Motivation for data mining

Necessity is the mother of invention. —Plato

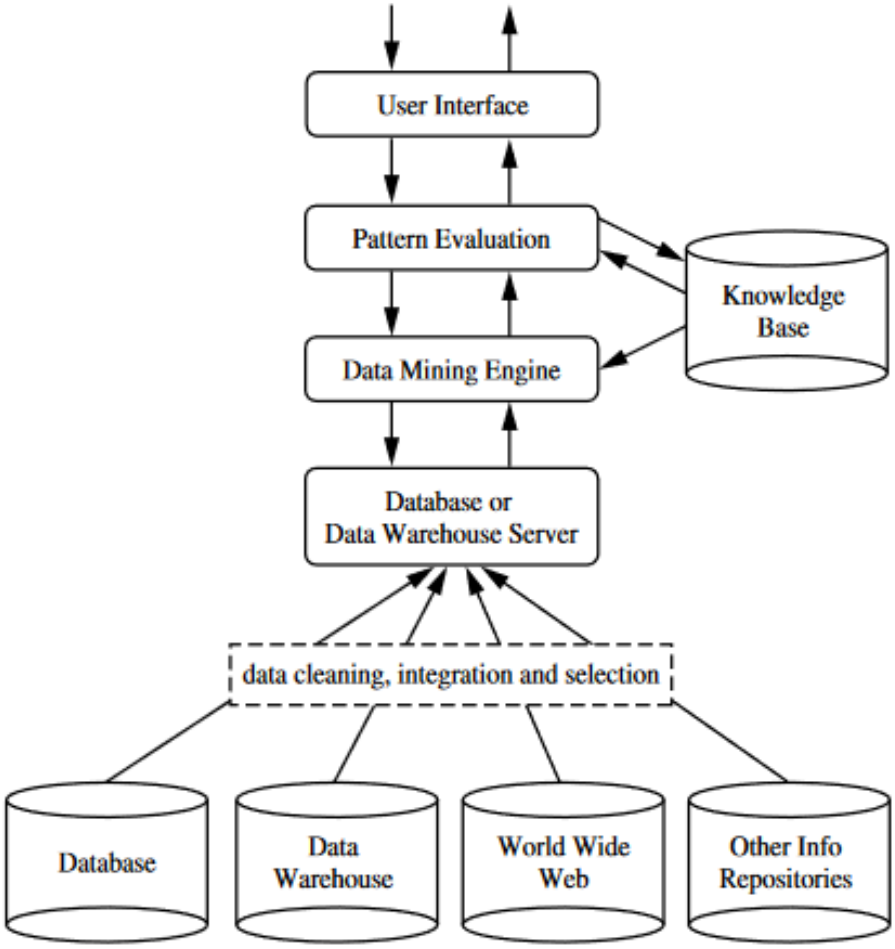
- The Explosive Growth of Data: from terabytes to petabytes
- We are drowning in data, but starving for knowledge
- Data Mining - Automated analysis of massive data sets

Introduction to data mining system

The process of Discovering meaningful patterns & trends often previously unknown, from large amount of data, using pattern recognition, statistical and mathematical techniques. In other words, data mining refers to extracting or "mining" knowledge from large amounts of data.

Architecture of a Data mining System

The architecture of a typical data mining system has six major components:



Database, data warehouse, World Wide Web, or other information repository:

This is one or a set of databases, data warehouses, spreadsheets, or other kinds of information repositories. Data cleaning and data integration techniques may be performed on the data.

Database 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.

Knowledge base:

This is the domain knowledge that is used to **guide the search or evaluate the interestingness of resulting patterns**. Such knowledge can include **concept hierarchies,** used to organize **attributes or attribute values** into different levels of abstraction. Knowledge such as user beliefs, pattern's interestingness (based on its unexpectedness), thresholds, and metadata (e.g., describing data from multiple heterogeneous sources) may also be included.

Data mining engine:

This is essential to the data mining system and ideally consists of a set of functional modules for tasks such as **characterization, association and correlation analysis, classification, prediction, cluster analysis, outlier analysis, and evolution analysis.**

Pattern evaluation module:

This component typically employs **interestingness measures** and interacts with the data mining modules so as to focus the search toward interesting patterns. It may use **interestingness thresholds to filter out discovered patterns**. For efficient data mining, it is highly recommended to push the evaluation of **pattern interestingness as deep as possible** into the mining process so as to confine the search to only the interesting patterns.

