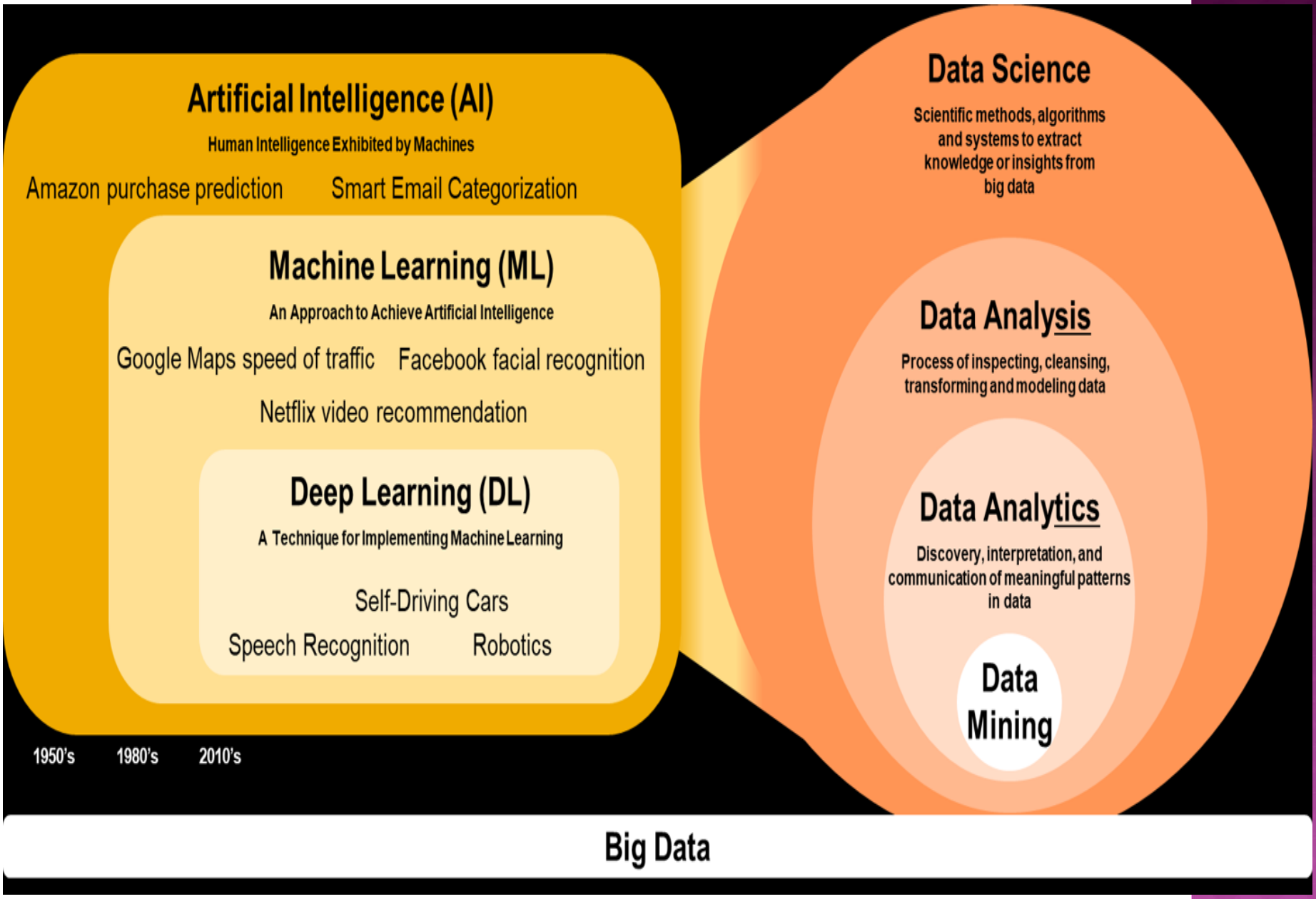
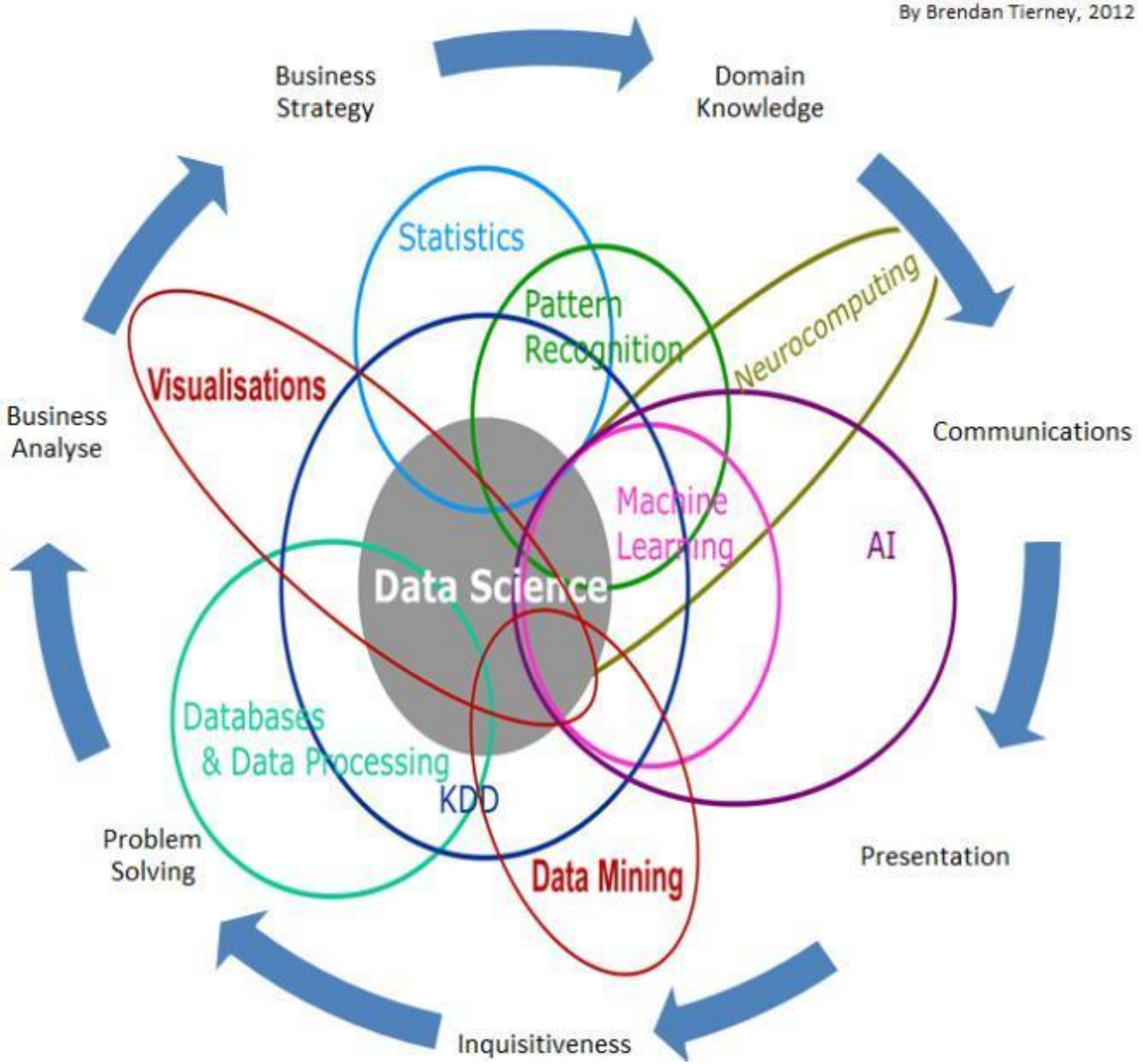
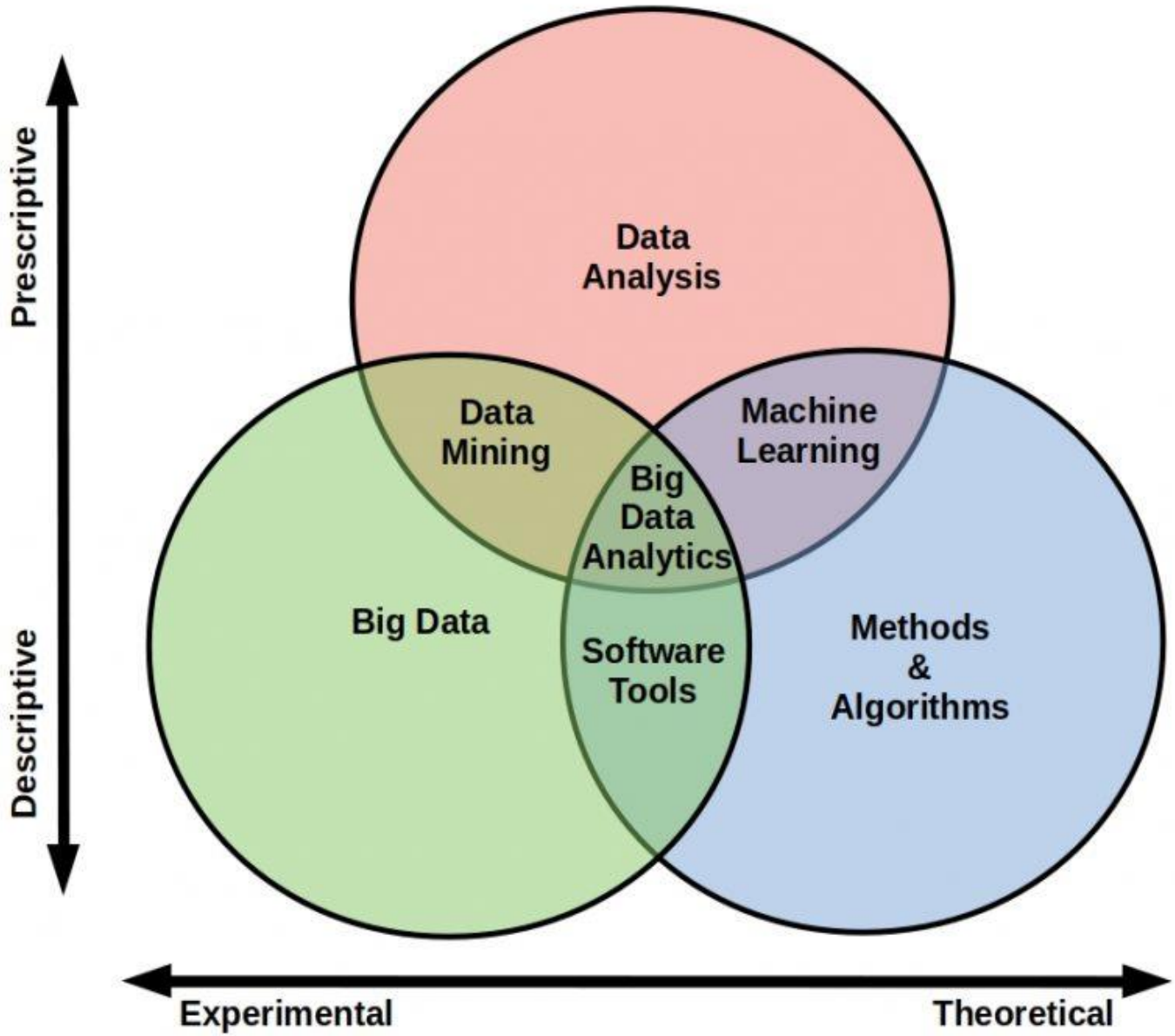
DATA MINING

