

Data Mining and Analytics(18CSE355T)

Unit-1

Data Mining

Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called ***Knowledge Discovery in Database (KDD)***. The knowledge discovery process includes Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation. The process of extracting information to identify patterns, trends, and useful data that would allow the business to take the data-driven decision from huge sets of data is called Data Mining.

In other words, we can say that Data Mining is the process of investigating hidden patterns of information to various perspectives for categorization into useful data, which is collected and assembled in particular areas such as data warehouses, efficient analysis, data mining algorithm, helping decision making and other data requirement to eventually cost-cutting and generating revenue. Data mining is the act of automatically searching for large stores of information to find trends and patterns that go beyond simple analysis procedures. Data mining utilizes complex mathematical algorithms for data segments and evaluates the probability of future events. Data Mining is also called Knowledge Discovery of Data (KDD). Data Mining is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Data Analysis vs. Data Analytics

- The main difference between data analysis and analytics lies in their approach, as analysis looks towards the past while analytics towards the future. As per any English dictionary, Analysis is the division of a whole into small components and analytics is the science of logical analysis.

- While analysis looks backward over time and works on the facts and figures of what has happened, analytics work towards modeling the future or predicting a result.
- In other words, the analysis restructures existing available information or data. And, the analytics uses this analyzed information to predict what may happen. Let's take an example of an apparel brand. The business/brand owner analyzes last year's sales data to gain insight into profit trends and sales trends as per seasons, months, and weeks. This analysis of what has happened is basically an in-depth review of the past facts. Whereas, analytics combines the results from the analysis of last year's sales data with logical reasoning to predict future sales pattern and design and plan accordingly. In practice, this means the brand will employ advanced (machine learning) tools and algorithms to make the best use of the historical review and predict future sales patterns. With analytics, the apparel brand can design its plan of when to launch new products over the coming weeks and months to get the maximum profit.
- Data analysis is the process of studying a given data set (in close detail), dividing them into small components, and studying the subcomponents individually and their relationship with each another. Data analytics, on the other hand, is a more comprehensive term referring to a discipline that comprises the complete management of data, including collection, cleaning, organizing, storing, administering, and analysis of data with the help of specialized tools and techniques. In other words, data analysis is a process or method, whereas data analytics is an overarching discipline (science).
- It is apparent by the definition itself, that data analytics is a broader term and comprises data analysis as a necessary subcomponent. It is the science or the cognitive process that an analyst uses to recognize problems and examine data in the most meaningful ways. Both analysis and analytics are highly significant and help businesses to estimate customers accurately, approach the right audience and get the best results using their marketing budget. Both of these help businesses explore and analyze the customer data to understand unknown patterns, grab opportunities and gain insights and transform that into productive decision making.

- Data analysis is a process involving the collection, manipulation, and examination of data for getting a deep insight. Data analytics is taking the analyzed data and working on it in a meaningful and useful way to make well-versed business decisions.
- Data analysis helps design a strong business plan for businesses, using its historical data that tell about what worked, what did not and what was expected from a product or service. Data analytics helps businesses in utilizing the potential of the past data and in turn identifying new opportunities that would help them plan future strategies. It helps in business growth by reducing risks, costs, and making the right decisions.
- Tools used for data analysis are Open Refine, Rapid Miner, KNIME, Google Fusion Tables, Node XL, Wolfram Alpha, Tableau Public, etc. Tools used in Data analytics are Python, Tableau Public, SAS, Apache Spark, Excel, etc.
- Data analytics is more extensive in its scope and encompasses data analysis as a sub-component. The life cycle of data analytics also comprises data analysis as one of the significant steps.
- Through data analytics and data analysis, both are essential to understand the data as the first one is useful in estimating future demands and the second one is necessary for gaining insight by analyzing the details of the past data. Data analysis is actually studying past data to understand ‘what happened?’ Whereas data analytics predicts ‘what will happen next or what is going to be next?’

Data Mining vs. Machine Learning

Data Mining relates to extracting information from a large quantity of data. Data mining is a technique of discovering different kinds of patterns that are inherited in the data set and which are precise, new and useful data. Data Mining is working as a subset of business analytics and similar to experimental studies. Data Mining's origins are databases, statistics.

Machine learning includes an algorithm that automatically improves through data-based experience. Machine learning is a way to find a new algorithm from experience. Machine

learning includes the study of an algorithm that can automatically extract the data. Machine learning utilizes data mining techniques and another learning algorithm to construct models of what is happening behind certain information so that it can predict future results.

What is Data Mining?

Data Mining is the method of extraction of data or previously unknown data patterns from huge sets of data. Hence as the word suggests, we 'Mine for specific data' from the large data set. Data mining is also called Knowledge Discovery Process, is a field of science that is used to determine the properties of the datasets. **Gregory Piatetsky-Shapiro** founded the term "**Knowledge Discovery in Databases**" (KDD) in 1989. The term "data mining" came in the database community in 1990. Huge sets of data collected from data warehouses or complex datasets such as time series, spatial, etc. are extracted in order to extract interesting correlations and patterns between the data items. For Machine Learning algorithms, the output of the data mining algorithm is often used as input.

What is Machine learning?

Machine learning is related to the development and designing of a machine that can learn itself from a specified set of data to obtain a desirable result without it being explicitly coded. Hence Machine learning implies 'a machine which learns on its own. **Arthur Samuel** invented the term Machine learning an American pioneer in the area of *computer gaming* and *artificial intelligence* in 1959. He said that "**it gives computers the ability to learn without being explicitly programmed.**"

Machine learning is a technique that creates complex algorithms for large data processing and provides outcomes to its users. It utilizes complex programs that can learn through experience and make predictions.

The algorithms are enhanced by themselves by frequent input of training data. The aim of machine learning is to understand information and build models from data that can be understood and used by humans.

Machine learning algorithms are divided into two types:

- Unsupervised Learning
- Supervised Learning

1. Unsupervised Machine Learning:

Unsupervised learning does not depend on trained data sets to predict the results, but it utilizes direct techniques such as clustering and association in order to predict the results. Trained data sets are defined as the input for which the output is known.

2. Supervised Machine Learning:

As the name implies, supervised learning refers to the presence of a supervisor as a teacher. Supervised learning is a learning process in which we teach or train the machine using data which is well leveled implies that some data is already marked with the correct responses. After that, the machine is provided with the new sets of data so that the supervised learning algorithm analyzes the training data and gives an accurate result from labeled data.

Major Difference between Data mining and Machine learning

1. Two-component is used to introduce data mining techniques first one is the database, and the second one is machine learning. The database provides data management techniques, while machine learning provides methods for data analysis. But to introduce machine learning methods, it used algorithms.

2. Data Mining utilizes more data to obtain helpful information, and that specific data will help to predict some future results. For example, In a marketing company that utilizes last year's data to predict the sale, but machine learning does not depend much on data. It uses algorithms. Many transportation companies such as OLA, UBER machine learning techniques to calculate ETA (Estimated Time of Arrival) for rides is based on this technique.

3. Data mining is not capable of self-learning. It follows the guidelines that are predefined. It will provide the answer to a specific problem, but machine learning algorithms are self-defined and

can alter their rules according to the situation and find out the solution for a specific problem and resolve it in its way.

4. The main and most important difference between data mining and machine learning is that without the involvement of humans, data mining can't work, but in the case of machine learning human effort only involves at the time when the algorithm is defined after that it will conclude everything on its own. Once it implemented, we can use it forever, but this is not possible in the case of data mining.

5. As machine learning is an automated process, the result produces by machine learning will be more precise as compared to data mining.

6. Data mining utilizes the database; data warehouse server, data mining engine, and pattern assessment techniques to obtain useful information, whereas machine learning utilizes neural networks, predictive models, and automated algorithms to make the decisions.

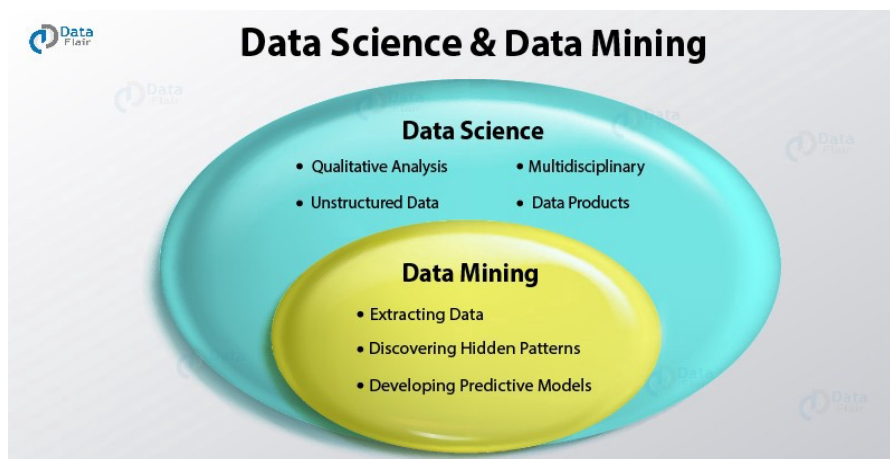
Data Mining Vs Machine Learning

Factors	Data Mining	Machine Learning
Origin	Traditional databases with unstructured data.	It has an existing algorithm and data.
Meaning	Extracting information from a huge amount of data.	Introduce new Information from data as well as previous experience.
History	In 1930, it was known as knowledge discovery in databases (KDD).	The first program, i.e., Samuel's checker playing program, was established in 1950.
Responsibility	Data Mining is used to obtain the rules from the existing data.	Machine learning teaches the computer, how to learn and comprehend the rules.
Abstraction	Data mining abstract from the data warehouse.	Machine learning reads machine.

Applications	In compare to machine learning, data mining can produce outcomes on the lesser volume of data. It is also used in cluster analysis.	It needs a large amount of data to obtain accurate results. It has various applications, used in web search, spam filter, credit scoring, computer design, etc.
Nature	It involves human interference more towards the manual.	It is automated, once designed and implemented; there is no need for human effort.
Techniques involve	Data mining is more of research using a technique like machine learning.	It is a self-learned and train system to do the task precisely.
Scope	Applied in the limited fields.	It can be used in a vast area.

Data Mining vs. Data Science

Data Mining is about finding the trends in a data set and using these trends to identify future patterns. It is an important step in the Knowledge Discovery process. It often includes analyzing the vast amount of historical data which was previously ignored. Data Science is a field of study which includes everything from Big Data Analytics, Data Mining, Predictive Modeling, Data Visualization, Mathematics, and Statistics. Data Science has been referred to as the fourth paradigm of Science. (the other three being Theoretical, Empirical and Computational). Academia often conducts exclusive research in Data Science.



Data Science vs. Data Mining Comparison Table

Basis for comparison	Data Mining	Data Science
What is it?	A technique	An area
Focus	Business process	Scientific study
Goal	Make data more usable	Building Data-centric products for an organization
Output	Patterns	Varied
Purpose	Finding trends previously not known	Social analysis, building predictive models, unearthing unknown facts, and more
Vocational Perspective	Someone with a knowledge of navigating across data and statistical understanding can conduct data mining	A person needs to understand Machine Learning, Programming, info-graphic techniques and have the domain knowledge to become a data scientist
Extent	Data mining can be a subset of Data Science as Mining activities are part of the Data Science pipeline	Multidisciplinary – Data Science consists of Data Visualizations, Computational Social Sciences, Statistics, Data Mining, Natural Language Processing, et cetera

Deals with (the type of data)	Mostly structured	All forms of data – structured, semi-structured and unstructured
Other less popular names	Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction	Data-driven Science

Example

Consider a scenario where you are a major retailer in India. You have 50 stores operating in 10 major cities in India and you have been operational for 10 years.

Let's say, you want to study the last 8 years' data to find the number of sales of sweets during festive seasons of 3 cities. If that's your objective, I would recommend you employ a person with Data Mining expertise. A Data Miner would probably go through historical information stored in legacy systems and employ algorithms to extract trends.

Consider another case where you want to know which sweets have received more positive reviews. In this case, your sources of data may not be limited to databases; they could extend to social websites or customer feedback messages. In this case, my suggestion to you would be to employ a Data Scientist. A person employed as a Data Scientist is more suited to apply algorithms and conduct this socio-computational analysis.

Below is the key difference between data science and data mining.

- Data Mining is an activity which is a part of a broader Knowledge Discovery in Databases (KDD) Process while Data Science is a field of study just like Applied Mathematics or Computer Science.
- Often Data Science is looked upon in a broad sense while Data Mining is considered a niche.

- Some activities under Data Mining such as statistical analysis, writing data flows and pattern recognition can intersect with Data Science. Hence, Data Mining becomes a subset of Data Science.
- Machine Learning in Data Mining is used more in pattern recognition while in Data Science it has a more general use.
- Data Science and Data Mining should not be confused with Big Data Analytics and one can have both Miners and Scientists working on big datasets.

Evolution/History of Data Mining

In the **1990s**, the term "Data Mining" was introduced, but data mining is the evolution of a sector with an extensive history. Early techniques of identifying patterns in data include Bayes theorem (**1700s**) and the evolution of regression (**1800s**). The generation and growing power of computer science have boosted data collection, storage, and manipulation as data sets have broad in size and complexity level. Explicit hands-on data investigation has progressively been improved with indirect, automatic data processing, and other computer science discoveries such as neural networks, clustering, genetic algorithms (**1950s**), decision trees (**1960s**), and supporting vector machines (**1990s**).

Data mining origins are traced back to three family lines: Classical statistics, Artificial intelligence, and Machine learning.

- **Classical statistics:**

Statistics are the basis of most technology on which data mining is built, such as regression analysis, standard deviation, standard distribution, standard variance, discriminatory analysis, cluster analysis, and confidence intervals. All of these are used to analyze data and data connection.

