Data Mining

Prerequisite

- Knowledge of
  - Discrete Mathematics & Probability Theory
  - Data Structures and Algorithms

Teaching Materials

- Text Book
  - Data mining: Concepts and Techniques, by Jiawei Han and Micheline Kamber, Morgan Kaufmann, ISBN 1-55860-489-8.
- Reference Books
  - Data Mining and Analysis: Fundamental Concepts and Algorithms, M. J. Zaki & Wagner M. Jr., Cambridge Press.
  - Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, Pearson/Addison Wesley, ISBN 0-321-32136-7.
  - Machine Learning, by Tom M. Mitchell, McGraw-Hill, ISBN 0-07-042807-7

Topics

- Introduction
- Data Pre-processing
- Association Rule Mining
- Classification Techniques (Supervised Learning)
- Clustering Techniques (Unsupervised Learning)
- Semi-Supervised Learning
- Applications
  - Social network analysis
  - Opinion mining and sentiment analysis
  - Recommender systems and collaborative filtering

Data

- Data is the Latin plural of datum
- Used to represent unprocessed facts and figures without any added interpretation or analysis.
- Generally associated with some entity and often viewed as the lowest level of abstraction from which information and knowledge are derived.
- Data may be unstructured, semi-structured, and structured
- Example: The price of petrol is Rs. 90 per liter

Information

- Information is interpreted (processed) data so that it has meaning for the user.
- "The price of petrol has risen from Rs. 70 to Rs. 90 per liter" – is information for a person who tracks petrol prices.
- Data becomes information when it is processed for some purpose and adds value for the recipient.
- A set of raw sales figures – Data
- Sales report (chart plotting, trend analysis) – Information

Knowledge

- Knowledge is a fluid mix of information, experience and insight that may benefit the individual or the organization.
- "When petrol prices go up by Rs. 20 per liter, it is likely that bus fare will rise by 25%" is knowledge.
- The boundaries between data, information, and knowledge is fuzzy
- What is data to one person is information to someone else.

Summarized View

- Data – as in databases, spreadsheets, text files…
- Information – Processed data
- knowledge – **Fluid mix of information, experience, and insight**

OR, knowledge is a meta information about the patterns hidden in the data

The patterns must be discovered automatically!!!

Data Categories & Mining Terminologies

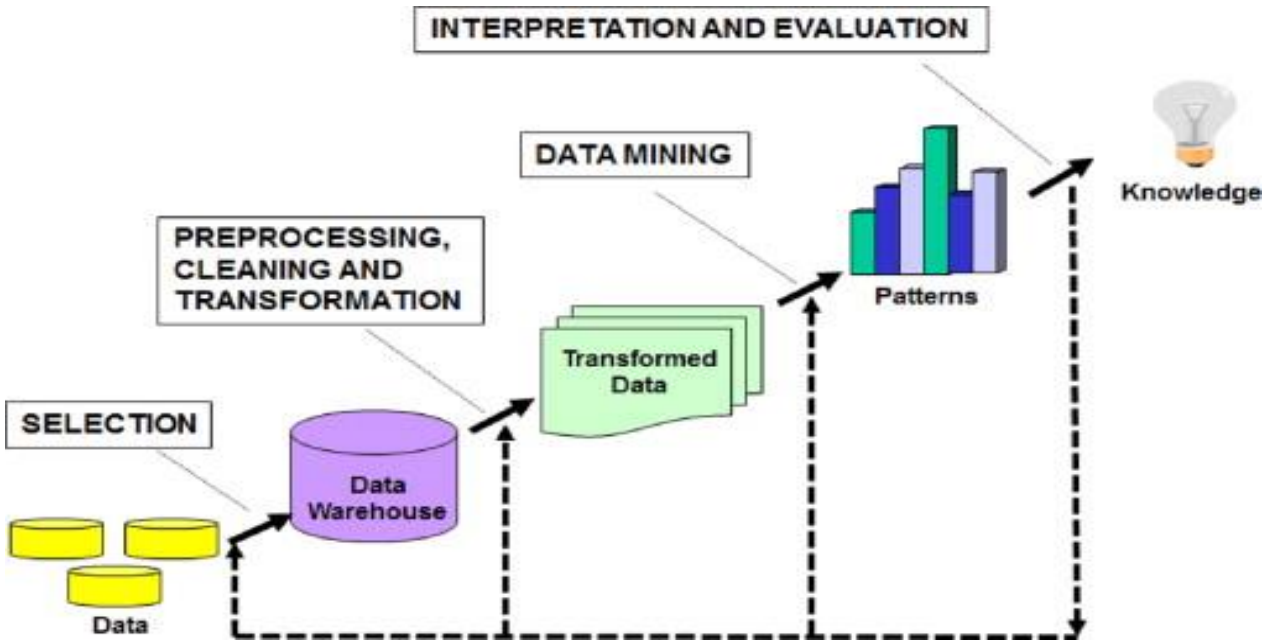
Data are stored in Documents (A file)