By Dr. Bidisha Bhabani


Artificial Intelligence (AI)
Human Intelligence Exhibited by Machines
Amazon purchase prediction
Smart Email Categorization
Machine Learning (ML)
An Approach to Achieve Artificial Intelligence
Google Maps speed of traffic
Facebook facial recognition
Netflix video recommendation
Deep Learning (DL)
A Technique for Implementing Machine Learning
Self-Driving Cars
Speech Recognition
Robotics
1950's
1980's
2010's
Data Science
Scientific methods, algorithms and systems to extract knowledge or insights from big data
Data Analysis
Process of inspecting, cleansing, transforming and modeling data
Data Analytics
Discovery, interpretation, and communication of meaningful patterns in data
Data Mining
Big Data

Data Science Is Multidisciplinary




Prescriptive
Descriptive
Data Analysis
Data Mining
Machine Learning
Big Data Analytics
Big Data
Software Tools
Methods & Algorithms
Experimental
Theoretical

- **Artificial Intelligence** (AI) Human Intelligence Exhibited by Machines
  - Intelligence exhibited by machines
  - Broadly defined to include any simulation of human intelligence
  - Expanding and branching areas of research, development, and investment
  - Includes robotics, rule-based reasoning, natural language processing (NLP), knowledge representation techniques (knowledge graphs) …
- **Machine Learning** (ML) An Approach to Achieve Artificial Intelligence
  - Subfield of AI that aims to teach computers the ability to do tasks with data, without explicit programming
  - Uses numerical and statistical approaches, including artificial neural networks to encode learning in models
  - Models built using "training" computation runs or through usage
- **Deep Learning** (DL) A Technique for Implementing Machine Learning
  - Subfield of ML that uses specialized techniques involving multi-layer (2+) artificial neural networks
  - Layering allows cascaded learning and abstraction levels (e.g. line -> shape -> object -> scene)
  - Computationally intensive enabled by clouds, GPUs, and specialized HW such as FPGAs, TPUs, etc.

- **Data Science** - Scientific methods, algorithms and systems to extract knowledge or insights from big data
  - Also known as Predictive or Advanced Analytics
  - Algorithmic and computational techniques and tools for handing large data sets
  - Increasingly focused on preparing and modeling data for ML & DL tasks
  - Encompasses statistical methods, data manipulation and streaming technologies (e.g. Spark, Hadoop)
  - Key skill and tools behind building modern AI technologies
- **Data Analysis** - Process of inspecting, cleansing, transforming and modeling data
- **Data Analytics** - Discovery, interpretation, and communication of meaningful patterns in data
- **Data Mining** - Process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems

INTRODUCTION

- Data mining, also known as knowledge discovery in data (KDD), is the process of uncovering patterns and other valuable information from large data sets using techniques such as machine learning and statistical analysis.
- The goal of data mining is to extract useful information from large datasets and use it to make predictions or inform decision-making.
- Data mining is important because it allows organizations to uncover insights and trends in their data that would be difficult or impossible to discover manually.

DATA MINING ARCHITECTURE

- ***Data Sources*:** Data sources are the sources of data that are used in data mining. These can include structured and unstructured data from databases, files, sensors, and other sources. Data sources provide the raw data that is used in data mining and can be processed, cleaned, and transformed to create a usable data set for analysis.

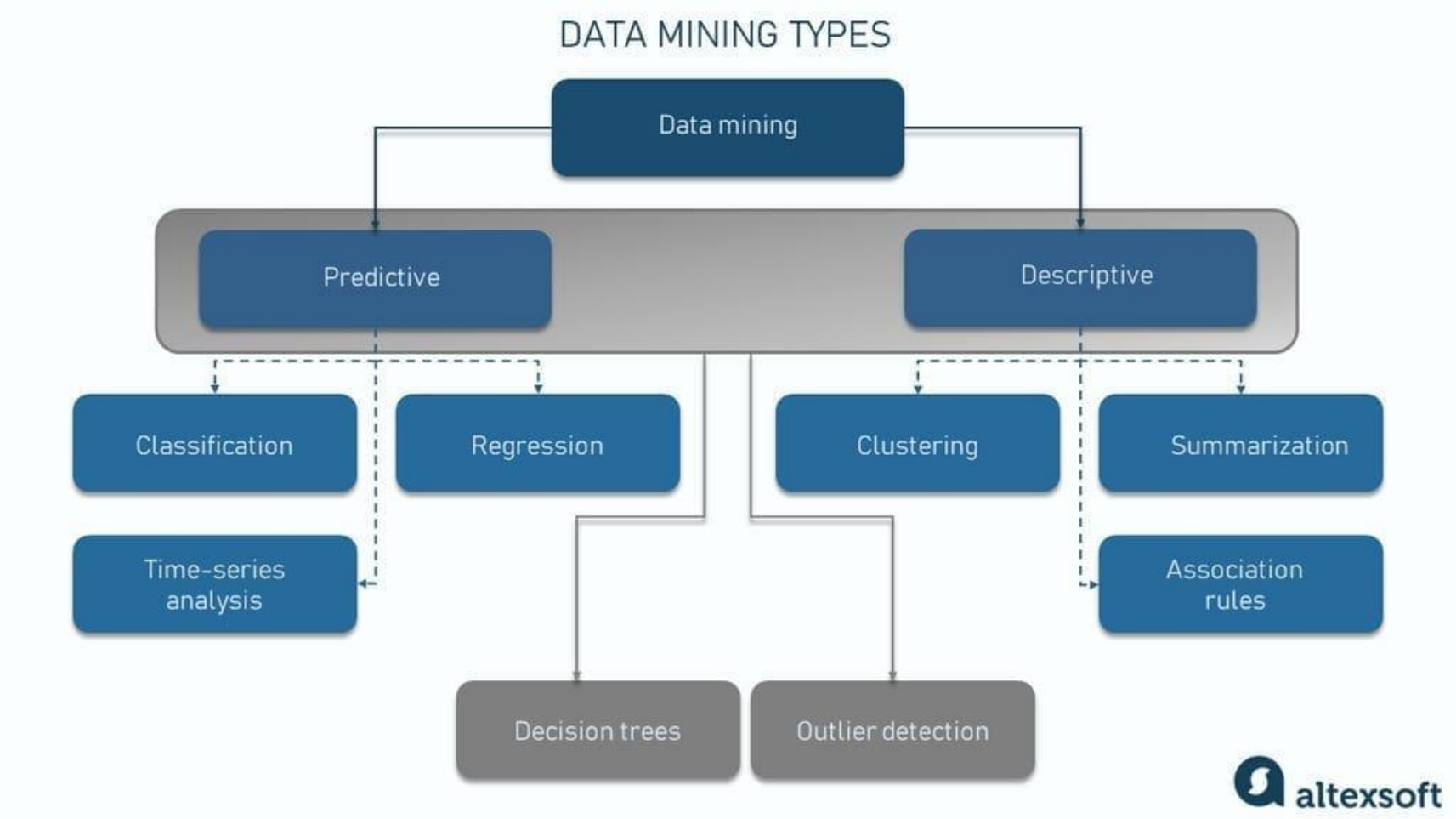
- ***Data Preprocessing*:** Data preprocessing is the process of preparing data for analysis. This typically involves cleaning and transforming the data to remove errors, inconsistencies, and irrelevant information, and to make it suitable for analysis. Data preprocessing is an important step in data mining, as it ensures that the data is of high quality and is ready for analysis.

DATA MINING ARCHITECTURE

- **Data Mining Algorithms:** Data mining algorithms are the algorithms and models that are used to perform data mining. These algorithms can include supervised and unsupervised learning algorithms, such as regression, classification, and clustering, as well as more specialized algorithms for specific tasks, such as association rule mining and anomaly detection. Data mining algorithms are applied to the data to extract useful insights and information from it.