- **Artificial Intelligence:**

AI or Artificial intelligence is based on heuristics as opposed to statistics. It tries to apply human- thought like processing to statistical problems. A specific AI concept was

adopted by some high-end commercial products, such as query optimization modules for **Relational Database Management System (RDBMS)**.

- **Machine Learning:**

Machine learning is a combination of statistics and AI. It might be considered as an evolution of AI because it mixes AI heuristics with complex statistical analysis. Machine learning tries to enable computer programs to know about the data they are studying so that programs make a distinct decision based on the characteristics of the data examined. It uses statistics for basic concepts and adding more AI heuristics and algorithms to accomplish its target.

Why we need Data Mining?

Volume of information is increasing everyday that we can handle from business transactions, scientific data, sensor data, Pictures, videos, etc. So, we need a system that will be capable of extracting essence of information available and that can automatically generate report, views or summary of data for better decision-making.

Why Data Mining is used in Business?

Data mining is used in business to make better managerial decisions by:

- Automatic summarization of data
- Extracting essence of information stored.
- Discovering patterns in raw data.

Advantages of Data Mining

- The Data Mining technique enables organizations to obtain knowledge-based data.
- Data mining enables organizations to make lucrative modifications in operation and production.
- Compared with other statistical data applications, data mining is a cost-efficient.
- Data Mining helps the decision-making process of an organization.

- It facilitates the automated discovery of hidden patterns as well as the prediction of trends and behaviors.
- It can be induced in the new system as well as the existing platforms.
- It is a quick process that makes it easy for new users to analyze enormous amounts of data in a short time.

Disadvantages of Data Mining

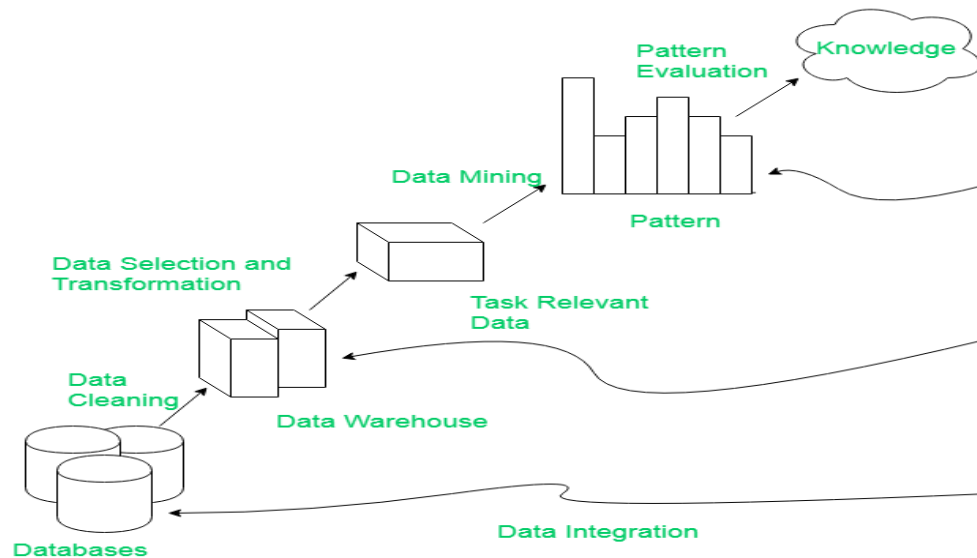
- There is a probability that the organizations may sell useful data of customers to other organizations for money. As per the report, American Express has sold credit card purchases of their customers to other organizations.
- Many data mining analytics software is difficult to operate and needs advance training to work on.
- Different data mining instruments operate in distinct ways due to the different algorithms used in their design. Therefore, the selection of the right data mining tools is a very challenging task.
- The data mining techniques are not precise, so that it may lead to severe consequences in certain conditions.

KDD (Knowledge Discovery from Data)

Data Mining refers to extracting or mining knowledge from a large amount of data. Many people treat Data Mining as a synonym for another popular used term Knowledge Discovery from Data (KDD). Data Mining is the process of discovering interesting knowledge from large amount of data stored in databases, data warehouses or other information repositories.

Data Mining also known as Knowledge Discovery in Databases, refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data stored in databases.

Steps Involved in KDD Process:



Data Cleaning: Data cleaning is defined as removal of noisy, inconsistent and irrelevant data from collection.

- Cleaning in case of **Missing values**.
- Cleaning **noisy** data, where noise is a random or variance error.
- Cleaning with **Data discrepancy detection** and **Data transformation tools**.

Data Integration: Data integration is defined as heterogeneous data from multiple sources combined in a common source (Data Warehouse).

- Data integration using **Data Migration tools**.
- Data integration using **Data Synchronization tools**.
- Data integration using **ETL** (Extract-Load-Transformation) process.

Data Selection: Data selection is defined as the process where data relevant to the analysis is decided and retrieved from the data collection.

- Data selection using **Neural network**.
- Data selection using **Decision Trees**.
- Data selection using **Naive bayes**.
- Data selection using **Clustering, Regression**, etc.

Data Transformation: Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure. Data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.

Data Transformation is a two step process:

- **Data Mapping:** Assigning elements from source base to destination to capture transformations.
- **Code generation:** Creation of the actual transformation program.

Data Mining: Data mining is defined as clever techniques that are applied to extract patterns potentially useful. It is an essential process where intelligent methods are applied in order to extract data pattern.

- Transforms task relevant data into **patterns**.
- Decides purpose of model using **classification** or **characterization**.

Pattern Evaluation: Pattern Evaluation is defined as identifying strictly increasing patterns representing knowledge based on given measures.

- Find **interestingness score** of each pattern.
- Uses **summarization** and **Visualization** to make data understandable by user.

Knowledge representation: Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.

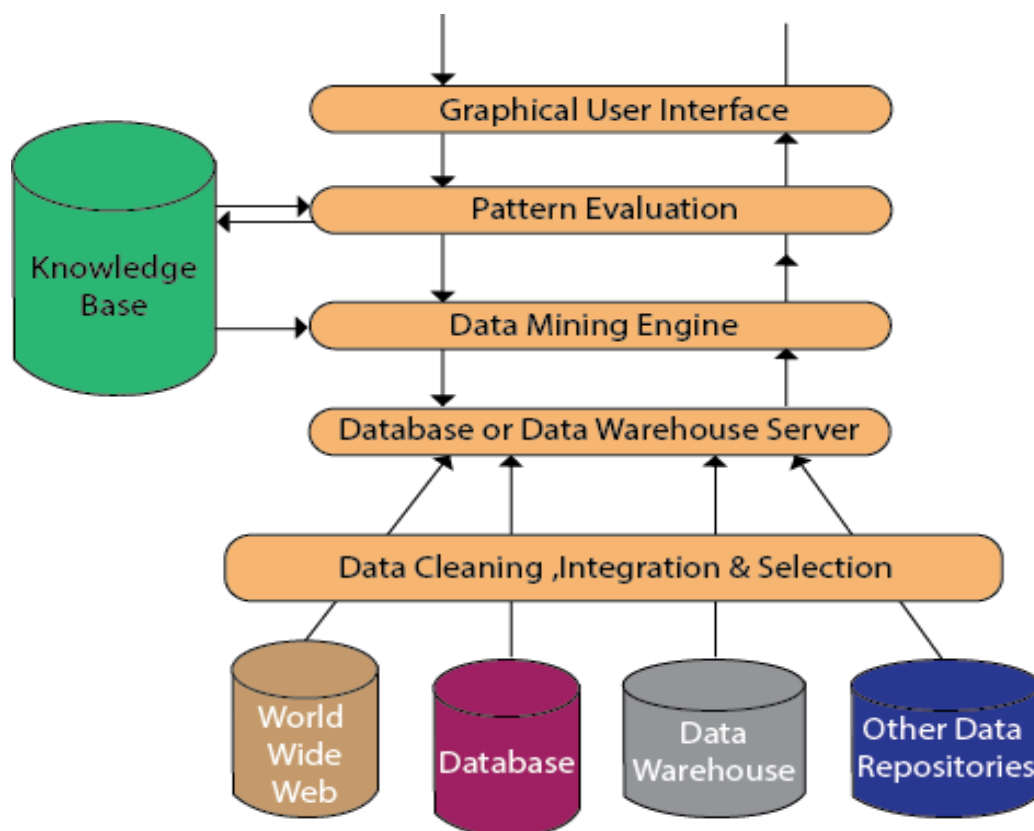
- Generate **reports**.
- Generate **tables**.
- Generate **discriminant rules, classification rules, characterization rules**, etc.

Note:

- KDD is an **iterative process** where evaluation measures can be enhanced, mining can be refined, new data can be integrated and transformed in order to get different and more appropriate results.
- **Preprocessing of databases** consists of **Data cleaning** and **Data Integration**.

Data Mining Architecture

Data mining is a significant method where previously unknown and potentially useful information is extracted from the vast amount of data. The data mining process involves several components and these components constitute a data mining system architecture. The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface and knowledge base.



Data Source:

The actual source of data is the Database, data warehouse, World Wide Web (WWW), text files, and other documents. You need a huge amount of historical data for data mining to be successful. Organizations typically store data in databases or data warehouses. Data warehouses may comprise one or more databases; text files spreadsheets, or other repositories of data. Sometimes, even plain text files or spreadsheets may contain information. Another primary source of data is the World Wide Web or the internet.

Different processes:

Before passing the data to the database or data warehouse server, the data must be cleaned, integrated and selected. As the information comes from various sources and in different formats, it can't be used directly for the data mining procedure because the data may not be complete and accurate. So, the first data requires to be cleaned and unified. More information than needed will be collected from various data sources and only the data of interest will have to be selected and passed to the server. Several methods may be performed on the data as part of selection, integration, and cleaning.

Database or Data Warehouse Server:

The database or data warehouse server consists of the original data that is ready to be processed. Hence, the server is cause for retrieving the relevant data that is based on data mining as per user request.

Data Mining Engine:

The data mining engine is a major component of any data mining system. It contains several modules for operating data mining tasks, including association, characterization, classification, clustering, prediction, time-series analysis, etc.

In other words, we can say data mining is the root of our data mining architecture. It comprises instruments and software used to obtain insights and knowledge from data collected from various data sources and stored within the data warehouse.

Pattern Evaluation Module:

The Pattern evaluation module is primarily responsible for the measure of investigation of the pattern by using a threshold value. It collaborates with the data mining engine to focus the search on exciting patterns.

This segment commonly employs stake measures that cooperate with the data mining modules to focus the search towards fascinating patterns. It might utilize a stake threshold to filter out discovered patterns. On the other hand, the pattern evaluation module might be coordinated with the mining module, depending on the implementation of the data mining techniques used. For efficient data mining, it is abnormally suggested to push the evaluation of pattern stake as much as possible into the mining procedure to confine the search to only fascinating patterns.

Graphical User Interface:

The graphical user interface (GUI) module communicates between the data mining system and the user. This module helps the user to easily and efficiently use the system without knowing the complexity of the process. This module cooperates with the data mining system when the user specifies a query or a task and displays the results.

Knowledge Base:

The knowledge base is helpful in the entire process of data mining. It might be helpful to guide the search or evaluate the stake of the result patterns. The knowledge base may even contain user views and data from user experiences that might be helpful in the data mining process. The data mining engine may receive inputs from the knowledge base to make the result more accurate and reliable. The pattern assessment module regularly interacts with the knowledge base to get inputs, and also update it.

Kinds of Data meant for mining

Data mining can be performed on the following types of data:

1. Flat Files: Flat files is defined as data files in text form or binary form with a structure that can be easily extracted by data mining algorithms. Data stored in flat files have no relationship or path among themselves, like if a relational database is stored on flat file, then there will be no relations between the tables. Flat files are represented by data dictionary. Eg: CSV file. Another example of a flat file is a name-and-address list with the fields Name, Address, and Phone Number. A list of names, addresses, and phone numbers written by hand on a sheet of paper is a flat-file database. Databases created in spreadsheet applications (like Microsoft Excel) are **flat file databases**. An old fashioned example of a flat file or two-dimensional database is the old printed telephone directory. **Application:** Used in DataWarehousing to store data, Used in carrying data to and from server, etc.

2. Relational Databases: A relational database is a collection of multiple data sets formally organized by tables, records, and columns from which data can be accessed in various ways without having to recognize the database tables. Tables convey and share information, which facilitates data searchability, reporting, and organization. Physical schema in Relational databases is a schema which defines the structure of tables. Logical schema in Relational databases is a schema which defines the relationship among tables. Standard API of relational database is SQL. **Application:** Data Mining, ROLAP model, etc.

3. Data Warehouse: A Data Warehouse is the technology that collects the data from various sources within the organization to provide meaningful business insights. The huge amount of data comes from multiple places such as Marketing and Finance. The extracted data is utilized for analytical purposes and helps in decision- making for a business organization. The data warehouse is designed for the analysis of data rather than transaction processing. There are three types of data warehouse: **Enterprise** data warehouse, **Data Mart** and **Virtual Warehouse**. Two approaches can be used to update data in Data Warehouse: **Query-driven** Approach and **Update-driven** Approach. **Application:** Business decision making, Data mining, etc.

4. Object-Relational Database:

A combination of an object-oriented database model and relational database model is called an object-relational model. It supports Classes, Objects, Inheritance, etc.

One of the primary objectives of the Object-relational data model is to close the gap between the Relational database and the object-oriented model practices frequently utilized in many programming languages, for example, C++, Java, C#, and so on.

5. Transactional Databases: A transactional database refers to a database management system (DBMS) that has the potential to undo a database transaction if it is not performed appropriately. Even though this was a unique capability a very long while back, today, most of the relational database systems support transactional database activities. Transactional databases are a collection of data organized by time stamps, date, etc to represent transaction in databases. This type of database has the capability to roll back or undo its operation when a transaction is not completed or committed. Highly flexible system where users can modify information without changing any sensitive information. It follows ACID property of DBMS. **Application:** Banking, Distributed systems, Object databases, etc.