Unstructured	Semi-structured	Structured
A file stored on your PC	A web page stored on WWW	A database
(Text Mining)	(Web Mining)	(Data Mining)

What is Data Mining?

- An important step of KDD (Knowledge Discovery from Databases) process
- Data mining is the automatic extraction of interesting knowledge (rules, regularities, patterns, constraints) from large data sources, e.g., databases, texts, web, images, etc.
- Identified patterns must be:
  - Valid, novel (non-trivial), potentially useful, and understandable

The KDD Process



Data Mining Objectives

- Identification of data as a source of useful information
- Automatic extraction of valid and novel patters from the data source
- Use of discovered patterns for competitive advantages when working in business environment

Why Data Mining?

- Data Explosion (Information Overload) problem
  - We are drowning in data, but starving for knowledge!
  - Data data everywhere nor any drop of insight!

    (water water everywhere nor any drop to drink)
- Explosive growth of data: from Terabytes to Petabytes
- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses, and other data repositories

Why Data Mining? Cont…

- Major sources of abundant data
  - Business: Web, e-commerce, transactions, stocks, …
  - Science: Remote sensing, bioinformatics, scientific simulation
  - Society and everyone: news, digital cameras,
- The computing power is not an issue.
- Data mining tools are available
- The competitive pressure is very strong.
  - Almost every company is doing (or has to do) it

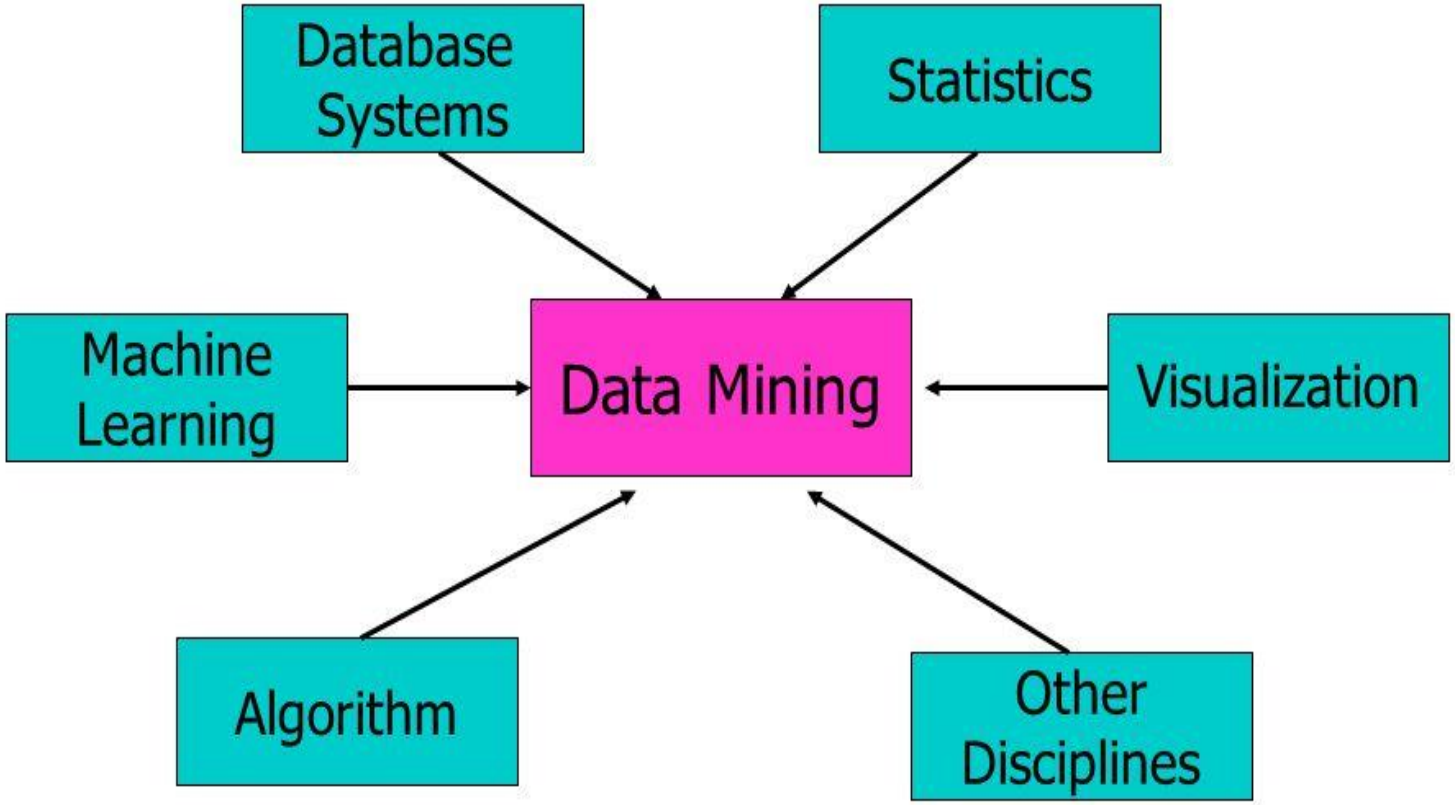
Why Data Mining Important?

