TELKOM
DIGITAL TALENT
INCUBATOR 2020

# Modul 6: Clustering

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Module Overview

**Topics**

- Clustering Model
- 3 Types of Clustering
- Clustering Method
- Evaluating Clustering Model

**Activities**

- Group Discussion
- Coding Practice

TELKOM DIGITAL TALENT INCUBATOR 2020

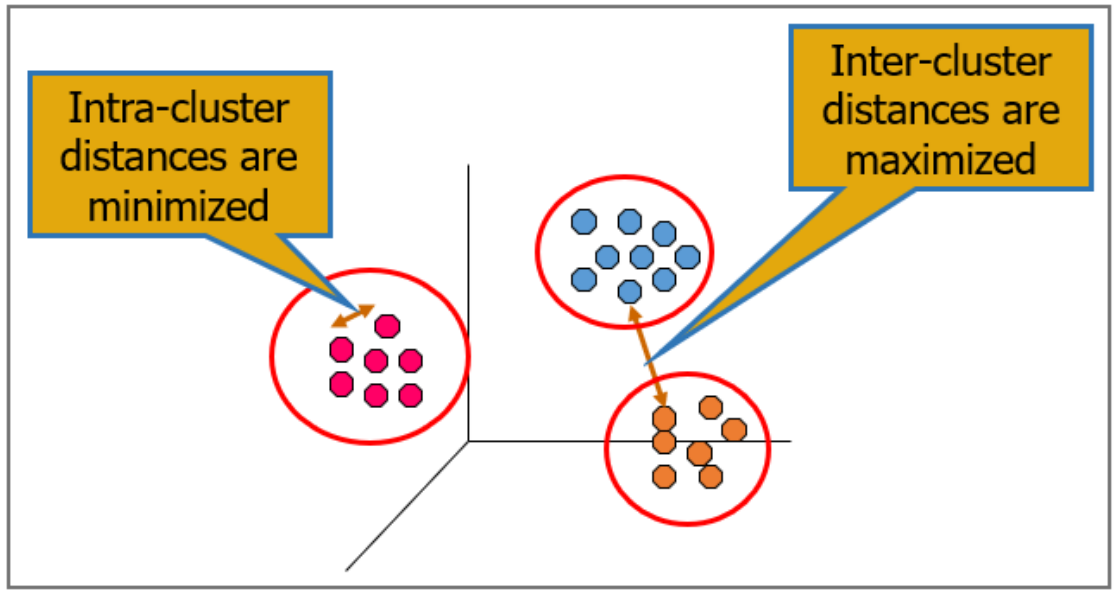
# Module Objectives

- Understand Clustering
- Understand type of Clustering
- Understand technique and Method of Clustering
- Understand how to evaluate clustering model

TELKOM DIGITAL TALENT INCUBATOR 2020

# Types of Clusters

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups.




Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Types of Clusters

- **Partitional clustering:** Partitional algorithms determine all clusters at once. They include:
  - K-means and derivatives
  - Fuzzy c-means clustering
  - QT clustering algorithm
- **Hierarchical algorithms:** these find successive clusters using previously established clusters.
  - Agglomerative ("bottom-up"): Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters.
  - Divisive ("top-down"): Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters

TELKOM DIGITAL TALENT INCUBATOR 2020

# Common Distance Measure

Distance measure will determine how the similarity of two elements is calculated and it will influence the shape of the clusters. They include:

1. The Euclidean distance (also called 2-norm distance) is given by:

$$d(x,y) = \sum_{i=1}^{p} |x_i - y_i|$$

1. The Manhattan distance (also called taxicab norm or 1-norm) is given by:

$$d(x,y) = \sqrt[2]{\sum_{i=1}^{p} |x_i - y_i|^2}$$

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Common Distance Measure

3. The maximum norm is given by:

$$d(x, y) = \max_{1 \le i \le p} | x_i - y_i |$$

3. The Mahalanobis distance corrects data for different scales and correlations in the variables.
4. Inner product space: The angle between two vectors can be used as a distance measure when clustering high dimensional data
5. Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

TELKOM DIGITAL TALENT INCUBATOR 2020

# K-Means Clustering

- The **k-means algorithm** is an algorithm to cluster n objects based on attributes into k partitions, where k < n.
- It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data.
- It assumes that the object attributes form a vector space.