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, providing information to help focus the search, and performing exploratory data mining based on the intermediate data mining results. In addition, 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 functionalities

Data mining system has wide range of functions to perform. In the baseline, those functionalities can be categorized as:

1. Class/Concept Description: Characterization and Discrimination

Data is associated with classes or concepts so they can be correlated with results.

a. *Data characterization*

When you summarize the general features of the data, it is called data characterization. Attribute-oriented induction technique is also used to generalize or characterize the data with minimal user interaction.

b. *Data discrimination*

Data discrimination is one of the functionalities of data mining. It compares the data between the two classes. Generally, it maps the target class with a predefined group or class. It compares and contrasts the characteristics of the class with the predefined class using a set of rules called discriminant rules.

2. Classification

Classification is probably one of the most important data mining functionalities. It uses data models to predict the trends in data. For example, the spending chart our internet banking or mobile application shows based on our spend patterns. This is sometimes used to define our risk of getting a new loan. IF-THEN rules, Decision Trees and Neural Networks are widely used for classification purposes.

3. Prediction

Prediction data mining functionality finds the missing numeric values in the data. It uses regression analysis to find the unavailable data. If the class label is missing, then the prediction is done using classification. Prediction is popular because of its importance in business intelligence. There are two ways one can predict data:

1. Predicting the unavailable or missing data using prediction analysis
2. Predicting the class label using the previously built class model.

4. Association Analysis

Association Analysis is a functionality of data mining. It relates two or more attributes of the data. It discovers the relationship between the data and the rules that are binding them. It finds its application widely in retail sales. For example, that is if mobile phones are bought with headphones: support is 2% and confidence is 40%. This means that 2% of the time that customers bought mobile phones with headphones. 40% of confidence is the probability of the same association happening again.

5. Cluster Analysis

Unsupervised classification is called cluster analysis. It is similar to the classification functionality of data mining where the data are grouped. Unlike classification, in cluster analysis, the class label is unknown. Data are grouped based on clustering algorithms. K-means clustering algorithm**, K-medoid clustering, Mean-shift algorithm, Gaussian Mixture Modal** are some of the methods used to for cluster analysis.

6. Outlier (Deviants, Abnormalities, Discordant, Anomalies) Analysis

When data that cannot be grouped in any of the class appears, we use outlier analysis. There will be occurrences of data that will have different attributes to any of the other classes or general models. These outstanding data are called outliers. They are usually considered noise or exceptions, and the analysis of these outliers is called outlier mining. For example: a potential credit card fraudulent activity.

