Machine Learning

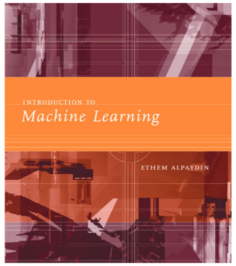
Berlin Chen 2006

References:

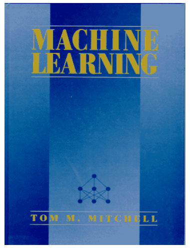
1. E. Alpaydin, *Introduction to Machine Learning*, Chapter 1
2. Kantard, *Data Mining: Concepts, Models, Methods and Algorithms*, Chapter 1
3. Mitchell, *Machine Learning*, Chapter 1
4. Han and Kamber, *Data Mining: Concepts and Techniques*, Chapter 1

Main Textbooks

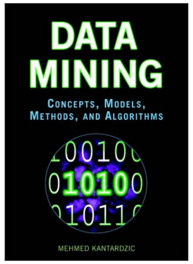
1. Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2004
   http://www.cmpe.boun.edu.tr/~ethem/i2ml/



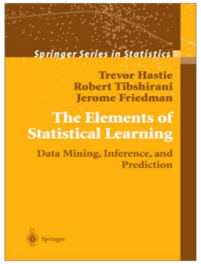
2. Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997



3. Mehmed M. Kantard, *Data Mining: Concepts, Models, Methods and Algorithms*, Wiley-IEEE Press, 2002



4. T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*, Springer-Verlag, 2001


Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Springer

Reference Textbooks

1. Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2001

2. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Prentice-Hall, 2003

3. Nils J. Nilsson, *Artificial Intelligence: A New Synthesis, Morgan Kaufmann*, 1998

4. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval, Addison Wesley Longman*, 1999

5. I. H. Witten and E. Frank, *Data Mining*, Morgan Kaufmann, 2000.

Goals

- Know the basic concepts and fundamentals of machine learning and data mining

- Theoretically understand a variety of models and algorithms that can be employed in the fields such as data mining, information retrieval, pattern recognition, speech processing, image processing, …

Machine Learning

- Address the question of how to build computer programs that improve their performance at some task through experience
  - Learning is a process → algorithm/program
- Can be viewed as searching a very large space of possible hypotheses to determine one that best fits the observed data and any prior knowledge held by the learner, and also can correctly generalize to unseen examples
  - Search strategies
  - Underlying structures of the hypothesis space

Different learning methods searching different hypothesis spaces

Machine Learning: Why?

- Recent progress in algorithms and theory
- Growing flood of online data
- Computational power is available
- Budding industry

Machine Learning: When ?

- Human expertise does not exist (navigating on Mars),
- Humans are unable to explain their expertise (speech recognition)
- Solution changes in time (routing on a computer network)
- Solution needs to be adapted to particular cases (user biometrics)

Machine Learning: Applications

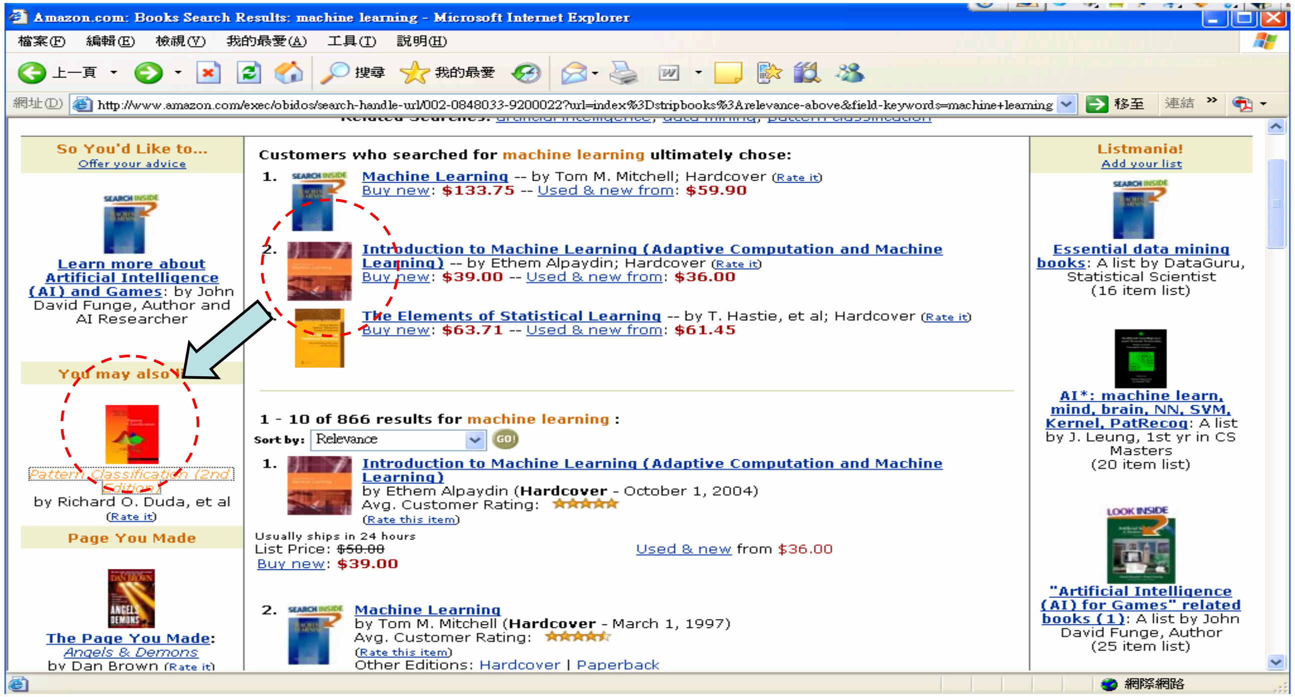
- Association
- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Reinforcement Learning

