# Data Warehousing and Mining

## — Module 1.2—
## Introduction to Data Mining

# Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, sale transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras, YouTube
- We are drowning in data, but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

# Evolution of Sciences

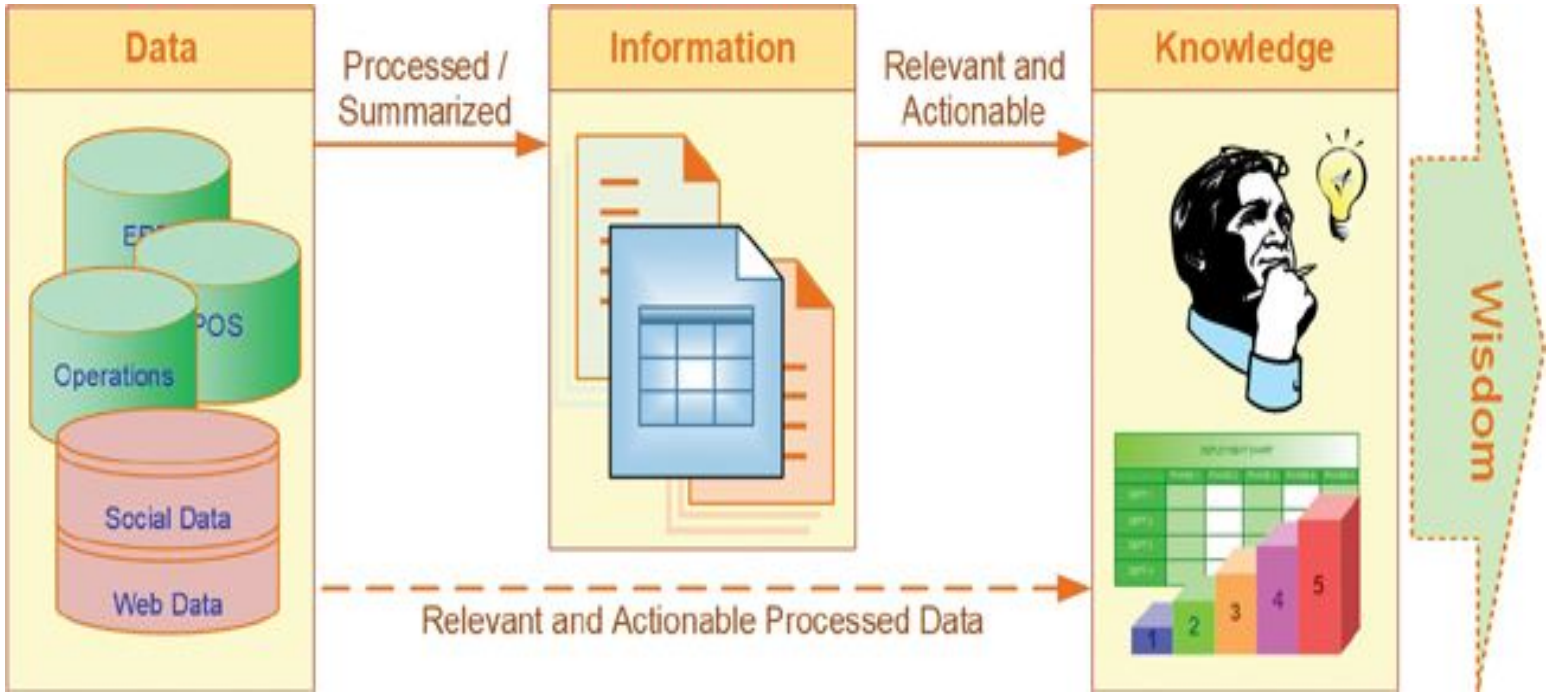
- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!
- Jim Gray and Alex Szalay, *The World Wide Telescope: An Archetype for Online Science*, Comm. ACM, 45(11): 50-54, Nov. 2002

# Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

# What Is Data Mining?

- Data mining is the process of converting data into information and then into knowledge.
- Knowledge is very distinct from data and information
- Knowledge is information that is contextual, relevant, and actionable.
- knowledge has strong experiential and reflective elements that distinguish it from information in a given context.

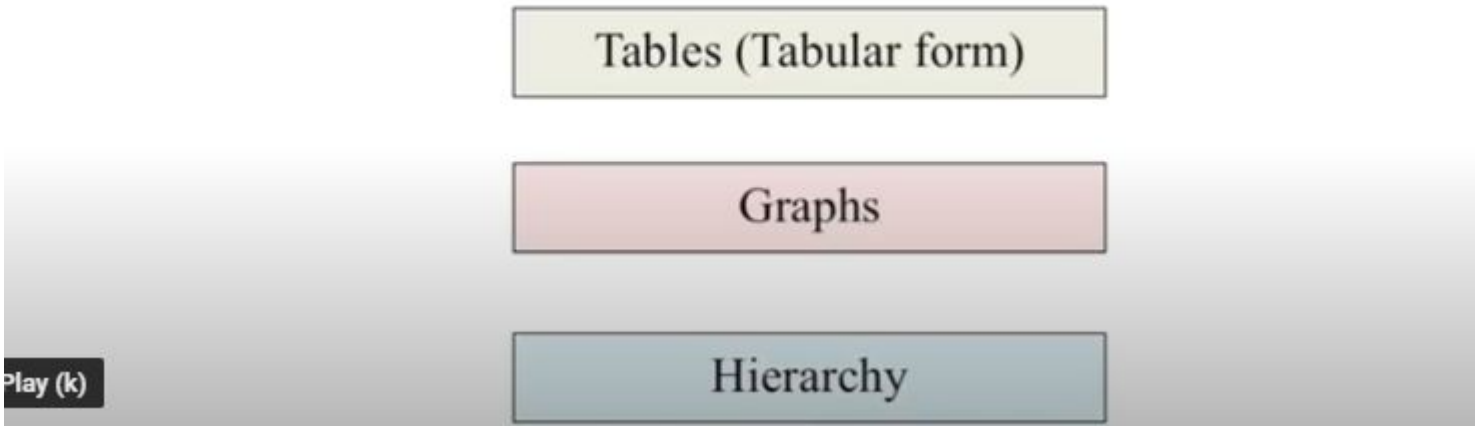


- Nowadays, data mining is used in almost all places where a large amount of data is stored and processed.
- For example, banks typically use ‘data mining’ to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well.
- Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

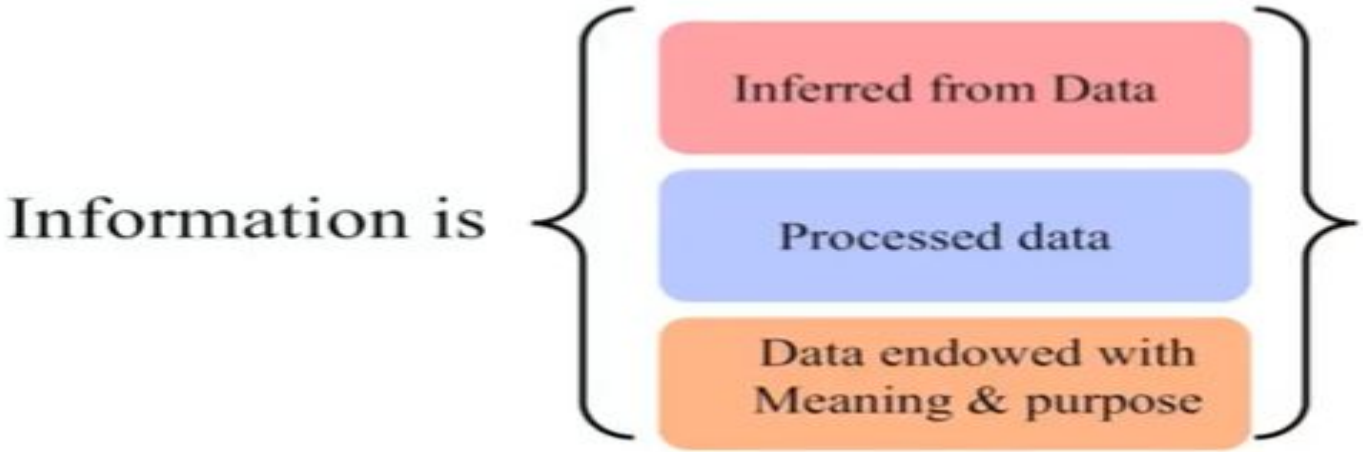
- Data is defined "as being discrete, objective facts or observations, which are unorganized and unprocessed and therefore have no meaning or value because of lack of context and interpretation"


Known facts about entity , Event , transaction etc.
Data is unorganized and unprocessed facts

- Data is generally presented in the form of:



**Information** is "organized or structured data, which has been processed in such a way that the information now has relevance for a specific purpose or context, and is therefore meaningful, valuable, useful and relevant"



Data are the raw alphanumeric values obtained through different acquisition methods. Data in their simplest form consist of raw alphanumeric values.

Information is created when data are processed, organized, or structured to provide context and meaning. Information is essentially processed data.

Knowledge is what we know. Knowledge is unique to each individual and is the accumulation of past experience and insight that shapes the lens by which we interpret, and assign meaning to, information. For knowledge to result in action, an individual must have the authority and capacity to make and implement a decision. Knowledge (and authority) are needed to produce actionable information that can lead to impact.

