

# **Lesson 8: Cluster Analysis: Concepts, Algorithms and Methods**

Version 2024-25T1.1

Instructor: Dr. Kam Tin Seong  
Associate Professor of Information Systems (Practice)  
School of Information Systems  
Singapore Management University

## Lesson Objectives

- Basic concepts of Cluster Analysis
- Hierarchical Cluster Analysis Techniques: Concepts and algorithms

Welcome to the lesson on Cluster Analysis. In this lesson, we tackle a canonical multivariable data analysis problem: how to group records, observations or cases into classes of similar objects or more popularly known as cluster. The method you will be introduced to is called Hierarchical Cluster Analysis.

## Learning Outcomes

Upon completion of this section, student will be able to:

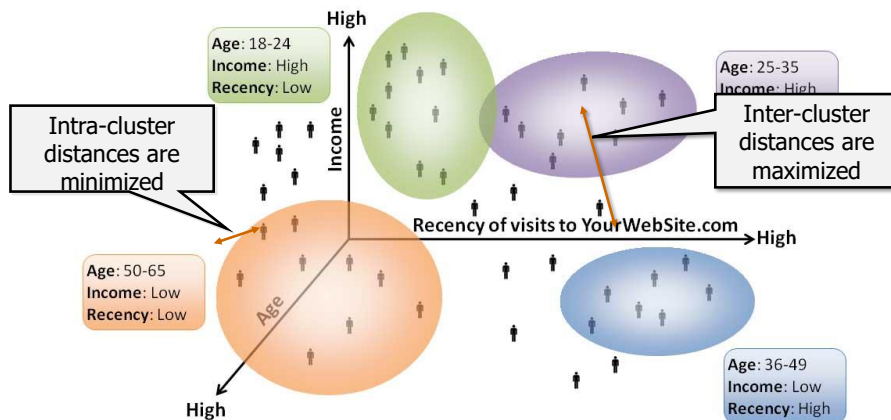
- understand the basic concepts of cluster analysis
- state the rationale of cluster analysis
- recognise the role of cluster analysis in business and policy planning

Upon completion of the lesson, you will be able to:

- Understand the basic concepts and challenges of cluster analysis.
- Understand hierarchical cluster analysis algorithms and how to use them to perform cluster analysis.
- Prepare data meet the different clustering algorithms needs
- Describe the characteristics of cluster by using appropriate graphical methods.

## What is Cluster Analysis?

- An exploratory data analysis technique.



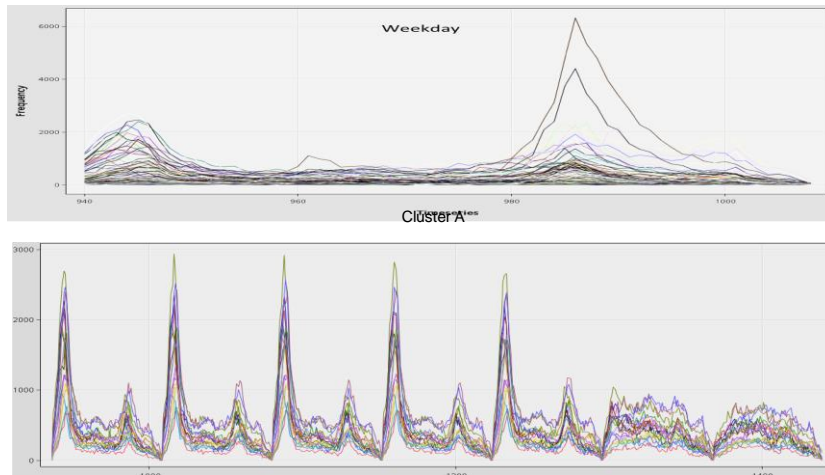
Cluster analysis is an exploratory data analysis technique for organizing observed data (e.g. people, things, events, brands, companies) into meaningful taxonomies, groups, or clusters, based on combinations of input variables, which maximizes the similarity of cases within each cluster while maximizing the dissimilarity between groups that are initially unknown. It provides no explanation as to why the clusters exist nor is any interpretation made. Each cluster thus describes, in terms of the data collected, the class to which its members belong. Items in each cluster are similar in some ways to each other and dissimilar to those in other clusters. Figure above shows clusters defined by using three variables, namely: income, age and recency of visits to YourWebSite.com, an online retail company. Four clusters were identified. The green cluster comprises visitor with age range of 18-24, high income and low recency to YourWebSite.com. The green cluster comprises customers within the age range of 18-24, high income and with low recency. The blue cluster comprises customer within the age range of 36-49, low income but high recency. Lastly, the purple cluster comprises customers within the age range of 25-35, high income and high recency. It is also important to note that there are also customers who are not belong to any one of these clusters.

As an unsupervised learning technique, cluster analysis finds patterns in unlabeled data, or data that lacks a defined response measure. Examples of unlabeled data include a bit-mapped photograph, a series of comments from social media, and a

battery of psychographic data gathered from a number of subjects. In each case, it may be possible to classify the objects through an external process: For example, you can ask a panel of oncologists to review a set of breast images and classify them as possibly malignant (or not), but the classification is not a part of the raw source data. Unsupervised learning techniques help the analyst identify data-driven patterns that may warrant further investigation.

## What is Cluster Analysis?

- An unsupervised machine learning technique



As an unsupervised machine learning technique, cluster analysis finds patterns in unlabelled data, or data that lacks a defined response measure. Examples of unlabelled data include a bit-mapped photograph, a series of comments from social media, and a battery of psychographic data gathered from a number of subjects.

Figure above shows weekday commuting patterns of MRT passengers. By using time-series cluster analysis technique, 11 clusters had been identified. For example Cluster A shows strong morning peak and moderate evening peak. They are mainly MRT stations located at or near to residential newtowns, namely: Admiralty MRT Station, Sembawang MRT Station, Yew Tew MRT Station, Lakeside MRT Station, Khatib MRT Station, Yishun MRT Station, Kovan MRT Station, Ang Mo Kio MRT Station, Choa Chu Kang MRT Station, Bedok MRT Station, Pasir Ris MRT Station, Simei MRT Station, Tampines MRT Station, Kembangan MRT Station, Sengkang LRT Station, Sengkang MRT Station and Serangoon MRT Station.

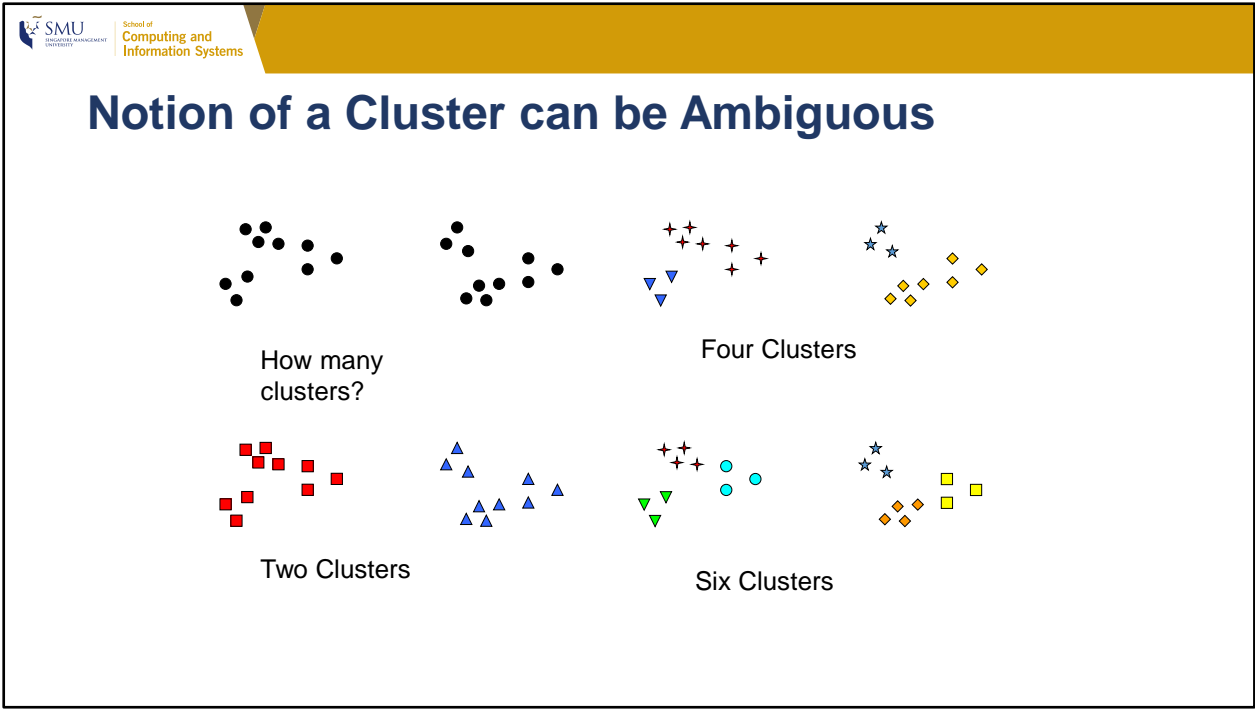
## What is not Cluster Analysis?

- Supervised classification
  - Have class label information
- Simple segmentation
  - Dividing students into different registration groups alphabetically, by last name
- Results of a query
  - Groupings are a result of an external specification
- Graph partitioning
  - Some mutual relevance and synergy, but areas are not identical

Cluster analysis is not a simple segmenting technique such as dividing students into different registration groups alphabetically order according to their last name.

Cluster analysis is not a query based on a collection of criteria. Members of a cluster exhibit a common statistical measurements.

Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included.



In many applications, the notion of cluster is not well defined. To better understand the difficulty of deciding what constitutes a cluster, let us consider the figure above. It shows twenty data points and three different ways of dividing them into clusters. The shape of the markers indicate cluster memberships. Figure (b) and (d) divide the data points into two and six parts, respectively. However, the apparent division of each of the two larger clusters into three sub-clusters may simply be an artifact of the human visual system. Also, it may not be unreasonable to say that the data points form four clusters, as shown in Figure (c). Hence, it is important to note that the definition of a cluster is imprecise and that the best definition depends on the nature of data and the desired results.



SMU

UNIVERSITY

School of

Computing and

Information Systems

# Cluster Analysis in Business



Capital One  
what's in your wallet?



