Unit 1 – Data Mining and Business Perspective

Introduction

- **Moore's law** – Processing speed – doubles every 18 months
- Storage capacity – doubles every 9 months
- Computer databases expand and fill available storage space
- It can be valuable resource – we can extract valuable knowledge from the data resource
- A process of converting a large amount of **data into knowledge** – **KDD** (knowledge discovery in databases)
- Data mining not a well-ordered discipline
- Major opportunities for improvement in data mining technology – scalability and compatibility with database systems, usability and accuracy of the techniques.

Why Mine Data?

- Lots of data is collected and warehoused
  - Web data, e-commerce, purchases at departmental stores, bank transactions, sensors, social media…
- Computers have become cheaper and powerful
- Competitive pressure is strong
  - Provide better customized services for an edge
- We are drowning in data, but starving for knowledge
- Solution: Data Mining
  - Extraction of interesting knowledge(rules, regularities, patterns, constraints) from the data in large datasets.

Data Mining and KDD

- KDD – A non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data and data mining as the extraction of patterns from the observed data.
- Data mining is the subset of KDD process
- Data mining techniques provide the algorithms to fuel the KDD process
- Alternative names for Data Mining
  - Knowledge mining, knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting

What is not data mining?

- Query processing – data retrieval from databases
- Expert system – uses AI to simulate judgement and behaviour
- Statistical program – mean, median, correlation….

Data Mining: On what kind of data?

- Relational databases
- Data warehouses
- Transactional databases
- Information repositories
  - Spatial databases
  - Time series and temporal data
  - Text and multimedia databases
  - Heterogeneous and legacy databases
  - www

Can we find all interesting patterns?

- Find all the interesting patterns: Completeness
- Search only for interesting patterns: optimization.

Steps in KDD process

1. Identify and develop an understanding of the application domain
2. Select the dataset to be studied
3. Select the complimentary datasets. Integrate the data sets.
4. Code the data. Clean the data of duplicates and errors, Transform the data.
5. Develop model and build hypotheses
6. Select appropriate data mining algorithms
7. Interpret results. View results using appropriate visualization tools

8. Test results in terms of simple proportions and complex predictions.
9. Manage the discovered knowledge.

Data mining vs Data analysis

- Data mining – is a dynamic process that enables more intelligent use of database/data warehouse than data analysis
- Builds models to make predictions without additional SQL Queries
- DM applies for both smaller and larger datasets
- Many attributes – width
- Many entities – volume
- Many records – depth
- Data Analysis – inspect the data and suggest conclusions

Data Mining Vs Statistics

- Data mining is data driven
  - Patterns and hypothesis are automatically extracted from data
- Statistics is human driven
  - Formal statistical inference is assumption driven
  - i.e, we formulate a hypothesis and validate against the data

DM – Success stories

- Bell Atlantic – When a customer reports a telephone problem, a decision is made about what type of technician to dispatch to resolve a problem
- American Express – loan application to be accepted, rejected or required a human expert to judge
- British petroleum corporation – machine learning algorithm to control the crude oil mixing with natural gas
- Flight simulation and learning –program logs the actions of a human driven aircraft. Later the same actions is formed as rules and the flight run is tested in auto pilot mode.

- Making robot learn - target pursuit and obstacle avoidance
- Computer controlled vehicles on road
- Learning to win game – using reinforcement learning technique
- Cattle farming – which cows to retain and which to sell
- Molecular biology – sequencing of genomes
- Drug discovery – determine structure activity relationships in drugs
- Pharma – to ascertain the class of proteins
- Astronomy- identify and classify the objects in galaxy
- Medicine – direct and indirect dependencies for tuberculosis
- Geophysics – new history of natural disasters
- Fraud Detection – patterns of terrorist activity
- Intrusion detection – anomalous logs in computer usage

Data Mining Research

- Main reasons for growth of data mining research
  - Developing algorithms to mine large, massive and high dimensional data sets
  - Developing algorithms to mine new types of data
  - Developing algorithms, protocols and other infrastructure to mine distributed data
- In order to respond these challenges, multidisciplinary and interdisciplinary research is required.

