Day 3

Data Mining and Machine Learning

Session 3

Introduction to Data Mining

Session Outcome

From this session, learners should be able to Understand

- The Concept of Data Mining
- History of Data Mining
- Purposes and Steps of Data Mining
- Different Techniques and Applications of Data Mining
- Different Advantages and Disadvantages of Data Mining

Session Outline

- Definition and Concept of Data Mining
- What Do you Get from Data Mining?
- History of Data Mining
- Purposes of Data Mining
- Steps of Data Mining
- Techniques of Data Mining
- Application of Data Mining
- Advantages and Disadvantages of Data Mining

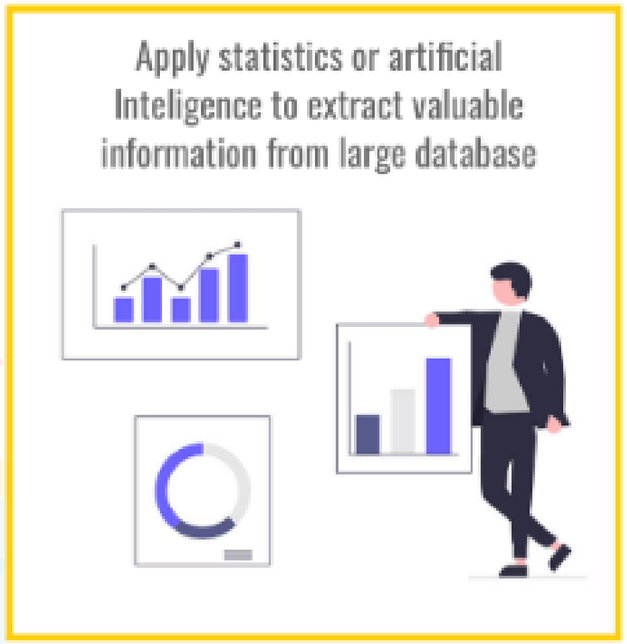
Data Mining?

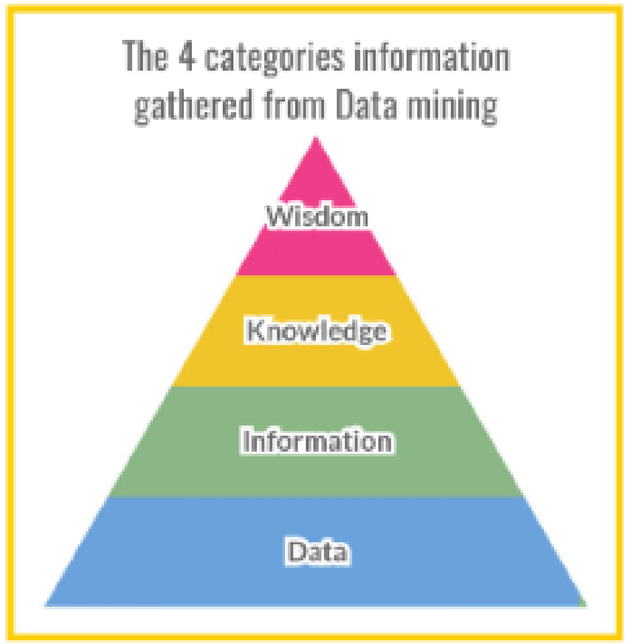
As the words seem to indicate, data mining is literally the method of digging into large volume of data to extract information and elaborate assumptions.

The more data we produce, the more difficult it becomes to make sense of all that data and derive meaningful insights from it. Imagine standing among trillions of data points; where do you start analyzing the big data landscape? Data mining provides a solution to this issue, one that shapes the ways businesses make decisions, reduce costs, and grow revenue. As a result, a variety of data science roles leverage mining as part of their daily responsibilities.

What is Data Mining?

Data mining is the process of searching and analyzing a large batch of raw data in order to identify patterns and extract useful information. It is most commonly defined as the process of using computers and automation to search large sets of data for patterns and trends, turning those findings into business insights and predictions, it uses data to evaluate future probabilities and develop actionable analyses.





What Do you Get from Data Mining?