P&G



Marriott



PROGRESSIVE  
#1 AUTO INSURANCE WEB SITE - 2008 KEYNOTE



AMERICAN EXPRESS



AXA



TOYOTA



ups



DHL



SINGAPORE  
POOLS



iNTUC  
INCOME



DBS



citi



The Walt Disney Company



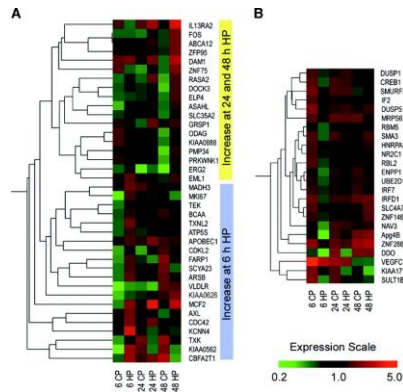
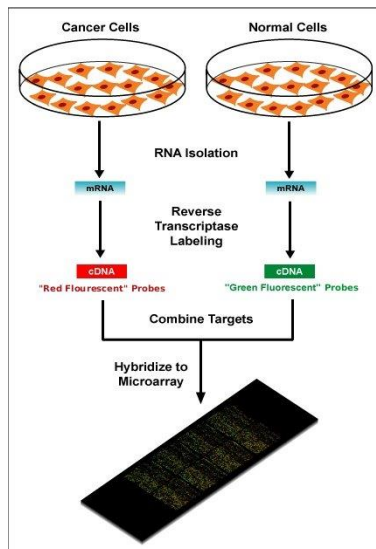
HARRAH'S  
ENTERTAINMENT

Hallmark

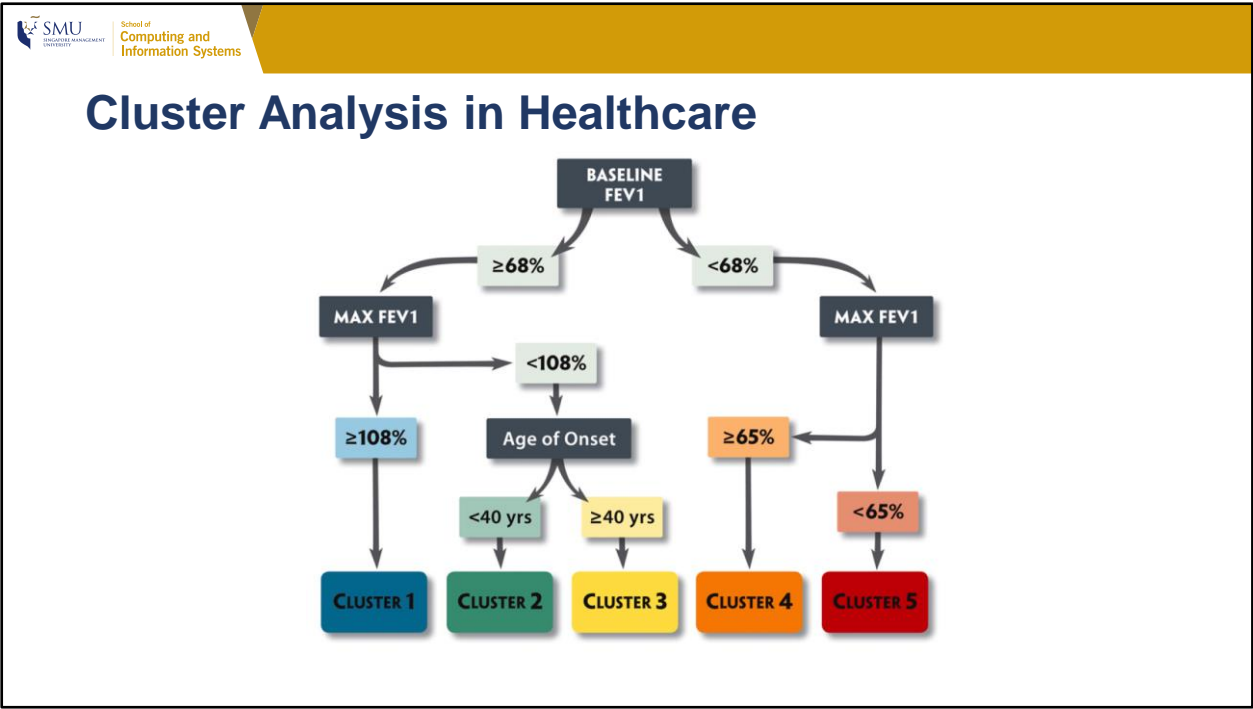
Several example on how cluster analysis had been used in business and policy planning will be discussed. This will allow you to recognise the role of cluster analysis in business and policy planning.

4-8

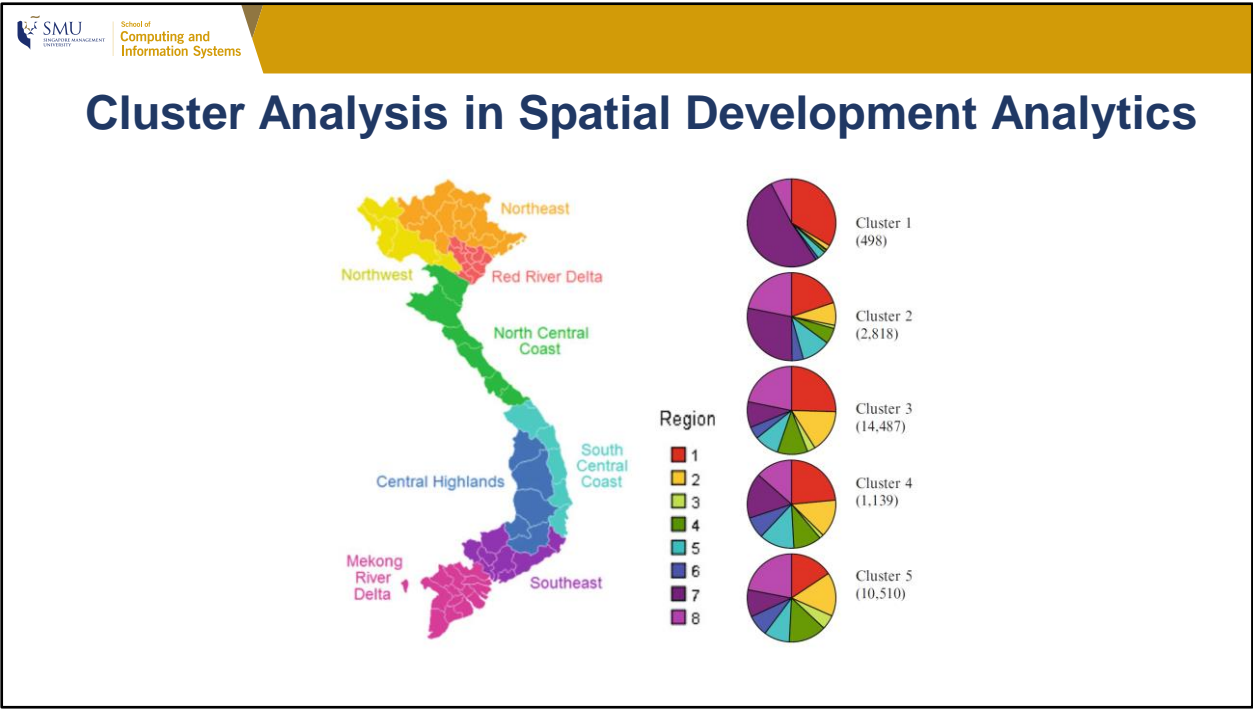
# Cluster Analysis in Bioinformatics



The past decade had witnessed a tremendous growth in Bioinformatics, which is the coming together of molecular biology, computer sciences, mathematics and statistics. Cluster analysis are used in many different application areas. One research area whereby cluster analysis have been playing important role is DNA microarrays. For example, Selinski and Ickstadt (2008) use cluster analysis of single-nucleotide polymorphisms to detect differences between diseased and control individuals in case-control studies, and Eisen et. al. (1998) use clustering of genome-wide expression data to identify cancer subtypes associated with survival; Witten and Tibshirani (2010) describe a similar application of clustering top renal cell carcinoma data.



Cluster analysis also have been widely used in healthcare and clinical studies. For example, Wendy C. Moore et al., (2010) used unsupervised hierarchical cluster analysis to identify novel asthma phenotypes. Their study had identified five distinct clinical phenotypes of asthma, which supports clinical heterogeneity in asthma and the need for new approaches for the classification of disease severity in asthma. The used of cluster analysis is not confined to clinical research but also in healthcare policy. For example, Sakiko Fukui et al., (2014) perform a cluster analysis on data from two national databases administered by the Ministry of Health, Labour and Welfare, Japan to identify five categories for home-visit nursing agencies based on the type of services delivery system.



Cluster analysis is also commonly employed in spatial development planning to group geographical areas such as communes, villages, districts and provinces into homogenous regions so that more targeted development strategies and programmes can be implemented. For examine in Vietnam, cluster analysis have been used to determine development regions by using socio-demographic and economic data of Household Living Standard Survey.

In Australia, Abbas Valadkhani et al., (2014) used a hierarchical cluster analysis to classify 109 retail petrol locations into six heterogeneous groups with homogeneous content. Their study had provided important policy implications for both consumers and regulators.

## Typology of clustering approaches

- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree
- Partitioning clustering (also known as k-means)
  - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset

There are a number of different methods that can be used to carry out a cluster analysis. These methods can be classified as follows:

- Hierarchical methods
  - A set of nested clusters organized as a hierarchical tree
- Non-hierarchical methods (often known as k-means clustering methods)

In this course, we will focus our discussion on hierarchical clustering.

## **Hierarchical Clustering: Basic Concepts and Algorithms**

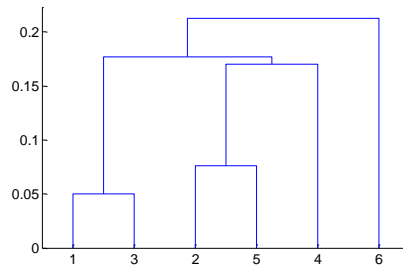
Upon completion of this section, student will be able to:

- understand the basic concepts of hierarchical clustering.
- explain the concepts proximity matrix.
- compare the differences between euclidean distance, city-block distance and chebychev distance.
- explain how Min, Max, Group Average, Centroid and Ward algorithms work.
- appreciate the advantages and limitation of hierarchical clustering techniques.

In this section, you will learn the basic concepts and methods of hierarchical clustering. The discussions include the concept of proximity matrix, different types of distances and their computation, and hierarchical clustering algorithm.

