Lecture 7: Introduction to Data Mining

Introduction

UNCOVER HIDDEN INFORMATION
DATA MINING

- Data is growing at a phenomenal rate.
- Users expect more sophisticated information.
- How?

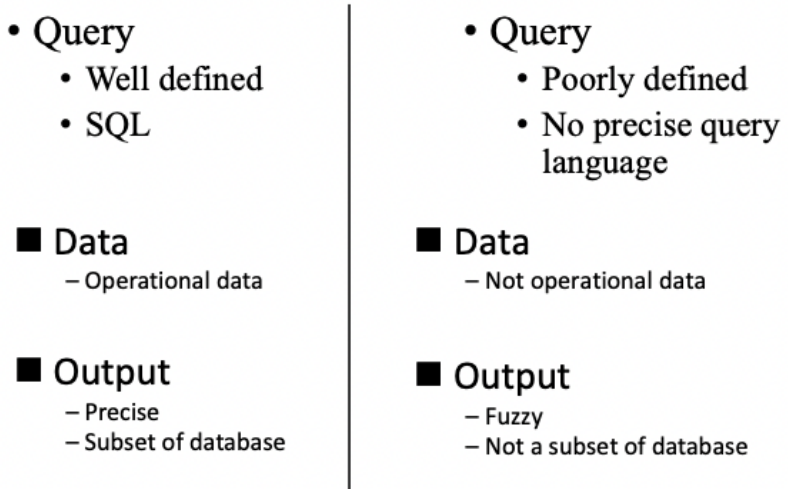
Data Mining – Definition

- Finding hidden information in a database.
- Fit data to a model.
- Similar terms:
    - Exploratory data analysis.
    - Data driven discovery.
    - Deductive learning.

Data Mining Algorithm

- Objective: Fit Data to a Model:
    - Descriptive.
    - Predictive.
- Preference – Technique to choose the best model.
- Search – Technique to search the data.
    - "Query".

Database Processing vs. Data Mining Processing



- Query
    - Well defined
    - SQL

■ Data
- Operational data

■ Output
- Precise
- Subset of database

- Query
    - Poorly defined
    - No precise query language

■ Data
- Not operational data

■ Output
- Fuzzy
- Not a subset of database

Query Examples

- Database:
    - Find all credit applicants with last name of Smith.
    - Identify customers who have purchased more than $10.000 in the last month.
- Data Mining:
    - Find all credit applicants who are poor credit risks (classification).
    - Identify customers with similar buying habits (clustering).
    - Find all items which are frequently purchased with milk (association rules).

Basic Data Mining Tasks

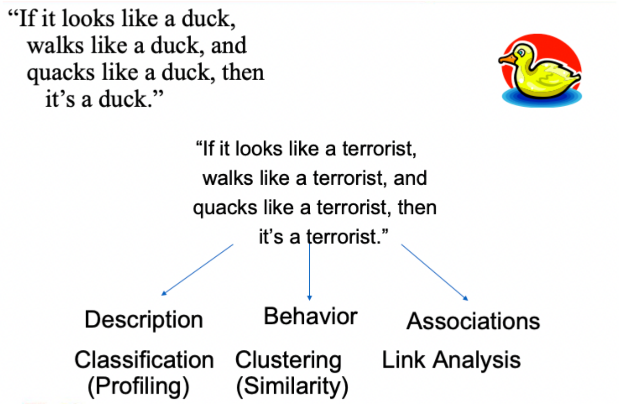
- Classification maps data into predefined groups or classes.
  - Supervised learning.
  - Prediction.
  - Regression.
- Clustering groups similar data together into clusters.
  - Unsupervised learning.
  - Segmentation.
  - Partitioning.
- Link Analysis uncovers relationships among data.
  - Affinity Analysis.
  - Association Rules.
  - Sequential Analysis determines sequential patterns.

Classification

- Assign data into predefined groups or classes.

But It Isn't Magic

- You must know what you are looking for.
- You must know how to look for you.
- Suppose you know that a specific cave had gold:
  - What would you look for?
  - How would you look for it?
  - Might need an expert miner.



"If it looks like a duck,
walks like a duck, and
quacks like a duck, then
it's a duck."

"If it looks like a terrorist,
walks like a terrorist, and
quacks like a terrorist, then
it's a terrorist."