Associations

- Example: Basket Analysis
  - Customer transaction to consumer behavior
    - P (Y | X) probability that somebody who buys X also buys Y where X and Y are products/services
    - Examples:

      P(buying “chips” | buying “beer”) = ?

      P(buying “Pattern Classification”| buying “Machine Learning”) = ?



http://www.amazon.com/exec/obidos/search-handle-url/002-0848033-9200022?url=index%3Dstripbooks%3Arelevance-above&field-keywords=machine+learning

So You'd Like to...
Offer your advice

Learn more about Artificial Intelligence (AI) and Games: by John David Funge, Author and AI Researcher

You may also li

Pattern Classification (2nd Edition)
by Richard O. Duda, et al
(Rate it)

Page You Made

The Page You Made: *Angels & Demons*
by Dan Brown (Rate it)

Customers who searched for machine learning ultimately chose:

1. Machine Learning -- by Tom M. Mitchell; Hardcover (Rate it)
   Buy new: **$133.75** -- Used & new from: **$59.90**
2. Introduction to Machine Learning (Adaptive Computation and Machine Learning) -- by Ethem Alpaydin; Hardcover (Rate it)
   Buy new: **$39.00** -- Used & new from: **$36.00**
3. The Elements of Statistical Learning -- by T. Hastie, et al; Hardcover (Rate it)
   Buy new: **$63.71** -- Used & new from: **$61.45**

1 - 10 of 866 results for machine learning :

Sort by: Relevance GO!

1. Introduction to Machine Learning (Adaptive Computation and Machine Learning)
   by Ethem Alpaydin (**Hardcover** - October 1, 2004)
   Avg. Customer Rating: ★★★★★
   (Rate this item)

Usually ships in 24 hours
List Price: ~~$50.00~~
Buy new: **$39.00**
Used & new from $36.00

2. Machine Learning
   by Tom M. Mitchell (**Hardcover** - March 1, 1997)
   Avg. Customer Rating: ★★★★☆
   (Rate this item)
   Other Editions: Hardcover | Paperback

Listmania!
Add your list

Essential data mining books: A list by DataGuru, Statistical Scientist
(16 item list)

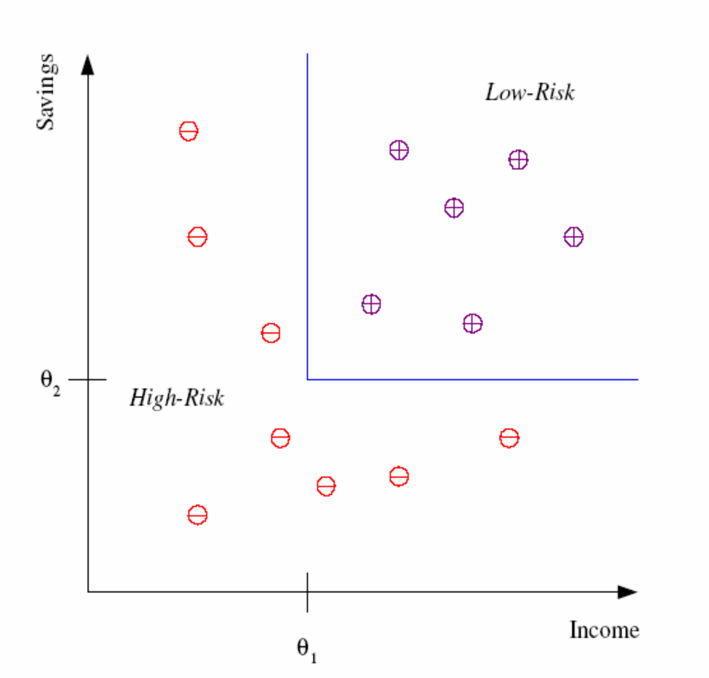
AI*: machine learn, mind, brain, NN, SVM, Kernel, PatRecog: A list by J. Leung, 1st yr in CS Masters
(20 item list)

"Artificial Intelligence (AI) for Games" related books (1): A list by John David Funge, Author
(25 item list)

Classification (1/2)

- Also known as Pattern Recognition
- Example 1: Credit Scoring
  - Differentiating between low-risk and high-risk customers from their income and savings
  - Discriminant:

IF income > θ 1 AND savings > θ 2 THEN low-risk ELSE high-risk



Classification (2/2)