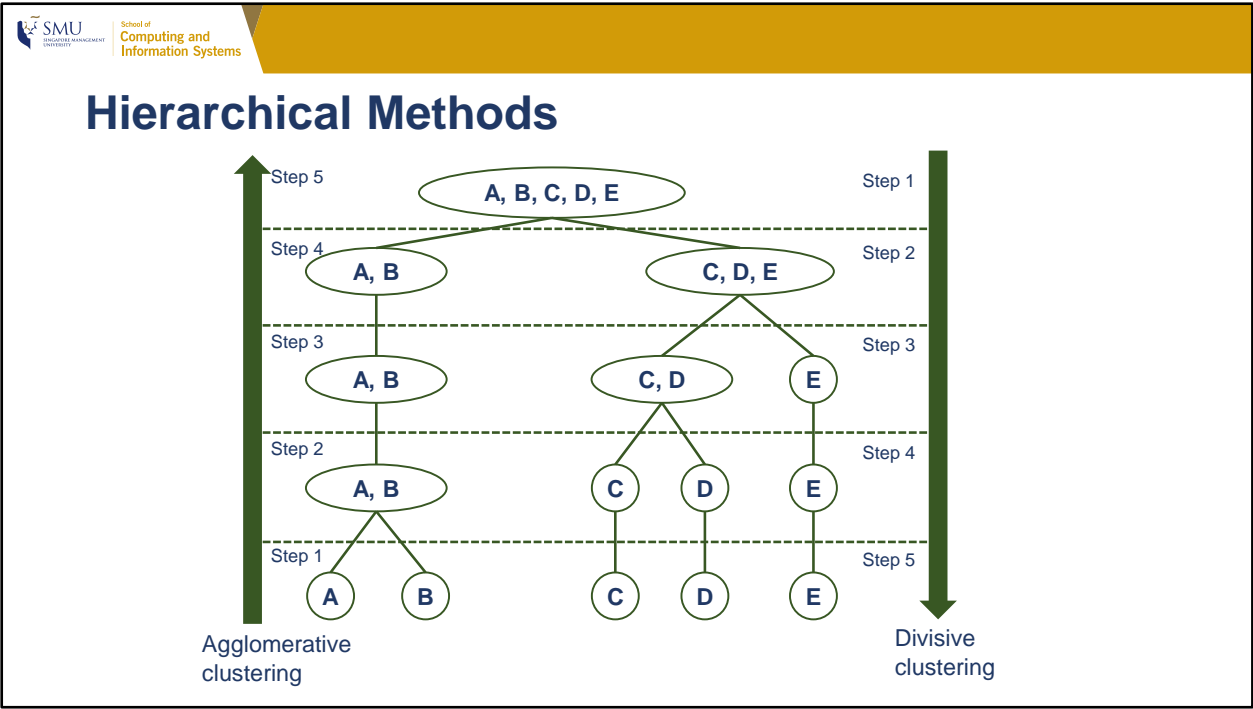
## Hierarchical Clustering

- Produces a set of nested clusters organized as a hierarchical tree



- Two main types of hierarchical clustering
  - Agglomerative
  - Divisive

**Hierarchical Clustering** refers to clustering methods that attempt to break data into a hierarchy of clusters. The procedures are characterised by the tree-like structure established in the course of the analysis. Generally, this can be done in two directions, namely: agglomerative and divisive.



Let us learn more about these two clustering procedures with reference to the figure above.

The procedures on the left show the flow of a typical **agglomerative clustering**. In this case, each item begins in its own cluster and clusters are subsequently joined until a criterion is met. Generally, in the case of agglomerative clustering, a distance metric is chosen such that the two clusters which are most near each other initially are chosen for the join.

The procedures on the right show the flow of a typical **divisive clustering**. The procedures begin with all items in one giant cluster and divide (split) them into subsequent clusters until some criteria are met. Generally, a distance metric is chosen and the split that maximises the distance between the two newly formed clusters is the split that is chosen.

Divisive procedures are quite rarely used in business analytics. We therefore concentrate on the agglomerative clustering procedures in the subsequent discussions.




## Basic Agglomerative Hierarchical Clustering Algorithm

- Basic algorithm is straightforward

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*

In general, agglomerative hierarchical clustering techniques are variations on a single approach: starting with individual data points such as clusters, successively merge the closest clusters until only one cluster remains.

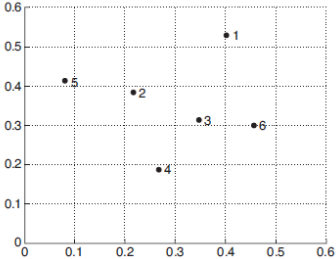
Generally speaking, there are two concepts at play in cluster analysis. Firstly, one must compute the distances between each data objects and store these distance values in a form of proximity matrix for the subsequent analysis. Secondly, one must determine the linkage. This is referred to the method that will be used to describe the relationship of a group of objects to another group of objects.



School of  
Computing and  
Information Systems

# Starting Situation


- A set of 6 two-dimensional data points



Point	<i>x</i> Coordinate	<i>y</i> Coordinate
p1	0.40	0.53
p2	0.22	0.38
p3	0.35	0.32
p4	0.26	0.19
p5	0.08	0.41
p6	0.45	0.30

Let’s learn about the steps involved in a basic agglomerative hierarchical clustering algorithm via a sample 6 two-dimensional data points. The data points are shown in the scatter plot and table.

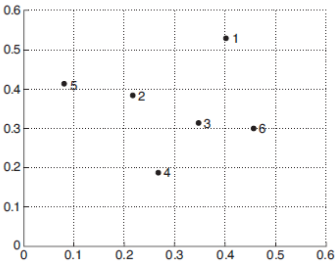
The key operation of agglomerative hierarchical clustering algorithm is the computation of the proximity between two clusters.



School of  
Computing and  
Information Systems

## What is Proximity Matrix?

- Measures of Similarity or Dissimilarity




	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

While cluster analysis sometimes uses the original data matrix, many clustering algorithms use a similarity matrix,  $S$ , or a dissimilarity matrix,  $D$ . For convenience, both matrices are commonly referred to as a proximity matrix,  $P$ . A proximity matrix,  $P$ , is an  $m$  by  $m$  matrix containing all the pairwise dissimilarities or similarities between the objects being considered. If  $x_i$  and  $x_j$  are the  $i$ th and  $j$ th objects, respectively, then the entry at the  $i$ th row and  $j$ th column of the proximity matrix is the similarity,  $s_{ij}$ , or the dissimilarity,  $d_{ij}$ , between  $x_i$  and  $x_j$ . For simplicity, we will use  $p_{ij}$  to represent either  $s_{ij}$  or  $d_{ij}$ . Figure above shows six data points and the corresponding data table.

Table above shows a typical proximity matrix (or distance matrix). Note that the diagonal is always empty and **NOT** zero. This is because it is not possible to have two data objects with similar variables' values. The non-diagonal values show the distances between pairs of objects . In this example, the proximity matrix is an 6 x 6 table and cell p1p2 indicates the distance between data object p1 and p2 is 0.23. In cluster analysis, there are three commonly used measures of dis(similarity), they are: Euclidean distance, City-block distance and Chebychev distance.





School of  
Computing and  
Information Systems

## Proximity matrix: Euclidean distance


- Euclidean distance formula:

$$D_{euclidean}(p_1, p_2) = \sqrt{(x_{p_1} - x_{p_2})^2 + (y_{p_1} - y_{p_2})^2}$$

- Proximity matrix of Euclidean distance

	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

A straightforward way to assess two data points' proximity is by drawing a straight line between them. For example, when we look at the scatter plot in the previous slide, we can easily see that the length of the line connecting data point p1 and p6 is much shorter than the line connecting p1 and p5. This type of distance is called Euclidean distance and is most commonly used type when it comes to analysing ratio or interval-scaled data.




School of  
Computing and  
Information Systems

## Proximity matrix: City-block distance

- City-block formula:
$$D_{City-block}(p_1, p_2) = |x_{p_1} - x_{p_2}| + |y_{p_1} - y_{p_2}|$$
- Proximity matrix of city-block distance

	p1	p2	p3	p4	p5	p6
p1		0.33	0.26	0.48	0.44	0.28
p2	0.33		0.19	0.23	0.17	0.31
p3	0.26	0.19		0.22	0.36	0.12
p4	0.48	0.23	0.22		0.40	0.30
p5	0.44	0.17	0.36	0.40		0.48
p6	0.28	0.31	0.12	0.30	0.48	

The city-block distance uses the sum of the variables’ absolute differences. This is often called the Manhattan metric as it is akin to the walking distance between two points in a city like New York’s Manhattan district.



School of  
Computing and  
Information Systems

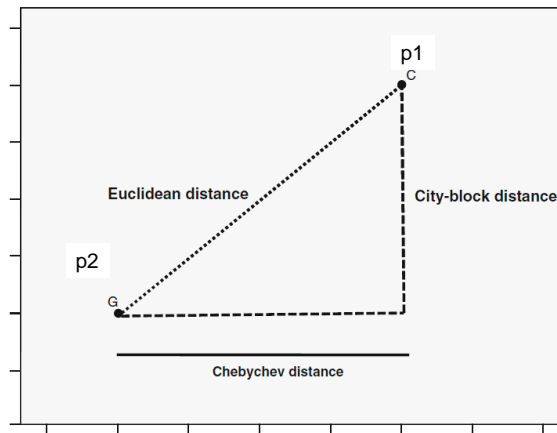
## Proximity matrix: Chebychev distance

- Chebychev distance formula:
$$D_{Chebychev}(p_1, p_2) = \max(|x_{p_1} - x_{p_2}|, |y_{p_1} - y_{p_2}|)$$
- Proximity matrix of Chebychev distance

	p1	p2	p3	p4	p5	p6
p1		0.18	0.21	0.34	0.32	0.23
p2	0.18		0.13	0.19	0.14	0.23
p3	0.21	0.13		0.13	0.27	0.10
p4	0.34	0.19	0.13		0.22	0.19
p5	0.32	0.14	0.27	0.22		0.37
p6	0.23	0.23	0.10	0.19	0.37	

When working with metric (or ordinal) data, data analysts frequently use the Chebychev distance, which is the maximum of the absolute difference in the clustering variables' values.


## Comparing Distance Measures



This figure summarises the interrelation between these three distance measures regarding two data points, p1 and p2 from our example.

- The Euclidean distance represents the slope distance between p1 and p2.
- The City-block distance represents the summation of the vertical and horizontal distances from p1 to p2.
- Lastly, the Chebychev distance represents the horizontal distance from p2 to the base of p1. Important to note that this is only true if the horizontal distance is longer than the vertical distance.