TELKOM DIGITAL TALENT INCUBATOR 2020

# K-Means Clustering

- An algorithm for partitioning (or clustering) N data points into K disjoint subsets Sj containing data points so as to minimize the sum-of-squares criterion

$$J = \sum_{j=1}^{K} \sum_{n \in S_j} |x_n - \mu_j|^2,$$

where xn is a vector representing the the nth data point and uj is the geometric centroid of the data points in Sj.

Organized by:
CAE

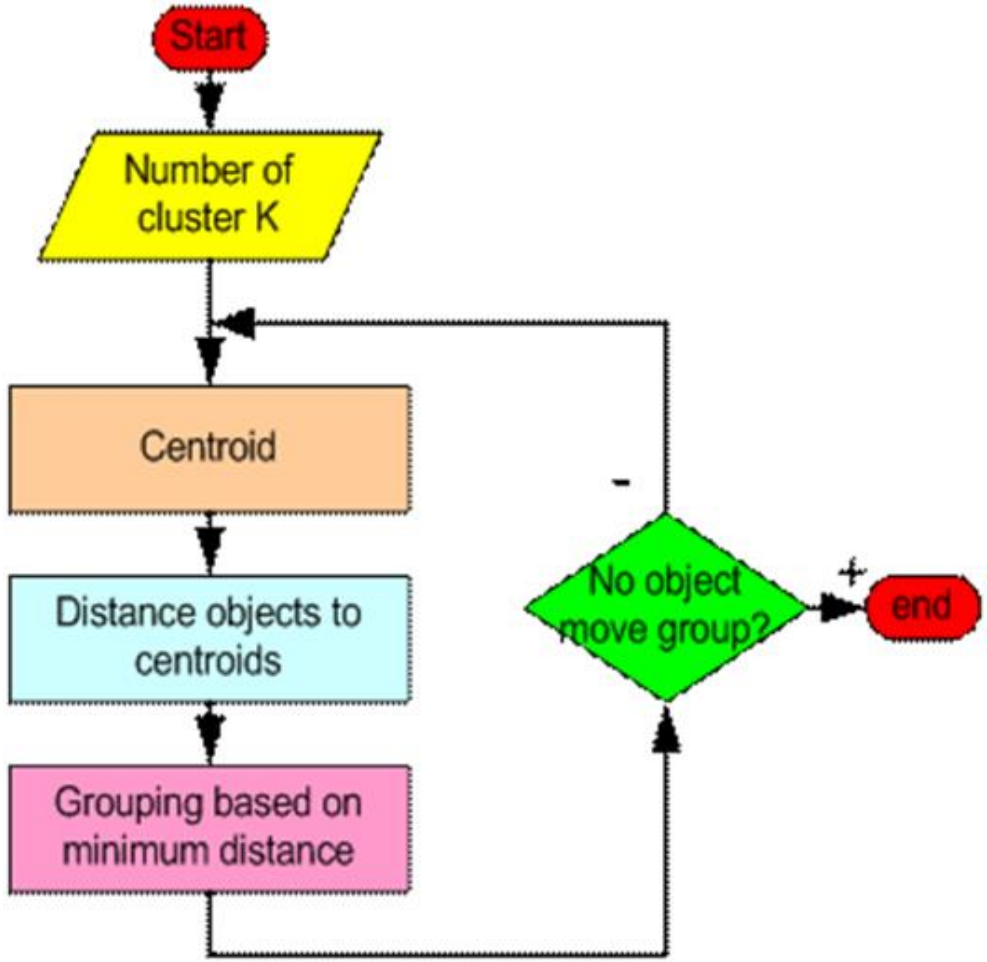
TELKOM DIGITAL TALENT INCUBATOR 2020

# K-Means Clustering

- Simply speaking k-means clustering is an algorithm to classify or to group the objects based on attributes/features into K number of group.
- K is positive integer number.
- The grouping is done by minimizing the sum of squares of distances between data and the corresponding cluster centroid.

TELKOM DIGITAL TALENT INCUBATOR 2020

# How it works?




Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# How it works?

- **Step 1:** Begin with a decision on the value of k = number of clusters .
- **Step 2:** Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly,or systematically as the following:
  - 1. Take the first k training sample as single - element clusters
  - 2. Assign each of the remaining (N-k) training sample to the cluster with the nearest centroid. After each assignment, recompute the centroid of the gaining cluster.

