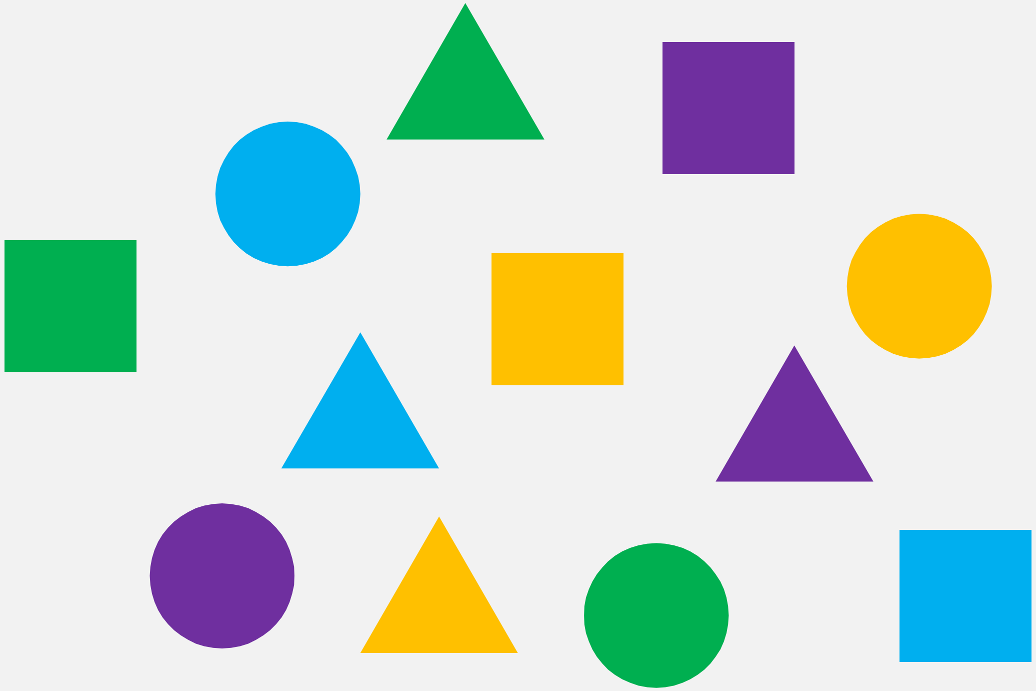
CLUSTER ANALYSIS

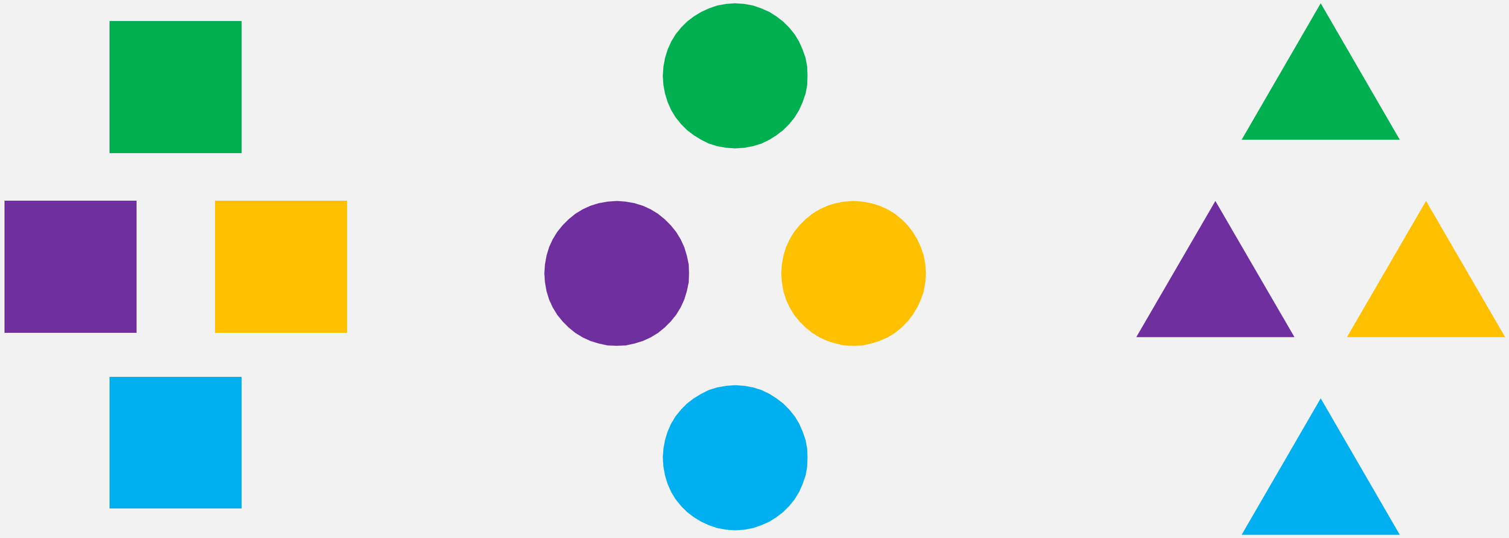
Introduction

- Clustering is usually one of the first tasks performed in most analytics projects.
- It helps data scientists to analyze individual clusters further.
- A cluster refers to a collection of data points aggregated together because of certain similarities.
- Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often according to some defined distance measure.