6. Multimedia Databases: Multimedia databases consists audio, video, images and text media. They can be stored on Object-Oriented Databases. They are used to store complex information in pre-specified formats. **Application:** Digital libraries, video-on demand, news-on demand, musical database, etc.

7. Spatial Database: Spatial databases are databases that, in addition to usual data, store geographical information like maps, and global or regional positioning. It stores data in the form of coordinates, topology, lines, polygons, etc. **Application:** Maps, Global positioning, etc.

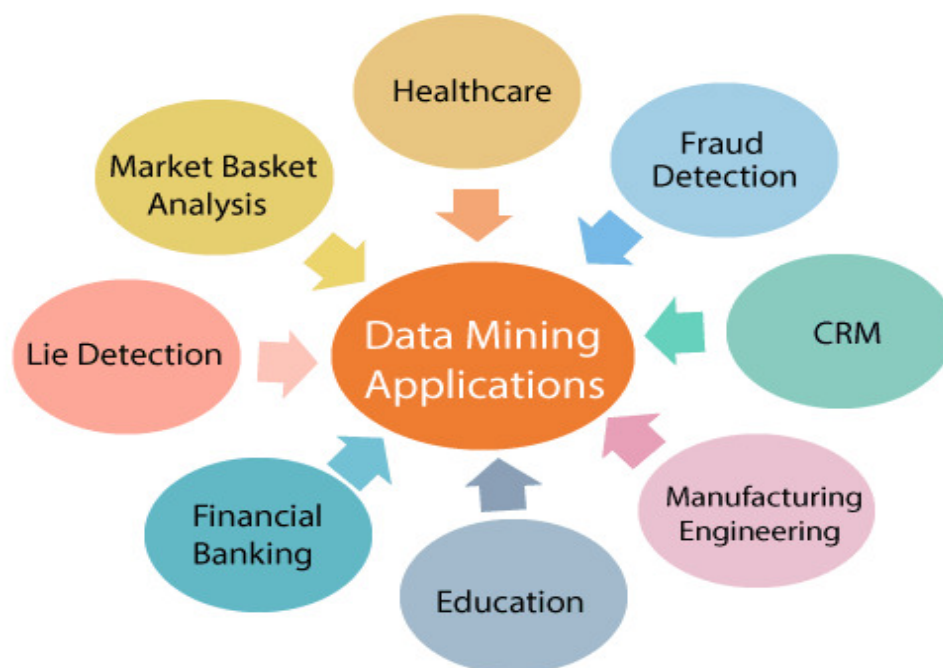
8. Time-series Databases: Time-series databases contain time related data such as stock market data or logged activities. These databases usually have a continuous flow of new data coming in, which sometimes causes the need for a challenging real time analysis. Data mining in such databases commonly includes the study of trends and correlations between evolutions of different variables, as well as the prediction of trends and movements of the variables in time. Handles array of numbers indexed by time, date, etc. It requires real-time analysis. **Application:** eXtremeDB, Graphite, InfluxDB, etc.

9. WWW: WWW refers to World wide web is a collection of documents and resources like audio, video, text, etc which are identified by Uniform Resource Locators (URLs) through web

browsers, linked by HTML pages, and accessible via the Internet network. It is the most heterogeneous repository as it collects data from multiple resources. It is dynamic in nature as Volume of data is continuously increasing and changing. **Application:** Online shopping, Job search, Research, studying, etc.

Data Mining Applications

Data Mining is primarily used by organizations with intense consumer demands- Retail, Communication, Financial, marketing company, determine price, consumer preferences, product positioning, and impact on sales, customer satisfaction and corporate profits. Data mining enables a retailer to use point-of-sale records of customer purchases to develop products and promotions that help the organization to attract the customer.



These are the following areas where data mining is widely used:

Data Mining in Healthcare:

Data mining in healthcare has excellent potential to improve the health system. It uses data and analytics for better insights and to identify best practices that will enhance health care services

and reduce costs. Analysts use data mining approaches such as Machine learning, Multi-dimensional database, Data visualization, Soft computing, and statistics. Data Mining can be used to forecast patients in each category. The procedures ensure that the patients get intensive care at the right place and at the right time. Data mining also enables healthcare insurers to recognize fraud and abuse.

Data Mining in Market Basket Analysis:

Market basket analysis (also known as association analysis or frequent itemset mining) is a modeling method based on a hypothesis. If you buy a specific group of products, then you are more likely to buy another group of products. This technique may enable the retailer to understand the purchase behavior of a buyer. This data may assist the retailer in understanding the requirements of the buyer and altering the store's layout accordingly. Using a different analytical comparison of results between various stores, between customers in different demographic groups can be done. Market basket analysis can also be used to cross-sell products. Amazon famously uses an algorithm to suggest items that you might be interested in, based on your browsing history or what other people have purchased.

Data mining in Education:

Education data mining is a newly emerging field, concerned with developing techniques that explore knowledge from the data generated from educational Environments. EDM objectives are recognized as affirming student's future learning behavior, studying the impact of educational support, and promoting learning science. An organization can use data mining to make precise decisions and also to predict the results of the student. With the results, the institution can concentrate on what to teach and how to teach. For example, several learning management systems (LMSs) track information such as when each student accessed each learning object, how many times they accessed it, and how many minutes the learning object was displayed on the user's computer screen. As another example, intelligent tutoring systems record data every time a learner submits a solution to a problem. They may collect the time of the submission, whether or not the solution matches the expected solution, the amount of time that has passed since the last submission, the order in which solution components were entered into the interface, etc.

Data Mining in Manufacturing Engineering:

Knowledge is the best asset possessed by a manufacturing company. Data mining tools can be beneficial to find patterns in a complex manufacturing process. Data mining can be used in system-level designing to obtain the relationships between product architecture, product portfolio, and data needs of the customers. It can also be used to forecast the product development period, cost, and expectations among the other tasks.

Data Mining in CRM (Customer Relationship Management):

Customer Relationship Management (CRM) is all about obtaining and holding Customers, also enhancing customer loyalty and implementing customer-oriented strategies. To get a decent relationship with the customer, a business organization needs to collect data and analyze the data. With data mining technologies, the collected data can be used for analytics.

Data Mining in Fraud detection:

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are a little bit time consuming and sophisticated. Data mining provides meaningful patterns and turning data into information. An ideal fraud detection system should protect the data of all the users. Supervised methods consist of a collection of sample records, and these records are classified as fraudulent or non-fraudulent. A model is constructed using this data, and the technique is made to identify whether the document is fraudulent or not.

Data Mining in Lie Detection:

Apprehending a criminal is not a big deal, but bringing out the truth from him is a very challenging task. Law enforcement may use data mining techniques to investigate offenses, monitor suspected terrorist communications, etc. This technique includes text mining also, and it seeks meaningful patterns in data, which is usually unstructured text. The information collected from the previous investigations is compared, and a model for lie detection is constructed.

Data Mining Financial Banking:

The Digitalization of the banking system is supposed to generate an enormous amount of data with every new transaction. The data mining technique can help bankers by solving business-related problems in banking and finance by identifying trends, casualties, and correlations in business information and market costs that are not instantly evident to managers or executives because the data volume is too large or are produced too rapidly on the screen by experts. The manager may find these data for better targeting, acquiring, retaining, segmenting, and maintain a profitable customer.

Data Mining Functionalities—What Kinds of Patterns Can Be Mined?

The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: *descriptive data mining* tasks that describe the general properties of the existing data. It includes certain knowledge to understand what is happening within the data without a previous idea. The common data features are highlighted in the data set. For examples: count, average etc. and *predictive data mining* tasks that attempt to do predictions based on inference on available data. It helps developers to provide unlabeled definitions of attributes. Based on previous tests, the software estimates the characteristics that are absent. For example: Judging from the findings of a patient's medical examinations that is he suffering from any particular disease.

The data mining functionalities and the variety of knowledge they discover are briefly presented in the following list:

- **Characterization:** Data characterization is a summarization of general features of objects in a target class (class under study) and produces what is called *characteristic rules*. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions. For example, to study the characteristics of software products with sales that increased by 10% in the previous year, the data related to such products can be collected

by executing an SQL query on the sales database. Another example, one may want to characterize the OurVideoStore customers who regularly rent more than 30 movies a year. With concept hierarchies on the attributes describing the target class, the *attribute-oriented induction* method can be used, for example, to carry out data summarization. Note that with a data cube containing summarization of data, simple OLAP operations fit the purpose of data characterization. The output of data characterization can be presented in various forms. Examples include pie charts, bar charts, curves, multidimensional data cubes and multidimensional tables, including crosstabs. A customer relationship manager at *AllElectronics* may order the following data mining task: *Summarize the characteristics of customers who spend more than \$5000 a year at AllElectronics*. The result is a general profile of these customers, such as that they are 40 to 50 years old, employed, and have excellent credit ratings. The data mining system should allow the customer relationship manager to drill down on any dimension, such as on *occupation* to view these customers according to their type of employment.

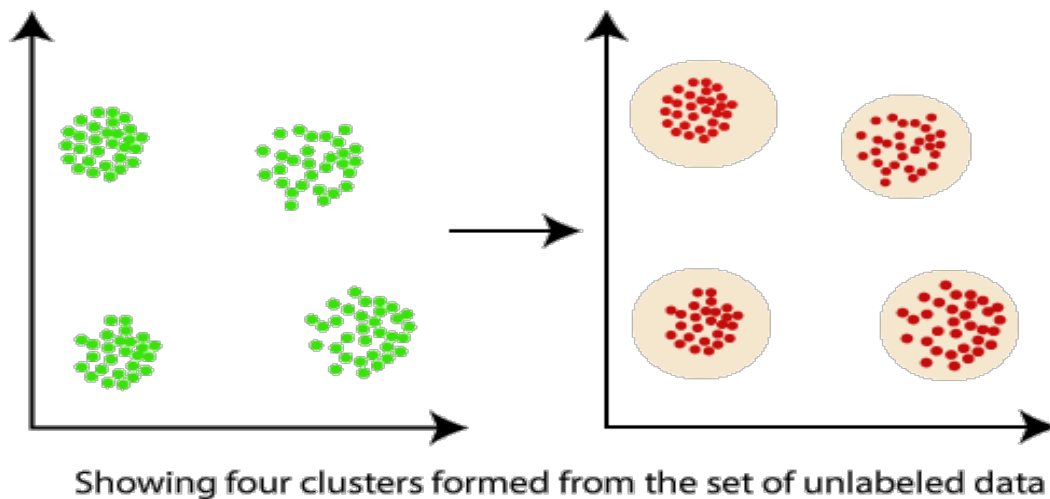
- **Discrimination:** Data discrimination is a comparison of the general features of the target class data objects against the general features of objects from one or multiple contrasting classes. Data discrimination produces what are called *discriminant rules* and is basically the comparison of the general features of objects between two classes referred to as the *target class* and the *contrasting class*. For example, user may want to compare the general features of software products with sales that increased by 10% last year against those with sales that decreased by at least 30% during the same period. Another example, one may want to compare the general characteristics of the customers who rented more than 30 movies in the last year with those whose rental account is lower than 5. The techniques used for data discrimination are very similar to the techniques used for data characterization with the exception that data discrimination results include comparative measures. The forms of output presentation are similar to those for characteristic descriptions, although discrimination descriptions should include comparative measures that help to distinguish between the target and contrasting classes.

- Association analysis:** Association analysis is the discovery of what are commonly called *association rules*. It studies the frequency of items occurring together in transactional databases, and based on a threshold called *support*, identifies the frequent item sets. Another threshold, *confidence*, which is the conditional probability that an item appears in a transaction when another item appears, is used to pinpoint association rules. Support represents the popularity of that product of all the product transactions. Support of the product is calculated as the ratio of the number of transactions includes that product and the total number of transactions. Confidence can be interpreted as the likelihood of purchasing both the products A and B. Association analysis is commonly used for market basket analysis. For example, it could be useful for the OurVideoStore manager to know what movies are often rented together or if there is a relationship between renting a certain type of movies and buying popcorn or pop. The discovered association rules are of the form: $P \rightarrow Q [s,c]$, where P and Q are conjunctions of attribute value-pairs, and s (for support) is the probability that P and Q appear together in a transaction and c (for confidence) is the conditional probability that Q appears in a transaction when P is present. For example, the hypothetical association rule: *RentType(X, "game") AND Age(X, "13-19") \rightarrow Buys(X, "pop")* [$s=2\%$, $c=55\%$] would indicate that 2% of the transactions considered are of customers aged between 13 and 19 who are renting a game and buying a pop, and that there is a certainty of 55% that teenage customers who rent a game also buy pop.
- Classification:** It is a data analysis task, i.e. the process of finding a model that describes and distinguishes data classes and concepts. Classification is the problem of identifying to which of a set of categories (subpopulations), a new observation belongs to, on the basis of a training set of data containing observations and whose categories membership is known. Classification analysis is the organization of data in given classes. Also known as *supervised classification*, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a *training set* where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model. The model is used to classify new objects. For example, after starting a credit policy, the OurVideoStore managers could analyze the customers behaviours vis-à-vis their credit, and label accordingly the

customers who received credits with three possible labels "safe", "risky" and "very risky". The classification analysis would generate a model that could be used to either accept or reject credit requests in the future. Suppose there is a person who is sitting under a fan and the fan starts falling on him, he should get aside in order not to get hurt. So, this is his training part to move away. While testing if the person sees any heavy object coming towards him or falling on him and moves aside then the system is tested positively and if the person does not move aside then the system is negatively tested. Same is the case with the data; it should be trained in order to get the accurate and best results.