TELKOM DIGITAL TALENT INCUBATOR 2020

# How it works?

- **Step 3:** Take each sample in sequence and compute its distance from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.
- **Step 4:** Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

TELKOM DIGITAL TALENT INCUBATOR 2020

# Example

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

TELKOM DIGITAL TALENT INCUBATOR 2020

# Step 1

- Initialization: Randomly we choose following two centroids (k=2) for two clusters. In this case the 2 centroid are: m1=(1.0,1.0) and m2=(5.0,7.0).

| Individual | Variable 1 | Variable 2 |
|---|---|---|
| 1 | 1.0 | 1.0 |
| 2 | 1.5 | 2.0 |
| 3 | 3.0 | 4.0 |
| 4 | 5.0 | 7.0 |
| 5 | 3.5 | 5.0 |
| 6 | 4.5 | 5.0 |
| 7 | 3.5 | 4.5 |

| | Individual | Mean Vector |
|---|---|---|
| Group 1 | 1 | (1.0, 1.0) |
| Group 2 | 4 | (5.0, 7.0) |

TELKOM DIGITAL TALENT INCUBATOR 2020

# Step 2

- Thus, we obtain two clusters containing: {1,2,3} and {4,5,6,7}. Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0+1.5+3.0), \frac{1}{3}(1.0+2.0+4.0)\right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0+3.5+4.5+3.5), \frac{1}{4}(7.0+5.0+5.0+4.5)\right)$$

$$= (4.12, 5.38)$$

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 0 | 7.21 |
| 2 (1.5, 2.0) | 1.12 | 6.10 |
| 3 | 3.61 | 3.61 |
| 4 | 7.21 | 0 |
| 5 | 4.72 | 2.5 |
| 6 | 5.31 | 2.06 |
| 7 | 4.30 | 2.92 |

$$d(m_1, 2) = \sqrt{|1.0-1.5|^2 + |1.0-2.0|^2} = 1.12$$

$$d(m_2, 2) = \sqrt{|5.0-1.5|^2 + |7.0-2.0|^2} = 6.10$$

Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Step 3

- Now using these centroids we compute the Euclidean distance of each object, as shown in table.
- Therefore, the new clusters are: {1,2} and {3,4,5,6,7}
- Next centroids are: m1=(1.25,1.5) and m2 = (3.9,5.1)

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 1.57 | 5.38 |
| 2 | 0.47 | 4.28 |
| 3 | 2.04 | 1.78 |
| 4 | 5.64 | 1.84 |
| 5 | 3.15 | 0.73 |
| 6 | 3.78 | 0.54 |
| 7 | 2.74 | 1.08 |

TELKOM DIGITAL TALENT INCUBATOR 2020

# Step 4

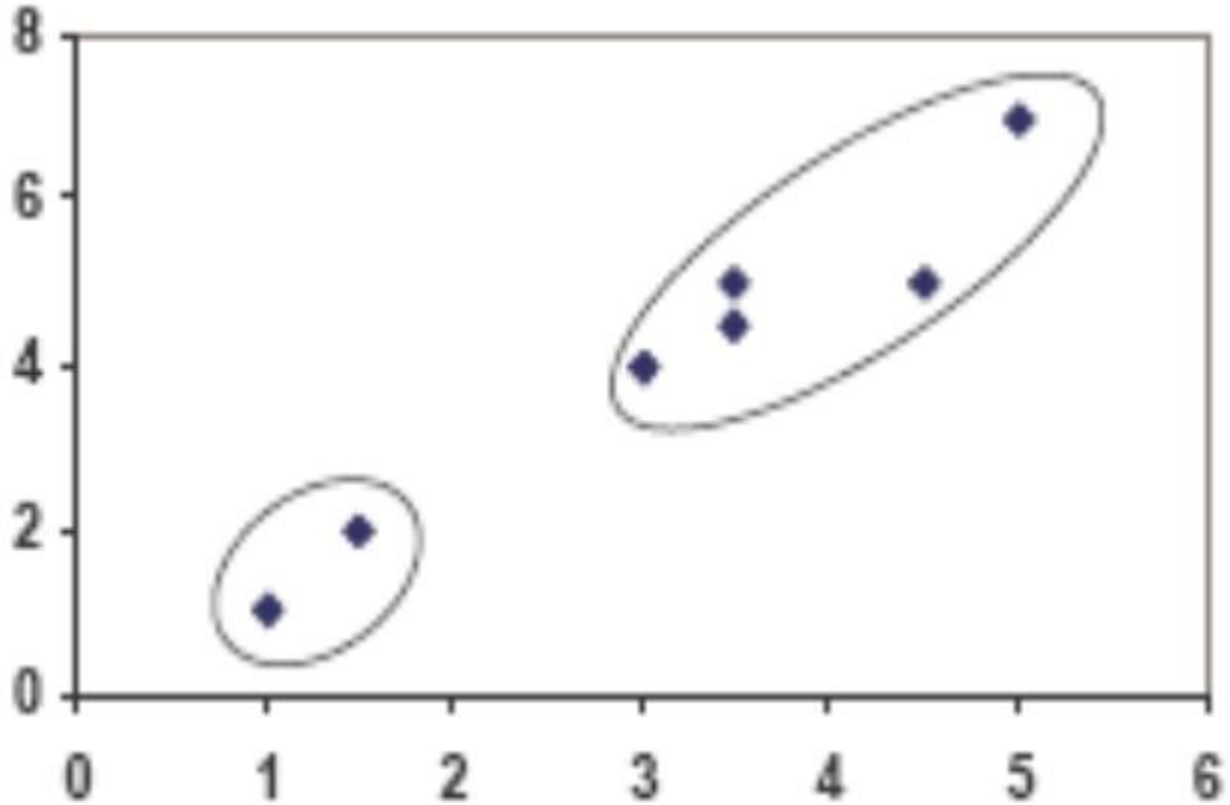
- The clusters obtained are: {1,2} and {3,4,5,6,7}
- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt here and final result consist of 2 clusters {1,2} and {3,4,5,6,7}.