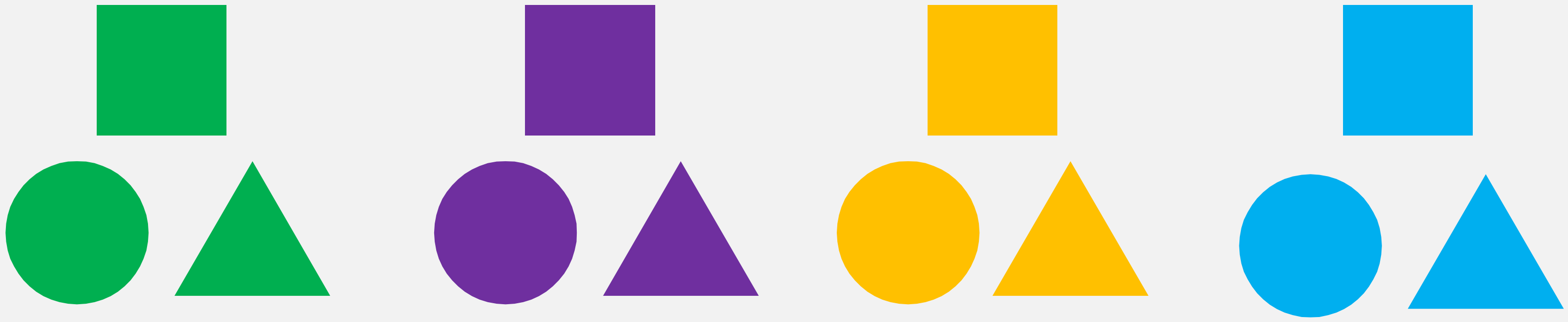
Classification



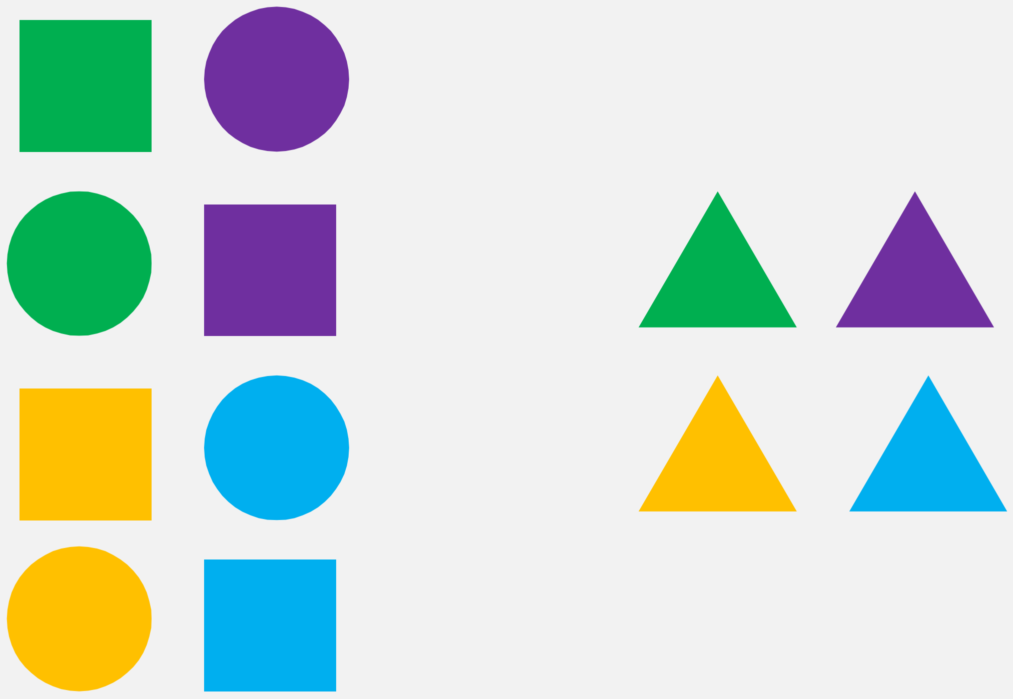
You may classify them by Shape



You may classify themby Colour



You may classify them by Sum of internal angles



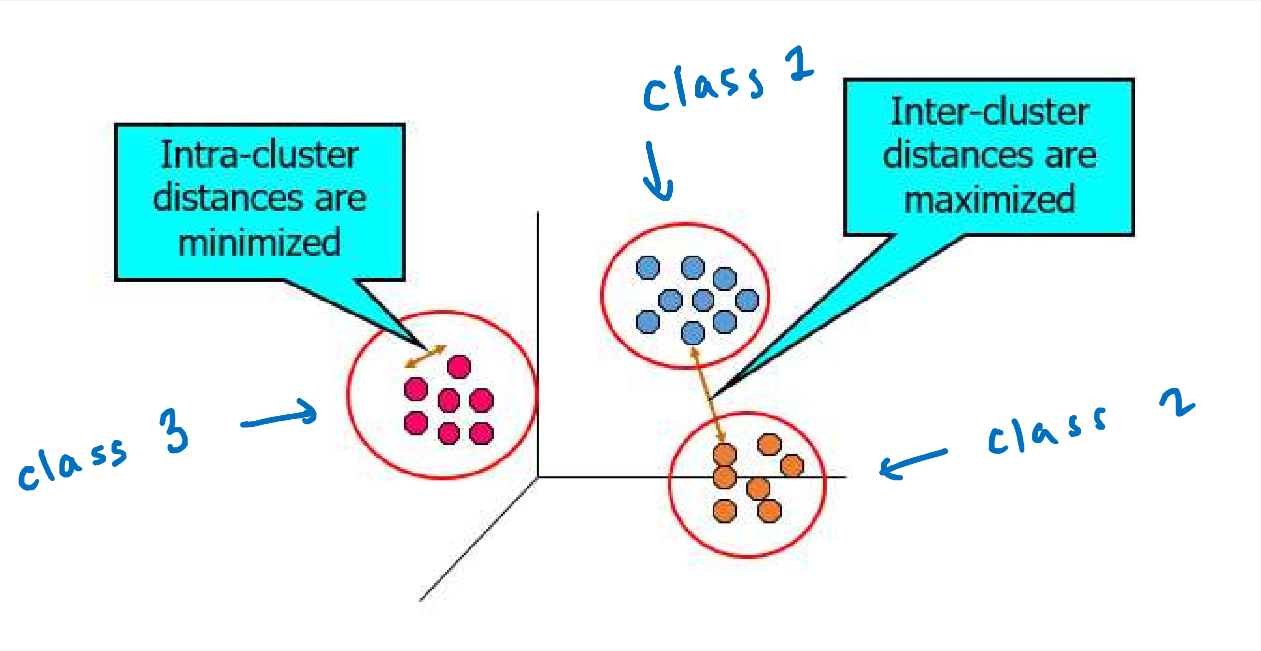
Classification

- □ **Shape**
- □ **Colour**
- □ **Sum of Internal angles**

→ based on similar characteristics

What is Cluster Analysis?

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


class 1
Inter-cluster distances are maximized
Intra-cluster distances are minimized
class 3
class 2

What is Cluster Analysis?

Clustering: the act of grouping "similar" object into sets.

A clustering problem can be viewed as unsupervised classification. → No labelling

Clustering is appropriate when there is no a priori knowledge about the data.

Grouping is based on the distance (proximity).

You don't know who or what belongs to which group. Not even the number of groups.

What is Cluster Analysis?

Clustering: the act of grouping “similar” object into sets.

A clustering problem can be viewed as unsupervised classification.

Clustering is appropriate when there is no a priori knowledge about the data.

Grouping is based on the distance (proximity).

You don’t know who or what belongs to which group. Not even the number of groups.

Example:

A bank wants to give credit card offers to its customers. Currently, they look at the details of each customer and based on this information, decide which offer should be given to which customer.

Now, the bank can potentially have millions of customers. Does it make sense to look at the details of each customer separately and then make a decision? Certainly not! It is a manual process and will take a huge amount of time.

So what can the bank do? One option is to segment its customers into different groups. For instance, the bank can group the customers based on their income

Application of Clustering:

- Marketing: Help marketers discover distinct groups in their customer bases, and then use this knowledge to develop targeted marketing programs.

- Land use: Identification of areas of similar land use in an earth observation database.

- Insurance: Identifying groups of motor insurance policy holders with a high average claim cost.

- City-planning: Identifying groups of houses according to their house type, value, and geographical location

- The field of psychiatry: The characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropriate form of therapy.

- Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults.