- **Data Visualization:** Data visualization is the process of presenting data and insights in a clear and effective manner, typically using charts, graphs, and other visualizations. Data visualization is an important part of data mining, as it allows data miners to communicate their findings and insights to others in a way that is easy to understand and interpret.

TYPES OF DATA MINING



TYPES OF DATA MINING

- **Descriptive data mining** involves summarizing and describing the characteristics of a data set. This type of data mining is often used to explore and understand the data, identify patterns and trends, and summarize the data in a meaningful way. “What has happened?”

- **Predictive data mining** involves using data to build models that can make predictions or forecasts about future events or outcomes. This type of data mining is often used to identify and model relationships between different variables, and to make predictions about future events or outcomes based on those relationships. “What could happen?”

- **Prescriptive data mining** involves using data and models to make recommendations or suggestions about actions or decisions. This type of data mining is often used to optimize processes, allocate resources, or make other decisions that can help organizations achieve their goals. “What should we do?”

HOW DOES DATA MINING WORK?

Data Mining Phases / Steps

Define the Problem

Identify business goals

Identify data mining goals

Identify Required Data

Assess needed data

Collect and understand data

Prepare and Pre-process

Select required data

Cleanse/format data as necessary

Model the Data

Select algorithms

Build predictive models

Train and Test

Train the model with sample data sets

Test and iterate

Verify and Deploy

Verify final model

Prepare visualizations and deploy

HOW DOES DATA MINING WORK?

- Data mining is the process of extracting useful information and insights from large data sets. It typically involves several steps, including defining the problem, preparing the data, exploring the data, modeling the data, validating the model, implementing the model, and evaluating the results.
  - **Identify the problem:** The process of data mining typically begins with defining the problem or question that you want to answer with your data. This involves understanding the business context and goals and identifying the data that is relevant to the problem.
  - **Prepare the data:** Next, the data is prepared for analysis. This involves cleaning the data, transforming it into a usable format, and checking for errors or inconsistencies.
  - **Explore the data:** Once the data is prepared, you can begin exploring it to ***gain insights and understand its characteristics***. This typically involves using visualization and summary statistics to understand the distribution, patterns, and trends in the data.

- **Model the data:** The next step is to ***build models that can be used to make predictions or forecasts based on the data***. This involves choosing an appropriate modeling technique, fitting the model to the data, and evaluating its performance.
- **Validate the model:** After the model is built, ***it is important to validate its performance to ensure that it is accurate and reliable***. This typically involves using a separate data set (called a validation set) to evaluate the model's performance and make any necessary adjustments.
- **Implement the model:** Once the model has been validated, it can be ***implemented in a production environment to make predictions or recommendations***. This involves deploying the model and integrating it into the organization's existing systems and processes.
- **Evaluate the results:** The final step in the data mining process is to ***evaluate the results of the model and determine its effectiveness in solving the problem or achieving the goals***. This involves measuring the model's performance, comparing it to other models or approaches, and making any necessary changes or improvements.

Data Preprocessing

Data Cleaning

- **Missing Data**
  1. Ignore The Tuplet
  2. Fill The Missing Values (manually, by mean or by most probable value
- **Noisy Data**
  1. Binning Method
  2. Regression
  3. Clustering

Data Transformation

- **Normalization**
- **Atribute Selection**
- **Discretization**
- **Concept Hiererchy Generation**

Data Reduction

- **Data Cube Aggregation**
- **Attribute Subset Selection**
- **Numerosity Reduction**
- **Dimensionality Reduction**

DATA PREPROCESSING IN DATA MINING

- Data preprocessing is an important step in the data mining process that involves cleaning and transforming raw data to make it suitable for analysis. Some common steps in data preprocessing include:

- **Data Cleaning:** Data cleaning also goes by a few other names, including data wrangling, scrubbing, and remediation. This involves identifying and correcting errors or inconsistencies in the data, such as missing values, outliers, and duplicates. Various techniques can be used for data cleaning, such as imputation, removal, and transformation.

- **Data Integration:** This involves combining data from multiple sources to create a unified dataset. Data integration can be challenging as it requires handling data with different formats, structures, and semantics. Techniques such as record linkage and data fusion can be used for data integration.

DATA PREPROCESSING IN DATA MINING

- **Data Transformation:** This involves converting the data into a suitable format for analysis. Common techniques used in data transformation include normalization, standardization, and discretization. Normalization is used to scale the data to a common range, while standardization is used to transform the data to have zero mean and unit variance. Discretization is used to convert continuous data into discrete categories.

- **Data Reduction:** This involves reducing the size of the dataset while preserving the important information. Data reduction can be achieved through techniques such as feature selection and feature extraction. Feature selection involves selecting a subset of relevant features from the dataset, while feature extraction involves transforming the data into a lower-dimensional space while preserving the important information.

STEPS INVOLVED IN DATA PREPROCESSING:

- **1. Data Cleaning:**

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Handling Missing Data:**

This situation arises when some data is missing in the data. It can be handled in various ways.

Some of them are:

- **Ignore the tuples:**
  This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
- **Fill the Missing values:**
  There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- Assigning a constant to all missing values
- Imputation- Done by replacing the mean of all samples, the mean of samples with similar classification or resultant values, or any other logical manner.
- Replacing the missing value by the most probable value can be done using the Bayesian formula, decision tree, etc.

STEPS INVOLVED IN DATA PREPROCESSING:

- **(b) Reformatting:** This involves making data format changes into a standard format to ensure that the attributes such as date have a similar format throughout, performing binning of numerical values, and detecting and handling errors.
- **(c) Attribute Conversions:** Since some methods require only numerical inputs, different strategies need to be employed to handle binary, ordered, multi-valued nominal fields. These may include using 0 and 1 for binary fields, numbers preserving order for ordered nominal attributes, and integer or one-hot encoding for unordered attributes.
- **(d) Outlier Identification and Smoothening Out Noisy Data:** Noisy data is a meaningless data that can't be interpreted by machines. It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

- **Binning Method:**
  This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

- **Regression:**
  Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

- **Clustering:**
  This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

- **2. Data Transformation:**
  This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:
  - **Normalization:**
    It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)
  - **Attribute Selection:**
    In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
  - **Discretization:**
    This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
  - **Concept Hierarchy Generation:**
    Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

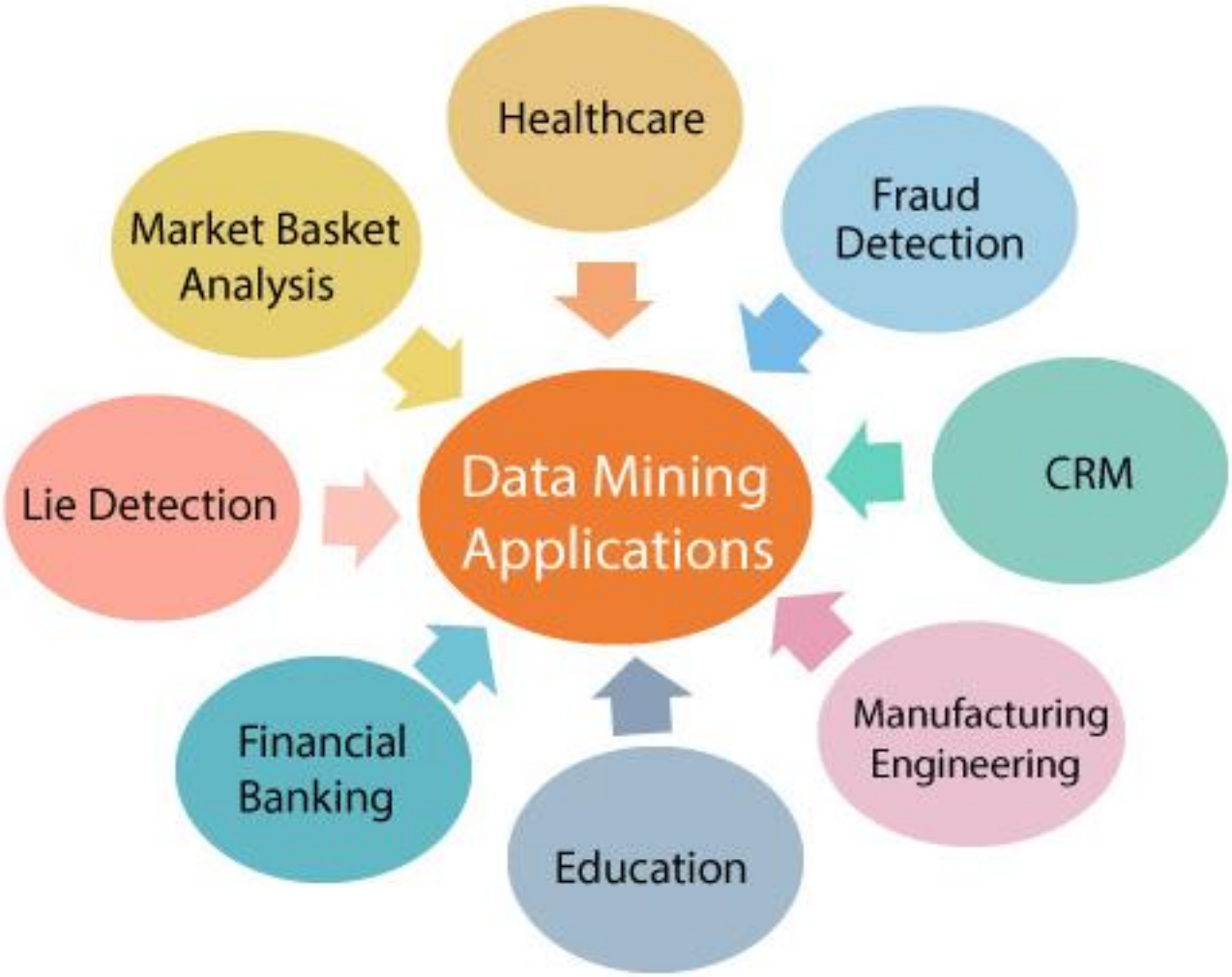
3. Data Reduction:

Data reduction techniques aim to derive a reduced representation of the data in terms of volume while closely maintaining the integrity of the original data. The various data reduction strategies include:

- **Dimensionality Reduction:** Dimensionality reduction is done by reducing the number of attributes to be considered. Some dimensionality reduction methods are:
  - **Wavelet transforms-** This involves converting the data vector into a vector of wavelet coefficients. This wavelet transformed data can then be truncated to obtain a compressed approximation of the original data by storing only a fraction of the strongest of the wavelet coefficients.
  - **Principal component analysis** - Principal component analysis or PCA works by searching for a set of orthogonal vectors, which is smaller than the original attribute vectors, that can best represent the data, thus resulting in dimensionality reduction. PCA compounds the original attributes into an alternative, smaller set.
  - **Attribute subset selection**- Involves selecting a set of features such that the weakly relevant or redundant features are removed. You can use heuristic methods such as stepwise forward selection, stepwise backward elimination, or a combination of the two, and decision tree induction to arrive at the subset of attributes.