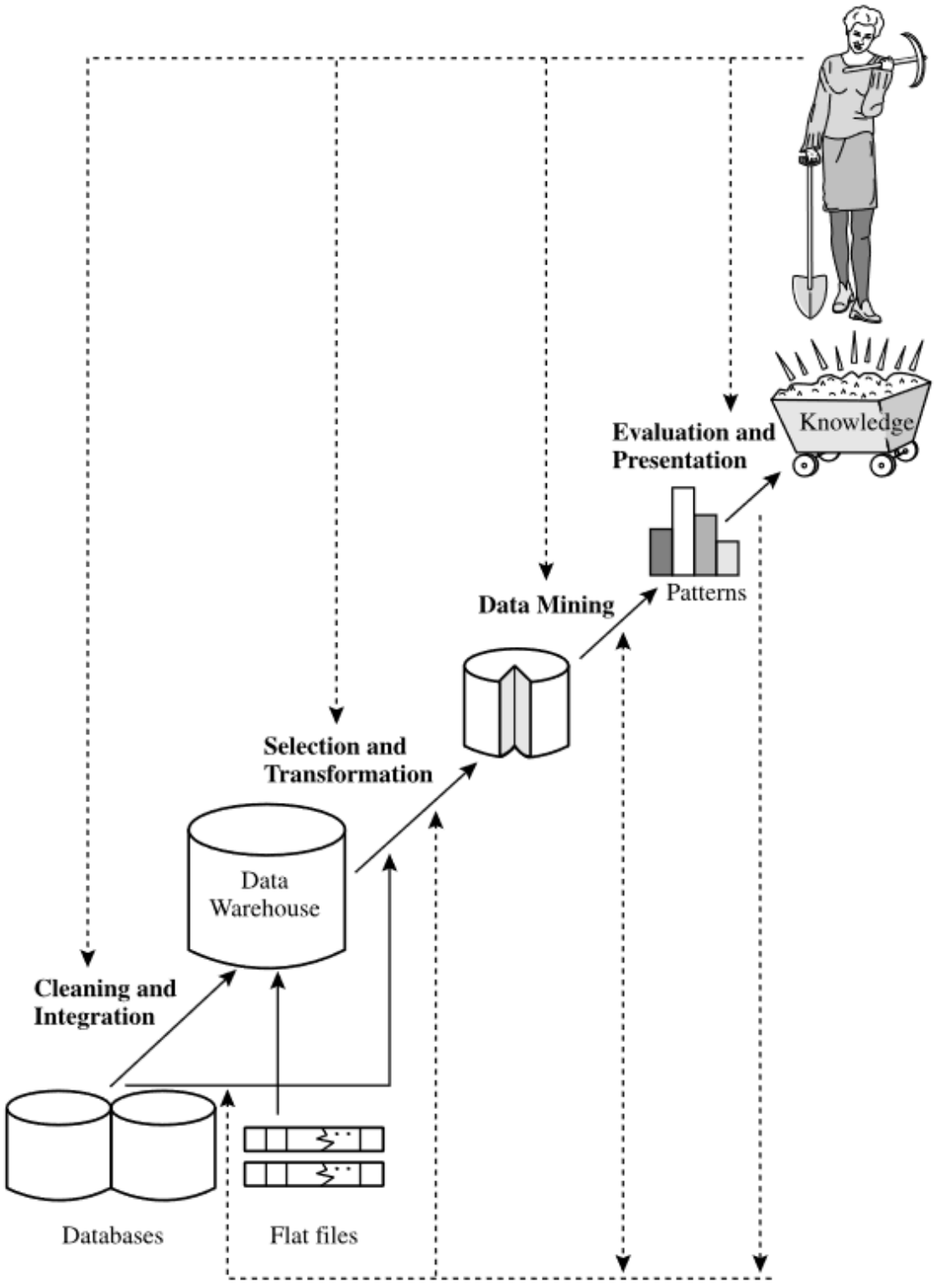
7. Evolution & Deviation Analysis

With evolution analysis being another data mining functionalities in data mining, we get time-related clustering of data. We can find trends and changes in behavior over a period. We can find features like time-series data, periodicity, and similarity in trends with such distinct analysis.

KDD

Data mining is often treated as a synonym for another popularly used term, Knowledge Discovery from Data, or KDD. KDD is an iterative sequence of the steps:

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



Data object and attribute types

Data sets are made up of data objects. A **data object** represents an entity. For example:

- In a sales database, the objects may be customers, store items, and sales
- In a medical database, the objects may be patients
- In a university database, the objects may be students, professors, and courses.

Data objects are typically described by attributes. Attribute is a data field, representing a characteristic or feature of a data object. For example: CUSTOMER_ID, name, address.

There are various kinds of Attributes:

- Nominal
- Binary
- Ordinal
- Numeric
    - Interval
    - Ratio

- Nominal

categories, states, or “names of things”

Hair_color = {*auburn, black, blond, brown, grey, red, white*}

marital status, occupation, ID numbers, zip codes

- Binary

Nominal attribute with only 2 states (0 and 1)

Symmetric binary: both outcomes equally important

e.g., gender

Asymmetric binary: outcomes not equally important.

e.g., medical test (positive vs. negative)

Convention: assign 1 to most important outcome (e.g., HIV positive)

- Ordinal:

Values have a meaningful order (ranking) but magnitude between successive values is not known.

Size = {*small, medium, large*}, grades, army rankings

- Numeric: quantitative

Quantity (integer or real-valued)

- Interval-scaled

  Measured on a scale of **equal-sized units**

  Values have order

  E.g., *calendar dates*

  No true zero-point

- Ratio-scaled

  Inherent **zero-point**

  We can speak of values as being an order of magnitude larger than the unit of measurement (10m is twice as long as 5m).

  e.g., *length, counts, monetary quantities*

Discrete vs Continuous Attributes

Discrete Attribute

Has only a finite or countably infinite set of values

E.g., zip codes, profession, or the set of words in a collection of documents

Sometimes, represented as integer variables

Note: Binary attributes are a special case of discrete attributes

Continuous Attribute

Has real numbers as attribute values

E.g., temperature, height, or weight

Practically, real values can only be measured and represented using a finite number of digits