| Individual | Centroid 1 | Centroid 2 |
|---|---|---|
| 1 | 0.58 | 5.02 |
| 2 | 0. 58 | 3.92 |
| 3 | 3.05 | 1.42 |
| 4 | 6.66 | 2.20 |
| 5 | 4.16 | 0.41 |
| 6 | 4.78 | 0.61 |
| 7 | 3.75 | 0.72 |

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Plot




Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Selecting optimal number of cluster

**Elbow Method**

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the total within-cluster sum of square (wss).
- Plot the curve of wss according to the number of clusters k.
- The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.

TELKOM DIGITAL TALENT INCUBATOR 2020

# Selecting optimal number of cluster

**Average silhouette method**

- Compute clustering algorithm (e.g., k-means clustering) for different values of k. For instance, by varying k from 1 to 10 clusters.
- For each k, calculate the average silhouette of observations (avg.sil).
- Plot the curve of avg.sil according to the number of clusters k.
- The location of the maximum is considered as the appropriate number of clusters.

TELKOM DIGITAL TALENT INCUBATOR 2020

# Weaknesses of K-Mean Clustering

- When the numbers of data are not so many, initial grouping will determine the cluster significantly.
- The number of cluster, K, must be determined beforehand. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.
- We never know the real cluster, using the same data, because if it is inputted in a different order it may produce different cluster if the number of data is few.
- It is sensitive to initial condition. Different initial condition may produce different result of cluster. The algorithm may be trapped in the local optimum.

TELKOM DIGITAL TALENT INCUBATOR 2020

# Applications of K-Mean Clustering

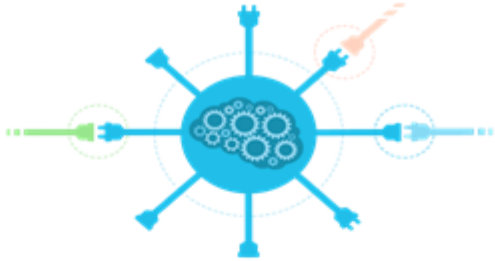
- It is relatively efficient and fast. It computes result at O(tkn), where n is number of objects or points, k is number of clusters and t is number of iterations.
- k-means clustering can be applied to machine learning or data mining
- Used on acoustic data in speech understanding to convert waveforms into one of k categories (known as Vector Quantization or Image Segmentation).
- Also used for choosing color palettes on old fashioned graphical display devices and Image Quantization.

TELKOM DIGITAL TALENT INCUBATOR 2020

# K-Means in Spotify

Spotify®

Artist on Spotify

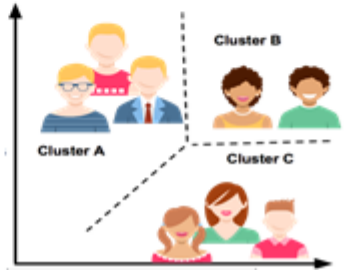


Get the data from Spotify API.