- **Prediction:** Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Once a classification model is built based on a training set, the class label of an object can be foreseen based on the attribute values of the object and the attribute values of the classes. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values. Prediction is nothing but finding out the knowledge or some pattern from the large amounts of data. For example, In credit card fraud detection, history of data for a particular person's credit card usage has to be analyzed. If any abnormal pattern was detected, then it should be reported as 'fraudulent action'.
- **Clustering:** Clustering is similar to classification in that data is grouped. However, unlike classification, the groups are not predefined. Instead, the grouping is accomplished by finding similarities between data according to characteristics found in the actual data. The groups are called Clusters. Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes. Clustering is also called *unsupervised classification*, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (*intra-class similarity*) and minimizing the

similarity between objects of different classes (*inter-class similarity*). Let's understand this with an example, suppose we are a market manager, and we have a new tempting product to sell. We are sure that the product would bring enormous profit, as long as it is sold to the right people. So, how can we tell who is best suited for the product from our company's huge customer base?



Clustering, falling under the category of **unsupervised machine learning**, is one of the problems that machine learning algorithms solve.

Clustering only utilizes input data, to determine patterns, anomalies, or similarities in its input data.

A good clustering algorithm aims to obtain clusters whose:

- The intra-cluster similarities are high; It implies that the data present inside the cluster is similar to one another.
- The inter-cluster similarity is low, and it means each cluster holds data that is not similar to other data.
- **Outlier analysis:** Outliers are data elements that cannot be grouped in a given class or cluster. Also known as *exceptions* or *surprises*, they are often very important to identify. For example, in a normal distribution, outliers may be values on the tails of the

distribution. The process of identifying outliers has many names in data mining and machine learning such as outlier mining, outlier modeling and novelty detection and anomaly detection. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis valuable. “Outlier Analysis is a process that involves identifying the anomalous observation in the dataset.” Outliers are caused due to the incorrect entry or computational error, mis-reporting, sampling error, Exceptional but true value error. For example, displaying a person’s weight as 1000kg could be caused by a program default setting of an unrecorded weight. Alternatively, outliers may be a result of indigenous data changeability. Many algorithms are used to minimize the effect of outliers or eliminate them. This may be able to result in the loss of important hidden information because one person’s noise could be another person’s signal. In some instances like fraud detection, the outlier indicates a fraudulent activity.

- **Evolution and deviation analysis:** Evolution and deviation analysis pertain to the study of time related data that changes in time. Evolution Analysis pertains to study of data sets that undergo change over a time period. Evolution analysis models are designed to capture evolutionary trends in data helping in characterizing, classifying, clustering or discrimination of time-related data. Deviation analysis, on the other hand, considers differences between measured values and expected values, and attempts to find the cause of the deviations from the anticipated values.

What are the issues in Data Mining?

Data mining algorithms embody techniques that have sometimes existed for many years, but have only lately been applied as reliable and scalable tools that time and again outperform older classical statistical methods. While data mining is still in its infancy, it is becoming a trend and ubiquitous. Before data mining develops into a conventional, mature and trusted discipline, many still pending issues have to be addressed. Some of these issues are addressed below. Note that these issues are not exclusive and are not ordered in any way.

Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected

for customer profiling, user behaviour understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control.

User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels.

Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs (when known), the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently.

Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information.

More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve.

Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to re-analyze the complete dataset.

Data source issues: There are many issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has

helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community.

Challenges of Implementation in Data mining

Although data mining is very powerful, it faces many challenges during its execution. Various challenges could be related to performance, data, methods, and techniques, etc. The process of data mining becomes effective when the challenges or problems are correctly recognized and adequately resolved.



Incomplete and noisy data:

The process of extracting useful data from large volumes of data is data mining. The data in the real-world is heterogeneous, incomplete, and noisy. Data in huge quantities will usually be inaccurate or unreliable. These problems may occur due to data measuring instrument or because of human errors. Suppose a retail chain collects phone numbers of customers who spend more than \$500, and the accounting employees put the information into their system. The person may make a digit mistake when entering the phone number, which results in incorrect data. Even some customers may not be willing to disclose their phone numbers, which results in incomplete data. The data could get changed due to human or system error. All these consequences (noisy and incomplete data) make data mining challenging.

Data Distribution:

Real-worlds data is usually stored on various platforms in a distributed computing environment. It might be in a database, individual systems, or even on the internet. Practically, it is a quite tough task to make all the data to a centralized data repository mainly due to organizational and technical concerns. For example, various regional offices may have their servers to store their data. It is not feasible to store, all the data from all the offices on a central server. Therefore, data mining requires the development of tools and algorithms that allow the mining of distributed data.

Complex Data:

Real-world data is heterogeneous, and it could be multimedia data, including audio and video, images, complex data, spatial data, time series, and so on. Managing these various types of data and extracting useful information is a tough task. Most of the time, new technologies, new tools, and methodologies would have to be refined to obtain specific information.

Performance:

The data mining system's performance relies primarily on the efficiency of algorithms and techniques used. If the designed algorithm and techniques are not up to the mark, then the efficiency of the data mining process will be affected adversely.

Data Privacy and Security:

Data mining usually leads to serious issues in terms of data security, governance, and privacy. For example, if a retailer analyzes the details of the purchased items, then it reveals data about buying habits and preferences of the customers without their permission.

Data Visualization:

In data mining, data visualization is a very important process because it is the primary method that shows the output to the user in a presentable way. The extracted data should convey the exact meaning of what it intends to express. But many times, representing the information to the end-user in a precise and easy way is difficult. The input data and the output information being complicated, very efficient, and successful data visualization processes need to be implemented to make it successful.

Data Objects and Attribute Types

Data objects are the essential part of a database. A data object represents the entity. Data Objects are like a group of attributes of an entity. For example, a sales data object may represent customers, sales, or purchases. When a data object is listed in a database they are called data tuples.

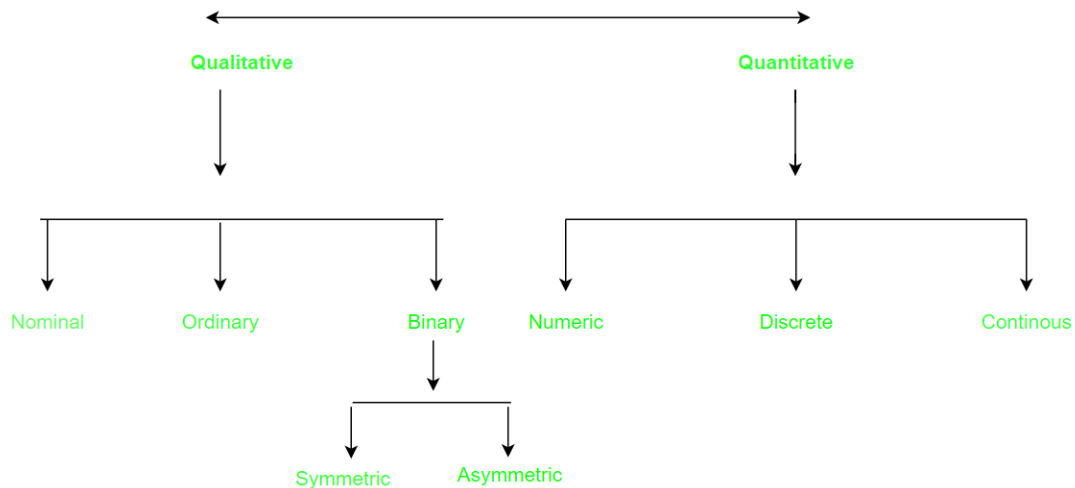
Attribute:

It can be seen as a data field that represents the characteristics or features of a data object. For a customer, object attributes can be customer Id, address, etc. We can say that a **set of attributes used to describe a given object are known as attribute vector or feature vector.**

Type of attributes:

This is the First step of Data Data-preprocessing. We differentiate between different types of attributes and then preprocess the data. So here is the description of attribute types.

1. Qualitative (Nominal (N), Ordinal (O), Binary(B)).
2. Quantitative (Numeric, Discrete, Continuous)



Qualitative Attributes:

1. Nominal Attributes – related to names: The values of a Nominal attribute are names of things, some kind of symbols. Values of Nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and there is no order (rank, position) among values of the nominal attribute.

Example:

Attribute	Values
Colours	Black, Brown, White
Categorical Data	Lecturer, Professor, Assistant Professor

2. Binary Attributes: Binary data has only 2 values/states. For Example, yes or no, affected or unaffected, true or false.

- **Symmetric:** Both values are equally important (Gender).

- **Asymmetric:** Both values are not equally important (Result).

Attribute		Attribute	Values
Attribute	Values	Cancer detected	Yes, No
Gender	Male , Female	result	Pass , Fail

3. Ordinal Attributes : The Ordinal Attributes contains values that have a meaningful sequence or ranking (order) between them, but the magnitude between values is not actually known, the order of values that shows what is important but don't indicate how important it is.

Attribute	Value
Grade	A,B,C,D,E,F
Basic pay scale	16,17,18

Quantitative Attributes:

1. Numeric: A numeric attribute is quantitative because, it is a measurable quantity, represented in integer or real values. Numerical attributes are of 2 types, **interval**, and **ratio**.

An **interval-scaled** attribute has values, whose differences are interpretable, but the numerical attributes do not have the correct reference point, or we can call zero points. Data can be added and subtracted at an interval scale but cannot be multiplied or divided. E.g. calendar dates, Consider an example of temperature in degrees Centigrade. If a day's temperature of one day is twice of the other day we cannot say that one day is twice as hot as another day.

A **ratio-scaled** attribute is a numeric attribute with a fix zero-point. If a measurement is ratio-scaled, we can say of a value as being a multiple (or ratio) of another value. The values are ordered, and we can also compute the difference between values, and the mean, median, mode, Quantile-range, and Five number summary can be given.

2. Discrete: Discrete data have finite values it can be numerical and can also be in categorical form. These attributes has finite or countable infinite set of values.

Example:

Attribute	Value
Profession	Teacher, Business man, Peon
ZIP Code	301701, 110040

3. Continuous: Continuous data have an infinite no of states. Continuous data is of float type. There can be many values between 2 and 3.

Example:

Attribute	Value
Height	5.4, 6.2 ...etc
weight	50.33etc

Basic Statistical Descriptions of Data

Basic Statistical descriptions can be used to identify properties of the data and highlight which data values should be treated as noise or outliers.

Three areas of basic statistical descriptions are as follows:

1. Measures of central tendency: which measures the location of the middle or center of a data distribution.

MEAN, MEDIAN, MODE, MIDDLE RANGE

2. Dispersion of the data: how are the data spread out? useful in identifying outliers

Most common data dispersion measures are:

RANGE, QUANTILES, INTER-QUARTILE RANGE, The Five-Number Summary and Boxplots; and the Variance and Standard deviation of the data.

3. Graphic displays of basic statistical descriptions: to visually inspect our data.

Includes Bar Charts, Pie Charts, and Line Graphs. Other popular displays of data summaries and distributions include Quantile Plots, Quantile–Quantile Plots, Histograms and Scatter Plots.

Measures of central tendency: which measure the location of the middle or center of a data distribution.

MEAN: Let x_1, x_2, \dots, x_N be the set of N observed values or observations for X , like X is salary. The mean of this set of values is: The mean of this set of values is $\bar{x} = (x_1 + x_2 + \dots + x_N)/N$

Example 1: Suppose we have the following values for salary (in thousands of dollars), shown in increasing order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = (30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110)/12 = 696/12 = 58.$$

Thus, the mean salary is 58,000.

MEDIAN: For skewed (asymmetric) data, a better measure of the center of data is the median, which is the middle value in a set of ordered data values. It is the value that separates the higher half of a data set from the lower half.

In probability and statistics, the median generally applies to numeric data; however, we may extend the concept to ordinal data. Suppose that a given data set of N values for an attribute X is sorted in increasing order. If N is odd, then the median is the middle value of the ordered set. If N is even, then the median is not unique; it is the two middlemost values and any value in between. If X is a numeric attribute in this case, by convention, the median is taken as the average of the two middlemost values.

EXAMPLE: Let's find the median of the data from Example 1. The data are already sorted in increasing order. There is an even number of observations (i.e., 12); therefore, the median is not

unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is, $(52+56)/2 = 108/2 = 54$.

Thus, the median is \$54,000.

Now, suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of \$52,000.

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70

Median=52

MODE: The mode is another measure of central tendency. The mode for a set of data is the value that occurs most frequently in the set. Therefore, it can be determined for qualitative and quantitative attributes.

It is possible for the greatest frequency to correspond to several different values, which results in more than one mode. Data sets with one, two, or three modes are respectively called unimodal, bimodal, and trimodal. In general, a data set with two or more modes is multimodal.

At the other extreme, if each data value occurs only once, then there is no mode.

EXAMPLE: The data from Example 1 are bimodal. The two modes are \$52,000 and \$70,000.

Midrange: It is the average of the largest and smallest values in the set. This measure is easy to compute using the SQL aggregate functions, `max()` and `min()`.

The midrange of the data of Example 1 is $(30,000+110,000)/2 = \$70,000$.

Measuring the Dispersion of Data: Range, Quartiles, Variance, Standard Deviation and Interquartile Range

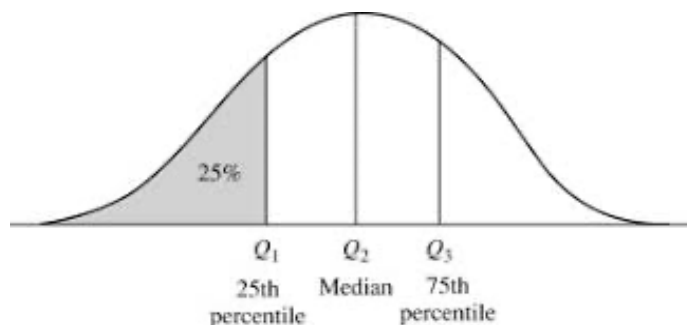
We now look at measures to assess the dispersion or spread of numeric data. The measures include range, quantiles, quartiles, percentiles, and the interquartile range. The five-number

summary, which can be displayed as a boxplot is useful in identifying outliers. Variance and standard deviation also indicate the spread of a data distribution.

