Data Mining

(CS C 415/CS F 415/IS C415)

BITS Pilani
K K Birla Goa Campus

Mrs.Aruna.G
Dept. of CS and IS

Text and Reference Books

Prescribed Text Book (S)

T1. Tan,Pang-Ning and other “Introduction to Data Mining” Pearson Education,2006.

Reference Book (S)

R1. Han J & Kamber M, “**Data Mining: Concepts and Techniques”,** Morgan Kaufmann Publishers, 2001

R2. Hand D, Mannila H, & Smyth P, “***Principles of Data Mining***”, MIT Press, 2001.

R3. Pujari A K, “Data Mining Techniques”, University Press (India), 2001.

R4. Kimball R, “The Data Warehouse Toolkit”, 2e, John Wiley, 2002.

Evaluation Components

- Test 1 : 20%
- Test 2 : 20%
- Project/Presentation : 30%
- Compre : 30%

innovate achieve lead

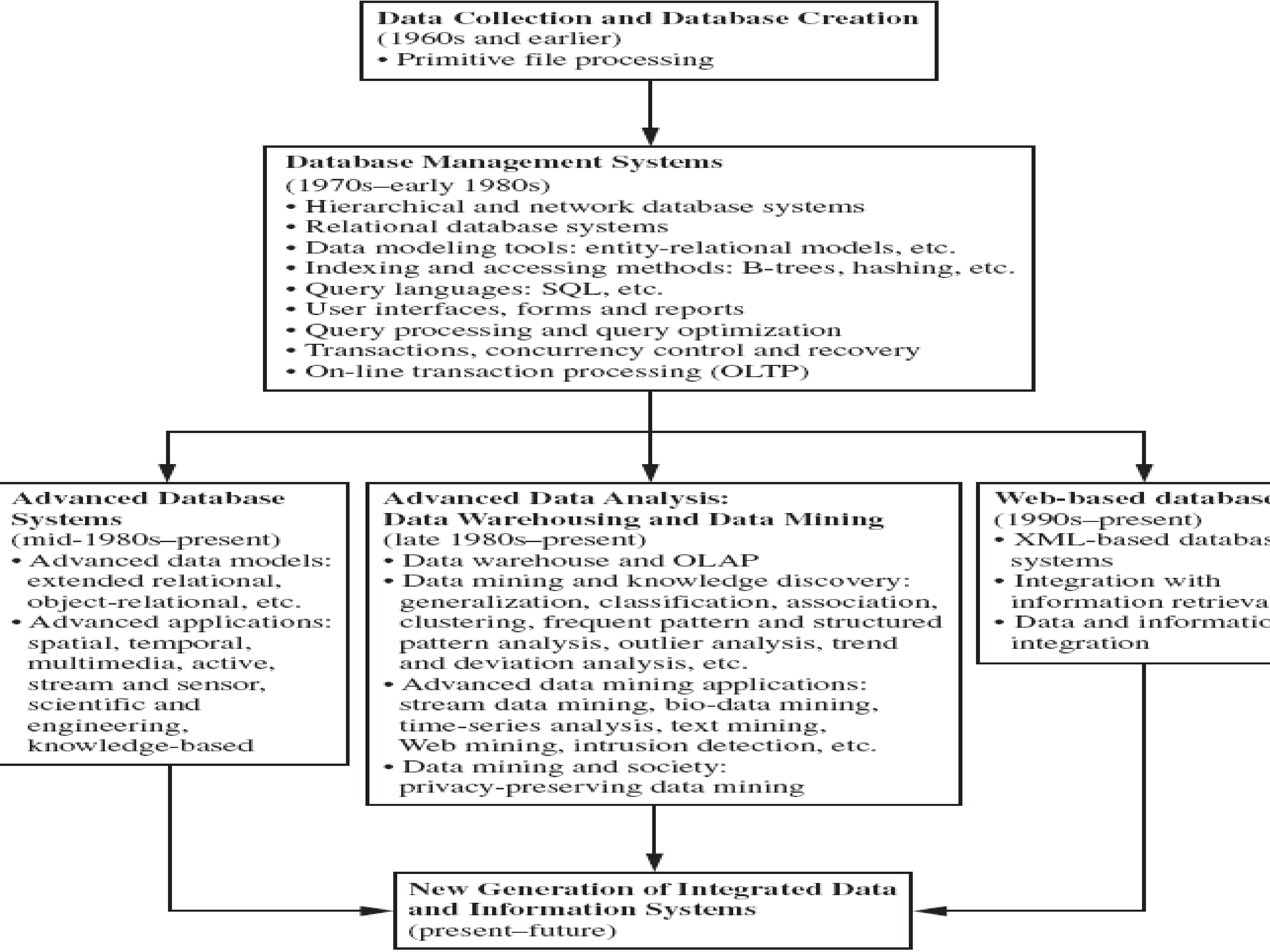
BITS Pilani
K K Birla Goa Campus

Lecture 1: Introduction

Road Map

- Motivation
- What is DM?
- DM Tasks
- Applications
- Issues / Challeneges of DM

Evolution of Database Technology



Data Collection and Database Creation
(1960s and earlier)
- Primitive file processing

