PART 1
BASIC CONCEPTS

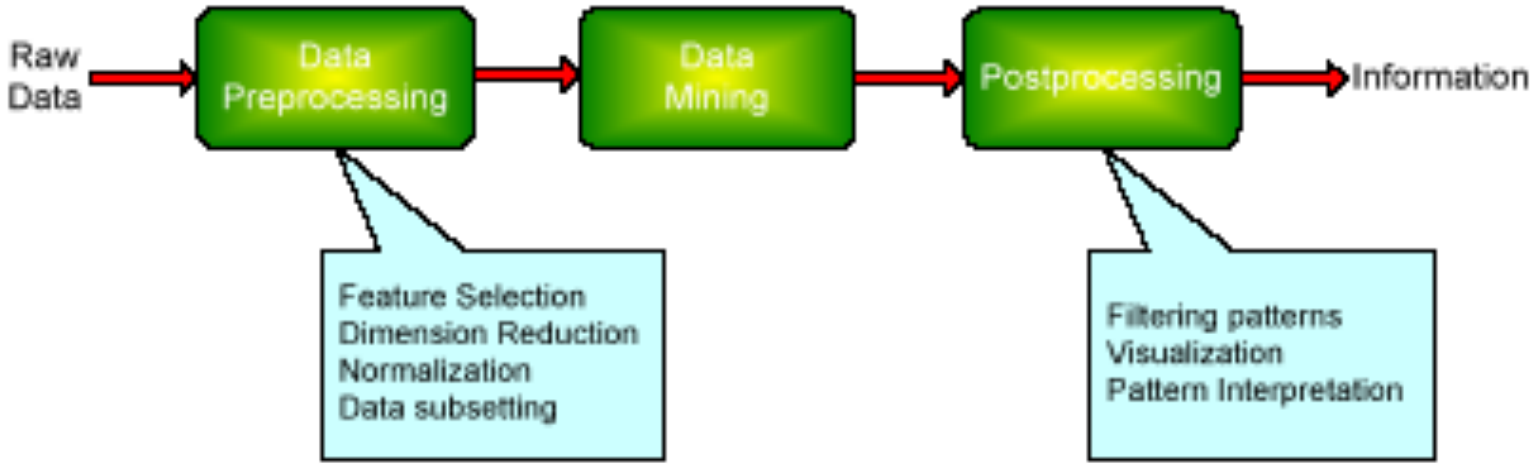
NGURAH AGUS SANJAYA ER, S.KOM, M.KOM, PH.D
E-mail: agus.sanjaya@cs.unud.ac.id

Evolution of Database Technology

- 1960s:
  - **Data collection, database creation, IMS and network DBMS**
- 1970s:
  - **Relational data model, relational DBMS implementation**
- 1980s:
  - **RDBMS, advanced data models (extended-relational, OO, deductive, etc.) and application-oriented DBMS (spatial, scientific, engineering, etc.)**
- 1990s—2000s:
  - **Data mining and data warehousing, multimedia databases, and Web databases**

What is Data Mining ?

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Data mining is an integral part of Knowledge Discovery in Databases (KDD)



Knowledge Discovery in Databases (KDD)

Motivation:
"Necessity is the Mother of Invention"

- Data explosion problem
  - **Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories**
- **We are drowning in data, but starving for knowledge!**
- Solution: Data warehousing and data mining
  - **Data warehousing and on-line analytical processing**
  - **Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases**

Data Explosion

"We are drowning in data, but starving for knowledge"

"The amount of data stored in various media has doubled in three years, from 1999 to 2002. The amount of data put into storage in 2002, five exabytes (one quintillion bytes), was equal to the contents of a half a million new libraries, each containing a digitised version of the print collection of the entire US Library of Congress" (Lyman and Varian, UC Berkeley, 2003)

