Chapter = DATA MINING + Datamining 1 + DM 2 Jaspreet Singh

1. Datamining refers to the mining or discovery of new information in terms of patterns or rules from vast amount of data.

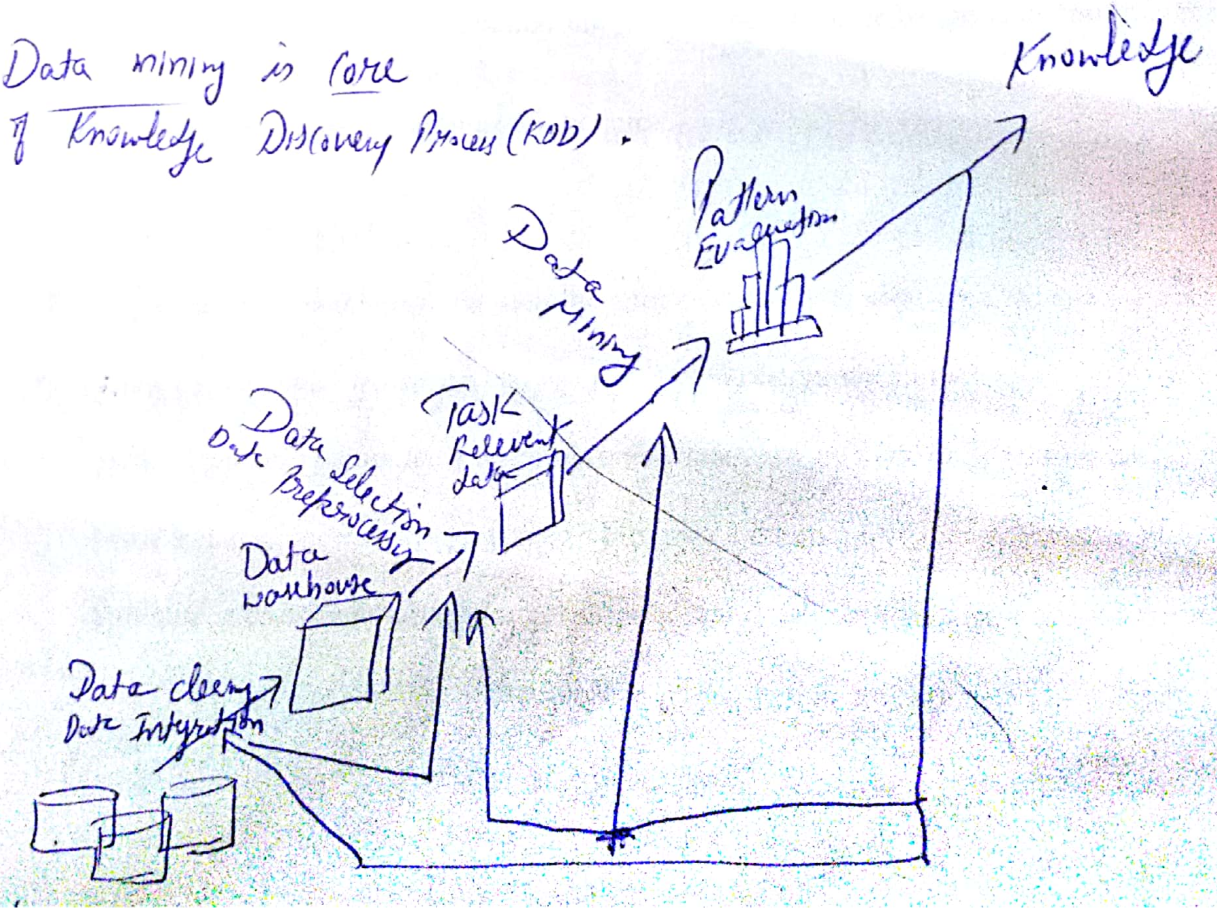
⇒ Extraction of interesting Knowledge (rules, regularities, patterns, constraints) from data in large database.

2. Knowledge Discovery in Database (KDD)

- Data Mining is a part of KDD process.
- The KDD process comprises Six Phases:
(Data Selection, Data ~~Cleansing~~ cleansing, enrichment, data transformation or encoding, data mining, and the reporting and display of the discovered information.

DATA Mining : A KDD Process.

- Data mining is core of Knowledge Discovery Process (KDD).



⇒ Goals of Data Mining and Knowledge Discovery.

Prediction, Identification, classification, optimization

1) Prediction:

Data mining can show how certain attributes within the data will behave in the future.

For example: Predictive data mining includes the analysis of buying transaction to predict what consumers will buy under certain discount, how much sales volume a store will generate in given period, profit.

2) Identification:

Data patterns can be used to identify the existence of an item, an event, or an activity.

For example: The area known as authentication is a form of identification. It verify whether a user is indeed a specific user or one from an authorized class, and involves a comparison of parameters or images or signals against a database.

3 Classification:

Data mining can partition the data so that different classes or categories can be identified based on combination of parameters.

For examples: customers in a supermarket can be categorized into. discount seeking shoppers, shoppers in a rush, regular shopper shopper attached to name brands. or infrequent shoppers.

Optimization:

An eventual goal of data mining may be to optimize the use of limited resources such as time, space or materials and to maximize output variables such as sales or profit under a given set of constraints.

4 → Types of Knowledge discovered during Data mining. (Results of Data Mining)

- In term of knowledge, there is a progression from raw data to information to knowledge as we go through additional processing.

Knowledge → ① Inductive knowledge
Knowledge → ② Deductive knowledge

① Inductive Knowledge:

It Discovers new rules and patterns from the supplied data. Eg. Jenny leaves for school at 7am. Jenny is always on time. Jenny assumes, she will always be on time if she leaves at 7am.

② Deductive Knowledge:

Deduce new information based on applying pre-specified logical rules of deduction on given data

Eg. all dolphins are mammals,
all mammals have kidneys.
∴, all dolphins have kidneys.

Knowledge discovered during datamining as follows:-

* Association rules, classification hierarchies, Sequential patterns, Patterns with in time series, clustering.

1) Association Rules:-

These rules correlates the presence of a set of items with another range of values for another set of variables.

Example: When a shopper(customer) buys bread, he is likely to buy butter.

, (Computer, UPS) , (AC, Stablizer)

2) Classification hierarchies:

The goal is to work from an existing set of events or transactions to create a hierarchy of classes.

1) A population may be divided into five ranges of credit worthiness based on a history of previous credit transactions.

Eg: customers can be classified → frequency of visits
→ financing used
→ amount of purchases.

3) Sequential Patterns:-

A sequence of actions or events.

A patient undervent cardiac bypass ~~surgery~~ surgery for blocked arteries and an aneurysm and later developed high blood urea within a year of surgery. He/she is likely to suffer from kidney failure with in next 18 months.

Pattern within time Series:

Similarities can be detected within positions of a time series of data, which is a sequence of data taken at regular regular intervals such as daily sales or daily closing stock prices.

Ex: Two products show the same selling pattern in summer but a different one in winter.

5] Clustering:

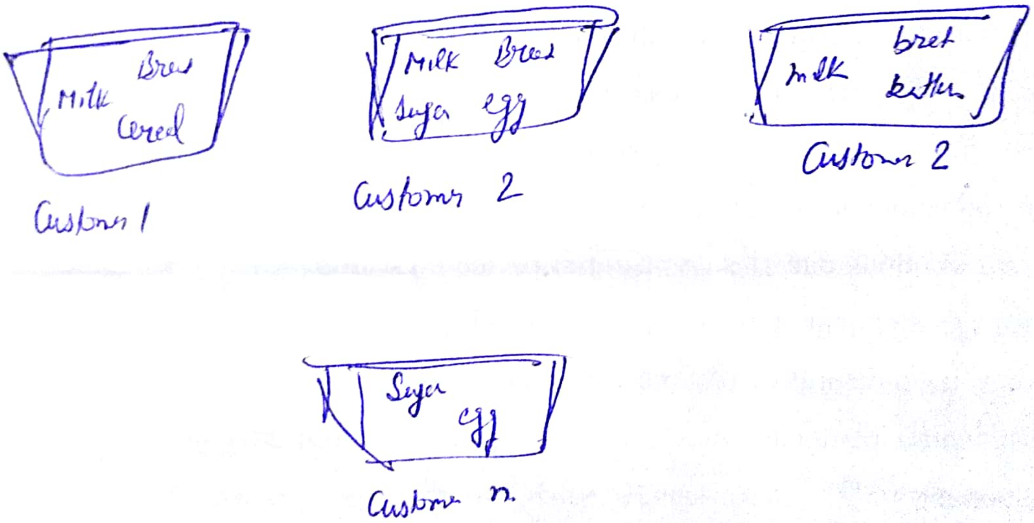
A given population of events or items can be partitioned (segmented) into sets of "similar" elements.

Ex An entire population of treatment data on a disease may be divided into groups bases on the similarity of symptoms (side effects) produced.

Market/Basket Analysis

Which items are frequently purchased together by customers? → To find This we have association Rules

Shopping Basket



Association Rules →

Antecedent ⇒ Consequent [Support, Confidence]

A ⇒ B [S, C]

LHS ⇒ RHS

Factors : 1) Support

2) Confidence

(Support & Confidence are used to measure Interestingness