Database Management Systems
(1970s—early 1980s)
- Hierarchical and network database systems
- Relational database systems
- Data modeling tools: entity-relational models, etc.
- Indexing and accessing methods: B-trees, hashing, etc.
- Query languages: SQL, etc.
- User interfaces, forms and reports
- Query processing and query optimization
- Transactions, concurrency control and recovery
- On-line transaction processing (OLTP)

Advanced Database Systems
(mid-1980s—present)
- Advanced data models: extended relational, object-relational, etc.
- Advanced applications: spatial, temporal, multimedia, active, stream and sensor, scientific and engineering, knowledge-based

Advanced Data Analysis: Data Warehousing and Data Mining
(late 1980s—present)
- Data warehouse and OLAP
- Data mining and knowledge discovery: generalization, classification, association, clustering, frequent pattern and structured pattern analysis, outlier analysis, trend and deviation analysis, etc.
- Advanced data mining applications: stream data mining, bio-data mining, time-series analysis, text mining, Web mining, intrusion detection, etc.
- Data mining and society: privacy-preserving data mining

Web-based database
(1990s—present)
- XML-based databas systems
- Integration with information retrieva
- Data and informatic integration

New Generation of Integrated Data and Information Systems
(present—future)

Road Map

- Motivation
- What is DM?
- DM Tasks
- Applications
- Issues in DM

What motivated DM?

- Necessity is the mother of invention – Plato

- The Explosive Growth of Data:
  - from terabytes to petabytes

  - Data collection and data
  - availability

  - Major sources of abundant
  - data

Contd……..

- Data collection and data availability
- Automated data collection tools, database systems, Web, computerized society
- Major sources of abundant data
- Business: Web, e-commerce, transactions, stocks, …
- Science: Remote sensing, bioinformatics, scientific simulation, …
- Society and everyone: news, digital cameras, YouTube

Contd........

- We are drowning in data, but starving for knowledge! (Much of the data is never analyzed at all)

Why DM?

- Consider the data of all educated people ...
- Consider the data of white collar crimes...
- To decide whether the crime rate is increasing or decreasing along with the rate of education......
- To be more specific , what is the rate of people from rural and urban areas who involve in the crimes.......

Contd .....

- Strategic Decision Making
- Wealth Generation
- Analyzing Trends

Road Map

- Motivation
- What is DM?
- DM Tasks
- Applications
- Issues in DM

What is DM?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Contd .....

- Alternative names

- Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

DM on what kind of Data?

- Relational Database
- Data Warehouse (is a repository of information collected from multiple sources, stored under a unified schema, and usually resides at a single site)

Contd......

- Advanced data and information systems
- Object-oriented database
- Temporal DB, Sequence DB and Time serious DB
- Spatial DB
- Text DB and Multimedia DB
- … and WWW

What is DM?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data.

Types of interestingness

- Frequancy
- Rarity
- Correlation
- Length Of Occurance (for sequance and temporal data)
- Consistency
- Repeating / Periodicity
- Abnormal Behaviour
- Other patterns ……

Contd .....

