DATA MINING
&
DATA WAREHOUSING

Module II

- **Data mining – What is?**
- **KDD vs data mining,**
- **DBMS vs data mining,**
- **DM Techniques,**
- **Issues and challenges,**
- **Applications.**

- **The database system industry has witnessed an evolutionary path in the development of the following functionalities (Figure in next slide)**
  - **data collection and database creation,**
  - **data management (including data storage and retrieval, and database transaction processing),**
  - **advanced data analysis (involving data warehousing and data mining)**

Data Collection and Database Creation
(1960s and earlier)
• Primitive file processing

Database Management Systems
(1970s–early 1980s)
• Hierarchical and network database systems
• Relational database systems
• Data modeling tools: entity-relational models, etc.
• Indexing and accessing methods: B-trees, hashing, etc.
• Query languages: SQL, etc.
• User interfaces, forms and reports
• Query processing and query optimization
• Transactions, concurrency control and recovery
• On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s–present)
• Advanced data models: extended relational, object-relational, etc.
• Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s–present)
• Data warehouse and OLAP
• Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
• Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
• Data mining and society: privacy-preserving data mining

Web-based databases
(1990s–present)
• XML-based database systems
• Integration with information retrieval
• Data and information integration

New Generation of Integrated Data and Information Systems
(present–future)

Success is the sum of small efforts repeated day in and day out

What is data mining

- **Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques**

Other Definitions

- **Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions.**
- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner”
- **Data mining is an interdisciplinary field bringing togther techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases**

Why Data Mining

- Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data.
- Within these masses of data lies hidden information of strategic importance.
- due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge.
- to control costs as well as contribute to revenue increases.

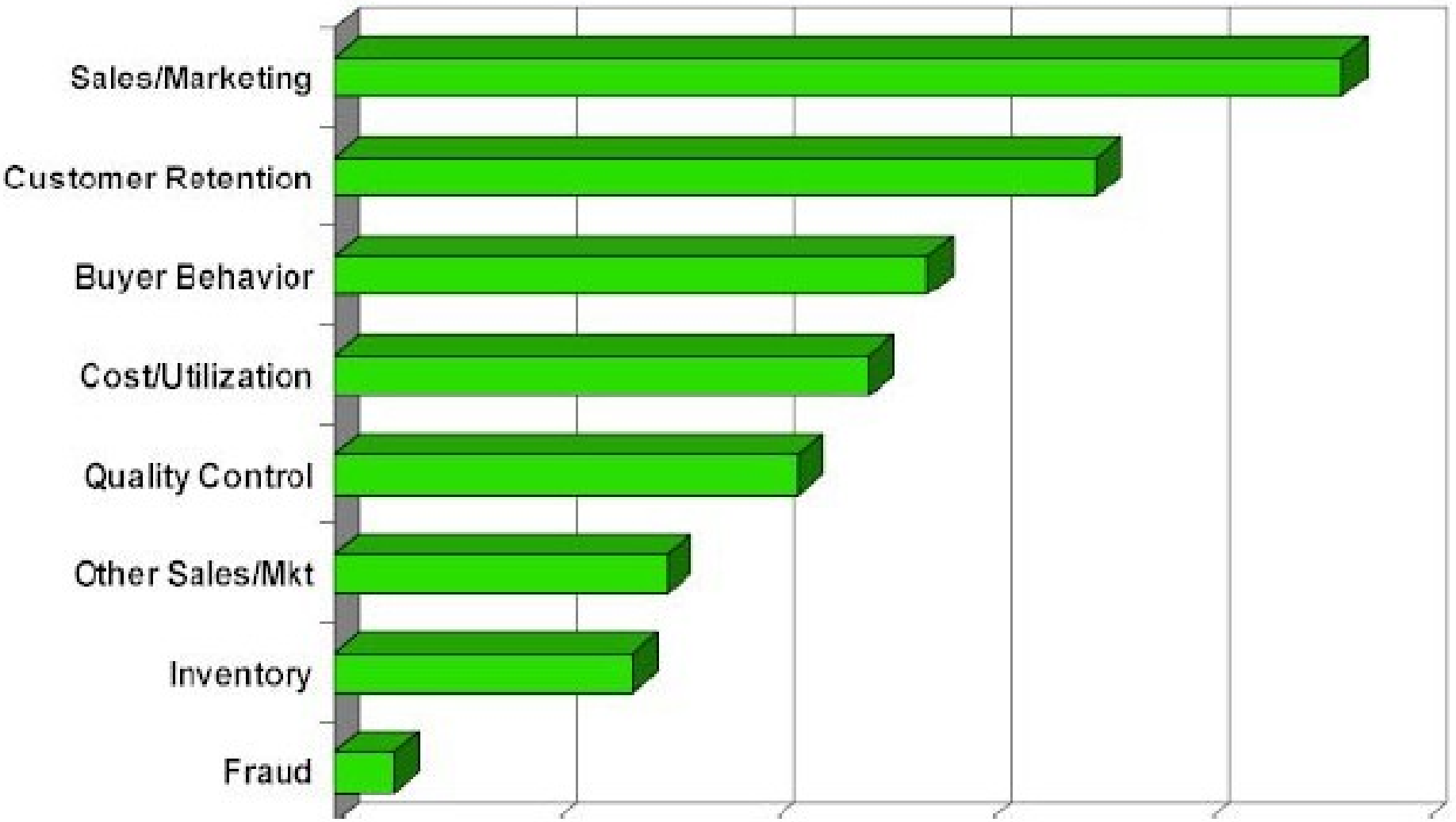
- **Many other terms carry a similar or slightly different meaning to data mining, such as**
  - **knowledge mining from data,**
  - **knowledge extraction,**
  - **data/pattern analysis,**
  - **data archaeology,**
  - **data dredging.**

- **On what kind of data , Data Mining is done**
  - **Relational Databases**
  - **Data Warehouses**
  - **Transactional Databases**
  - **Advanced Data and Information Systems and Advanced Applications**
  - **Object-Relational Databases**
  - **Temporal Databases, Sequence Databases & Time-Series Databases**
  - **Spatial Databases and Spatiotemporal Databases**
  - **Text Databases and Multimedia Databases**
  - **Heterogeneous Databases and Legacy Databases**
  - **Data Streams**
  - **The World Wide Web**

