UNIT - I

Introduction: What Motivated Data Mining? Why Is It Important, Data Mining—On What Kind of Data, Data Mining Functionalities—What Kinds of Patterns Can Be Mined? Are All of the Patterns Interesting? Classification of Data Mining Systems, Data Mining Task Primitives, Integration of a Data Mining System with a Database or Data Warehouse System, Major Issues in Data Mining. (Han & Kamber)

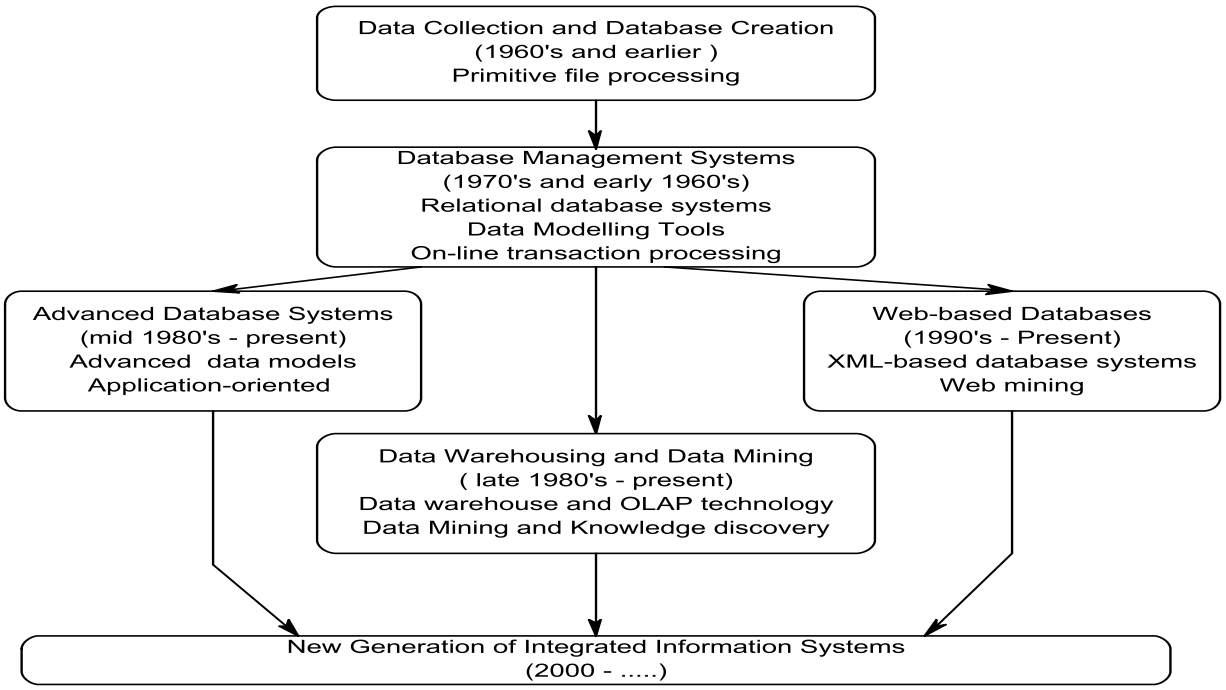
What Motivated Data Mining

Evolution of Database Technology:-

Necessity is the mother of invention.

The major reason for using data mining techniques is requirement of useful information and knowledge from huge amounts of data. The information and knowledge gained can be used in many applications such as business management, production control etc. Data mining came into existence as a result of the natural evolution of information technology. Evolutionary path in the database industry has developed the following functionalities. They are

1. Data collection and database creation.
2. Data management (including data storage and retrieval, and database transaction processing)
3. Data analysis and understanding (involving data warehousing and data mining).



During 1960's database and information technology has been evolving from primitive file processing systems to powerful database systems. During 1970's relational database systems were developed. In addition users access data through query languages. Efficient methods for on-line transaction processing (OLTP) were developed. During the mid-1980s many advanced database systems and application-

oriented database systems were developed. In 1990's Heterogeneous database systems and Internet-based global information systems such as the World-Wide Web (WWW) also emerged and play a vital role in the information industry.

Data can now be stored in many different types of databases. One database architecture that has recently emerged is the data warehouse. It is a repository of multiple heterogeneous data sources, organized under a unified schema to facilitate management decision making. Data warehouse technology includes data cleaning, data integration, and On-Line Analytical Processing (OLAP). Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis. The tremendous amount of data collected and stored in large and numerous databases, has lead to the development of data mining tools which perform data analysis to convert huge data into useful knowledge.