- Face Recognition

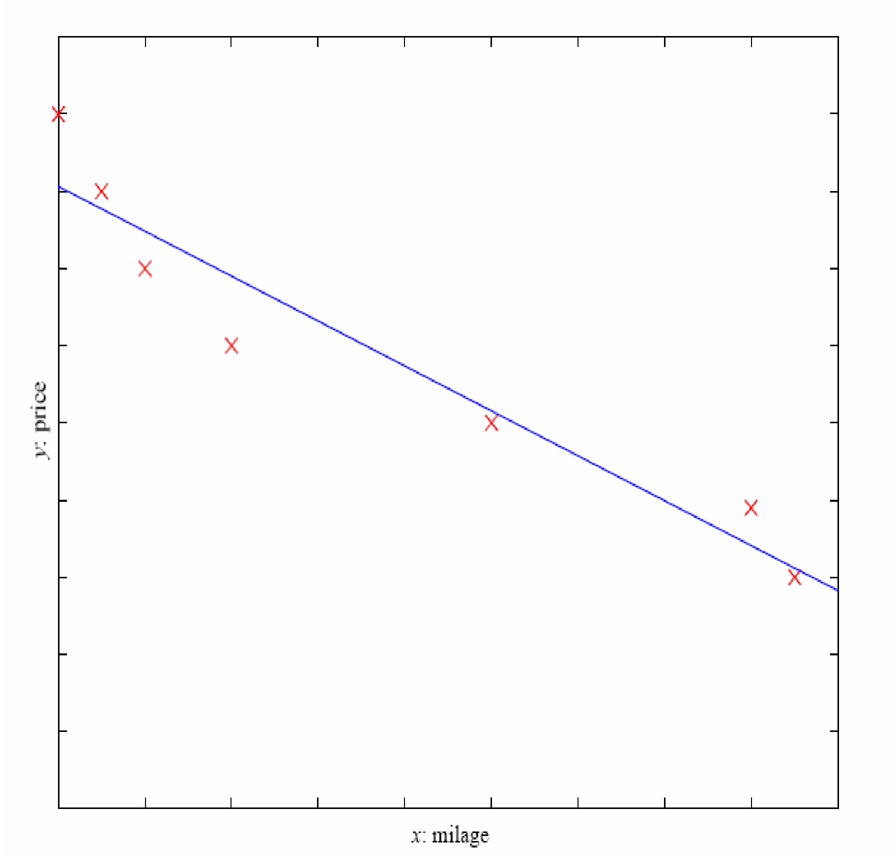
Training examples of a person

Test images

AT&T Laboratories, Cambridge UK
http://www.uk.research.att.com/facedatabase.html

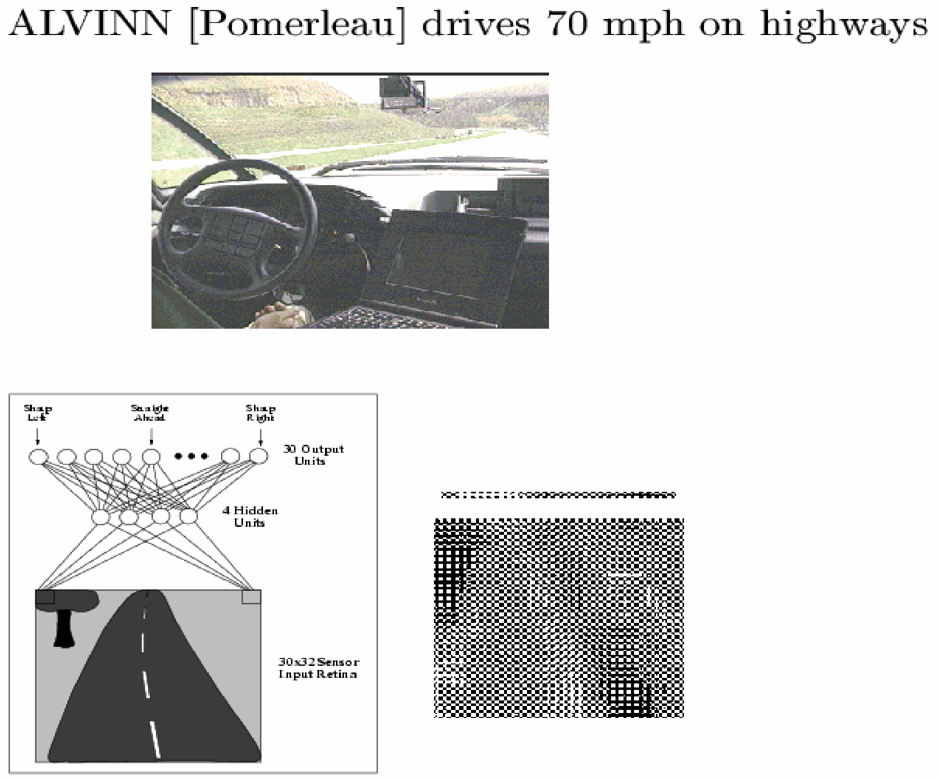
Regression (1/2)

- Example 1: Price of a Used Car
  - x : car attributes
  - y : price
  - y = g (x | θ)
  - g (•): model
  - Θ: parameters



Regression (2/2)