Applications of Data Mining

- **Marketing**
  - Analysis of consumer behavior
  - Advertising campaigns
  - Targeted mailings
  - Segmentation of customers, stores, or products
- **Finance**
  - Creditworthiness of clients
  - Performance analysis of finance investments
  - Fraud detection
- **Manufacturing**
  - Optimization of resources
  - Optimization of manufacturing processes
  - Product design based on customer requirements
- **Health Care**
  - Discovering patterns in X-ray images
  - Analyzing side effects of drugs
  - Effectiveness of treatments

Data Mining Applications



Advantages Of Data Mining

- **Marking/Retailing: Data mining can aid direct marketers by providing them with useful and accurate trends about their customers' purchasing behavior.**
- **Banking/Crediting: Data mining can assist financial institutions in areas such as credit reporting and loan information.**

Advantages Of Data Mining

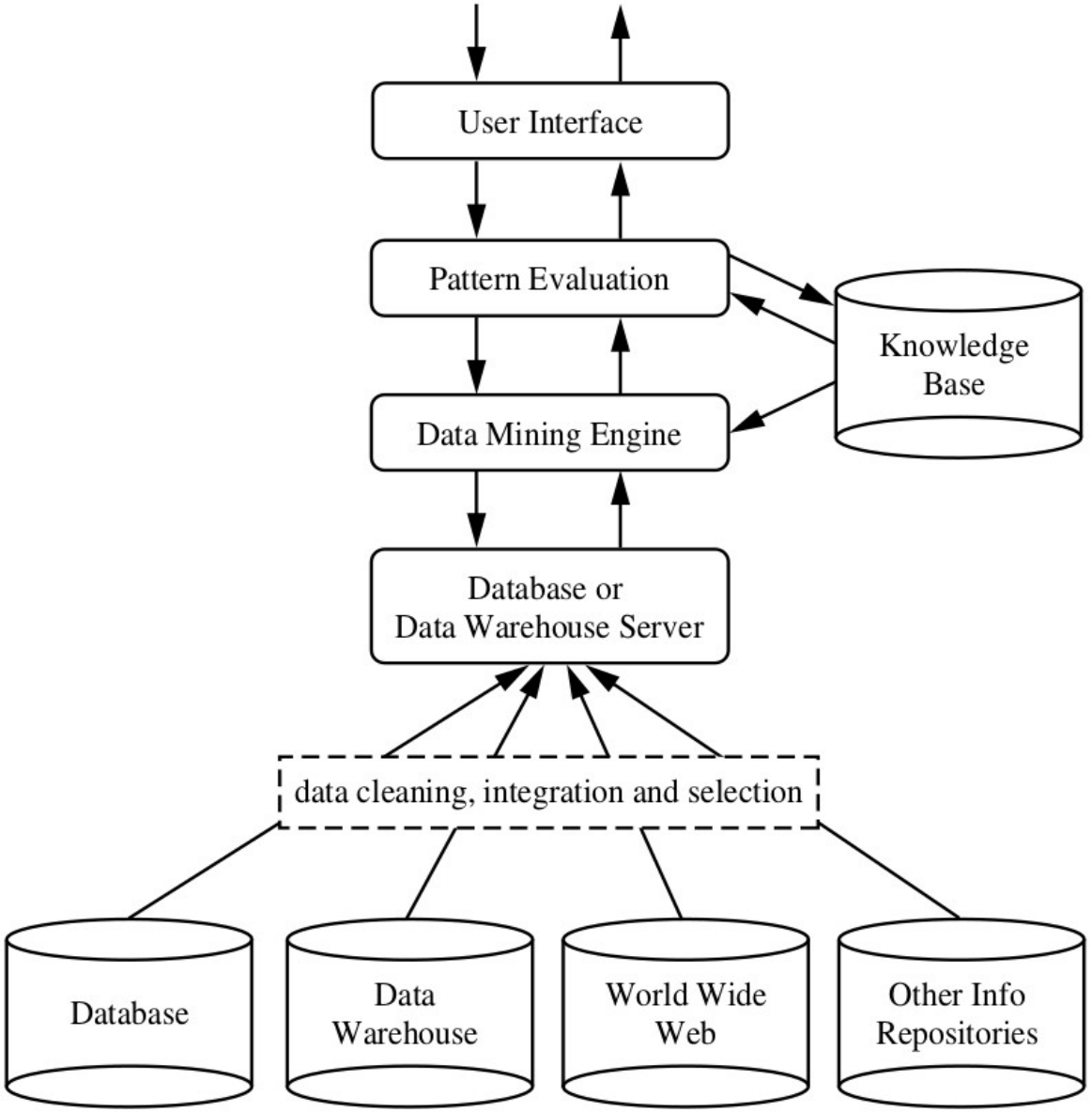
- **Law enforcement: Data mining can aid law enforcers in identifying criminal suspects as well as apprehending these criminals by examining trends in location, crime type, habit, and other patterns of behaviors.**
- **Researchers: Data mining can assist researchers by speeding up their data analyzing process; thus, allowing them more time to work on other projects.**

Disadvantages Of Data Mining

- **Privacy Issues:**
  - **Selling patient's prescription purchases to a different company**
  - **Selling customers' credit card purchases to another company.**
- **Security issues:**
  - **do not have sufficient security systems in place to protect that information.**
- **Misuse of information**

Typical architecture of data mining system

Architecture of a typical data mining system.



Typical data mining system may have these 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 collecting 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.
  - It includes concept hierarchies,

    used to organize attributes or attribute values into different levels of abstraction.
  - Other domain knowledge are additional constraints and metadata
- **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.