Research Achievements

- Neural Networks – a System developed with the inspiration of brain
- They are designed specifically for pattern recognition.
- An artificial neuron is a device with many inputs and one output.
- The neuron has two modes of operation;

  the training mode and

  the using mode.

- Consists of i/p, output and hidden nodes
- Initially the nodes are connected with random weights
- During the training, a gradient descent algorithm is used to adjust the weights, so that output nodes classify data presented to the input nodes.
- Once trained, it provides projections given new situations of interests and answers what if questions
- The trained network is viewed as a black box, since it does not give explanation of the results.

- In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.

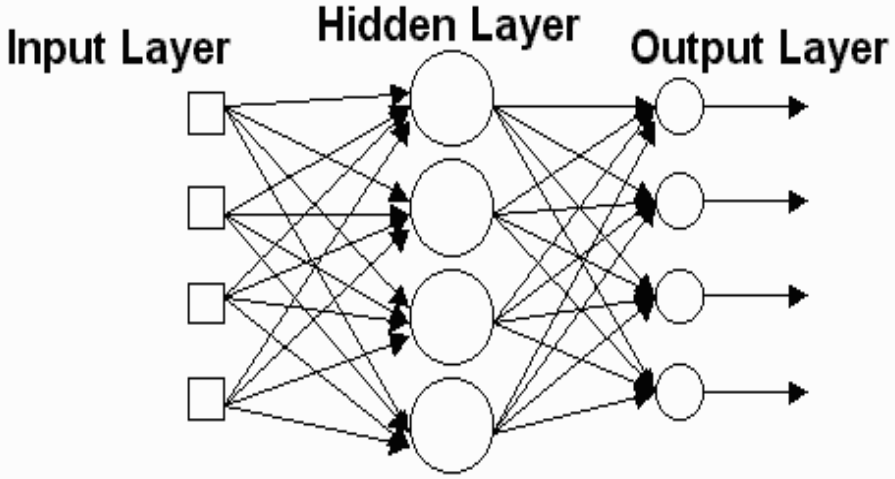
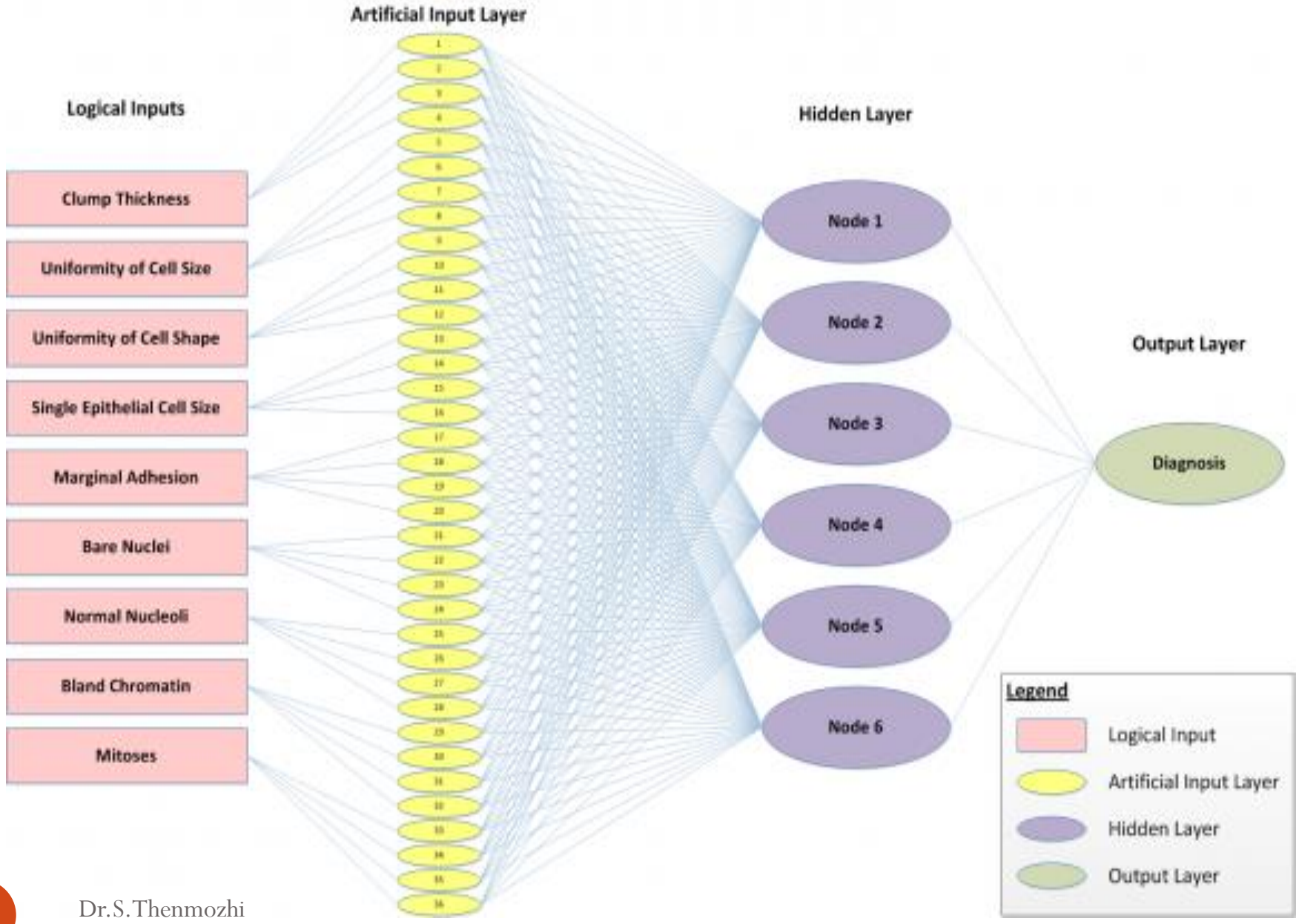
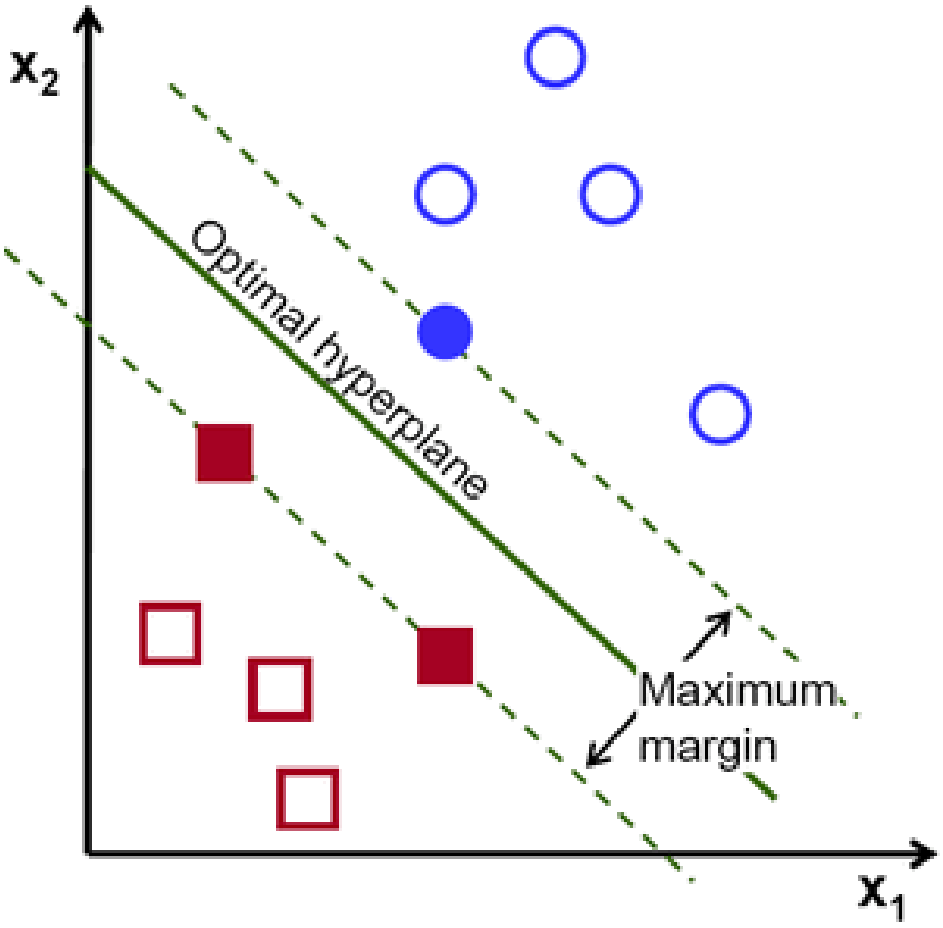


