PRESIDENCY UNIVERSITY

BENGALURU

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

MODULE 4

CLUSTERING AND FORECASTING WITH TIME-SERIES DATA.

CONTENT

- Partitioned Clustering
    - K-means Clustering
- Hierarchical Clustering techniques
- Cluster validity measures
- Components of Time series Data
- Association Rule Mining
- Collaborative Filtering
    - User based and item based similarity
    - closed and maximal frequent item sets.

CLUSTERING

- **Clustering:** the process of grouping a set of objects into classes of similar objects
    - Documents within a cluster should be similar.
    - Documents from different clusters should be dissimilar.
- The common form of unsupervised learning
        - Unsupervised learning = learning from raw data, as opposed to supervised data where a classification of examples is given
    - A common and important task that finds many applications in IR and other places
- But there may be many cases in which we do not have labeled data and need to find the hidden patterns from the given dataset. So, to solve such types of cases in machine learning, we need unsupervised learning techniques.
- “Unsupervised learning is a type of machine learning in which models are trained using unlabeled dataset and are allowed to act on that data without any supervision”.

Partitioned Clustering

- It is a type of clustering that divides the data into non-hierarchical groups. It is also known as the centroid-based method. The most common example of partitioning clustering is the K-Means Clustering algorithm.
- In this type, the dataset is divided into a set of k groups, where K is used to define the number of pre-defined groups.
- In the partitioning method when database(D) that contains multiple(N) objects then the partitioning method constructs user-specified(K) partitions of the data in which each partition represents a cluster and a particular region.
- There are many algorithms that come under partitioning method some of the popular ones are K-Mean, PAM(K-Medoids), CLARA algorithm (Clustering Large Applications) etc

K-Means Clustering

1. K-means is an unsupervised machine learning algorithm.

2. Group the dataset into different clusters

3. Here K defines the number of pre-defined clusters that need to be created in the process.

*Suppose K=2, then 2 clusters will be created

*K=3, then 3 clusters will be created.

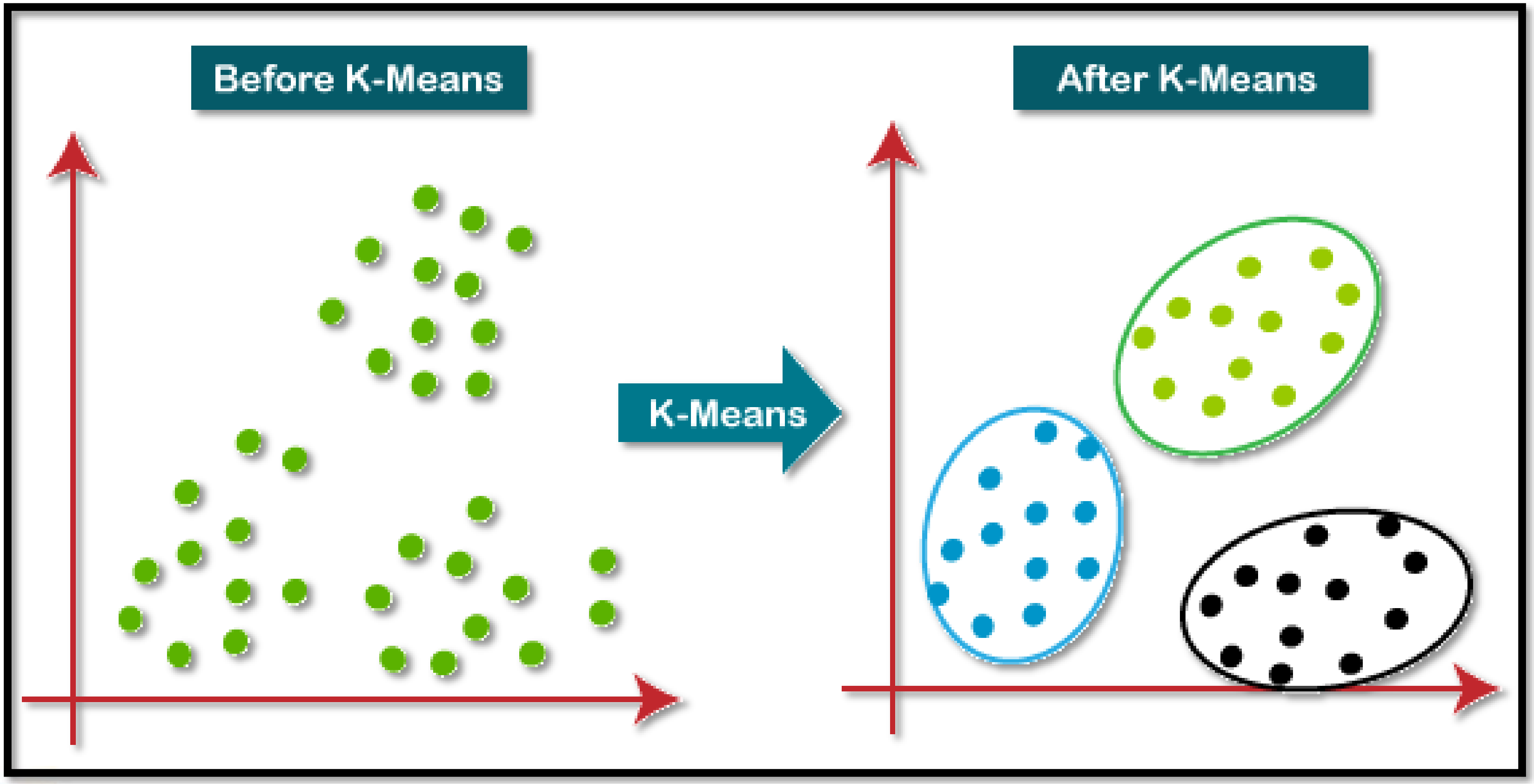
"It is an iterative algorithm that divides the unlabelled dataset into k different clusters in such a way that each dataset belongs to only one group that has similar properties"