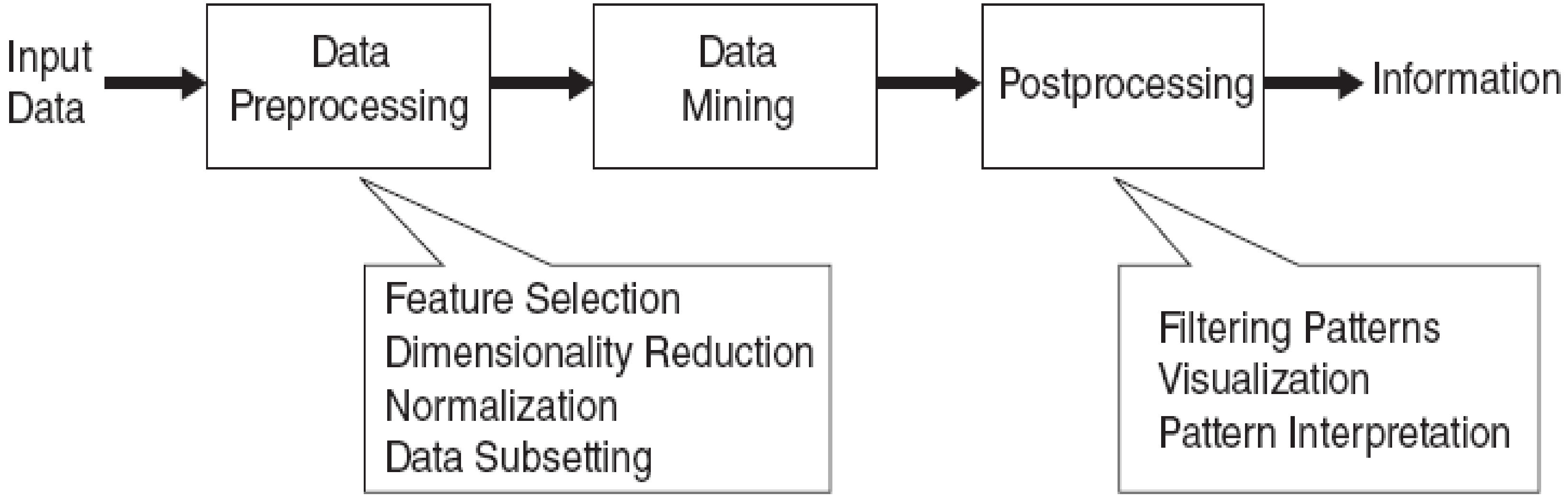
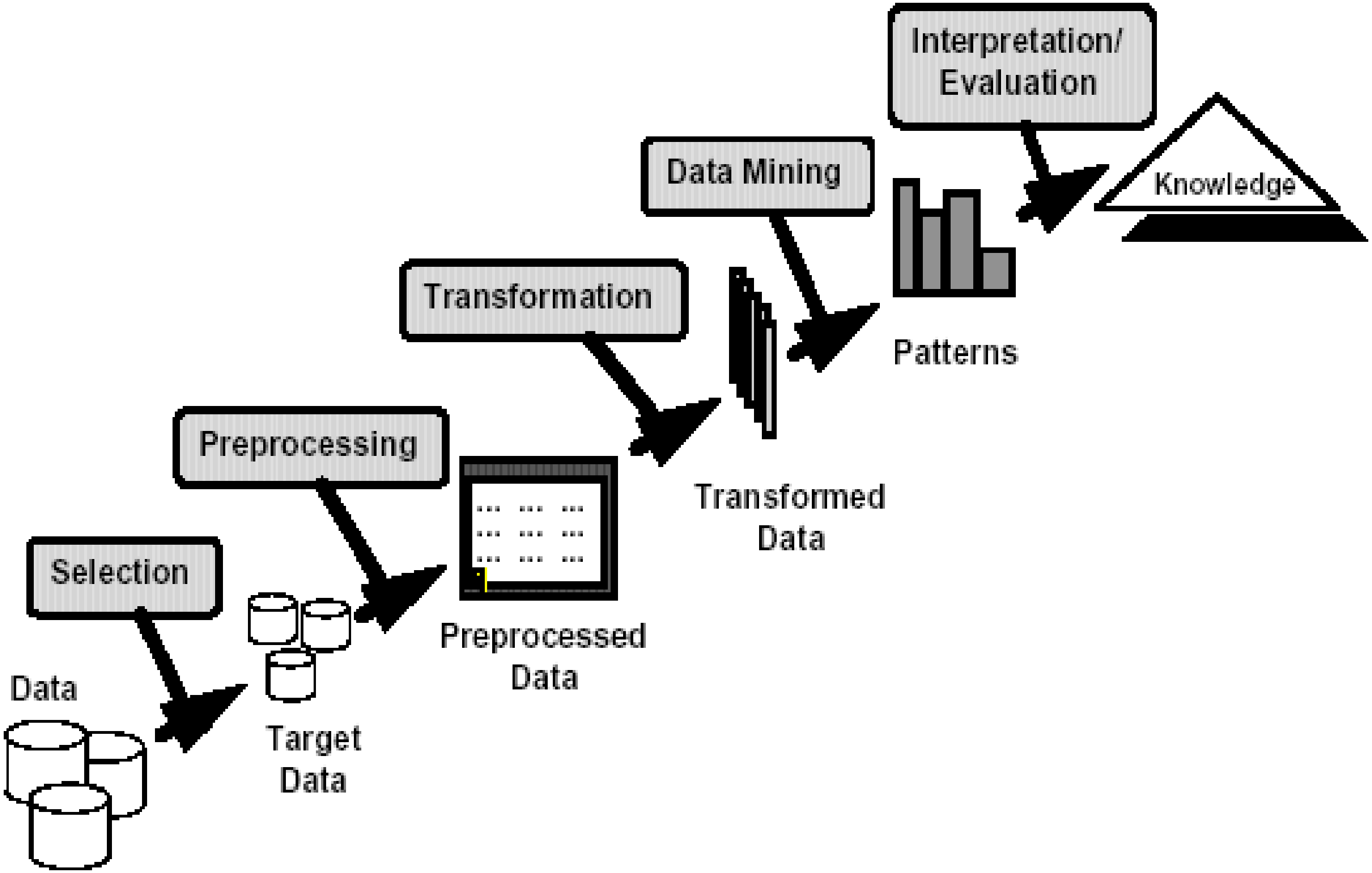
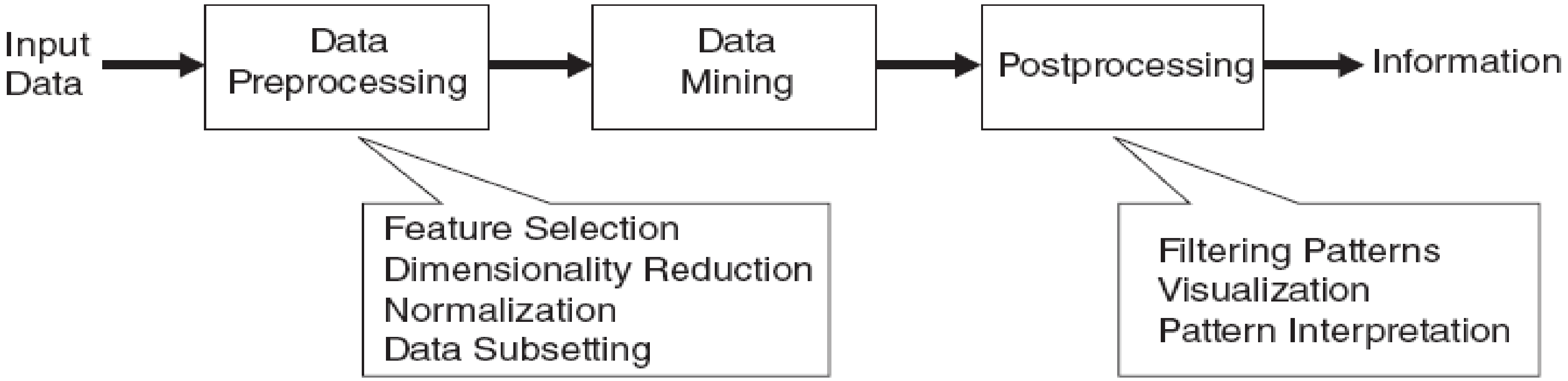