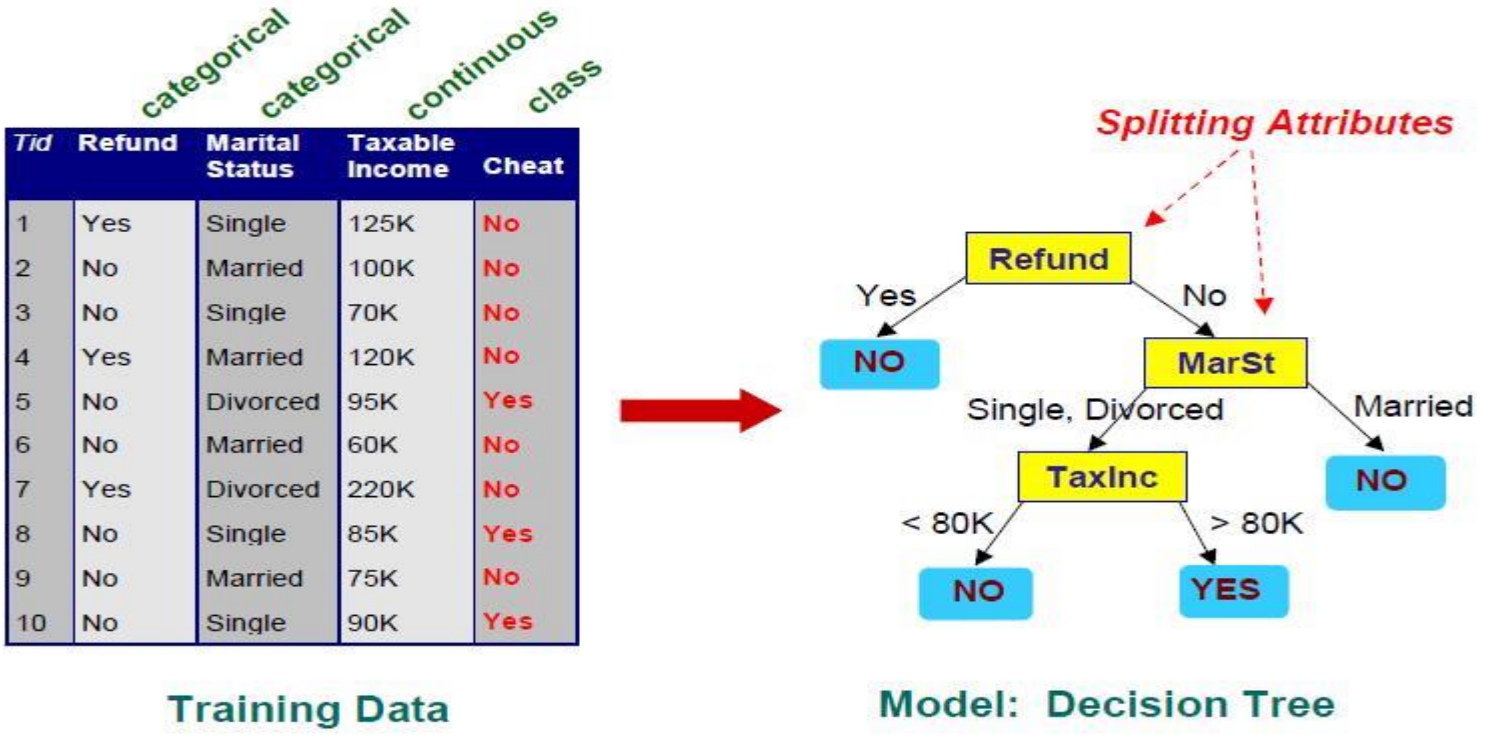
Figure 2 The anatomy of a neural network.