- Climate - Understanding the Earth's climate requires finding patterns in the atmosphere and ocean. To that end, cluster analysis has been applied to find patterns in the atmospheric pressure of Polar Regions and areas of the ocean that have a significant impact on land climate.

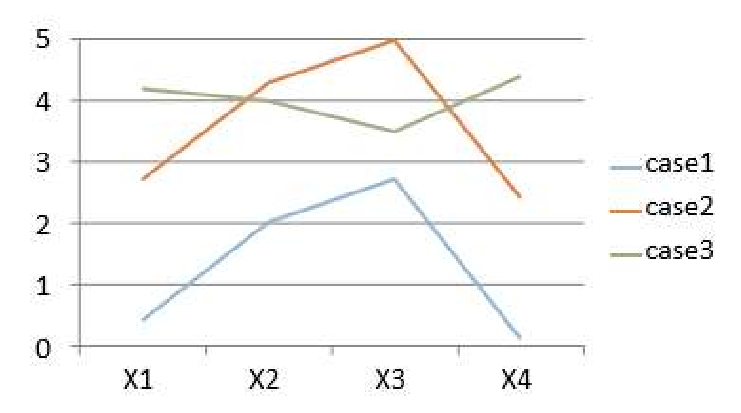
- Biology - used to find groups of genes that have similar functions.

Measuring similarity

- Similarity
  - The degree of correspondence among objects across all of the characteristics.
    - Correlational measures
    - Distance measures

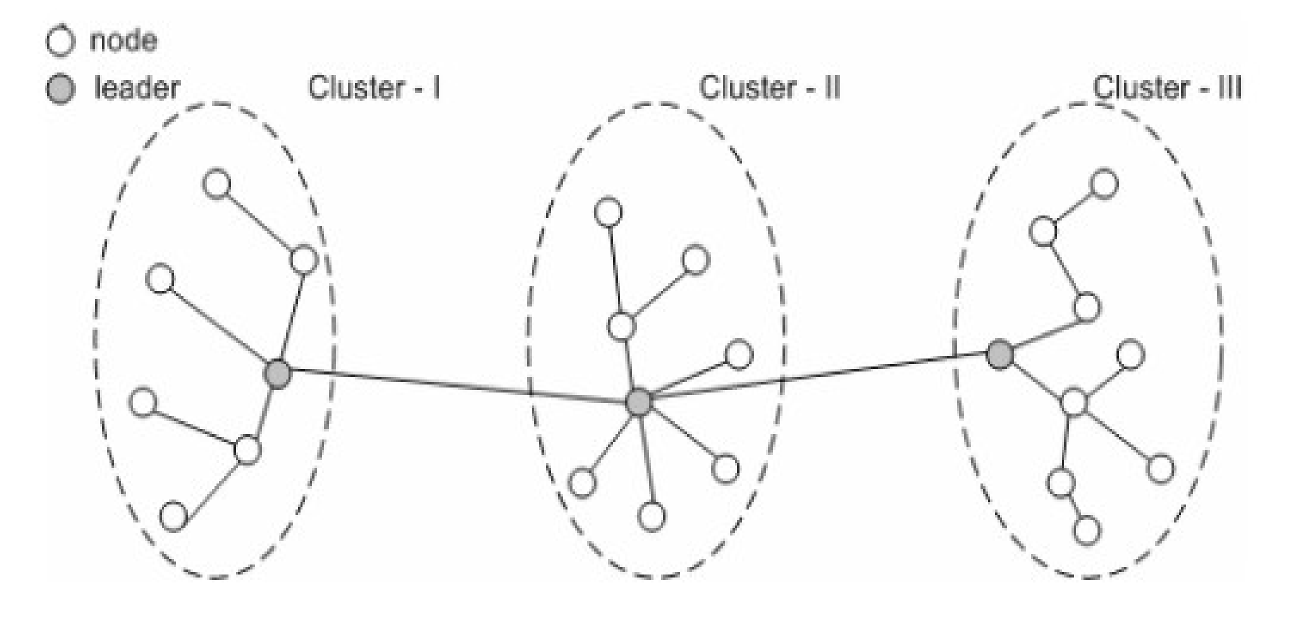
- Correlation measure
  - Grouping cases base on respondent pattern

- Distance measure
  - Grouping cases base on distance



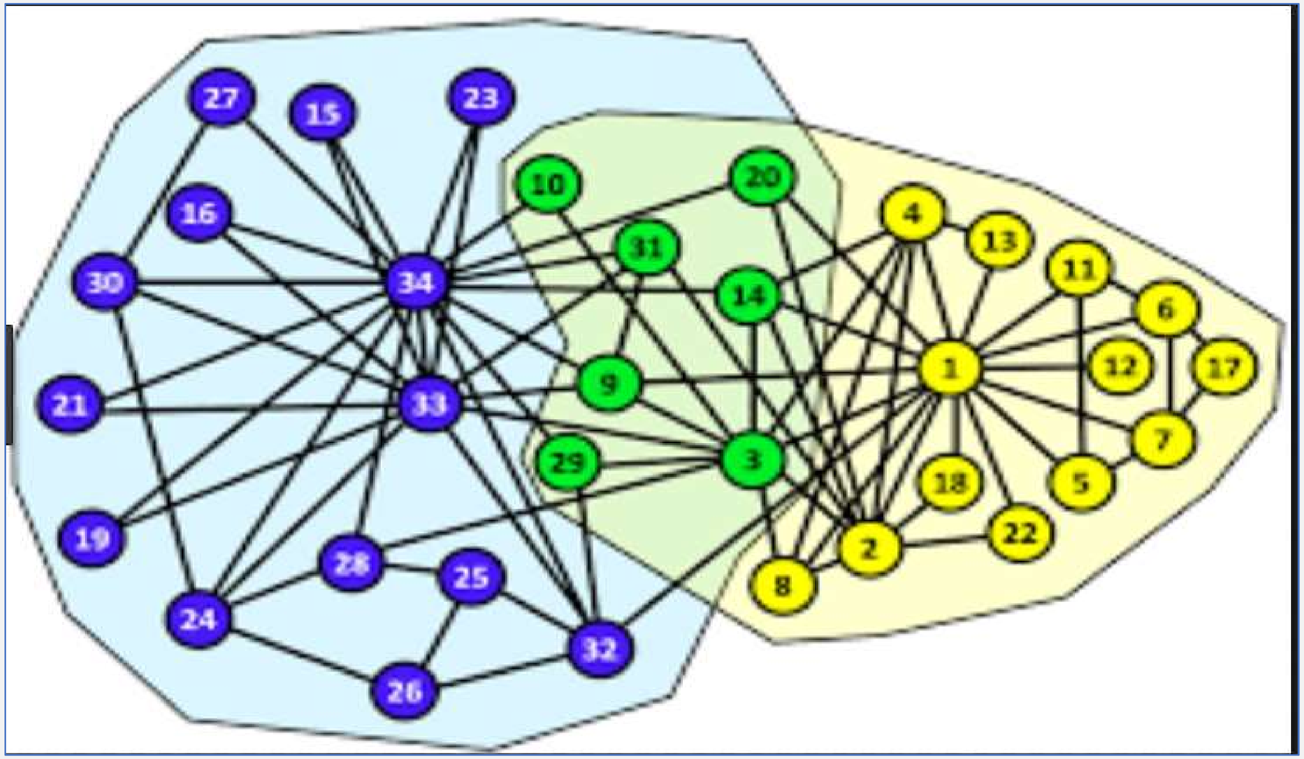
Non-overlapping Clusters

- **Cluster in which each observation belongs to only one cluster. Non-overlapping clusters are more frequently used clustering techniques in practice.**