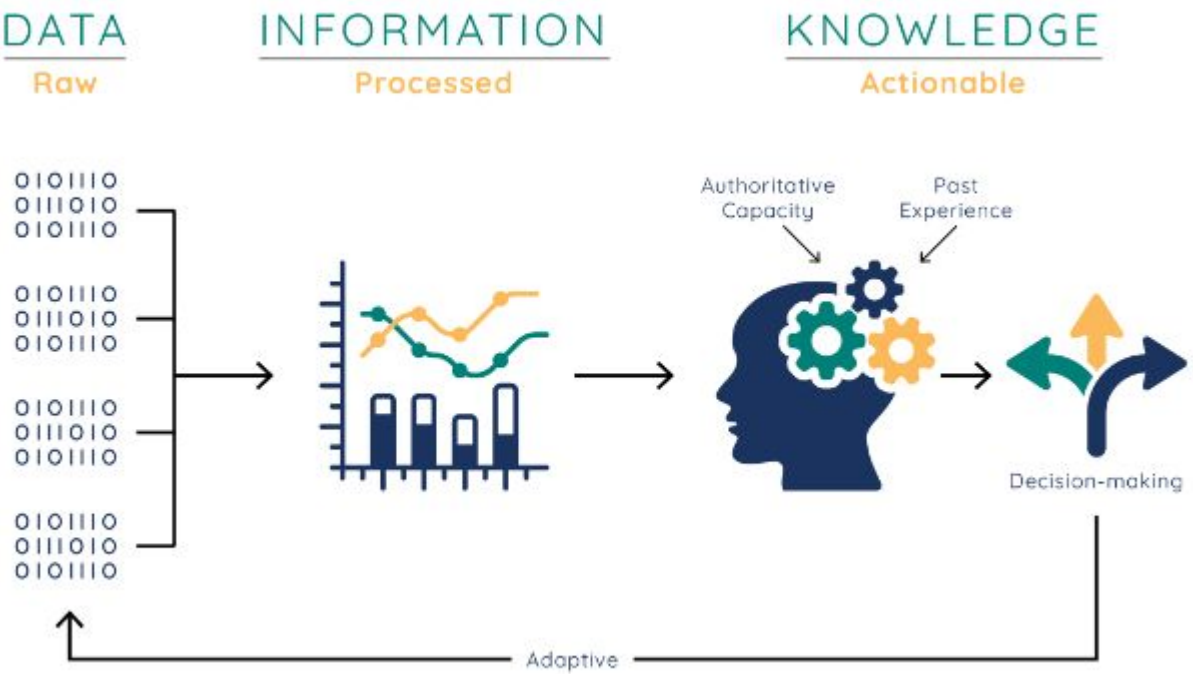


**Table 1:** Characteristics of data, information, and knowledge (adopted from de Vries 2018).

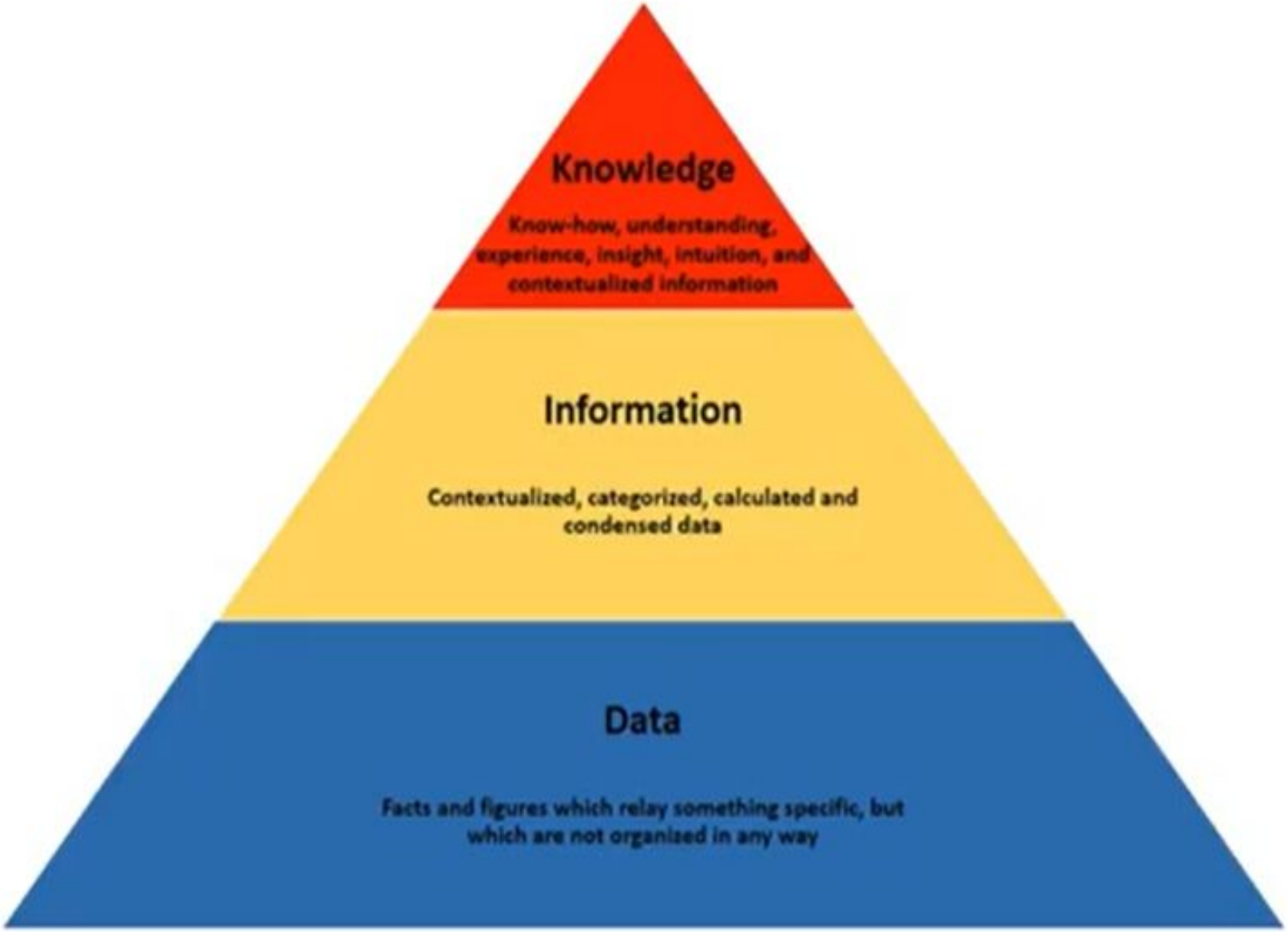
| Data | Information | Knowledge |
|---|---|---|
| Is objective | Should be objective | Is subjective |
| Has no meaning | Has a meaning | Has meaning for a specific purpose |
| Is unprocessed | Is processed | Is processed and understood |
| Is quantifiable, there can be data overload | Is quantifiable, there can be information overload | Is not quantifiable, there is no knowledge overload |

**Table 2:** Examples of transforming water data to information to knowledge that leads to action.

| Data | Information | Knowledge |
|---|---|---|
| Stream gage height | Convert gage height to stream flow estimates to provide summary stats for last 10 years | Restrict withdraws because streamflow is below 7Q10 |
| Amount of precipitation in rain gage | Assess whether annual precipitation is increasing, decreasing, or staying the same | Prioritize investing in floodplain mapping given increases in precipitation over last 20 years |
| Amount of lead in water samples | Combine lead level, customer, and drinking water standards data to locate violations | Alert customers with lead levels exceeding safe drinking water standards |
| Volume of treated water | Correlate volume of treated water with number of low-flush toilets installed over time | Continue investing in the low flush toilet rebate program give large water savings |

- *Heart Disease Detection: Problem Area*
- *Patients (weight,height,sugar,bp,BMI,HeartDisease)*
- *80 5 120 100 5.5 yes* ← *Information*
- *70 5.1 100 80 5.0 No*

- *weight,height,sugar,bp,BMI,HeartDisease*
- *60 4.8 100 95 5.0 ?* ← *Knowledge*


Knowledge
Know-how, understanding, experience, insight, intuition, and contextualized information
Information
Contextualized, categorized, calculated and condensed data
Data
Facts and figures which relay something specific, but which are not organized in any way

# What Is Data Mining?

- Data mining is a process that involves using statistical, mathematical, and artificial intelligence techniques and algorithms to extract and identify useful information and subsequent knowledge (or patterns) from large sets of data.

## Fayyad et al. (1996) defined data mining as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data stored in structured databases,"

Ø **Process** implies that data mining comprises many iterative steps.

Ø **Nontrivial** means that some experimentation-type search or inference is involved;

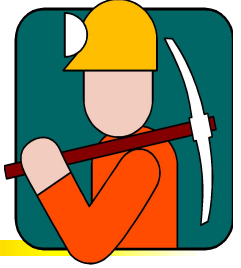
Ø **Valid** means that the discovered patterns should hold true on new data with a sufficient degree of certainty.

Ø **Novel** means that the patterns were not previously known to the user in the context of the system being analyzed.

Ø **Potentially useful** means that the discovered patterns should lead to some benefit to the user or task.