- **Numerosity Reduction:** This involves replacing the original data with smaller forms of data representation to achieve volume reduction. The two types of methods used for this purpose are:
  - **Parametric**: These methods involve regression and log-linear models, whose parameters need to be stored instead of the actual data.
  - **Nonparametric**: Nonparametric methods involve storing the data in representations like histograms, clusters, a smaller sample of the original dataset, or data cube aggregation.
- **Compression:** This involves compressing the dataset while preserving the important information. Compression is often used to reduce the size of the dataset for storage and transmission purposes. This involves applying transformations to obtain a compressed representation of the original data. Depending on whether the reconstruction can be done with or without the loss of information the technique is called lossless or lossy compression. Dimensionality reduction and numerosity reduction techniques are also considered forms of data compression.

APPLICATIONS OF DATA MINING



APPLICATION OF DATA MINING

- Data mining has a wide range of applications and uses cases across many industries and domains. Some of the most common use cases of data mining include:

- **Market Basket Analysis:** Market basket analysis is a common use case of data mining in the retail and e-commerce industries. It involves analyzing data on customer purchases to identify items that are frequently purchased together, and using this information to make recommendations or suggestions to customers.

- **Fraud Detection:** Data mining is widely used in the financial industry to detect and prevent fraud. It involves analyzing data on transactions and customer behavior to identify patterns or anomalies that may indicate fraudulent activity.

- **Customer Segmentation:** Data mining is commonly used in the marketing and advertising industries to segment customers into different groups based on their characteristics and behavior. This information can then be used to tailor marketing and advertising campaigns to specific segments of customers.

APPLICATION OF DATA MINING

- **Predictive Maintenance:** Data mining is increasingly used in the manufacturing and industrial sectors to predict when equipment or machinery is likely to fail or require maintenance. It involves analyzing data on the performance and usage of equipment to identify patterns that can indicate potential failures, and using this information to schedule maintenance and prevent downtime.

- **Network Intrusion Detection:** Data mining is used in the cybersecurity industry to detect network intrusions and prevent cyber attacks. It involves analyzing data on network traffic and behavior to identify patterns that may indicate an attempted intrusion, and using this information to alert security teams and prevent attacks.

- Overall, data mining has a wide range of applications and use cases across many industries and domains. It is a powerful tool for uncovering insights and information hidden in data sets and is widely used to solve a variety of business and technical challenges.

DATA MINING VS. DATA ANALYTICS AND DATA WAREHOUSING

- **Data mining** is the process of extracting useful information and insights from large data sets. It involves applying algorithms and techniques to uncover hidden patterns and relationships in the data and to generate predictions and forecasts.
- **Data analytics** is the process of analyzing data to extract insights and information. It involves applying statistical and mathematical methods to data sets in order to understand and describe the data and draw conclusions and make predictions.
- **Data warehousing** is the process of storing and managing large data sets. It involves designing and implementing a database or data repository that can efficiently store and manage data, and that can be queried and accessed by data mining and analytics tools.
- In summary, data mining, data analytics, and data warehousing are closely related fields that are often used together to extract useful information and insights from large data sets. Data mining focuses on applying algorithms and techniques to uncover hidden patterns and relationships in the data, data analytics focuses on applying statistical and mathematical methods to data sets, and data warehousing focuses on storing and managing large data sets.

BENEFITS OF DATA MINING

- Data mining is the process of extracting useful information and insights from large data sets. It is a powerful and flexible tool that has many benefits, including:
- **Better Decision Making:** Data mining helps to extract useful information from large datasets, which can be used to make informed and accurate decisions. By analyzing patterns and relationships in the data, businesses can identify trends and make predictions that help them make better decisions.
- **Improved Marketing:** Data mining can help businesses identify their target market and develop effective marketing strategies. By analyzing customer data, businesses can identify customer preferences and behavior, which can help them create targeted advertising campaigns and offer personalized products and services.
- **Increased Efficiency:** Data mining can help businesses streamline their operations by identifying inefficiencies and areas for improvement. By analyzing data on production processes, supply chains, and employee performance, businesses can identify bottlenecks and implement solutions that improve efficiency and reduce costs.
- **Fraud Detection:** Data mining can be used to identify fraudulent activities in financial transactions, insurance claims, and other areas. By analyzing patterns and relationships in the data, businesses can identify suspicious behavior and take steps to prevent fraud.

BENEFITS OF DATA MINING

- **Customer Retention:** Data mining can help businesses identify customers who are at risk of leaving and develop strategies to retain them. By analyzing customer data, businesses can identify factors that contribute to customer churn and take steps to address those factors.
- **Competitive Advantage:** Data mining can help businesses gain a competitive advantage by identifying new opportunities and emerging trends. By analyzing data on customer behavior, market trends, and competitor activity, businesses can identify opportunities to innovate and differentiate themselves from their competitors.
- **Improved Healthcare:** Data mining can be used to improve healthcare outcomes by analyzing patient data to identify patterns and relationships. By analyzing medical records and other patient data, healthcare providers can identify risk factors, diagnose diseases earlier, and develop more effective treatment plans.

LIMITATIONS OF DATA MINING

- **Data quality** – One of the main limitations of data mining is the quality of the data. Data mining can only be as accurate and reliable as the data that it is based on, and poor-quality data can lead to inaccurate or misleading results.
- **Model bias** – Another limitation of data mining is the potential for bias in the models that are built from the data. If the data is not representative of the population, or if there is bias in the way the data is collected or analyzed, the models that are built from the data may be biased, and may not accurately reflect the underlying relationships in the data.
- **Ethical considerations** – Data mining also raises ethical considerations. The data that is collected and analyzed may be sensitive or personal, and organizations must ensure that they handle this data responsibly and in compliance with relevant laws and regulations.
- **Technical challenges** – Data mining can also be technically challenging, especially when dealing with large and complex data sets. Extracting useful information and insights from data can require specialized skills and expertise, and can be time-consuming and resource-intensive.
- Overall, data mining is a powerful and flexible tool, but it has its limitations and challenges. Organizations must be aware of these limitations, and take steps to address them in order to ensure that their data mining efforts are accurate, reliable, and ethical.

DATA MINING AND SOCIAL MEDIA

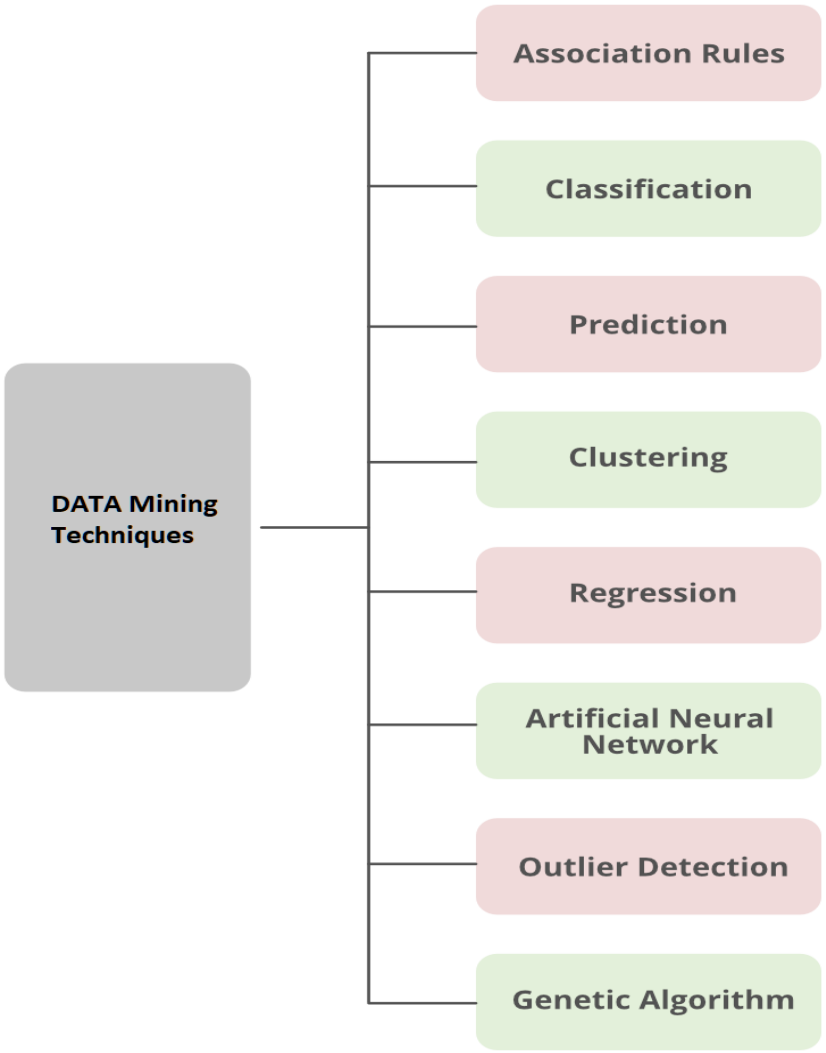
- Data mining is the process of extracting useful information and insights from large data sets, and social media is a rich source of data that can be mined for insights and information. By analyzing data from social media platforms, organizations can gain valuable insights into consumer behavior, preferences, and opinions, and use this information to inform and improve their marketing and advertising efforts.
- Some common examples of data mining in social media include:
- **Sentiment analysis** – Sentiment analysis is a common application of data mining in social media. By analyzing the text of social media posts and comments, organizations can determine the overall sentiment of users towards their products, services, or brand, and use this information to improve their marketing and customer service efforts.