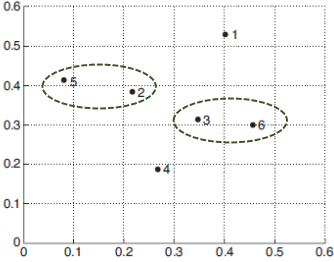




School of  
Computing and  
Information Systems

# Merging the clusters

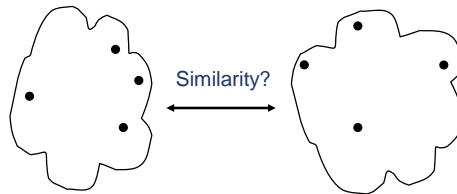
- Which clustering algorithm to apply?



	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	


After having chosen the distance matrix, we need to decide which clustering algorithm to apply in order to merge the individual object clusters into higher order clusters.

## Agglomerative Hierarchical Clustering Algorithms



- MIN
- MAX
- Group Average
- Distance Between Centroids
- Ward's minimum variance method

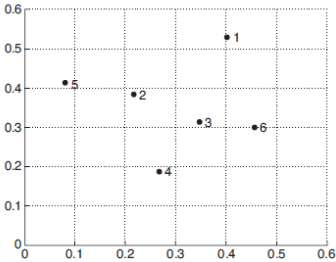
There are five commonly used hierarchical clustering algorithms. They are: Min, Max, Group Average, Centroid and Ward. None of these five methods is uniformly the best. In practice, it's advisable to try several methods and then compare the results to form an overall judgment about the final formation of clusters. In the subsequent slides, the implementation of each of this algorithm will be discussed in details.



School of  
Computing and  
Information Systems

## Agglomerative Hierarchical Clustering Algorithms

- Single linkage (Min) algorithm



	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

In **single linkage (MIN)**, we define the distance between two clusters to be the **minimum** distance between any single data point in the first cluster and any single data point in the second cluster. On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest single linkage distance.

Mathematically, the distance between cluster {3,6} and {2,5} is given by

$$\begin{aligned} dist(\{3,6\}, \{2,5\}) &= \min(dist(3,2), dist(6,2), dist(3,5), dist(6,5)) \\ &= \min(0.14, 0.24, 0.28, 0.39) \\ &= 0.14 \end{aligned}$$

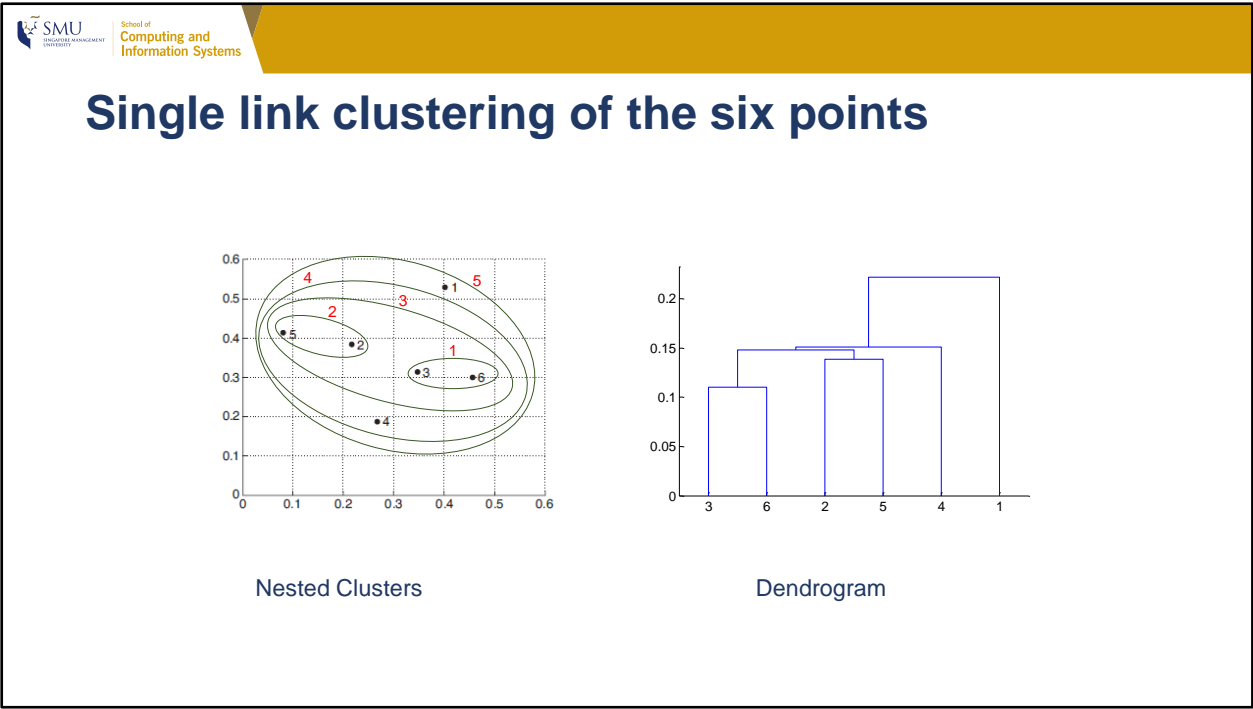
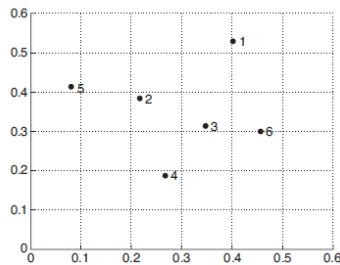


Figure above shows the results of applying the single link technique to the six points dataset.

The height at which two clusters are merged in the dendrogram reflects the distance of the two cluster. For example at stage 1, the distance between p3 and p6 is 0.11, and that is the height at which they are joined into one cluster in the dendrogram. At stage 2, p2 and p5 are then merge to form a cluster. These two newly formed clusters are merged to form a bigger cluster at stage 3. This cluster then merges with p4 and p1 at stage 4 and stage 5 respectively.

## Agglomerative Hierarchical Clustering Algorithms

- Complete linkage (or Max) algorithm



	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

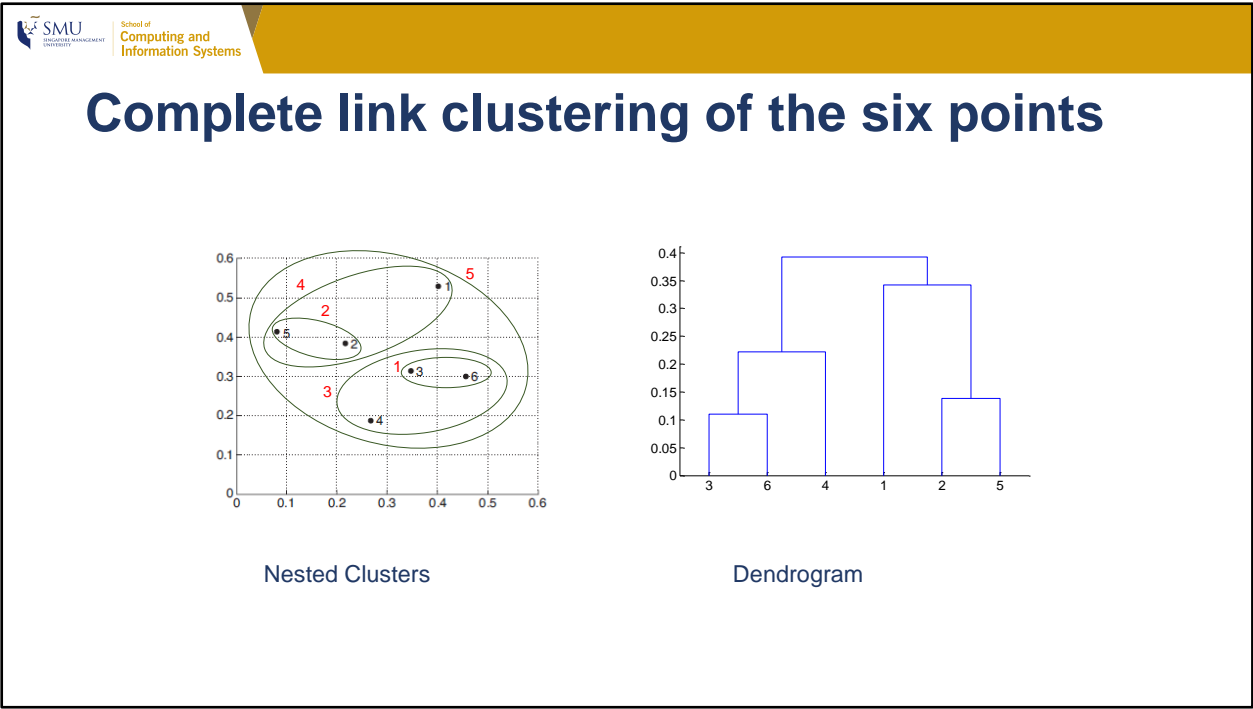
In **complete linkage (or MAX)**, we define the distance between two clusters to be the **maximum** distance between any single data point in the first cluster and any single data point in the second cluster. On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest complete linkage distance.

As with the single link, point 3 and 6 are merged first. However, {3,6} is merged with {4} instead of {2,5} or {1} because


$$\begin{aligned} \text{dist}(\{3,6\}, \{4\}) &= \max(\text{dist}(3,4), \text{dist}(6,4)) \\ &= \\ \max(0.16, 0.22) &= 0.22 \end{aligned}$$

$$\begin{aligned} \text{dis}(\{3,6\}, \{2,5\}) &= \max(\text{dist}(3,2), \text{dist}(6,2), \text{dist}(3,5), \text{dist}(6,5)) \\ &= \max(0.14, 0.24, 0.28, 0.39) \\ &= 0.39 \end{aligned}$$

$$\begin{aligned} \text{dist}(\{3,6\}, \{1\}) &= \max(\text{dist}(3,1), \text{dist}(6,1)) \\ &= \max(0.22, 0.24) \\ &= 0.24 \end{aligned}$$



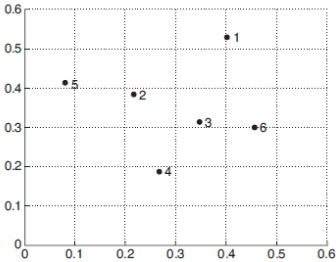
By using complete link algorithm, p3 and p6 merge to form a cluster in step 1. In step 2, p2 and p5 merge to form a cluster. However, instead of merging with the newly formed cluster of p2 and p5, The cluster of p3 and p6 merges with p4 and the cluster of p2 and p5 merges with p1 in step 3 and 4 respectively. At step five, theses two newly formed clusters merge into one big cluster.