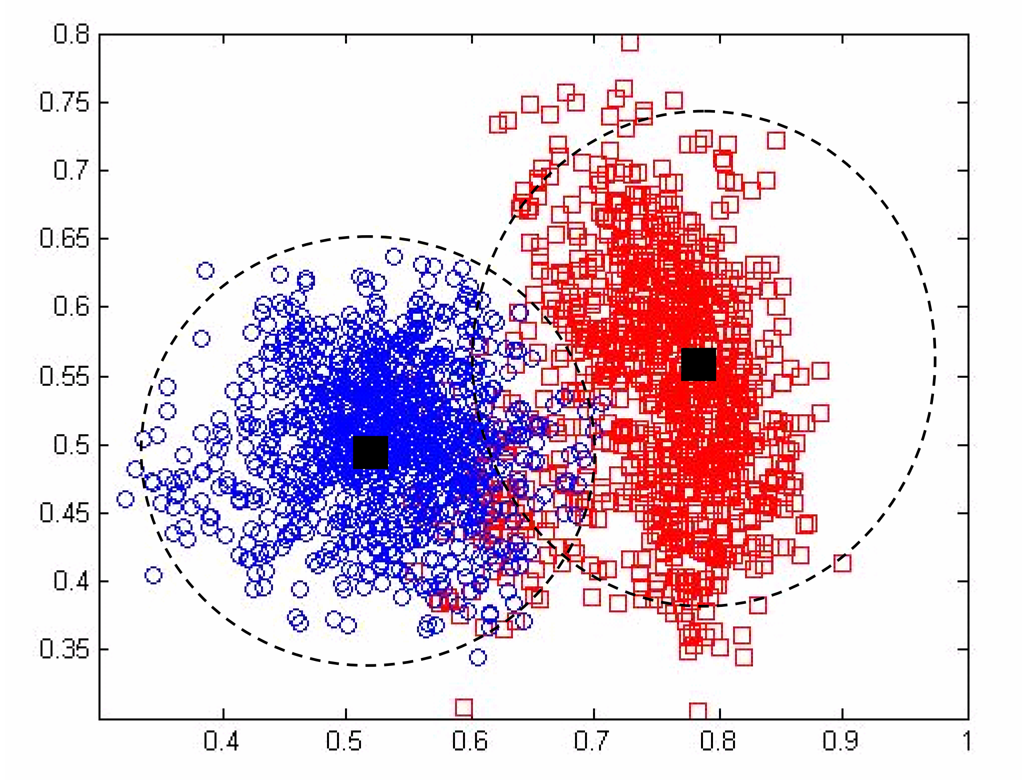
- Example 2: Navigating a Car
  - CMU Project: Driving Autonomous Vehicles (1989)
    - Autonomous Land Vehicle In a Neural Network (ALVINN)
  - Control the angle of the steering wheel, etc.



ALVINN [Pomerleau] drives 70 mph on highways

Clustering

- Clustering :A kind of unsupervised Learning
- Grouping similar instances



Reinforcement Learning

- Learning a policy: A sequence of outputs
- No supervised output but delayed reward
- Credit assignment problem
- Game playing
- Robot in a maze
- Multiple agents, partial observability

An action is good if it is part of a good game playing policy
(Is it on the way to win?)

Other Possible Applications

- Business Management
- Production Control
- Scientific/Medical Research
- …

Some Issues in Machine Learning

- What algorithms can approximate functions well (and when) ?
- How does number of training examples influence accuracy ? (Variance)
- How does complexity of hypothesis representation impact it ? (Bias)
- How can prior knowledge of learner help ?
- How does noisy data influence accuracy ?
- What are the theoretical limits of learnability ?
- How can systems alter their own representations ?

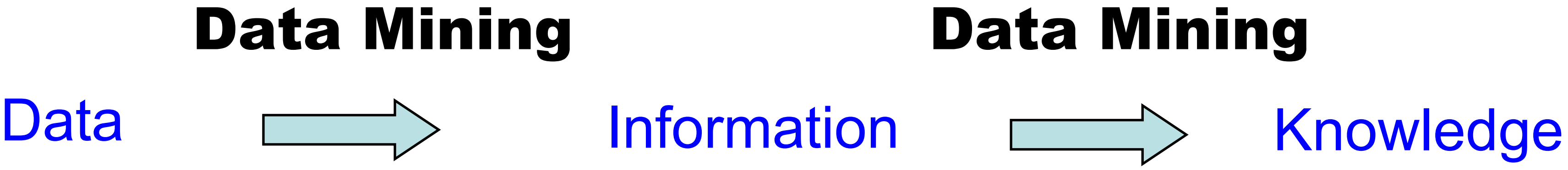
What is Data Mining ? (1/4)

- Also called *Knowledge Discovery in Databases* (KDD), *Information Extraction* (IE), *Knowledge Extraction* (KE) ..
- Emerged during the late 1980s, has made great strides during the 1990s, and continues to flourish into the new millennium
- Data-Mining? Information-Mining? Knowledge-Mining?
  - Cf. gold mining (but not rock or sand mining)
  - The term "Data Mining" is just a misnomer ?