- **Influencer identification** – Data mining can also be used to identify influencers on social media. By analyzing data on user engagement, reach, and influence, organizations can identify users who are influential and have a large audience and target their marketing and advertising efforts toward these users.

DATA MINING AND SOCIAL MEDIA

- **Trend analysis** – Data mining can also be used to analyze trends on social media. By analyzing data on user behavior and interactions, organizations can identify emerging trends and topics of interest, and use this information to tailor their content and messaging to be more relevant and engaging.
- Overall, data mining is a powerful tool for extracting useful information and insights from social media data. By analyzing data from social media platforms, organizations can gain valuable insights into consumer behavior, preferences, and opinions, and use this information to inform and improve their marketing and advertising efforts.

DATA MINING TECHNIQUES



ASSOCIATION RULE MINING

- Association rule mining is a data mining technique that is used to identify and explore relationships between items or attributes in a data set. In association rule mining, the goal is to identify patterns and rules that describe the co-occurrence or occurrence of items or attributes in the data set and to evaluate the strength and significance of these patterns and rules.
- There are many different algorithms and methods for association rule mining, including the Apriori algorithm and the Frequent Pattern (FP)-growth algorithm. These algorithms differ in the way that they generate and evaluate association rules, and in the assumptions that they make about the data.
- In general, association rule mining is used to answer questions such as:
  - What are the main patterns and rules in the data?
  - How strong and significant are these patterns and rules?
  - What are the implications of these patterns and rules for the data set and the domain?
- Overall, association rule mining is a powerful and widely used data mining technique that is used to identify and explore relationships between items or attributes in a data set. It is a crucial tool for many applications in the field of data mining and is commonly used in areas such as market basket analysis, recommendation systems, and fraud detection.

APRIORI ALGORITHM

- This algorithm was given by the **R. Agrawal** and **Srikant** in the year **1994**. It is mainly used for *market basket analysis* and helps to find those products that can be bought together. It can also be used in the healthcare field to find drug reactions for patients.
- Apriori algorithm refers to the algorithm which is used to calculate the association rules between objects. It means how two or more objects are related to one another. In other words, we can say that the apriori algorithm is an association rule leaning that analyzes that people who bought product A also bought product B.
- The primary objective of the apriori algorithm is to create the association rule between different objects. The association rule describes how two or more objects are related to one another. Apriori algorithm is also called frequent pattern mining. Generally, you operate the Apriori algorithm on a database that consists of a huge number of transactions. Let's understand the apriori algorithm with the help of an example; suppose you go to Reliance Smart Bazar and buy different products. It helps the customers buy their products with ease and increases the sales performance of the Reliance Smart Bazar.

APRIORI ALGORITHM

- We take an example to understand the concept better. You must have noticed that the Pizza shop seller makes a pizza, soft drink, and French fries combo together. He also offers a discount to their customers who buy these combos. Do you ever think why does he do so? He thinks that customers who buy pizza also buy soft drinks and French fries. However, by making combos, he makes it easy for the customers. At the same time, he also increases his sales performance.
- Similarly, you go to Reliance Smart Bazar, and you will find biscuits, chips, and Chocolate bundled together. It shows that the shopkeeper makes it comfortable for the customers to buy these products in the same place.
- The above two examples are the best examples of Association Rules in Data Mining. It helps us to learn the concept of apriori algorithms.

APRIORI ALGORITHM

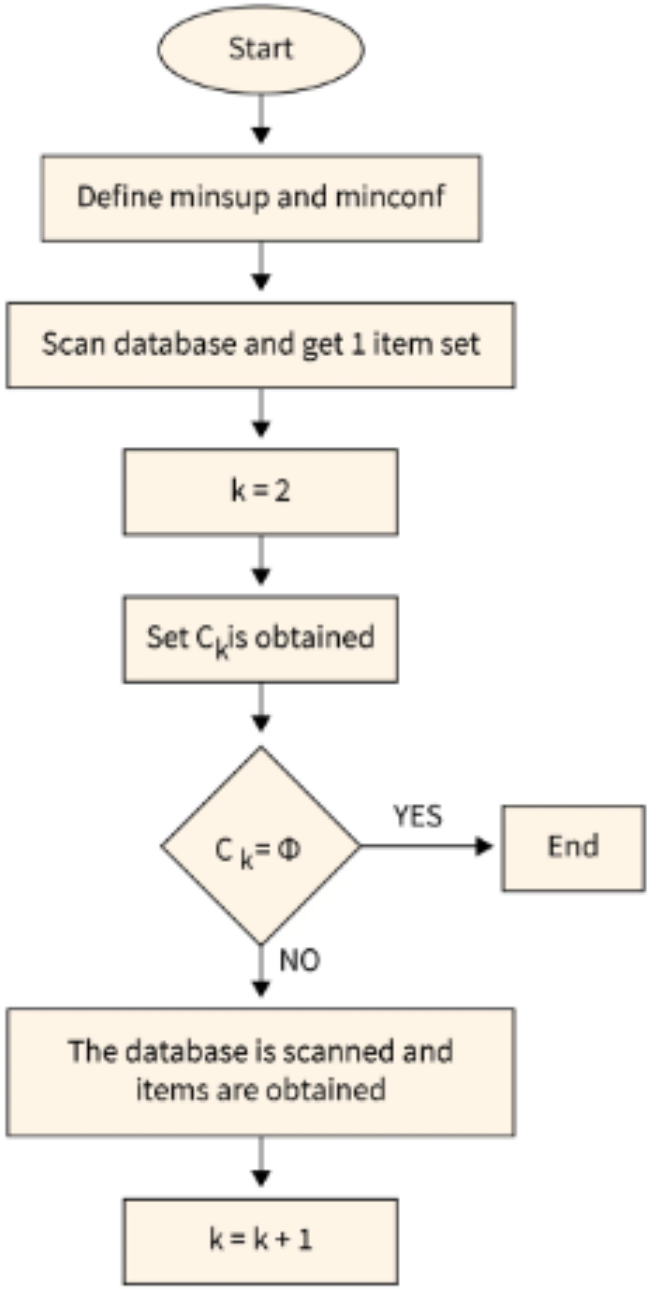
- Apriori algorithm refers to an algorithm that is used in mining frequent products sets and relevant association rules. Generally, the apriori algorithm operates on a database containing a huge number of transactions. For example, the items customers buy at a Reliance Smart Bazar.
- Apriori algorithm helps the customers to buy their products with ease and increases the sales performance of the particular store.
- Components of Apriori algorithm
  - Support
  - Confidence
  - Lift
- Let's take an example to understand this concept.
- Suppose you have 4000 customers transactions in a Reliance Smart Bazar. You have to calculate the Support, Confidence, and Lift for two products, and you may say Biscuits and Chocolate. This is because customers frequently buy these two items together.
- Out of 4000 transactions, 400 contain Biscuits, whereas 600 contain Chocolate, and these 600 transactions include a 200 that includes Biscuits and chocolates. Using this data, we will find out the support, confidence, and lift.

APRIORI ALGORITHM

- **Support**: Support refers to the default popularity of any product. You find the support as a quotient of the division of the number of transactions comprising that product by the total number of transactions. Hence, we get
- **Support** (Biscuits) = (Transactions relating biscuits) / (Total transactions) = 400/4000 = 10 percent.
- **Confidence**: Confidence refers to the possibility that the customers bought both biscuits and chocolates together. So, you need to divide the number of transactions that comprise both biscuits and chocolates by the total number of transactions to get the confidence.
- Hence, **Confidence** = (Transactions relating both biscuits and Chocolate) / (Total transactions involving Biscuits) = 200/400 = 50 percent.
- It means that 50 percent of customers who bought biscuits bought chocolates also.
- **Lift**: Consider the above example; lift refers to the increase in the ratio of the sale of chocolates when you sell biscuits. The mathematical equations of lift are given below.
- **Lift** = (Confidence (Biscuits - chocolates)/ (Support (Biscuits) = 50/10 = 5
- It means that the probability of people buying both biscuits and chocolates together is five times more than that of purchasing the biscuits alone. If the lift value is below one, it requires that the people are unlikely to buy both the items together. Larger the value, the better is the combination.