Process the data to extract audio features for each artist.

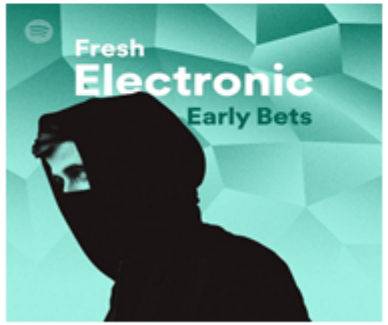
Apply k-means clustering to separate the artists into different groups.

K-Means




Analyze each feature for all the artists.

Well catered songs to each-user music preferences by categorizing the music



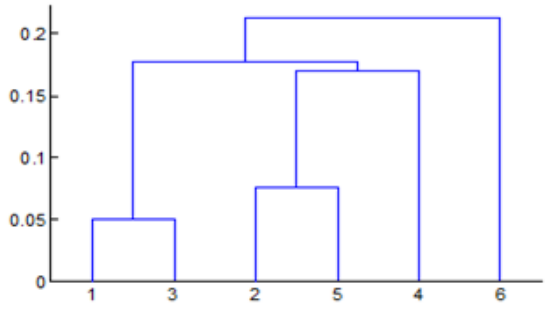

Organized by:

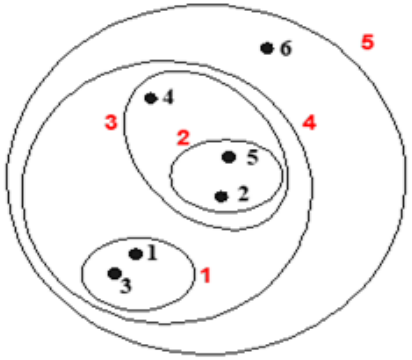
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# 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




Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Type of Hierarchical Clustering

**Agglomerative:**

- Start with the points as individual clusters
- At each step, merge the closest pair of clusters until only one cluster (or k clusters) left

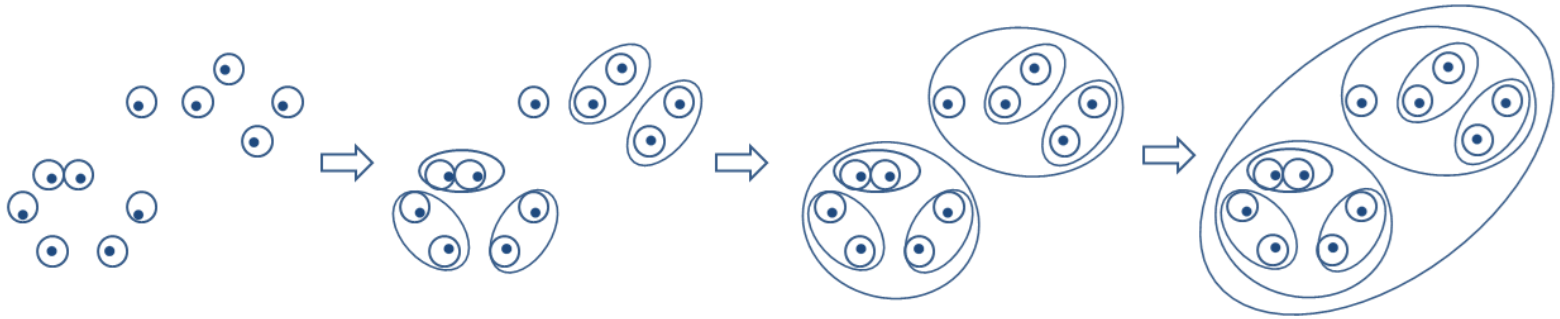
**Divisive:**

- Start with one, all-inclusive cluster
- At each step, split a cluster until each cluster contains a point (or there are k clusters)