Range: Let x_1, x_2, \dots, x_N be a set of observations for some numeric attribute, X . The range of the set is the difference between the largest ($\max()$) and smallest ($\min()$) values. The range is the difference in the maximum and minimum values of a data set. The maximum is the largest value in the dataset and the minimum is the smallest value.

$$\text{Range} = \text{maximum} - \text{minimum}$$

Quartile: A quartile divides data into three points—a lower quartile, median, and upper quartile—to form four groups of the dataset. The first quartile (Q_1) is defined as the middle number between the smallest number and the median of the data set. The second quartile, Q_2 , is also the median. The upper or third quartile, denoted as Q_3 , is the central point that lies between the median and the highest number of the distribution. Example: 1, 3, 3, 4, 5, 6, 6, 7, 8, 8. The numbers are already in order. Cut the list into quarters: In this case, Quartile 2 is half way between 5 and 6: $Q_2 = (5+6)/2 = 5.5$.



The quartiles give an indication of a distribution's center, spread, and shape. The first quartile, denoted by Q_1 , is the 25th percentile. It cuts off the lowest 25% of the data. The third quartile, denoted by Q_3 , is the 75th percentile—it cuts off the lowest 75% (or highest 25%) of the data. The second quartile is the 50th percentile. As the median, it gives the center of the data distribution.

Interquartile Range: The **interquartile range** is the difference between upper and lower quartiles and denoted as **IQR**.

$$\text{IQR} = Q3 - Q1 = \text{upper quartile} - \text{lower quartile} = 75\text{th percentile} - 25\text{th percentile}$$

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the Interquartile range (IQR).

Example: the quartiles for this data **30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110**.

are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, $Q1 = \$47,000$ and $Q3$ is $\$63,000$. Thus, the interquartile range is $\text{IQR} = 63 - 47 = \$16,000$. (Note that the sixth value is a median, $\$52,000$, although this data set has two medians since the number of data values is even.)

Five-Number Summary, Boxplots and Outliers

Five-Number Summary: Because $Q1$, the median, and $Q3$ together contain no information about the endpoints (e.g., tails) of the data, a fuller summary of the shape of a distribution can be obtained by providing the lowest and highest data values as well. This is known as the five-number summary.

The five-number summary of a distribution consists of the median ($Q2$), the quartiles $Q1$ and $Q3$, and the smallest and largest individual observations, written in the order of Minimum, $Q1$, Median, $Q3$, Maximum.

The five-number summary of the data set: 5, 7, 11, 13, 17, 19, **23**, 29, 31, 37, 41, 43, 47 is

5, 12, 23, 39, and 47.

Outliers: A common rule of thumb for identifying suspected outliers is to single out values falling at least $1.5 \times \text{IQR}$ above the third quartile or below the first quartile.

Lower and Upper Limits

The lower limit and upper limit of a data set are given by:

$$\text{Lower limit} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper limit} = Q_3 + 1.5 \times \text{IQR}$$

Data points that lie below the lower limit or above the upper limit are potential outliers.

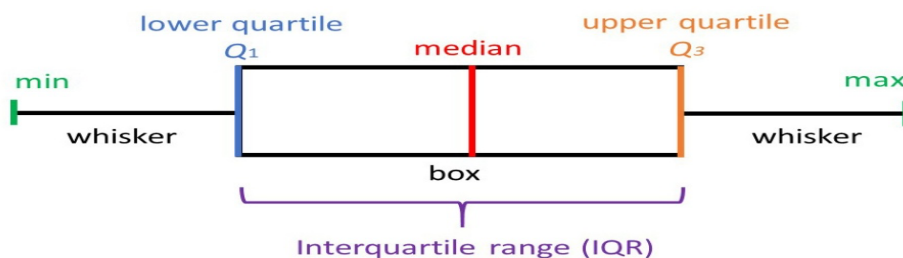
$$\text{The lower limit of the set is } 12 - (1.5 \times 27) = -28.5$$

$$\text{The upper limit of the set is } 39 + (1.5 \times 27) = 79.5$$

Since there are no elements of the set less than -28.5 or greater than 79.5, there are no outliers.

Boxplots: In descriptive statistics, a **box plot** or **boxplot** is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending from the boxes (*whiskers*) indicating variability outside the upper and lower quartiles, hence the terms **box-and-whisker plot** and **box-and-whisker diagram**. Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

- Typically, the ends of the box are at the quartiles so that the box length is the interquartile range.
- The median is marked by a line within the box.
- Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.



Example:

22, 25, 17, 19, 33, 64, 23, 17, 20, 18

Step-1 Arrange the data in ascending order

17, 17, 18, 19, 20, 22, 23, 25, 33, 64

Step-2 So, median is $(20+22)/2=21$ (Q2)

Step-3 17, 17, 18, 19, 20, 22, 23, 25, 33, 64

Q1=18 Q3=25

Step-4 interquartile range=Q3-Q1=25-18=7

Check for outliers

1. higher outlier=Q3+[1.5*IQR]

$$=25+[1.5*7]$$

$$=25+10.5=35.5$$

Any data > higher outlier

64 > higher outlier

So 64 is a outlier.

2. lower outlier=Q1-[1.5*IQR]

$$=18-[1.5*7]$$

$$=18-10.5$$

$$=7.5$$

Any data <7.5

No

So, $Q1=18$

$Q2=21$

$Q3=25$

Outlier=64

Min=17

Max=33

Variance and Standard Deviation

Variance and standard deviation are measures of data dispersion. They indicate how spread out a data distribution is. A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values. Unlike range and quartiles, the variance combines all the values in a data set to produce a measure of spread. The variance (symbolized by S^2) and standard deviation (the square root of the variance, symbolized by S) are the most commonly used measures of spread.

We know that variance is a measure of how spread out a data set is. It is calculated as the average squared deviation of each number from the mean of a data set. For example, for the numbers 1, 2, and 3 the mean is 2 and the variance is 0.667.

$$[(1 - 2)^2 + (2 - 2)^2 + (3 - 2)^2] \div 3 = 0.667$$

[squaring deviation from the mean] \div number of observations = variance

Variance (S^2) = average squared deviation of values from mean

Calculating variance involves squaring deviations, so it does not have the same unit of measurement as the original observations. For example, lengths measured in metres (m) have a variance measured in metres squared (m²).

Taking the square root of the variance gives us the units used in the original scale and this is the standard deviation.

Standard deviation (S) = square root of the variance

Standard deviation is the measure of spread most commonly used in statistical practice when the mean is used to calculate central tendency. Thus, it measures spread around the mean. Because of its close links with the mean, standard deviation can be greatly affected if the mean gives a poor measure of central tendency. Generally, the more widely spread the values are, the larger the standard deviation is. For example, imagine that we have to separate two different sets of exam results from a class of 30 students the first exam has marks ranging from 31% to 98%, the other ranges from 82% to 93%. Given these ranges, the standard deviation would be larger for the results of the first exam.

Example: Standard deviation

A hen lays eight eggs. Each egg was weighed and recorded as follows:

60 g, 56 g, 61 g, 68 g, 51 g, 53 g, 69 g, 54 g.

- a. First, calculate the mean:

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{472}{8} \\ &= 59\end{aligned}$$

- b. Now, find the standard deviation.

Table 1. Weight of eggs, in grams		
Weight (x)	(x - \bar{x})	(x - \bar{x})²
60	1	1

56	-3	9
61	2	4
68	9	81
51	-8	64
53	-6	36
69	10	100
54	-5	25
472		320

c. Using the information from the above table, we can see that

$$\sum (x - \bar{x})^2 = 320$$

In order to calculate the standard deviation, we must use the following formula:

$$\begin{aligned}
 s &= \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \\
 &= \sqrt{\frac{320}{8}} \\
 &= 6.32 \text{ grams}
 \end{aligned}$$

The basic properties of the standard deviation, σ , as a measure of spread are as follows:

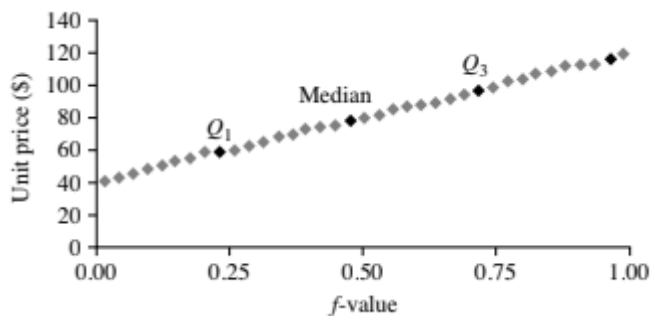
σ measures spread about the mean and should be considered only when the mean is chosen as the measure of center.

$\sigma = 0$ only when there is no spread, that is, when all observations have the same value. Otherwise, $\sigma > 0$.

Graphic Displays of Basic Statistical Descriptions of Data

Quantile Plot: A quantile plot is a simple and effective way to have a first look at a univariate data distribution. Let x_i , for $i = 1$ to N , be the data sorted in increasing order so that x_1 is the smallest observation and x_N is the largest for some ordinal or numeric attribute X . Each observation, x_i , is paired with a percentage, f_i , which indicates that approximately $f_i \times 100\%$ of the data are below the value, x_i . We say “approximately” because there may not be a value with exactly a fraction, f_i , of the data below x_i . Note that the 0.25 percentile corresponds to quartile Q_1 , the 0.50 percentile is the median, and the 0.75 percentile is Q_3 .

Let $f_i = (i - 0.5)/N$



Quantile–Quantile Plot: A quantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another. It is a powerful visualization tool in that it allows the user to view whether there is a shift in going from one distribution to another.

Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price (\$)	Count of items sold
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

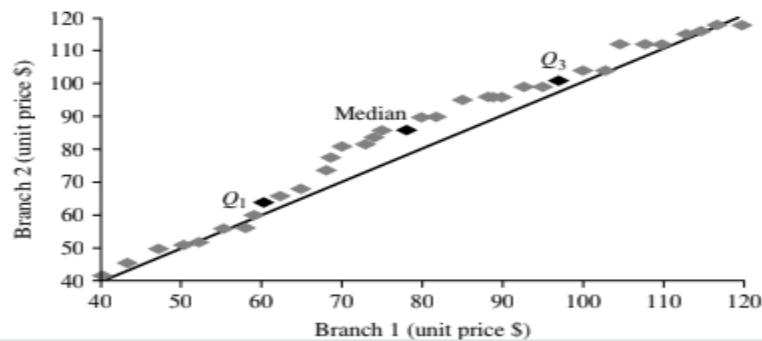


Figure shows a quantile–quantile plot for unit price data of items sold at two branches of AllElectronics during a given time period. Each point corresponds to the same quantile for each data set and shows the unit price of items sold at branch 1 versus branch 2 for that quantile. (To aid in comparison, the straight line represents the case where, for each given quantile, the unit price at each branch is the same. The darker points correspond to the data for Q_1 , the median, and Q_3 , respectively.)

We see, for example, that at Q_1 , the unit price of items sold at branch 1 was slightly less than that at branch 2. In other words, 25% of items sold at branch 1 were less than or equal to \$60, while 25% of items sold at branch 2 were less than or equal to \$64. At the 50th percentile (marked by the median, which is also Q_2), we see that 50% of items sold at branch 1 were less than \$78, while 50% of items at branch 2 were less than \$85.

In general, we note that there is a shift in the distribution of branch 1 with respect to branch 2 in that the unit prices of items sold at branch 1 tend to be lower than those at branch 2.

Histograms: Histograms (or frequency histograms) are at least a century old and are widely used. “Histos” means pole or mast, and “gram” means chart, so a histogram is a chart of poles. Plotting histograms is a graphical method for summarizing the distribution of a given attribute, X . If X is nominal, such as automobile model or item type, then a pole or vertical bar is drawn for each known value of X . The height of the bar indicates the frequency (i.e., count) of that X value. The resulting graph is more commonly known as a bar chart.

If X is numeric, the term histogram is preferred. The range of values for X is partitioned into disjoint consecutive subranges. The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X . The range of a bucket is known as the width. Typically, the buckets are of equal width. For example, a price attribute with a value range of \$1 to \$200 (rounded up to the nearest dollar) can be partitioned into subranges 1 to 20, 21 to 40, 41 to 60, and so on.

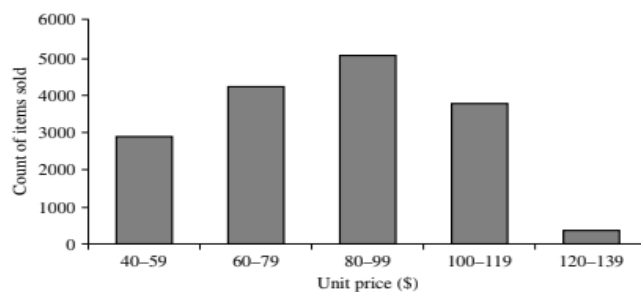


Figure 2.6 A histogram for the Table 2.1 data set.

Scatter Plots and Data Correlation

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes. To construct a scatter plot, each pair of values is treated as a pair of coordinates in an algebraic sense and plotted as points in the plane.

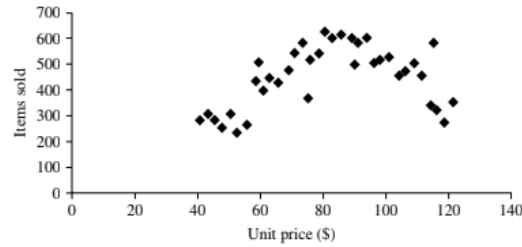


Figure 2.7 A scatter plot for the Table 2.1 data set.

Two attributes, X , and Y , are correlated if one attribute implies the other. Correlations can be positive, negative, or null (uncorrelated). Figure below shows the examples of positive and negative correlations between two attributes. If the plotted point's pattern slopes from lower left to upper right, this means that the values of X increase as the values of Y increase, suggesting a positive correlation (Figure 2.8a). If the pattern of plotted points slopes from upper left to lower right, the values of X increase as the values of Y decrease, suggesting a negative correlation (Figure 2.8b). A line of best fit can be drawn to study the correlation between the variables.