Fundamentals of data mining:-Data mining refers to extracting or "mining" knowledge from large amounts of data. Data mining can be named as "knowledge mining from data". But the name is quite lang. The shorter name "Knowledge mining" may not reflect the emphasis on mining from large amounts of data. Thus, such a misnomer which carries both "data" and "mining" became a popular choice.

Many people treat data mining as a synonym for another popularly used term, "Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery is a process which consists of an iterative sequence of steps. They are

Data cleaning and Preprocessing Stage:-Data cleaning is a process of removing unnecessary and inconsistent data from the databases. The main purpose of preprocessing is to improve the quality of the data by filling the missing values, configuring the data to make sure that it in consistent format.

Data Integration Stage: - In this stage multiple data sources may be combined (i.e integrated) to form a large database.

Data Selection Stage :- Data which is required for data mining process can be extracted from multiple and heterogeneous data sources such as databases, files etc., Data selection is a process where the appropriate data required for analysis is fetched from the databases.

Data Transformation and Reduction Stage: - In the transformation stage data extracted from multiple data sources are converted into an appropriate format for data mining process. Data reduction is used to decrease the number of possible values of data without affecting the integrity of data.

Data Mining Stage:-Data mining is an important process where expert techniques are applied to extract the hidden patterns from large volume of data stored in databases.

Pattern Evaluation:- In this stage, patterns generated during data mining stage are transformed into knowledge which is used in decision support system.

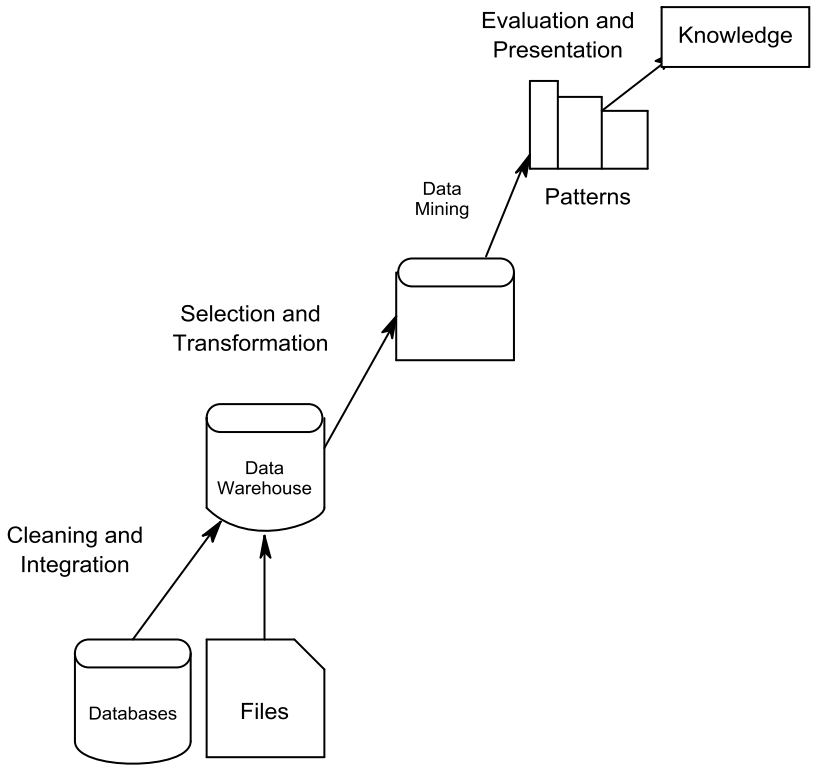


Fig:- Steps in the process of Knowledge Discovery

Knowledge presentation:- In this stage where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

Architecture of Data Mining System:- The architecture of a data mining system may have the following components

1. *Data Sources or Repositories:-* This component represents multiple data sources such as database, data warehouse, or any other information repository. Data cleaning and data integration techniques may be performed on the data.
2. *Database server or data warehouse server:-* The database or data warehouse server is responsible for fetching the relevant data, based on the user's data mining request.
3. *Knowledge base:-* . It is the area of knowledge that is used to guide the search, or to perform analysis of the resulting patterns.
4. *Data mining engine:-* This is core component to the data mining system and consists of a set of functional modules for tasks such as characterization, association analysis, classification, evolution and deviation analysis.
5. *Pattern evaluation module:-* This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns.