4. It is a centroid-based algorithm.

5. Each cluster will be associated with a centroid.

6.The aim of this algorithm is to minimize the sum of distance between the data points and the corresponding cluster.

K-MEANS CLUSTERING



PRESIDENCY UNIVERSITY

Private University Estd. in Karnataka State by Act No. 41 of 2013

K-MEANS CLUSTERING ALGORITHM

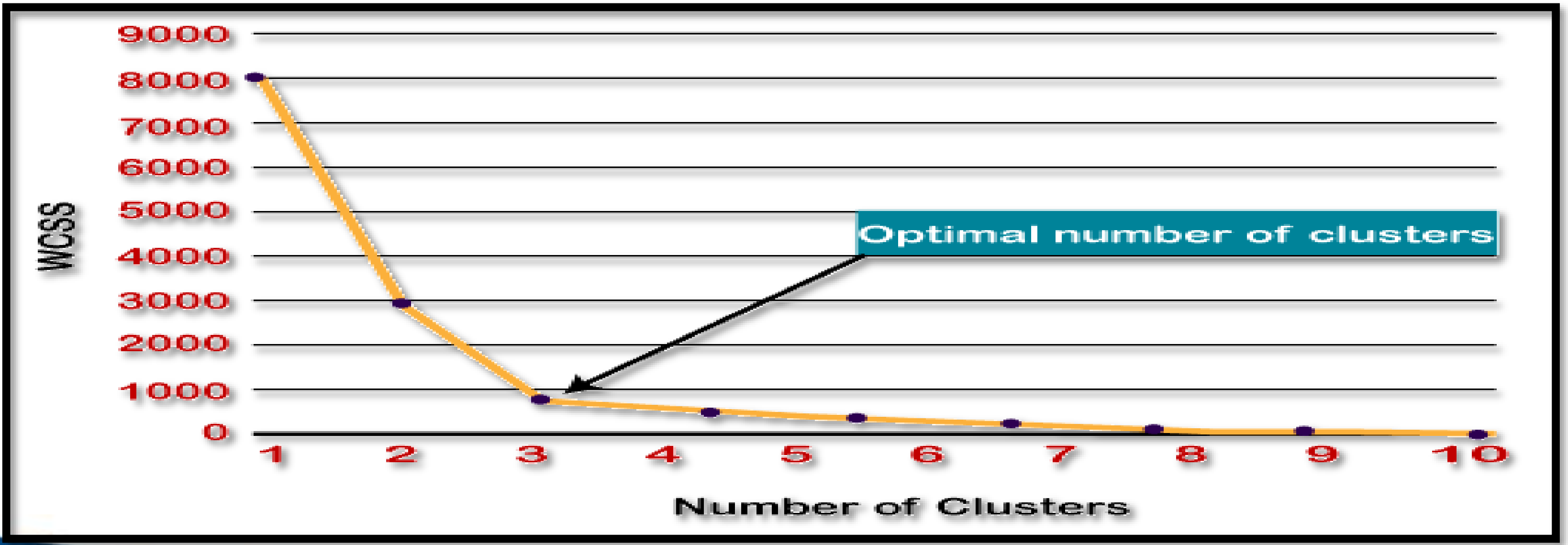
Step 1: Select the number K to decide the number of clusters.
Step 2: Select random K points or centroids. (It can be other than the input dataset).
Step 3: Assign each data point to its closest centroid, which will form the predefined K clusters.
Step 4: Calculate the variance and place a new centroid of each cluster.
Step-5: Repeat the third steps, which means reassigning each datapoint to the new closest centroid of each cluster.
Step-6: If any reassignment occurs, then go to step 4 else go to FINISH.
Step-7: The model is ready.

HOW TO CHOOSE THE VALUE OF "K NUMBER OF CLUSTERS" IN K-MEANS CLUSTERING?