Overlapping Clusters

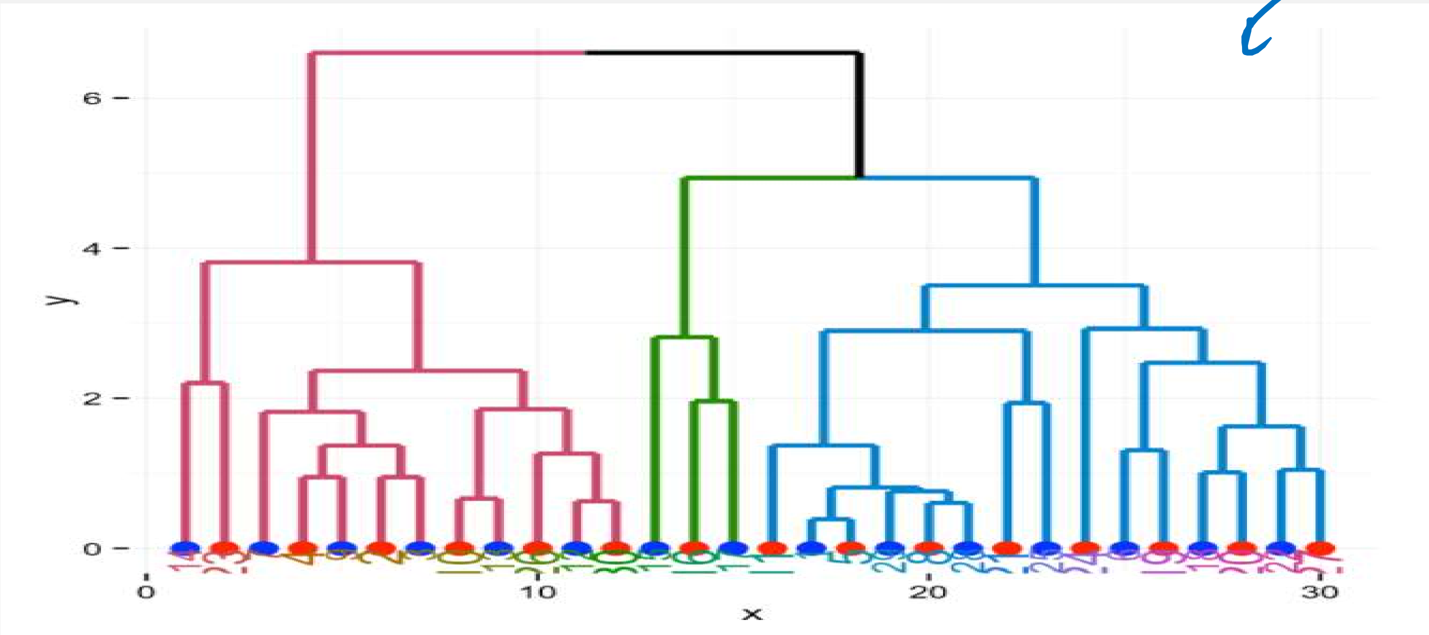
- An observation may belong to more than one cluster



Hierarchical clustering

- Hierarchical clustering creates subsets of data similar to a tree-like structure in which the root node corresponds to the complete set of data. Branches are created from the root node to split the data into heterogeneous subsets (clusters).

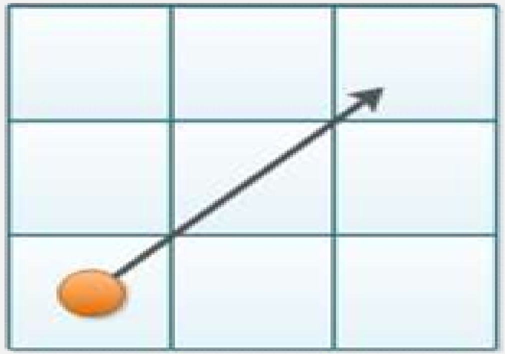
Dendogram



Distance Measures

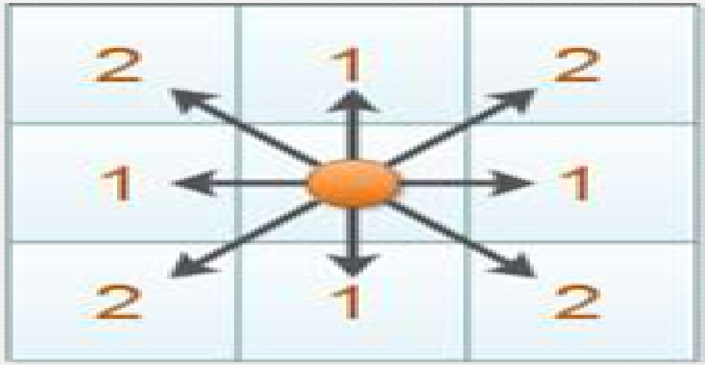
- Euclidean distance
- City-block (Manhattan) distance
- Chebyshev's distance

Euclidean Distance



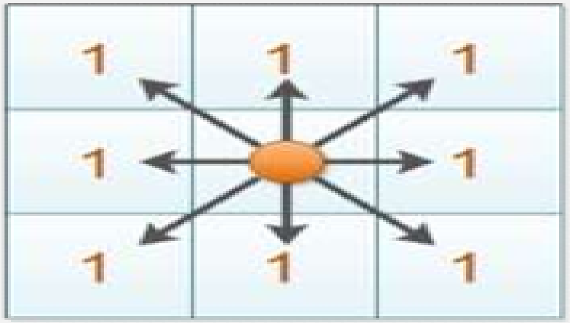
$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Manhattan Distance



$$|x_1 - x_2| + |y_1 - y_2|$$

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

Euclidean Distance

- Euclidean is one of the frequently used distance measures when the data are either in interval or ratio scale.