What is Data Mining ? (2/4)

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories
  - Extract/Mine interesting information or knowledge (rules, regularities, patterns, constraints) from *huge amounts of data* stored in databases, data warehouse, and other information repositories
    - Explore hidden and nontrivial facts
  - "knowledge mining" from data

What is Data Mining ? (3/4)



Data Mining

Data Mining

Data

Information

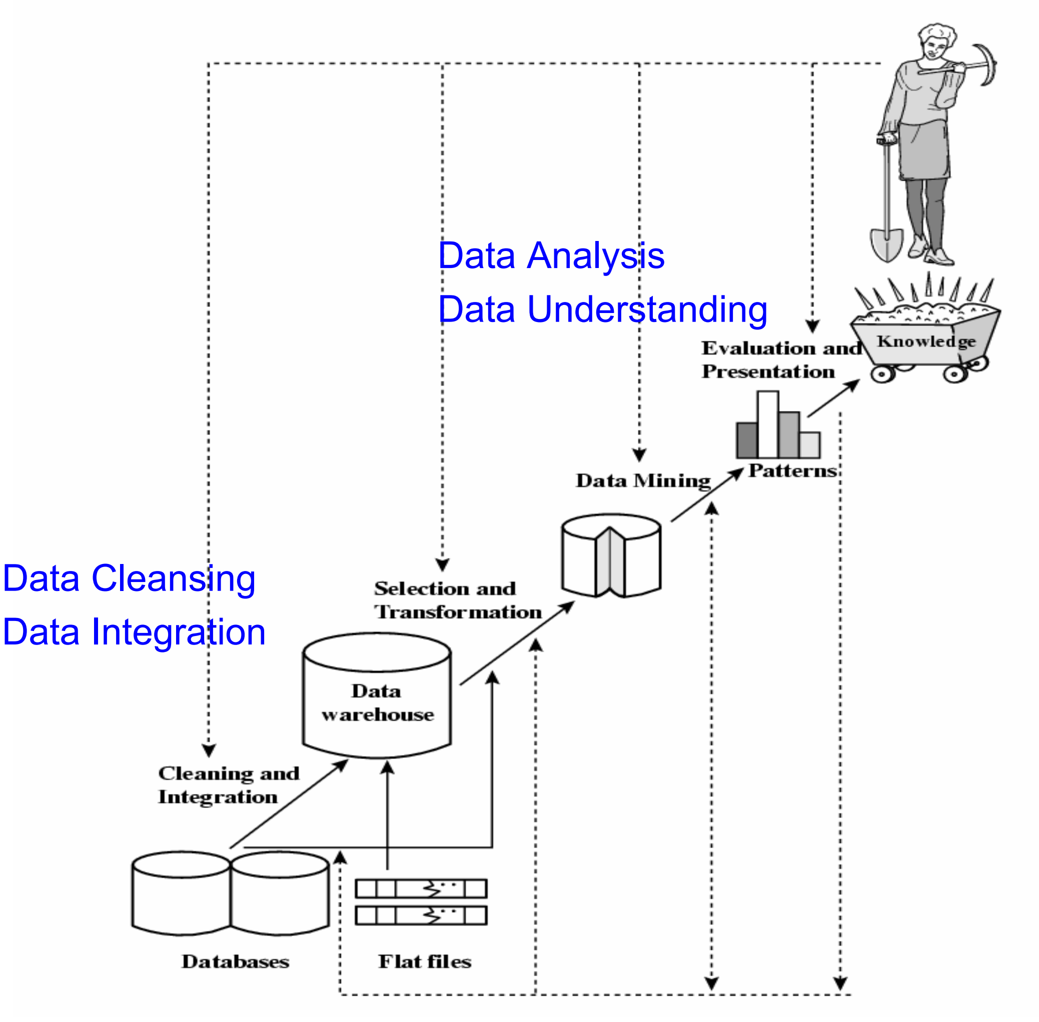
Knowledge

Data Tombs ?

Golden Nuggets ?

What is Data Mining ? (4/4)

- Data mining is an essential step in knowledge discovery



Data Analysis
Data Understanding

Data Cleansing
Data Integration

Categories of Data Mining

- Predictive Data Mining
  - Produce the model of the system described by the given data set
  - I.e., perform inference on the current data to make predictions
    - Classification
    - Regression

- Descriptive Data Mining
  - Produce new, nontrivial information (uncover patterns and relationships) based on the available data set
  - Namely, characterize the general properties of the data
    - Clustering
    - Summarization, or Concept/Class Interpretation
    - Dependency/Association Modeling $age(X, \text{"20...29"}) \wedge Income(X, \text{"20K...29K"}) \Rightarrow Buy(X, \text{"}CD\ Player\text{"})$
    - Change and Deviation Detection evolution, outlier detection