There are fundamentally 4 categories of knowledge that derive from the practice of data mining:

- **Data:** Numerical values that have not been organized or classified
- **Information:** Data organized and classified
- **Knowledge:** Trends and findings obtained from information
- **Wisdom:** People make decisions based on knowledge

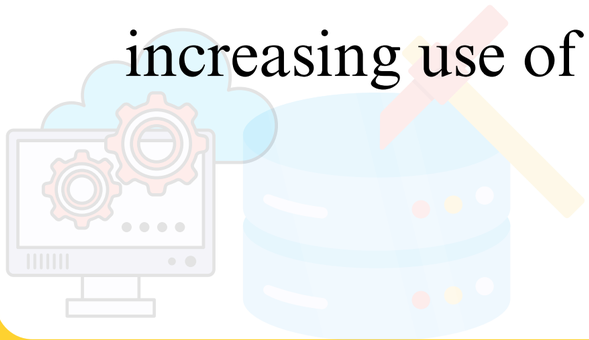
The information obtained by data mining is called “DIKW model”, the acronym of the 4 categories above. Information is more useful than data, knowledge is more useful than information, and wisdom is more useful than knowledge.

Data mining can be used to collect, organize, classify, and acquire knowledge, but it requires human judgment to utilize the acquired knowledge as wisdom.

History of Data Mining

Data Mining Evolution and Impact

- Originated before computers with Bayes' Theorem in 1763 and regression analysis discovery in 1805.
- Advancements in computer processors, data storage, and technology in the 1990s and 2000s made data mining more powerful and prolific.
- The book Moneyball introduced data mining to a broader audience in 2003.
- Today, data mining plays a critical role in countless industries due to the increasing use of big data solutions.

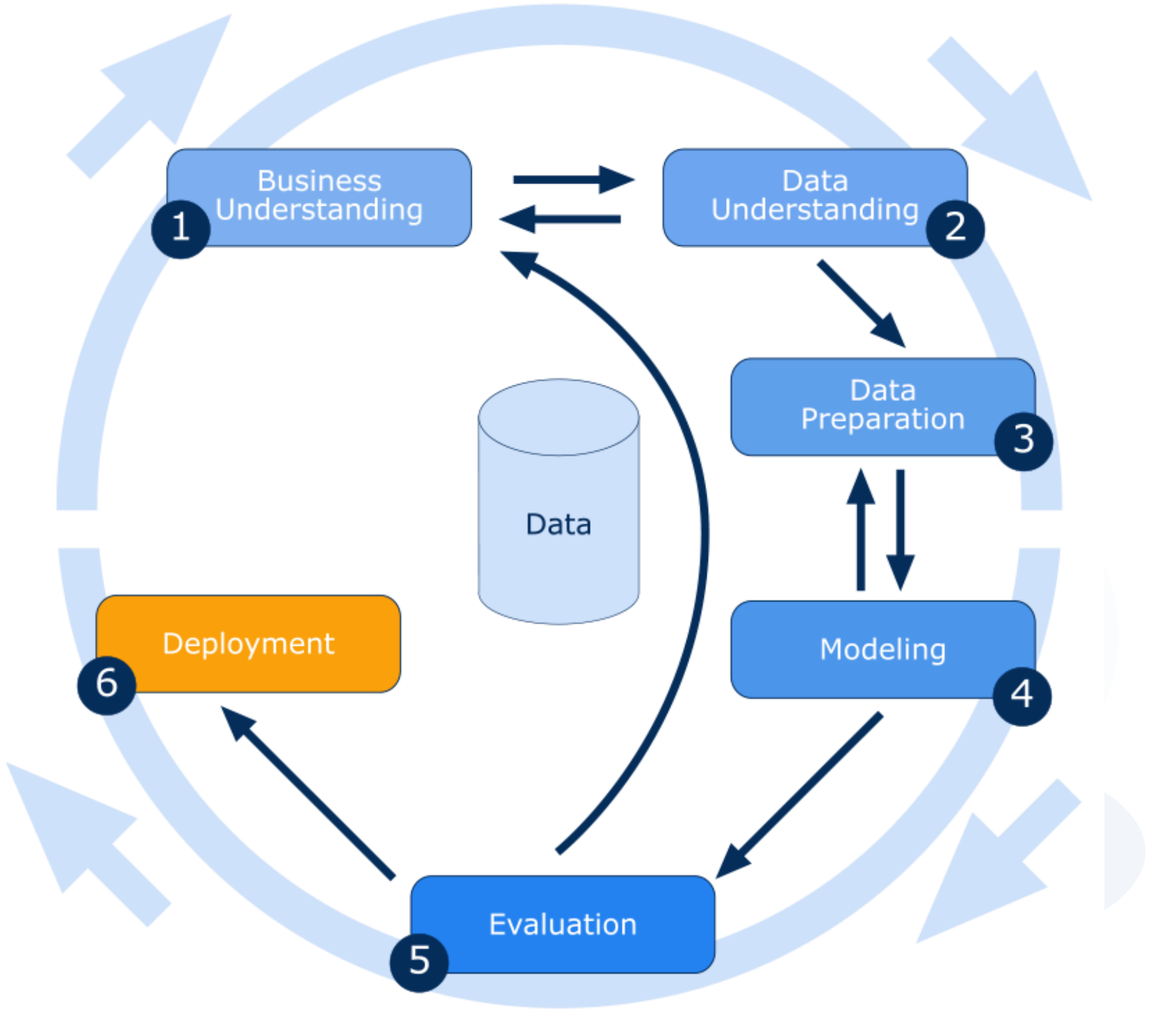


Purpose of Data Mining

The purpose of data mining is twofold:

- The creation of predictive power using the current information for predicting future values,
- Finding descriptive power for a better description of patterns in the present data.