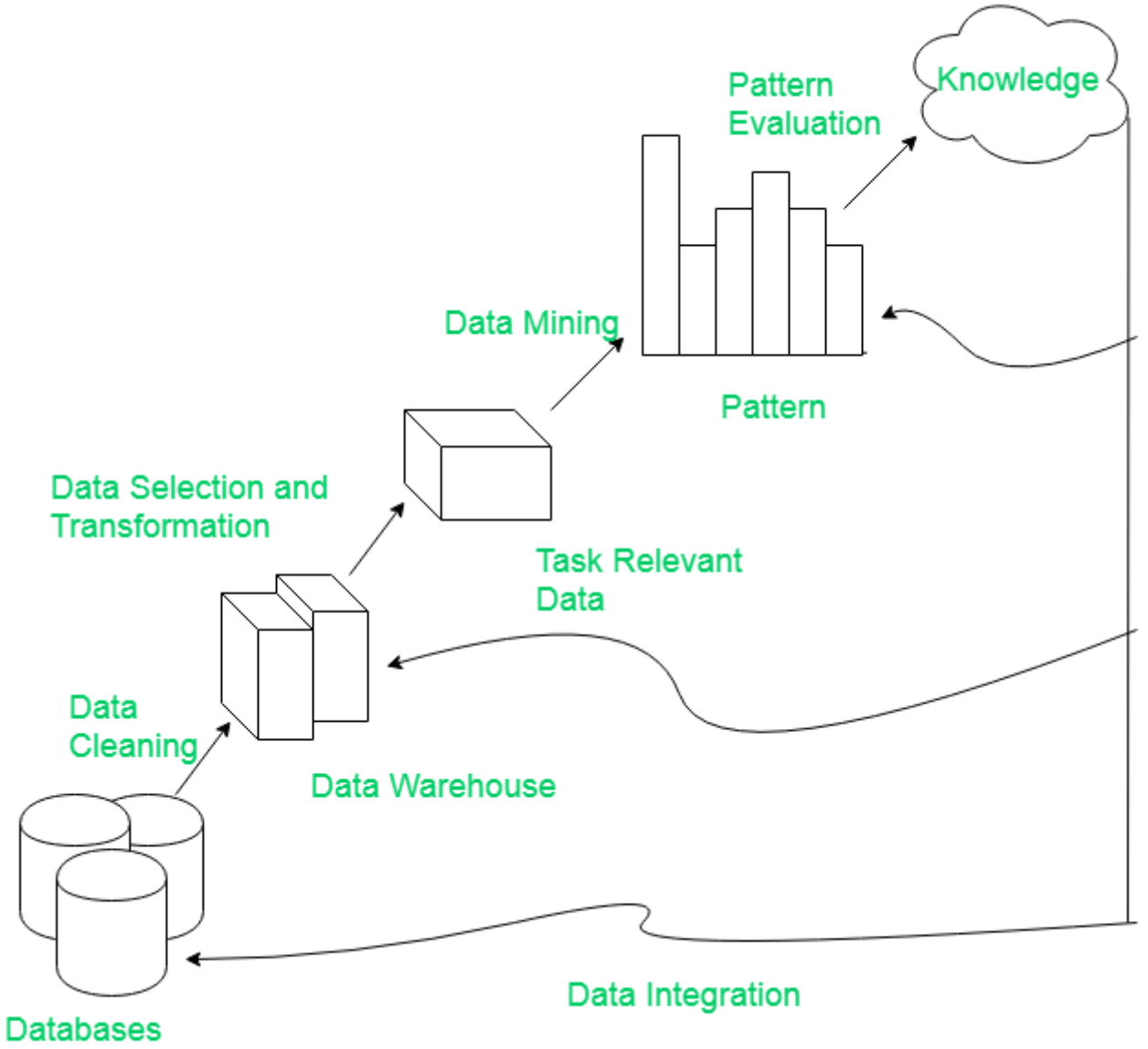
- **Pattern evaluation module:**
  - **This employs interestingness measures and interacts with the data mining modules so as to focus the search toward interesting patterns.**
  - **It may use interestingness thresholds to filter out discovered patterns.**
- **User interface:**
  - **This module communicates between users and the data mining system**
  - **allowing the user to interact with the system**