STEPS IN APRIORI ALGORITHM

- **Define minimum support threshold** - This is the minimum number of times an item set must appear in the dataset to be considered as frequent. The support threshold is usually set by the user based on the size of the dataset and the domain knowledge.
- **Generate a list of frequent 1-item sets** - Scan the entire dataset to identify the items that meet the minimum support threshold. These item sets are known as frequent 1-item sets.
- **Generate candidate item sets** - In this step, the algorithm generates a list of candidate item sets of length k+1 from the frequent k-item sets identified in the previous step.
- **Count the support of each candidate item set** - Scan the dataset again to count the number of times each candidate item set appears in the dataset.
- **Prune the candidate item sets** - Remove the item sets that do not meet the minimum support threshold.
- Repeat steps 3-5 until no more frequent item sets can be generated.
- **Generate association rules** - Once the frequent item sets have been identified, the algorithm generates association rules from them. Association rules are rules of form A -> B, where A and B are item sets. The rule indicates that if a transaction contains A, it is also likely to contain B.
- **Evaluate the association rules** - Finally, the association rules are evaluated based on metrics such as confidence and lift.


Start
Define minsup and minconf
Scan database and get 1 item set
k = 2
Set C_k is obtained
$C_k = \Phi$
YES
End
NO
The database is scanned and items are obtained
k = k + 1

EXAMPLE

- Let's try to understand the Apriori algorithm implementation using an example. In this example, we will use a minimum support threshold of 3. This means an item set must appear in at least three transactions to be considered frequent.
- Let's consider the transaction dataset of a retail store as shown in the below table.

TID	Items
T1	{milk, bread}
T2	{bread, sugar}
T3	{bread, butter}
T4	{milk, bread, sugar}
T5	{milk, bread, butter}
T6	{milk, bread, butter}
T7	{milk, sugar}
T8	{milk, sugar}
T9	{sugar, butter}
T10	{milk, sugar, butter}
T11	{milk, bread, butter}

EXAMPLE

- Let's calculate support for each item present in the dataset. As shown in the below table, support for all items is greater than 3. It means that all items are considered as frequent 1-itemsets and will be used to generate candidates for 2-itemsets.

- Below table represents all candidates generated from frequent 1-itemsets identified from the previous step and their support value.

Item	Support (Frequency)
milk	8
bread	7
sugar	5
butter	7

Candidate Item Sets	Support (Frequency)
{milk, bread}	5
{milk, sugar}	3
{milk, butter}	5
{bread, sugar}	2
{bread, butter}	3
{sugar, butter}	2

EXAMPLE

- Now remove candidate item sets that do not meet the minimum support threshold of 3. After this step, frequent 2-itemsets would be - {milk, bread}, {milk, sugar}, {milk, butter}, and {bread, butter}. In the next step, let's generate candidates for 3-itemsets and calculate their respective support values. It is shown in the below table.

Candidate Item Sets	Support (Frequency)
{milk, bread, sugar}	1
{milk, bread, butter}	3
{milk, sugar, butter}	1

EXAMPLE

- As we can see in the above table, only one candidate item set exceeds the minimum defined support threshold - {milk, bread, butter}. As there is only one 3-itemset exceeding minimum support, we can't generate candidates for 4-itemsets. So, in the next step, we can write the association rules and their respective metrics, as shown in the below table.

Candidate Item Sets	Support (Frequency)
{milk, bread}	{butter} (Confidence - 60%)1
{bread, butter}	{milk} (Confidence - 100%)
{milk, butter}	{bread} (Confidence - 60%)

Based on association rules mentioned in the above table, we can recommend products to the customer or optimize product placement in retail stores.

ADVANTAGES OF APRIORI ALGORITHM

- Apriori algorithm is simple and easy to implement, making it accessible even to those without a deep understanding of data mining or machine learning.
- Apriori algorithm can handle large datasets and run on distributed systems, making it scalable for large-scale applications.
- Apriori algorithm is one of the most widely used algorithms for association rule mining and is supported by many popular data mining tools.

LIMITATIONS OF APRIORI ALGORITHM

- Apriori algorithm can be computationally expensive, especially for large datasets with many item sets. For example, if a dataset contains 10^4 from frequent 1-item sets, it will generate more than 10^7 2-length candidates, which makes this algorithm computationally expensive.
- Apriori algorithm can generate a large number of rules, making it difficult to sift through and identify the most important ones.
- The algorithm requires multiple database scans to generate frequent item sets, which can be a limitation in systems where data access is slow or expensive.
- Apriori algorithm is sensitive to data sparsity, meaning it may not perform well on datasets with a low frequency of item sets.

FREQUENT PATTERN-GROWTH ALGORITHM

- The **FP Growth algorithm** is a popular method for frequent pattern mining in data mining. It works by constructing a **frequent pattern tree (FP-tree)** from the input dataset. The **FP-tree** is a compressed representation of the dataset that captures the frequency and association information of the items in the data.

- The algorithm first scans the dataset and maps each transaction to a path in the tree. Items are ordered in each transaction based on their frequency, with the most frequent items appearing first. Once the FP tree is constructed, frequent item sets can be generated by recursively mining the tree. This is done by starting at the bottom of the tree and working upwards, finding all combinations of item sets that satisfy the minimum support threshold.

- The FP Growth algorithm in data mining has several advantages over other frequent pattern mining algorithms, such as Apriori.
  The **Apriori algorithm** is not suitable for handling large datasets because it generates a large number of candidates and requires multiple scans of the database to my frequent items. In comparison, the FP Growth algorithm requires only a single scan of the data and a small amount of memory to construct the FP tree. It can also be **parallelized to improve performance**.

WORKING ON FP GROWTH ALGORITHM

- The working of the FP Growth algorithm in data mining can be summarized in the following steps:
- **Scan the database:**
  In this step, the algorithm scans the input dataset to determine the frequency of each item. This determines the order in which items are added to the FP tree, with the most frequent items added first.
- **Sort items:**
  In this step, the items in the dataset are sorted in descending order of frequency. The infrequent items that do not meet the minimum support threshold are removed from the dataset. This helps to reduce the dataset's size and improve the algorithm's efficiency.
- **Construct the FP-tree:**
  In this step, the FP-tree is constructed. The FP-tree is a compact data structure that stores the frequent item sets and their support counts.
- **Generate frequent item sets:**
  Once the FP-tree has been constructed, frequent item sets can be generated by recursively mining the tree. Starting at the bottom of the tree, the algorithm finds all combinations of frequent item sets that satisfy the minimum support threshold.
- **Generate association rules:**
  Once all frequent item sets have been generated, the algorithm post-processes the generated frequent item sets to generate association rules, which can be used to identify interesting relationships between the items in the dataset.

FP-TREE

- The **FP-tree (Frequent Pattern tree)** is a data structure used in the FP Growth algorithm for frequent pattern mining. It represents the frequent item sets in the input dataset compactly and efficiently. The FP tree consists of the following components:
- **Root Node:**
  The root node of the FP-tree represents an empty set. It has no associated item but a pointer to the first node of each item in the tree.
- **Item Node:**
  Each item node in the FP-tree represents a unique item in the dataset. It stores the item name and the frequency count of the item in the dataset.
- **Header Table:**
  The header table lists all the unique items in the dataset, along with their frequency count. It is used to track each item's location in the FP tree.
- **Child Node:**
  Each child node of an item node represents an item that co-occurs with the item the parent node represents in at least one transaction in the dataset.
- **Node Link:**
  The node-link is a pointer that connects each item in the header table to the first node of that item in the FP-tree. It is used to traverse the conditional pattern base of each item during the mining process.
- The FP tree is constructed by scanning the input dataset and inserting each transaction into the tree one at a time. For each transaction, the items are sorted in descending order of frequency count and then added to the tree in that order. If an item exists in the tree, its frequency count is incremented, and a new path is created from the existing node. If an item does not exist in the tree, a new node is created for that item, and a new path is added to the tree. We will understand in detail how FP-tree is constructed in the next section.

EXAMPLE

Transaction ID	Items
T1	{M, N, O, E, K, Y}
T2	{D, O, E, N, Y, K}
T3	{K, A, M, E}
T4	{M, C, U, Y, K}
T5	{C, O, K, O, E, I}

- Let's scan the above database and compute the frequency of each item as shown in the below table.

Item	Frequency
A	1
C	2
D	1
E	4
I	1
K	5
M	3
N	2
O	3
U	1
Y	3

EXAMPLE

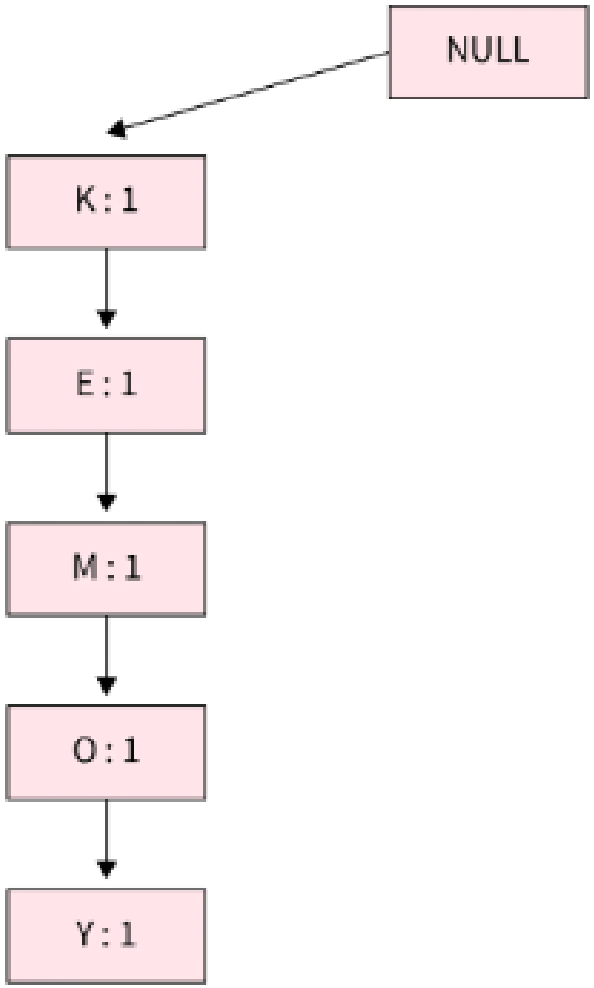
- Let's consider minimum support as 3. After removing all the items below minimum support in the above table, we would remain with these items - {K: 5, E: 4, M : 3, O : 3, Y : 3}. Let's re-order the transaction database based on the items above minimum support. In this step, in each transaction, we will remove infrequent items and re-order them in the descending order of their frequency, as shown in the table below.

