Data Mining Definitions

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data stored in database, data warehouse or other information repositories.

It is also defined in different five terms:

(Information Perspective)

- Data mining or knowledge discovery in databases, as it is also known, is the nontrivial extraction of implicit, previously unknown and potentially useful information from data.

(Relationship from large databases)

- Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among vast amount of data.

(DM techniques related)

- Data mining refers to using a variety of techniques to identify nuggets of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation.

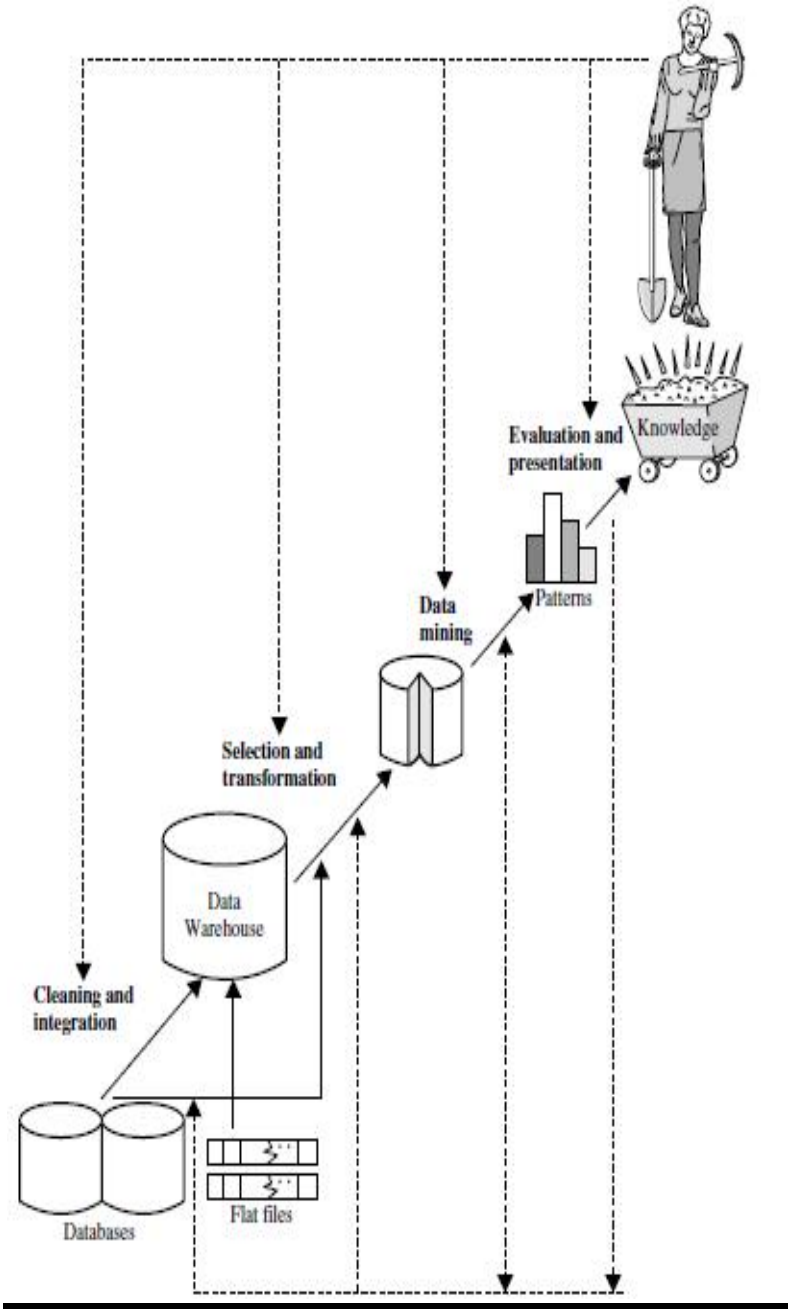
(Discovery of rules)

- Discovering relations that connect variables in a database is the subject of mining. The data mining system self-learns from previous history of the investigated system, formulating and testing hypothesis, etc.

(Correlation)

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

KDD 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 and consolidated into forms appropriate for mining by performing summary or aggregation operations.

5) **Data mining**
   - an essential process where intelligent methods are applied to extract data patterns.

6) **Pattern evaluation**
   - to identify the truly interesting patterns representing knowledge based on interestingness measures.

7) **Knowledge presentation**
   - where visualization and knowledge representation techniques are used to present mined knowledge to users.

KDD vs DATA MINING

KDD–

- Knowledge Data Discovery , seeking knowledge from data

DM–

- Technique to which is used to represent and analyze data for decision makers.

DM–

- DM is one step of KDD.

KDD

- It contains data selection, data cleaning, data transformation, data mining, data interpretation and visualization.

KDD–

- KDD process tends to be highly iterative and interactive.

DM–

- DM tends to work up from data and develop best technique suitable for large volume of data, decision making and conclusions.

KDD–

- It is a process of identifying valid, potentially useful and understandable structure of in data. This process involves all different steps.

DM–

- Step in KDD process concerned with the algorithmic approach, where patterns or structures are defines from data under acceptable computational limitations.

DBMS Vs DM

DBMS - It supports query languages which are useful for query triggered data exploration. **DM -** It supports automatic data exploration.
DBMS - It supports when exact data is known to explore from database. **DM -** It is used when correlation, patterns are required to find out where exact data is not known.
DBMS - SQL **DM -** DMQL

Overall it is defined that Data mining can use DBMS at different Ways.

1) No use of DBMS System.
2) Loosely –Coupled Approach
3) Tightly – Coupled Approach

No Use of DBMS System

- Many DM system do not use any DBMS, they have their own memory and storage management.
- Data is **downloaded** from repository into own memory structure.
- **Adv:** one can optimize memory management specific to DM algorithm.
- **Disadv:** system ignores certain features of DBMS, such as recovery, concurrency, etc.

Loosely – Coupled Approach

- DBMS is used only for data storage.
- Front end of the application is implemented in a host programming language, with embedded SQL for storing and retrieving data.
- **Adv:** data retrieval is easy.
- **Disadv:** this approach does not use query capability of DBMS

Tightly – Coupled Approach

- Portions of application programs are selectively pushed to the database system to perform the necessary computation (e.g. sum, avg, sort, etc.)
- Much processing is done at database end.
- It is different than bringing data from database to data mining area. (i.e. the DM application goes where data naturally resides.)
- **Adv:** avoids performance degradation, and takes full advantage of DBMS technology

DATA MINING TASK PRIMITIVES

- 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.
- LIST OF DATA MINING TASK Primitives:
  1) Task-relevant data
  2) Knowledge type to be mined
  3) Background knowledge
  4) Pattern interestingness measures
  5) Visualization of discovered patterns

Task-relevant data

- It 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.

Knowledge type 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.

Background knowledge

- 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.

Pattern interestingness measures

- 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.

(different methods)

Visualization of 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.

Issues and Challenges in Data Mining

- **Difficulties in data mining can be categorized as:**

  - **Limited information**
    - DB is designed for the purpose other than DM
    - Some times essential attributes may be missing
    - Difficult to discover significant pattern

  - **Noise or missing data**
    - Missing data can be treated as: discard missing values, omit records, infer missing values, add special value
    - Data should be cleaned so that it can be free from errors and missing data

  - **User integration and prior knowledge**
    - Analyst is usually not KDD expert
    - Challenging to provide proper interface to assist users as well as proper selection of appropriate techniques to achieve goal
    - Use of domain knowledge is imp in all steps of KDD process
    - It is convenient to design KDD tool – interactive and iterative

  - **Uncertainty**
    - Severity of error and degree of noise

  - **Size, updates and irrelevant data**
    - DB is large and dynamic – information is added, modified and removed
    - Challenging to keep knowledge up-to-date