Figure 1.1. The process of knowledge discovery in databases (KDD).

Contd .....


Interpretation/ Evaluation
Data Mining
Knowledge
Transformation
Patterns
Preprocessing
Transformed Data
Selection
Preprocessed Data
Data
Target Data

Contd .....



Input Data

Data Preprocessing

Data Mining

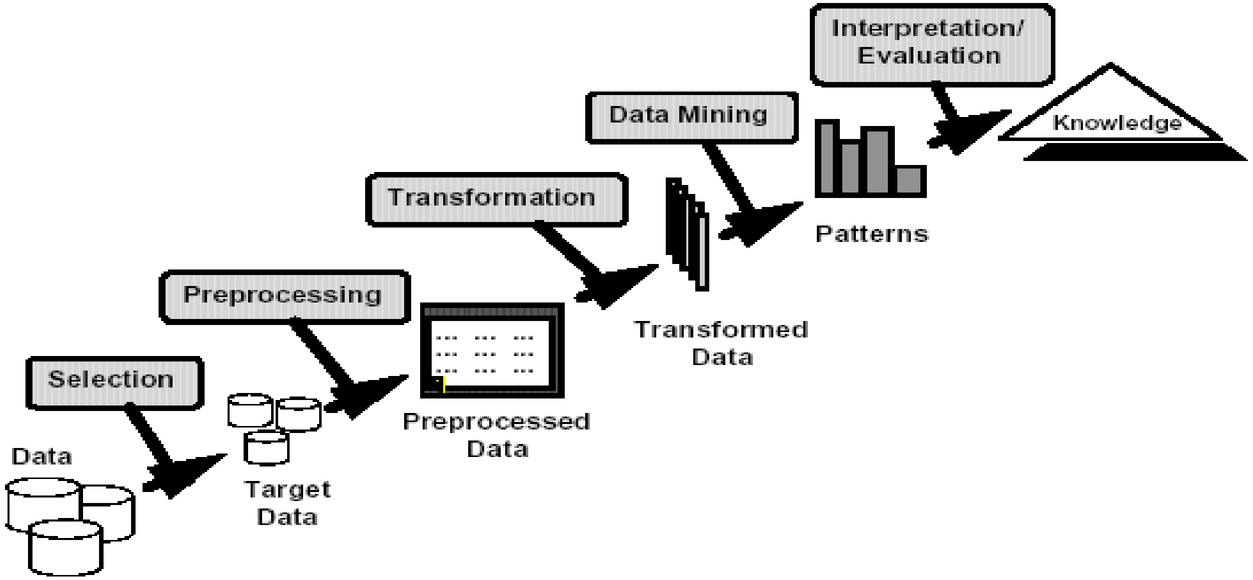
Postprocessing

Information

Feature Selection
Dimensionality Reduction
Normalization
Data Subsetting

Filtering Patterns
Visualization
Pattern Interpretation

Figure 1.1. The process of knowledge discovery in databases (KDD).



Contd .....

- The process of converting raw data into useful information.

- We always try to represent in a model so that the end user can interpret the reults in a meaningful manner.

What is (not) DM?

- **What is not Data Mining?**
  - Look up phone number in phone directory
  - Query a Web search engine for information about “Amazon”

- **What is Data Mining?**
  - Certain names are more prevalent in certain North India locations (Srivastava Jain, Chawla…)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Data Mining:
Confluence of Multiple Disciplines