- The performance of the K-means clustering algorithm depends upon the highly efficient clusters that it forms.
- But choosing the optimal number of clusters is a big task.
- There are some different ways to find the optimal number of clusters, but here we are discussing the most appropriate method to find the number of clusters or value of K.
- We use the Elbow method to select the optimal value of K in K-means clustering.

ELBOW METHOD

- The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. WCSS stands for Within Cluster Sum of Squares, which defines the total variations within a cluster.



K-MEANS CLUSTERING EXAMPLE

- **Example 1:** Use K means Clustering algorithm to divide the following data into two clusters.

X1	1	2	2	3	4	5
X2	1	1	3	2	3	5

- **Step 1:** Choosing randomly 2 Cluster centers

  Say V1=(2,1) V2=(2,3)

- **Step 2:** Find the distance between the Cluster Center and each data points & assign minimum as assigned Center

Data point	Distance from V1(2,1)	Distance from V2(2,3)	Assigned center
a1(1,1)	1	2.24	V1
a2(2,1)	0	2	V1
a3(2,3)	2	0	V2
a4(3,2)	1.41	1.41	V1
a5(4,3)	2.83	2	V2
a6(5,5)	5	3.61	V2

K-Means Clustering Example by Euclidean Distance

- a1(1,1) V1(2,1) =√(1-2)2 + (1-1)2 = √(1) = 1
- a2(2,1) V1(2,1) =√(2-2)2 + (1-1)2 = √0 = 0
- a3(2,3) V1(2,1) =√(2-2)2 + (3-1)2 = √0+4 = 2
- a4(3,2) V1(2,1) =√(3-2)2 + (2-1)2 = √2 = 1.414 = 1.41
- a5(4,3) V1(2,1) =√(4-2)2 + (3-1)2 = √8 = 2.828 = 2.83
- a6(5,5) V1(2,1) =√(5-2)2 + (5-1)2 = √25 = 5
- a1(1,1) V2(2,3) =√(1-2)2 + (1-3)2 = √5 = 2.236 = 2.24
- a2(2,1) V2(2,3) =√(2-2)2 + (1-3)2 = √4 = 2
- a3(2,3) V2(2,3) =√(2-2)2 + (3-3)2 = √0 = 0
- a4(3,2) V2(2,3) =√(3-2)2 + (2-3)2 = √2 = 1.414 = 1.41
- a5(4,3) V2(2,3) =√(4-2)2 + (3-3)2 = √4 = 2
- a6(5,5) V2(2,3) =√(5-2)2 + (5-3)2 = √13 = 3.606 = 3.61

K-Means Clustering Example by Euclidean Distance

- **Step 4**: Cluster1 of V1: {a1, a2, a4}

  Cluster2 of V2: {a3, a5, a6}

  V1 = 1/3[a1+a2+a4] = 1/3[(1,1) + (2,1) + (3,2)] = 1/3[(6,4)] = (2,1.33)

  V2 = 1/3[a3+a5+a6] = 1/3[(2,3) + (4,3) + (5,5)] = 1/3[(11,11)] = (3.67,3.67)
- **Step 5:** Repeat from step 2 until we get same cluster center or same cluster elements as in the previous iteration.

Data point	Distance from V1(2,1.33)	Distance from V2(3.67,3.67)	Assigned center
a1(1,1)	1.05	3.78	V1
a2(2,1)	0.33	3.15	V1
a3(2,3)	1.67	1.8	V1
a4(3,2)	1.204	1.8	V1
a5(4,3)	2.605	0.75	V2
a6(5,5)	4.74	1.88	V2

K-MEANS CLUSTERING EXAMPLE BY EUCLIDEAN DISTANCE

Step 6: Cluster1 of V1: {a1, a2, a3, a4}

Cluster2 of V2: {a5, a6}

V1 = 1/4[a1+a2+a3 +a4] = 1/4[(1,1) + (2,1) + (2,3) + (3,2)] = 1/4[(8,7)]
= (2,1.75)

V2 = 1/2[a5+a6] = 1/2[(4,3) + (5,5)] = 1/2[(9,8)] = (4.5,4)

Thus, Cluster elements and centers are not same so repeat the steps.

Step 7: Repeat from step 2 until we get same cluster center or same cluster elements as in the previous iteration.