Multi-Dimensional View of Data Mining

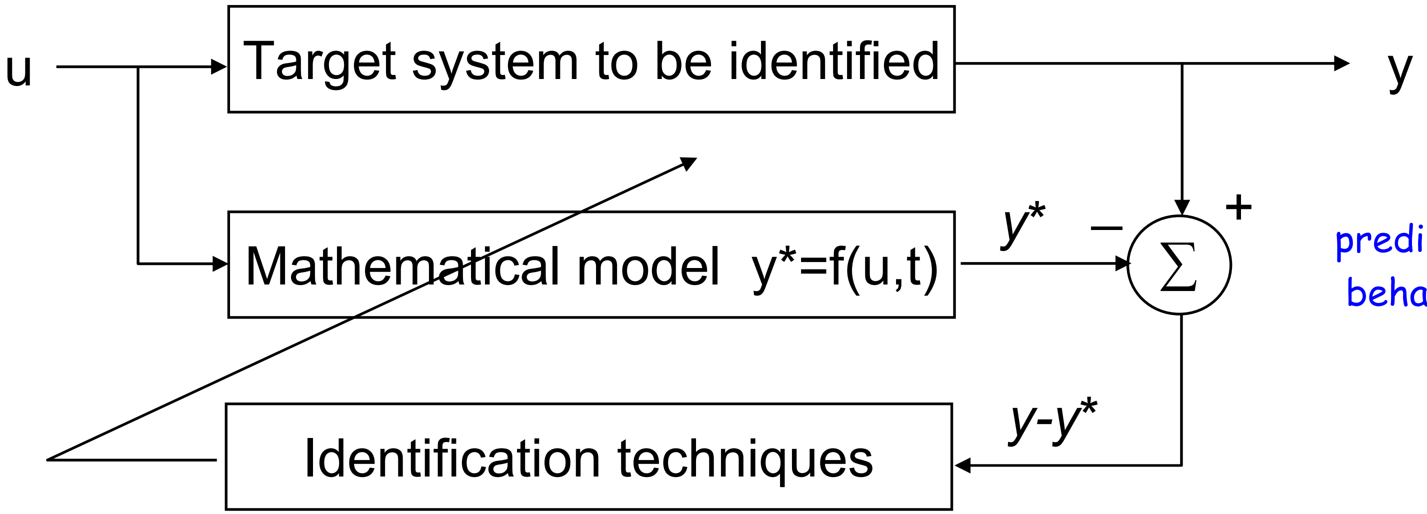
- Databases to be mined
  - Relational, transactional, object-oriented, object-relational, active, spatial, time-series, text, multi-media, heterogeneous, legacy, WWW, etc.
- Knowledge to be mined
  - Characterization, discrimination, association, classification, clustering, trend, deviation and outlier analysis, etc.
  - Granularity: mining at multiple levels of abstraction
- Techniques utilized
  - Machine learning, statistics, visualization, neural network, database-oriented, data warehouse (OLAP), etc.
- Applications adapted
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Roots of Data Mining (1/2)

- Statistics, Mathematics
  - Models
- Machine Learning
  - Algorithms
- Control theory
  - System identification

Roots of Data Mining (2/2)

- System Identification (an iterative process)
  - Structure Identification
  - Parameter Identification


u
Target system to be identified
y
Mathematical model y*=f(u,t)
y*
–
+
Σ
Identification techniques
y-y*

predict system's behaviors

Linear ? Nonlinear ?

Phases of Data Mining (1/7)

1. State the Problem and Formulate the Hypothesis
   - The problem statement should be established based on domain-specific knowledge and experience

   - But application studies tend to focus on the data-mining technique at the expanse of a clear problem statement

   - Cooperation between data-mining expertise and application expertise

experience and treatment

Phases of Data Mining (2/7)

2. Collect the Data

- Two possible approaches
  - Designed experiment
    - Data generation process is under control of an expert
  - Observational approach (random data generation)
    - The expert can not influence the data generation process

- A prior knowledge can be very useful for modeling and final interpretation of results

- Data respective for estimating a model and testing should come from the same, unknown, sampling distribution

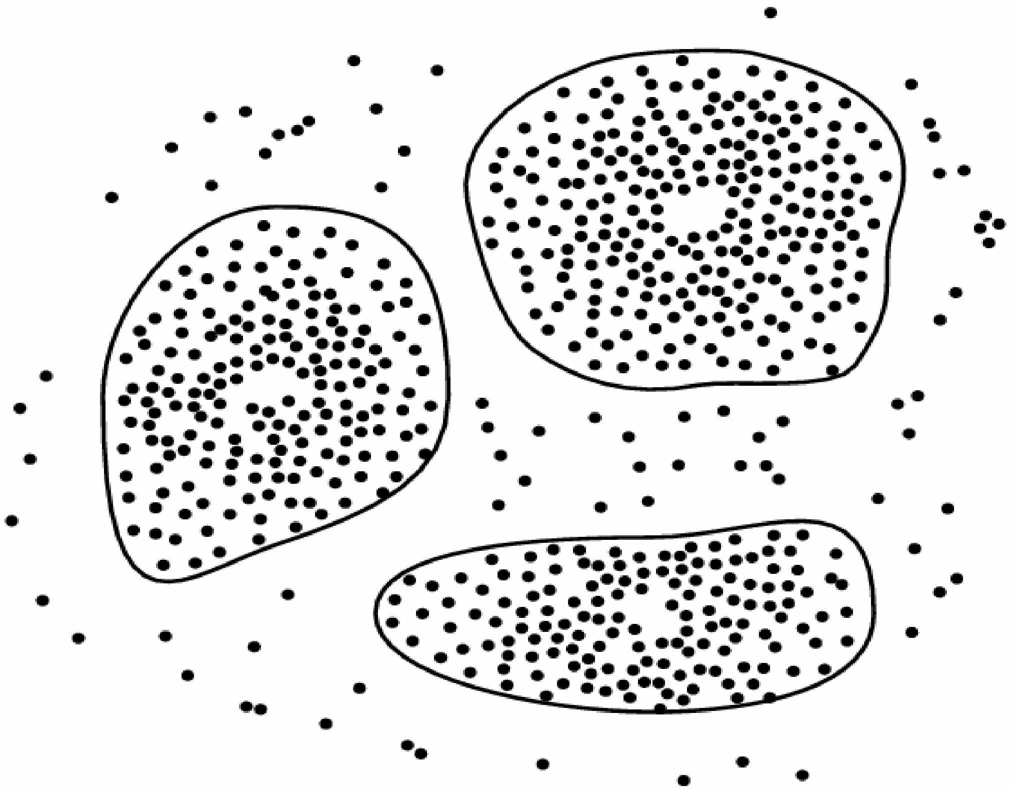
Phases of Data Mining (3/7)

