**Introduction to Data Mining**

Online analytical processing (OLAP) is an analysis technique with functionalities such as summarization, consolidation, and aggregation- and the ability to view information from different angles.

Online analytical processing (OLAP) is a software technology you can use to analyze business data from different points of view.

Although OLAP tools support multidimensional analysis and decision-making, additional data analysis tools are required for in-depth analysis.

For example, data mining tools that provide data classification, clustering, outlier/anomaly detection, and the characterization of changes in data over time.

**KDD versus Data Mining**

Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery.

The knowledge discovery process is an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
5. Data mining (an essential process where intelligent methods are applied to extract data patterns)
6. Pattern evaluation: to identify the truly interesting patterns representing knowledge based on interestingness measures
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Steps 1 through 4 are different forms of data preprocessing, where data are prepared for mining. The data mining step may interact with the user or a knowledge base.

The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base.

The preceding view shows data mining as one step in the knowledge discovery process, albeit an essential one because it uncovers hidden patterns for evaluation.

However, in industry, media, and the research milieu, the term data mining is often used to refer to the entire knowledge discovery process (perhaps because the term is shorter than knowledge discovery from data).

**Data Mining**

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data.

The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

*It is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making.*

From this definition, the important take aways are:

- Data mining is a process of automated discovery of previously unknown patterns in large volumes of data.
- This large volume of data is usually the historical data of an organization known as the data warehouse.
- Data mining deals with large volumes of data, in Gigabytes or Terabytes of data and sometimes as much as Zetabytes of data (in case of big data).
- Patterns must be valid, novel, useful and understandable.
- Data mining allows businesses to determine historical patterns to predict future behaviour.
- Although data mining is possible with smaller amounts of data, the bigger the data the better the accuracy in prediction.
- There is considerable hype about data mining at present, and the Gartner Group has listed data mining as one of the top ten technologies to watch.

### ***The need for data mining***

- Growth in generation and storage of corporate data
- Need for sophisticated decision making
- Evolution of technology
- Availability of much cheaper storage, easier data collection and better database management for data analysis and understanding
- Point of sale terminals and bar codes on many products, railway bookings, educational institutions, mobile phones, electronic gadgets, e-commerce, etc., all generate data.
- Great volumes of data generated with the recent prevalence of Internet banking, ATMs, credit and debit cards; medical data, hospitals; automatic toll collection on toll roads, growing air travel; passports, visas, etc.
- Decline in the costs of hard drives
- Growth in worldwide disk capacities

### **What Can Data Mining Do and Not Do?**

- Data mining is a powerful tool that helps to determine the relationships and patterns within data.
- However, it does not work by itself and does not eliminate the requirement for understanding data, analytical methods and to know business.
- Data mining extracts hidden information from the data, but it is not able to assess the value of the information.
- One should know the important patterns and relationships to work with data over time.

- In addition to discovering new patterns, data mining can also highlight other empirical observations that are not instantly visible through simple observation.
- It is important to note that the relationships or patterns predicted through data mining are not necessarily causes for an action or behavior.

**What Kinds of Data Can Be Mined?**

The most basic forms of data for mining applications are:

- Database data,
- Data warehouse data, and
- Transactional data
- Time-related or sequence data (e.g., historical records, stock exchange data, and time-series and biological sequence data)
- Data streams (e.g., video surveillance and sensor data, which are continuously transmitted),
- Spatial data (e.g., maps),
- Engineering design data (e.g., the design of buildings, system components, or integrated circuits),
- Hypertext and multimedia data (including text, image, video, and audio data),
- Graph and networked data (e.g., social and information networks), and
- The Web (a huge, widely distributed information repository made available by the Internet).

**Data Mining Applications**

The applications of data mining exist in almost every field.

**Loan/Credit card approvals**

- Banks are able to assess the credit worthiness of their customers by mining a customer's historical records of business transactions.
- So, credit agencies and banks collect a lot of customer' behavioural data from many sources.
- This information is used to predict the chances of a customer paying back a loan.

**Market segmentation**

- A huge amount of data about customers is available from purchase records.
- This data is very useful to segment customers on the basis of their purchase history.
- Let us suppose a mega store sells multiple items ranging from grocery to books, electronics, clothing et al.
- That store now plans to launch a sale on books. Instead of sending SMS texts to all the customers it is logical and more efficient to send the SMS only to those customers who have demonstrated interest in buying books, i.e., those who had earlier purchased books from the store.
- In this case, the segmentation of customers based on their historical purchases will help to send the message to those who may find it relevant.
- It will also give a list of people who are prospects for the product.

**Fraud detection**

- Fraud detection is a very challenging task because it's difficult to define the characteristics for detection of fraud.
- Fraud detection can be performed by analyzing the patterns or relationships that deviate from an expected norm.
- With the help of data mining, we can mine the data of an organization to know about outliers as they may be possible locations for fraud.

**Better marketing**

- Usually all online sale web portals provide recommendations to their users based on their previous purchase choices, and purchases made by customers of similar profile.
- Such recommendations are generated through data mining and help to achieve more sales with better marketing of their products.
- For example, amazon.com uses associations and provides recommendations to customers on the basis of past purchases and what other customers are purchasing.
- To take another example, a shoe store can use data mining to identify the right shoes to stock in the right store on the basis of shoe sizes of the customers in the region.

**Trend analysis**

- In a large company, not all trends are always visible to the management.
- It is then useful to use data mining software that will help identify trends.
- Trends may be long term trends, cyclic trends or seasonal trends.

**Market basket analysis**

- Market basket analysis is useful in designing store layouts or in deciding which items to put on sale.
- It aims to find what the customers buy and what they buy together.