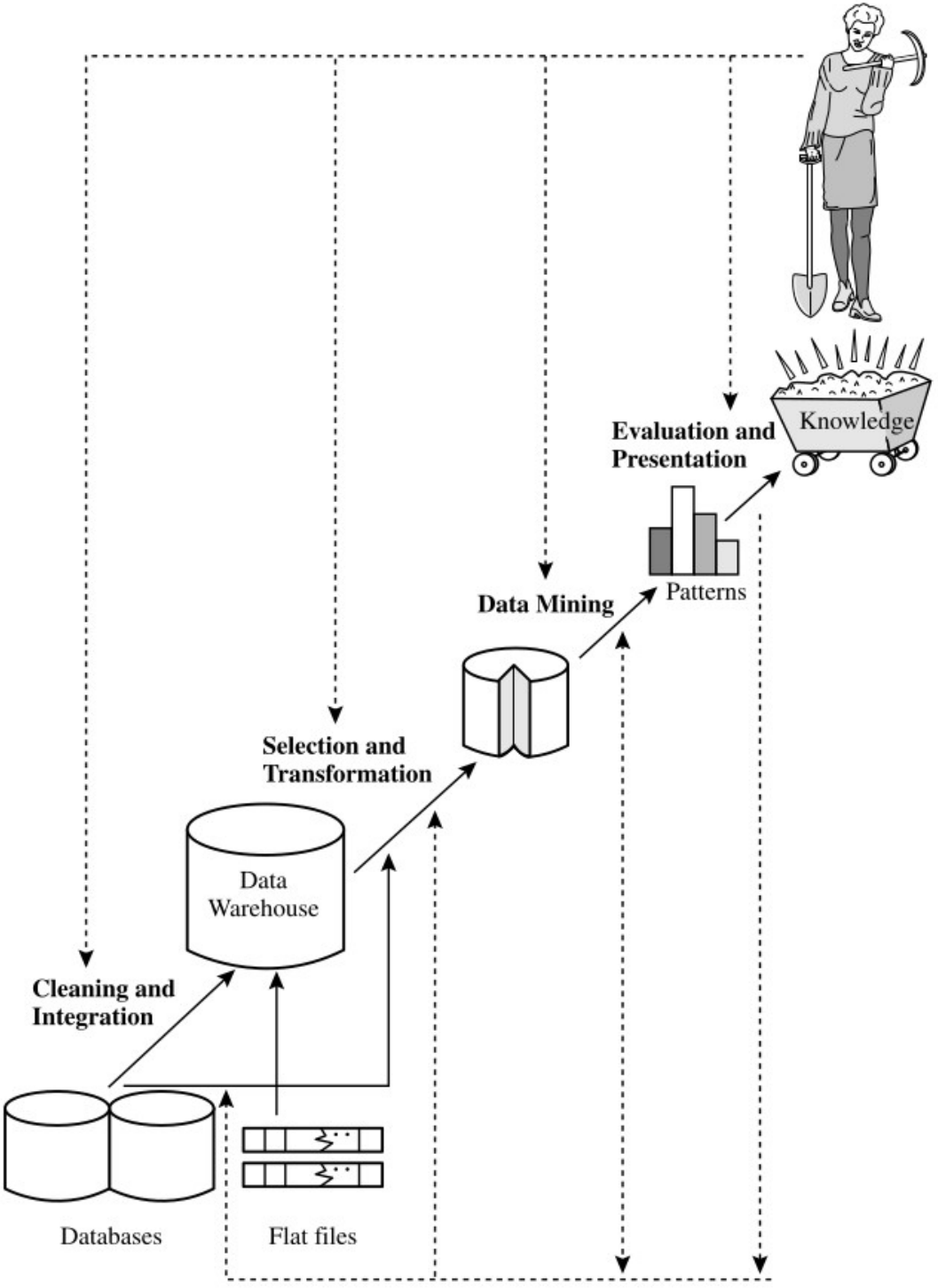
KDD vs Data Mining

DBMS vs Data Mining

KDD

-
-
- **Knowledge discovery as a process**
- **is depicted in Figure**


Knowledge
Pattern Evaluation
Data Mining
Pattern
Data Selection and Transformation
Task Relevant Data
Data Cleaning
Data Warehouse
Data Integration
Databases



Success is the sum of small efforts repeated day in and day out

KDD consists of an iterative sequence of the following steps .....

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

- **Data integration (where multiple data sources may be combined)**
  - **It is defined as heterogeneous data from multiple sources combined in a common source(DataWarehouse).**
  - **Data integration using Data Migration tools.**
  - **Data integration using Data Synchronization tools.**
  - **Data integration using ETL(Extract-Load-Transformation) process.**

- **Data selection (where data relevant to the analysis task are retrieved from the database)**
  - **It is the process where data relevant to the analysis is decided and retrieved or segmented from the data collection.**
  - **Data selection using Neural network.**
  - **Data selection using Decision Trees.**
  - **Data selection using Naive bayes.**
  - **Data selection using Clustering, Regression, etc.**

- **Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)**
  - **Data Transformation is defined as the process of transforming data into appropriate form required by mining procedure.**
  - **Data Transformation is a two step process:**
    - **Data Mapping: Assigning elements from source base to destination to capture transformations.**
    - **Code generation: Creation of the actual transformation program.**

- **Data mining (an essential process where intelligent methods are applied in order to extract data patterns)**
  - **Data mining is defined as techniques that are applied to extract useful patterns**
  - **Transforms task relevant data into patterns.**
  - **Decides purpose of model using classification or characterization**

- **Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)**
  - **It is defined as as identifying strictly increasing patterns representing knowledge based on given measures.**
  - **Find interestingness score of each pattern.**
  - **Uses summarization and Visualization to make data understandable by user.**

- **Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)**
  - **Knowledge representation is defined as technique which utilizes visualization tools to represent data mining results.**
  - **Generate reports.**
  - **Generate tables.**
  - **Generate discriminant rules, classification rules, characterization rules, etc.**

Knowledge Discovery in Databases is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting or sampling data from a data warchouse, cleaning or preprocessing it, transforming or reducing it (if needed), applying a data mining component to produce a structure, and then evaluating the derived structure.

Data Mining is a step in the KDD process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations.

Thus, the structures that are the outcome of the data mining process must meet certain conditions so that these can be considered as knowledge. These conditions are: *validity, understandability, utility, novelty* and *interestingness*.

- **DBMS :**
  - **It is a Database Management System.**
  - **It’s basically a product or software which is used to manage data. Eg. SQL Server.**
- **Data Mining:**
  - **Data Mining is the procedure to extract the information from huge amount of raw data**
    - **which can be used to take business decisions or to decide future strategies.**

- **DBMS**
  - **is a full-fledged system for housing and managing a set of digital databases.**
  - **There are four important elements in any DBMS. They are the modeling language, data structures, query language and mechanism for transactions**
    - **The modeling language defines the language of each database hosted in the DBMS. Popular approaches like hierarchal, network, relational and object are in practice.**
    - **Data structures help organize the data such as individual records, files, fields and their definitions and objects such as visual media.**
    - **Data query language maintains the security of the database by monitoring login data, access rights to different users, and protocols to add data to the system. SQL is a popular query language that is used in RDBMS**
    - **Finally, the mechanism that allows for transactions help concurrency and multiplicity.**

- **Data Mining**
  - **is a technique for extracting useful and previously unknown information from raw data.**
  - **these raw data are stored in very large databases.**
  - **Data Mining use the existing functionalities of DBMS to handle, manage and even preprocess raw data before and during the Data mining process.**
  - **DM is very important tool to convert this large wealth of data in to business intelligence**
    - **Bcos manual extraction of patterns has become seemingly impossible in the past few decades.**

Database

The database is the organized collection of data. Most of the times, these raw data are stored in very large databases.

A Database may contain different levels of abstraction in its architecture.

Typically, the three levels: external, conceptual and internal make up the database architecture.

Data mining

Data mining is analyzing data from different information to discover useful knowledge.

Data mining deals with extracting useful and previously unknown information from raw data.

The data mining process relies on the data compiled in the data warehousing phase in order to detect meaningful patterns.