- The Euclidian distance between two n-dimensional observations X1 (x11, x12, …, x1n) and X2 (x21, x22, …, x2n) is given by

$$\boxed{D(X_1, X_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \cdots + (x_{1n} - x_{2n})^2}}$$

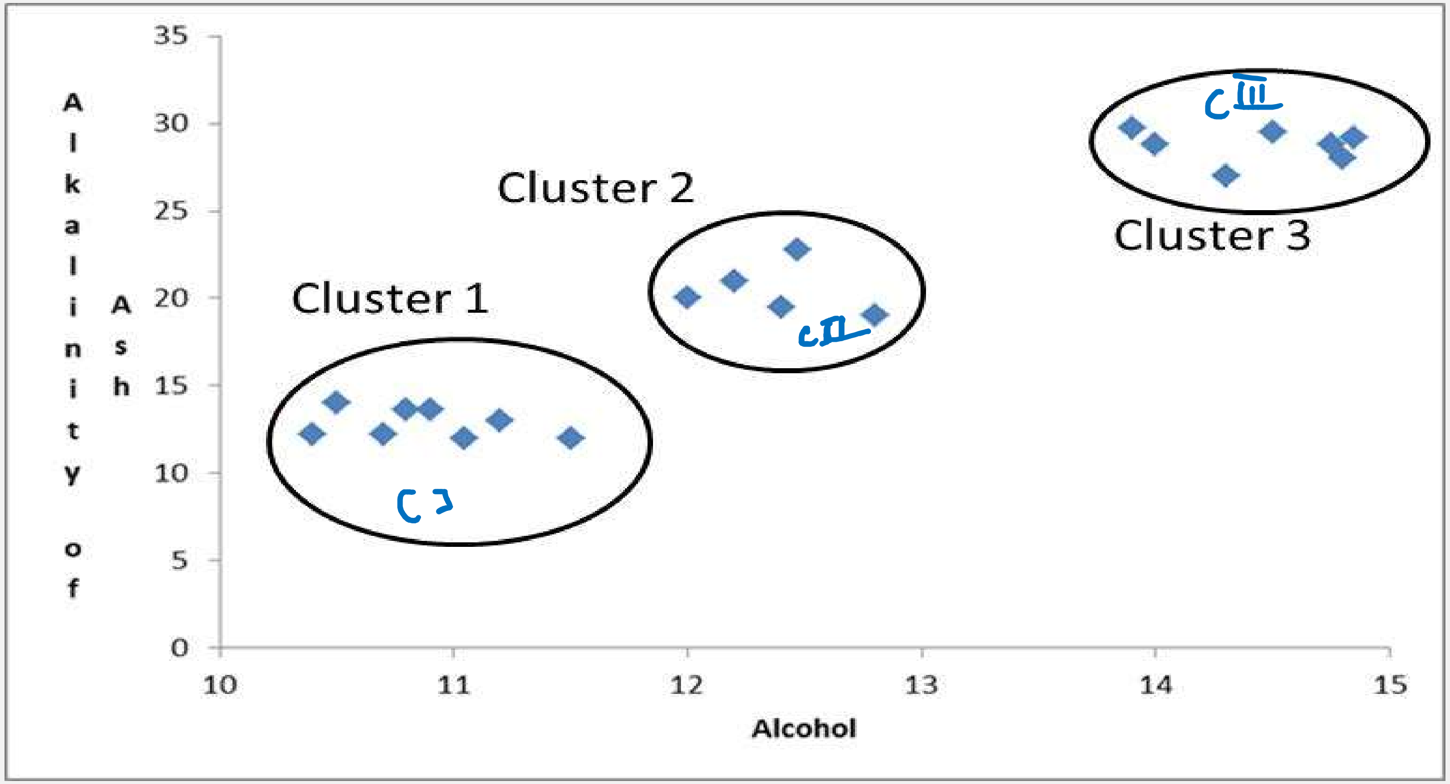
Example

- **The below table has information about 20 wines sold in the market along with their alcohol and alkalinity of ash content**

Wine	Alcohol	Alkalinity of Ash	Wine	Alcohol	Alkalinity of Ash
1	14.8	28	11	10.7	12.2
2	11.05	12	12	14.3	27
3	12.2	21	13	12.4	19.5
4	12	20	14	14.85	29.2
5	14.5	29.5	15	10.9	13.6
6	11.2	13	16	13.9	29.7
7	11.5	12	17	10.4	12.2
8	12.8	19	18	10.8	13.6
9	14.75	28.8	19	14	28.8
10	10.5	14	20	12.47	22.8

Example

- **Clusters of wine based on alcohol and ash content.**



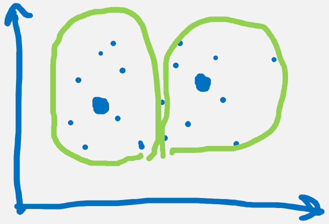
Types of Clustering

$$\underbrace{\text{KNN- Algorithm}}_{\text{unsupervised ML Algo}}$$

- A clustering is a set of clusters
- Important distinction between hierarchical and partitional sets of clusters
    - Partitional Clustering[k means clustering] ✓✓
        - A division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
    - Hierarchical clustering ✓✓
        - A set of nested clusters organized as a hierarchical tree