Description

Classification
(Profiling)

Behavior

Clustering
(Similarity)

Associations

Link Analysis

Classification Example

- Grading.
- Given a collection of annotated data (in this case 5 instances of Katydids and five of Grasshoppers), decide what type of insect the unlabeled example is.
- The classification problem can now be expressed as:
  - Given a training database, predict the class label of previously unseen instance.
- Facial Recognition.
- Handwriting Recognition.
- Anomaly Detection.

Clustering

- Partition data into previously undefined groups.

Two Types of Clustering

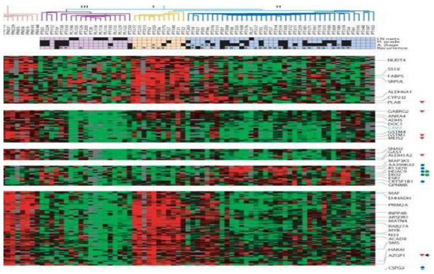
- Hierarchical.
- Partitional.

Hierarchical Clustering Example

- Iris Data Set.

Microarray Data Analysis

- Each probe location associated with gene.
- Color indicates degree of gene expression.
- Compare different samples (normal/disease).
- Track same sample over time.
- Questions:
  - Which genes are related to this disease?
  - Which genes behave in a similar manner?
  - What is the function of a gene?
- Clustering:
  - Hierarchical.
  - K-means.
- Gene Expression Profiling identifiers clinically relevant subtypes of prostate cancer.



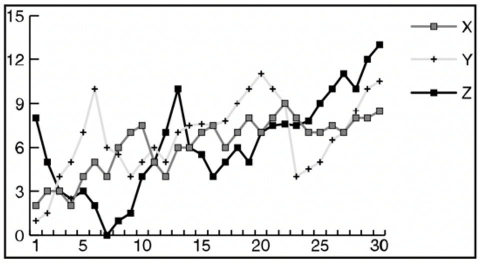
Association Rules/Link Analysis

- Find relationships between data.

Association Rules Examples

- People who buy diapers also buy beer.
- If gene A is highly expressed in this disease, then gene A is also expressed.
- Relationships between people.
- Book Stores.
- Department Stores.
- Advertising.
- Product Placement.

Example: Stock Market Analysis

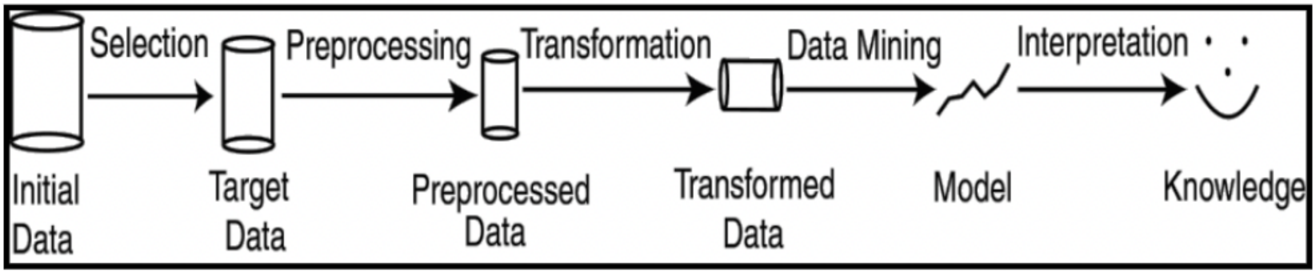


- Example: Stock Market.
- Predict future values.
- Determine similar patterns over time.
- Classify behavior.

Data Mining vs. KDD

- Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data.
- Data Mining: Use of algorithms to extract the information and patterns derived by the KDD process.

KDD Process



- Selection: Obtain data from various sources.
- Preprocessing: Cleanse data.
- Transformation: Convert to common format. Transform to new format.
- Data Mining: Obtain desired results.
- Interpretation/Evaluation: Present results to user in meaningful manner.