Steps of Data Mining



Steps of Data Mining

Data mining is a systematic process of discovering previously unknown findings that hide within large datasets. The data mining process generally involves six main phases:

- Business understanding (Problem Statement),
- Data understanding,
- Data preparation,
- Data analysis,
- Evaluation,
- Deployment

In each stage useful insights are gathered to support the development of an effective data mining strategy.

Steps of Data Mining

Business Understanding

To get started, first ask these questions: What is our objective? What problem are we trying to solve? What data do we need to solve it?

Without a clear understanding of the proper data to mine, the project can produce errors, inaccurate results, or results that don't answer the correct questions.

Data Understanding

Once the overall objective is determined, proper data needs to be collected. The data must be relevant to subject matter and usually comes from a variety of sources such as sales records, customer surveys, and geolocation data. This phase's goal is to ensure the data correctly encompasses all necessary data sets to address the objective.

Steps of Data Mining

Data Preparation

The most time-consuming phase, the preparation phase, consists of three steps: **extraction, transformation, and loading** — also referred to as ETL. First, data is extracted from various sources and deposited into a staging area. Next, during the transformation step: the data is cleaned, null sets are populated, duplicative data is removed, errors are resolved, and all data is allocated into tables. In the final step, loading, the formatted data is loaded into the database for use.

Modeling

Data modeling addresses the relevant data set and considers the best statistical and mathematical approach to answering the objective question(s). It's also not uncommon to use different models on the same data to address specific objectives.