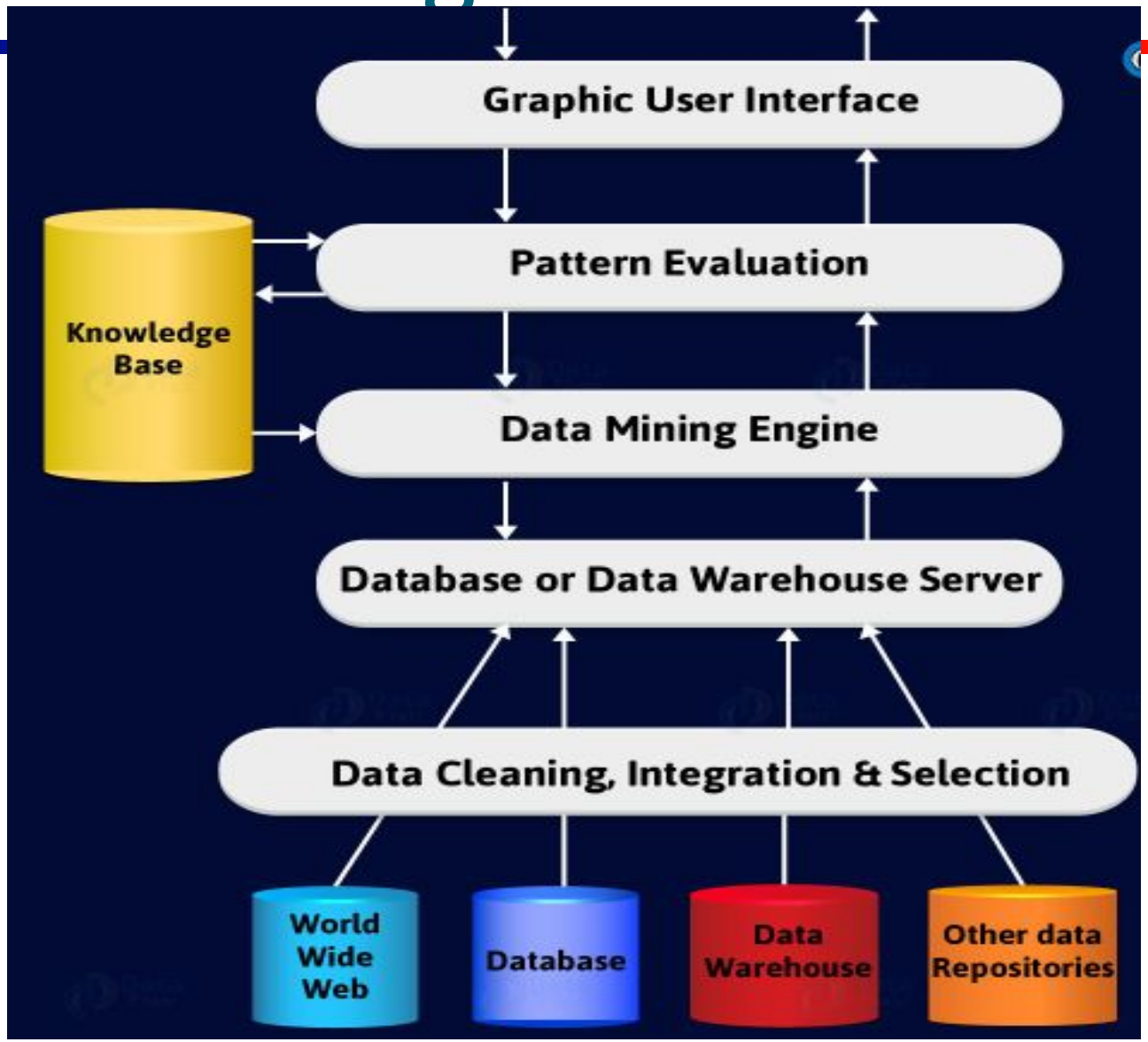
Ø **Ultimately understandable** means that the pattern should make business sense that leads to users saying "This makes sense. Why didn't I think of that?"—if not immediately at least after some processing.

# Data Mining: Alternative names



- Alternative names
  - Knowledge discovery (mining) in databases (KDD),
  - knowledge extraction,
  - data/pattern analysis,
  - data archeology, data dredging,
  - information harvesting,
  - business intelligence, etc.

# Data Mining Architecture



- ***Data mining Architecture system contains  many components.***
- ***That is a data source, data warehouse server, data mining engine, and knowledge base.***

***a. Data Sources***

- *There are so many documents present.*
- *That is a database, data warehouse, World Wide Web (WWW).*
- *That are the actual sources of data. Sometimes, data may reside even in plain text files or spreadsheets.*
- *World Wide Web or the Internet is another big source of data.*

***b. Database or Data Warehouse Server***

- *The database server contains the actual data that is ready to be processed.*
- *Hence, the server handles retrieving the relevant data. That is based on the data mining request of the user.*

***c. Data Mining Engine***

- *In data mining system data mining engine is the core component.*
- *As It consists a number of modules.*
- *That we used to perform data mining tasks. That includes association, classification, characterization, clustering, prediction, etc.*

***d. Pattern Evaluation Modules***

- *This module is mainly responsible for the measure of interestingness of the pattern. For this, we use a threshold value.*
- *Also, it interacts with the data mining engine. That's main focus is to search towards interesting patterns.*

***e. Graphical User Interface***

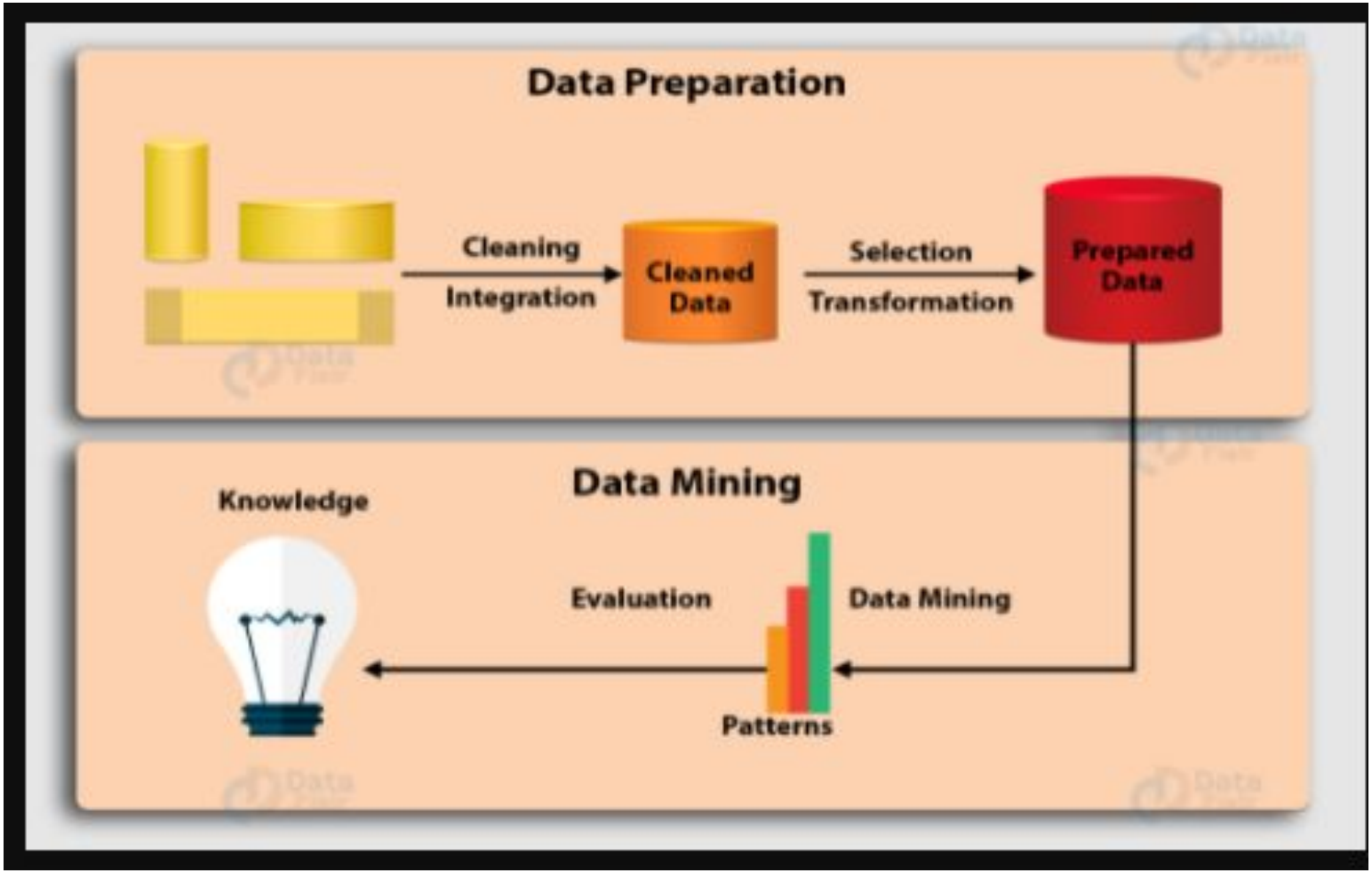
- *We use this interface to communicate between the user and the data mining system.*
- *Also, this module helps the user use the system easily and efficiently. They don't know the real complexity of the process.*
- *When the user specifies a query, this module interacts with the data mining system.*
- *Thus, displays the result in an easily understandable manner.*

- ***What is Data Mining?***
- ***It is a process of discovering hidden valuable knowledge by analyzing a large amount of data. Also, we have to store that data in different databases.***
- ***As data mining is a very important process.***
- ***It becomes an advantage for various industries. Such as manufacturing, marketing, etc. to increase their business efficiency. Therefore, the needs for a standard data mining process increased dramatically.***

- ***Stages of Data Mining Process***
- ***Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining***

# Stages of Data Mining Process

Data Mining Process is classified into two stages: Data preparation or data preprocessing and data mining