- Digitization of businesses produce huge amount of data
  - How to make best use of data?
  - Knowledge discovered from data can be used for competitive advantage.
- E-businesses are generating huge amount of datasets
  - Online retailers (e.g., amazon.com) are largely driving by data mining.
  - Web search engines are information retrieval (text mining) and data mining companies

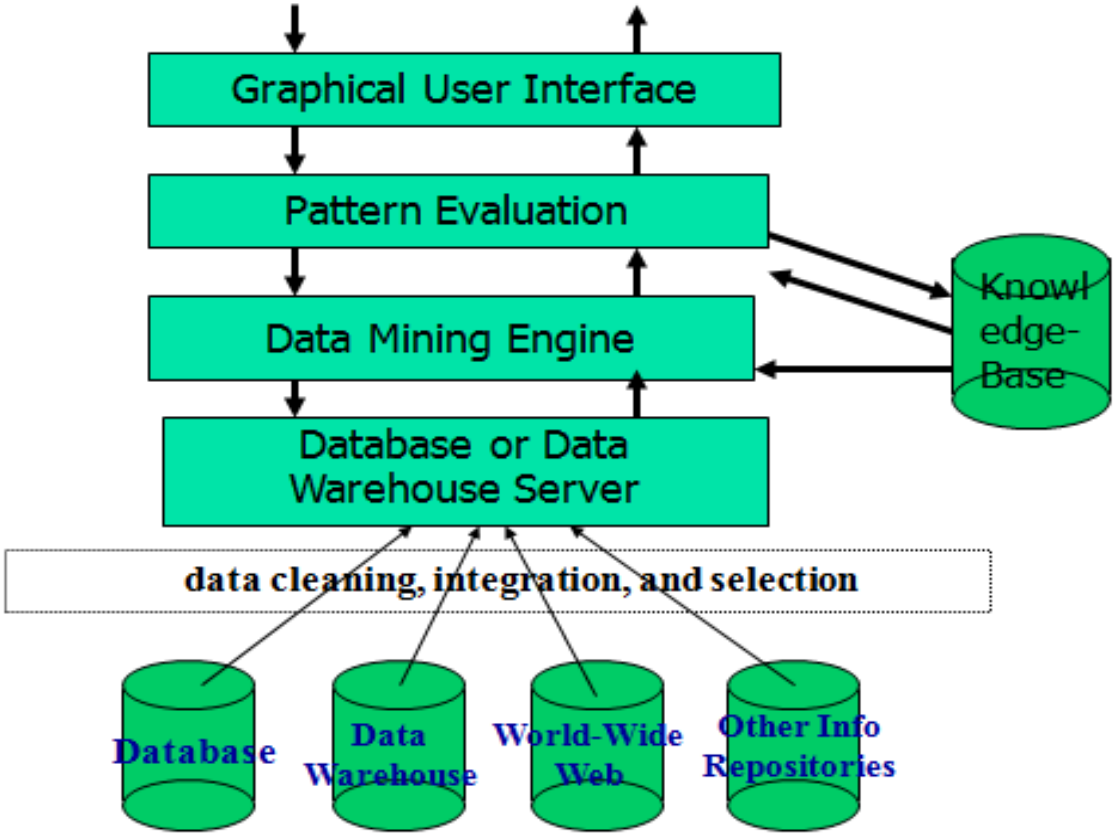
Why is Data Mining Necessary?