Steps of Data Mining

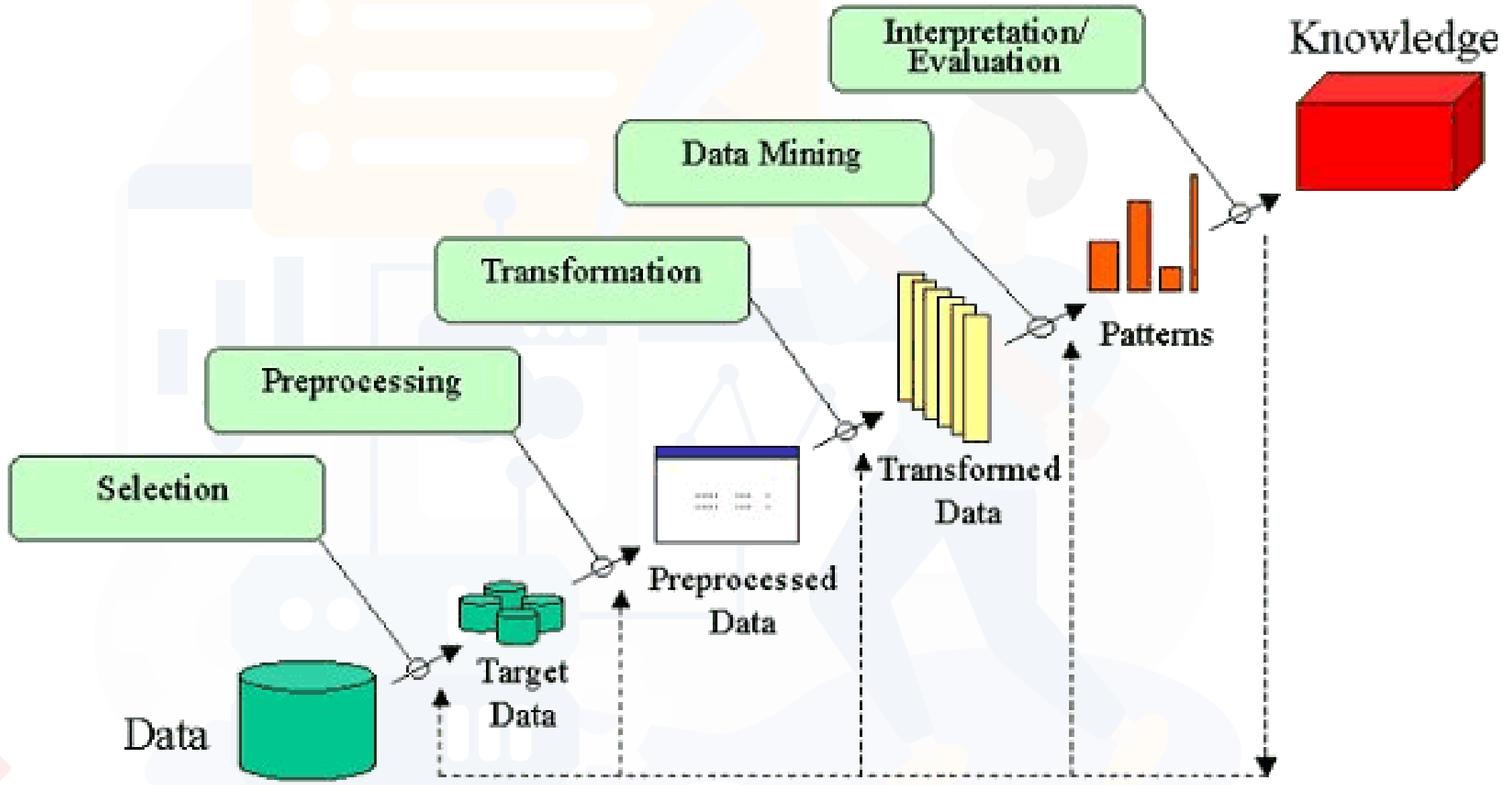
Evaluation

After the models are built and tested, it's time to evaluate their efficiency in answering the question identified during the business understanding phase. This is a human-driven phase, as the individual running the project must determine whether the model output sufficiently meets their objectives. If not, a different model can be created, or different data can be prepared.

Deployment

Once the data mining model is deemed accurate and successful in answering the objective question, it's time to put it to use. Deployment can occur in the form of a visual presentation or a report sharing insights. It also can lead to action such as generating a new sales strategy or implementing risk-reduction measures.