Data point	Distance from V1(2,1.75)	Distance from V2(4.5,4)	Assigned center
a1(1,1)	1.25	4.61	V1
a2(2,1)	0.75	3.9	V1
a3(2,3)	1.25	2.69	V1
a4(3,2)	1.03	2.5	V1
a5(4,3)	2.36	1.12	V2
a6(5,5)	4.42	1.12	V2

Same cluster elements as in the previous iteration. So, final Cluster will be

Cluster1 : {a1,a2,a3,a4}

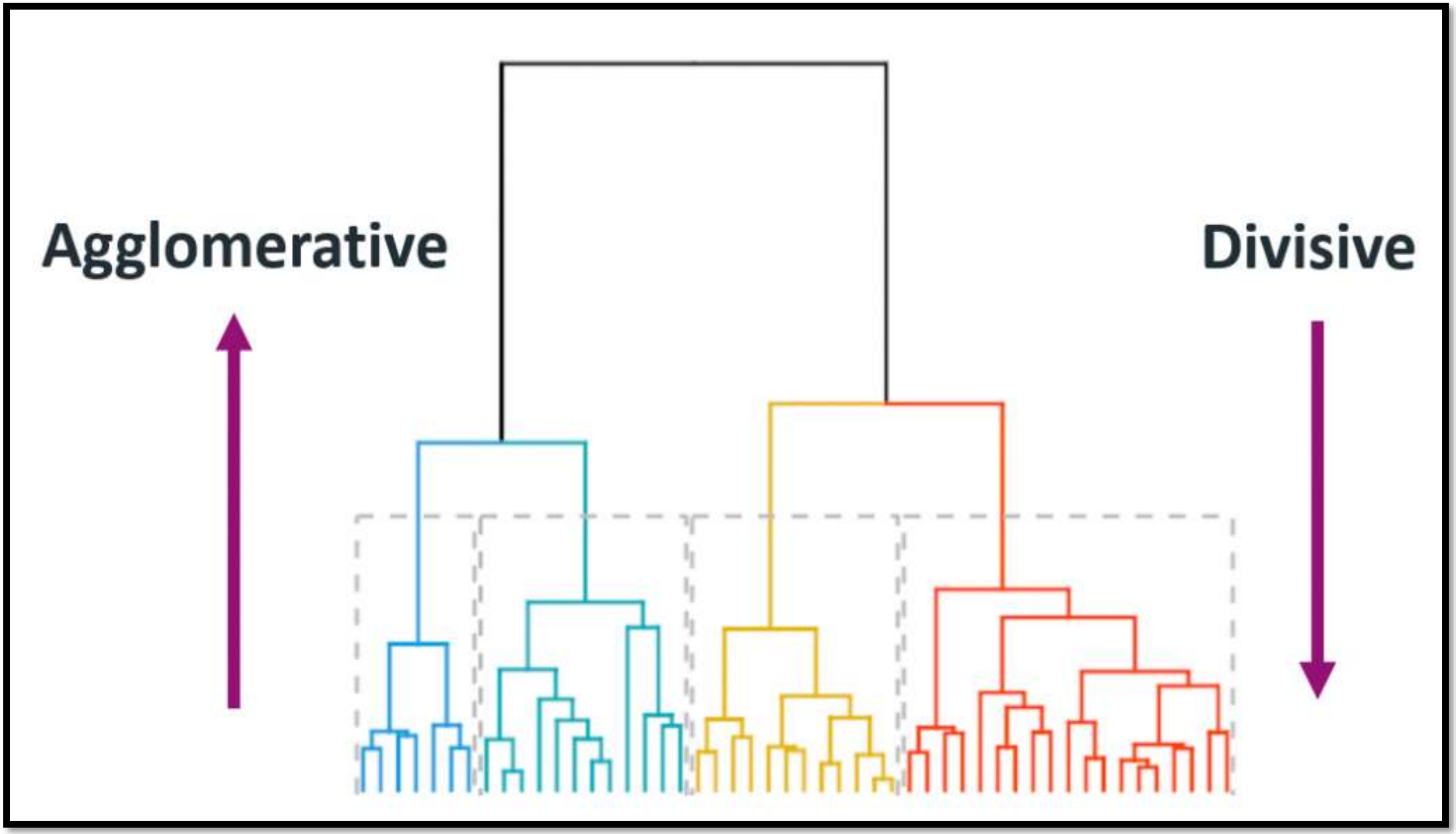
Cluster 2: {a5, a6}

HIERARCHICAL CLUSTERING

- Hierarchical clustering is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster and also known as **hierarchical cluster analysis** or HCA.
- In this algorithm, we develop the hierarchy of clusters in the form of a tree, and this tree-shaped structure is known as the **dendrogram.**

Hierarchical clustering technique has two approaches:

- **Agglomerative:** Agglomerative is a **bottom-up** approach, in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left.
- **Divisive:** Divisive algorithm is the reverse of the agglomerative algorithm as it is a **top-down approach.**



Agglomerative

Divisive

AGGLOMERATIVE HIERARCHICAL CLUSTERING WORK?

Step-1: Create each data point as a single cluster. Let's say there are N data points, so the number of clusters will also be N.