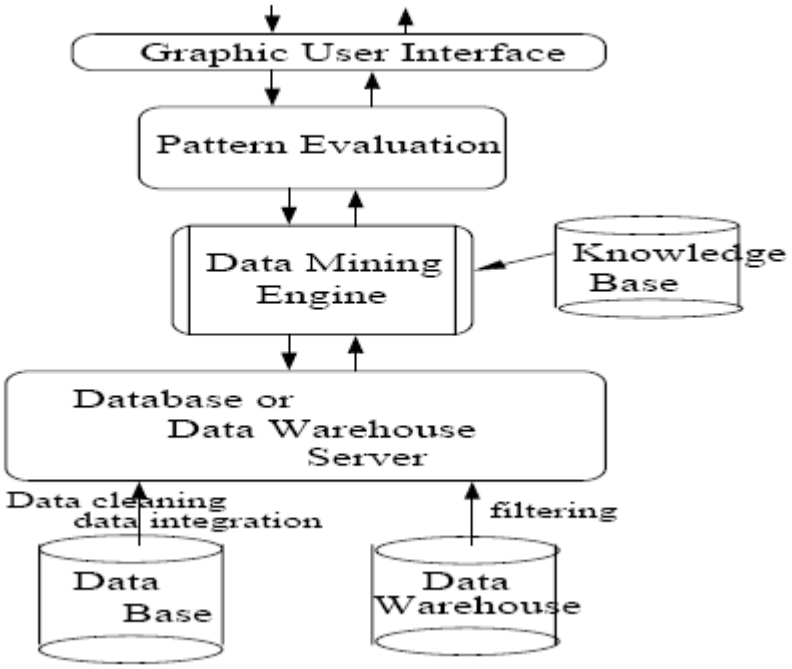


Fig:- Architecture of Data Mining System

6. *Graphical user interface*:- This module communicates between users and the data mining system, allowing the user to interact with the system by specifying a data mining query or task. This component allows the user to browse database and data warehouse schemas or data structures, evaluate mined patterns, and visualize the patterns in different forms.

Data Mining—On What Kind of Data: - There are number of different data stores on which mining can be performed. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World-Wide Web. Advanced database systems include object-oriented and object-relational databases, and specific application-oriented databases such as spatial databases, time-series databases, text databases, and multimedia databases.

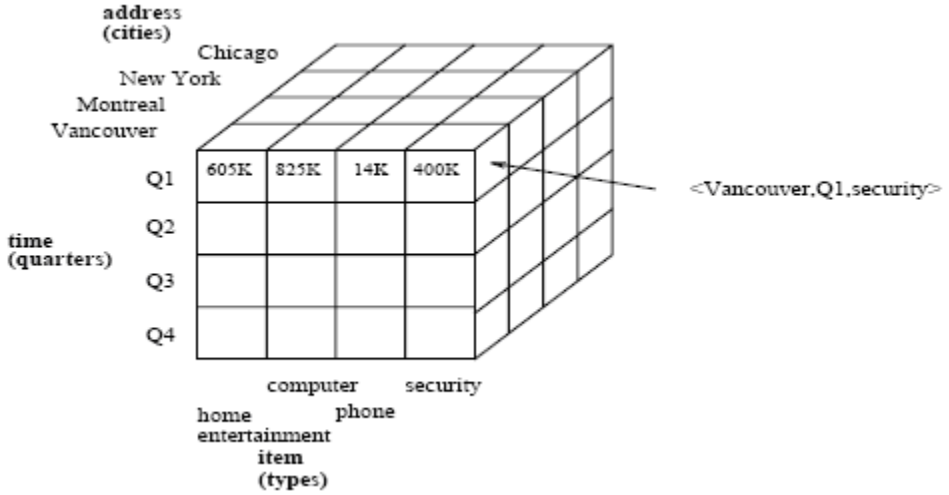
Relational databases:- A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. A relational database is a collection of tables, each of which is assigned a unique name. Each table consists of a set of attributes (columns or fields) and usually stores a large number of tuples (records or rows). Each tuple in a relational table represents an record identified by a unique key and described by a set of attribute values.

Ex:-

St_ID	St_Name	Address
101	Sai	Ramireddypeta
102	Meghana	Kukatpally

St_ID	Marks
101	90
102	85

Data warehouses:- A data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleansing, data transformation, data integration, data loading, and periodic data refreshing. A data



warehouse is usually modeled by a multidimensional database structure, where each dimension corresponds to an attribute or a set of attributes in the schema, and each cell stores the value of some aggregate measure, such as count or sales amount. The actual physical structure of a data warehouse may be a relational data store or a multidimensional data cube. It provides a multidimensional view of data and allows the precomputation and fast accessing of summarized data.