KDD Process Example: Web Log

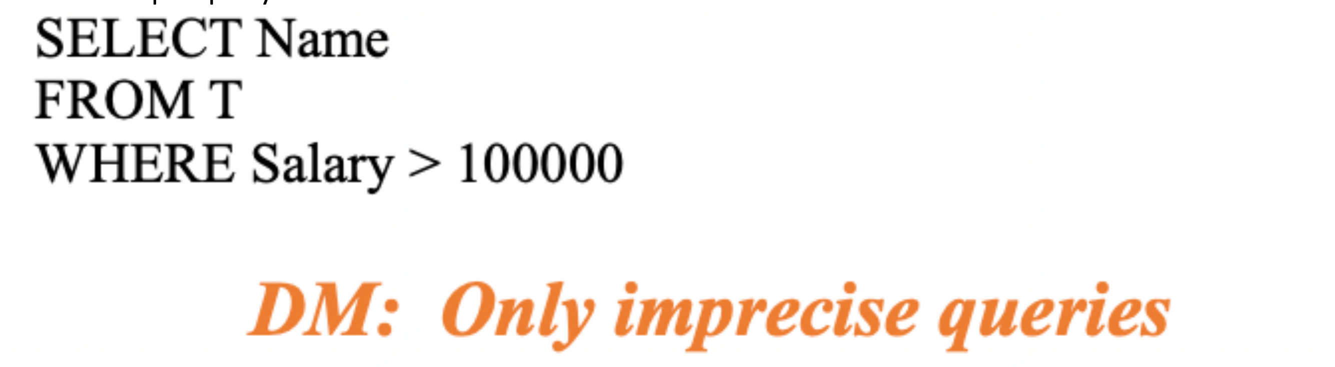
- Selection:
  - Select log data (dates and locations) to use.
- Preprocessing:
  - Remove identifying URLs. Remove error logs.
- Transformation:
  - Sectionize (Sort and Group).
- Data Mining:
  - Identify and count patterns. Construct data structure.
- Interpretation/Evaluation:
  - Identify and display frequently accessed sequences.
- Potential User Applications:
  - Cache prediction.
  - Personalization.

Related Topics

- Databases.
- OLTP.
- OLAP.
- Information Retrieval.

DB & OLTP Systems

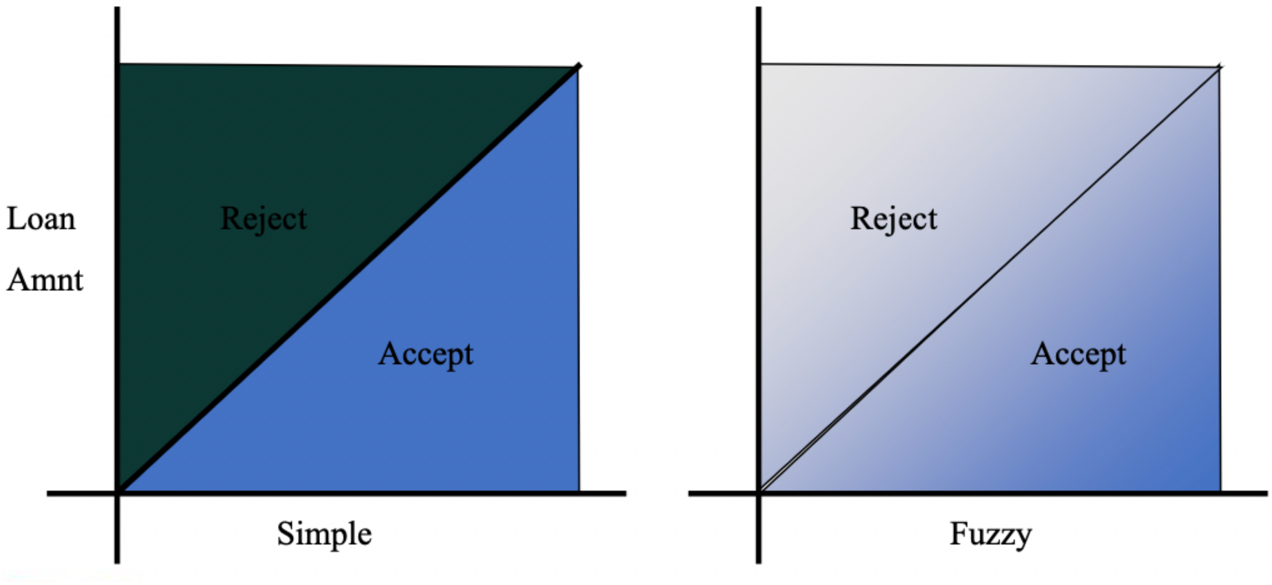
- Schema:
  - (ID, Name, Address, Salary, JobNo).
- Data Model:
  - Entity Relationship.
  - Relational.
- Transaction.
- Sample query:



```
SELECT Name
FROM T
WHERE Salary > 100000
```