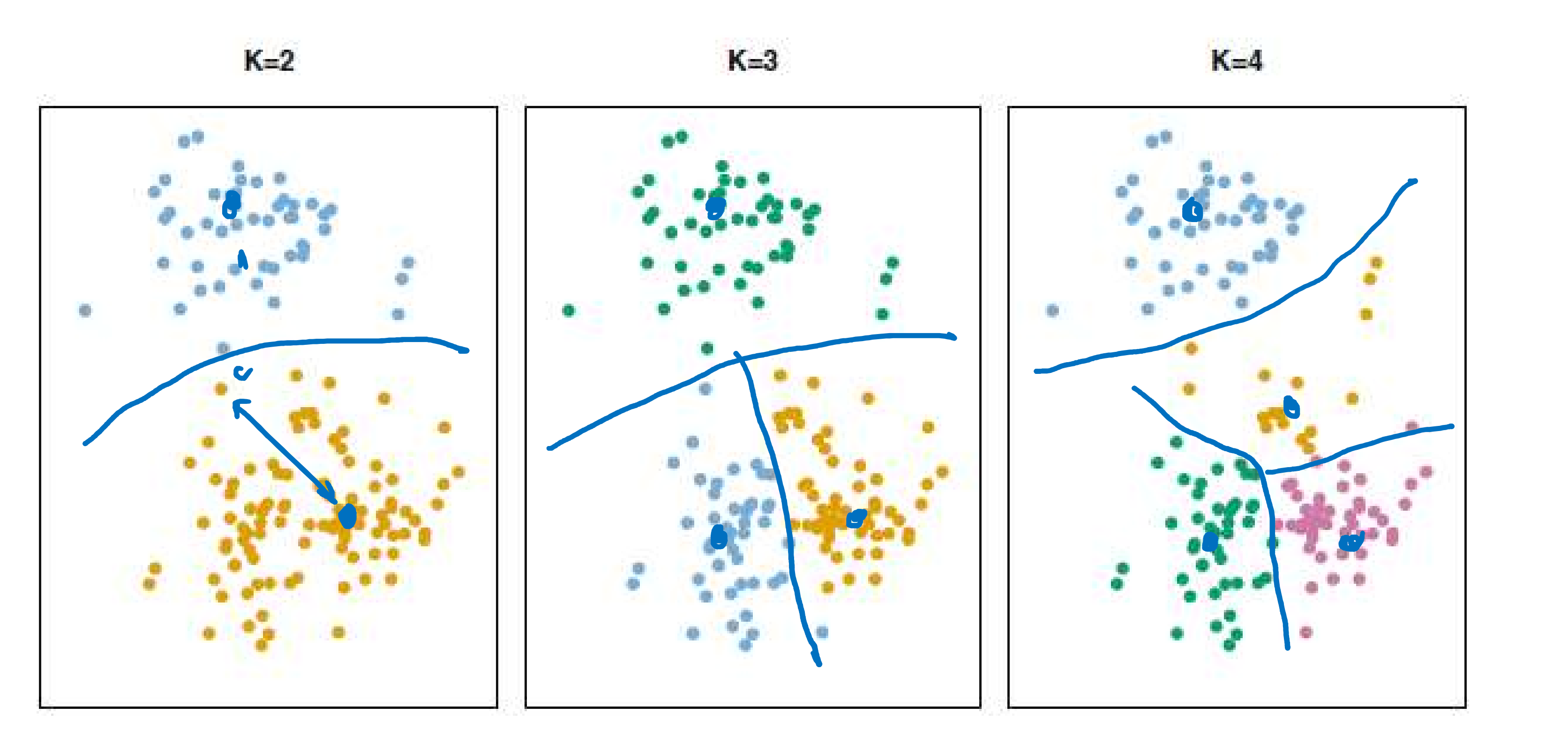
K-Means Clustering

- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- Data set into K distinct, non-overlapping clusters.
- To perform k-means Clustering, we must first specify the desired number of clusters K; then the K-means algorithm will assign each observation to exactly one of the K Clusters.
- You'll define a target number k, which refers to the number of centroids you need in the dataset. A centroid is the imaginary or real location representing the center of the cluster.
- Every data point is allocated to each of the clusters through reducing the in-cluster sum of squares.
- In other words, the K-means algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible.



K-Means Clustering

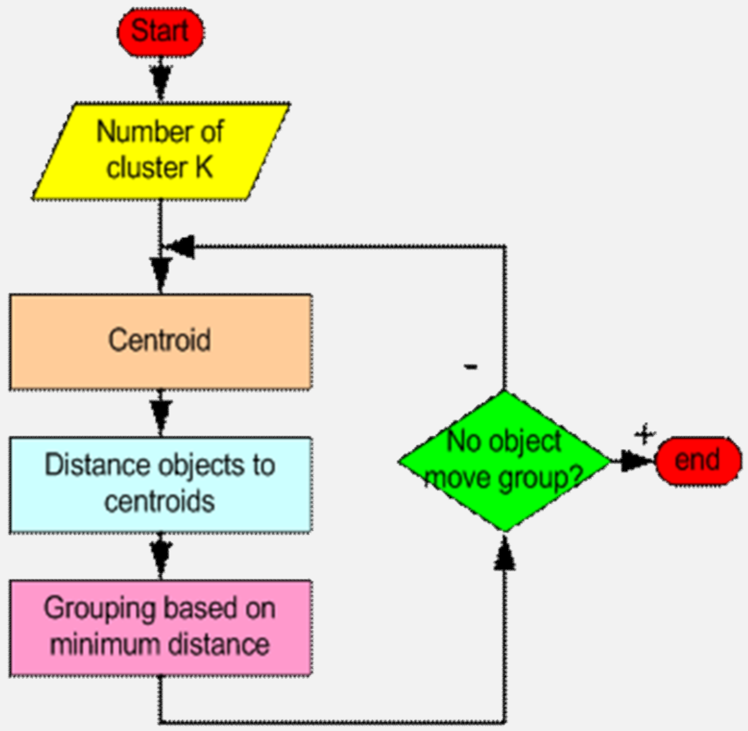
Figure shows the results obtained from performing *K*-means clustering on a simulated example consisting of 150 observations in two dimensions, using three different values of *K*.



Note that there is no ordering of the clusters, so the cluster coloring is arbitrary.

K-Means Clustering

How the K-Mean Clustering algorithm works?


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object move group?
-
+
end

The following data set consisting of the scores of two variables on each of seven individuals:

	Subject	A	B	
✓	1	1	1	K_1
	2	1.5	2	K_1
	3	3	4	K_1
✓	4	5	7	K_2
	5	3.5	5	K_2
	6	4.5	5	K_2
	7	3.5	4.5	K_2

Apply KNN-Algo using $K=2$.

Sol^n:- pick 2 rows randomly for k_1 & k_2 respectively

$k_1 = (1, 1)$ $k_2 = (5, 7)$

Euclidean Distance Formula = $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

step 1:- for row 2, $(1.5, 2)$

$K_1 \rightarrow \sqrt{(1.5-1)^2 + (2-1)^2} = 1.11803$

$K_2 \rightarrow \sqrt{(1.5-5)^2 + (2-7)^2} = 6.1032$

$K_1 < K_2$

$K_1 = \{1, 2\}$, $K_2 = \{4\}$

$$K_1 \Rightarrow \left\{ \frac{1.5+1}{2} , \frac{2+1}{2} \right\} = (1.25 , 1.5)$$

step 2 :- for row 3, (3 , 4)

$$K_1 = \sqrt{(3-1.25)^2+(4-1.5)^2} = 3.05$$

$$K_2 = \sqrt{(3-5)^2 +(4-7)^2} = 3.6055$$

$$K_1 < K_2$$

$\longrightarrow$ $K_1 = \{1, 2, 3\}$, $K_2 \{4\}$

$$k_1 = \left(\frac{3+1.25}{2} , \frac{4+1.5}{2} \right) = (2.125 , 2.75)$$

step 3 :- for row 5 , (3.5 , 5)

$$k_1 = \sqrt{(3.5-2.125)^2 + (5-2.75)^2} = 2.368$$

$$k_2 = \sqrt{(3.5-5)^2 + (5-7)^2} = 2.5$$

$k_1 = \{1, 2, 3\}$ $k_2 = \{4, 5\}$

$$K_2 = \left(\frac{3.5+5}{2} , \frac{5+7}{2} \right) = (4.25 , 6)$$

step 4 :- for row 6 , (4.5 , 5)