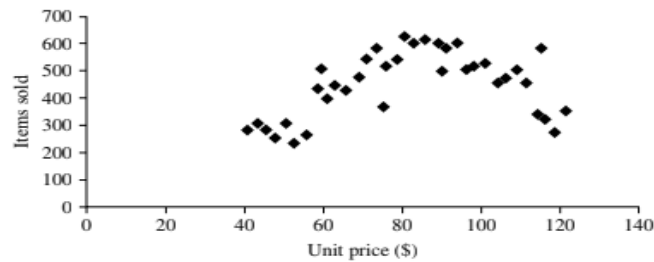


Figure 2.7 A scatter plot for the Table 2.1 data set.

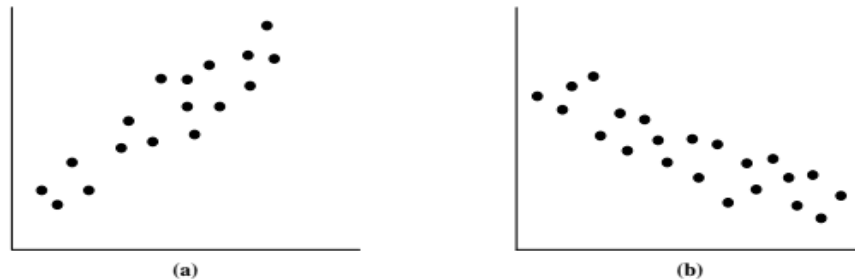


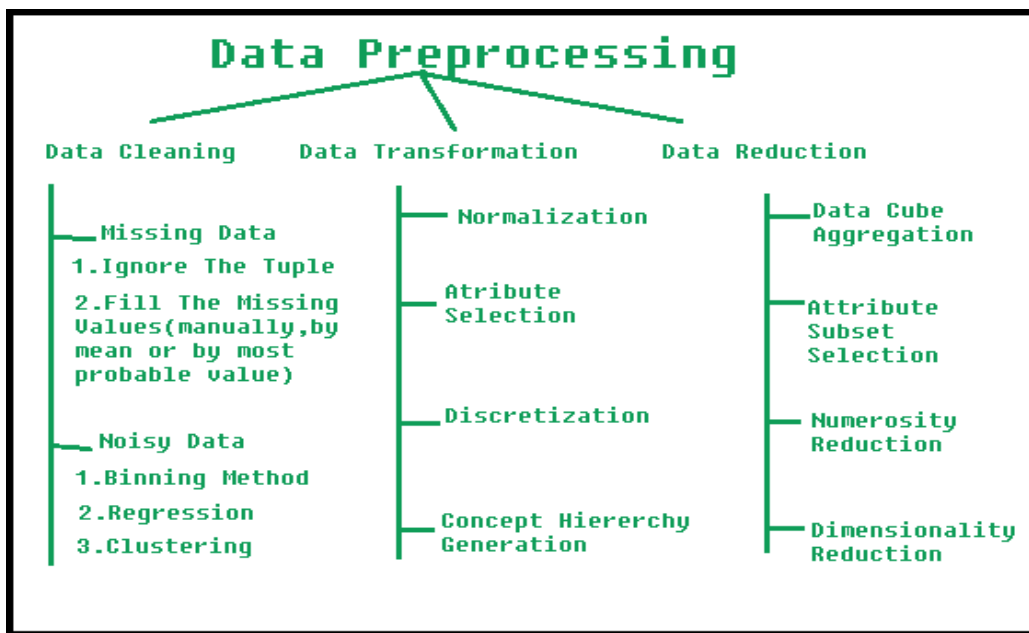
Figure 2.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

Need for Data Preprocessing and Data Quality

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, lacking in certain behaviors or trends, and is likely to contain many errors. In the real world, data is frequently unclear – missing key values, containing inconsistencies or displaying “noise” (containing errors and outliers). Without data preprocessing, these data mistakes will survive and detract from the quality of data mining. Data preprocessing is a proven method of resolving such issues.

Preprocessing in Data Mining:

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.



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) Missing Data:

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

(i) **Ignore the tuples:** This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

(ii) **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.

(b) 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:

(i) **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.

(ii) **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).

(iii) **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:

(a) **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)

(b) **Attribute Selection:** In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

(c) **Discretization:** This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

(d) **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:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

(a) **Data Cube Aggregation:** Aggregation operation is applied to data for the construction of the data cube.

(b) **Attribute Subset Selection:** The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute, the attribute having p-value greater than significance level can be discarded.

(c) **Numerosity Reduction:** This enables to store the model of data instead of whole data, for example: Regression Models.

(d) **Dimensionality Reduction:** This reduces the size of data by encoding mechanisms. It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Component Analysis).

Data Quality: Why Preprocesses the data

Data quality is a measure of the condition of data based on factors such as accuracy, completeness, consistency, timeliness, believability and interpretability whether it's up to date. Measuring data quality levels can help organizations identify data errors that need to be resolved and assess whether the data in their IT systems is fit to serve its intended purpose. Factors used for data quality assessment are:

- **Accuracy:** There are many possible reasons for flawed or inaccurate data here. i.e. having incorrect values of properties that could be human or computer errors. The data collection instruments used may be faulty. There may have been human or computer errors occurring at data entry. Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information, e.g., by choosing the default value 'January 1' displayed for birthday. (This is known as disguised missing data.)
- **Completeness:** For some reasons, incomplete data can occur, attributes of interest such as customer information for sales & transaction data may not always be available. Other data may not be included simply because they were not considered important at the time of entry. Relevant data may not be recorded due to a misunderstanding, or because of equipment malfunctions. Data that were inconsistent with other recorded data may have been deleted.
- **Consistency:** Incorrect data can also result from inconsistencies in naming convention or data codes, or from input field incoherent format. Duplicate tuples need cleaning of details, too.
- **Timeliness:** It also affects the quality of the data. At the end of the month, several sales representatives fail to file their sales record on time. These are also several corrections & adjustments which flow into after the end of the month. Data stored in the database are incomplete for a time after each month.

- **Believability:** It is reflective of how much users trust the data.
- **Interpretability:** It is a reflection of how easy the users can understand the data.

Suppose that a database, at one point, had several errors, all of which have since been corrected. The past errors, however, had caused many problems for users in the sales department, and so they no longer trust the data. The data also use many accounting codes, which the sales department does not know how to interpret. Even though such a database is now accurate, complete, consistent, and timely, users from the sales department may regard it as of low quality due to poor believability and interpretability.

Data Integration

Data integration merges data from several heterogeneous sources to attain meaningful data. The source involves several databases, flat files, multiple files or data cubes. The integrated data must exempt inconsistencies, discrepancies, redundancies and disparity. Data integration is important as it provides a unified view of the scattered data not only this it also maintains the accuracy of data. This helps the data-mining program in mining the useful information which in turn helps the executive and managers in taking the strategic decisions for the betterment of the enterprise.

Issues in Data Integration

While integrating the data we have to deal with several issues which are discussed below.

1. Entity Identification Problem

As we know the data is unified from the heterogeneous sources then how can we ‘match the real-world entities from the data’? For example, we have customer data from two different data source. An entity from one data source has customer_id and the entity from the other data source has customer_number. Now how the data analyst or the system does would understand that these two entities refer to the same attribute?

Well, here the **schema integration** can be achieved using metadata of each attribute. Metadata of an attribute incorporates its name, what does it mean in the particular scenario, what is its data

type, up to what range it can accept the value. What rules does the attribute follow for the null value, blank, or zero? Analyzing this metadata information will prevent error in schema integration.

Structural integration can be achieved by ensuring that the functional dependency of an attribute in the source system and its referential constraints matches the functional dependency and referential constraint of the same attribute in the target system.

This can be understood with the help of an example suppose in the one system, the discount would be applied to an entire order but in another system, the discount would be applied to every single item in the order. This difference must be caught before the data from these two sources are integrated into the target system.

2. Redundancy and Correlation Analysis

Redundancy is one of the big issues during data integration. Redundant data is an unimportant data or the data that is no longer needed. It can also arise due to attributes that could be derived using another attribute in the data set.

For example, one data set has the customer age and other data set has the customer's date of birth then age would be a redundant attribute as it could be derived using the date of birth.

Inconsistencies in the attribute also raise the level of redundancy. The redundancy can be discovered using correlation analysis. The attributes are analyzed to detect their interdependency on each other thereby detecting the correlation between them.

3. Tuple Duplication

Along with redundancies data integration has also deal with the duplicate tuples (e.g. where there are two or more identical tuples for a given unique data entry case). Duplicate tuples may come in the resultant data if the denormalized table has been used as a source for data integration. For example, if a purchase order database contains attributes for the purchaser's name and address instead of a key to this information in a purchaser database, discrepancies can occur, such as the same purchaser's name appearing with different addresses within the purchase order database.

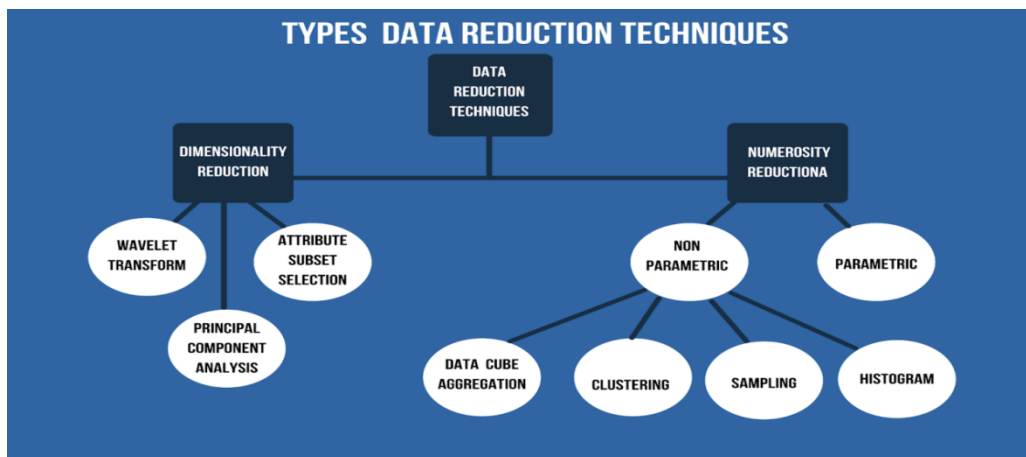
4. Data Conflict Detection and Resolution

Data conflict means the data merged from the different sources do not match. Like the attribute values may differ in different data sets for the same real world entity. The difference may be because they are represented differently in the different data sets. For suppose the price of a hotel room may be represented in different currencies in different cities. For a hotel chain, the *price* of rooms in different cities may involve not only different currencies but also different services (e.g., free breakfast) and taxes. This kind of issues is detected and resolved during data integration.

Data Reduction

Data reduction is the preprocessing technique that helps in obtaining reduced representation of dataset from the available dataset. It is the transformation of numerical or alphabetical digital information derived empirically or experimentally into a corrected, ordered, and simplified form. In simple terms, it simply means large amounts of data are cleaned, organized and categorized based on prerequisite criteria to help in driving business decisions. A database or data warehouse may store terabytes of data. So, it may take very long to perform data analysis and mining on such huge amounts of data. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume but yet produces the same (or almost the same) analytical results.

Data Reduction Techniques



There are two primary methods of Data Reduction, Dimensionality Reduction and Numerosity Reduction.

1. Dimensionality Reduction

Dimensionality Reduction is the process of reducing the number of dimensions the data is spread across. It means the attributes or features that the data set carries as the number of dimensions increases the sparsity. This sparsity is critical to clustering, outlier analysis and other algorithms. With reduced dimensionality, it is easy to visualize and manipulate data. There are three types of Dimensionality reduction.

(i) Wavelet Transform

Wavelet Transform is a lossy method for dimensionality reduction, where a data vector X is transformed into another vector X' , in such a way that both X and X' still represent the same length. The result of wavelet transform can be truncated, unlike its original, thus achieving dimensionality reduction. Wavelet transforms are well suited for data cube, sparse data or data which is highly skewed. Wavelet transform is often used in image compression. The discrete wavelet transform (DWT) is a linear signal processing technique. It transforms a vector into a numerically different vector (D to D') of wavelet coefficients. The two vectors are of the same length. However it is useful for compression in the sense that wavelet-transformed data can be truncated. A small compressed approximation of the data can be retained by storing only a small fraction of the strongest wavelet coefficient e.g., retain all wavelet coefficients larger than some particular threshold and the remaining coefficients are set to zero. The resulting data representation is sparse. Computations that can take advantage of sparsity are very fast if performed in wavelet space. Given a set of coefficients, an approximation of the original data can be got by applying the inverse DWT. The DWT is closely related to the discrete Fourier transform (DFT) a signal processing technique involving sine's and cosines. The general procedure for applying a discrete wavelet transform uses a hierarchical pyramid algorithm that halves the data in each iteration, resulting in fast computational speed. The method is as follows:

1. The length, L , of the input data vector must be an integer power of 2. This condition can be met by padding the data vector with zeros as necessary.

2. Each transform involves applying two functions. The first applies some data smoothing, such as sum or weighted average. The second performs a weighted difference, which acts to bring out the detailed features of the data.
3. The two functions are applied to pairs of input data, resulting in two sets of data of length $L/2$. In general these represent a smoothed or low frequency version of the input data and the high frequency content of it.
4. The two functions are recursively applied to sets of data obtained in the previous loop, until the resulting data sets obtained are of length 2.
5. A selection of values from the data sets obtained in the above iterations are designated the wavelet coefficients of the transformed data.

Wavelet transforms can be applied to multidimensional data such as data cubes. Wavelet transforms have many real world applications, including the compression of fingerprint images, computer vision, and analysis of time-series data and data cleaning.