- ***a. Data Cleaning***
- In the phase of data mining process, data gets cleaned.
- As we know data in the real world is noisy, inconsistent and incomplete.
- It includes a number of techniques. Such as filling in the missing values, combined compute.
- The output of the data cleaning process is adequately cleaned data.
- **b. Data Integration**
- In this phase of Data Mining process data in integrated from different data sources into one. As data lies in different formats in a different location.
- We can store data in a database, text files, spreadsheets, documents, data cubes, and so on.
- Although, we can say data integration is so complex, tricky and difficult task. That is because normally data doesn't match the different sources.
- We use metadata to reduce errors in the data integration process. Another issue faced is data redundancy.
- In this case, the same data might be available in different tables in the same database.
- Data integration tries to reduce redundancy to the maximum possible level. As without affecting the reliability of data

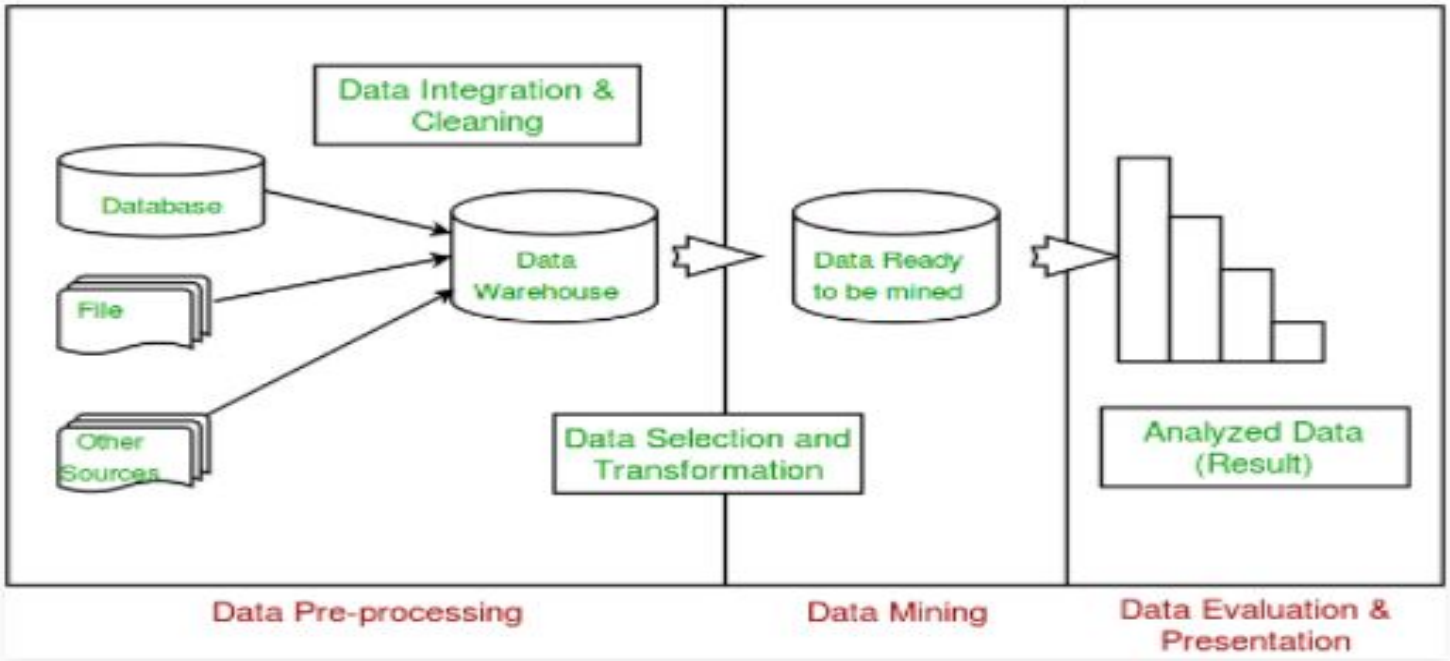
- ***c. Data Selection***
- This is the process by which data relevant to the analysis is retrieved from the database.
- As this process requires large volumes of historical data for analysis.
- So, usually, the data repository with integrated data contains much more data than actually required.
- From the available data, data of interest needs to be selected and stored.
- ***d. Data Transformation***
- In this process, we have to transform and consolidate the data into different forms. That must be suitable for mining.
- Normally this process includes normalization, aggregation, generalization etc.
- For example, a data set available as "-5, 37, 100, 89, 78" can be transformed as "-0.05, 0.37, 1.00, 0.89, 0.78". Here data becomes more suitable for data mining. After data integration, the available data is ready for data mining.
- ***e. Data Mining***
- In this phase of Data Mining process, we have applied methods to extract patterns from the data.
- As these methods are complex and intelligent. Also, this mining includes several tasks. Such as classification, prediction, clustering, time series analysis and so on.

## *f. Pattern Evaluation*

- The pattern evaluation identifies the truly interesting patterns.
- That is representing knowledge based on different types of interesting measures.
- A pattern is considered to be interesting if it is potentially useful.
- Also, easily understandable by humans.

- ***g. Knowledge Representation***
- In the phase of Data Mining process, we have to represent data to the user in an appealing way.
- Also, that information is mined from the data. To generate output different techniques are need to be applied.

- ***Data Mining phases:***
- ***The whole process of Data Mining consists of three main phases:***
- ***Data Pre-processing – Data cleaning, integration, selection, and transformation takes place***
- ***Data Extraction – Occurrence of exact data mining***
- ***Data Evaluation and Presentation – Analyzing and presenting results***



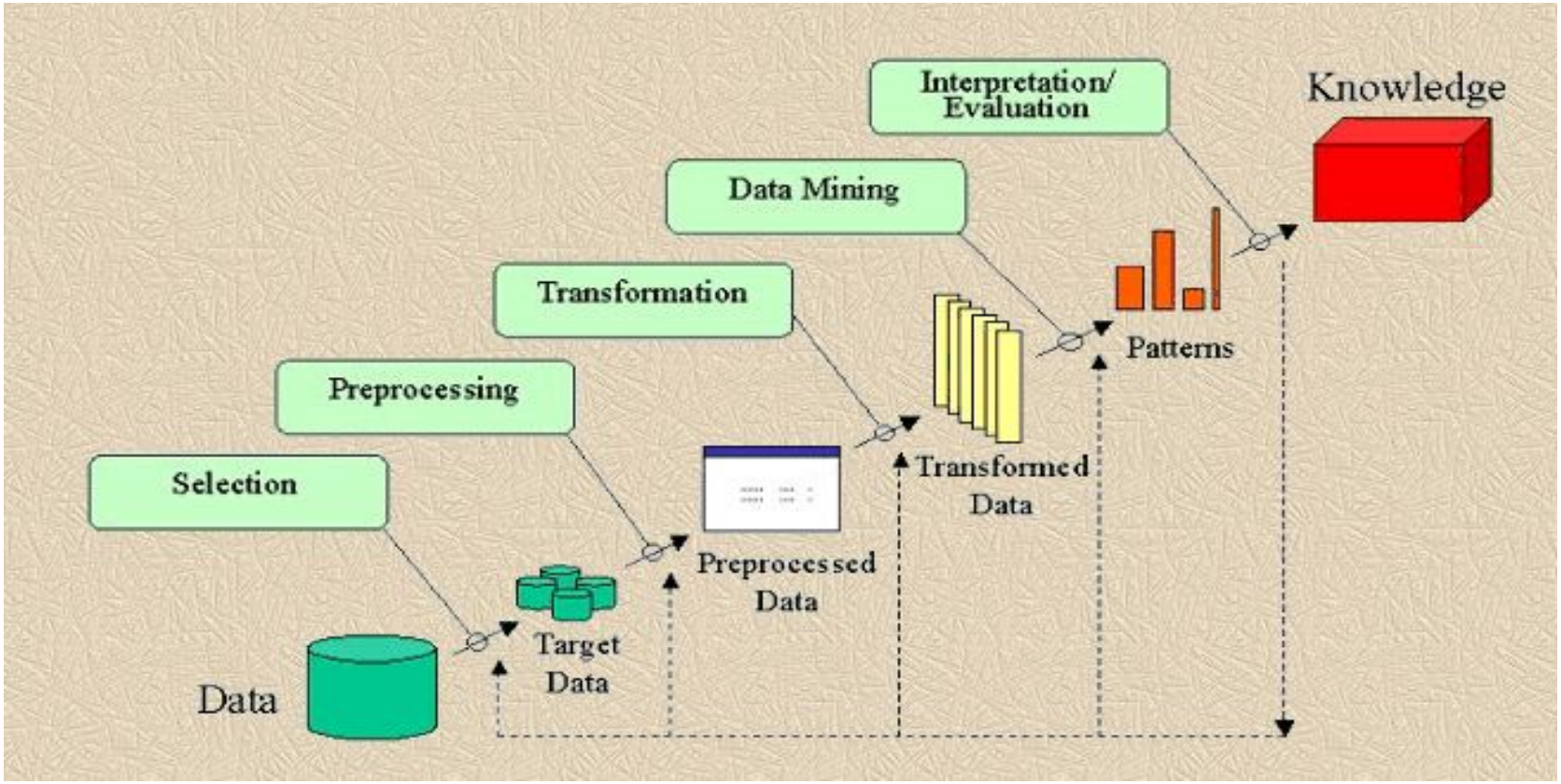
- ***What is the KDD Process?***

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

- ***KDD process is to extract knowledge from data in the context of large databases.***

- ***It does this by using data mining methods (algorithms) to extract (identify) what is deemed knowledge, according to the specifications of measures and thresholds, using a database along with any required preprocessing, subsampling, and transformations of that database.***