KDD and Data Mining



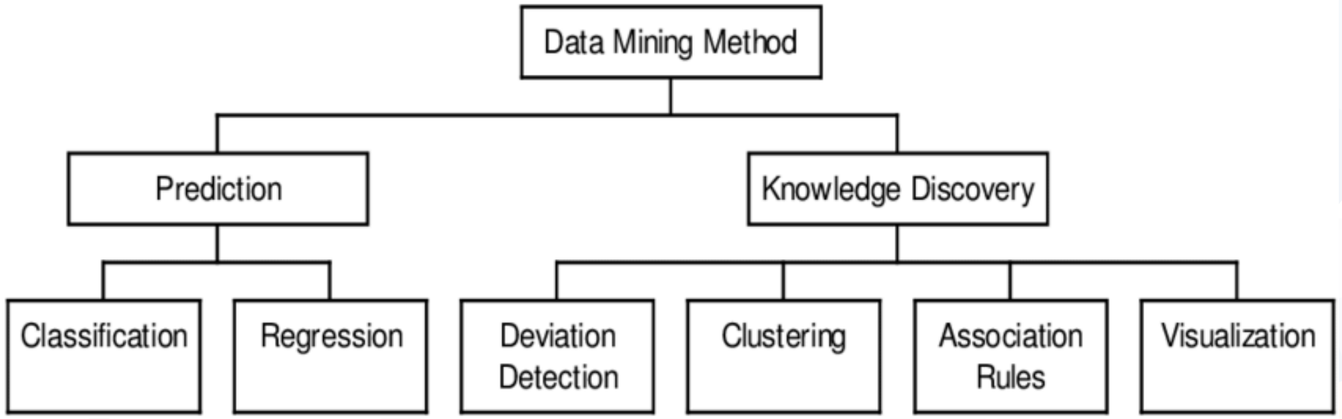
KDD and Data Mining

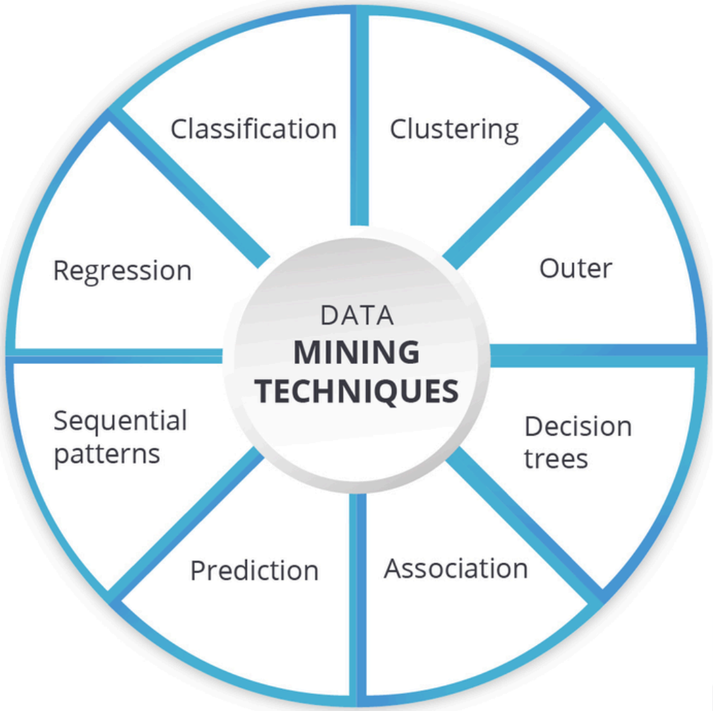
The term Knowledge Discovery in Databases, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization.

- The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.
- It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge.

Techniques or Types of Data Mining

Data mining uses algorithms and various other techniques to convert large collections of data into useful output. The most popular types of data mining techniques include





Techniques of Data Mining

Classification

This data analysis is implemented to regain vital and actual information. It's considered to be a complex data method among other data mining techniques. Information is classified into different classes. For instance, credit customers can be classified according to three risk categories: "low," "medium," or "high."

Clustering

Cluster analysis is a bit different classifying in the sense that here the pieces are grouped according to their similarities. For instance, different groups of customers are clustered together to find similarities and dissimilarities between the strands of information about them.

Techniques of Data Mining

Regression

This data mining tool is designed to pinpoint and analyze the interactions between different variables. It's used for identification of the probability of a particular variable from other variables' existence. This method is also known as predictive power. Regression techniques are rather advantageous, due to the power of neural networks which is a unique method that emulates the neural signals in the brain.

Association

This mining data technique is used to find an association between two or more events or properties. It drills down to an underlying model in the database systems. Somewhat similar to buying a laptop – you are immediately offered to buy a bag to go with it.

Techniques of Data Mining

Outer detection (Outlier analysis)

This a process of identifying certain anomalies (outliers) in the data set. You need to be able to explain why there are these outliers amidst the all-encompassing pattern. For example, among your male audience of buyers, you have a sudden peak in female buying activity.

Prediction

Prediction is considered to be an essential data mining technique. We all want to know the future value of our investments and to be protected from fraudulent crooks while online shopping. So it's applied to forecast different types of data mining in the days to come. Analysis of the previous events can help to project more or less accurate predictions tomorrow.

Techniques of Data Mining

Sequential patterns

This type of data analysis seeks to find out the same models, regularities or transaction tendencies in informational strands over a specified period. In sales, businesses can identify when some items are bought together during a particular season of the year. Based on this, companies offer better deals to those clients that have an actual purchasing history.