Transaction ID	Items	Ordered Itemset
T1	{M, N, O, E, K, Y}	{K, E, M, O, Y}
T2	{D, O, E, N, Y, K}	{K, E, O, Y}
T3	{K, A, M, E}	{K, E, M}
T4	{M, C, U, Y, K}	{K, M, Y}
T5	{C, O, K, O, E, I}	{K, E, O}

Now we will use the ordered item set in each transaction to build the FP tree. Each transaction will be inserted individually to build the FP tree, as shown below -

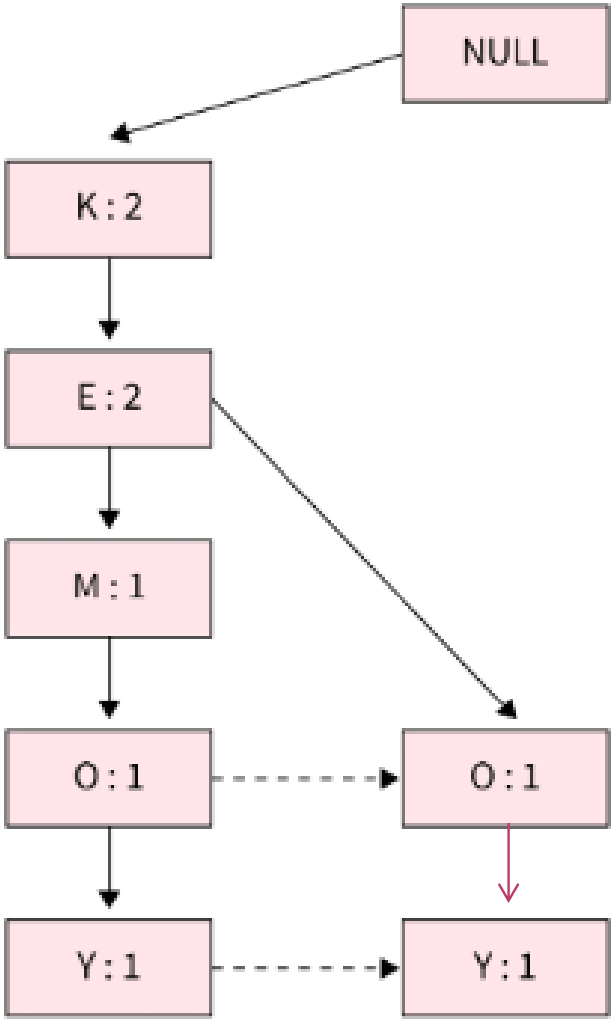
EXAMPLE

- **First Transaction {K, E, M, O, Y}:**
  In this transaction, all items are simply linked, and their support count is initialized as 1.



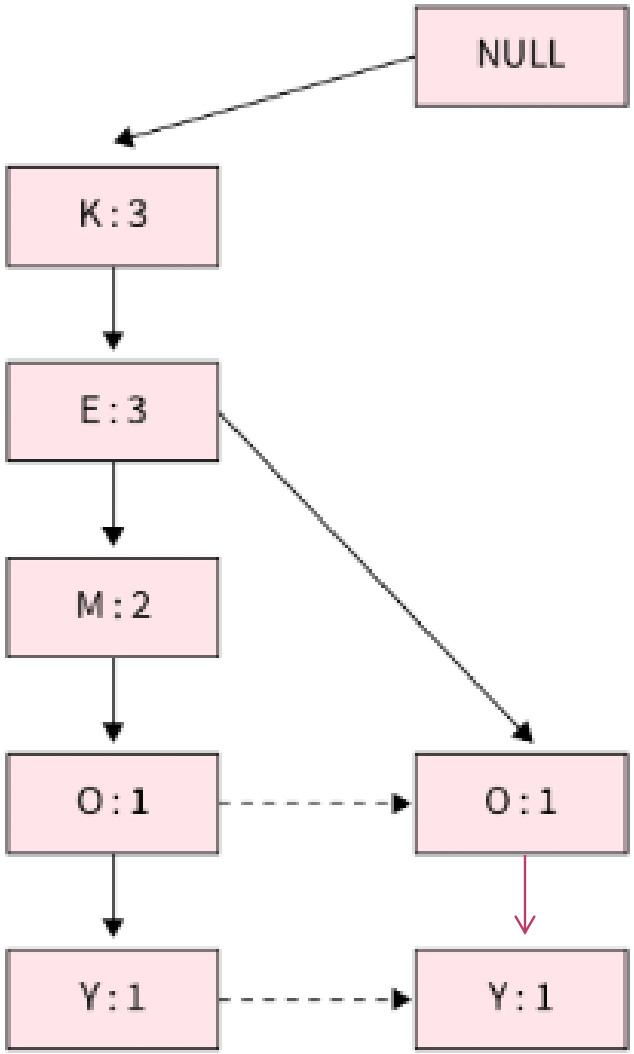
EXAMPLE

- **Second Transaction {K, E, O, Y}:**
  In this transaction, we will increase the support count of K and E in the tree to 2. As no direct link is available from E to O, we will insert a new path for O and Y and initialize their support count as 1.



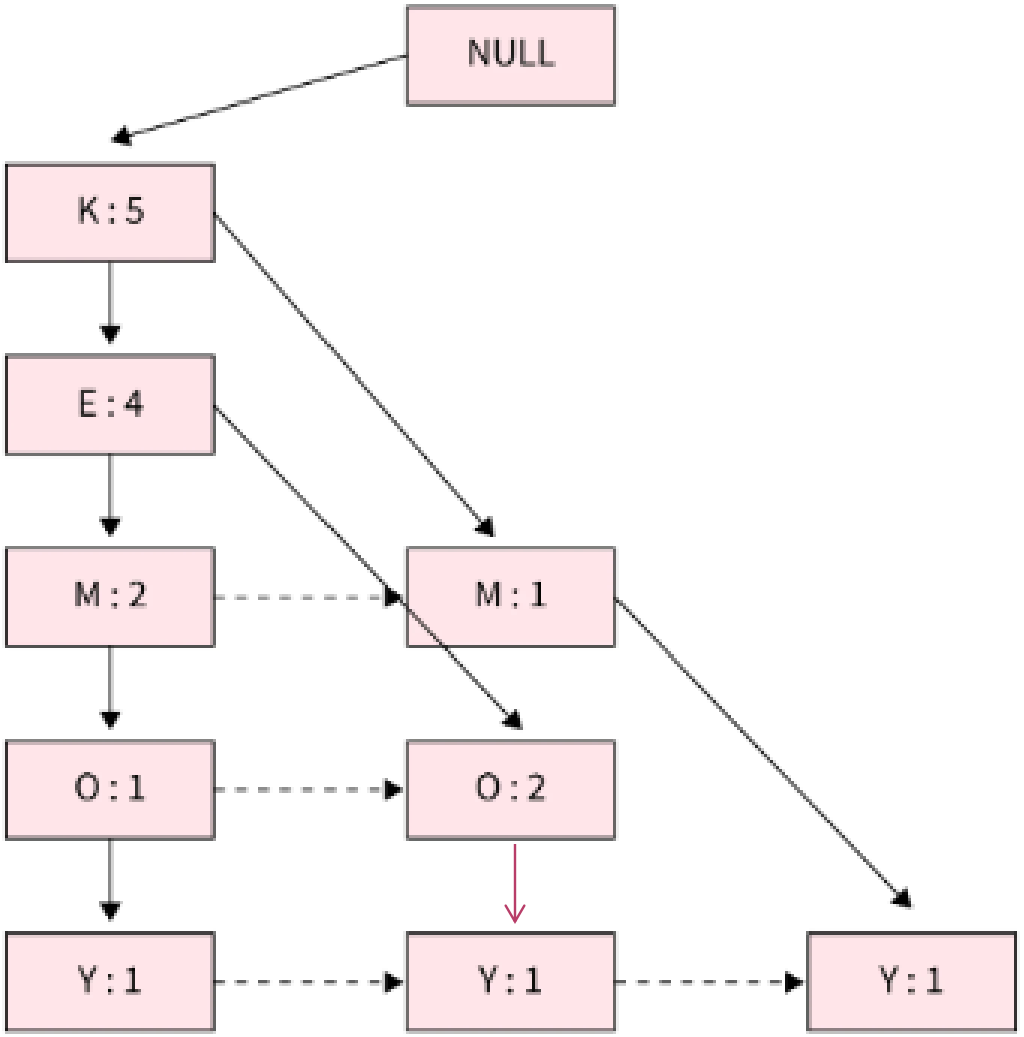
EXAMPLE

- **Third Transaction {K, E, M}:** After inserting this transaction, the tree will look as shown below. We will increase the support count for K and E to 3 and for M to 2.