Continuous attributes are typically represented as floating-point variables

Statistical description of data

Mean (algebraic measure) (sample vs. population):

Note: *n* is sample size and *N* is population size. $\bar{x} = \frac{1}{n} \Sigma_{i=1}^{n} x_i$, $\mu = \frac{\Sigma x}{N}$

Weighted arithmetic mean: $\bar{x} = \frac{1}{n} \frac{(\Sigma_{i=1}^{n} w_i x_i)}{(\Sigma_{i=1}^{n} w_i)}$

Trimmed mean: chopping extreme values

- Median:

Middle value if odd number of values, or average of the middle two values otherwise

Estimated by interpolation (for *grouped data*):

$$median = L_1 + \frac{\frac{n}{2} - (\Sigma\, freq) * l}{freq_{median}} * width$$

- Mode

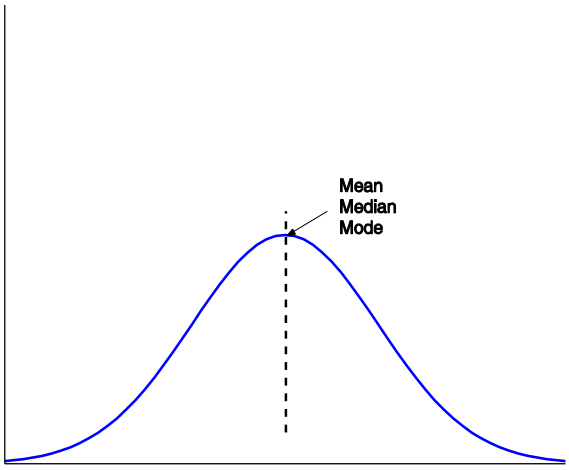
Value that occurs most frequently in the data

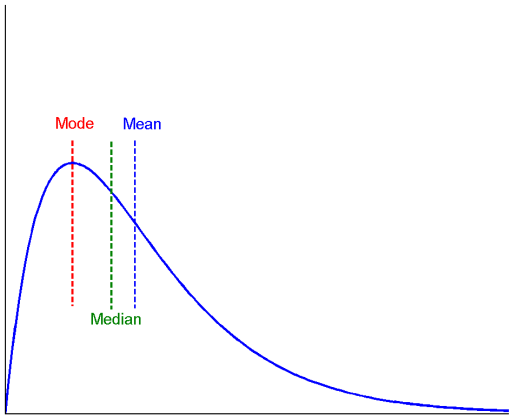
Unimodal, bimodal, trimodal

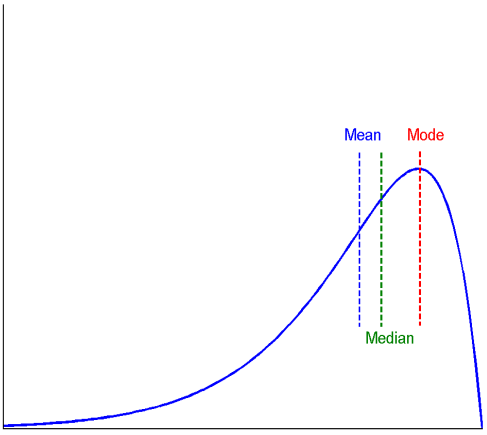
Empirical formula: $mean - mode -= 3 * (mean - median)$

Symmetric vs. Skewed Data

Median, mean and mode of symmetric, positively and negatively skewed data







- **Quartiles, outliers and boxplots**

Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)

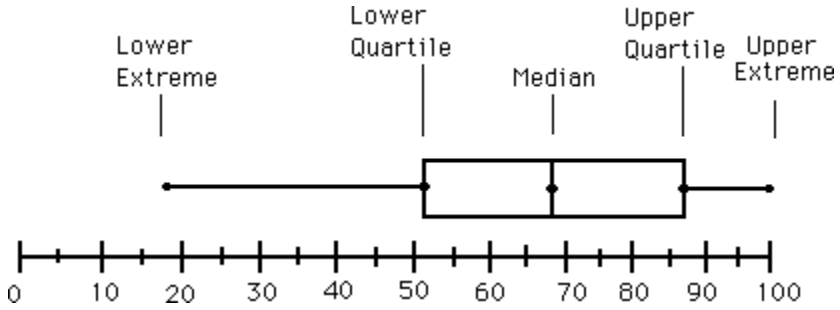
Inter-quartile range: $IQR = Q_3 - Q_1$

Five number summaries: min, Q_1, median, Q_3, max

Boxplot:

- Data is represented with a box

- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- Whiskers: two lines outside the box extended to Minimum and Maximum
- Outliers: points beyond a specified outlier threshold, plotted individually



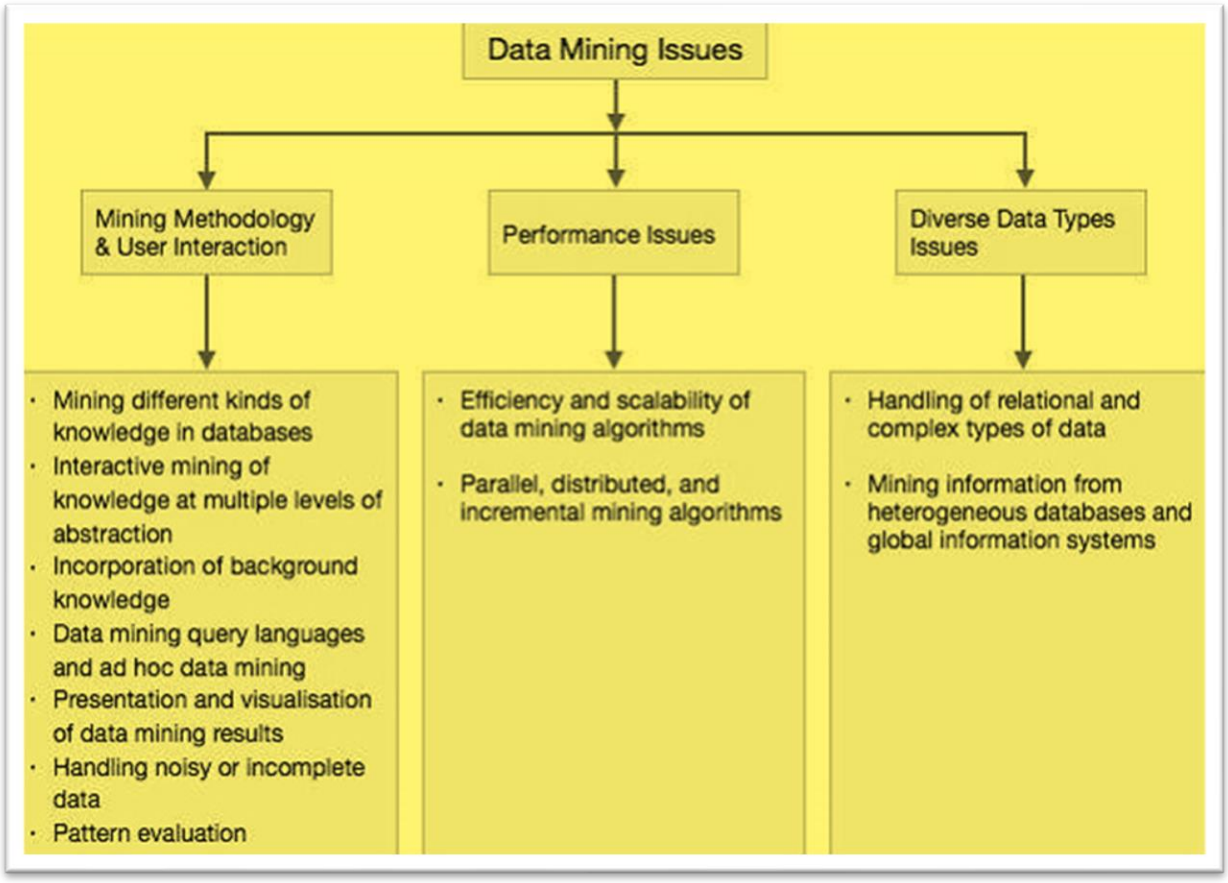
- **Variance and standard deviation (*sample: s, population: σ)***
  - **Variance:**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^{\,2} - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2]$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^{\,2} - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Issues in Data Mining



Mining Methodology and User Interaction Issues

It refers to the following kinds of issues −

- **Mining different kinds of knowledge in databases** − Different users may be interested in different kinds of knowledge. Therefore, it is necessary for data mining to cover a broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction** − The data mining process needs to be interactive because it allows users to focus the search for patterns, providing and refining data mining requests based on the returned results.
- **Incorporation of background knowledge** − To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to express the discovered patterns not only in concise terms but at multiple levels of abstraction.