# KDD Process:


Interpretation/
Evaluation
Knowledge
Data Mining
Transformation
Patterns
Preprocessing
Selection
Transformed
Data
Preprocessed
Data
Target
Data
Data

- *The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:*

- ***Developing an understanding of***
- *the application domain*
- *the relevant prior knowledge*
- *the goals of the end-user*
- ***Creating a target data set:***
- *selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.*
- ***Data cleaning and preprocessing.***
- *Removal of noise or outliers.*
- *Collecting necessary information to model or account for noise.*
- *Strategies for handling missing data fields.*
- *Accounting for time sequence information and known changes.*
- ***Data reduction and projection.***
- *Finding useful features to represent the data depending on the goal of the task.*

- Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
- ***Choosing the data mining task.***
- *Deciding whether the goal of the KDD process is classification, regression, clustering, etc.*
- ***Choosing the data mining algorithm(s).***
- *Selecting method(s) to be used for searching for patterns in the data.*
- *Deciding which models and parameters may be appropriate.*
- *Matching a particular data mining method with the overall criteria of the KDD process.*
- ***Data mining.***
- *Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.*
- ***Interpreting mined patterns.***
- ***Consolidating discovered knowledge.***

- **KDD** refers to the overall process of discovering useful knowledge from data.
- It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge.
- It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.
  **Data mining** refers to the application of algorithms for extracting patterns from data without the additional steps of the **KDD process.**
- ***Definitions Related to the KDD Process***
- ***Knowledge discovery in databases is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.***

| | |
|---|---|
| Data | A set of facts, $F$. |
| Pattern | An expression $E$ in a language $L$ describing facts in a subset $F_E$ of $F$. |
| Process | KDD is a *multi-step process* involving data preparation, pattern searching, knowledge evaluation, and refinement with iteration after modification. |
| Valid | Discovered patterns should be true on new data with some degree of certainty.<br>Generalize to the future (other data). |
| Novel | Patterns must be novel (should not be previously known). |
| Useful | Actionable; patterns should potentially lead to some useful actions. |
| Understandable | The process should lead to human insight.<br>Patterns must be made understandable in order to facilitate a better understanding of the underlying data. |

- ***Applications of Data Mining***

- ***Financial Analysis***
- ***Biological Analysis***
- ***Scientific Analysis***
- ***Intrusion Detection***
- ***Fraud Detection***
- ***Research Analysis***

- ***Data Mining Applications & Use Cases***
- ***Following are the applications of data mining in various sectors:***
- ***a. Data Mining in Finance***
- *We have to Increase customer loyalty by collecting and analyzing customer behavior data. Also, one needs to help banks that predict customer behavior and launch relevant services and products.*
- *Helps in Discovering hidden correlations between various financial indicators that need to detect suspicious activities with a high potential risk.*
- *Generally, it identifies fraudulent or non-fraudulent actions. As it done by collecting historical data. And then turning it into valid and useful information.*
- ***b. Data Mining in Healthcare***
- *Basically, it provides government, regulatory and competitor information that can fuel competitive advantage. Although, it supports the R&D process. And then go-to-market strategy with rapid access to information at every phase.*

- *Generally, it discovers the relationships between diseases and the effectiveness of treatments.*
- *That is to identify new drugs or to ensure that patients receive appropriate, timely care.*
- *Also, it supports healthcare insurers in detecting fraud and abuse.*

***c. Data Mining for Intelligence***

- *Generally, it reveals hidden data related to money laundering, narcotics trafficking, etc.*
- *Also, helps in Improving intrusion detection with a high focus on anomaly detection. And identify suspicious activity from a day one.*
- *Basically, convert text-based crime reports into word processing files.*
- *That can be used to support the crime-matching process.*

- ***d. Data Mining in Telecommunication***
- *In this, data mining gains a competitive advantage and reduce customer churn by understanding demographic characteristics and predicting customer behavior.*
- *Increases customer loyalty and improve profitability by providing customized services.*
- *As it supports customer strategy by developing appropriate marketing campaigns and pricing strategies.*

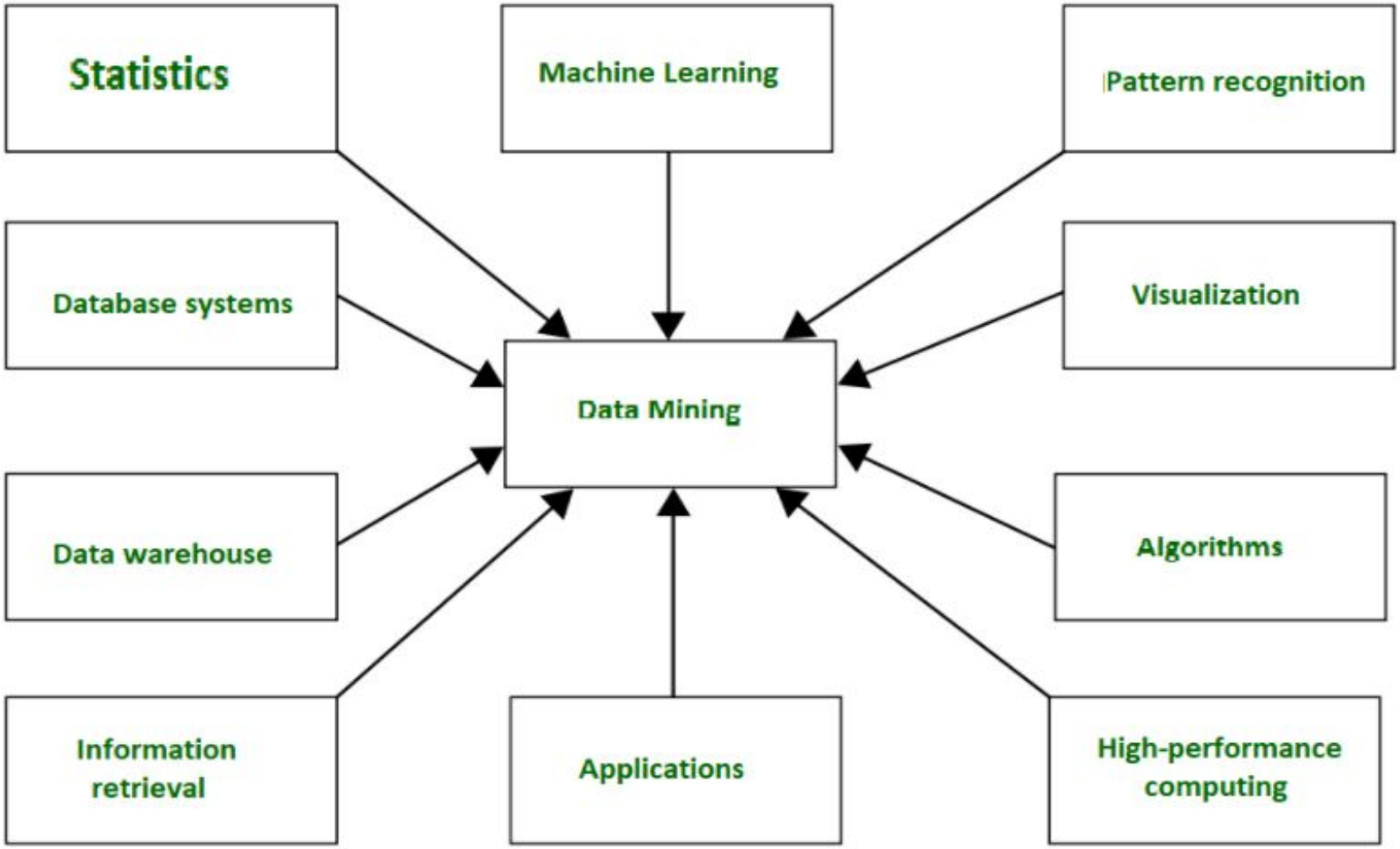
- ***f. Data Mining in Marketing and Sales***
- *Basically, it enables businesses to understand the hidden patterns inside historical purchasing transaction data. Thus helping in planning and launching new marketing campaigns.*
- *Generally, the following illustrates several data mining applications in sale and marketing.*
- *We use it for market basket analysis. That is to provide information on what product combinations have to purchased together. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products.*
- *Retail companies use data mining to identify customer's behavior buying patterns.*
- ***g. Data Mining in E-commerce***
- *Many E-commerce companies are using data mining business Intelligence to offer cross-sells through their websites.*
- *One of the most famous of these is, of course, Amazon. They use sophisticated mining techniques to drive their 'People who viewed that product. Also liked this' functionality.*

- ***h. Data Mining in Education***
- There is a newly emerging field, called Educational Data Mining. As it concerns with developing methods. That discover knowledge from data originating from educational Environments.