- Make use of your data assets (knowledge-based economy)
- Big gap from stored data to knowledge
  - Transition won't occur automatically.
- Many interesting things can't be found using database queries
  - Customers likely to buy my products?
  - Why sale was down after demonitization?
  - Which items should be recommended to a person purchasing computer?

Data Mining: Confluence of Multiple Disciplines



Architecture: Typical Data Mining System



Data Mining: On what kind of data?

- Relational Databases
- Data Warehouses
- Transactional Databases
- Advanced DB and Data Repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - Heterogeneous and legacy databases
  - Web database

Data Mining Functionalities:

Characterization (1)

- A data mining process aims to find rules that describe the properties of a concept.
- Standard form:

If concept then characteristics

- C=1 → A=1 & B=3 (Support: 25%, i.e., there are 25% records for which the rule is true)
- C=1 → A=1 & B=4 (Support: 17%)
- C=1 → A=0 & B=2 (Support: 16%)

Data Mining Functionalities:
Discrimination (2)

- A data mining process which aims is to find rules that allow us to discriminate the objects (records) belonging to a given concept (one class) from the rest of records (classes)
- Standard form:

  If characteristics then concept
- A=0 & B=1 → C=1 (Support: 33%, Confidence: 83%)
  - Confidence: The conditional probability of the concept given the characteristics
- A=2 & B=0 → C=1 (27%, 80%)
- A=1 & B=1 → C=1 (12%, 76%)

Data Mining Functionalities:
Classification and Prediction (3)

- Finding models (rules) that describe (characterize) and/or distinguish (discriminate) classes or concepts for future prediction.
  - Classify countries based on climate (characteristics)
  - Classify cars based on gas mileage and use it to predict classification of a new car
- Presentation:
  - Decision Tree
  - Classification Rules
  - Neural Network
  - Bayes Network

Data Mining Functionalities:
Prediction (statistical) (4)

- A Data Mining process to predict some unknown or missing numerical values.
- Output space: continuous

Data Mining Functionalities:
Association Analysis (5)

- A Data Mining process which aims to identify patterns (aka frequent itemsets) in data
- For example:
  - Buy(X, Printer) → Buy (X, Cartridge)
  - Buy (X, Bread) → Buy (X, Butter) ∧ Buy (X, Milk)

Data Mining Functionalities:
Cluster Analysis (6)

- Unsupervised learning
- Aims to group data to form new classes
  - Cluster houses to find distribution patterns
- Basic principle: Maximizing the intra-class similarity and minimizing the inter-class similarity

Data Mining Functionalities:
Outlier Analysis (7)

- Outlier: A data object that does not comply with the general behavior of the data
- It can be considered as noise or exception, but is quite useful in fraud detection, rare events analysis, etc.

Major issues in Data Mining

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages
- Expression and visualization of data mining results
- Handling noise and incomplete data
- Pattern evaluation: the interestingness problem
- Efficiency and scalability of data mining algorithms
- Parallel, distributed, and incremental mining methods

Major issues in Data Mining (cont…)

- Handling relational and complex types of data
- Mining information from heterogeneous databases and global information systems (WWW)
- Application of discovered knowledge
  - Domain-specific data mining tools
  - Intelligent query answering
  - Process control and decision making
- Integration of the discovered knowledge with existing knowledge: A knowledge fusion problem
- Protection of data security, integrity, and privacy

Data Mining Applications (1)

- Target marketing, customer relation management, market basket analysis, cross selling,
- Forecasting, customer retention, quality control, competitive analysis
- Text mining (news group, email, documents) and Web analysis.
- Intelligent query answering
- Buying patterns
- Decision support
- Fraud detection

Data Mining Applications (2)

- Scientific Applications
  - Networks failure detection
  - Controllers design
  - Geographic Information Systems
  - Genome - Bioinformatics
  - Intelligent robots