(ii) Principal Component Analysis

This method involves the identification of a few independent tuples with 'n' attributes that can represent the entire data set. This method can be applied to skewed and sparse data. Let the data to be compressed consist of data vectors, from k dimensions. Principal Component Analysis or PCA searches for c k-dimensional orthogonal vectors that can be best used to represent the data, where $c \leq k$. The original data are thus projected onto a much smaller space. PCA can be used to perform dimensionality reduction. The basic procedure is as follows:

1. The input data are normalized, so that each attribute falls within the same range. This step helps ensure that attributes with large domains will not dominate attributes with smaller domains.
2. PCA computes c orthonormal vectors that provide a basis for the normalized input data. These are unit vectors that each point in a direction perpendicular to the others. These vectors are referred to as the principal components. The input data are a linear combination of principal components.

3. The principal components are sorted in order of decreasing “significance” or strength . The principal components essentially serve as a new set of axes for the data, by providing important information about variance.

4. Since the components are sorted according to decreasing order of “ significance”, eliminating the weaker components can reduce the size of the data. Using the strongest principal components it should be possible to reconstruct the original data to a good approximation.

PCA is computationally inexpensive and it can be ordered or unordered

(iii) **Attribute Subset Selection**

Here, attributes irrelevant to data mining or redundant ones are not included in a core attribute subset. The core attribute subset selection reduces the data volume and dimensionality. Attribute subset Selection is a technique which is used for data reduction in data mining process. Data reduction reduces the size of data so that it can be used for analysis purposes more efficiently.

Need of Attribute Subset Selection-

The data set may have a large number of attributes. But some of those attributes can be irrelevant or redundant. The goal of attribute subset selection is to find a minimum set of attributes such that dropping of those irrelevant attributes does not much affect the utility of data and the cost of data analysis could be reduced. Mining on a reduced data set also makes the discovered pattern easier to understand.

Process of Attribute Subset Selection-

The brute force approach can be very expensive in which each subset (2^n possible subsets) of the data having n attributes can be analysed.

The best way to do the task is to use the statistical significance tests such that best (or worst) attributes can be recognized. Statistical significance test assumes that attributes are independent of one another. This is a kind of greedy approach in which a significance level is decided (statistically ideal value of significance level is 5%) and the models are tested again and again until p-value (probability value) of all attributes is less than or equal to the selected significance level. The attributes having p-value higher than significance level are discarded.

This procedure is repeated again and again until all the attribute in data set has p-value less than or equal to the significance level. This gives us the reduced data set having no irrelevant attributes.

Methods of Attribute Subset Selection-

1. Stepwise Forward Selection.
2. Stepwise Backward Elimination.
3. Combination of Forward Selection and Backward Elimination.
4. Decision Tree Induction.

All the above methods are greedy approaches for attribute subset selection.

Stepwise Forward Selection: This procedure start with an empty set of attributes as the minimal set. The most relevant attributes are chosen (having minimum p-value) and are added to the minimal set. In each iteration, one attribute is added to a reduced set.

Stepwise Backward Elimination: Here all the attributes are considered in the initial set of attributes. In each iteration, one attribute is eliminated from the set of attributes whose p-value is higher than significance level.

Combination of Forward Selection and Backward Elimination: The stepwise forward selection and backward elimination are combined so as to select the relevant attributes most efficiently. This is the most common technique which is generally used for attribute selection.

Decision Tree Induction: This approach uses decision tree for attribute selection. It constructs a flow chart like structure having nodes denoting a test on an attribute. Each branch corresponds to the outcome of test and leaf nodes are a class prediction. The attribute that is not the part of tree is considered irrelevant and hence discarded.

2. Numerosity Reduction

This method uses alternate, small forms of data representation, thus reducing data volume. There are two types of Numerosity reduction: Parametric and Non-Parametric.

Parametric

This method assumes a model into which the data fits. Data model parameters are estimated, and only those parameters are stored, and the rest of the data is discarded. For example, a regression model can be used to achieve parametric reduction if the data fits the Linear Regression model.

Linear Regression models a linear relationship between two attributes of the data set. Let's say we need to fit a linear regression model between two attributes, x and y , where y is the dependent attribute, and x is the independent attribute or predictor attribute. The model can be represented by the equation $y = wx + b$, Where w and b are regression coefficients. A multiple linear regression model lets us express the attribute y in terms of multiple predictor attributes.

Another method, the Log-Linear model discovers the relationship between two or more discrete attributes. Assume, we have a set of tuples in n -dimensional space; the log-linear model helps to derive the probability of each tuple in this n -dimensional space.

Non-Parametric

A non-parametric numerosity reduction technique does not assume any model. The non-Parametric technique results in a more uniform reduction, irrespective of data size, but it may not achieve a high volume of data reduction like the parametric one. There are at least four types of Non-Parametric data reduction techniques: Histogram, Clustering, Sampling, Data Cube Aggregation, Data Compression.

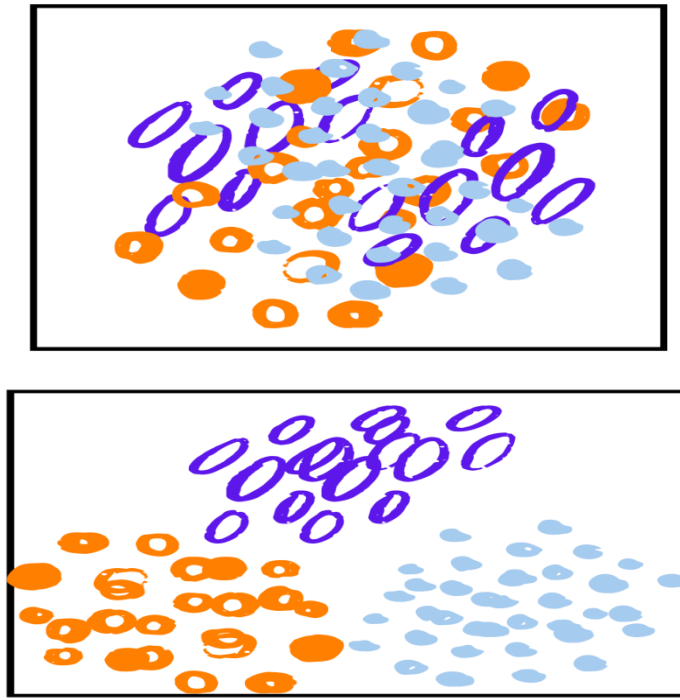
(i) Histogram

A histogram can be used to represent dense, sparse, skewed or uniform data, involving multiple attributes, effectively up to 5 together.

(ii) Clustering

In Clustering, the data set is replaced by the cluster representation, where the data is split between clusters depending on similarities to each other within-cluster and dissimilarities to other clusters. The more the similarity within-cluster, the closer they appear within the cluster.

The quality of the cluster depends on the maximum distance between any two data items in the cluster.



(iii) Sampling

Sampling is capable of reducing large data set into smaller sample data sets, reducing it to a representation of the original data set. There are four types of sampling data reduction methods.

- Simple Random Sample without Replacement of size s
- Simple Random Sample with Replacement of size s
- Cluster Sample
- Stratified Sample

(iv) Data Cube Aggregation

Data Cube Aggregation is a multidimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction. Data Cube

Aggregation, where the data cube is a much more efficient way of storing data, thus achieving data reduction, besides faster aggregation operations.

(v) Data Compression

It employs modification, encoding or converting the structure of data in a way that consumes less space. Data compression involves building a compact representation of information by removing redundancy and representing data in binary form. Data that can be restored successfully from its compressed form is called Lossless compression while the opposite where it is not possible to restore the original form from the compressed form is Lossy compression.

Data reduction achieves a reduction in volume, making it easy to represent and run data through advanced analytical algorithms. Data reduction also helps in the reduplication of data reducing the load on storage and the algorithms serving data science techniques downstream. It can be achieved in two principal ways. One by reducing the number of data records, or the features and the other by generating summary data and statistics at different levels.

Data Transformation

Data transformation in data mining is done for combining unstructured data with structured data to analyze it later. It is also important when the data is transferred to a new cloud data warehouse. When the data is homogeneous and well-structured, it is easier to analyze and look for patterns. For example, a company has acquired another firm and now has to consolidate all the business data. The smaller company may be using a different database than the parent firm. Also, the data in these databases may have unique IDs, keys and values. All this needs to be formatted so that all the records are similar and can be evaluated. Here are the steps involved:

1. Smoothing: It is a process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset. It helps in predicting the patterns. When collecting data, it can be manipulated to eliminate or reduce any variance or any other noise form.

The concept behind data smoothing is that it will be able to identify simple changes to help predict different trends and patterns. This serves as a help to analysts or traders who need to look at a lot of data which can often be difficult to digest for finding patterns that they wouldn't see otherwise.

2. Aggregation: Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description. This is a crucial step since the accuracy of data analysis insights is highly dependent on the quantity and quality of the data used. Gathering accurate data of high quality and a large enough quantity is necessary to produce relevant results.

The collection of data is useful for everything from decisions concerning financing or business strategy of the product, pricing, operations, and marketing strategies. For **example**, Sales, data may be aggregated to compute monthly& annual total amounts.

3. Discretization: It is a process of transforming continuous data into set of small intervals. Most Data Mining activities in the real world require continuous attributes. Yet many of the existing data mining frameworks are unable to handle these attributes. Also, even if a data mining task can manage a continuous attribute, it can significantly improve its efficiency by replacing a constant quality attribute with its discrete values. For **example**, (1-10, 11-20) (age:- young, middle age, senior).

4. Attribute Construction: In the attribute construction method, new attributes are created from an existing set of attributes. For example, in a dataset of employee information, the attributes can be employee name, employee ID and address. These attributes can be used to construct another dataset that contains information about the employees who have joined in the year 2019 only. This method of reconstruction makes mining more efficient and helps in creating new datasets quickly.

5. Generalization: In this process, low-level data attributes are transformed into high-level data attributes using concept hierarchies. This conversion from a lower level to a higher conceptual level is useful to get a clearer picture of the data. For example, age data can be in the form of (20, 30) in a dataset. It is transformed into a higher conceptual level into a categorical value (young, old). Another example, Categorical attributes, such as house addresses, may be generalized to higher-level definitions, such as town or country.

6. Normalization: Also called data pre-processing, this is one of the crucial techniques for data transformation in data mining. Here, the data is transformed so that it falls under a given range. When attributes are on different ranges or scales, data modelling and mining can be difficult. Normalization helps in applying data mining algorithms and extracting data faster.

The popular normalization methods are:

- Min-max normalization
- Decimal scaling
- Z-score normalization

Min-Max Normalization

What is easier to understand – the difference between 200 and 1000000 or the difference between 0.2 and 1? Indeed, when the difference between the minimum and maximum values is less, the data becomes more readable. The min-max normalization functions by converting a range of data into a scale that ranges from 0 to 1.

Min-Max Normalization Formula

$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

To understand the formula, here is an example. Suppose a company wants to decide on a promotion based on the years of work experience of its employees. So, it needs to analyze a database that looks like this:

Employee Name	Years of Experience
ABC	8
XYZ	20
PQR	10
MNO	15

The minimum value is 8

The maximum value is 20

As this formula scales the data between 0 and 1,

The new min is 0

The new max is 1

Here, V stands for the respective value of the attribute, i.e., 8, 10, 15, 20

After applying the min-max normalization formula, the following are the V' values for the attributes:

For 8 years of experience: $v' = 0$

For 10 years of experience: $v' = 0.16$

For 15 years of experience: $v' = 0.58$

For 20 years of experience: $v' = 1$

So, the min-max normalization can reduce big numbers to much smaller values. This makes it extremely easy to read the difference between the ranging numbers.

Decimal Scaling Normalization

Decimal scaling is another technique for normalization in data mining. It functions by converting a number to a decimal point.

Decimal Scaling Formula

$$v' = \frac{v}{10^J}.$$

Here:

V' is the new value after applying the decimal scaling

V is the respective value of the attribute

Now, integer J defines the movement of decimal points. So, how to define it? It is equal to the number of digits present in the maximum value in the data table. Here is an example:

Suppose a company wants to compare the salaries of the new joiners. Here are the data values:

Employee Name	Salary
ABC	10,000
XYZ	25,000
PQR	8,000
MNO	15,000

Now, look for the maximum value in the data. In this case, it is 25,000. Now count the number of digits in this value. In this case, it is '5'. So here 'j' is equal to 5, i.e 100,000. This means the V (value of the attribute) needs to be divided by 100,000 here.

After applying the zero decimal scaling formula, here are the new values:

Name	Salary	Salary after Decimal Scaling
ABC	10,000	0.1
XYZ	25, 000	0.25
PQR	8, 000	0.08
MNO	15,000	0.15

Thus, decimal scaling can tone down big numbers into easy to understand smaller decimal values. Also, data attributed to different units becomes easy to read and understand once it is converted into smaller decimal values.

Z-Score Normalization

Z-Score value is to understand how far the data point is from the mean. Technically, it measures the standard deviations below or above the mean. It ranges from -3 standard deviation up to +3 standard deviation. Z-score **normalization in data mining** is useful for those kinds of data analysis wherein there is a need to compare a value with respect to a mean (average) value, such as results from tests or surveys.

For example, a person's weight is 150 pounds. Now, if there is a need to compare that value with the average weight of a population listed in a vast table of data, Z-score normalization is needed to study such values, especially if someone's weight is recorded in kilograms.

Formula:

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

v' , v is the new and old of each entry in data respectively. σ_A , \bar{A} is the standard deviation and mean of A respectively.