School of  
Computing and  
Information Systems

# Agglomerative Hierarchical Clustering Algorithms

- Average linkage algorithm



	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

In **average linkage**, we define the distance between two clusters to be the **average** distance between data points in the first cluster and data points in the second cluster. On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest average linkage distance.

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

$$proximity(Cluster_i, Cluster_j) = \frac{\sum_{p_i \in Cluster_i, p_j \in Cluster_j} proximity(p_i, p_j)}{m_i * m_j}$$

Where  $m_i$  and  $m_j$  are the size of cluster  $i$  and  $j$  respectively.

Figure above shows how the group average algorithm works.

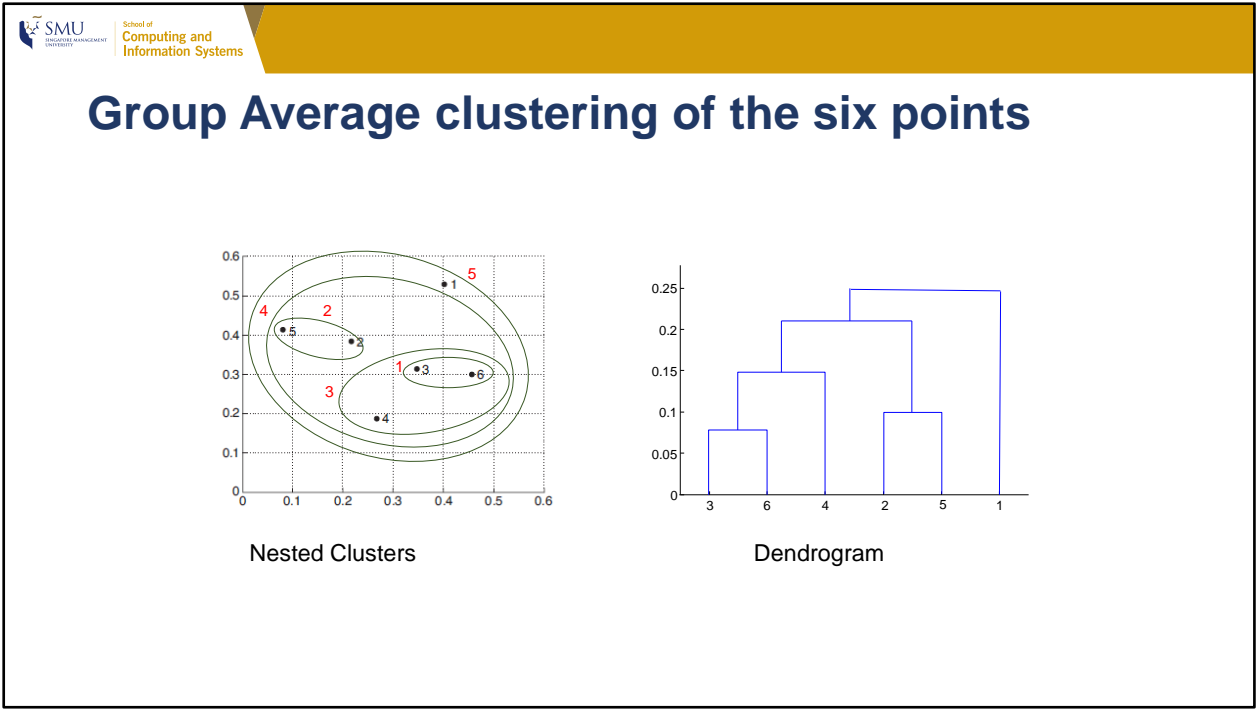
$$\begin{aligned} dist(\{3,6,4\}, \{1\}) &= \frac{(0.22 + 0.37 + 0.24)}{3 \times 1} \\ &= 0.28 \end{aligned}$$

$$\begin{aligned} dist(\{2,5\}, \{1\}) &= \frac{(0.23 + 0.34)}{2 \times 1} \\ &= 0.29 \end{aligned}$$


$$\begin{aligned} dist(\{3,6,4\}, \{2,5\}) &= \frac{0.14 + 0.28 + 0.24 + 0.39 + 0.19 + 0.28}{3 \times 2} \\ &= 0.26 \end{aligned}$$

Because  $dist(\{3, 6, 4\}, \{2, 5\})$  is smaller than  $dist(\{3, 6, 4\}, \{1\})$  and  $dist(\{3, 6, 4\}, \{1\})$  and  $dist(\{2, 5\}, \{1\})$ , cluster  $\{3, 6, 4\}$  and  $\{2, 5\}$  are merged at the fourth stage.





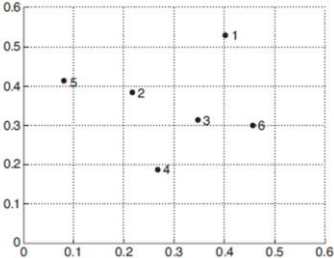
By using the group average algorithm, p3, p6 and p2, p5 merge to form clusters at stage 1 and 2 respectively. At stage 3, cluster (p3,p6) merges with p4. Then, this newly formed cluster merges with cluster (p2, p5). Finally, the cluster (p3, p6, p4, p2, p5) merges with p1 to form the biggest cluster.



School of  
Computing and  
Information Systems

# Agglomerative Hierarchical Clustering Algorithms

- Ward algorithm



	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

This method does not directly define a measure of distance between two points or clusters. It is an **ANOVA** based approach. At each stage, those two clusters merge, which provides the smallest increase in the combined error sum of squares from one-way univariate ANOVAs that can be done for each variable with groups defined by the clusters at that stage of the process.

Ward’s method joins clusters to maximise the likelihood at each level of the hierarchy under the assumptions of multivariate normal mixtures, spherical covariance matrices, and equal sampling probabilities.

Ward’s method tends to join clusters with a small number of observations and is strongly biased towards producing clusters with approximately the same number of observations. It is also very sensitive to outliers.

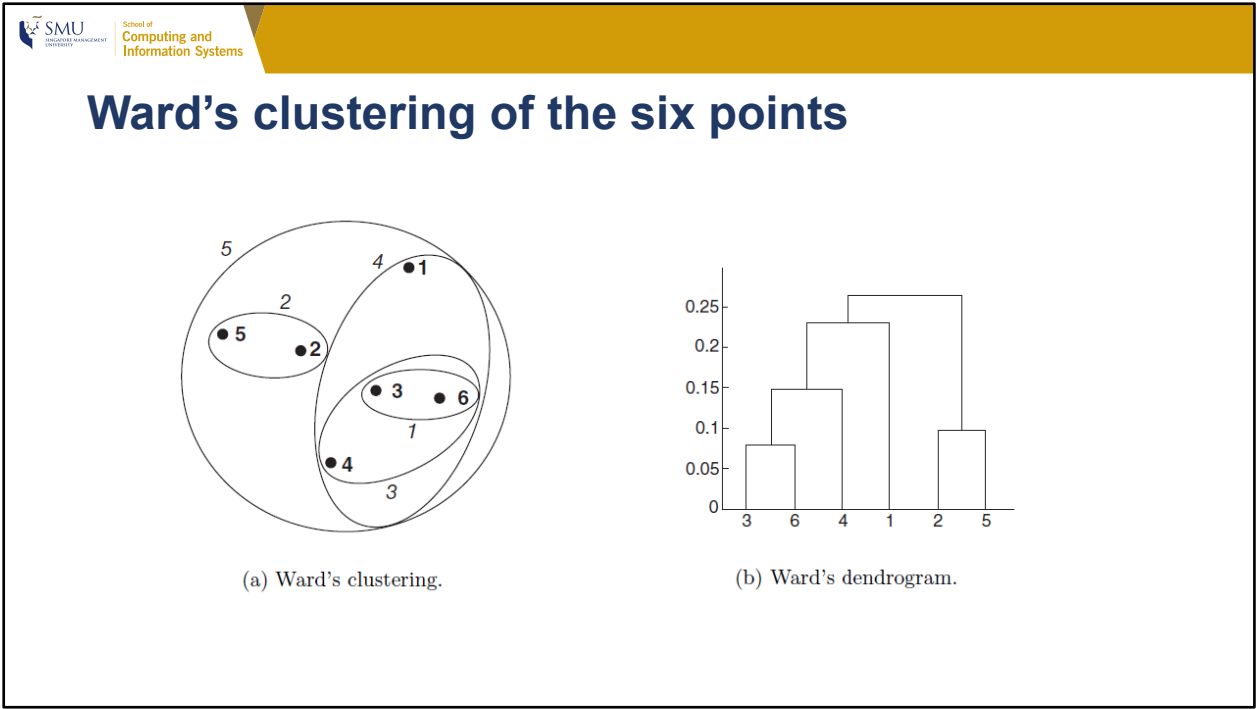



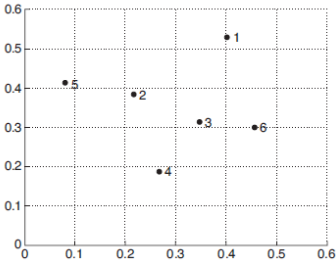
Figure above shows the results of applying Ward's method to the 6-point dataset. The clustering that is produced is different from those produced by single link, complete link, and group average.