$$K_1 = \sqrt{(4.5 - 2.125)^2 + (5 - 2.75)^2} = 2.49$$

$$K_2 = \sqrt{(4.5 - 4.25)^2 + (5 - 6)^2} = 1.03$$

$$K_2 < K_1$$

$K_1 = \{1, 2, 3\}$ $K_2 = \{4, 5, 6\}$

$$K_2 = \left(4.375 , 5.5 \right)$$

step 5:- for no 7 $\Rightarrow (3.5, 4.5)$

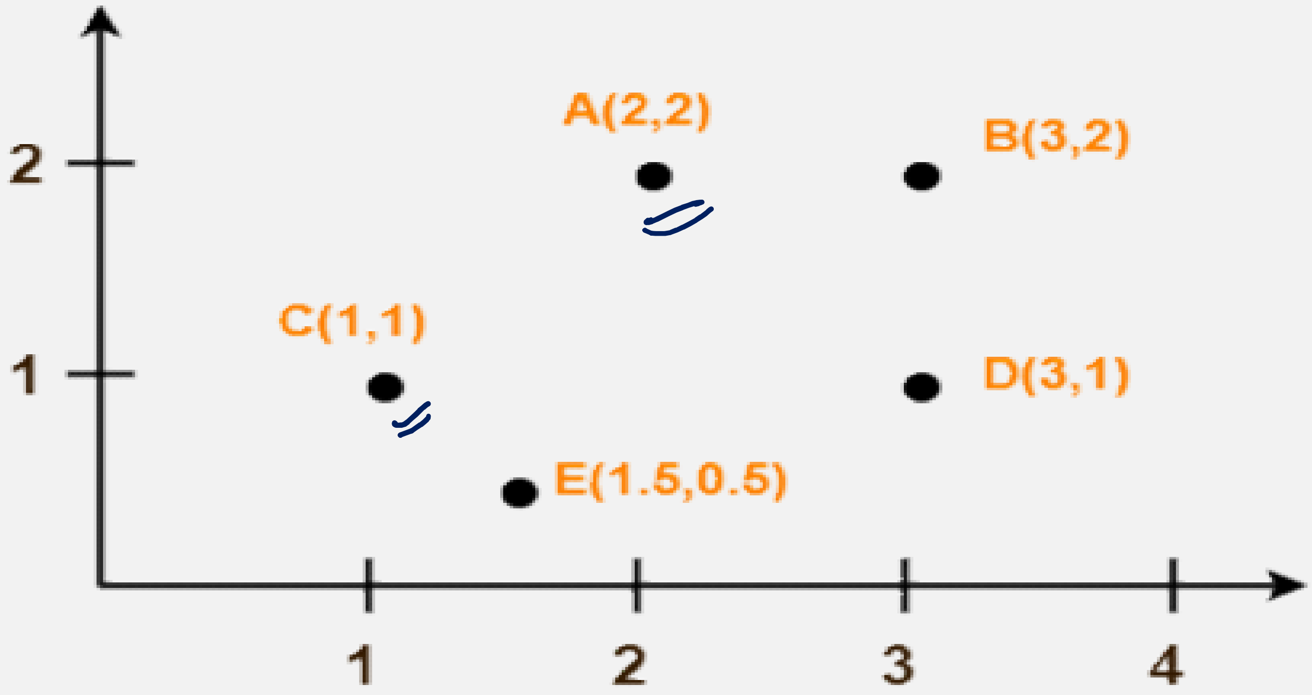
$$K_1 = \sqrt{(3.5 - 2.125)^2 + (4.5 - 2.75)^2} = 2.225$$

$$K_2 = \sqrt{(3.5 - 4.375)^2 + (4.5 - 5.5)^2} = 1.32$$

$$K_2 < K_1$$

$K_1 = \{1, 2, 3\}$ $\quad$ $K_2 = \{4, 5, 6, 7\}$

Cluster the following points (with (x, y) representing locations) into two clusters:



Given	D(2, 2)	D(1, 1)	point belong to cluster
A(2, 2)	0	1.41	C1
B(3, 2)	1	2.24	C1
C(1, 1)	1.41	0	C2
D(3, 1)	1.41	2	C1
E[1.5, 0.5]	1.58	0.71	C2

C1 = {A, B, D}   C2 = {C, E}

Hierarchical clustering

Hierarchical clustering will help to determine the optimal number of clusters.

1. It starts by putting every point in its own cluster, so each cluster is a singleton

2. It then merges the 2 points that are closest to each other based on the distances from the distance matrix. The consequence is that there is one less cluster

3. It then recalculates the distances between the new and old clusters and save them in a new distance matrix which will be used in the next step

4. Finally, steps 1 and 2 are repeated until all clusters are merged into one single cluster including all points.

Hierarchical clustering

There exists 5 main methods to measure the distance between clusters, referred as linkage methods:

1. Simple linkage: computes the minimum distance between clusters before merging them.

2. Complete linkage: computes the maximum distance between clusters before merging them.

3. Average linkage: computes the average distance between clusters before merging them.

4. Centroid linkage: calculates centroids for both clusters, then computes the distance between the two before merging them.

5. Ward's (minimum variance) criterion: minimizes the total within-cluster variance and find the pair of clusters that leads to minimum increase in total within-cluster variance after merging.

Hierarchical clustering

Dendogram: -

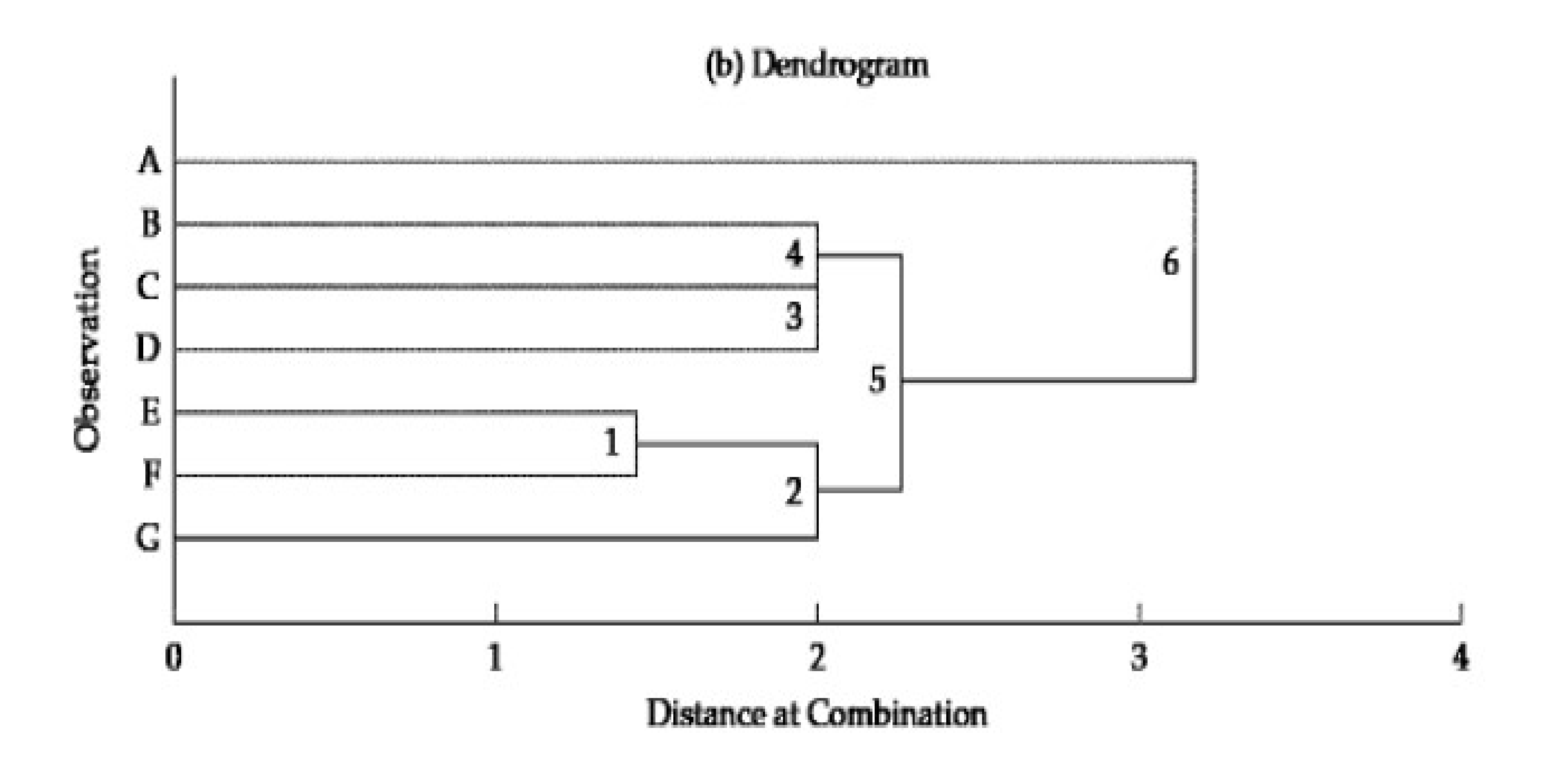
- A Dendogram, or tree graph, is a graphical device for displaying clustering results. Vertical lines represent clusters that are joined together.