Artificial Input Layer
Logical Inputs
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Single Epithelial Cell Size
Marginal Adhesion
Bare Nuclei
Normal Nucleoli
Bland Chromatin
Mitoses
Hidden Layer
Node 1
Node 2
Node 3
Node 4
Node 5
Node 6
Output Layer
Diagnosis
Legend
Logical Input
Artificial Input Layer
Hidden Layer
Output Layer

- Support Vector Machines – a new type of algorithm that compete with NN in application like classification, regression and clustering
- It is a discriminative classifier formally defined by separating hyperplane
- The algorithm produces the optimal hyperplane which categorises new examples


x_2
Optimal hyperplane
Maximum
margin
x_1

- Tree based classifiers – A tree is a convenient way to break large datasets into smaller datasets
- A learning set is defined at the root. Ask questions at each interior node, the data at the leaves can often be analysed



Graphical models required..

- Ensemble learning – build collection of models and select efficient by voting strategy, possibly the overall error of the model.
- Large scale optimization – problems too large to be solved exactly
- High performance computing and Communication – what is possible for wide area high performance networks moving large amount of data between geographically separated sites
- Databases, Data warehouses and digital library – integrating the data in a much easier way
- Visualization of massive data sets – visualization tools to be developed

Trends that affect Data mining

- Data trends
  - Data grows six to ten orders of magnitude
  - To be accessible via networks
- Hardware trends
  - High performance computing may be required to process data
- Network trends
  - 100 times faster than the current connectivity in networks
- Scientific computing trends
  - Simulation on large data sets
- Business trends
  - Using fewer people at lower cost

Research Challenges

- Scaling data mining algorithms
  - No of records increases, attributes increases, predictive models increases, interactivity and real-time responses increases
- Extending data mining algorithms to new data types
  - Time series, unstructured, multimedia, hierarchical and multi-scale data
- Developing distributed data mining algorithms
  - Data mining algorithms should work making the data to work in place,
  - Develop metadata and has to be used
- Ease of use
  - Collaborative methods for the ease of use
- Privacy and security
  - Develop privacy and security models and protocols to prevent misuse of data

Test beds and infrastructure

- Data mining test beds are disk oriented and processor-oriented
- Necessary infrastructure should be available

Data mining from Business Perspective

- KDD is the process whereas Data mining is technique to discover knowledge
- Now, in today's context both can be used interchangeably
- Evolution of Business Data Mining
  - 1960's – Data Collection
  - 1980's – Data access
  - 1990's- Data warehousing and DSS
  - 1990's – Data Mining
- ERP systems helps in a greater context as it makes all the data from various departments and braches available at one place

Data Mining for Business

- *Market segmentation* - Identify the common characteristics of customers who buy the same products from your company.
- *Customer churn* - Predict which customers are likely to leave your company and go to a competitor.
- *Fraud detection* - Identify which transactions are most likely to be fraudulent.
- *Direct marketing* - Identify which prospects should be included in a mailing list to obtain the highest response rate.
- *Interactive marketing* - Predict what each individual accessing a Web site is most likely interested in seeing.
- *Market basket analysis* - Understand what products or services are commonly purchased together; e.g., beer and diapers.
- *Trend analysis* - Reveal the difference between a typical customer this month and last.

Evolution of Data Mining Systems

- 1st Generation –all are research driven tools such as NN , a decision tree classifier
- Using more than one tools was complicated which was not easy to achieve for expert users
- 2nd Generation – Data mining developers developed data mining systems called suites in 1995. They can do several tasks such as classification, clustering and visualization. Ex: SPSS clementine, Mineset, Intelligent miner
- 3rd Generation – Customized applications were developed along with data mining suites.