School of  
Computing and  
Information Systems

## Agglomerative Hierarchical Clustering Algorithms

- Centroid algorithm

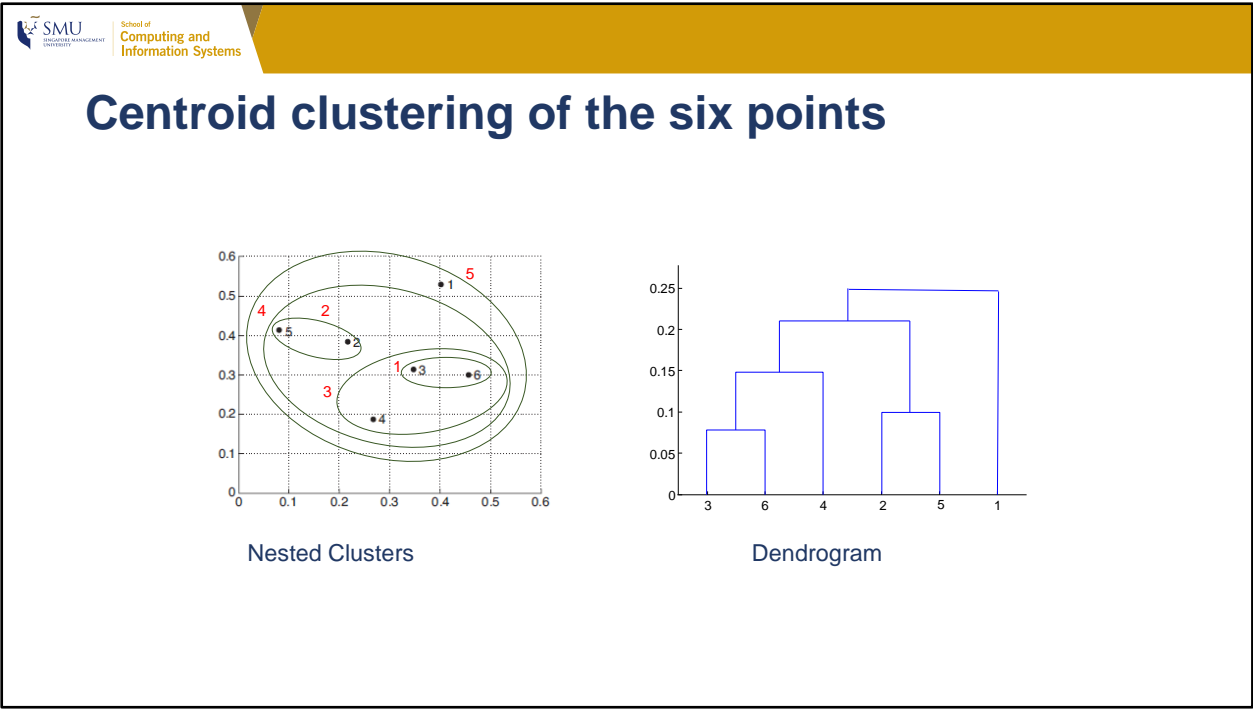


	p1	p2	p3	p4	p5	p6
p1		0.23	0.22	0.37	0.34	0.24
p2	0.23		0.14	0.19	0.14	0.24
p3	0.22	0.14		0.16	0.28	0.10
p4	0.37	0.19	0.16		0.28	0.22
p5	0.34	0.14	0.28	0.28		0.39
p6	0.24	0.24	0.10	0.22	0.39	

In **centroid method**, the distance between two clusters is the distance between the two **mean vectors** of the clusters. At each stage of the process, we combine the two clusters that have the smallest centroid distance.

When centroid algorithm is used, it is important for us to note the possibility of inversions. Two clusters that are merged may be more similar (less distant) than the pair of clusters that were merged in a previous step.

The centroid method is more robust to **outliers**.



By using the group average algorithm, p3, p6 and p2, p5 merge to form clusters at stage 1 and 2 respectively. At stage 3, cluster (p3,p6) merges with p4. Then, this newly formed cluster merges with cluster (p2, p5). Finally, the cluster (p3, p6, p4, p2, p5) merges with p1 to form the biggest cluster.

## Strengths of Hierarchical Clustering

- Do not have to assume any particular number of clusters
  - Any desired number of clusters can be obtained by 'cutting' the dendrogram at the proper level
- They may correspond to meaningful taxonomies
  - Example in biological sciences (e.g., animal kingdom, phylogeny reconstruction, ...)

## Hierarchical Clustering: Problems and Limitations

- Once a decision is made to combine two clusters, it cannot be undone
- No objective function is directly minimized
- 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

One downside of hierarchical clustering is that they have large storage requirements, and they can be computationally intensive. This is especially true for big data. These complex algorithms are about four times the size of the k-means clustering. Also, merging can't be reversed, which can create a problem if you have noisy, high-dimensional data.

## Cluster Analysis Best Practice

Upon completion of this section, student will be able to:

- Describe the cluster analysis process.
- Select appropriate clustering variables.
- Investigate the distribution of the cluster variables
- Perform variable standardisation
- Perform variable transformation
- Interpret clusters

In this section, you will learn the best practices in cluster analysis. Upon completion of this chapter, you will be able to

- understand the cluster analysis process framework and to implement a cluster analysis project by using this framework.
- learn key considerations in selecting clustering variable and be able to apply these understanding to select appropriate cluster variables.
- investigate the distribution of the cluster variables, to detect outliers and multi-collinearity by using appropriate interactive data exploration and analysis techniques.
- Perform variable standardisation and/or variable transformation
- Interpret the clusters using dendrogram, heatmap and parallel coordinates.



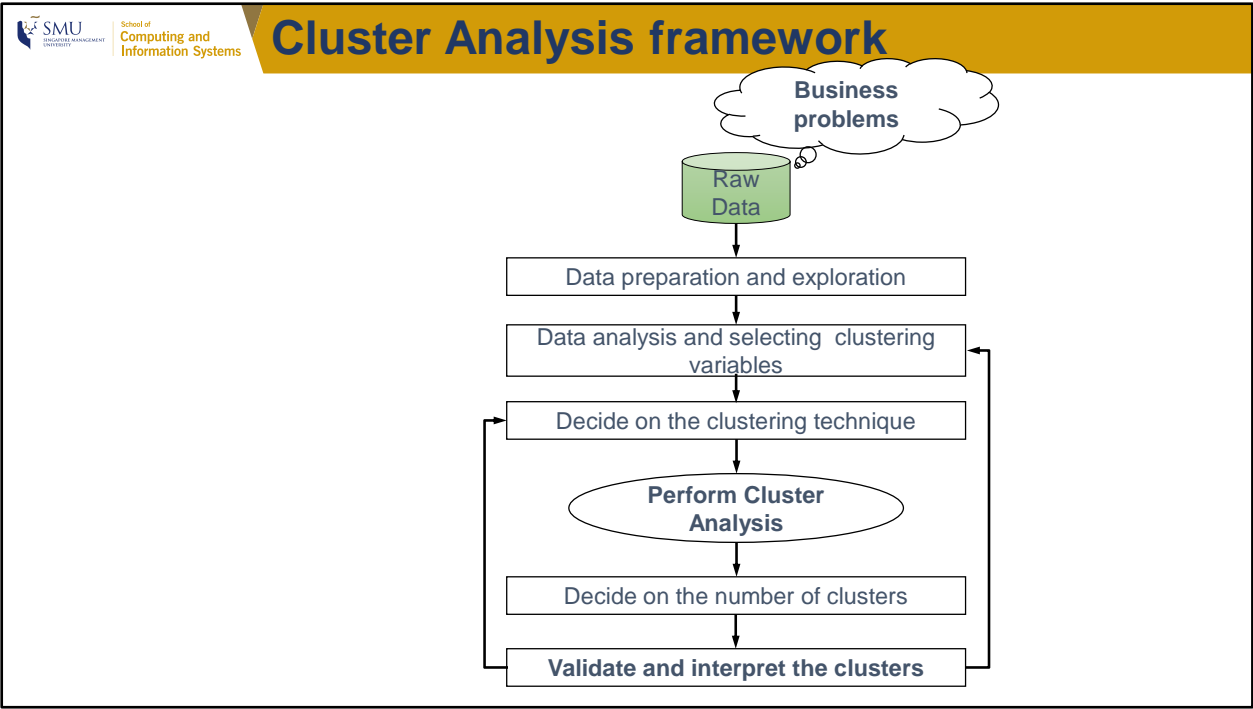


Figure above shows the steps involved with a cluster analysis. One of the greatest challenge in applied cluster analysis is to find actionable business outcome. It is not particularly difficult to find groups (or clusters) within a given dataset; indeed, in this lesson we see several ways to do this, all of which “succeed” according to one statistical criterion or another. Rather, the difficulty is to ensure that the outcome is meaningful for a particular business need.

Before embarking into a cluster analysis project, we strongly urge you ask a few question along the following lines. If you were to find clusters, what would you do about them? Would anyone in your organization use them? Why and how? Are the differences found large enough to be meaningful for your business? Among various solutions you might find, are there organizational efforts or politics that would make one solution more or less influential than another?

Another very important message from the framework above is that statistical methods such as cluster analysis are only part of the answer. It often happens that a “stronger” statistical solution poses complexity that makes it impossible to implement in business context while slightly “weaker” solution illuminates the data with a clear story and fits the business context so well that it can have a broad influence.

To maximize chances of finding such as model, we recommend that you, as a data analyst always expects, and prepares management to understand, the need to

iterate analyses. A cluster analysis project is not a matter of “running a clustering study” or “doping cluster analysis on data”. Rather, it is likely to take multiple rounds of data collection and analysis to determine the important data that should be collected in the first place, to refine and test the solutions, and to conduct rounds of interpretation with business stakeholders to ensure that the results are actionable.

## Decide on the Clustering Variables

Avoid “Garbage-In, Garbage-Out”

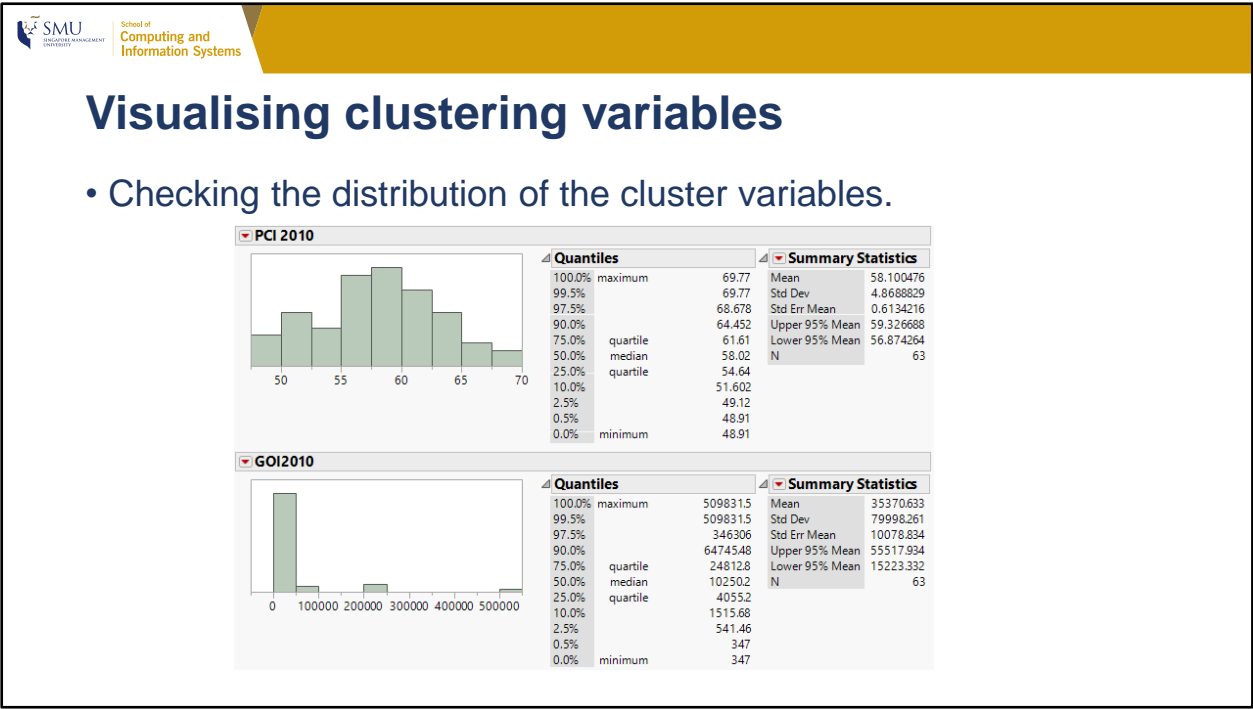


Cluster analysis has no mechanism for differentiating between relevant and irrelevant variables. Therefore the choice of variables included in a cluster analysis must be underpinned by conceptual considerations. This is very important because the clusters formed can be very dependent on the variables included.