3. Preprocessing the Data

- Two tasks involved
    - Outlier detection (and removal)
        - Outliers are unusual data values that are not consistent with most observations which can seriously affect modeling accuracy
        - Two strategies for dealing with outliers
            - » Removal of outliers
            - » Robust modeling methods
    - Scaling (normalization), encoding (discretization), and selecting features (dimensionality reduction)
- The prior knowledge of application domain should be considered in data-preprocessing steps

Phases of Data Mining (4/7)

- Clusters and Outliers

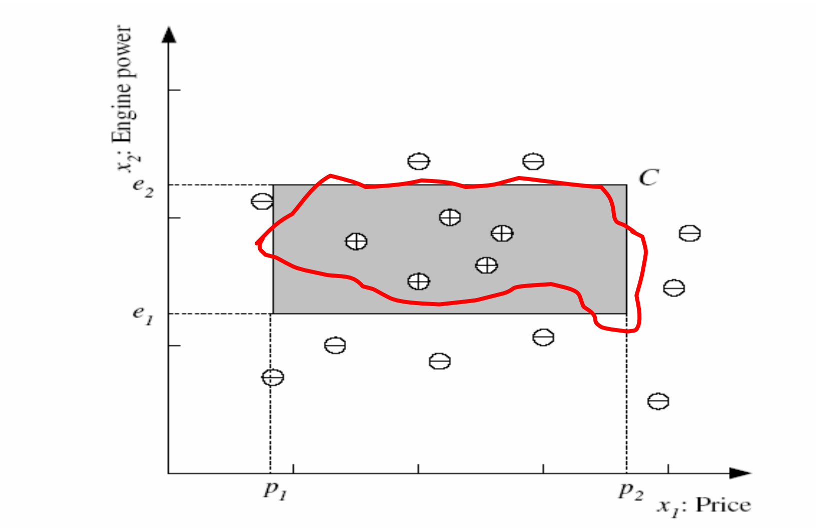


Phases of Data Mining (5/7)

4. Estimate the Model

- Select and implement the appropriate data-mining technique
  - The implementation is based on several models

- Use the technique to learn and discovery information from large volumes of data sets

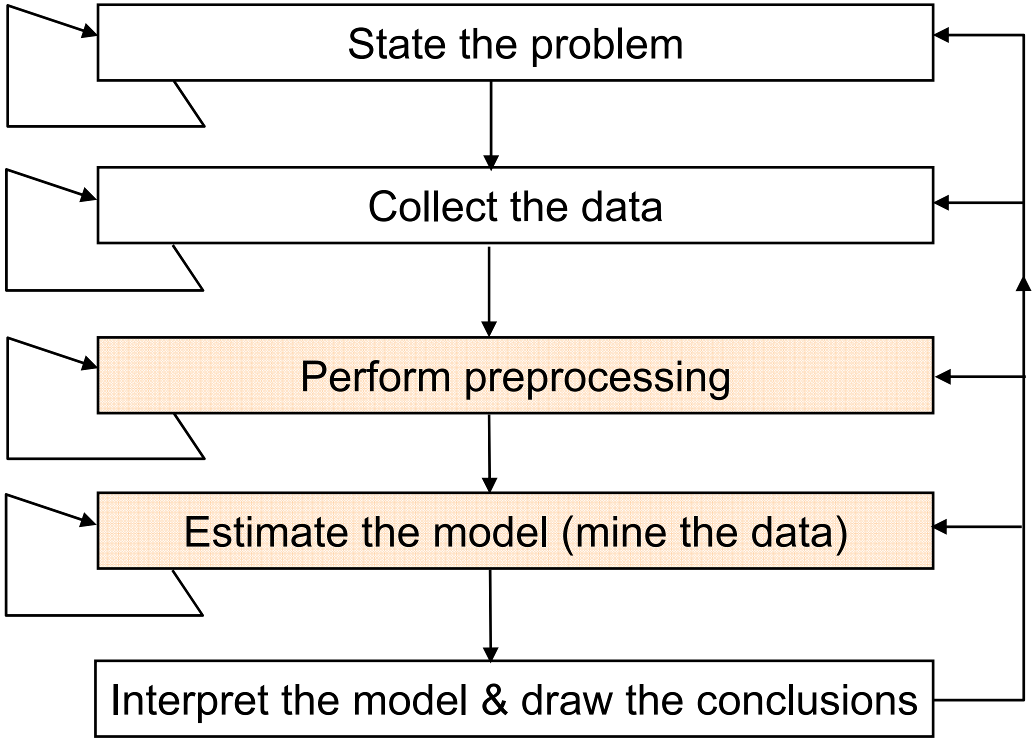


Phases of Data Mining (6/7)