Data Mining Techniques

TASKS done by DATA MINING

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association

Goals of Data Mining and KDD

- **Prediction** e.g. sales volume, earthquakes
- **Identification** e.g. existence of genes, system intrusions
- **Classification** of different categories e.g. discount-seeking shoppers or loyal regular shoppers in a supermarket

Goals of Data Mining and KDD

- Optimization of limited resources such as
  - time, space, money or materials
- maximization of outputs such as
  - sales or profits

DM Techniques

- **Two major goals of DM**
  - **Prediction**
  - **Description**

DM Techniques

- **Broadly classified into**
  - **Verification-driven or user-guided**
  - **Discovery-driven or automatic discovery**

DM Techniques

- **Verification model**
  - **Hypothsis is the emphasis**
    - **User makes hypothesis**
    - **And test the hypothesis on data to verify its validity**
  - **The emphasis is on the user who is responsible for formulating the hypothesis**
  - **Refinements of hypothesis until the required limit is reached**
  - **New hypothesis or refine existing one and verify against the DB**

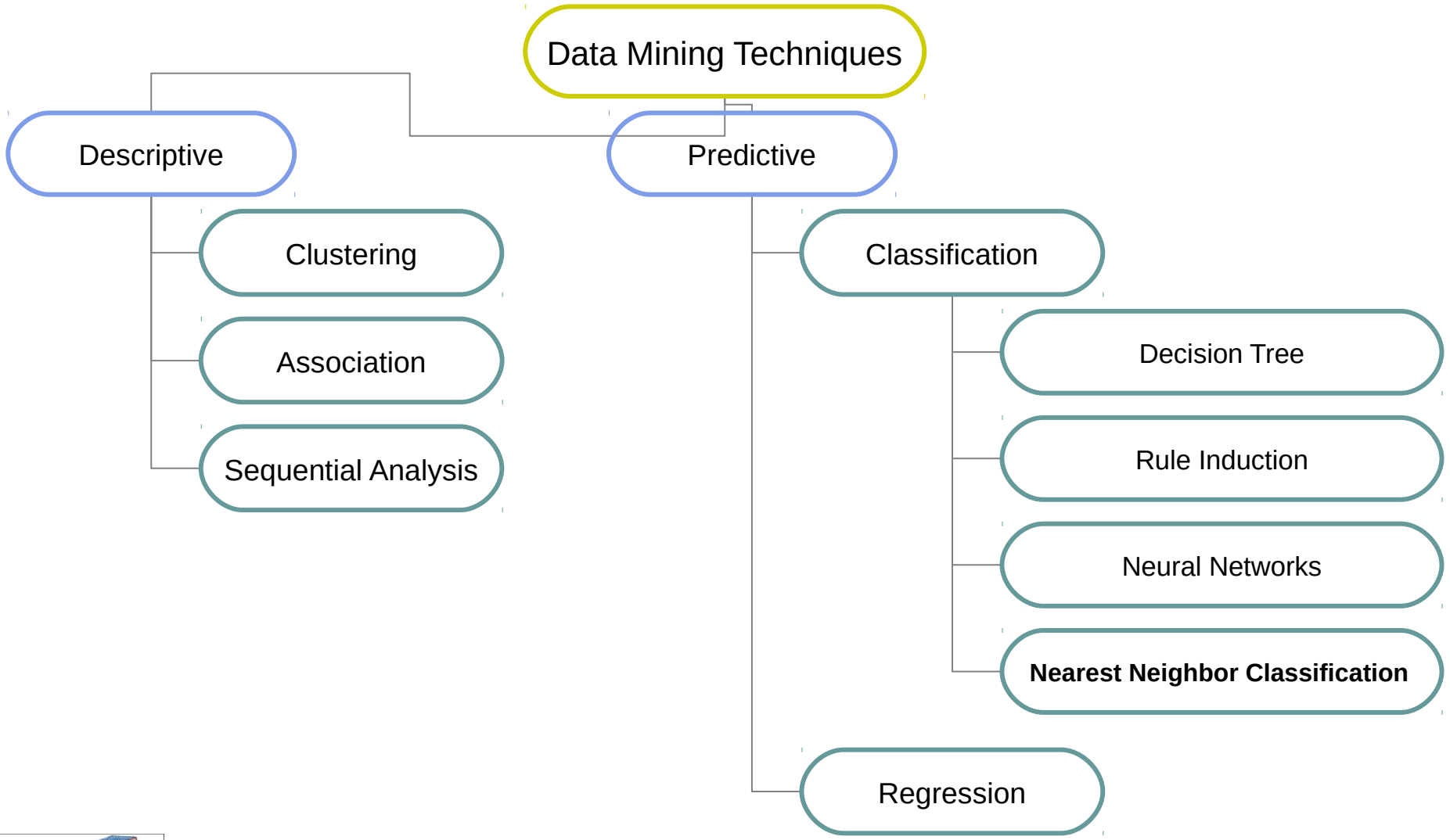
DM Techniques

- **Discovery-driven**
  - **Automatically discovers the info hidden in the data**
  - **Discover hidden information by searching the**
    - **Fequent patterns & Trends**
    - **Generalizations about the data**
  - **No intervention of user**
  - **Discovery depends on the DM applications**
    - **Typical discovery tasks are .....**

Discovery driven tasks

- **Discovery of Association rules**
- **Discovery of Classification rules**
- **Clustering**
- **Discovery of Frequent episodes**
- **Deviation detection**

Data Mining Techniques



Discovery of Association rule

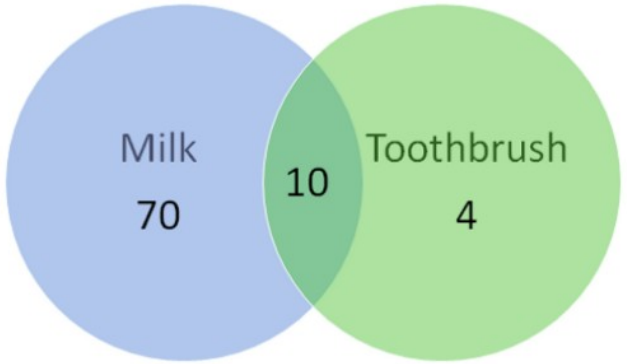
- Market-basket analysis is a well-known example of association discovery
- Associations are written as A ==> B, where A & B are the sets of items
  - Means
    - The transactions of the DB which contanin A tends to contain B
    - Given a DB, the goal is to discover all the rules that have the support and confidence >= min-support & min-confidence
- Set of items, Database, Transactions
- Support and Confidence – the measures used in this technique
- Frequent itemsets – one of the inputs to be generated
- Low Intermediate High ---------- values