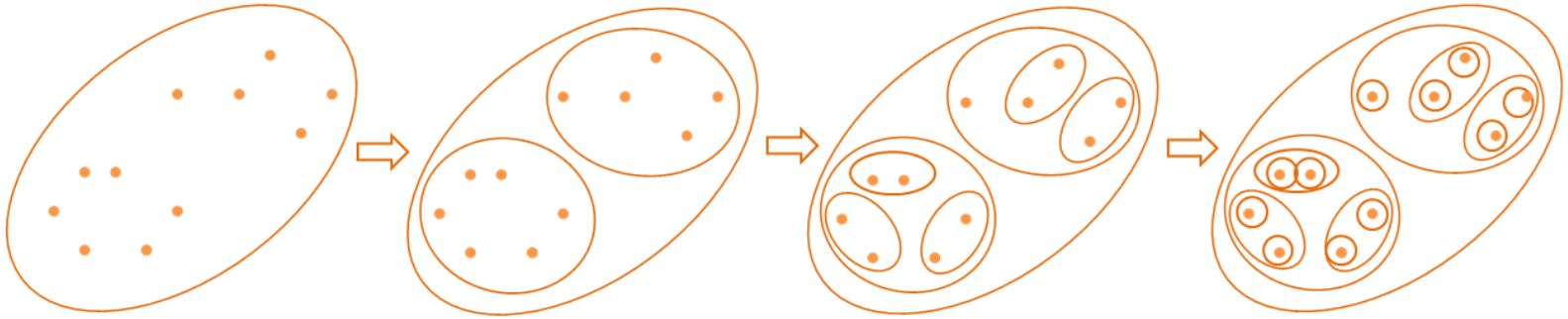
TELKOM DIGITAL TALENT INCUBATOR 2020

# Type of Hierarchical Clustering

Agglomerative Hierarchical Clustering



Divisive Hierarchical Clustering



Organized by:

CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Strength 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, …)

TELKOM DIGITAL TALENT INCUBATOR 2020

# Limitation of Hierarchical Clustering

- 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

TELKOM DIGITAL TALENT INCUBATOR 2020

# Cluster Validity

- For supervised classification we have a variety of measures to evaluate how good our model is Accuracy, precision, recall
- For cluster analysis, the analogous question is how 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

TELKOM DIGITAL TALENT INCUBATOR 2020

# Different Aspects of Cluster Validation

1. Determining the clustering tendency of a set of data, i.e., distinguishing whether non-random structure actually exists in the data.
2. Comparing the results of a cluster analysis to externally known results, e.g., to externally given class labels.
3. Evaluating how well the results of a cluster analysis fit the data without reference to external information.  - Use only the data
4. Comparing the results of two different sets of cluster analyses to determine which is better.
5. Determining the ‘correct’ number of clusters.

**For 2, 3, and 4, we can further distinguish whether we want to evaluate the entire clustering or just individual clusters.*

TELKOM DIGITAL TALENT INCUBATOR 2020

# Practice with Python

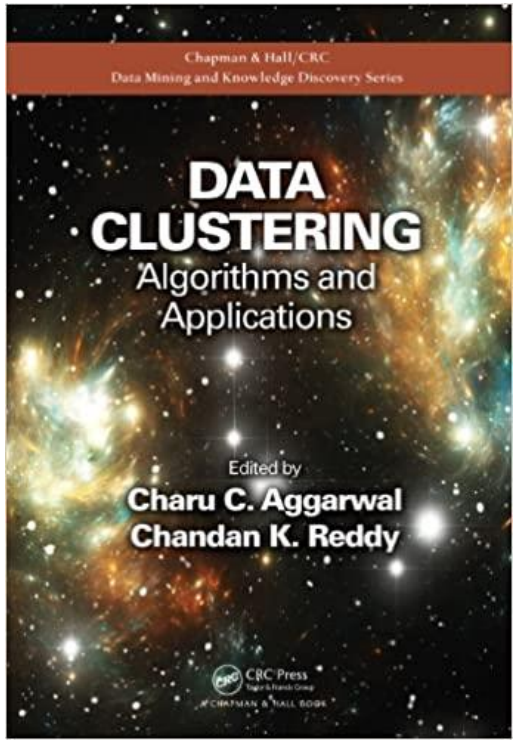
- Practice link: https://github.com/rc-dbe/dti

Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Further Readings

- Data Clustering: Algorithms and Applications (Chapman & Hall/CRC Data Mining and Knowledge Discovery Series Book 31)




Organized by:
CAE

TELKOM DIGITAL TALENT INCUBATOR 2020

# Assignment

- Create a clustering model using “Mall Customer Segmentation Data” dataset from Kaggle https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python
- Use Google Collab (or Jupyter Notebook if you want)
- Find insights from the clustering model that you created (Find the hidden spending patterns)
- Put the code in your GitHub
- Make it informative as possible