Transactional databases:- *A* transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans ID), and a list of the items making up the transaction. The

sales

trans_ID	list of item_ID's
T100	I1, I3, I8, I16
...	...

transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person, and of the branch at which the sale occurred, and so on.

Advanced database systems and advanced database applications:- Relational database systems have been widely used in business applications. With the advances of database technology, various kinds of advanced database systems have emerged and are undergoing development to address the requirements of new database applications. The new database applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), time-related data (such as historical records or stock exchange data), and the World-Wide Web (a huge, widely distributed information repository made

available by Internet). These applications require efficient data structures. In response to these needs, advanced database systems and specific application-oriented database systems have been developed.

Object-Relational Databases

Object-relational databases are constructed based on an object-relational data model. This model extends the relational model by providing a rich data type for handling complex objects and object orientation.

Conceptually, the object-relational data model inherits the essential concepts of object-oriented databases, where, in general terms, each entity is considered as an object.

Temporal Databases, Sequence Databases, and Time-Series Databases

A temporal database typically stores relational data that include time-related attributes. These attributes may involve several timestamps, each having different semantics.

A sequence database stores sequences of ordered events, with or without a concrete notion of time. Examples include customer shopping sequences, Web click streams, and biological sequences.

A time-series database stores sequences of values or events obtained over repeated measurements of time (e.g., hourly, daily, weekly). Examples include data collected from the stock exchange, inventory control, and the observation of natural phenomena (like temperature and wind).

These include object-oriented and object-relational database systems, spatial database systems, temporal and time-series database systems, text and multimedia database systems etc.,

Spatial Databases and Spatiotemporal Databases

Spatial databases contain spatial-related information. Examples include geographic (map) databases, very large-scale integration (VLSI) or computed-aided design databases, and medical and satellite image databases. Spatial data may be represented in raster format, consisting of n-dimensional bit maps or pixel maps.

A spatial database that stores spatial objects that change with time is called a spatiotemporal database, from which interesting information can be mined. For example, we may be able to group the trends of moving objects and identify some strangely moving vehicles

Text Databases and Multimedia Databases

Text databases are databases that contain word descriptions for objects. Text databases may be highly unstructured (such as some Web pages on the World Wide Web). Some text databases may be somewhat structured, that is, semi structured such as e-mail messages and many HTML/XML Web pages), whereas others are relatively well structured (such as library catalogue databases).

Multimedia databases store image, audio, and video data. They are used in applications such as picture content-based retrieval, voice-mail systems, video-on-demand systems, the World Wide Web, and speech-based user interfaces that recognize spoken commands.

Heterogeneous Databases and Legacy Databases

A heterogeneous database consists of a set of interconnected, autonomous component databases. The components communicate in order to exchange information and answer queries.

A legacy database is a group of heterogeneous databases that combines different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases, or file systems. The heterogeneous databases in a legacy database may be connected by intra or inter-computer networks.

Data Streams

Many applications involve the generation and analysis of a new kind of data, called stream data, where data flow in and out of an observation platform (or window) dynamically.

TheWorld WideWeb

The World Wide Web and its associated distributed information services, such as Yahoo!, Google, America Online, and AltaVista, provide rich, worldwide, on-line information services, where data objects are linked together to facilitate interactive access.

Data Mining Functionalities—What Kinds of Patterns Can Be Mined?:- Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

Concept/Class Description: Data characterization and discrimination:- Data can be associated with classes or concepts. Data characterization refers to summarizing the data of the class under study (often called the target class) in general terms. Data discrimination refers to comparison of the target class with one or a set of comparative classes (often called the contrasting classes).