Transaction DB

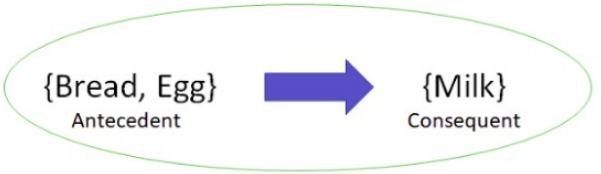
TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Association Rule

```
{Milk, Diaper}->{Beer}
```



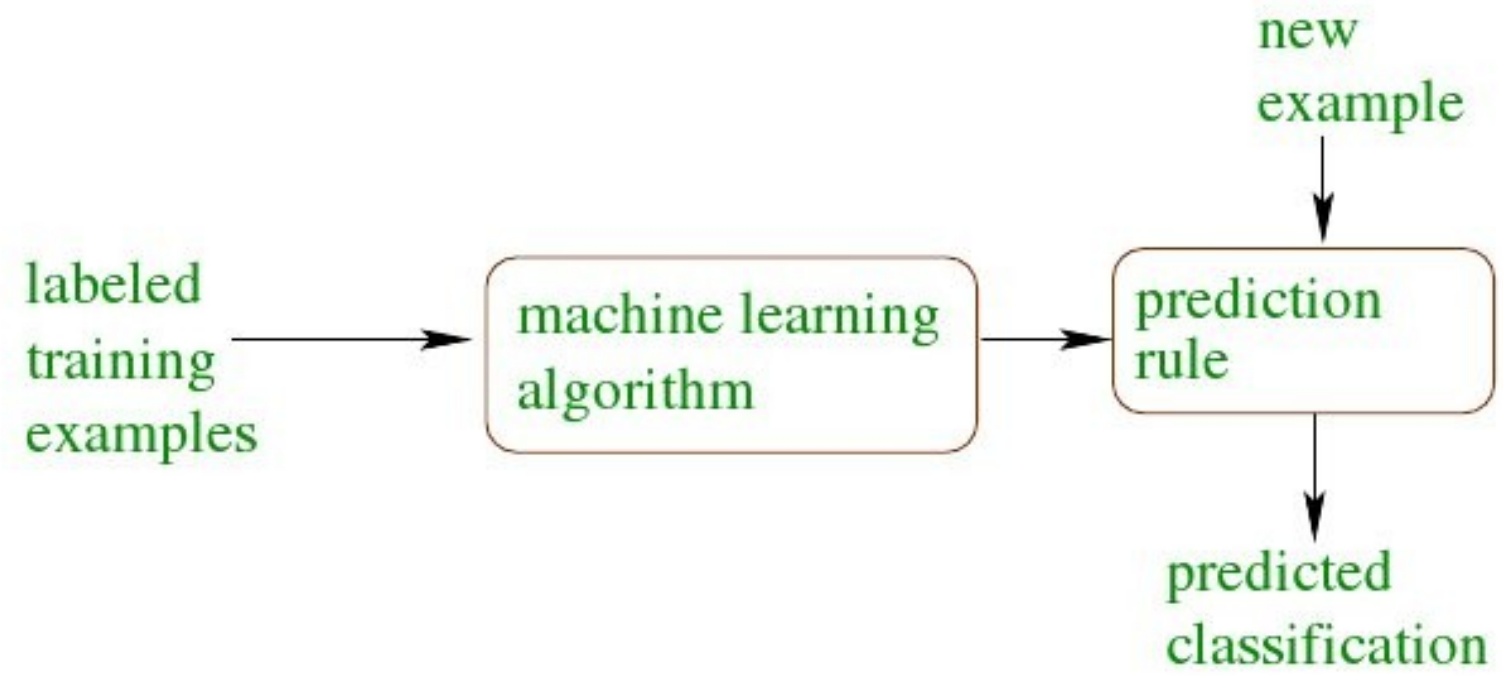
Total transactions = 100. 10 of them have both milk and toothbrush, 70 have milk but no toothbrush and 4 have toothbrush but no milk.