There are several type of clustering variables and these can be classified into:

- General, independent of products, services or circumstances such as demographic and socio-economic profile of customers.
- Specific, related to both the customer and product, services and/or particular circumstance such as customer status, purchasing frequency, store and brand loyalty.

Whichever clustering variables are chosen, it is important to select those that provide a clear-cut differentiation between the segments regarding a specific managerial objective. More precisely, criterion validity is of special interest; that is, the extent to which the “independent” clustering variables are associated with one or more “dependent” variables not included in the analysis. Criterion variables usually relate to some aspect of behaviour, such as purchase intention or usage frequency.



## Variable Standardisation Techniques

- Z-score

$$Z(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

- Min-Max

$$MM(x_{ij}) = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}}$$

- Decimal scaling

$$DS(x_{ij}) = \frac{x_{ij}}{10^c}$$

*where  $c$  is the smallest integer  
such that  $\max[|DS(x_{ij})|] < 1$*

Before you perform a cluster analysis on coordinate data, it is necessary to consider scaling or transforming the variables since variables with large variances tend to have a larger effect on the resulting clusters than variables with small variances do. There are at least three commonly used variable standardisation methods, they are:

- Z-score: The Z-score is a form of standardisation used for transforming normal variants to standard score form.
- Min-max normalisation: Min-Max normalisation is the process of taking data measured in its engineering units and transforming it to a value between 0.0 and 1.0. Type equation here.
- Decimal scaling normalisation: normalise by moving the decimal points of values of feature  $x$ . The number of decimal points moved depends on the maximum absolute value of  $x$ . For example, suppose the range of attribute  $x$  is -500 to 45. The maximum absolute value of  $x$  is 500. To normalise by decimal scaling, we will divide each value by 1000 ( $c=3$ ). In this case, -500 becomes -0.5 while 45 will becomes 0.045.

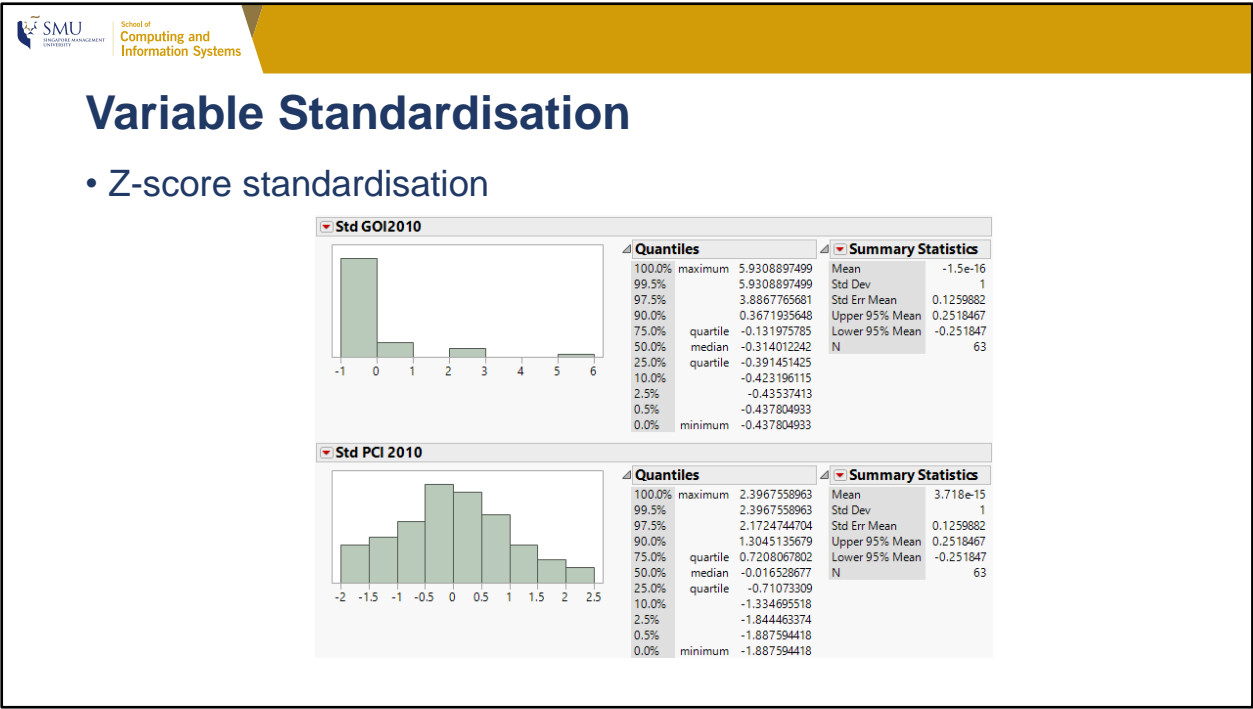
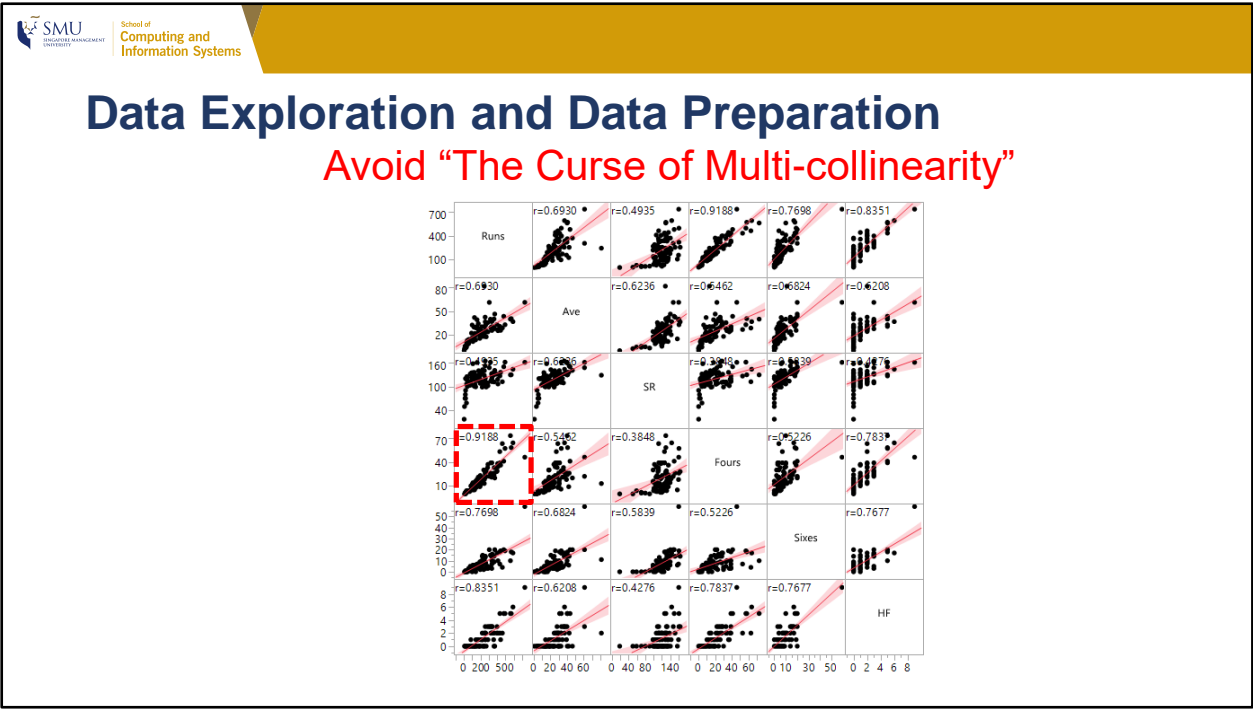
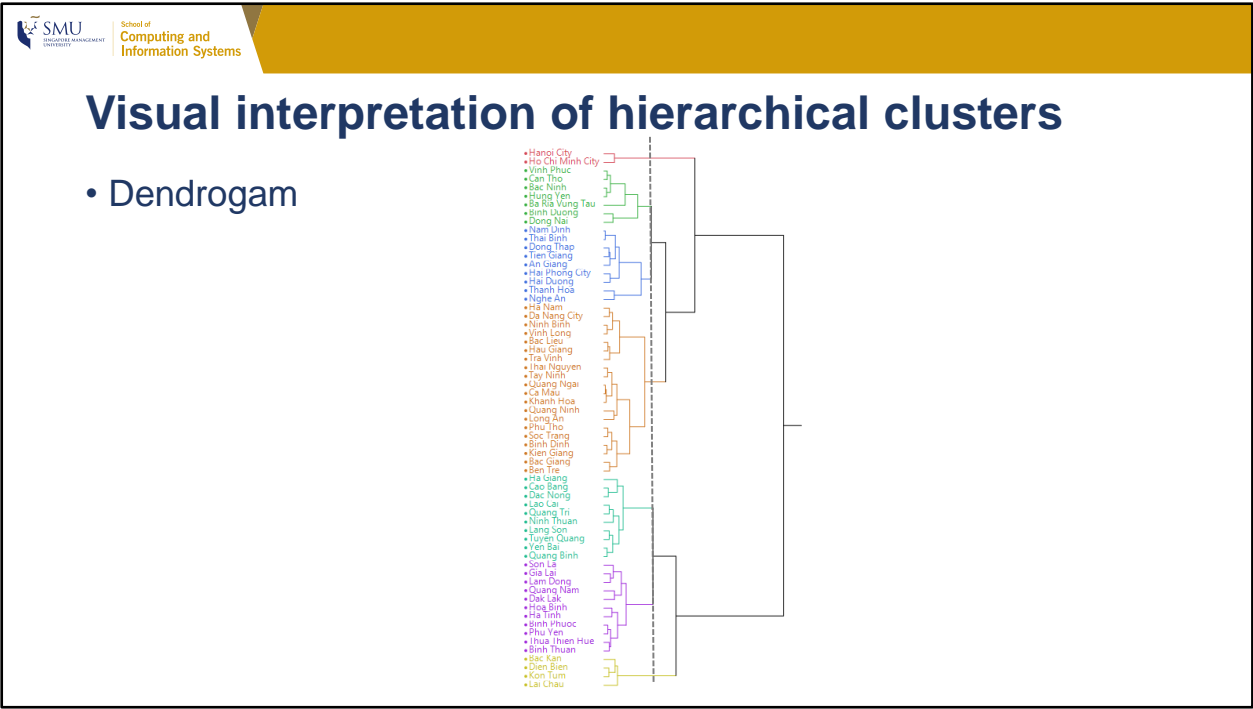


Figure above shows the distributions of GIO2010 and CPI2010 after z-score transformation. Notice that the mean and standard deviation of both variables are 0 and 1 respectively. Both of them are in comparable measurement range now.