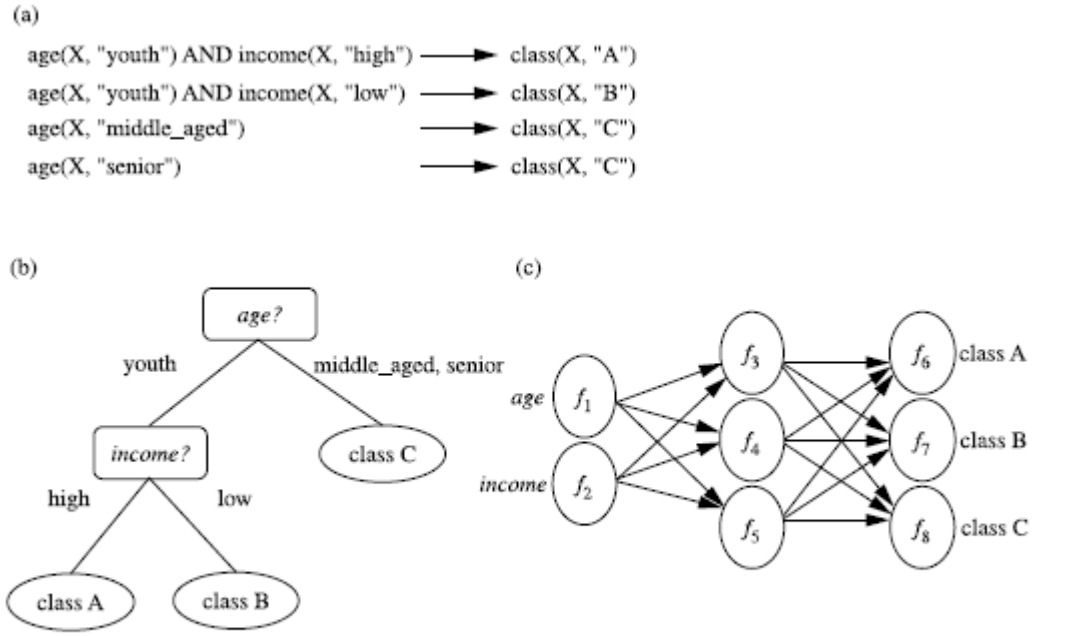
Mining Frequent Patterns, Associations, and Correlations:- Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

Ex:- The association rules may be specified as

age(X; “20…29") ^ income(X; “20K…30K") => buys(X, “CD player")

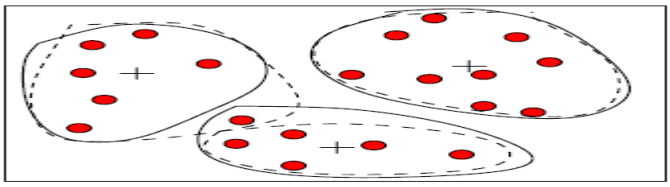
[support = 2%; confidence = 60%]

Classification and prediction:- Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known). Classification can be used for predicting the class label of data objects. Prediction may refer to both data value prediction and class label prediction, it is usually confined to data value prediction and thus is distinct from classification.



A classification model can be represented in various forms, such as (a) IF-THEN rules, (b) a decision tree, or a (c) neural network.

Clustering analysis:- Clustering analyzes data objects without consulting a known class label. Clustering can be used to generate class labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived.



Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of extremely large amounts for a given account number in comparison to regular charges incurred by the same account.

Evolution and deviation analysis:- Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-

series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

Are All of the Patterns Interesting?

A data mining system has the potential to generate thousands or even millions of patterns, or rules. "So," you may ask, "are all of the patterns interesting?" Typically not—only a small fraction of the patterns potentially generated would actually be of interest to any given user. This raises some serious questions for data mining.

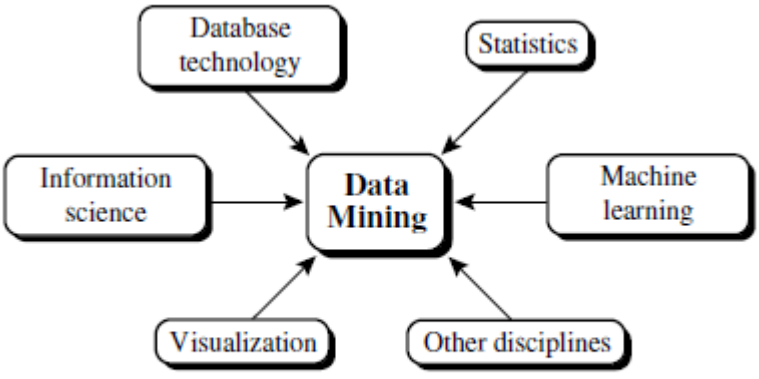
1. "What makes a pattern interesting?
2. Can a data mining system generate all of the interesting patterns?
3. Can a data mining system generate only interesting patterns?"

To answer the first question, a pattern is interesting if it is
(1) Easily understood by humans
(2) Valid on new or test data with some degree of certainty
(3) Potentially useful and
(4) Novel.

The second question—"Can a data mining system generate all of the interesting patterns?" refers to the completeness of a data mining algorithm. It is often unrealistic and inefficient for data mining systems to generate all of the possible patterns. Instead, user-provided constraints and interestingness measures should be used to focus the search.