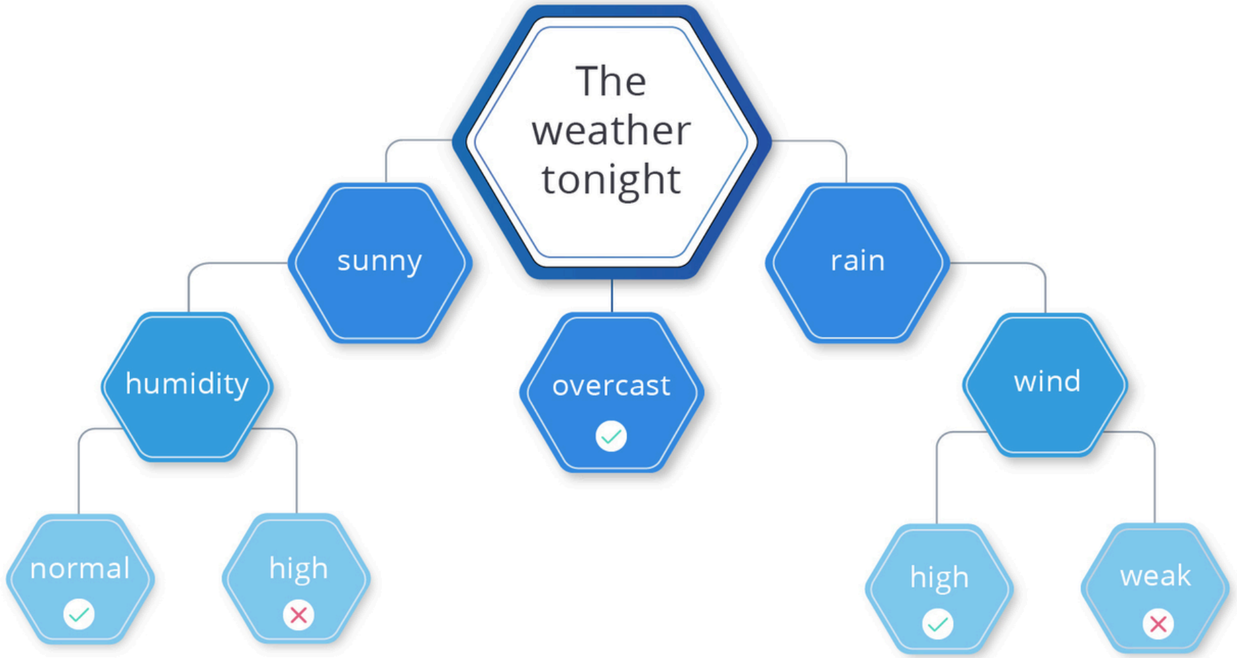
Decision trees

This type of data mining tool is used quite often as it's the simplest for understanding. At the root of such decision trees, there is a simple question with many possible answers. Based on the responses, we can get the final answer to the central question.

Techniques of Data Mining

Decision trees

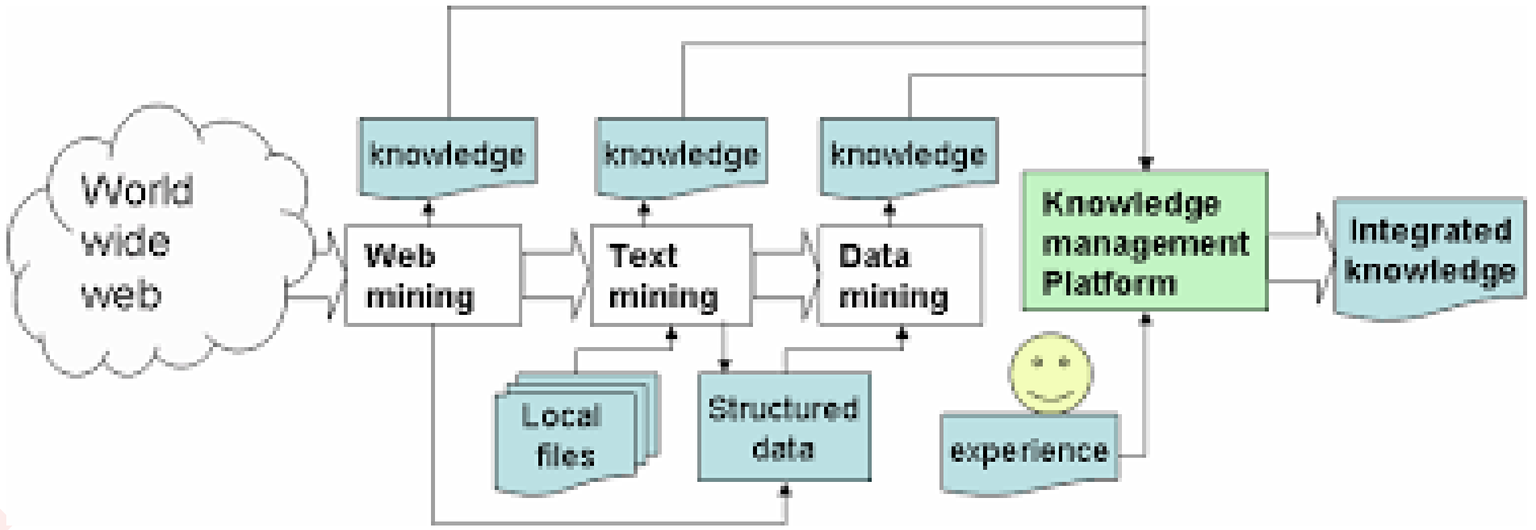
For example, we can attempt to respond to the following question: Should we play golf today?



Kicking off at the root box, if the weather forecast promises to be overcast, then might play golf today. If it's going to be rainy, but it's not raining yet, we could play provided it's not too windy. If the weather is sunny, we should play golf if the humidity isn't high. Such schematization helps to choose the best options among the good ones.