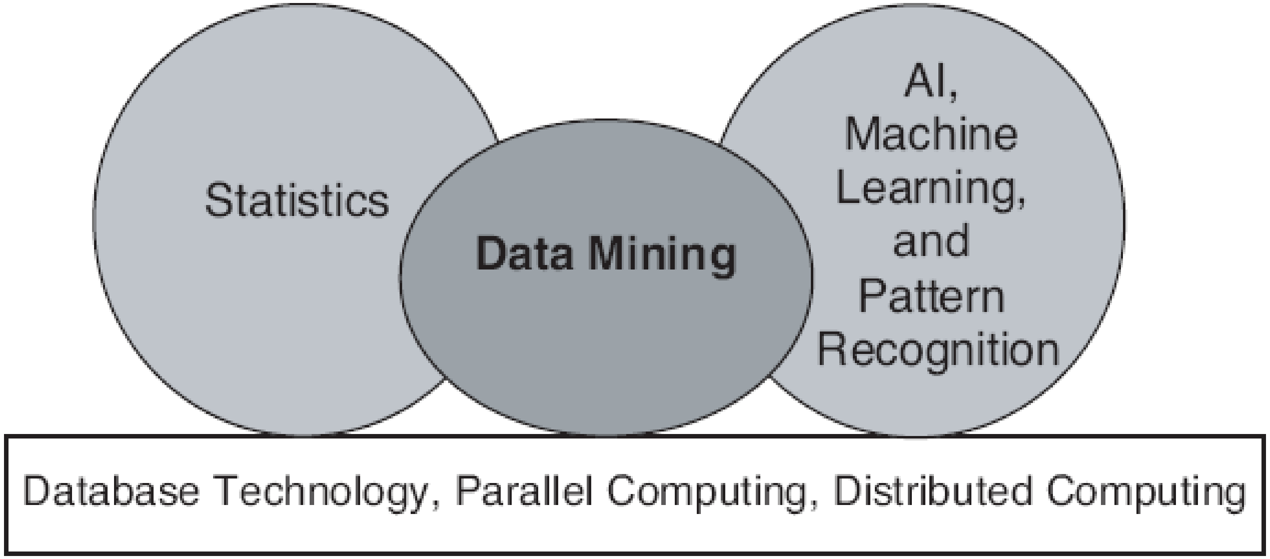

Statistics
Data Mining
AI, Machine Learning, and Pattern Recognition
Database Technology, Parallel Computing, Distributed Computing

Figure 1.2. Data mining as a confluence of many disciplines.

Road Map

Motivation

What is DM?

DM Tasks

Applications

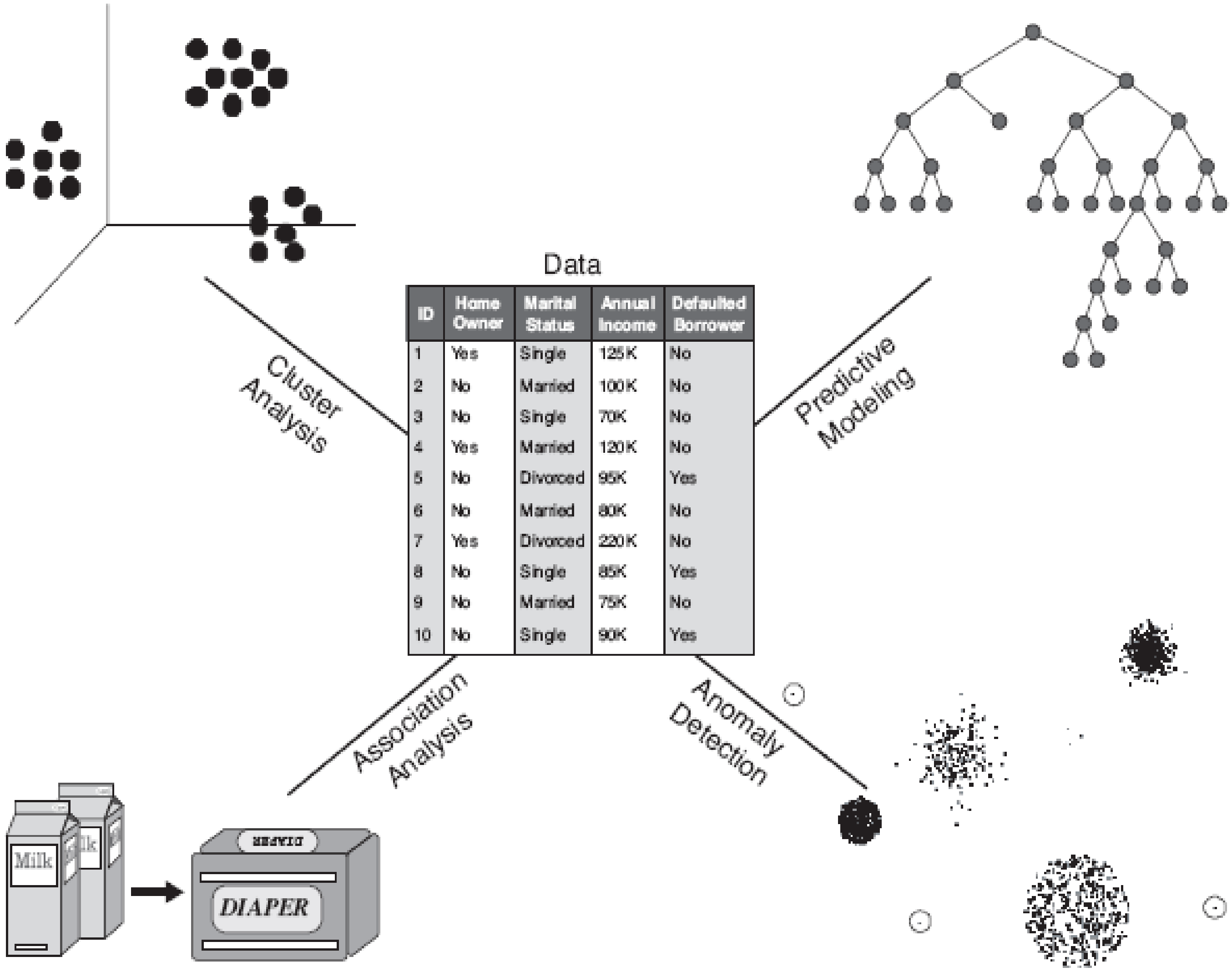
Issues in DM

DM Tasks

- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
- Description Methods
  - Find human-interpretable patterns that describe the data.