Finally, the third question—"Can a data mining system generate only interesting patterns?" is an optimization problem in data mining. It is highly desirable for data mining systems to generate only interesting patterns.

Classification of data mining systems:- Data mining is an interdisciplinary field, including database systems, statistics, machine learning, visualization, and information science. Data mining systems can be categorized according to various criteria, as follows.



Data mining as a confluence of multiple disciplines.

1. ***Classification according to the kinds of databases mined:-*** A data mining system can be classified according to the kinds of databases mined. Data mining systems can be classified according to data models such as relational, transactional, object-oriented, object-relational, or data warehouse mining

system. If classifying according to the special types of data handled, we may have a spatial, time-series, text, or multimedia data mining system, or a World-Wide Web mining system. Other system types include heterogeneous data mining systems, and legacy data mining systems.

2. ***Classification according to the kinds of knowledge mined:-***Data mining systems can be categorized according to the kinds of knowledge they mine, i.e., based on data mining functionalities, such as characterization, discrimination, association, classification, clustering, trend and evolution analysis, deviation analysis, similarity analysis, etc.
3. ***Classification according to the kinds of techniques utilized:-*** Data mining systems can also be categorized according to the underlying data mining techniques employed. These techniques can be described according to the degree of user interaction involved, or the methods of data analysis employed. A sophisticated data mining system will often adopt multiple data mining techniques .
4. ***Classification according to the applications adapted:-***The classification of data mining system can be done based on the type of applications they use. Some of the applications where data mining systems are used include finance , e-mail, stock-markets etc.

Data Mining Task Primitives:-

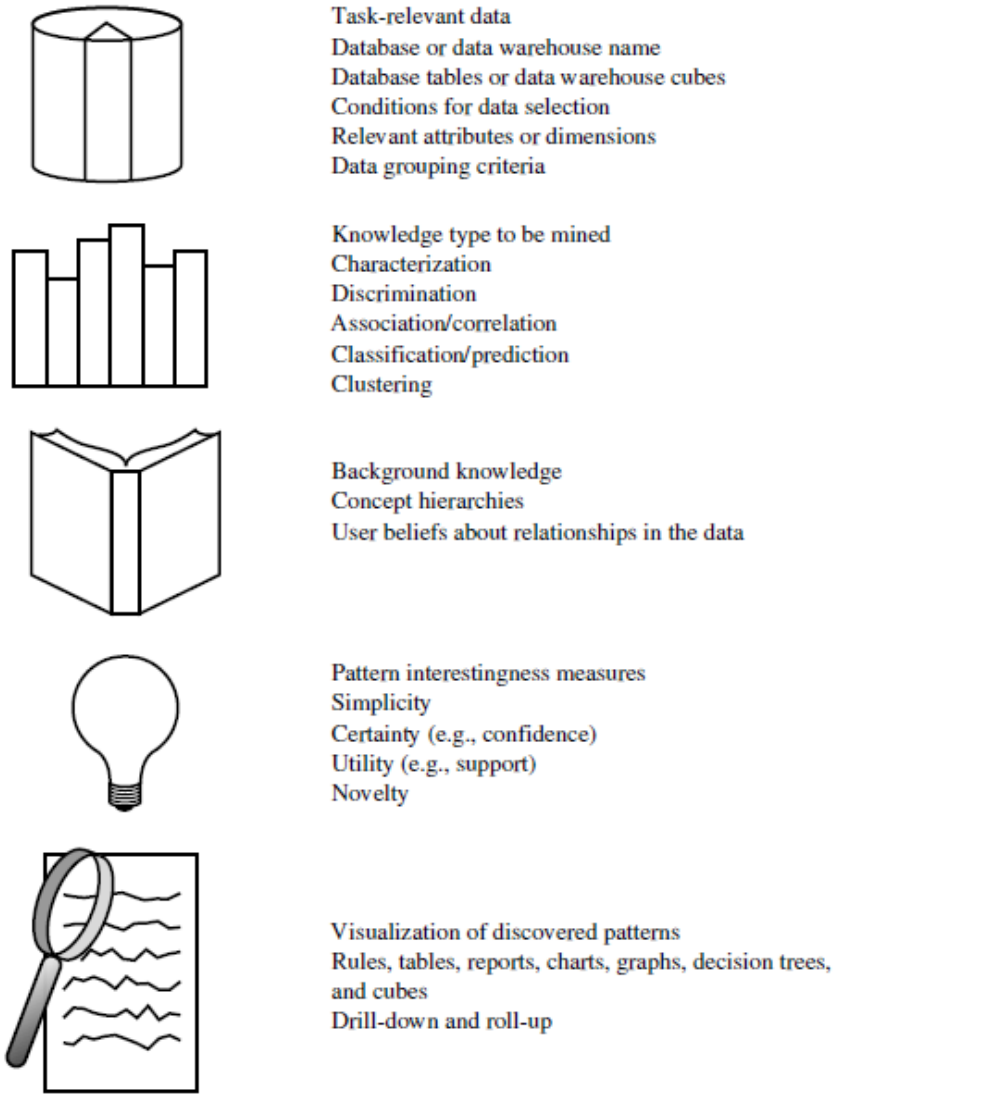
A data mining task can be specified in the form of a data mining query, which is input to the data mining system. A data mining query is defined in terms of data mining task primitives. These primitives allow the user to interactively communicate with the data mining system during discovery in order to direct the mining process, or examine the findings from different angles or depths. The data mining primitives specify the following, as illustrated in following Figure.