You should also avoid using an abundance of cluster variables, as this increases the odds that the variables are no longer dissimilar. If there is a high degree of collinearity between the variables, they are not sufficiently unique to identify distinct clusters. If highly correlated variables are used for cluster analysis, specific aspects covered by these variables will be overrepresented in the clustering solution. In this regard, absolute correlations above 0.85 are always problematic and should be avoided.

In this figure, scatter plot matrix, an exploratory data analysis technique, is used to detect variables that are highly correlated. The figure reveals that Runs and Fours are highly correlated with a  $r$  value of 0.9188.



A dendrogram is a tree-structured graph used to visualize the result of a hierarchical clustering calculation. The result of a clustering is presented either as the distance or the similarity between the clustered rows or columns depending on the selected distance measure.

The row dendrogram shows the distance or similarity between rows and which nodes each row belongs to, as a result of clustering. An example of a row dendrogram is shown above.

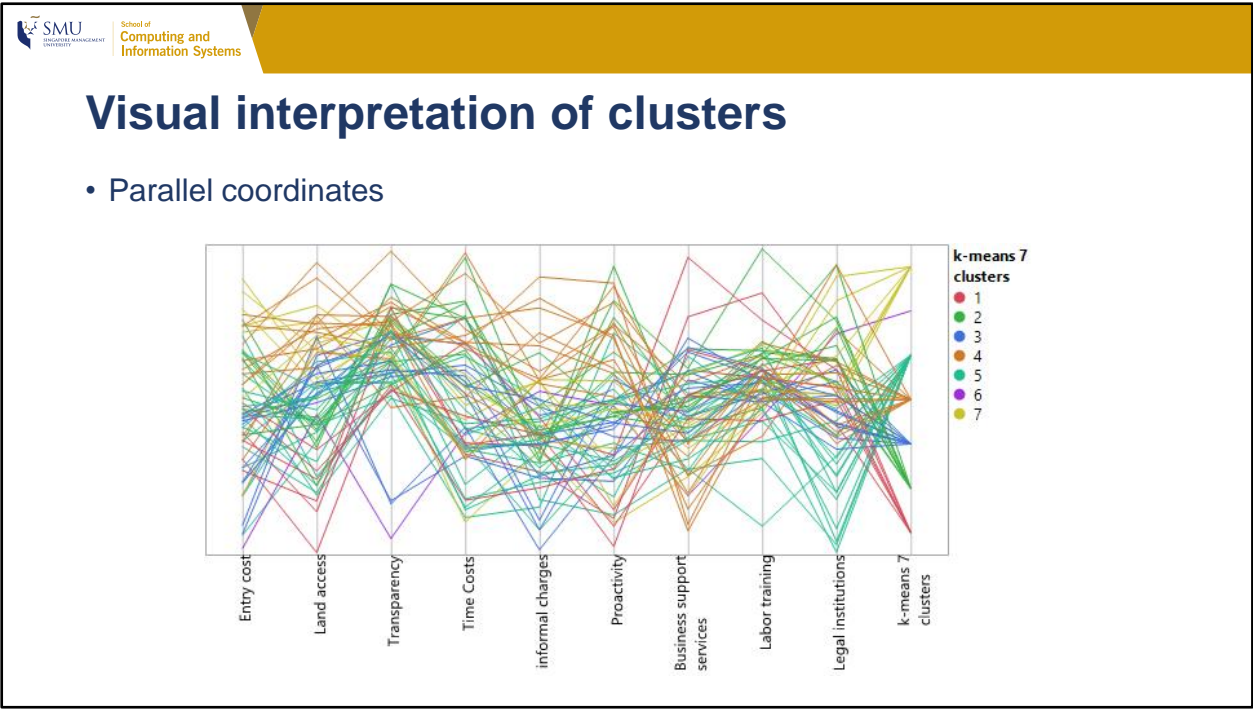
The individual rows in the clustered data are represented by the left-most nodes, the leaf nodes, in the row dendrogram. Each node in the dendrogram represents a cluster of all rows that lie to the left of it in the dendrogram. The right-most node in the dendrogram is therefore a cluster that contains all rows. The vertical dotted line is the pruning line, which can be dragged sideways in the dendrogram. The values next to the pruning line indicate the number of clusters starting from the current position of the line, as well as the calculated distance or similarity at that position. In the example above, there are seven clusters starting at the position of the pruning line. The top most cluster, indicated by red circles, contain two nodes, while the lower cluster contains only four nodes.

The horizontal axis of the dendrogram represents the distance or dissimilarity between clusters. The vertical axis represents the objects and clusters. The dendrogram is fairly simple to interpret. Remember that our main interest is in

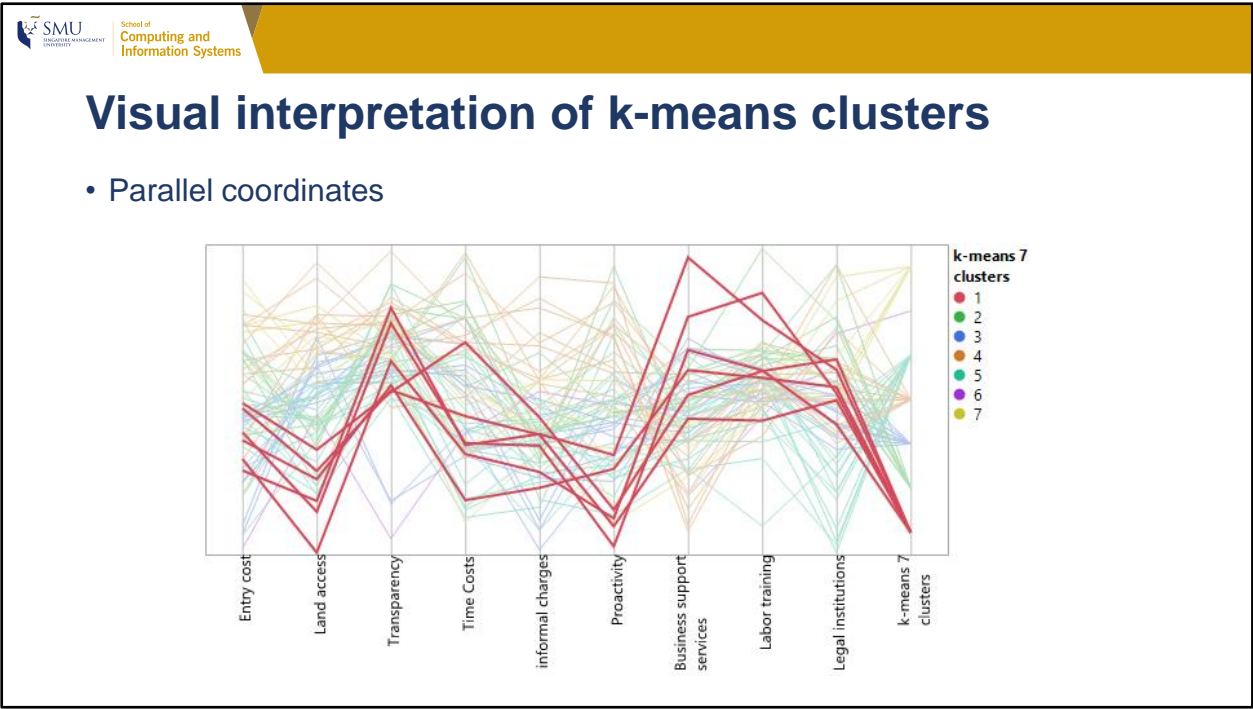


similarity and clustering. Each joining (fusion) of two clusters is represented on the graph by the splitting of a horizontal line into two horizontal lines. The horizontal position of the split, shown by the short vertical bar, gives the distance (dissimilarity) between the two clusters.

Looking at this dendrogram, you can see the three clusters as three branches that occur at about the same horizontal distance. The two outliers, 6 and 13, are fused in rather arbitrarily at much higher distances. This is the interpretation.



Parallel coordinates, a visualisation technique used to plot individual data elements across many dimensions. Each of the dimensions corresponds to a vertical axis and each data element is displayed as a series of connected points along the dimensions/axes.



Interactive is the key to the success used of parallel coordinates to discover meaning of a k-means clustering result. For example, by brushing Cluster 1, the parallel coordinates chart reveals that this cluster mainly represented by provinces with relatively low Land Access and Proactivity but relatively high in Business Support Services indices.

## Reference

- Cluster Analysis@wiki ([https://en.wikipedia.org/wiki/Cluster\\_analysis](https://en.wikipedia.org/wiki/Cluster_analysis))
- Hierarchical clustering@wiki ([https://en.wikipedia.org/wiki/Hierarchical\\_clustering](https://en.wikipedia.org/wiki/Hierarchical_clustering))
- Cluster Analysis: Basic Concepts and Algorithms (<http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf>)
- Brian S. Everitt, et. al. (2011) Cluster Analysis (5<sup>th</sup> Edition) John Wiley. (This book is available in SMU eLibrary)
- [Understanding Chebyshev Distance: A Comprehensive Guide](#)