DM supporting Technologies overview

- Statistics
- DSS
- Machine learning
- Visualization
- Parallel processing

Main Source of Data for DM

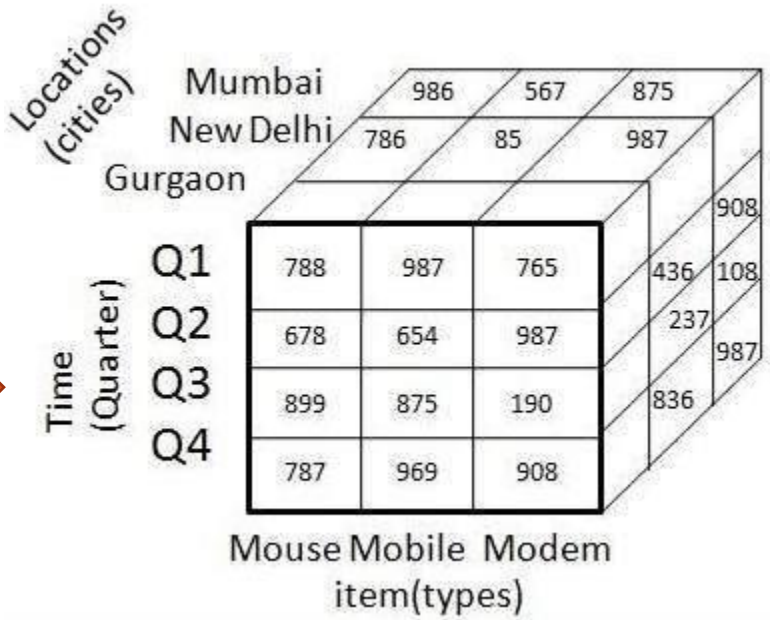
- DBMS and Datawarehouse
- DBMS – software that stores the data of day-to-day transactions
- Datawarehouse – Data store / repository of data

Characteristics of Data warehouse

- Platform of integrated, historical data
- 4 main characteristics
  - **Subject oriented** - analyse a particular subject area eg: Sales
  - **Integrated** - integrates data from multiple data sources
  - **Time-variant** - one can retrieve data from 3 months, 6 months, 12 months, or even older data
  - **Non-volatile** - data in the data warehouse will not change

Location="New Delhi"				
Time(quarter)	Item(type)			
	Entertainment	Keyboard	Mobile	Locks
Q1	500	700	10	300
Q2	769	765	30	476
Q3	987	489	18	659
Q4	666	976	40	539

Time	Location="Gurgaon"			Location="New Delhi"			Location="Mumbai"		
	Item			Item			Item		
	Mouse	Mobile	Modem	Mouse	Mobile	Modem	Mouse	Mobile	Modem
Q1	788	987	765	786	85	987	986	567	875
Q2	678	654	987	659	786	436	980	876	908
Q3	899	875	190	983	909	237	987	100	1089
Q4	787	969	908	537	567	836	837	926	987



OLTP System - Online Transaction Processing (Operational System)
OLAP System - Online Analytical Processing (Data Warehouse)

Source of data

OLTP: Operational data; OLTPs are the original source of the data.

OLAP: Consolidation data; OLAP data comes from the various OLTP Databases

Purpose of data

OLTP: To control and run fundamental business tasks

OLAP: To help with planning, problem solving, and decision support

What the data

OLTP: Reveals a snapshot of on-going business processes

OLAP: Multi-dimensional views of various kinds of business activities

Inserts and Updates

OLTP: Short and fast inserts and updates initiated by end users

OLAP: Periodic long-running batch jobs refresh the data

Queries

OLTP: Relatively standardized and simple queries Returning relatively few records

OLAP: Often complex queries involving aggregations

Processing Speed

OLTP: Typically very fast

OLAP: Depends on the amount of data involved; batch data refreshes and complex queries may take many hours; query speed can be improved by creating indexes

Space Requirements
OLTP: Can be relatively small
OLAP: Larger due to the existence of aggregation structures and history data; requires more indexes than OLTP

Database Design
OLTP: Highly normalized with many tables
OLAP: Typically de-normalized with fewer tables; use of star and/or snowflake schemas

Backup and Recovery
OLTP: Backup religiously; operational data is critical to run the business, data loss is likely to entail significant monetary loss and legal liability
OLAP: Instead of regular backups, some environments may consider simply reloading the OLTP data as a recovery method

Data Mining Process

- Goal of Data mining
  - The hidden information must be extracted
  - The obtained information must be organized to enable decision making
- 4 important steps of DM
  - Data Selection
  - Data Transformation
  - Mining the data
  - Interpretation of results

DM Techniques

- Classes of Tools
  - Association
  - Classification
  - Clustering
  - Regression
- Advanced Tools
  - Prediction
  - Estimation and Forecasting
  - Dependency networks