustering algorithms to help you judge the quality of the cluster analysis.
    - The cluster **silhouette** is a combined measure of **internal cohesion** and **external separation** to gauge the quality of the cluster solution.

Clustering Issues

- Utilising the clusters
  - Analyse clusters for knowledge discovery
  - Use of clusters as predictive (or other) models
    - E.g. to generate a mailing list given a list of responders to a previous mailing campaign



Contact eGL

Singapore e-Government Leadership Centre
National University of Singapore
29 Heng Mui Keng Terrace
Block D & E
Singapore 119620

Tel	:	**(65) 6516 1156**
Fax	:	**(65) 6778 2571**
URL	:	**www.egl.sg**
Email	:	**egl-enquiries@nus.edu.sg**

Transform
Lead
Inspire