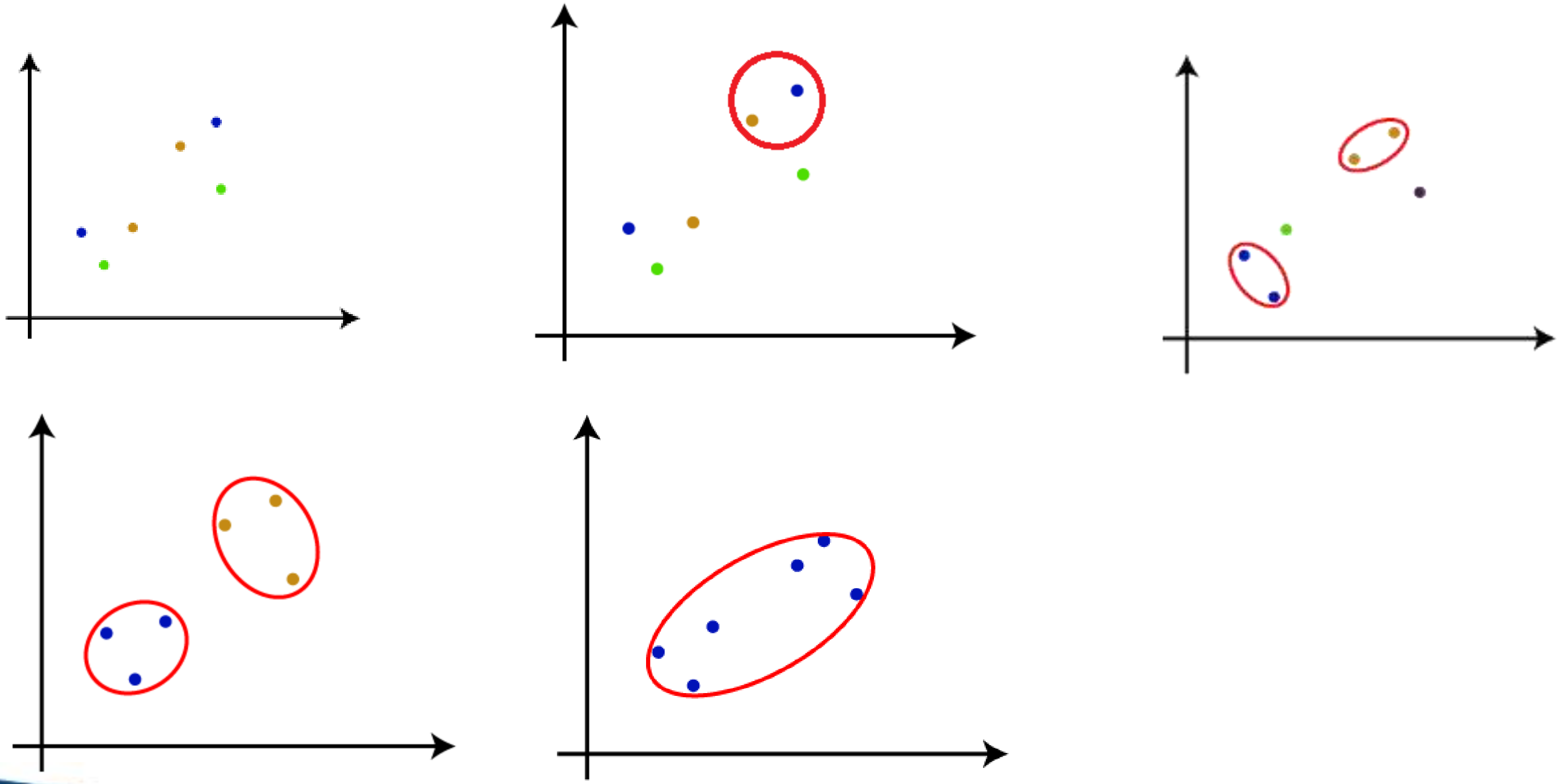
Step-2: Take two closest data points or clusters and merge them to form one cluster. So, there will now be N-1 clusters.

Step-3: Again, take the two closest clusters and merge them together to form one cluster. There will be N-2 clusters.

Step-4: Repeat Step 3 until only one cluster left. So, we will get the following clusters.

Step-5: Once all the clusters are combined into one big cluster, develop the dendrogram to divide the clusters as per the problem

AGGLOMERATIVE HIERARCHICAL CLUSTERING WORK?



CLUSTER VALIDITY MEASURES

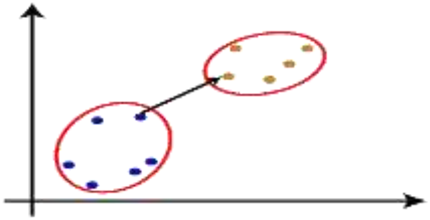
- For Cluster analysis, the analogous question is how to evaluate the “goodness” of the resulting clusters? Why do we need cluster validity measures?
  - To compare clustering algorithms.
  - To compare two sets of clusters.
  - To compare two clusters i.e which one is better in terms of compactness and connectedness.
  - To determine whether random structure exists in the data due to noise.