Itemset = {Bread, Egg, Milk}

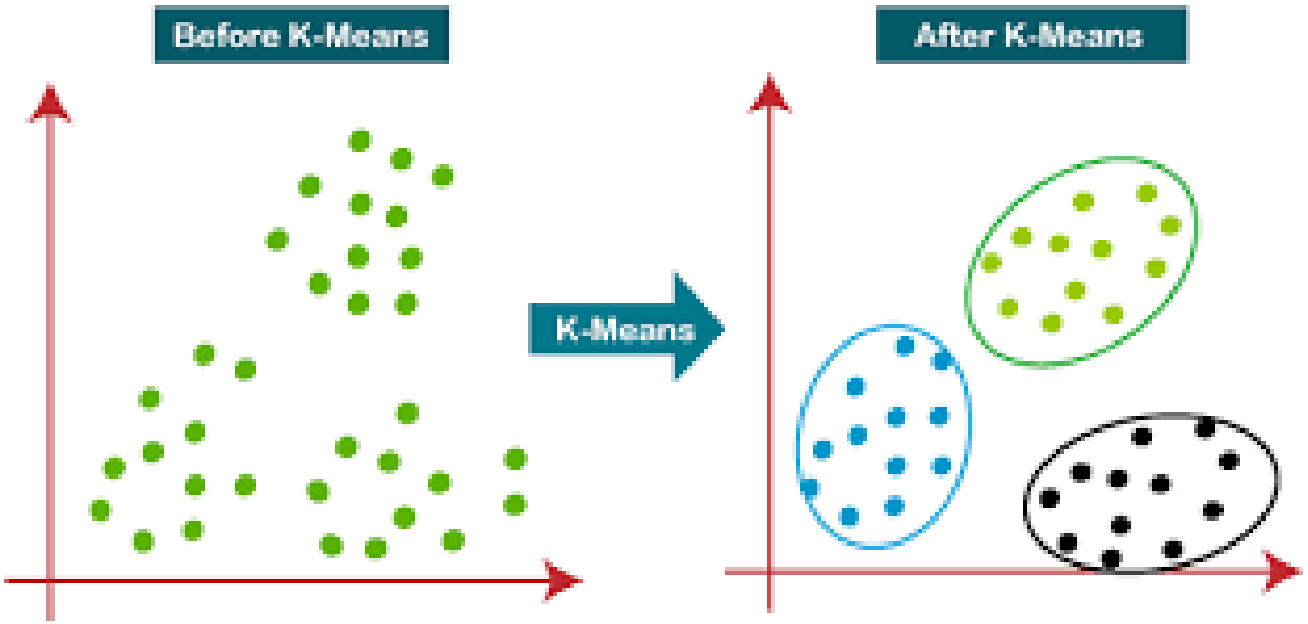
Discovery of Classification rule

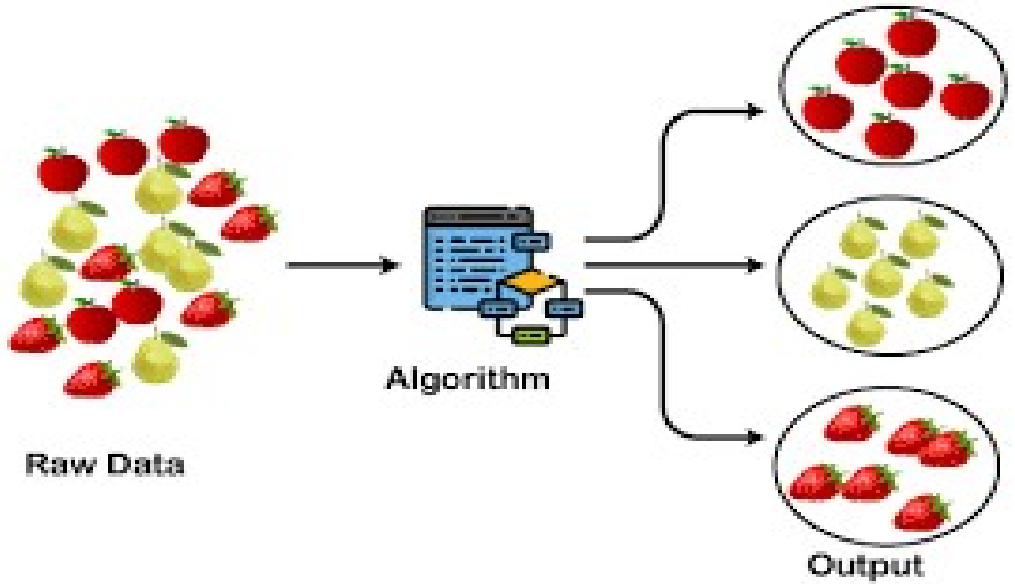
- Finding the rules that Partiotion the data into disjoint groups
- For this process
  - Input is the Training data set
    - Whose class label is already known
  - Then construct a model to predict the class of a new object
- This is Supervised learning model
- Some of the classification discovery models are
  - Decion trees, neural networks, genetic algo and sttistical models
- e.g credit card analysis, banking, medical applications


new
example
labeled
training
examples
machine learning
algorithm
prediction
rule
predicted
classification

Clustering

- Grouping data that share
  - Similar patterns, similar trends
- The clustering algo attempts to
  - Automatically partition the data into regions or clusters
  - Which is done either deterministic/probability-wise
- For partiotioning ,
  - similarities and differences of data will be identified
  - the following approaches are used
    - Measure of similarity
      - Many techniques are there for clustering
    - Set of functions that measure some particular property or groups (this approach is known as optimal partitioning)
- Objectives of clustering are
  - Uncover natual grouping
  - Initiate hypothesis about the data
  - Find consistent and valid organization of data


Before K-Means
After K-Means
K-Means


Raw Data
Algorithm
Output

Discovery of Frequent episodes

- What are Frequent episodes
  - Sequence of events occuring frquently close to each other
  - Are extracted from the time sequences
- How frequent
  - is Domain dependent
  - given by the user as the input
  - Pediction rules for the time sequence is the output
- Starting and ending time of a sequence of event

R is a set of event types

A is a particular type of event

Therefore $A \in R$

An event is defined as a pair (A, t) , where as above

$A \in R$

A sequence of events (also called event sequence) S of R is a triple (Ts, Tc, S)

Where Ts = starting time

Tc = ending time

S= {(A1,t1), (A2,t2), … … … (An, tn) } is the ordered sequence of events, such that

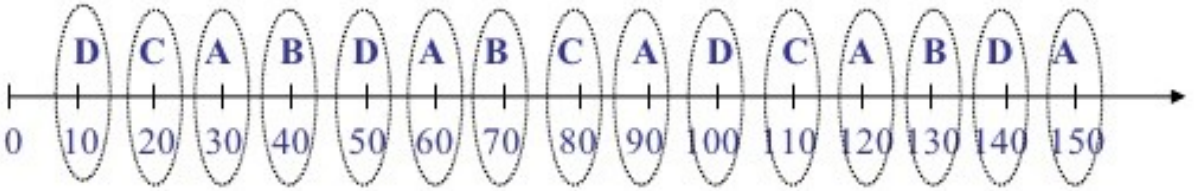
$Ai \in R$

and

$Ts <= ti <= Tc$ for all i = 1,2, … … … n-1

Example of Episode Mining

- **Alarm data sequence:**



- **Here:**
  - *A, B, C* and *D* are event (or here alarm) types
  - *10…150* are occurrence times
  - $s = \langle (D, 10), (C, 20), \ldots, (A, 150) \rangle$
  - T_s (starting time) = 10 and T_e (ending time) = 150
- **Note: There needs <u>not</u> to be events on every time slot!**

- **Eg**
- **In the telecommunication alarm management, where thousands of alarms accumulate daily;**
  - **there can be hundreds of different alarm types.**
- **When discovering episodes in a telecommunication network alarm log, the goal is to find relationships between alarms.**
- **Such relationships can then be used in the on-line analysis of the incoming alarm stream**
  - **To better explain the problems that cause alarms,**
  - **To suppress redundant alarms,**
  - **To predict severe faults**
  -

- **Episodes can be**
  -

a) Serial episodes: Which occur in sequence.

b) Parallel episodes: No constraints on the occurrence of event types.