Data Mining and Web Mining

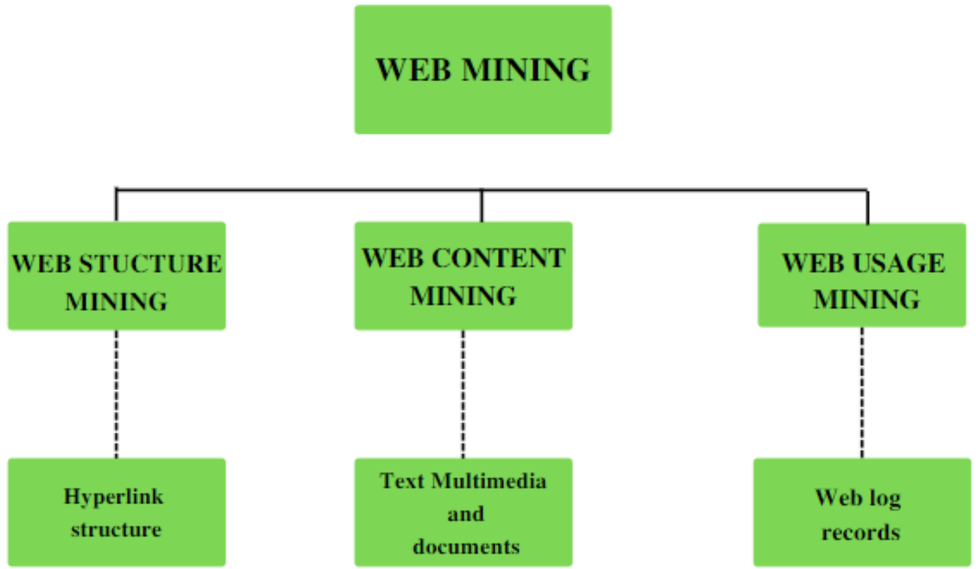
Web Mining is the process of Data Mining techniques to automatically discover and extract information from Web documents and services. The main purpose of web mining is to discover useful information from the World Wide Web and its usage patterns.



Process of Web Mining

Types of Web Mining

Web mining can be broadly divided into three different types of techniques of mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. These are explained as following below.



Types of Web Mining

Web Content Mining: Extracts valuable information from various types of web content such as text, images, audio, and video. It helps identify user needs and patterns, often involving techniques from text mining, machine learning, and natural language processing.

Web Structure Mining: Uncovers structural information from the web, focusing on the relationships between web pages through hyperlinks. It provides a structured summary of websites, revealing connections and associations between pages, which is useful for understanding website organization and navigation.

Web Usage Mining: Identifies usage patterns from large datasets, allowing insights into user behaviors and preferences. By analyzing user access logs, it helps understand how users interact with websites, aiding in website optimization and personalization efforts.

Application of Data Mining

Sales

• Data mining aids in efficient capital use for revenue growth.

• For instance, a coffee shop uses data to craft its product line.

Marketing

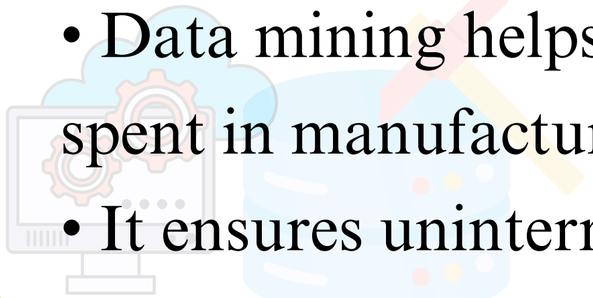
• Data mining helps in understanding client demographics, ad placement, and customer-resonating marketing strategies.

• Aligning marketing campaigns, promotional offers, cross-sell offers, and programs to data mining findings is crucial.

Manufacturing

• Data mining helps in analyzing raw material costs, efficient use of materials, and time spent in manufacturing processes.

• It ensures uninterrupted flow of goods.



Application of Data Mining

Fraud Detection

- Data mining identifies patterns, trends, and correlations linking data points.
- It can identify outliers or correlations in cash flow analysis.

Human Resources

- Data mining correlates data on retention, promotions, salary ranges, benefits, and employee satisfaction surveys.

Customer Service

- Data mining gathers operational information about customer interactions to pinpoint weak points and highlight strengths.