The set of task-relevant data to be mined: This specifies the portions of the database or the set of data in which the user is interested. This includes the database attributes or data warehouse dimensions of interest (referred to as the relevant attributes or dimensions).

The kind of knowledge to be mined: This specifies the data mining functions to be performed, such as characterization, discrimination, association or correlation analysis, classification, prediction, clustering, outlier analysis, or evolution analysis.

The background knowledge to be used in the discovery process: This knowledge about the domain to be mined is useful for guiding the knowledge discovery process and for evaluating the patterns found. Concept hierarchies are a popular form of background knowledge, which allow data to be mined at multiple levels of abstraction.

The interestingness measures and thresholds for pattern evaluation: They may be used to guide the mining process or, after discovery, to evaluate the discovered patterns. Different kinds of knowledge may have different interestingness measures. For example, interestingness measures for association rules include support and confidence.



Primitives for specifying a data mining task.

The expected representation for visualizing the discovered patterns: This refers to the form in which discovered patterns are to be displayed, which may include rules, tables, charts, graphs, decision trees, and cubes.

A data mining query language can be designed to incorporate these primitives, allowing users to flexibly interact with data mining systems.

Mining classification rules:. Suppose, as a marketing manager of AllElectronics, you would like to classify customers based on their buying patterns. You are especially interested in those customers whose salary is no less than $40,000, and who have bought more than $1,000 worth of items, each of which is priced at no less than $100. In particular, you are interested in the customer's age, income, the types of items purchased, the purchase location, and where the items were made. You would like to view the resulting classification in the form of rules. This data mining query is expressed in DMQL3 as follows, where each line of the query has been enumerated to aid in our discussion.

(1) use database AllElectronics db

(2) use hierarchy location hierarchy for T.branch, age hierarchy for C.age
(3) mine classification as promising customers
(4) in relevance to C.age, C.income, I.type, I.place made, T.branch
(5) from customer C, item I, transaction T
(6) where I.item ID = T.item ID and C.cust ID = T.cust ID
and C.income >= 40,000 and I.price >= 100
(7) group by T.cust ID

Integration of a Data Mining System with a Database or DataWarehouse System:-

A good system architecture will facilitate the data mining system to make best use of the software environment, accomplish data mining tasks in an efficient and timely manner, interoperate and exchange information with other information systems, be adaptable to users' diverse requirements, and evolve with time. To integrate or couple the DM system with a database (DB) system and/or a data warehouse (DW) system possible integration schemes include no coupling, loose coupling, semitight coupling, and tight coupling.

No coupling: No coupling means that a DM system will not utilize any function of a DB or DW system. It may fetch data from a particular source (such as a file system), process data using some data mining algorithms, and then store the mining results in another file.

Such a system, though simple, suffers from several drawbacks.

First, a DB system provides a great deal of flexibility and efficiency at storing, organizing, accessing, and processing data. Without using a DB/DW system, a DM system may spend a substantial amount of time finding, collecting, cleaning, and transforming data. In DB and/or DW systems, data tend to be well organized, indexed, cleaned, integrated, or consolidated, so that finding the task-relevant, high-quality data becomes an easy task.

Most data have been or will be stored in DB/DW systems. Without any coupling of such systems, a DM system will need to use other tools to extract data, making it difficult to integrate such a system into an information processing environment. Thus, no coupling represents a poor design.

Loose coupling: Loose coupling means that a DM system will use some facilities of a DB or DW system, fetching data from a data repository managed by these systems, performing data mining, and then storing the mining results either in a file or in a designated place in a database or data warehouse.