- **Data mining query languages and ad hoc data mining** − Data Mining Query language that allows the user to describe ad hoc mining tasks, should be integrated with a data warehouse query language and optimized for efficient and flexible data mining.
- **Presentation and visualization of data mining results** − Once the patterns are discovered it needs to be expressed in high level languages, and visual representations. These representations should be easily understandable.
- **Handling noisy or incomplete data** − The data cleaning methods are required to handle the noise and incomplete objects while mining the data regularities. If the data cleaning methods are not there then the accuracy of the discovered patterns will be poor.
- **Pattern evaluation** − The patterns discovered should be interesting because either they represent common knowledge or lack novelty.

Performance Issues

There can be performance-related issues such as follows −

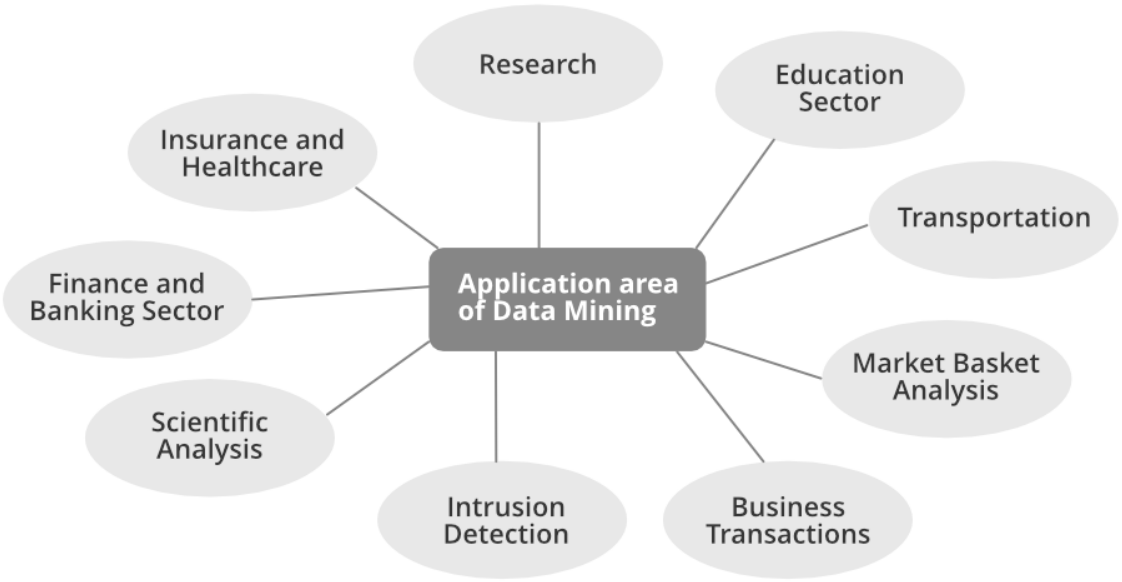
- **Efficiency and scalability of data mining algorithms** − In order to effectively extract the information from huge amount of data in databases, data mining algorithm must be efficient and scalable.
- **Parallel, distributed, and incremental mining algorithms** − The factors such as huge size of databases, wide distribution of data, and complexity of data mining methods motivate the development of parallel and distributed data mining algorithms. These algorithms divide the data into partitions which is further processed in a parallel fashion. Then the results from the partitions is merged. The incremental algorithms, update databases without mining the data again from scratch.

Diverse Data Types Issues

- **Handling of relational and complex types of data** − The database may contain complex data objects, multimedia data objects, spatial data, temporal data etc. It is not possible for one system to mine all these kinds of data.
- **Mining information from heterogeneous databases and global information systems** − The data is available at different data sources on LAN or WAN. These data source may

be structured, semi structured or unstructured. Therefore, mining the knowledge from them adds challenges to data mining.

Data Mining Applications:



Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations

- Misuse Detection
- Anomaly Detection

Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example:

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things) and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify ‘Loyal’ customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.