- **The goals of EDM are identified as predicting students' future learning behavior, studying. We use data mining by an institution to take accurate decisions. And also to predict the results of the student.**

- With the results, the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured. And used to develop techniques to teach them.
- https://data-flair.training/blogs/data-mining-applications/

# Main Purpose of Data Mining



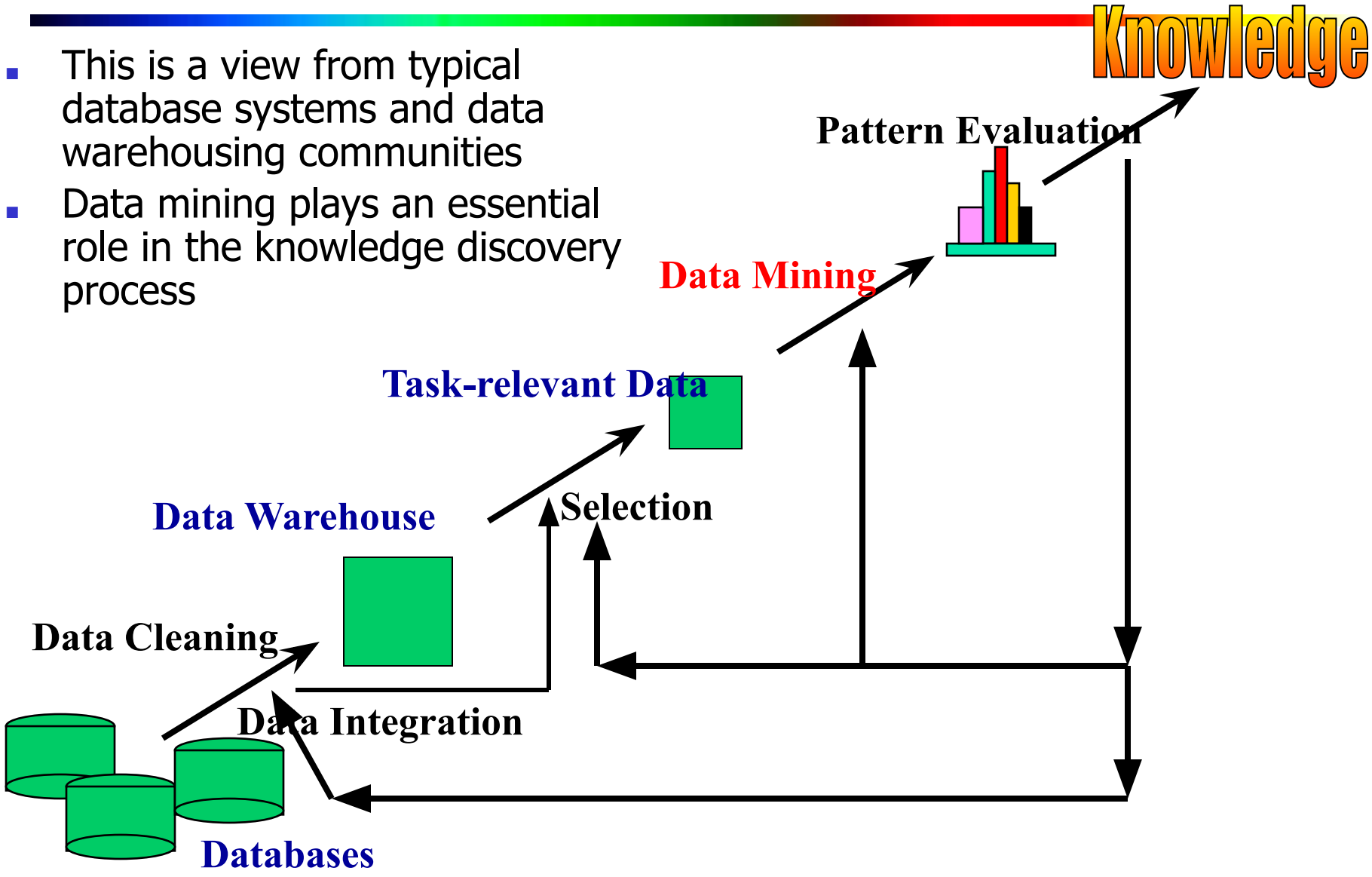
- ***Basically, Data mining has been integrated with many other techniques from other domains such as statistics, machine learning, pattern recognition, database and data warehouse systems, information retrieval, visualization, etc. to gather more information about the data and to helps predict hidden patterns, future trends, and behaviors and allows businesses to make decisions.***
- Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information.

# Getting to Know Your Data

- ***Data Objects and Attribute Types***
- ***Basic Statistical Descriptions of Data***
- ***Measuring Data Similarity and Dissimilar***

# Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities
- Data mining plays an essential role in the knowledge discovery process



# Data Mining: On What Kinds of Data?

- Database-oriented data sets and applications
  - Relational database, data warehouse, transactional database
- Advanced data sets and advanced applications
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Object-relational databases
  - Spatial data
  - Multimedia database
  - Text databases
  - The World-Wide Web

# What Kinds of Patterns Can Be Mined?

- Data mining functionalities are used to specify the kinds of patterns to be found in data mining tasks.

  - **Descriptive mining tasks :** Deals with the General characteristics and converts them into relevant and useful information

  - **Predictive mining tasks:** Predicts future values by analyzing data patterns and their outcomes based on past data.

# Descriptive DM Functionalities

## 1. Class/Concept Description:

- Data entries can be associated with the classes or concepts.
- These descriptions can be derived using

(1) *data characterization*, by summarizing the data of the class under study (often called the **target class**) in general terms,

- *Example: At electronic store a Customer relationship manager asks to Summarize the characteristics of customers who spend more than Rs.10000 a year at the store.*

or

(2) *data discrimination*, by comparison of the target class with one or a set of comparative classes (often called the **contrasting classes**),

- *Example: A customer relationship manager at Electronics store may want to compare two groups of customers—those who shop for computer products regularly (e.g., more than twice a month) and those who rarely shop for such products (e.g. less than three times a year).*

or

(3) both data characterization and discrimination.

# Descriptive DM Functionalities

## 2. Mining of frequent patterns:

- Patterns that occur frequently in data.
- It includes--
  - **Frequent itemset** : refers to a set of items that often appear together in a transactional data set
  - **Frequent subsequences (also known as sequential patterns):** A frequently occurring subsequence like laptop digital camera memory card
  - **Frequent substructures:**
  - A substructure can refer to different structural forms (e.g., graphs, trees, or lattices) that may be combined with itemsets or subsequences.
  - If a substructure occurs frequently, it is called a (*frequent*) *structured pattern*.

- Mining frequent patterns leads to the discovery of interesting associations and correlations within data.

# Descriptive DM Functionalities

## 3. Association Analysis

- Defines relationships between the data and predefined association rules.
- Suppose that, as a marketing manager at *Electronics store*, you want to know which items are frequently purchased together
- A rule, ı $buys(X, \text{“}computer\text{”}) \Rightarrow buys(X, \text{“}software\text{”})\ [support = 1\%, confidence = 50\%]$
- Association rules that contain a single predicate are referred to as **single-dimensional association rules**.
- Suppose, instead, that we are given the *Electronics* relational database related to purchases. A data mining system may find association rules like

$$age(X, \text{“}20..29\text{”}) \wedge income(X, \text{“}40K..49K\text{”}) \Rightarrow buys(X, \text{“}laptop\text{”})$$
$$[support = 2\%, confidence = 60\%].$$

- Association rules that contain a more than one predicate/attributes are referred to as **multi-dimensional association rules**.

# Descriptive DM Functionalities

**4. Clustering :** can be used to generate class labels for a group of data. The objects are clustered or grouped based on the principle of *maximizing the intraclass similarity and minimizing the interclass similarity*.

i.e. clusters of objects are formed so that objects within a cluster have high similarity , but are rather dissimilar to objects in other clusters.

- Clustering can also facilitate taxonomy formation  Organization of observations into a hierarchy of classes that group similar events together.
- Example: Cluster analysis can be performed on *Electronics store* customer data to identify homogeneous subpopulations of customers. These clusters may represent individual target groups for marketing

# Descriptive Mining

- This term is basically used to produce correlation, cross-tabulation, frequency etc.
- These technologies are used to determine the similarities in the data and to find existing patterns.
- One more application of descriptive analysis is to develop the captivating subgroups in the major part of the data available.
- This analytics emphasis on the summarization and transformation of the data into meaningful information for reporting and monitoring.
- Examples of descriptive data mining include clustering, association rule mining, and anomaly detection. Clustering involves grouping similar objects together, while association rule mining involves identifying relationships between different items in a dataset. Anomaly detection involves identifying unusual patterns or outliers in the data.

# Predictive  Data mining functionalities

- Predicts future values by analyzing data patterns and their outcomes based on past data.

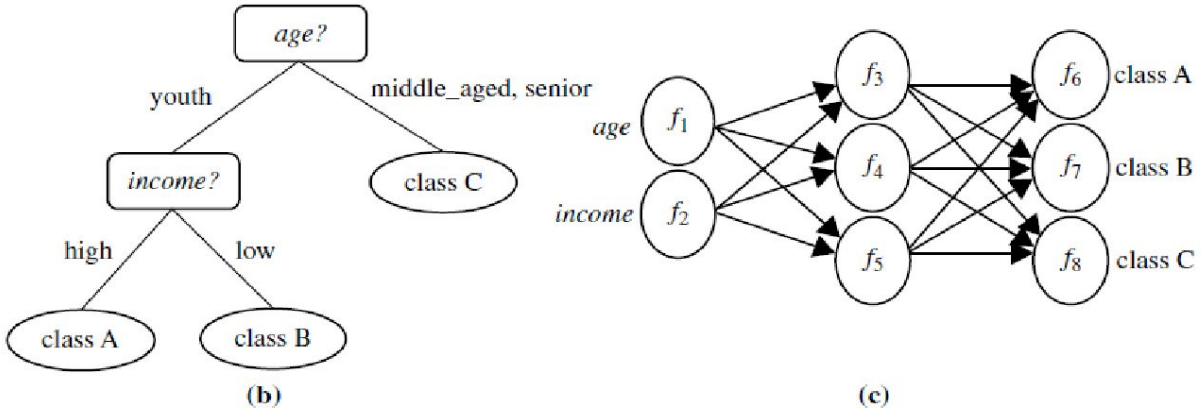
  - Classification
  - Regression
  - Outlier analysis

# Predictive Data mining functionalities

1. **Classification:** is the process of finding a **model** (or function) that describes and distinguishes data classes or concepts.
   - The models are derived based on the analysis of a set of **training data** (i.e., data objects for which the class labels are known).
   - The model is used to predict the class label of objects for which the class label is unknown.
   - The derived model may be represented in various forms, such as *classification rules* (i.e., *IF-THEN rules*), *decision trees*, *mathematical formulae*. or *neural networks*

*age(X, "youth") AND income(X, "high")* ⟶ *class(X, "A")*
*age(X, "youth") AND income(X, "low")* ⟶ *class(X, "B")*
*age(X, "middle_aged")* ⟶ *class(X, "C")*
*age(X, "senior")* ⟶ *class(X, "C")*

(a)



# Predictive Data mining functionalities

**2. Regression:**

- Whereas classification predicts categorical (discrete, unordered) labels, **regression** models continuous-valued functions. That is, regression is used to predict missing or unavailable *numerical data values* rather than (discrete) class labels.
- The term *prediction* refers to both numeric prediction and class label prediction.
- **Regression analysis** is a statistical methodology that is most often used for numeric prediction.
- Regression also encompasses the identification of distribution *trends* based on the available data.