For **example**:

Let mean of an attribute $P = 60,000$, Standard Deviation = 10,000, for the attribute P . Using z-score normalization, a value of 85000 for P can be transformed to:

$$\frac{85000 - 60000}{10000} = 2.50$$

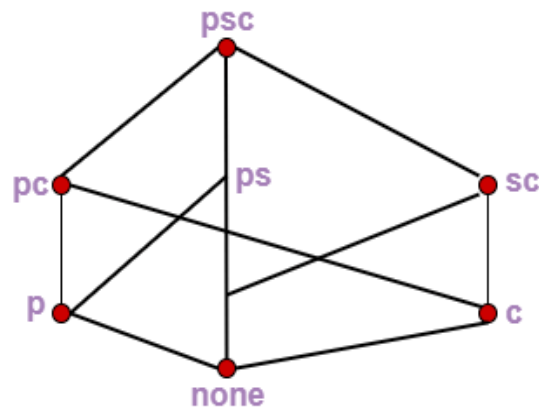
And hence we get the value of v' to be 2.5

Data Cube

When data is grouped or combined in multidimensional matrices called Data Cubes. The data cube method has a few alternative names or a few variants, such as "Multidimensional databases," "materialized views," and "OLAP (On-Line Analytical Processing)."

The general idea of this approach is to materialize certain expensive computations that are frequently inquired.

For example, a relation with the schema sales (part, supplier, customer, and sale-price) can be materialized into a set of eight views as shown in fig, where **psc** indicates a view consisting of aggregate function value (such as total-sales) computed by grouping three attributes part, supplier, and customer, **p** indicates a view composed of the corresponding aggregate function values calculated by grouping part alone, etc.



Eight views of data cubes for sales information.

A data cube is created from a subset of attributes in the database. Specific attributes are chosen to be measure attributes, i.e., the attributes whose values are of interest. Another attributes are selected as dimensions or functional attributes. The measure attributes are aggregated according to the dimensions.

For example, XYZ may create a sales data warehouse to keep records of the store's sales for the dimensions time, item, branch, and location. These dimensions enable the store to keep track of things like monthly sales of items, and the branches and locations at which the items were sold. Each dimension may have a table identify with it, known as a dimensional table, which describes the dimensions. For example, a dimension table for items may contain the attributes item_name, brand and type.

Data cube method is an interesting technique with many applications. Data cubes could be sparse in many cases because not every cell in each dimension may have corresponding data in the database.

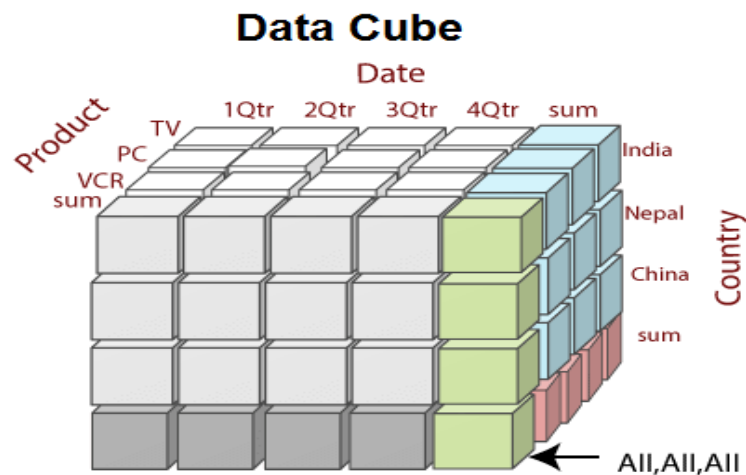
Techniques should be developed to handle sparse cubes efficiently.

If a query contains constants at even lower levels than those provided in a data cube, it is not clear how to make the best use of the precomputed results stored in the data cube.

The model view data in the form of a data cube. OLAP tools are based on the multidimensional data model. Data cubes usually model n-dimensional data.

A data cube enables data to be modeled and viewed in multiple dimensions. A multidimensional data model is organized around a central theme, like sales and transactions. A fact table represents this theme. Facts are numerical measures. Thus, the fact table contains measure (such as Rs_sold) and keys to each of the related dimensional tables.

Dimensions are a fact that defines a data cube. Facts are generally quantities, which are used for analyzing the relationship between dimensions.



Example: In the **2-D representation**, we will look at the All Electronics sales data for **items sold per quarter** in the city of Vancouver. The measured display in dollars sold (in thousands).

2-D view of Sales Data

location ="Vancouver"				
time (quarter)	item (type)			
	home entertainment	computer	phone	security
Q1	605	825	14	400
Q2	680	952	31	512
Q3	812	1023	30	501
Q3	927	1038	38	580

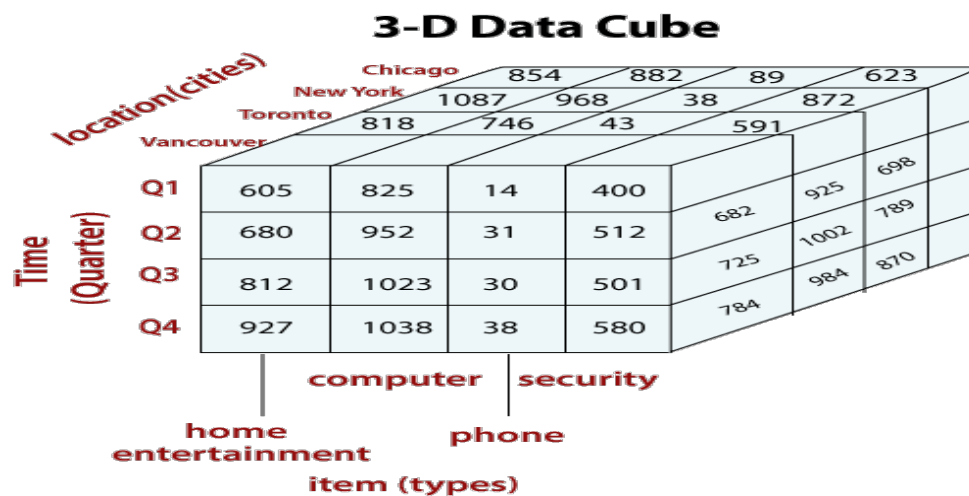
3-Dimensional Cuboids

Let suppose we would like to view the sales data with a third dimension. For example, suppose we would like to view the data according to time, item as well as the location for the cities Chicago, New York, Toronto, and Vancouver. The measured display in dollars sold (in thousands). These 3-D data are shown in the table. The 3-D data of the table are represented as a series of 2-D tables.

3-D view of Sales Data

location ="Chicago"					location ="New York"					location ="Toronto"				
item					item					item				
home					home					home				
time	ent.	comp.	phone	sec.	time	comp.	phone	sec.	time	ent.	comp.	phone	sec.	
Q1	854	882	89	623	1087	968	38	872	818	746	43	591		
Q2	943	890	64	698	1130	1024	41	925	894	769	52	682		
Q3	1032	924	59	789	1034	1048	45	1002	940	795	58	728		
Q4	1129	992	63	870	1142	1091	54	984	978	864	59	784		

Conceptually, we may represent the same data in the form of 3-D data cubes, as shown in fig:



Let us suppose that we would like to view our sales data with an additional fourth dimension, such as a supplier.

In data warehousing, the data cubes are n-dimensional. The cuboid which holds the lowest level of summarization is called a **base cuboid**.

For example, the **4-D cuboid** in the figure is the base cuboid for the given time, item, location, and supplier dimensions.

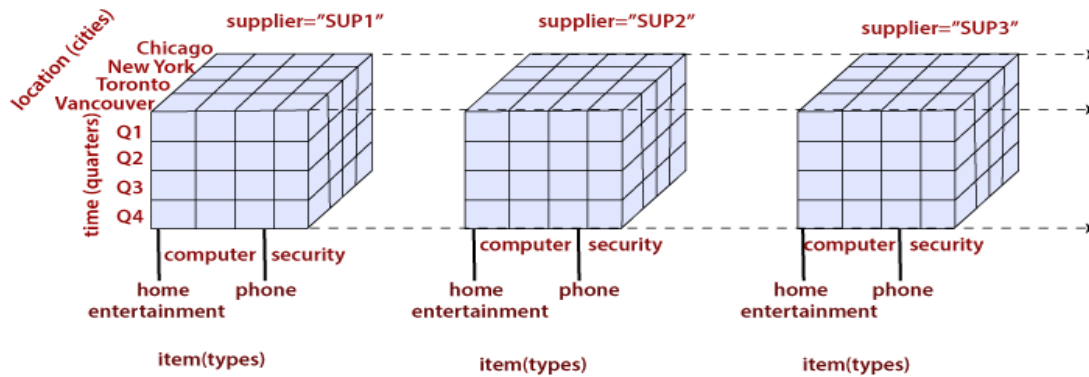
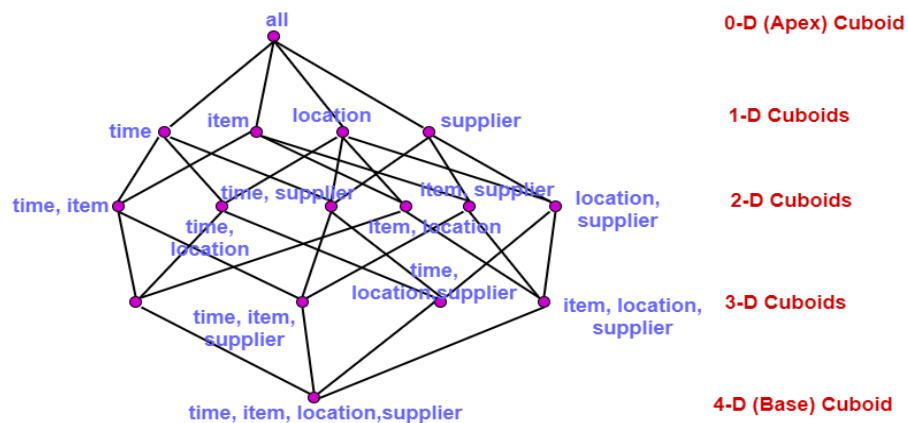


Figure is shown a **4-D data cube** representation of sales data, according to the dimensions time, item, location, and supplier. The measure displayed is dollars sold (in thousands).

The topmost **0-D cuboid**, which holds the highest level of summarization, is known as the apex cuboid. In this example, this is the total sales, or dollars sold, summarized over all four dimensions.

The lattice of cuboid forms a data cube. The figure shows the lattice of cuboids creating 4-D data cubes for the dimension time, item, location, and supplier. Each cuboid represents a different degree of summarization.



02

(336-029) Wk 49

DECEMBER

MONDAY

DECEMBER - 2019

M	T	W	T	F	S	S	M	T	W	T	F	S	S
						1	2	3	4	5	6	7	8
9	10	11	12	13	14	15	16	17	18	19	20	21	22
23	24	25	26	27	28	29	30	31					

Data preprocessing

It is done to improve the quality of data in data warehouse.

→ Increases efficiency

→ Ease of mining process

→ Removes Noisy Data, inconsistent data and incomplete data.

↳ data with missing values

Data cleaning

It cleans the data by filling in the missing values, smoothing noisy data, resolving the inconsistency (naming convention) and removing the outliers.

ways to Handle missing Data during cleaning.

i) Manual entry of missing data.

ii) Using attribute mean

iii) Using most probable value.

(Decision tree, Regression), we can determine the value. Basically we are predicting the value.

iv) Using Global constant
 (Suppose any value is missing then we put NA in place of it) or you can write unknown.

v) Ignore the tuple —

1	50
2	60
3	50
4	40

This value was unknown so we ignore the tuple.

04

(338-027) Wk 49

DECEMBER

WEDNESDAY

Noise in data \rightarrow It is a random error or variance in a measured variable.

Techniques to remove noise

1) Binning

eg - Consider the data
2, 10, 18, 18, 19, 20, 22, 25, 28

Step 1 Sort the data

2, 10, 18, 18, 19, 20, 22, 25, 28

Divide the data into bins / Buckets of range of values

2, 10, 18 \rightarrow 10
18, 19, 20 \rightarrow 19
22, 25, 28 \rightarrow 25

mean value

Bin size = 3

① Smoothing by BIN means

Value of bin is replaced by mean value (average)

10, 10, 10
 19, 19, 19
 25, 25, 25

② Bin Medians odd $\frac{n+1}{2}$
 even $\frac{n}{2}$

10 10 10
 19 19 19
 25 25 25

③ By bin boundaries \rightarrow max.
 \rightarrow min.
 All the bin values are replaced
 by closest max. or min value
 boundary

2, 2, 10
 10, 10, 20
 22, 22, 20

06

(340-025) Wk 49

DECEMBER

FRIDAY

 DECEMBER - 2016
 M T W T F S S M T W T F S S
 1 2 3 4 5 6 7 8
 9 10 11 12 13 14 15 16 17 18 19 20 21 22
 23 24 25 26 27 28 29 30 31

② Regression Data Mining
 technique which is used to fit
 an equation to a data set.

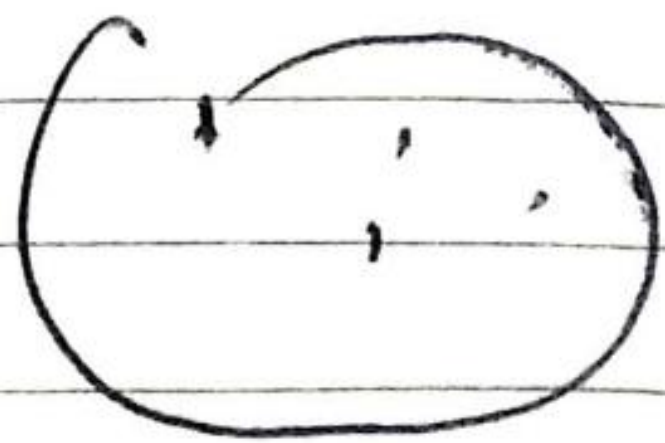
(Prediction)

Linear Regression $y = b + mx$
 Predicted value y given value x

[Data fitting function]

③ Clustering ↓

Groups (clusters) are formed
 from the data having similar
 values.



outliers removed