DM: Only imprecise queries

Classification/Prediction is Fuzzy



Information Retrieval

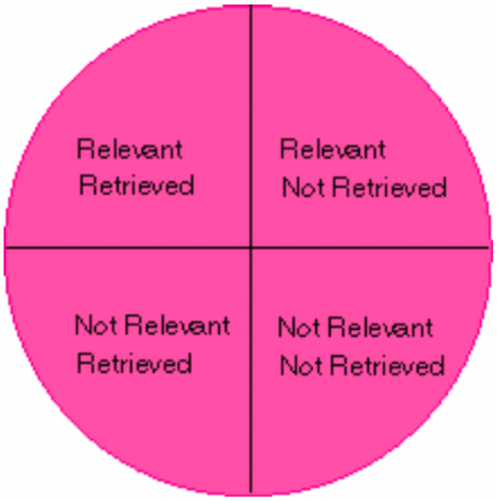
- Information Retrieval (IR): retrieving desired information from textual data.
- Library Science.
- Digital Libraries.
- Web Search Engines.
- Traditionally keyword based.
- Sample query:
    - Find all documents about “data mining”.

DM: Similarity measures;
Mine text/Web data.

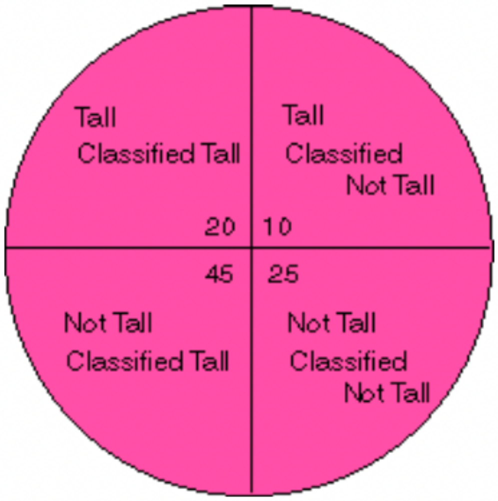
- Similarity: measure of how close a query is to a document.
- Documents which are “close enough” are retrieved.
- Metrics:

- $$\textit{Precision} = \frac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$$
- $$\textit{Recall} = \frac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$$

IR Query Result Measures and Classification



IR



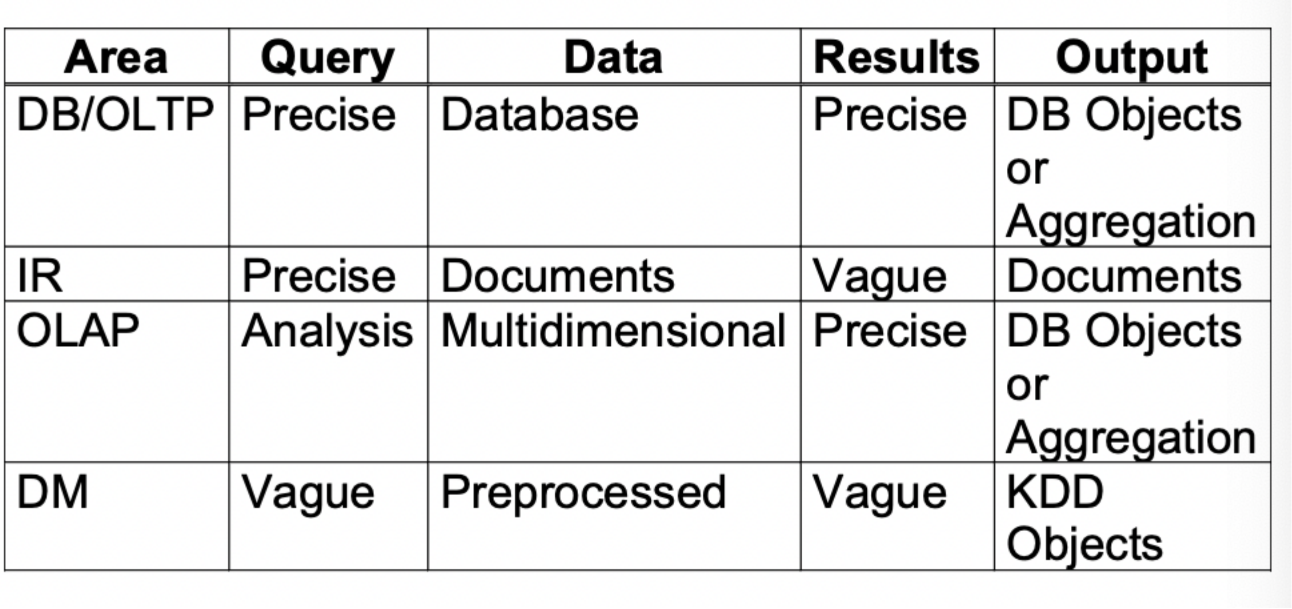
Classification

OLAP

- Online Analytic Processing (OLAP): provides more complex queries than OLTP.
- Online Transaction Processing (OLTP): traditional database/transaction processing.
- Dimensional data; cube view.
- Visualization of operations:
  - Slice: examine sub-cube.
  - Dice: rotate cube to look at another dimension.
  - Roll Up/Drill Down.