Application of Data Mining

Education

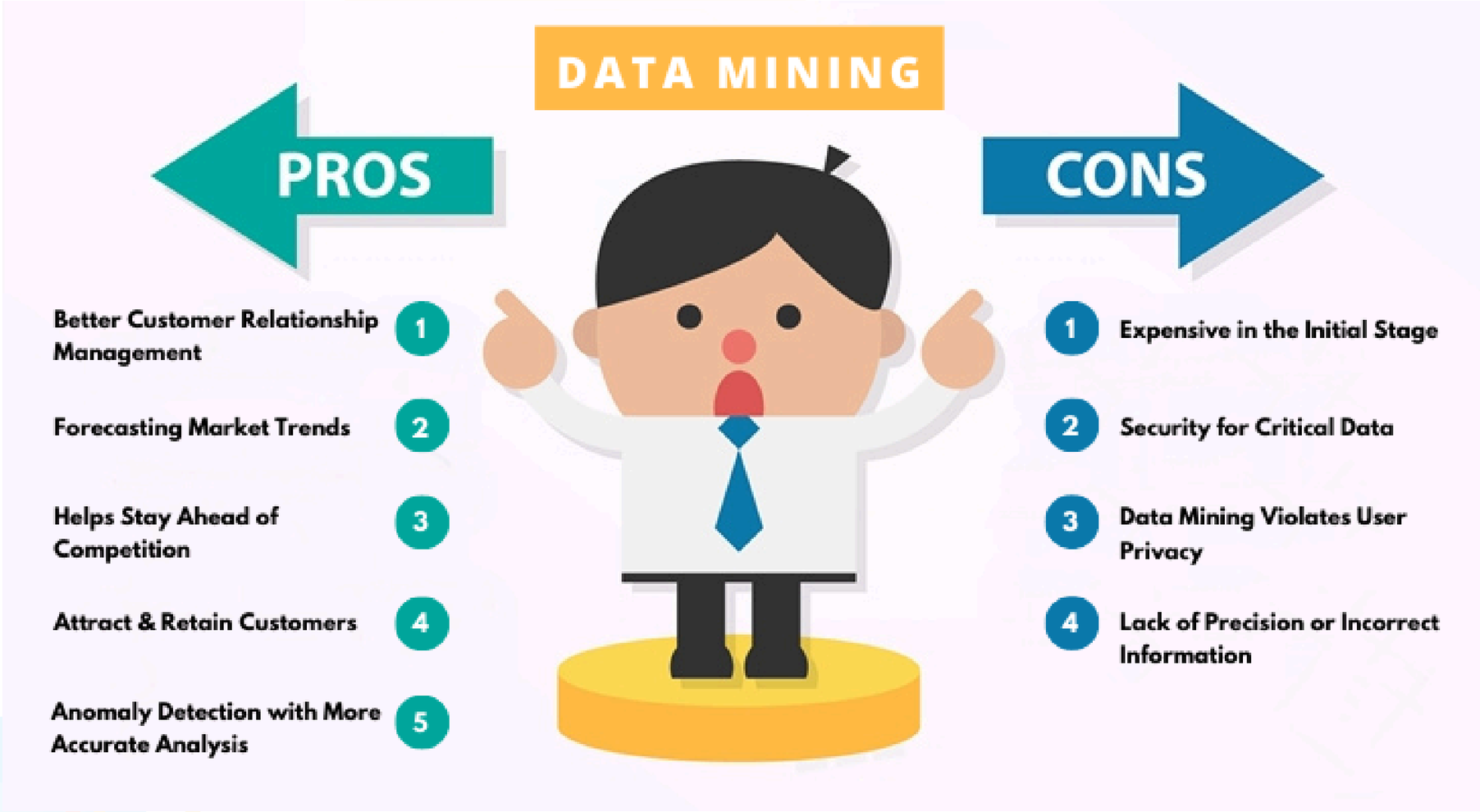
- By analyzing the progress of learning and the results of tests, it is possible to understand the subjects and contents that students are good at and weak at, and the degree of understanding of each student.
- It is also possible to group students according to their level of comprehension, provide guidance according to their level, and consider individual measures to improve their grades.

Application of Data Mining

Education

- By analyzing the progress of learning and the results of tests, it is possible to understand the subjects and contents that students are good at and weak at, and the degree of understanding of each student.
- It is also possible to group students according to their level of comprehension, provide guidance according to their level, and consider individual measures to improve their grades.

Advantages and Disadvantages of Data Mining



Difference Between Data Mining and Machine Learning

Here's a comparison table highlighting the differences between data mining and machine learning:

Aspect	Data Mining	Machine Learning
Definition	Process of finding patterns in data	Process of teaching a computer to learn
Goal	Identify patterns, trends, and insights	Make predictions or decisions based on data
Approach	Proactively identifies patterns through algorithms	Teaches computers to learn from data
Human Involvement	Interpretation and application of insights	Less frequent human involvement once trained
Type of Learning	Passive	Active
Examples	Market basket analysis, anomaly detection	Predictive modeling, classification, clustering
Use Cases	Business intelligence, marketing, fraud detection	Self-driving cars, recommendation systems


THANK
YOU