EXAMPLE

- **Fourth Transaction {K, M, Y}** and **Fifth Transaction {K, E, O}:** After inserting the last two transactions, the FP-tree will look like as shown below:



EXAMPLE

- Now we will create a **Conditional Pattern Base** for all the items. The conditional pattern base is the path in the tree ending at the given frequent item. For example, for item O, the paths {K, E, M} and {K, E} will result in item O. The conditional pattern base for all items will look like as shown below table:

Item	Conditional Pattern Base
Y	{K, E, M, O : 1}, {K, E, O : 1}, {K, M : 1}
O	{K, E, M : 1}, {K, E : 2}
M	{K, E : 2}, {K : 1}
E	{K : 4}
K	

EXAMPLE

- Now for each item, we will build a conditional frequent pattern tree. It is computed by identifying the set of elements common in all the paths in the conditional pattern base of a given frequent item and computing its support count by summing the support counts of all the paths in the conditional pattern base. The conditional frequent pattern tree will look like this as shown below table:

Item	Conditional Pattern Base	Conditional FP Tree
Y	{K, E, M, O : 1}, {K, E, O : 1}, {K, M : 1}	{K : 3}
O	{K, E, M : 1}, {K, E : 2}	{K, E : 3}
M	{K, E : 2}, {K: 1}	{K : 3}
E	{K: 4}	{K: 4}
K		

EXAMPLE

- From the above conditional FP tree, we will generate the frequent item sets as shown in the below table:

Item	Frequent Patterns
Y	{K, Y - 3}
O	{K, O - 3}, {E, O - 3}, {K, E, O - 3}
M	{K, M - 3}
E	{K, E - 4}

EXAMPLE

- The transaction which we consider here suppose consists of 5 items such as-
- Asparagus (A), Corn (C), Beans (B), Tomatoes (T) & Squash (S)
- Let us also consider the minimum support for this small transaction data to be 2. Hence, min_support = 2.

Transaction ID	List of items in the transaction
T1	B , A , T
T2	A , C
T3	A , S
T4	B , A , C
T5	B , S
T6	A , S
T7	B , S
T8	B , A , S , T
T9	B , A , S

FP GROWTH ALGORITHM VS. APRIORI ALGORITHM

Factor	FP Growth Algorithm	Apriori Algorithm
Working	FP Growth uses FP-tree to mine frequent itemsets.	Apriori algorithm mines frequent items in an iterative manner - 1-itemsets, 2-itemsets, 3-itemsets, etc.
Candidate Generation	Generates frequent item sets by constructing the FP-Tree and recursively generating conditional pattern bases.	Generates candidate itemsets by joining and pruning.
Data Scanning	Scans the database only twice to construct the FP-Tree and generate conditional pattern bases.	Scans the database multiple times for frequent itemsets.
Memory Usage	Requires less memory than Apriori as it constructs the FP-Tree, which compresses the database	Requires a large amount of memory to store candidate itemsets.
Speed	Faster due to efficient data compression and generation of frequent itemsets.	Slower due to multiple database scans and candidate generation.
Scalability	Performs well on large datasets due to efficient data compression and generation of frequent itemsets.	Performs poorly on large datasets due to a large number of candidate item sets.

ADVANTAGES OF FP GROWTH ALGORITHM

- The FP Growth algorithm in data mining has several advantages over other frequent item set mining algorithms, as mentioned below:
- **Efficiency:**
  FP Growth algorithm is faster and more memory-efficient than other frequent item set mining algorithms such as Apriori, especially on large datasets with high dimensionality. This is because it generates frequent item sets by constructing the FP-Tree, which compresses the database and requires only two scans.
- **Scalability:**
  FP Growth algorithm scales well with increasing database size and item set dimensionality, making it suitable for mining frequent item sets in large datasets.
- **Resistant to noise:**
  FP Growth algorithm is more resistant to noise in the data than other frequent item set mining algorithms, as it generates only frequent item sets and ignores infrequent item sets that may be caused by noise.
- **Parallelization:**
  FP Growth algorithm can be easily parallelized, making it suitable for distributed computing environments and allowing it to take advantage of multi-core processors.

DISADVANTAGES OF FP GROWTH ALGORITHM

- While the FP Growth algorithm in data mining has several advantages, it also has some limitations and disadvantages, as mentioned below:
- **Memory consumption:**
  Although the FP Growth algorithm is more memory-efficient than other frequent item set mining algorithms, storing the FP-Tree and the conditional pattern bases can still require a significant amount of memory, especially for large datasets.
- **Complex implementation:**
  The FP Growth algorithm is more complex than other frequent item set mining algorithms, making it more difficult to understand and implement.

CLASSIFICATION

- Classification is a data mining technique that is used to predict the class or category of an item or instance based on its characteristics or attributes. In classification analysis, the goal is to build a model that can accurately predict the class of an item based on its attributes and to evaluate the performance of the model.
- There are many different types of classification models, including decision trees, k-nearest neighbors, and support vector machines. These models differ in the way that they model the relationship between the classes and the attributes, and in the assumptions that they make about the data.
- In general, classification models are used to answer questions such as:
  - What is the relationship between the classes and the attributes
  - How well does the model fit the data?
  - How accurate are the predictions made by the model?
- Overall, classification is a powerful and widely used data mining technique that is used to predict the class or category of an item based on its characteristics. It is a crucial tool for many applications in the field of data mining and is commonly used in areas such as marketing, finance, and healthcare.

PREDICTION

- Data Prediction is a two-step process, similar to that of data classification. Although, for prediction, we do not utilize the phrasing of “Class label attribute” because the attribute for which values are being predicted is consistently valued(ordered) instead of categorical (discrete-esteemed and unordered). The attribute can be referred to simply as the predicted attribute. Prediction can be viewed as the construction and use of a model to assess the class of an unlabeled object, or to assess the value or value ranges of an attribute that a given object is likely to have.

CLUSTERING

- Clustering is a data mining technique that is used to group items or instances in a data set into clusters or groups based on their similarity or proximity. In clustering analysis, the goal is to identify and explore the natural structure or organization of the data, and to uncover hidden patterns and relationships.
- There are many different types of clustering algorithms, including k-means clustering, hierarchical clustering, and density-based clustering. These algorithms differ in the way that they define and measure similarity or proximity, and in the way that they group the items in the data set.
- In general, clustering is used to answer questions such as:
  - What is the natural structure or organization of the data?
  - What are the main clusters or groups in the data?
  - How similar or dissimilar are the items in the data set?
- Overall, clustering is a powerful and widely used data mining technique that is used to group items in a data set into clusters based on their similarity. It is a crucial tool for many applications in the field of data mining and is commonly used in areas such as market research, customer segmentation, and image analysis.

REGRESSION

- Regression can be defined as a statistical modeling method in which previously obtained data is used to predicting a continuous quantity for new observations. This classifier is also known as the Continuous Value Classifier. There are two types of regression models: Linear regression and multiple linear regression models.

ARTIFICIAL NEURAL NETWORK (ANN) CLASSIFIER METHOD

- An artificial neural network (ANN) also referred to as simply a “Neural Network” (NN), could be a process model supported by biological neural networks. It consists of an interconnected collection of artificial neurons. A neural network is a set of connected input/output units where each connection has a weight associated with it. During the knowledge phase, the network acquires by adjusting the weights to be able to predict the correct class label of the input samples. Neural network learning is also denoted as connectionist learning due to the connections between units. Neural networks involve long training times and are therefore more appropriate for applications where this is feasible. They require a number of parameters that are typically best determined empirically, such as the network topology or “structure”. Neural networks have been criticized for their poor interpretability since it is difficult for humans to take the symbolic meaning behind the learned weights. These features firstly made neural networks less desirable for data mining.
- The advantages of neural networks, however, contain their high tolerance to noisy data as well as their ability to classify patterns on which they have not been trained. In addition, several algorithms have newly been developed for the extraction of rules from trained neural networks. These issues contribute to the usefulness of neural networks for classification in data mining.
- An artificial neural network is an adjective system that changes its structure-supported information that flows through the artificial network during a learning section. The ANN relies on the principle of learning by example. There are two classical types of neural networks, perceptron and also multilayer perceptron.

OUTLIER DETECTION

- A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are Outliers. The investigation of OUTLIER data is known as OUTLIER MINING. An outlier may be detected using statistical tests which assume a distribution or probability model for the data, or using distance measures where objects having a small fraction of “close” neighbors in space are considered outliers. Rather than utilizing factual or distance measures, deviation-based techniques distinguish exceptions/outlier by inspecting differences in the principle attributes of items in a group.

GENETIC ALGORITHM

- Genetic algorithms are adaptive heuristic search algorithms that belong to the larger part of evolutionary algorithms. Genetic algorithms are based on the ideas of natural selection and genetics. These are intelligent exploitation of random search provided with historical data to direct the search into the region of better performance in solution space. They are commonly used to generate high-quality solutions for optimization problems and search problems. Genetic algorithms simulate the process of natural selection which means those species who can adapt to changes in their environment are able to survive and reproduce and go to the next generation. In simple words, they simulate “survival of the fittest” among individuals of consecutive generations for solving a problem. Each generation consist of a population of individuals and each individual represents a point in search space and possible solution. Each individual is represented as a string of character/integer/float/bits. This string is analogous to the Chromosome.

TIME SERIES ANALYSIS

- **Time-series analysis** is a specialized technique for analyzing and interpreting data collected at regular time intervals. This method is particularly useful for identifying trends, seasonal patterns, and cyclical behaviors. Unlike other data mining methods that deal with static information, time-series analysis focuses on data that changes over time.
- Airlines frequently use time-series analysis to forecast passenger demand. By examining historical data on flight bookings, cancellations, and passenger numbers over time, an airline can identify peak travel periods, seasonal variations, and long-term demand trends.