**Customer churn**

- If an organization knows its customers better then it can retain them longer.
- In businesses like telecommunications, companies very hard to retain their good customers and to perhaps persuade good customers of their competitors to switch to them.
- In such an environment, businesses want to rate customers (whether they are worthwhile to be retained), why customers switch over, and what makes customers loyal.
- With the help of this information some businesses may wish to get rid of customers that cost more than they are worth, e.g., credit card holders that don't use the card; bank customers with very small amounts of money in their accounts.

**Website design**

- A web site is effective only if the visitors easily find what they are looking for.
- Data mining can help discover the affinity of visitors to pages and the site layout may be modified based on data generated from their web clicks.

**Corporate analysis and risk management**

- Data mining can be used to perform cash flow analysis to plan finance and estimate assets.
- The analysis of data can be performed by comparing and summarizing the spending and planning of resources.
- It helps to analyze market directions and monitor competitors to analyze the competition.

**Data Mining Process**

Data mining process consists of six phases

**Problem definition phase**

- The main focus of the first phase of a data mining process is to understand the requirements and objectives of such a project.
- Once the project has been specified, it can be formulated as a data mining problem.
- After this, a preliminary implementation plan can be developed.
- Let us consider a business problem such as 'How can I sell more of my product to customers?'
- This business problem can be translated into a data mining problem such as 'Which customers are most likely to buy the product?'
- A model that predicts the frequent customers of a product must be built on the previous records of customers' data.
- Before building the model, the data must be assembled that consists of relationships between customers who have purchased the product and customers who have not purchased the product.
- The attributes of customers might include age, years of residence, number of children, owners/renters, and so on.

**Data understanding phase**

- The next phase of the data mining process starts with the data collection.
- In this phase, data is collected from the available sources
- In order to make data collection proper, some important activities such as data loading and data integration are performed.
- After this, the data is analyzed closely to determine whether the data will address the business problem or not.
- Therefore, additional data can be added or removed to solve the problem effectively.
- At this stage missing data is also identified.
- For example, if we require the AGE attribute for a record then column DATE_OF_ BIRTH can be changed to AGE. We can also consider another example in which average income can be inserted if the value of column INCOME is null.
- Moreover, new computed attributes can be added in the data in order to obtain better focused information. For example, a new attribute such as 'Number of Times Amount Purchased Exceeds Kshs. 5000 in a 12-month time period.' can be created instead of using the purchase amount.

**Data preparation phase**

- This phase generally consumes about 90% of the time of a project.
- Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form for further processing.

**Modeling phase**

- In this phase, different data mining algorithms are applied to build models.
- Appropriate data mining algorithms are selected and applied on given data to achieve the objectives of proposed solution.

**Evaluation phase**

- In the evaluation phase, the model results are evaluated to determine whether it satisfies the originally stated business goal or not.
- For this the given data is divided into training and testing datasets.
- The models are trained on training data and tested on testing data.
- If the accuracy of models on testing data is not adequate then one goes back to the previous phases to fine tune those areas that may be the reasons for low accuracy.
- Having achieved a satisfactory level of accuracy, the process shifts to the deployment phase

**Deployment phase**

- In the deployment phase, insights and valuable information derived from data need to be presented in such a way that stakeholders can use it when they want to.
- On the basis of requirements of the project, the deployment phase can be simple (just creating a report) or complex (requiring further iterative data mining processing).
- In this phase, Dashboards or Graphical User Interfaces are built to solve all the requirements of stakeholders.

**Data Mining Techniques**

Data mining can be classified into four major techniques as given below.

- Predictive modeling
- Database segmentation
- Link analysis
- Deviation detection

**Predictive modeling**

- Predictive modeling is based on predicting the outcome of an event.
- It is designed on a pattern similar to the human learning experience in using observations to form a model of the important characteristics of some task.
- It is developed using a supervised learning approach, where we have some labeled data and we use this data to predict the outcome of unknown instances.
- It can be of two types, i.e., classification or regression.

- Some of the applications of predictive modeling are: predicting the outcome of an event, predicting the sale price of a property, predicting placement of students, predicting the score of any team during a cricket match and so on.

**Database segmentation**

- Database segmentation is based on the concept of clustering of data and it falls under unsupervised learning, where data is not labeled.
- This data is segmented into groups or clusters based on its features or attributes.
- Segmentation is creating a group of similar records that share a number of properties.
- Applications of database segmentation include customer segmentation, customer churn, direct marketing, and cross-selling.

**Link analysis**

Link analysis aims to establish links, called associations, between the individual record, or sets of records, in a database. There are three specializations of link analysis.

- Associations discovery
- Sequential pattern discovery
- Similar time sequence discovery

- **Associations discovery** locates items that imply the presence of other items in the same event. There are association rules which are used to define association.
    - For example, ‘when a customer rents property for more than two years and is more than 25 years old, in 40% of cases, the customer will buy a property. This association happens in 35% of all customers who rent properties.
- **Sequential pattern discovery** finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time. For example, this approach can be used to understand long-term customer buying behavior.
- **Time sequence discovery** is used to determine whether links exist between two sets of data that are time-dependent.
    - For example, within three months of buying property, new homeowners will purchase goods such as cookers, freezers, and washing machines.
    - Applications of link analysis include market basket analysis, recommendation systems, direct marketing, and stock price movement.

**Deviation detection**

- Deviation detection is a relatively new technique in terms of commercially available data mining tools.
- It is based on identifying the outliers in the database, which indicates deviation from some previously known expectations and norms.
- This operation can be performed using statistics and visualization techniques.
- Applications of deviation detection include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing