Data mining – UNIT – I

Motivation

evolution of sciences

- before 1600 → empirical science
  - empirical data is produced by experiment & observation
- 1600–1950's → Theoretical science
  - each discipline has grown a theoretical component
  - These motivate experiments & generalize our understanding
- 1950–1990's → Computational science
  - Computational science involved simulation.
  - It grew out to our inability to find closed form solutions to complex mathematical models
- 1990–now → Data science
  - flood of data from new scientific instruments & simulations
  - ability to stores petabytes of data online
  - Accessibility of internet
  - Data Management

so Data mining is a new major challenge

Data Mining as an evolution of Information Technology

Data mining can be viewed as a result of natural evolution of Information technology. The database and data management industry evolved in development of several critical functionalities

- 1960's & earlier
  - Data collection, database creation, (IMS) Information management System and network DBMS
  - primitive file processing
- 1970's
  - DBMS, relational DBMS
  - Hierarchical & network database systems
  - Relational DBMS
  - SQL
  - Transaction control, concurrency & recovery
  - OLTP (online Transaction processing)
- 1980's
  - Advanced datamodels
  - Advanced DBMS: Spatial, temporal etc.
  - Advanced data Analysis OLAP (online analytical Processing) & data warehousing & data mining
- Data warehousing
  - Subject oriented
  - Integrated
  - non volatile
  - Time variant
- 1980's to 2000
  - Data mining & data warehousing
  - Web databases
  - XML-based data systems
- Present – Future:
  - New generation of Integrated data & Information system

Data Mining (knowledge extraction in database)

"extraction or mining interesting (implicit, previously unknown & useful) information / knowledge / patterns from large amounts of databases"

Alternate Names:

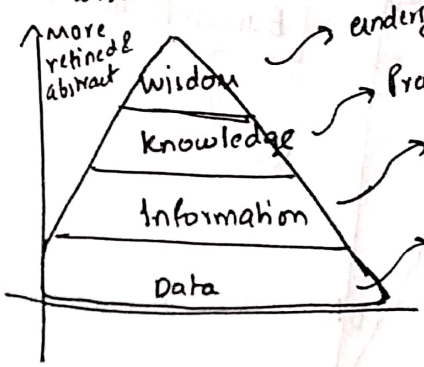
- knowledge discovery
- knowledge extraction
- data Analysis
- Information Harvesting
- Business Intelligence

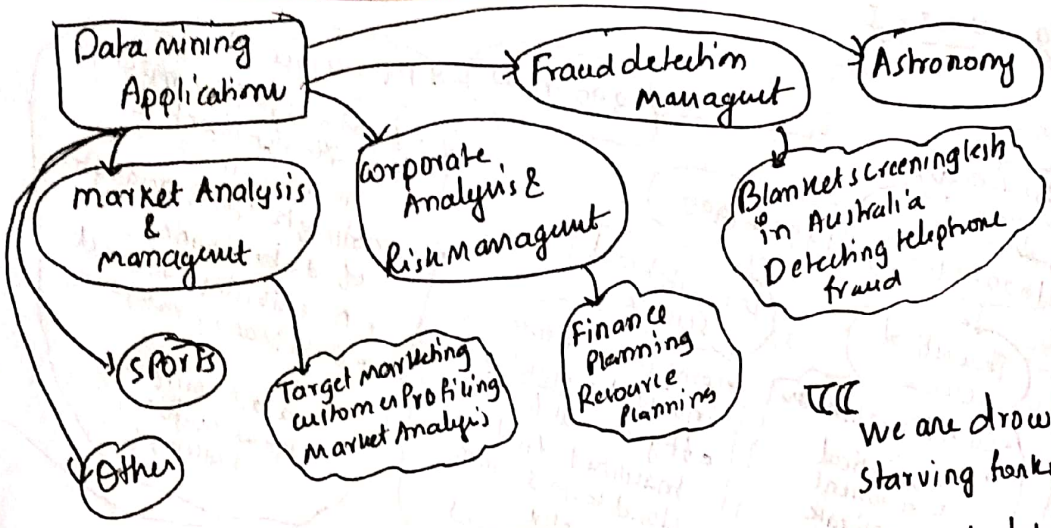
What is not data mining

- Deductive query processing
- Expert systems
- Statistical programs

Information Hierarchy / DIKW pyramid

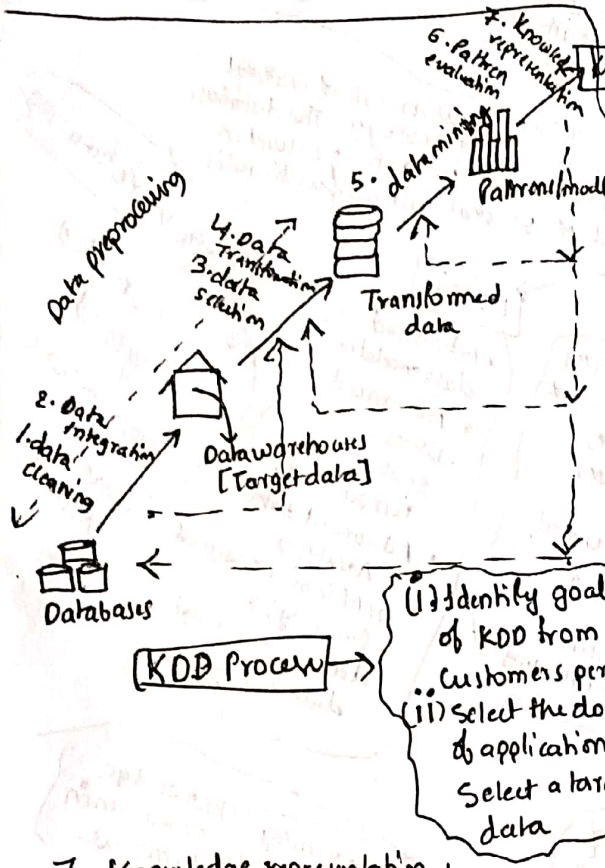
refers to a class of models for representing the structural or functional relationships between data, information, knowledge & wisdom





"We are drowning in data, Starving for knowledge"

"Explosive growth of data from terabytes to zetabytes is typically a data explosion"



1. Data cleaning :- To remove noise & inconsistent data.
2. Data Integration :- where multiple data sources may be combined.
3. Data selection :- where data relevant to analysis tasks are retrieved from databases.
4. Data Transformation :- where the data are transformed and consolidated into new and common formats which are appropriate for mining

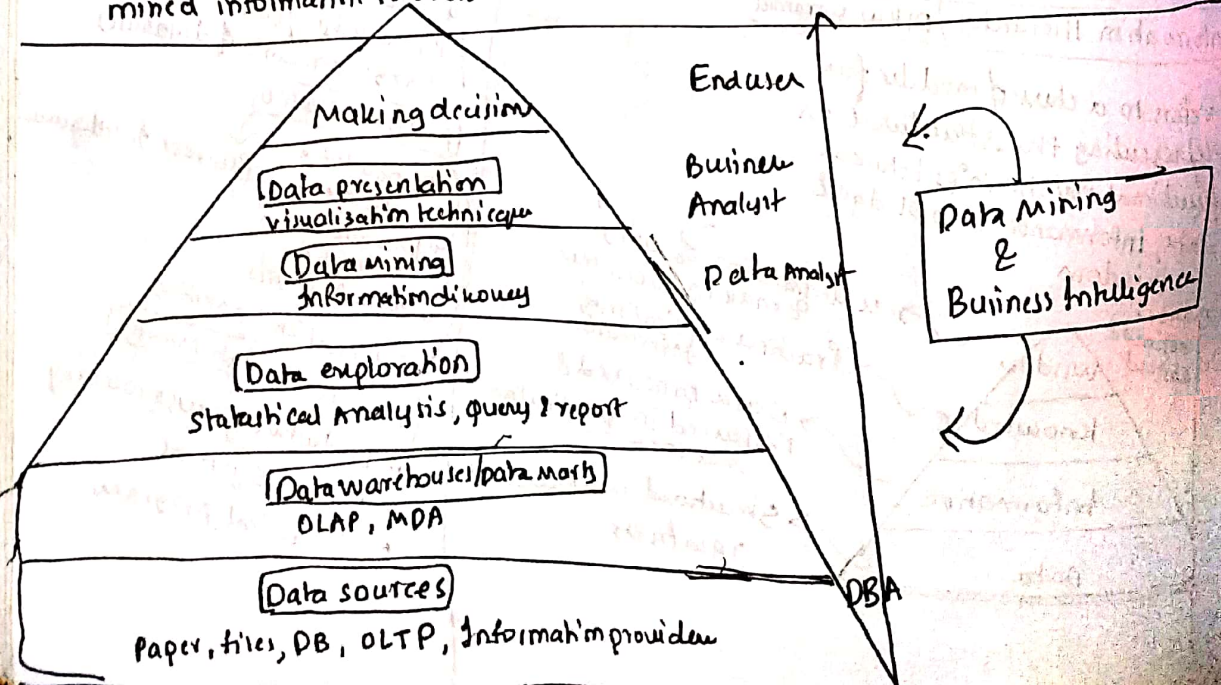
*Data Preprocessing : steps 1 to 4

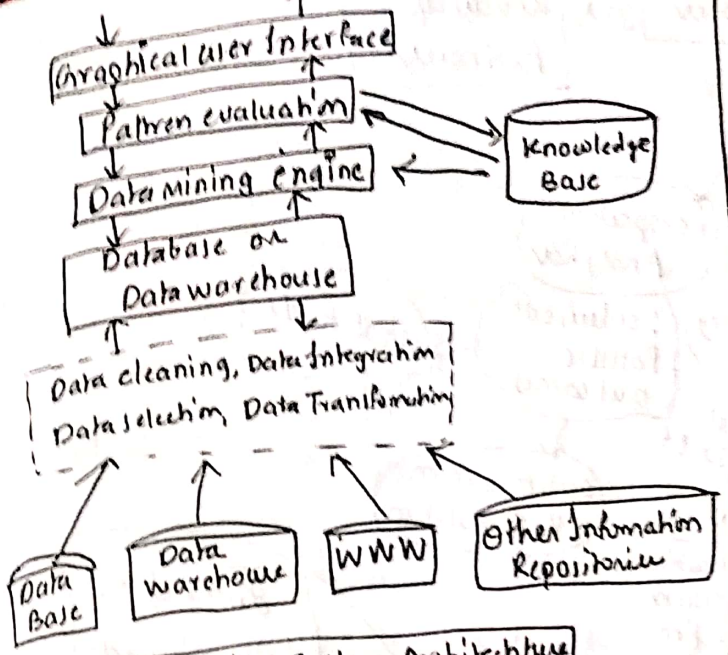
5. Data Mining :- an essential process where intelligent methods are applied to extract data patterns or desired results.

"data mining step may interact with a user or a database".

6. Pattern evaluation :- To identify truly interesting patterns and representing knowledge based on interesting measures.

7. Knowledge representation :- where visualization & knowledge representation techniques are used to present mined information to users





Data Mining System Architecture

- Knowledge base is domain knowledge that is used to guide the search or evaluate the interestingness of a resulting pattern
- Data mining engine consists of functional modules
  - → characterization
  - → Association & Co-relation analysis
  - → Prediction etc.

Kinds of data that can be Mined

1. Relational Database
   - Tables
   - attributes
   - Domains
   - Tuples
   - [ER] data model
2. Data warehouse
   - unified schema
   - OLAP
   - historical data
   - Multidimensional data mining
3. Transactional Database
   - Transactions
   - unique Transaction id
   - eg: flight booking [market analysis]
4. Object Relational Database (ORD)
   - (RDBMS)
   - objects, classes are supported in database
5. Temporal Databases
   - Provides a uniform & systematic way of dealing historical & future data
6. Spatial Databases
   - spatial related information
   - maps
   - VLSI
   - medical
   - CAD images
7. Text databases
   - word description for objects
   - large paragraphs
8. Heterogenous Database
   - consist of different database with different data models
   - called Legacy database
9. Data streams
   - video feed
   - Security cams
   - sensor data
10. WWW

Spatiotemporal Database
- stores spatial objects that change with time
- eg: google maps

Two types of Temporal Database

Sequence Database
- stores data as ordered events, with or without notion of time

Time-series Database
- stores sequences of values/events obtained over a time
- eg:- Temperatures collected with a particular time.

(5) DODH TTRSW

Data mining Functionalities (or) Kinds of Patterns

What data is telling!

are three types

- **Descriptive Analytics**
  - : Insight into Past
  - → **Business Intelligence**
    - cluster Analysis
    - Outlier Analysis
    - evolution Analysis
    - Mining frequent Patterns
    - Associations
    - co-relations
    - characterization & discrimination
- **Predictive Analytics**
  - : understanding the future
  - what might?
  - → **Forecasting**
    - classification
    - Regression
    - prediction
    - Time series Analysis
- **Perspective Analytics**
  - : advise on Possible outcomes
  - what to do?
  - → **Optimization & simulation**

Some of the data mining functionalities

1. characterization & description
2. Mining of frequent Patterns
3. Associations & co-relations
4. classification & regression
5. clustering analysis
6. Outlier Analysis and Evolution Analysis

Characterization & discrimination

↳ descriptions of a class or a concept is called class/concept descriptions

These descriptions can be derived from

(i) **Data characterization**: is a summarization of general characteristics/features of a target class of data
Output forms: pie-chart, barcharts

(ii) **Data discrimination**:- is a comparision of general features of the target class data objects with the general features of objects from one or a set of contrasting classes.
output forms: Same as above
Discrimination descriptions expressed in rule form are referred to as discriminant rules

what technologies are used →

Mining Frequent Patterns → Substructure

Frequent Patterns → Patterns (itemset) that occur frequently in the data.

Frequent Itemset → set of items that frequently appear together in the Transactional data

(in order (or) in any order) eg:- Milk-bread bat & ball etc.

Frequent Subsequence → A sequence of patterns that appear frequently

- A Frequent pattern mining searches for recurring relationships in a given dataset
- A frequent mining of patterns leads to a discovery of Associations & Correlations items between the items in the dataset → There are many Algorithms

Association rule is a data mining Algorithm discovered by IBM company

Association rule identifies the association between two items based on their occurrence

eg:-
- In a given set of transactions, find the rules that will predict the transaction on
- predict the occurrence of item based on the occurrence of another items.

→ Mostly used in market-Basket Transactions.

Market - Basket Analysis

eg:-

TID	itemset
1	Bread, milk
2	Bread, diaper, beer, coke
3	Milk, diaper, beer, coke.
4	Rc car, battery.

Itemset → A collection of one or more items eg:- milk, bread & diaper.

K-itemset → An itemset which contains K items.

Frequent itemset → An itemset whose support >= threshold >= minsup

- Any itemset whose support value >= minsupport >= minsup which is called as frequent pattern.

Association rules
{milk} → {bread}
{milk, bread} → {egg, coke}

Support count → A support count is represented using σ & frequency of an occurrence of an itemset.
eg:- σ(milk, bread, diaper) = 2

Support & Support count are synonymous

A fraction of transactions that contains an itemset is called as support
eg:- {milk, bread, diaper} = 2/5

22-12-18

- Types of databases
- Data mining functionality
  - descriptive
  - Predictive

③ Transactional
eg:- Booking, Buying

ID	Items
1	TV, mobile
1	i

④ Other kinds of data
→ Time-related/sequence data
→ data stream (video surveillence)
→ Spatial data (maps)
→ hypertext data etc

Classification and prediction

- A classification is used to find a model or a function which describes / distinguish between class & concept.
- Different approaches of classification

① If-then-else :-
> Based on the features
If (wing color == red)
classify as (type-a);

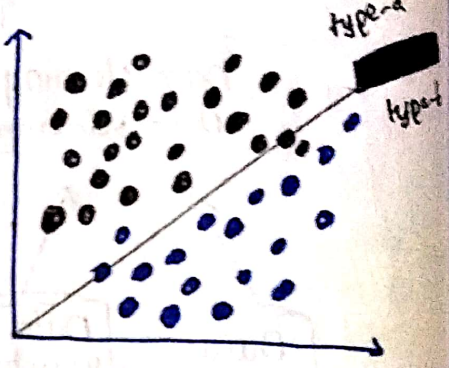
② Decision tree :-
> using a decision tree and an algorithm to classify.

③ Mathematical formula :-
> using mathematical formula & relations we can classify.

eg:-
Age(x, youth) && income(x, 10000) → (x, class A)
Age(x, mid-age) && income(x, 10000) → (x, class B)

④ Nueral Networks.

eg:- classification of types of butterfly



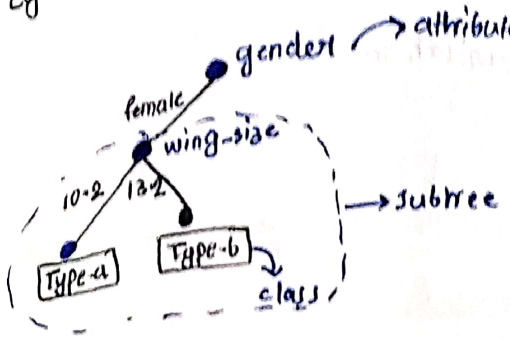
→ datapoints of type-a butterfly
→ datapoints of type-b butterfly

- Based on the training data, we extract the features
- These training data contains labelled examples of which type of butterfly
- This is given to the model To Train the model.

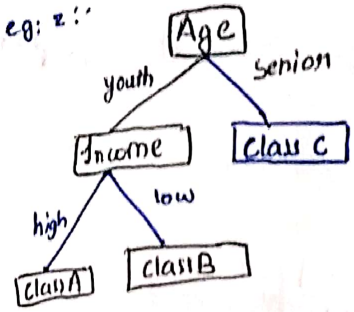
" The above problem is having a linearly seperable Solution, so it is a linearly seperable problem "

⇒ Decision Tree :- A decision tree may look like a flow chart or a tree diagram with each node containing an attribute value which is a test attribute value. And every sub-tree represents outcome of test value. And leaf nodes represent the classes (labels).

eg:- decision tree for type of butterflies



eg: 2:-



⇒ Prediction:-

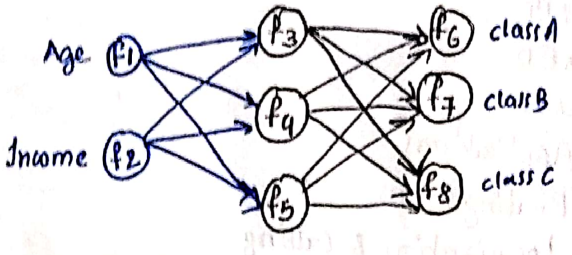
- whereas the prediction takes the continuous data and predict the future values.
- prediction is a continuous valued function.
- This can be used to predict unavailable, missing numerical values.

⇒ clustering analysis :-

- Data objects which are not known class labels or unknown class labels or without knowing class labels.
- class labels are not present in the Training data, for that classes generate the clusters.
- objects are clustered or grouped based on the similarities.
- Maximizing intra class similarity and minimizing the inter-class similarity.

⇒ Neural Networks :-

- It contains a collection of Neurons
- And they contains weights and connection between units

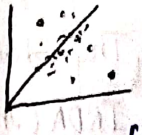


TT classification is performed only on

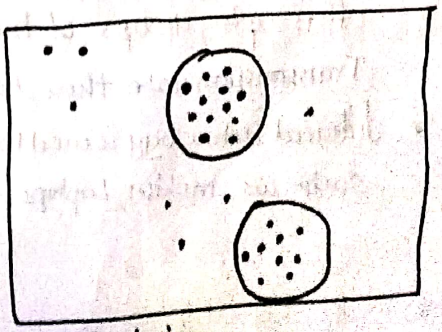
- Numerical datasets
- categorical data (eg: male or female)
- ordered events (eg:- historical events)
- Discrete data (NON-sequences)

⇒ Regression analysis :-

- A Regression Analysis is a Analysis in which we find the error rate using statistical methods.
- Regression analysis contains statistical methods which is used for Numerical Prediction
- Identification of trends on the available data
- Relevance analysis is done before classification & prediction to identify the attributes.



Find MSE in the regression analysis



Are all Patterns Interesting ?

1. It is easily understood by humans
2. valid on new or test data with some degree of certainity
3. Potentially useful
4. novel.
5. If it validates a hypothesis

& There are subjective vs objective interesting measures:

Classification of data Mining systems

- according to the type of data that is mined (spatial, text, www)
- according to data model drawn on (Relational DB, DODB, data warehouse, Transactional DB)
- according to the knowledge discovered (on based on functionalities used) (characterization, discrimination, clustering -etc)
- according to mining techniques used (machine learning, NN's, genetic, statistic, -query driven systems)

Five task primitives of specifying a data mining task

1. Task relevant data {relevant attributes, dimensions}
2. Knowledge to be mined { specifying which data mining functionality to be used }
3. Background knowledge { Domain Knowledge }
4. Pattern interestingness measure { interesting new measures for certainity, simplicity, utility & novelity }
5. Visualization of discovered patterns { rules, tables, pie, barcharts, decision trees }

Integration of Data Mining system with a database or a Data warehouse system

— difference between OLAP & OLTP

- **No coupling**:
  - → data mining system sources such as flat files to obtain data for mining.
  - → But no data ~~mining~~ base functions are implemented in Process
  - → This is a poor design choice.

- **Loose coupling**:
  - → The data mining system is not integrated with database / data warehouse beyond their usage as a source of data
  - → But still take advantage of DB/DW's indexing & other features
  - → It is hard for loose coupling to achieve high scalability & good performance with large datasets

- **Semi tight coupling**:
  - Some of the primitive operations such as aggregation, sorting & pre computation of statistical functions can be done within the database & during query. And these can be stored inside DB/DW's itself to promote high performance of data mining system

- **Tight coupling**:
  - → Complete Integration of DB/DW into Data mining System
  - → high scalability & performance

— neglecting Technical & Implementation details. It is the best Architecture

Major Issues in data mining

Mining methodology & user-interaction issues

- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query language & ad-hoc data mining
- Presentation & visualization of mining results
- Handling outlier or incomplete data
- Pattern evaluation

Performance issues

- efficiency & scalability of data mining algorithm
- Parallel, distributed & incremental algorithm

Issues related to diversity of database types

- Handling relational & complex data types
- Mining information from Heterogeneous database

Types of Data sets

- **Record**
  - relational
  - matrix
  - Transactional data
- **graph & Network**
  - WWW
  - social network
  - molecular structure
- **Ordered**
  - video data
  - Temporal data
  - sequential data
- **Spatial image & multimedia**

Important Characteristics of structured data

- dimensionality
- Sparsity
- Resolution
- distribution
  - centrality & dispersion

Data objects

represents an entity

data objects are described by attributes

Database rows ↓ data objects	columns ↓ attributes

Attributes / dimensions / features / variables

- **Nominal**
  - Categories
  - name of things
- **Binary**
  - nominal with only two states
  - **Symmetric Binary**
    - both outcomes are important
    - eg: gender
  - **Asymmetric Binary**
    - outcomes are not equally important (HIV +ve, HIV -ve)
- **Numeric**
  - **Interval Scaled**
    - equal spaced units
    - no true zero point
    - eg: temp in C° or F°
  - **Ratio scaled**
    - inherent zero point
    - eg K° in C° (Kelvin)
- **Ordinal**
  - values with a meaningful order & no magnitude
  - eg: (small, large, ...)

Discrete vs continuous Attributes

- **discrete**
  - finite or countable set of values
  - eg: zip codes, Binary attributes
- **continuous attributes**
  - has real finite numbers {represented using floating point variables}

Formulas

$$\text{Mean} = \frac{(x_1 + x_2 + \dots + x_n)}{n}$$

$$\text{midrange} = \frac{\text{Max} + \text{Min}}{2}$$

Median $= \frac{1}{2}(n+1)$ when n = number of data values

(If n is odd)

Other wise

If n = even

$$\text{median} = \frac{1}{2}\left[n/2 + n/2 + 1\right]$$

Mode : most repeated / most Frequent value

- If two values Occur frequently Bimodal
- If three " " Tvimodal
- If more " " Multimodal

"The mean, median mode are collectively called as central tendency"

Measures of dispersion

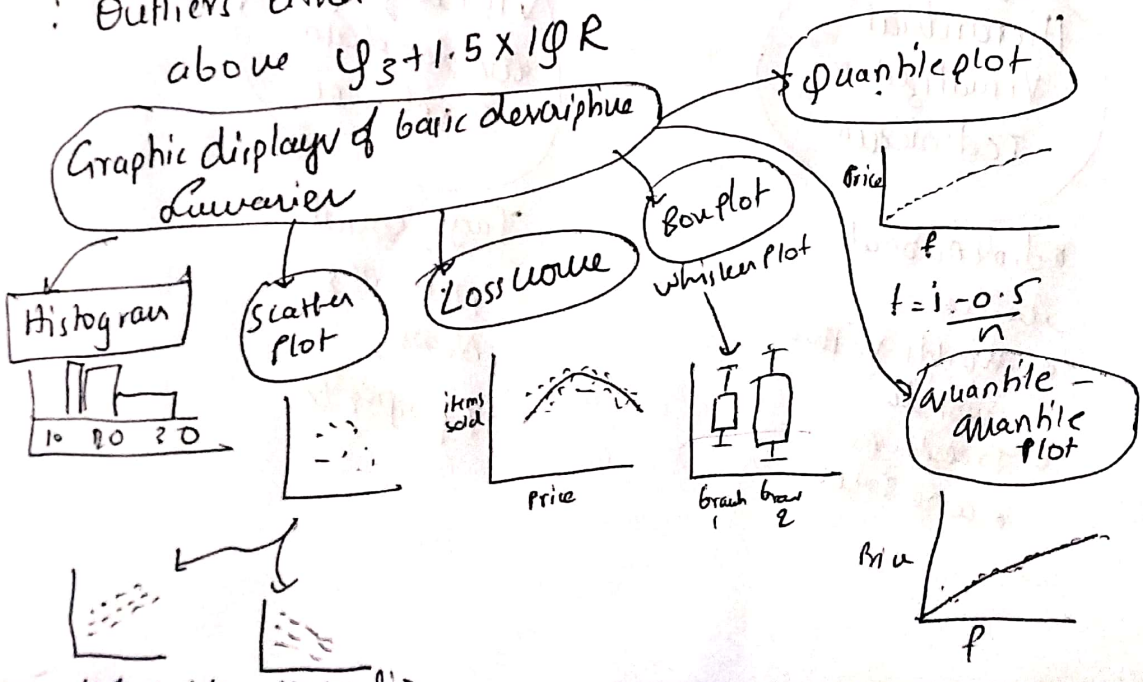
← Notes

: A Box plot used to represent range, median, quartiles & Inner quartile range (IQR) $[q_3 - q_1]$

: Five number summary is

(min), Q_1, median, Q_2, Q_3, max

: Outliers either fall below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$

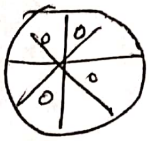


Data visualization Technique

Pixel oriented

- data of m dimensions create m windows on screen, one for each dimension
- values in dimensions are mapped to pixels
- colors represent the corresponding value

→ laying out pixels in circles



Geometric Projection visualization Technique

- direct visualizations
- scatter plot
- Landscapes
- hyperslice
- Parallel coordinates
- Projection Pursuit Technique

Icon Based Visualization Technique

- Chernoff faces
- stick figures
- Tile bars
- shape coding
- color icons

Hierarchical Visualization Technique

- dimensional scaling
- worlds within worlds
- Tree Maps
- cone Trees

visualizing complex data & relations

Tags, social networks, News, Typography

Measures of Similarity & dissimilarity

Measures of data quality

- Accuracy
- completeness
- consistency
- Timeliness
- Believability
- Interpretability

Major Tasks in data Preprocessing

- Data cleaning
  - fill in missing values,
  - smooth noisy data
  - identify or remove outliers
  - resolve inconsistencies
- Data Integration
  - Integration of multiple databases, data cubes or files
- Data Reduction
  - dimensionality reduction
  - numerosity reduction
  - data compression
- Data Transformation & data discretization
  - normalization
  - concept hierarchy generation