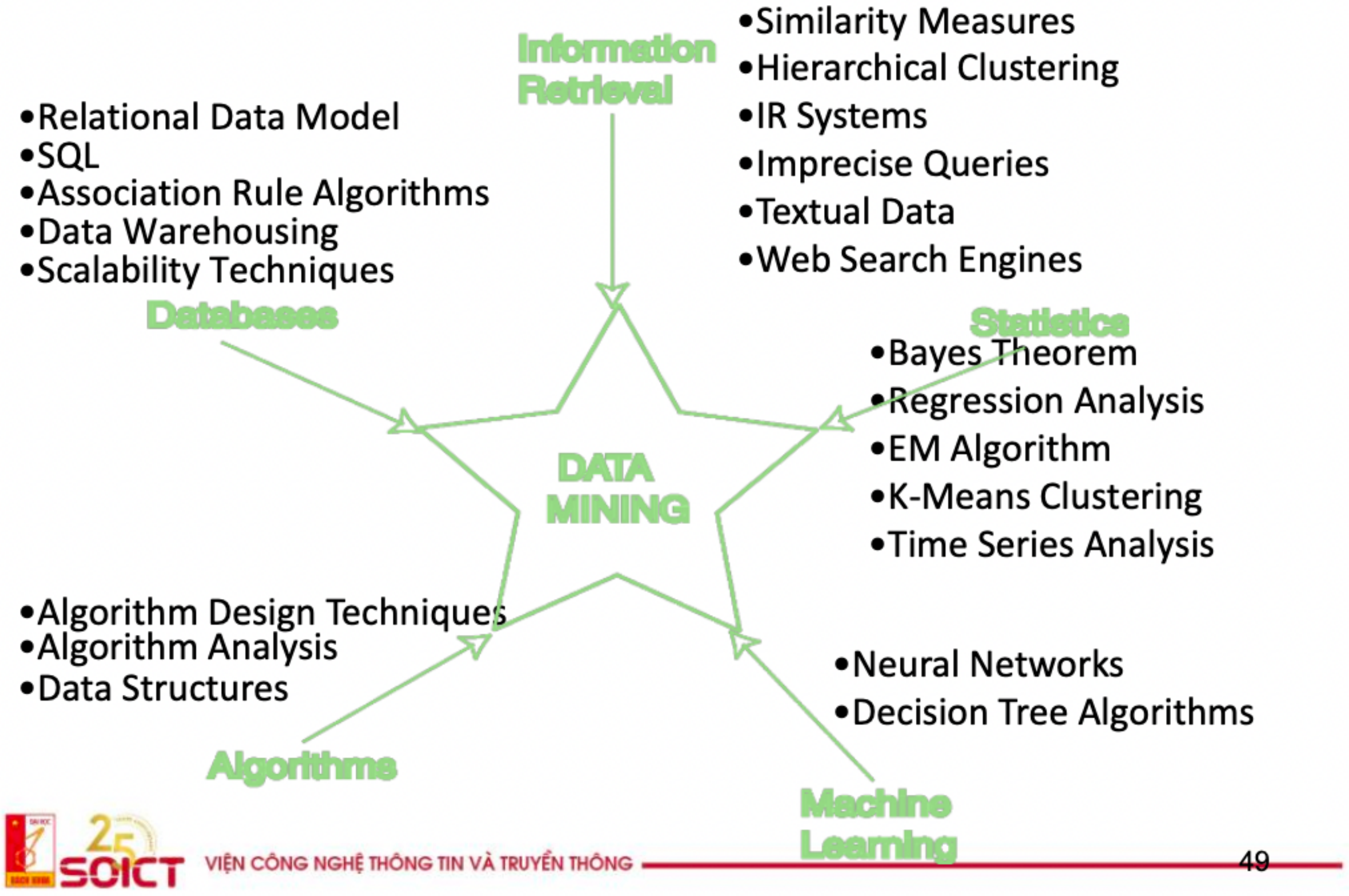
DM: May use OLAP queries.