Contd .....

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]



ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 1.3. Four of the core data mining tasks.

Road Map

- Motivation
- What is DM?
- DM Tasks
- Applications
- Issues in DM

Applications

- Commercial Applications / Business Apllications

  Financial Applications (Bank ,Stock Exchange)
  E-Commerce (Flipcart ,eBay...)

- Scientific Applications

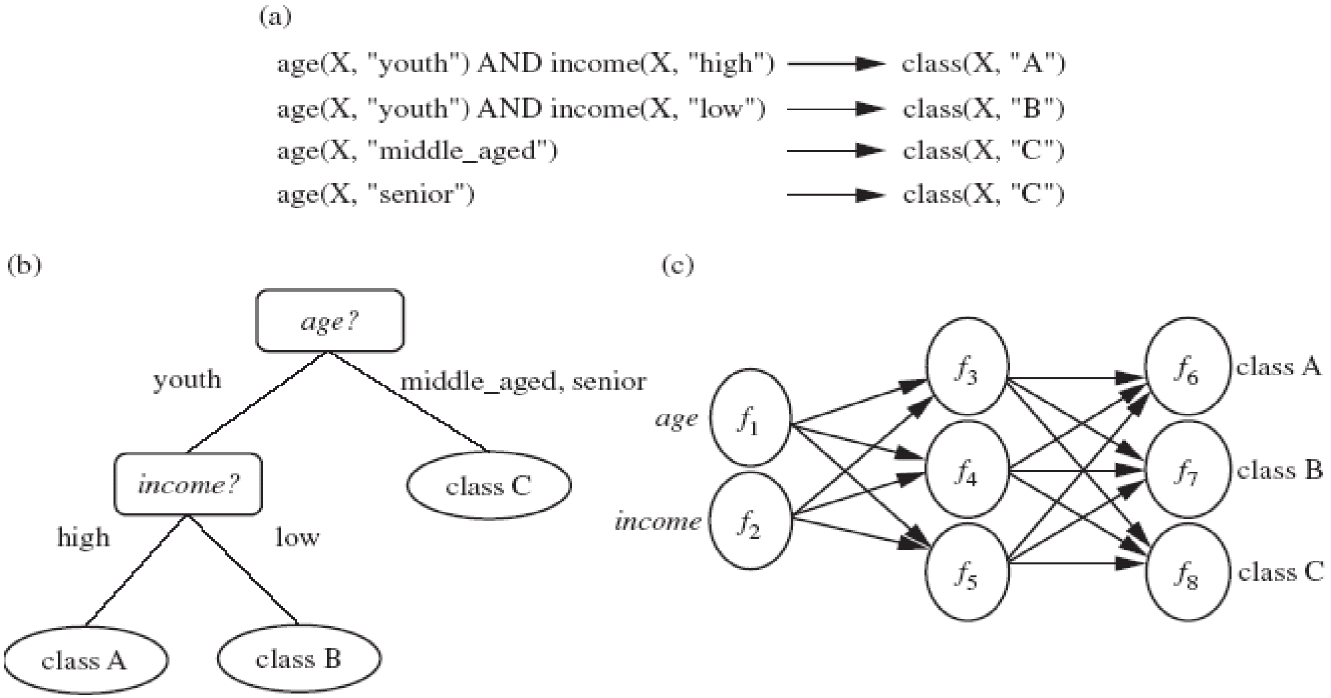
  Astronomical Applications
  Weather Information
  Earth ScienceApplications

- Social Applications

  - FaceBook
  - Economy of People

Classification .....

- **Classification** is the process of finding a MODEL that describes and distinguish data classes or concepts



(a)

age(X, "youth") AND income(X, "high") ⟶ class(X, "A")

age(X, "youth") AND income(X, "low") ⟶ class(X, "B")

age(X, "middle_aged") ⟶ class(X, "C")

age(X, "senior") ⟶ class(X, "C")

Classification Application 1

- Direct Marketing
  - Goal: Reduce cost of mailing by *targeting* a set of consumers likely to buy a new cell-phone product.
  - Approach:
  - Use the data for a similar product introduced before.
  - We know which customers decided to buy and which decided otherwise. This *{buy, don’t buy}* decision forms the *class attribute*.
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers.
  - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

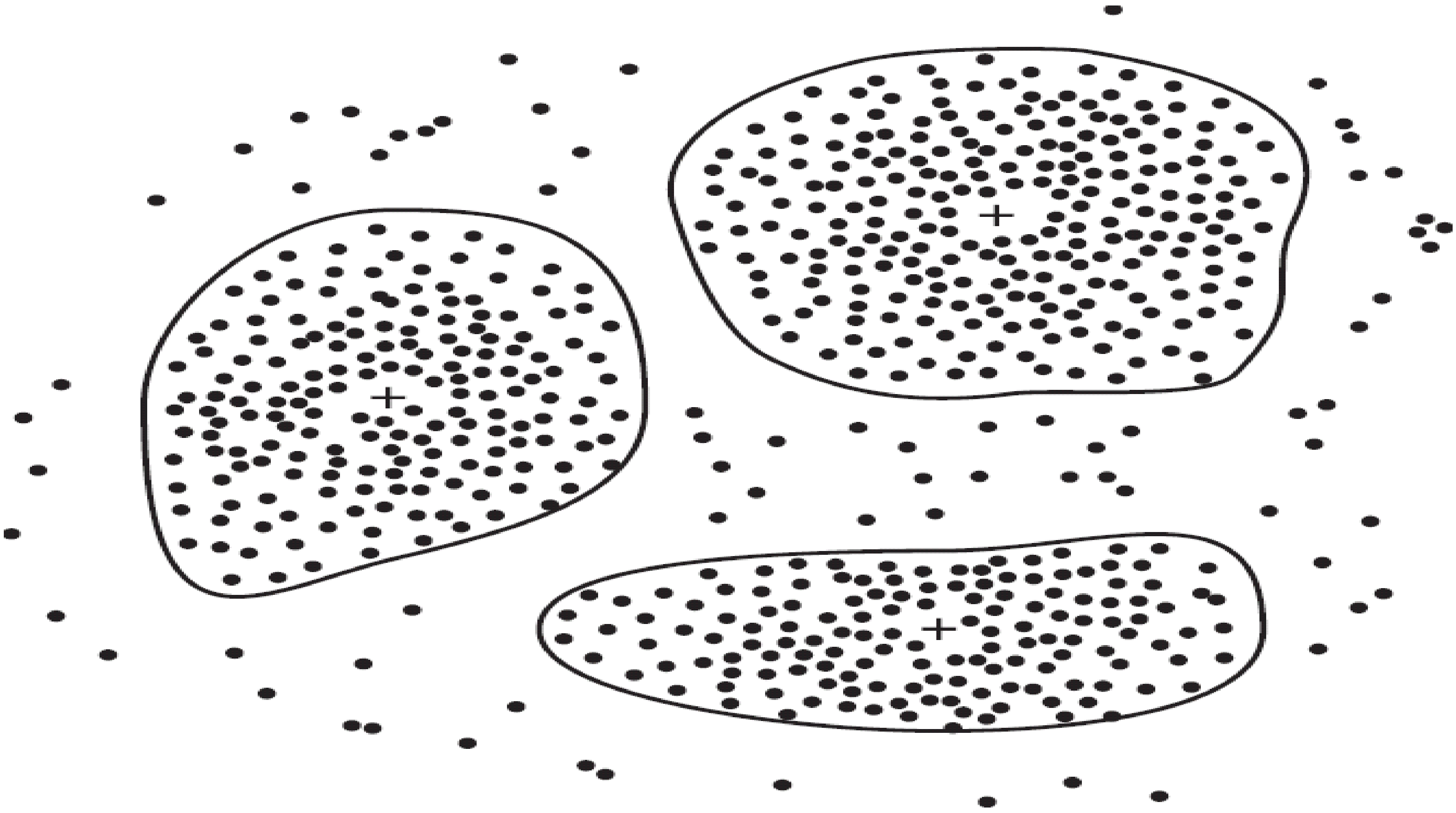
Classification Application 2

- Fraud Detection
  - Goal: Predict fraudulent cases in credit card transactions.
  - Approach:
  - Use credit card transactions and the information on its account-holder as attributes.
  - When does a customer buy, what does he buy, how often he pays on time, etc
  - Label past transactions as fraud or fair transactions. This forms the class attribute.
  - Learn a model for the class of the transactions.
  - Use this model to detect fraud by observing credit card transactions on an account.

Clustering

- In general, the class label are not present in the training data simply they are not known to begin with
- The objects are clustered or grouped based on the principle of *maximizing the intra-cluster similarity* and *minimizing the inter-cluster similarity*

Clustering



Clustering: Application 1

- Market Segmentation:
  - Goal: subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
  - Collect different attributes of customers based on their geographical and lifestyle related information.
  - Find clusters of similar customers.
  - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

Clustering: Application 2

- Document Clustering:
  - Goal: To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach: To identify frequently occurring terms in each document. Form a similarity measure based on the frequencies of different terms. Use it to cluster.
  - Gain: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents.

Association Rule Discovery

- **Frequent patterns** are patterns that occur frequently in data

- Association analysis:

  - Example: buys(X,”computer”) => buys(X,”software”) [support = 1%, confidence = 50%]

Association Rule Discovery: Application 1

- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
  - If a customer buys diaper and milk, then he is very likely to buy cheese.

Sequential Pattern Discovery

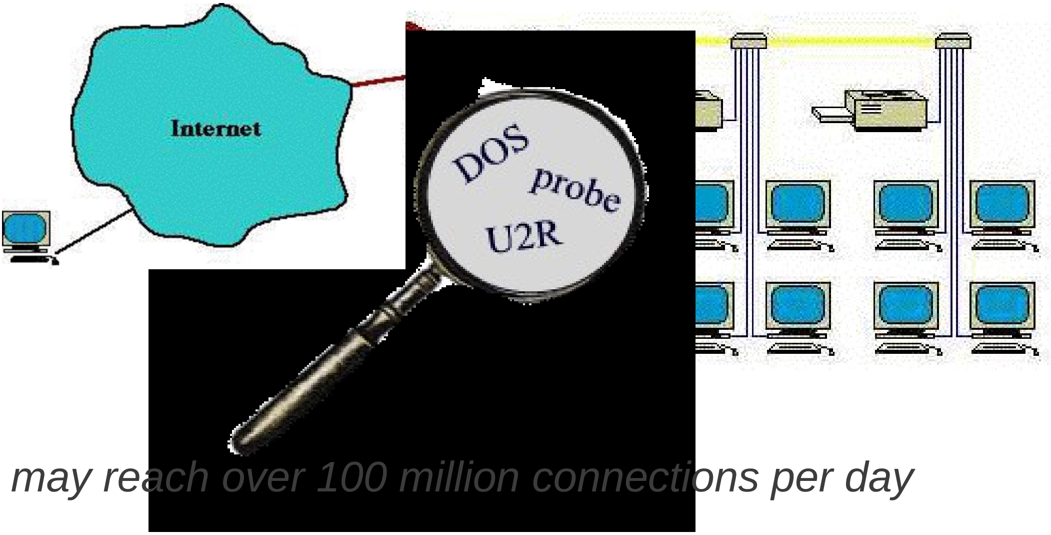
- Applications of sequential pattern mining
  - Customer shopping sequences:
  - First buy computer, then CD-ROM, and then digital camera, within 3 months.
  - Medical treatments, natural disasters (e.g., earthquakes), science & eng. processes, stocks and markets, etc.
  - Telephone calling patterns, Weblog click streams
  - DNA sequences and gene structures

Regression

- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Greatly studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation Detection

- Applications:
  - Credit Card Fraud Detection
- Network Intrusion
  - Detection



- *Typical network traffic at University level may reach over 100 million connections per day*

Road Map

- Motivation
- What is DM?
- DM Tasks
- Applications
- Issues in DM

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

Scalability

- Algorithms generally:
- Operate on data with assumption of in-memory processing of entire data set
- Operate under assumption of developers will address I/O and other performance scaling issues
- Or just don't address scalability within resource constraints at all

Dimensionality

LSST : 1000's of dimensions.

- Massive data stream: ~2Terabytes of image data per
- hour that must be mined in real time (for 10 years).

- Massive 20-Petabyte database: more than 50 billion objects need to be classified, and most will bemonitored for important variations in real time.

- Massive event stream: knowledge extraction in real time for 100,000 events each night.

Complex & Heterogenous data

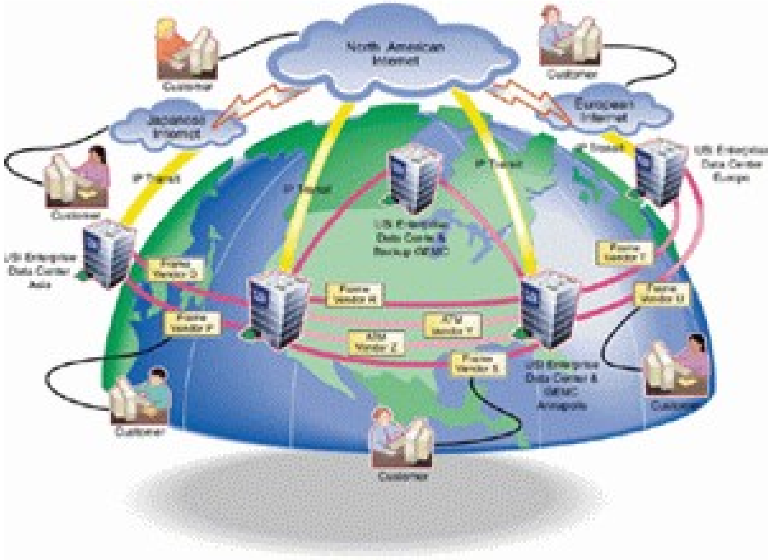
Data are usually heterogeneous (e.g., databases, images, catalogs, file systems, web interfaces, document libraries, binary, text, structured, unstructured, …)

Data Quality

- Data quality problems are expensive and pervasive
- DQ problems cost hundreds of billion $$$ each year.
- Resolving data quality problems is often the biggest
- effort in a data mining study.

Data Ownership & Distribution

- Distributed data are the norm
- (across people, institutions,
- projects, agencies, nations, …)



Privacy Preservation

- A Scenario in which two parties owning confidential
- databases wish to run a data mining algorithm on the
- union of their databases, without revealing any
- unnecessary information.

- The need to both protect privileged information and
- enable its use for research or other purpose.

Streaming Data

- Traditional data mining techniques usually require entire data set to be present.
- Random access (or multiple access) to the data.
- Impractical to store the whole data.
- Simple calculation per data due to time and space constraints.