- The position of the line on the scale indicates the distances at which clusters were joined.

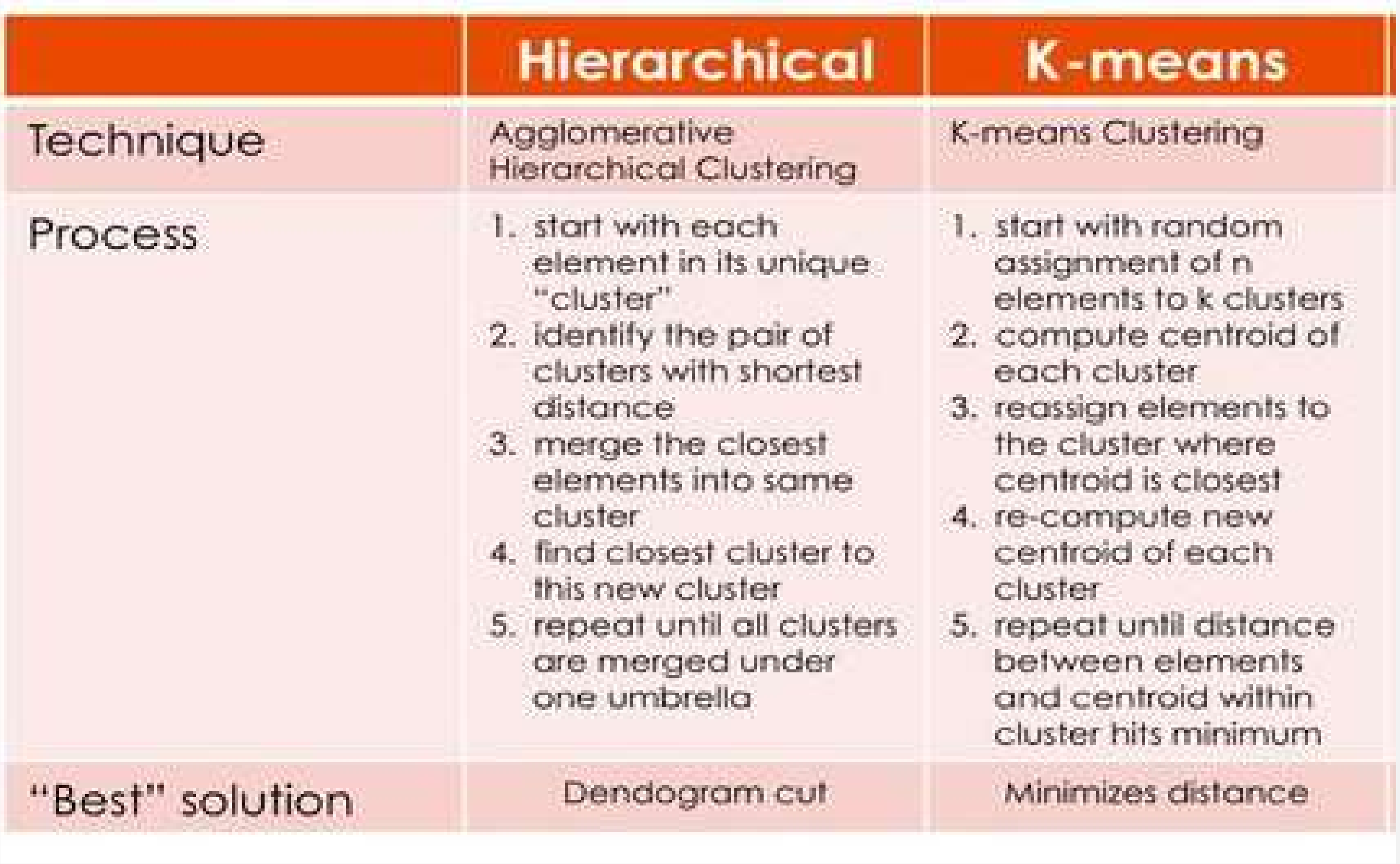
- The Dendogram is read from left to right.

- Graphical representation (tree graph) of the results of a hierarchical procedure. Starting with each object as a separate cluster, the Dendogram shows graphically how the clusters are combined at each step of the procedure until all are contained in a single cluster

Hierarchical clustering



Hierarchical clustering



	Hierarchical	K-means
Technique	Agglomerative Hierarchical Clustering	K-means Clustering
Process	1. start with each element in its unique "cluster" 2. identify the pair of clusters with shortest distance 3. merge the closest elements into same cluster 4. find closest cluster to this new cluster 5. repeat until all clusters are merged under one umbrella	1. start with random assignment of n elements to k clusters 2. compute centroid of each cluster 3. reassign elements to the cluster where centroid is closest 4. re-compute new centroid of each cluster 5. repeat until distance between elements and centroid within cluster hits minimum
"Best" solution	Dendogram cut	Minimizes distance

Simple Linkage Method

The daily expenditures on food (X1) and clothing (X2) of five persons are shown in Table

Person	X1	X2
A	2	4
B	8	2
C	9	3
D	1	5
E	8.5	1