Generally, cluster validity measures are categorized into 3 classes, they are

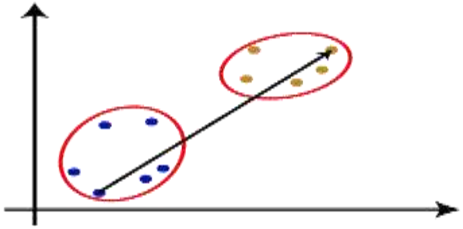
- **Internal cluster validation** : The clustering result is evaluated based on the data clustered itself (internal information) without reference to external information.
- **External cluster validation** : Clustering results are evaluated based on some externally known result, such as externally provided class labels.
- **Relative cluster validation** : The clustering results are evaluated by varying different parameters for the same algorithm (e.g. changing the number of clusters).

CLUSTER VALIDITY MEASURES

- There are various ways to calculate the distance between two clusters, and these ways decide the rule for clustering. These measures are called **Linkage methods**.
- **Single Linkage:** It is the Shortest Distance between the closest points of the clusters. Consider the below image:



- **Complete Linkage:** It is the farthest distance between the two points of two different clusters. It is one of the popular linkage methods as it forms tighter clusters than single-linkage.



- **Average Linkage:** It is the linkage method in which the distance between each pair of datasets is added up and then divided by the total number of datasets to calculate the average distance between two clusters. It is also one of the most popular linkage methods.

COMPONENTS OF TIME SERIES DATA

- **Time series data** is a sequence of data points recorded or collected at regular time intervals. It is a type of data that tracks the evolution of a variable over time, such as sales, stock prices, temperature, etc
- Time series data is commonly used in fields such as economics, finance, weather forecasting, and operations management, among others, to analyze trends, and patterns, and to make predictions or forecasts.

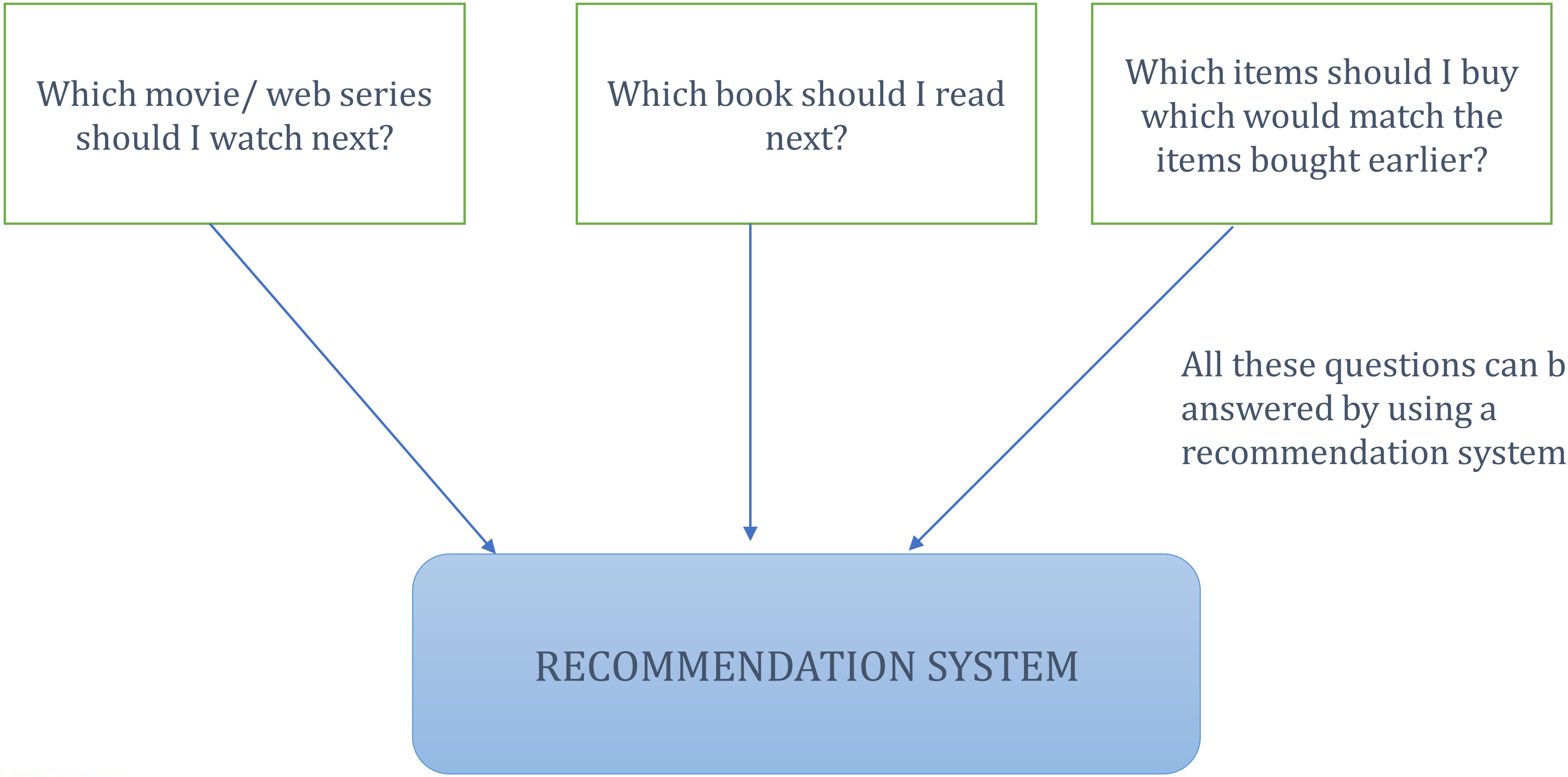
In time series data, there are several types of patterns that can occur:

- **Trend**: A long-term upward or downward movement in the data, indicating a general increase or decrease over time.
- **Seasonality**: A repeating pattern in the data that occurs at regular intervals, such as daily, weekly, monthly, or yearly.
- **Cycle**: A pattern in the data that repeats itself after a specific number of observations, which is not necessarily related to seasonality.
- **Irregularity**: Random fluctuations in the data that cannot be easily explained by trend, seasonality, or cycle.
- **Autocorrelation**: The correlation between an observation and a previous observation in the same time series.
- **Outliers**: Extreme observations that are significantly different from the other observations in the data.
- **Noise**: Unpredictable and random variations in the data.

Recommendation Systems

- The recommender system mainly deals with the likes and dislikes of the users.
- Its major objective is to recommend an item to a user which has a high chance of liking or is in need of a particular user based on his previous purchases.
- It is like having a personalized team who can understand our likes and dislikes and help us in making the decisions regarding a particular item without being biased by any means by making use of a large amount of data in the repositories which are generated day by day.
- The aim of recommender systems is to supply simply accessible, high-quality recommendations for the user community.

RECOMMENDATION SYSTEMS



Which movie/ web series should I watch next?

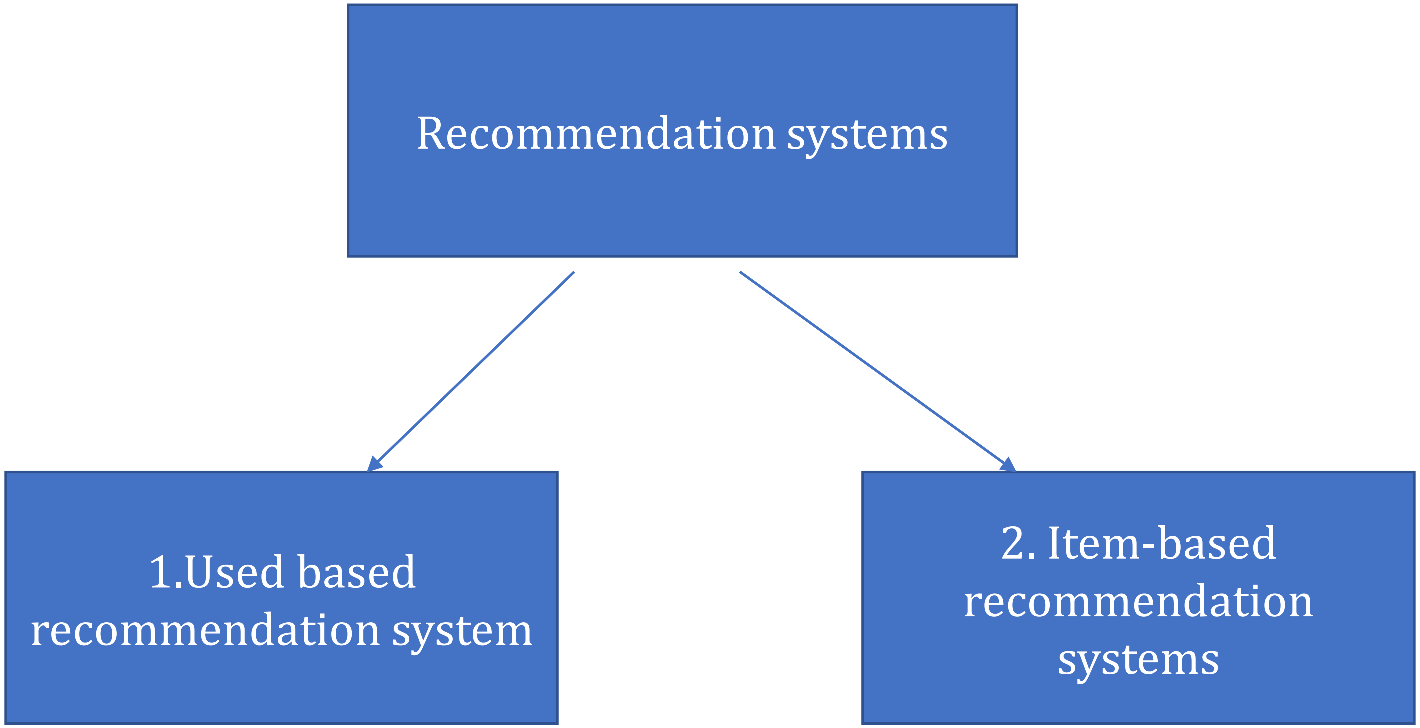
Which book should I read next?