# Predictive Data mining functionalities

## 3. Outlier Analysis :

- Outlier: A data object that does not comply with the general behavior of the data
- Many data mining methods discard outliers as noise or exceptions. However, in some applications (e.g., fraud detection), the rare events can be more interesting than the more regularly occurring ones.
- Outliers may be detected using statistical tests that assume a distribution or probability model for the data, or using distance measures where objects that are remote from any other cluster are considered outliers.
- Example: Outlier analysis may uncover fraudulent usage of credit cards by detecting purchases of unusually large amounts for a given account number in comparison to regular charges incurred by the same account.
  Outlier values may also be detected with respect to the locations and types of purchase, or the purchase frequency.
  - It is used in observing the change in trends of buying patterns of a customer.

# Predictive Data Mining:

- The main goal of this mining is to say something about future results not of current behaviour.
- It uses the supervised learning functions which are used to predict the target value.
- The methods come under this type of mining category are called classification, time-series analysis and regression.
- Modelling of data is the necessity of the predictive analysis, and it works by utilizing a few variables of the present to predict the future not known data values for other variables.
- Examples of predictive data mining include regression analysis, decision trees, and neural networks. Regression analysis involves predicting a continuous outcome variable based on one or more predictor variables. Decision trees involve building a tree-like model to make predictions based on a set of rules. Neural networks involve building a model based on the structure of the human brain to make predictions.

**The main differences between descriptive and predictive data mining are:**

**Purpose:** Descriptive data mining is used to describe the data and identify patterns and relationships. Predictive data mining is used to make predictions about future events.

**Approach:** Descriptive data mining involves analyzing historical data to identify patterns and relationships. Predictive data mining involves using statistical models and machine learning algorithms to identify patterns and relationships that can be used to make predictions.

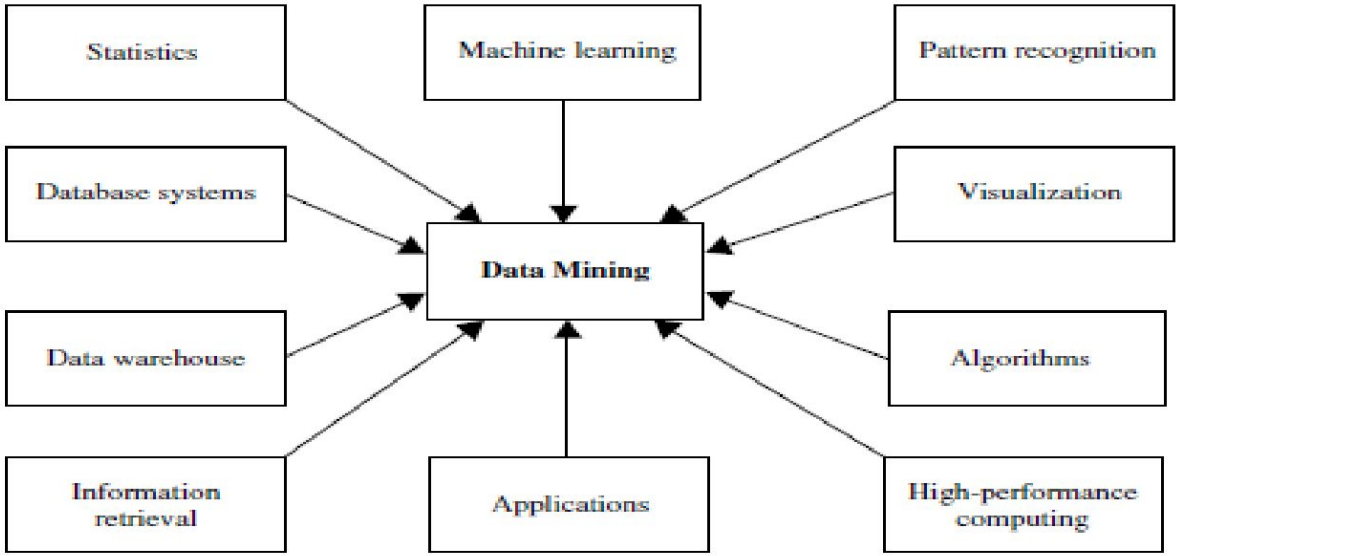
**Output:** Descriptive data mining produces summaries and visualizations of the data. Predictive data mining produces models that can be used to make predictions.

**Timeframe:** Descriptive data mining is focused on analyzing historical data. Predictive data mining is focused on making predictions about future events.

**Applications:** Descriptive data mining is used in applications such as market segmentation, customer profiling, and product recommendation. Predictive data mining is used in applications such as fraud detection, risk assessment, and demand forecasting.

| S.No. | Descriptive Data Mining | Predictive Data Mining |
|---|---|---|
| 1. | It tries to understand what happened in the past by analyzing the stored data. | It tries to understand what could happen in the future using past data analysis. |
| 2. | The data it provides is accurate. | It may not be accurate result. |
| 3. | It provides standard reporting. It also provides ad-hoc reporting. | It is used in predictive modelling, forecasting, simulation and alerts. |
| 4. | It uses data aggregation and data mining. | It uses statistics and forecasting methods. |
| 5. | It uses a reactive approach. | It uses a proactive approach. |

# Technologies used in Data Mining




Data mining adopts techniques from many domains.

## 1. Statistics:

- It uses the mathematical analysis to express representations, model and summarize empirical data or real world observations.
- Statistical analysis involves the collection of methods, applicable to large amount of data to conclude and report the trend.

# Technologies used in Data Mining

## 2. Machine learning

- **Arthur Samuel** defined machine learning as a field of study that gives computers the ability to learn without being programmed.
- When the new data is entered in the computer, algorithms help the data to grow or change due to machine learning.
- In machine learning, an algorithm is constructed to predict the data from the available database **(Predictive analysis).**
- It is related to computational statistics.

# Technologies used in Data Mining

**The four types of machine learning are:**

**1. Supervised learning**

It is based on the classification.

It is also called as **inductive learning**. In this method, the desired outputs are included in the training dataset.

**2. Unsupervised learning**

Unsupervised learning is based on clustering. Clusters are formed on the basis of similarity measures and desired outputs are not included in the training dataset.

**3. Semi-supervised learning**

Semi-supervised learning includes some desired outputs to the training dataset to generate the appropriate functions. This method generally avoids the large number of labeled examples (i.e. desired outputs) .

**4. Active learning**

Active learning is a powerful approach in analyzing the data efficiently.

The algorithm is designed in such a way that, the desired output should be decided by the algorithm itself (the user plays important role in this type).

# Technologies used in Data Mining

### 3. Database systems and data warehouses

- Databases are used for the purpose of recording the data as well as data warehousing.
- Online Transactional Processing (OLTP) uses databases for day to day transaction
  purpose.
- To remove the redundant data and save the storage space, data is normalized and stored in the form of tables.
- **E**ntity-**R**elational modeling techniques are used for relational database management system design.
- Data warehouses are used to store historical data which helps to take strategical decision for business.
- It is used for online analytical processing (OALP), which helps to analyze the data.

# Technologies used in Data Mining

**4. Information retrieval**

Information deals with uncertain representations of the semantics of objects (text, images).