Scale of Data

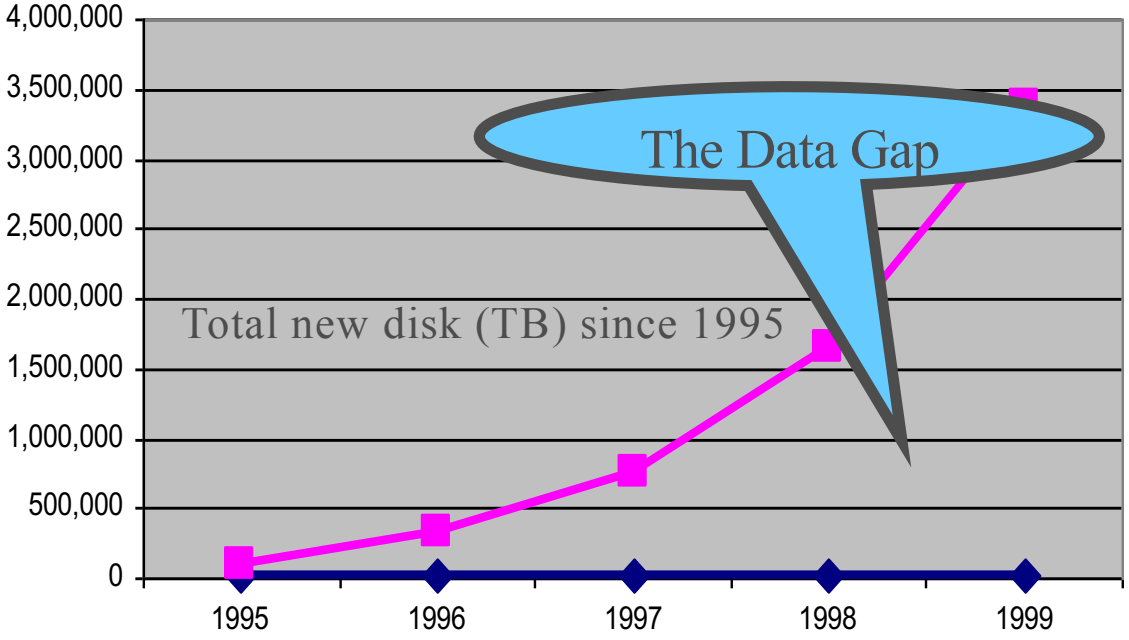
Organization	Scale of Data
Walmart	~ 20 million transactions/day
Google	~ 8.2 billion Web pages
Yahoo	~10 GB Web data/hr
NASA satellites	~ 1.2 TB/day
NCBI GenBank	~ 22 million genetic sequences
France Telecom	29.2 TB
UK Land Registry	18.3 TB
AT&T Corp	26.2 TB

"The great strength of computers is that they can reliably manipulate vast amounts of data very quickly. Their great weakness is that they don't have a clue as to what any of that data actually means"

(S. Cass, IEEE Spectrum, Jan 2004)

Why Mine Data ? - Motivation

- There is often information "*hidden*" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Why Mine Data?
- Commercial Viewpoint

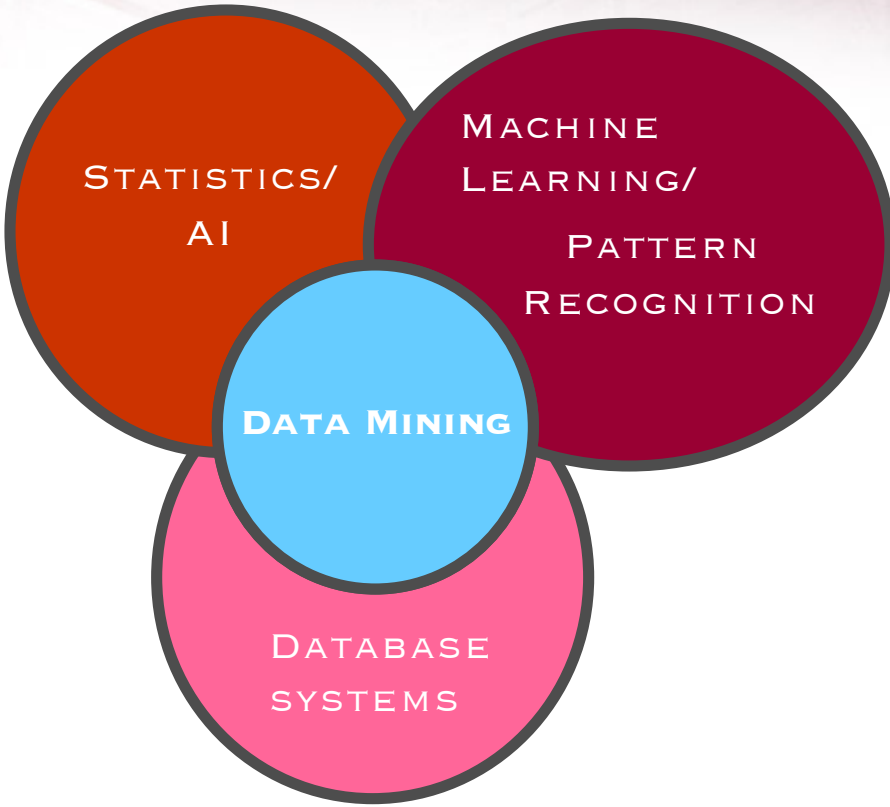
- Lots of data is being collected and warehoused
  - **Web data, e-commerce**
  - **purchases at department/grocery stores**
  - **Bank/Credit Card transactions**
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - **Provide better, customized services for an *edge* (e.g. in Customer Relationship Management)**

Why Mine Data? - Scientific Viewpoint

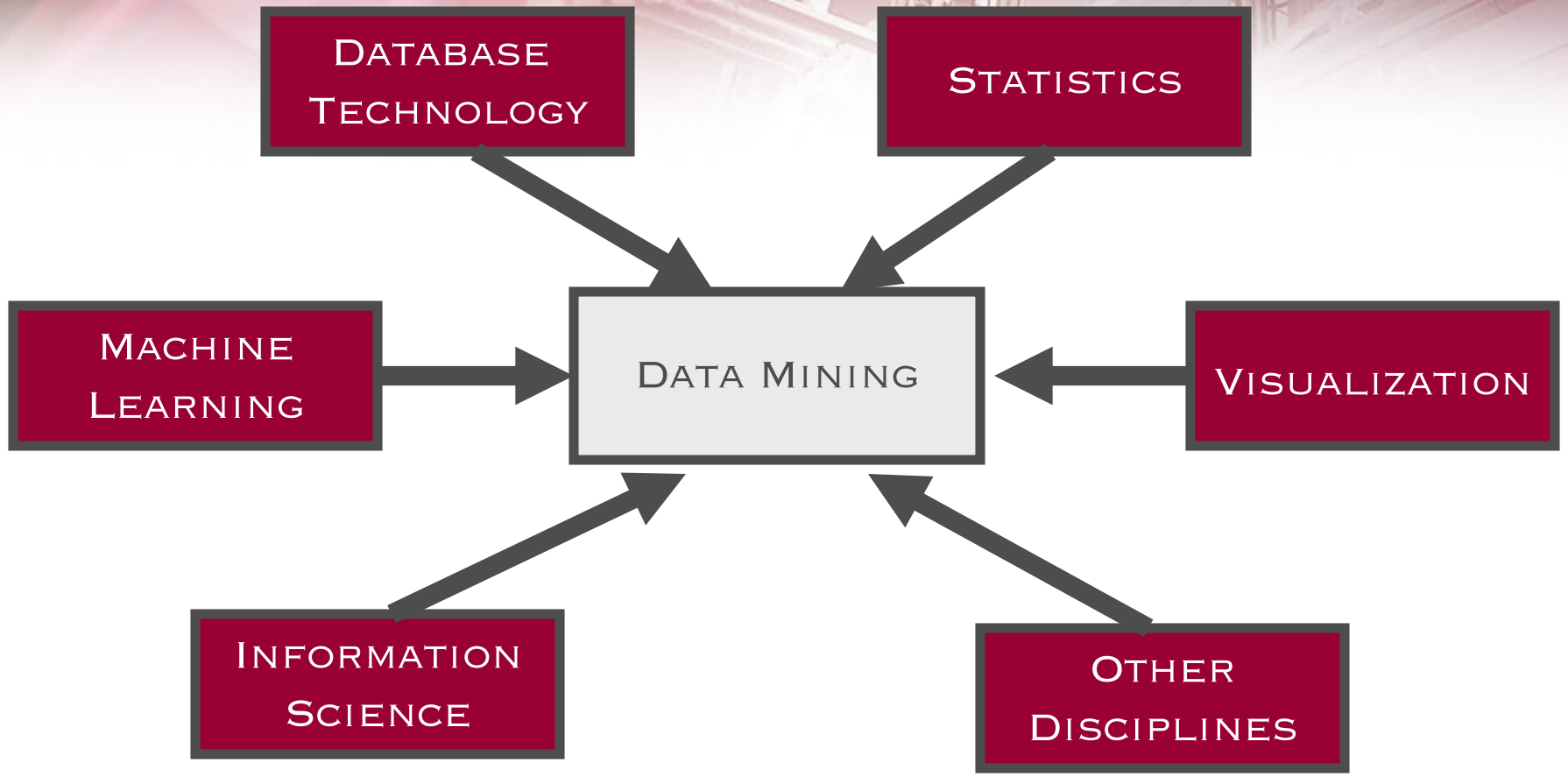
- Data collected and stored at enormous speeds (GB/hour)
  - **Remote sensors on a satellite**
  - **Telescopes scanning the skies**
  - **Micro-arrays generating gene expression data**
  - **Scientific simulations generating terabytes of data**
- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - **In classifying and segmenting data**
  - **In Hypothesis Formation**

Origins of Data Mining

- Draws ideas from machine learning / AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



Origins Data Mining: Confluence of Multiple Disciplines



Data Mining 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.**

Data Mining Tasks...

- Predictive
  - **Classification**
  - **Regression**
  - **Deviation detection**
- Descriptive
  - **Clustering**
  - **Association Rule Discovery**
  - **Sequential Pattern Discovery**

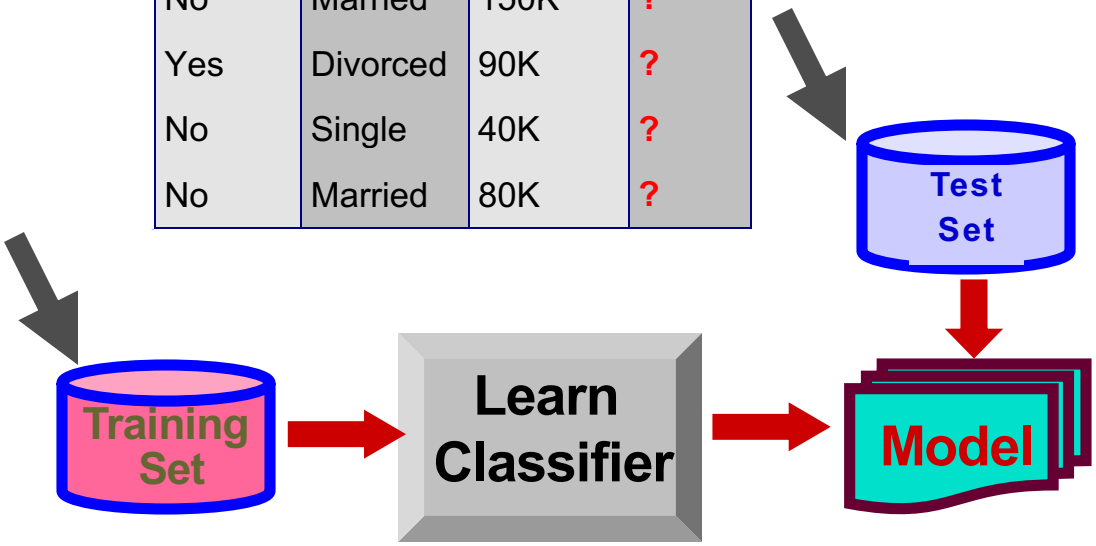
Classification: Definition

- Input: Given a collection of records (*training set*)
  - **Each record contains a set of *attributes*, one of the attributes is the *class*.**
- Output: Find a *model* for class attribute as a function of the values of other attributes.
- Goal: Use model to predict/assign the class for previously unseen records as accurately as possible.
  - **A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.**

Classification Example

Tid	Refund (categorical)	Marital Status (categorical)	Taxable Income (continuous)	Cheat (class)
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

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?



Classification: Applications

- Direct marketing
  - **Predict customers who will most likely buy a new product besed on their demographic, lifestyle and previous buying behavior**
- Spam detection
  - **Categorize email messages as spam or non-spam based on message header and content**
- Functional classification of proteins
  - **Assign sequences of unknown proteins to their respective functional classes**
- Galaxy classification
  - **Classify galaxy based on their image features**
- Automated target recognition
  - **Identify target objects (enemy tanks, trucks, etc) based on signals gathered from sensor arrays**

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

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

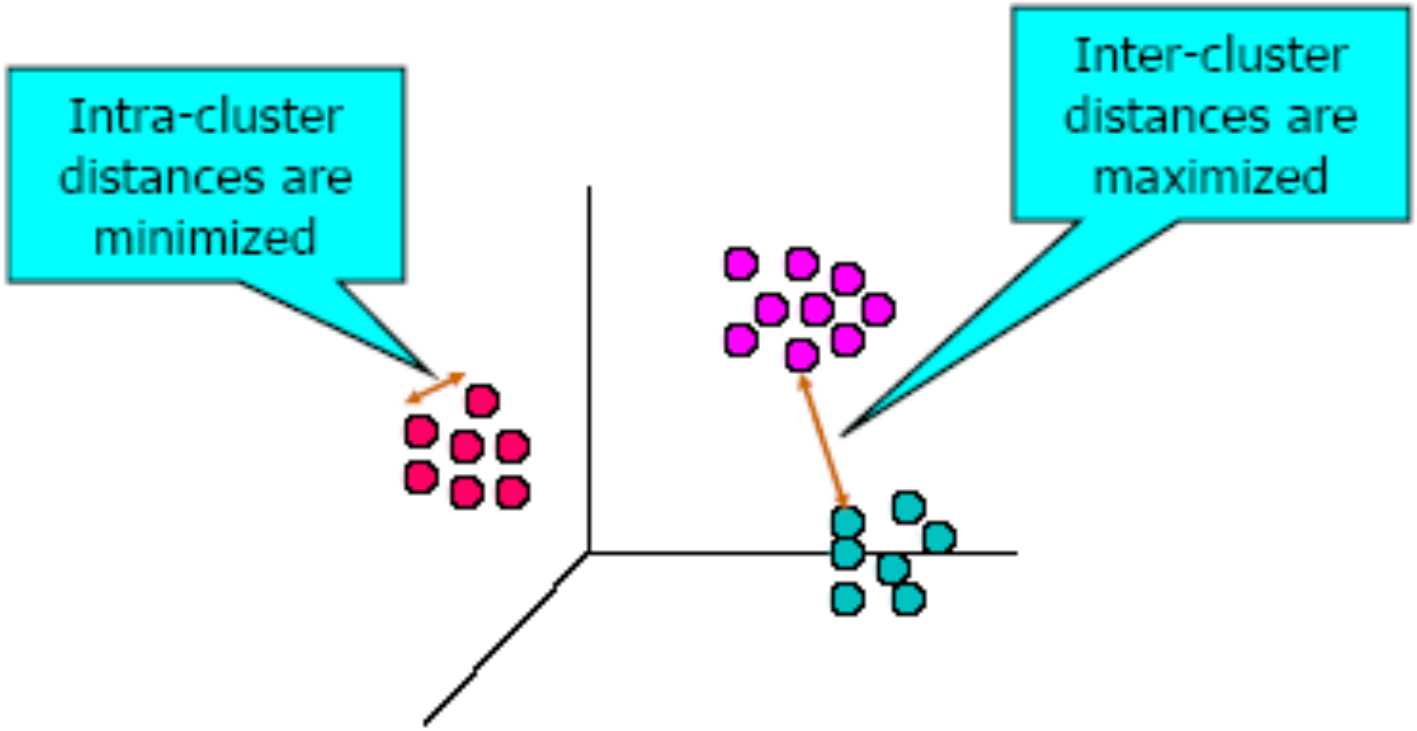
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - **Data points in one cluster are more similar to one another.**
  - **Data points in separate clusters are less similar to one another.**
- Similarity Measures:
  - **Euclidean Distance if attributes are continuous.**
  - **Other Problem-specific Measures.**

Clustering: Definition

- Input:
  - **A set of data points**
    - Each data point has a set of attributes
  - **A distance/similarity measure between data points**
    - e.g Euclidean distance, cosine similarity, and edit distance
- Output:
  - **Partition the data points into separate groups (clusters)**
- Goal:
  - **Data points that belong to the same cluster are very similar to one another**
  - **Data points that belong to different clusters are less similar to one another**

Clustering: Illustration

- Euclidean Distance Based Clustering in 3-D space.



Clustering: Applications

- Market segmentation
  - **Subdivide customers based on their geographical and lifestyle related information**
- Document clustering
  - **Find groups of documents that are similar to each other based on the important terms appearing in them**
- Time series clustering
  - **Find groups of similar time series (e.g stock prices, ECG, seismic waves) based on their shapes**
- Sequence clustering
  - **Find group of sequences (e.g web or protein sequnces) with similar features**

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.

Document Clustering: Illustration

- Clustering Points: 3204 Articles of Los Angeles Times.
- Similarity Measure: How many words are common in these documents (after some word filtering).

Category	Total Articles	Correctly Placed
Financial	555	364
Foreign	341	260
National	273	36
Metro	943	746
Sports	738	573
Entertainment	354	278

Association Analysis: Definition

- Input:
  - **A collection of transactions**
    - Each transaction contains a set of items
- Output:
  - **Patterns that represent strong relationships present in the data**
- Goal:
  - **Patterns must have high support, i.e occur frequently in the data**
  - **Patterns should also satisfy other interestingness criteria (e.g confidence, lift, etc)**

Association Rule Discovery: Illustration

- Given a set of records each of which contain some number of items from a given collection;
  - **Produce dependency rules which will predict occurrence of an item based on occurrences of other items.**

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

Association Analysis: Applicaions

- Market-basket analysis
  - **Rules are used for sales promotion, shelf management and inventory management**
- Telecommunication alarm diagnosis
  - **Rules are used to find combination of alarms that occur together frequently in the same time period**
- Medical informatics
  - **Rules are used to find combination of patient symptoms and complaints associated with certain diseases**

Association Rule Discovery: Application 1

- Marketing and Sales Promotion:
  - Let the rule discovered be

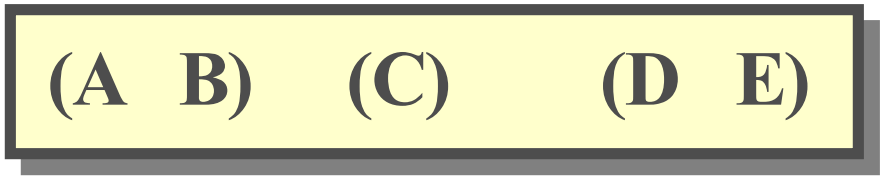
    {Bagels, … } --> {Potato Chips}
  - Potato Chips as consequent ➔Can be used to determine what should be done to boost its sales.
  - Bagels in the antecedent ➔ Can be used to see which products would be affected if the store discontinues selling bagels.
  - Bagels in antecedent *and* Potato chips in consequent ➔ Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

Association Rule Discovery: Application 2

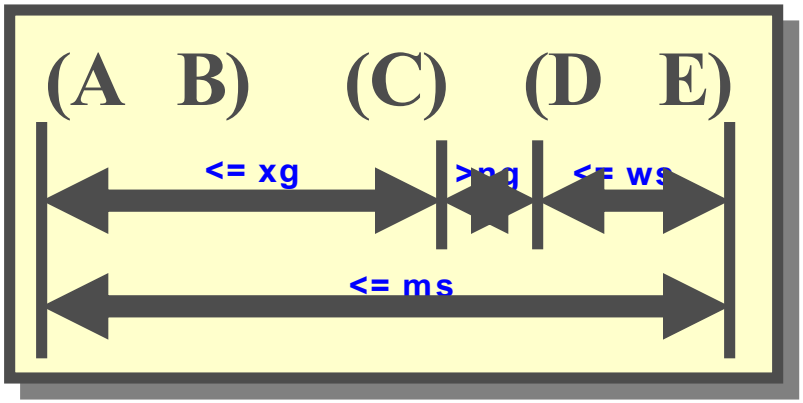
- Supermarket shelf management.
  - **Goal:** To identify items that are bought together by sufficiently many customers.
  - **Approach:** Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Sequential Pattern Discovery: Definition

- Input: a set of *objects*, with each object associated with its own *timeline of events*
- Output: find rules that predict strong sequential dependencies among different events.



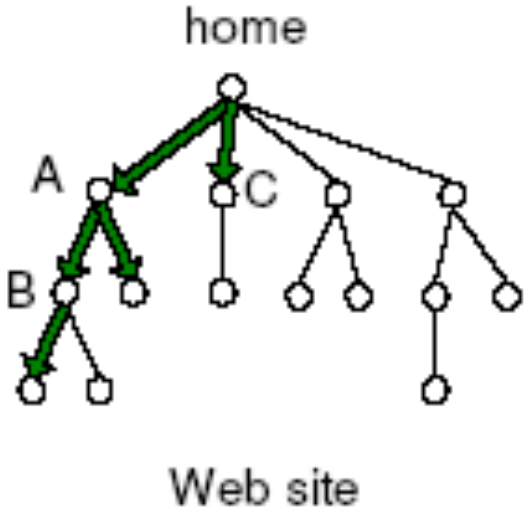
- Rules are formed by first discovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential Pattern Discovery: Applications

- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    (Intro_To_Visual_C) (C++_Primer)) (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    (Shoes) (Racket, Racketball)) (Sports_Jacket)

Sequential Pattern Discovery: Web Mining Example



Web site

User Id	Sequence of Pages Visited
0001	/home → /home/A → /home/A/B → /home/C
0002	/home → /home/D → /home/D/E
0003	/home → /home/A → /home/C

Pattern: /home → /home/A → /home/C

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.**

Regression: Definition

- Input:
  - **A collection of records (training set)**
    - Each record contains a set of attributes
    - One of the continuous-valued attributes is designated as target variable
- Output:
  - **A model for the target variable as a function of their attributes**
- Goal:
  - **Use the model to predict the value of the target variable, assuming a linear or nonlinear model of dependency**

Regression: Applications

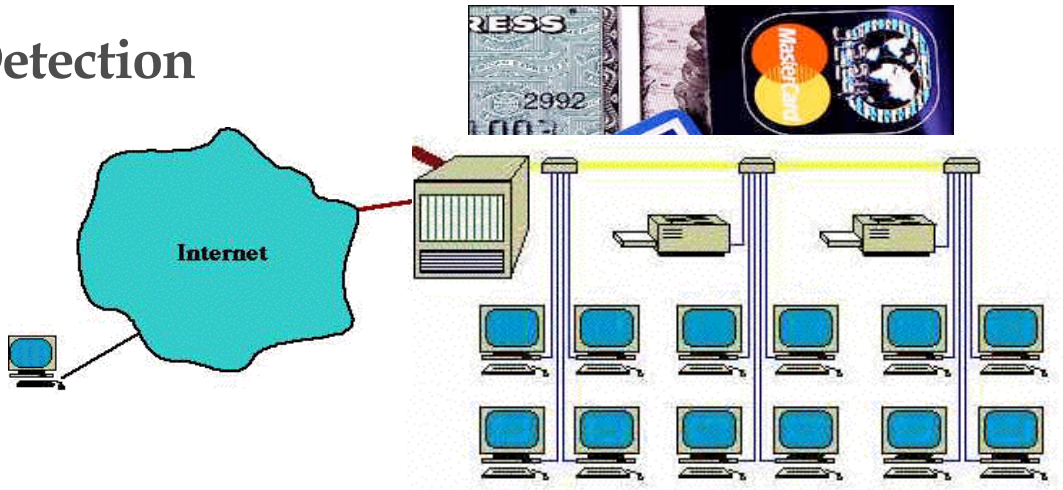
- Marketing
  - **Predicting sales amounts of new product based on advertising expenditure**
- Earth science
  - **Predicting wind velocities as a function of temperature, humidity, air pressure, etc**
- Finance
  - **Time series prediction of stock market indices**
- Agriculture
  - **Predicting crop yield based on soil fertility and wather information**
- Socio-economy
  - **Predicting electricity consumption in single family based on outdoor temperatures**

Anomaly Detection: Definition

- Input:
  - **A set of records**
    - Most of the records are assumed to be "normal"
- Output:
  - **A set of anomalous of records**
- Goal:
  - **High detection rate**
  - **Low false alarm rate**

Deviation/Anomaly Detection

- Detect significant deviations from normal behavior
- Applications:
  - **Credit Card Fraud Detection**
  - **Network Intrusion Detection**



Challenges of Data Mining

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