DM vs. Related Topics



Area	Query	Data	Results	Output
DB/OLTP	Precise	Database	Precise	DB Objects or Aggregation
IR	Precise	Documents	Vague	Documents
OLAP	Analysis	Multidimensional	Precise	DB Objects or Aggregation
DM	Vague	Preprocessed	Vague	KDD Objects

Data Mining Development



- Relational Data Model
- SQL
- Association Rule Algorithms
- Data Warehousing
- Scalability Techniques

- Similarity Measures
- Hierarchical Clustering
- IR Systems
- Imprecise Queries
- Textual Data
- Web Search Engines

- Bayes Theorem
- Regression Analysis
- EM Algorithm
- K-Means Clustering
- Time Series Analysis

- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Neural Networks
- Decision Tree Algorithms

KDD Issues

- Human Interaction.
- Overfitting.
- Outliers.
- Interpretation.
- Visualization.
- Large Datasets.
- High Dimensionality.
- Multimedia Data.
- Irrelevant Data.
- Noisy Data.
- Changing Data.
- Integration.
- Application.

Warning

- With data mining, you don’t always know what you are looking for.
- There is not one right answer.
- The data you are using is noisy.
- Data Mining is a very applied discipline.
- A data mining course provides you tools to use to analyze data.
- Experience provides you knowledge of how to use these tools.

Social Implications of DM

- Privacy.
- Unauthorized use.
- Profiling.
- Invalid results and claims.

Data Mining Metrics

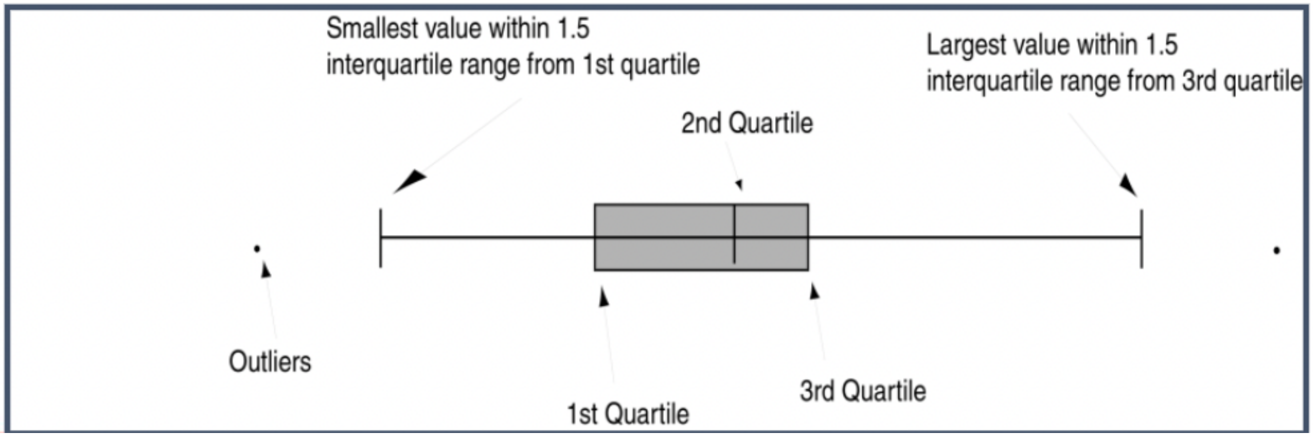
- Usefulness.
- Return on Investment (ROI).
- Accuracy.
- …
- Space/Time.

Visualization Techniques

- Graphical.
- Geometric.
- Icon-based.
- Pixel-based.
- Hierarchical.
- Hybrid.

Models Based on Summarization

- Visualization: Frequency distribution, mean, variance, median, mode, etc.
- Box Plot:


Smallest value within 1.5 interquartile range from 1st quartile
Largest value within 1.5 interquartile range from 3rd quartile
2nd Quartile
Outliers
1st Quartile
3rd Quartile

DM Tools

- XLMiner – Easy addin to Excel: http://www.solver.com/xlminer/index.html
- Webka – Open Source; Visualization, Functionality, Interface: http://www.cs.waikato.ac.nz/ml/weka/
- SAS (JMP) – Commercial Product.
- SPSS – Commercial Product.
- MATLAB – Statistical/Math Applications.
- R – Programming.

Table of Contents