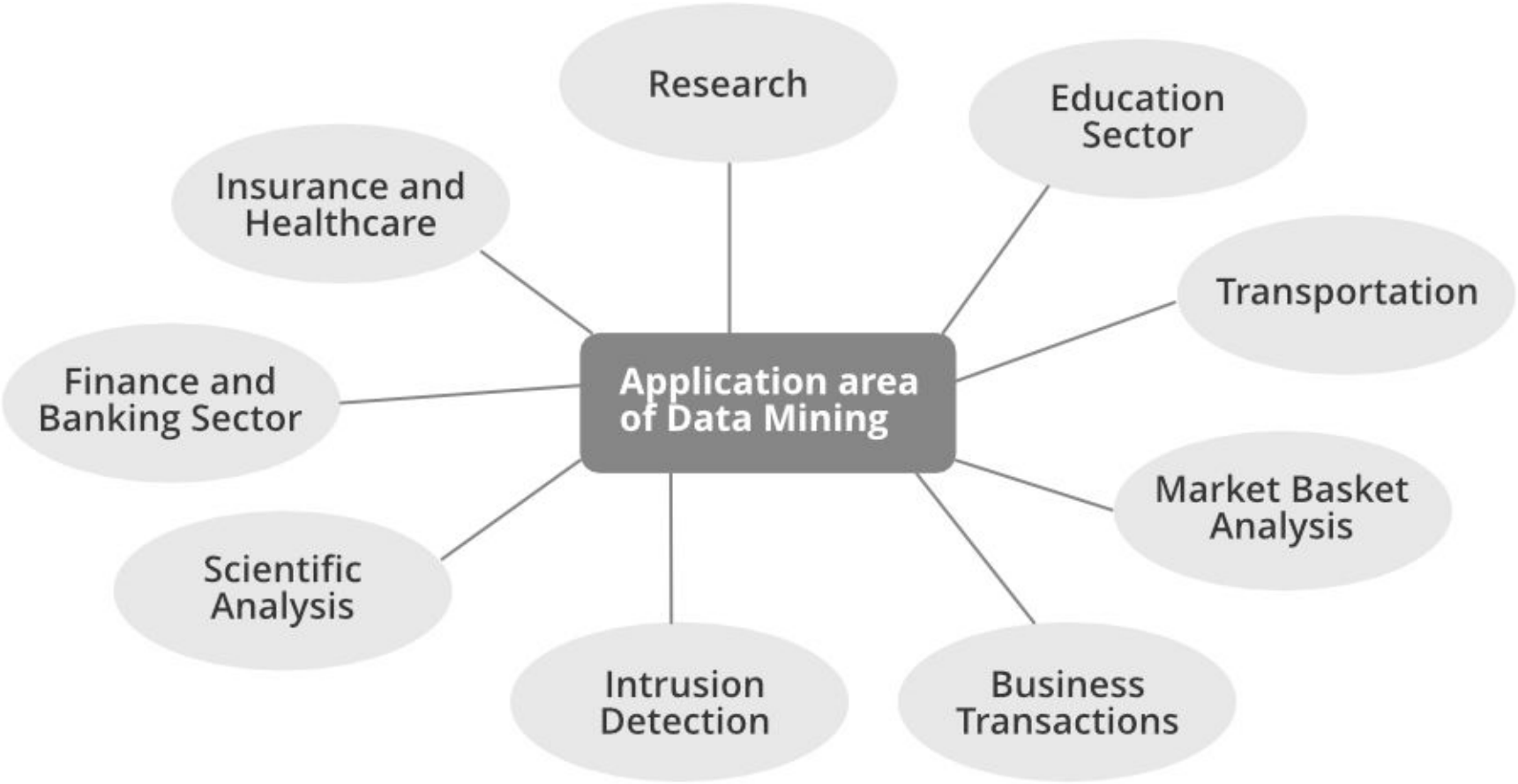
**For example:** Finding relevant information from a large document.

# Technologies used in Data Mining

**5. Decision support system**

- Decision support system is a category of information system. It is very useful in decision making for organizations.
- It is an interactive software based system which helps decision makers to extract useful information from the data, documents to make the decision.

# Data Mining: Applications



# Data Mining: Applications

1. **Marketing and CRM:**
   -To identify most likely buyers of new products
   -To identify root causes of customer attrition so as to improve customer retention
   -To discover time variant associations between products and services to maximize sale and find most profitable customers.
1. **Banking and Finance:**
   -To detect fraudulent credit card and online banking transactions
   -To optimize the cash return by forecasting cash flow on banking entities
   -To streamline and automate the processing of loan applications by accurately predicting most probable defaulters.
   -To maximize the customer value by identifying and selling the products and services that customers are most likely to buy.
1. **Retailing and Logistics:**
   -To identify accurate sales volume at specific retail locations in order to determine correct inventory levels.
   -To do MBA to improve store layout and optimize sales promotions
   -To forecast consumption levels for different product types.
   -To discover interesting patterns in the movement of products in a supply chain by analyzing sensory and RFID data.

# Data Mining: Applications

**4. Manufacturing:**

- To predict machine failures using sensory data
- To discover novel patterns to identify and improve product quality.

**5. Brokerages and Security Tradings:**

-To predict when and how much certain stock / bond prices will change.
-To forecast range of market fluctuations,direction of fluctuations
-To assess effect of particular issues/events on market movements.
-To identify and prevent fraudulent activities in security trading.

**6. Insurance:**

- To predict which customers will buy new policies
- Identify fraudulent behavior of customers
- Prevent incorrect claim payments

**7. Computer Hardware and Software:**

-To predict disk failure
- To identify and filter unwanted web contents and email messages
-To identify potentially unsecured software products

**8. Government and Defense:**

-To forecast the cost of moving military personnel and equipments.
-To predict resource consumption for better planning and budgeting

# Data Mining: Applications

**9. Travel and Lodging:**

- To predict sales of different services to optimally price these services.
- To forecast demand at different locations to better allocate limited organizational resources. .
- To identify most profitable customers and provide them with personalised services.
- To retain valuable employees by identifying and acting on the root causes for attrition

**9. Health and Healthcare:**

- To identify successful medical therapies for different illnesses.
- To identify people without health insurance and reasons behind it.
- To forecast the time of demand at different service locations to optimally allocate organizational resources.
- To retain valuable employees by identifying root causes for attrition

**9. Entertainment:**

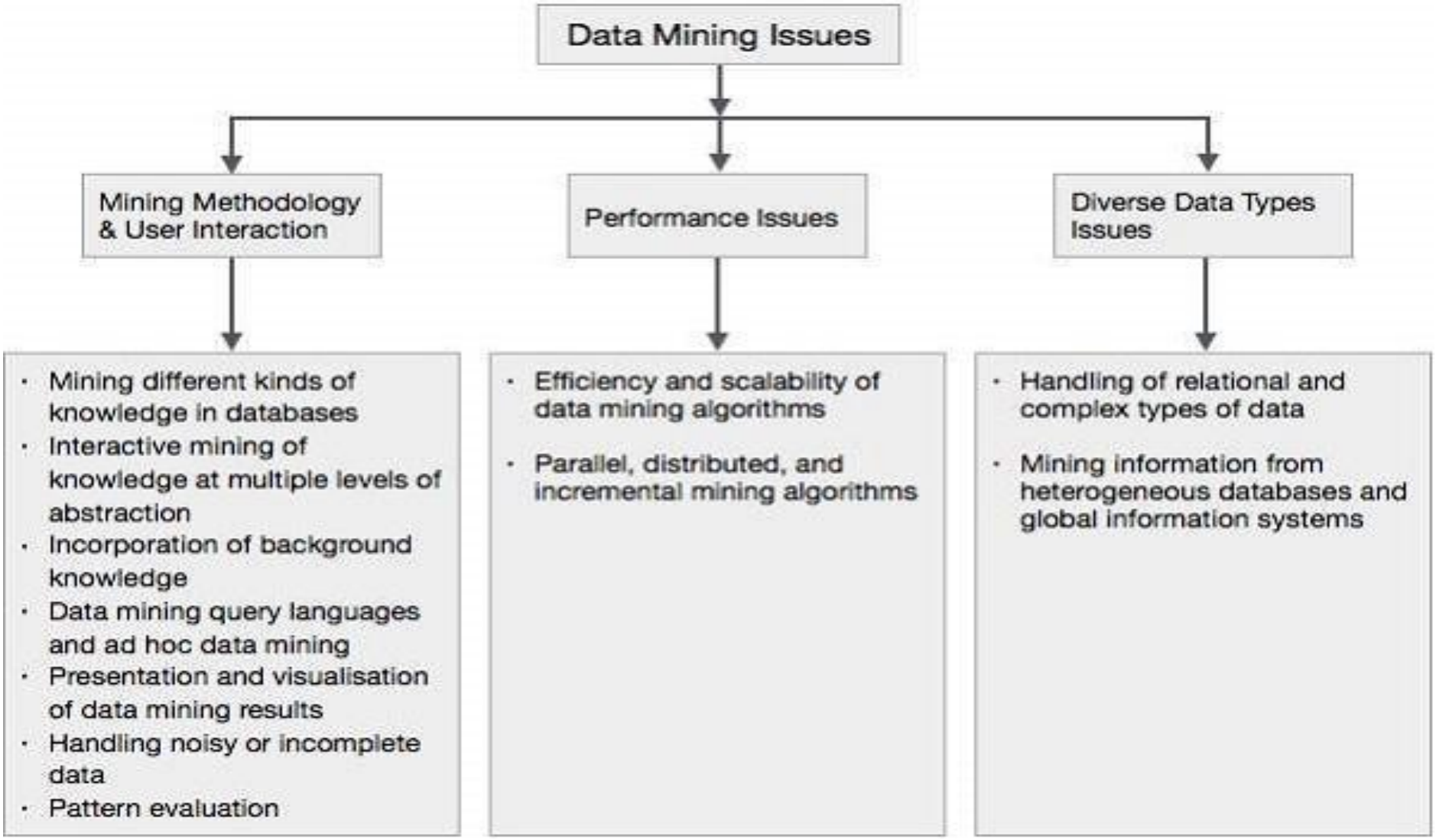
-To analyze viewer data to determine which programs to show during prime time.
-To decide where to insert advertisements so as to maximize the returns.
-To predict financial success of the movies before they are produced.

**9. Sports:**

-To improve performance of NBA teams in US
-To increase the chances of winning

*

# Data Mining Issues



# Self Learning Topics

1. Data Marts

2. Major issues in Data Mining