Loose coupling is better than no coupling because it can fetch any portion of data stored in databases or data warehouses by using query processing, indexing, and other system facilities.

However, many loosely coupled mining systems are main memory-based. So, it is difficult for loose coupling to achieve high scalability and good performance with large data sets.

Semi tight coupling: Semi tight coupling means that besides linking a DM system to a DB/DW system, efficient implementations of a few essential data mining primitives (identified by the analysis of frequently encountered data mining functions) can be provided in the DB/DW system. These primitives can include sorting, indexing, aggregation, histogram analysis, multi way join, and precomputation of some essential statistical measures, such as sum, count, max, min, standard deviation, and so on. Moreover, some frequently used intermediate mining results can be precomputed and stored in the DB/DW system. Because these intermediate mining results are either precomputed or can be computed efficiently, this design will enhance the performance of a DM system.

Tight coupling: Tight coupling means that a DM system is smoothly integrated into the DB/DW system. The data mining subsystem is treated as one functional component of an information system. Data mining queries and functions are optimized based on mining query analysis, data structures, indexing schemes, and query processing methods of a DB or DW system. With further technology advances, DM, DB, and DW systems will evolve and integrate together as one information system with multiple functionalities. This will provide a uniform information processing environment.

This approach is highly desirable because it facilitates efficient implementations of data mining functions, high system performance, and an integrated information processing environment.

Major Issues in Data Mining:- Major issues in data mining regarding mining methodology, user interaction, performance, and diverse data types. These issues are introduced below:

Mining methodology and user-interaction issues:- These react the kinds of knowledge mined, the ability to mine knowledge at multiple granularities, the use of domain knowledge, ad-hoc mining, and knowledge visualization.

i) **Mining different kinds of knowledge in databases:** Because different users can be interested in different kinds of knowledge, data mining should cover a wide spectrum of data analysis and knowledge discovery tasks, including data characterization, discrimination, association and correlation analysis, classification, prediction, clustering, outlier analysis, and evolution analysis

ii) **Interactive mining of knowledge at multiple levels of abstraction:** Interactive mining allows users to focus the search for patterns, providing and refining data mining requests based on returned results. Incorporation of background knowledge: Background knowledge, or information regarding the domain under study, may be used to guide the discovery process and allow discovered patterns to be expressed in concise terms and at different levels of abstraction.

iii) **Data mining query languages and ad hoc data mining:** Relational query languages (such as SQL) allow users to pose ad hoc queries for data retrieval.

iv) **Presentation and visualization of data mining results:** Discovered knowledge should be expressed in high-level languages, visual

representations, or other expressive forms so that the knowledge can be easily understood and directly usable by humans.

v) **Handling noisy or incomplete data:** The data stored in a database may reflect noise, exceptional cases, or incomplete data objects.

vi) **Pattern evaluation—the interestingness problem:** A data mining systemcan uncover thousands of patterns. Many of the patterns discovered may be uninteresting to the given user, either because they represent common knowledge or lack novelty.

Performance issues:- These include efficiency, scalability, and parallelization of data mining algorithms.

i) **Efficiency and scalability of data mining algorithms:** To effectively extract information from a huge amount of data in databases, data mining algorithms must be efficient and scalable. In other words, the running time of a data mining algorithm must be predictable and acceptable in large databases.

ii) **Parallel, distributed, and incremental mining algorithms:** The huge size of many databases, the wide distribution of data, and the computational complexity of some data mining methods are factors motivating the development of parallel and distributed data mining algorithms. Such algorithms divide the data into partitions, which are processed in parallel. The results from the partitions are then merged. Moreover, the high cost of some data mining processes promotes the need for incremental data mining algorithms that incorporate database updates without having to mine the entire data again "from scratch."

Issues relating to the diversity of database types:- These include Handling of relational and complex types of data.

i) **Handling of relational and complex types of data:** Because relational databases and data warehouses are widely used, the development of efficient and effective data mining systems for such data is important. However, other databases may contain complex data objects, hypertext and multimedia data, spatial data, temporal data, or transaction data. It is unrealistic to expect one system to mine all kinds of data, given the diversity of data types and different goals of data mining. Specific data mining systems should be constructed for mining specific kinds of data.

ii) **Mining information from heterogeneous databases and global information systems:** Local- and wide-area computer networks (such as the Internet) connect many sources of data, forming huge, distributed, and heterogeneous databases. The discovery of knowledge from different sources of structured, semi structured, or unstructured data with diverse data semantics poses great challenges to data mining.