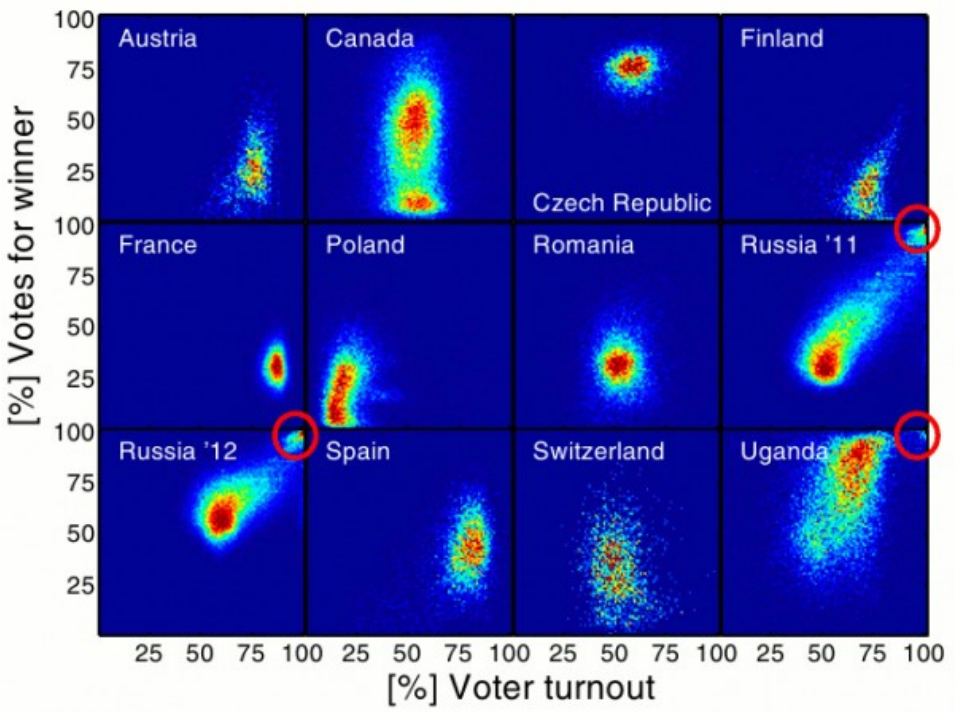
c) Non serial non parallel: If the occurrences of A and B preceed an occurrence of C, and there is no constraint on the occurrences of A and B

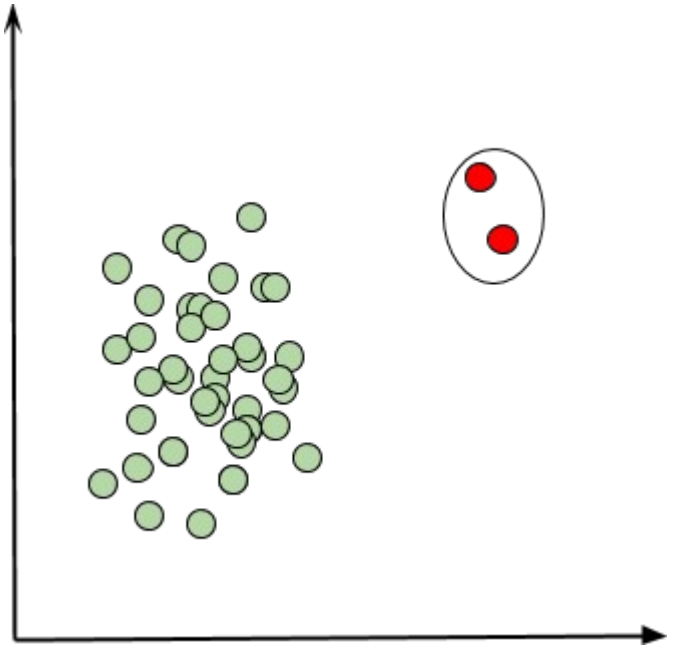
- **Used for temporal data**
  - **E.g Telecommunications, share market analysis**

Deviation detetion

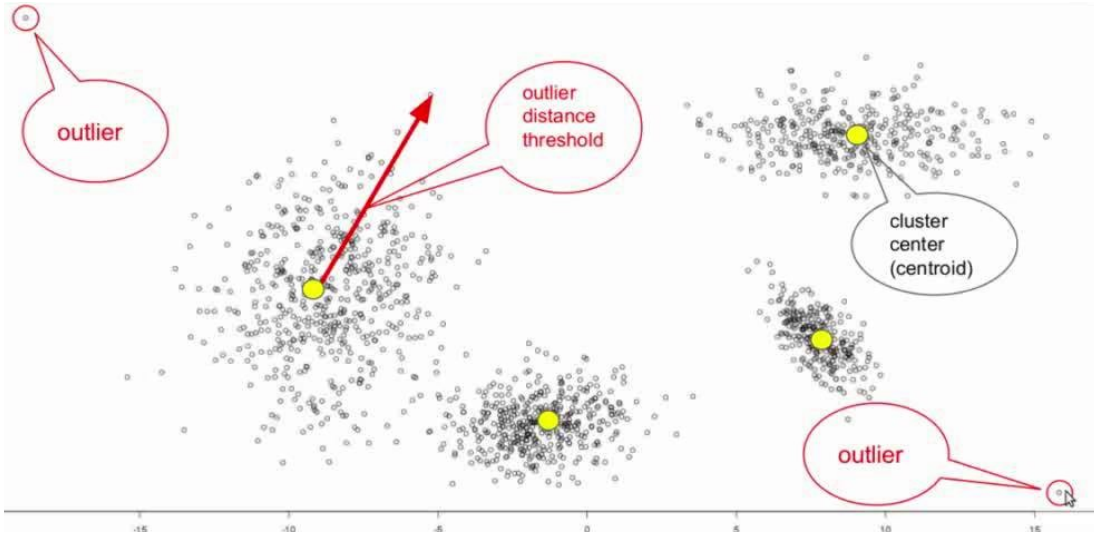
- Deviation detection is to identify outlying points in a particular data set, and explain whether they are due to noise or other impurities being present in the data or due to trivial reasons

Deviation detetion





Outlier
Anamoly



Success is the sum of small efforts repeated day in and day out

Other DM techniques

- Neural networks
- Support vector machines
- Genetic algorithms
- Rough sets techniques
- Regression
- Link Analysis