Which items should I buy which would match the items bought earlier?

RECOMMENDATION SYSTEM

All these questions can be answered by using a recommendation system

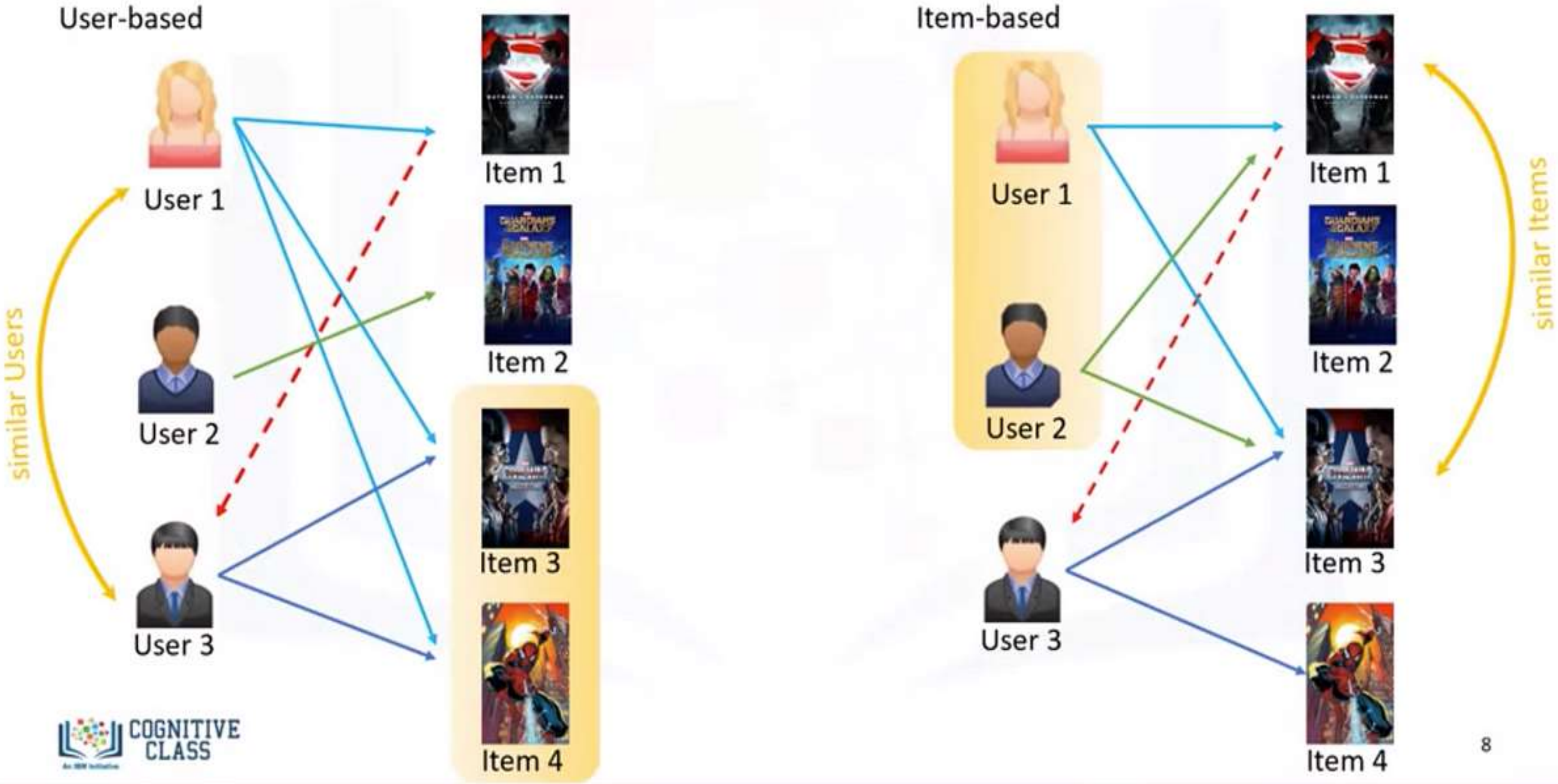
TYPES OF RECOMMENDER SYSTEM


Recommendation systems

1.Used based recommendation system

2. Item-based recommendation systems

Collaborative filtering



COGNITIVE CLASS

8

Self Study Topics:

- Association Rule Mining

Thank You