5. Interpret the Model and Draw Conclusions
   - Data-mining models should help in decision making
   - Data-mining models thus should be interpretable
   - Tradeoff between accuracy of model and accuracy of model's interpretation

Phases of Data Mining (7/7)

- All phases and the entire data-mining process are highly iterative


State the problem
Collect the data
Perform preprocessing
Estimate the model (mine the data)
Interpret the model & draw the conclusions

Large Data Sets (1/2)

- An exponential growth in information sources and information-storage units

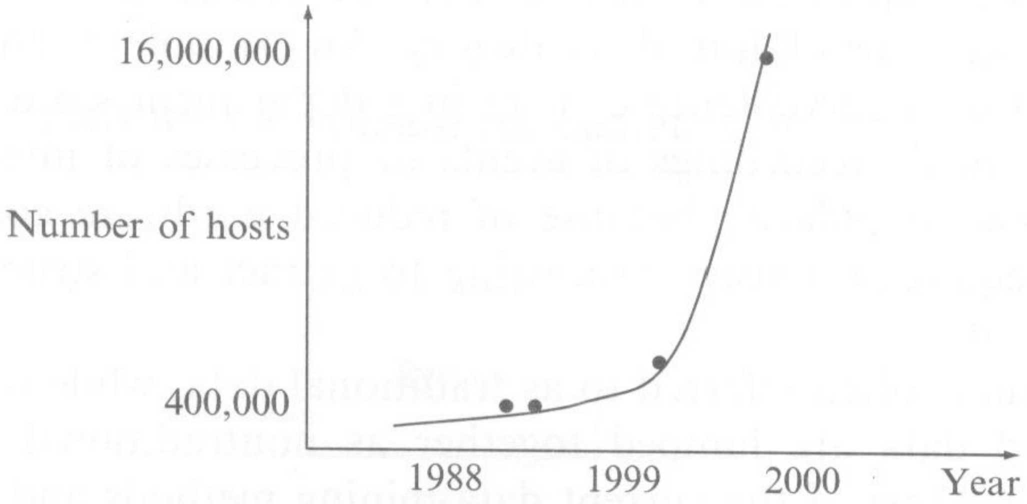


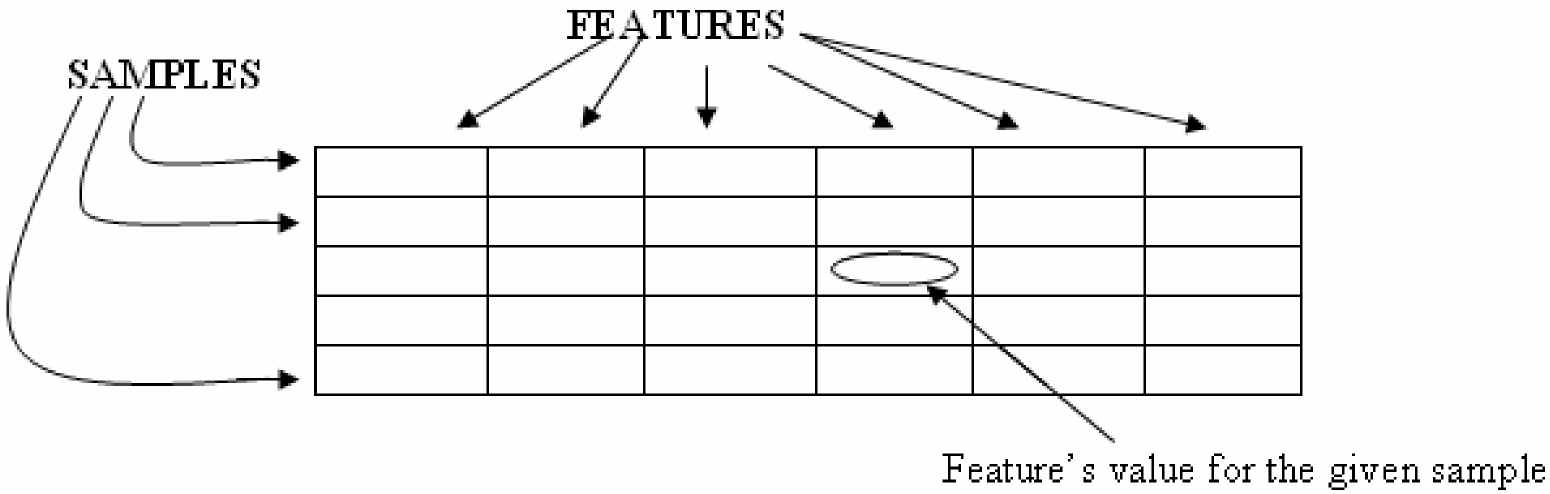
FIGURE 1.3 Growth of Internet hosts

  – The number of hosts are directly proportional to the amount of data stored on the Internet

Large Data Sets (2/2)

- Infer knowledge form huge volumes of raw datasets
  - Big data can lead to much stronger conclusions
  - A rapidly widening gap between data-collection and data-organization capabilities and the ability to analyze the data
  - Manual analysis and semiautomatic computer-based analysis can not deal with the large volumes of data sets
- Data as the sources for data mining can be classified into structured, semi-structured and unstructured data
  - Traditional data: structured data
  - Nontraditional data (multimedia): semi-structured and unstructured data

Structured Data

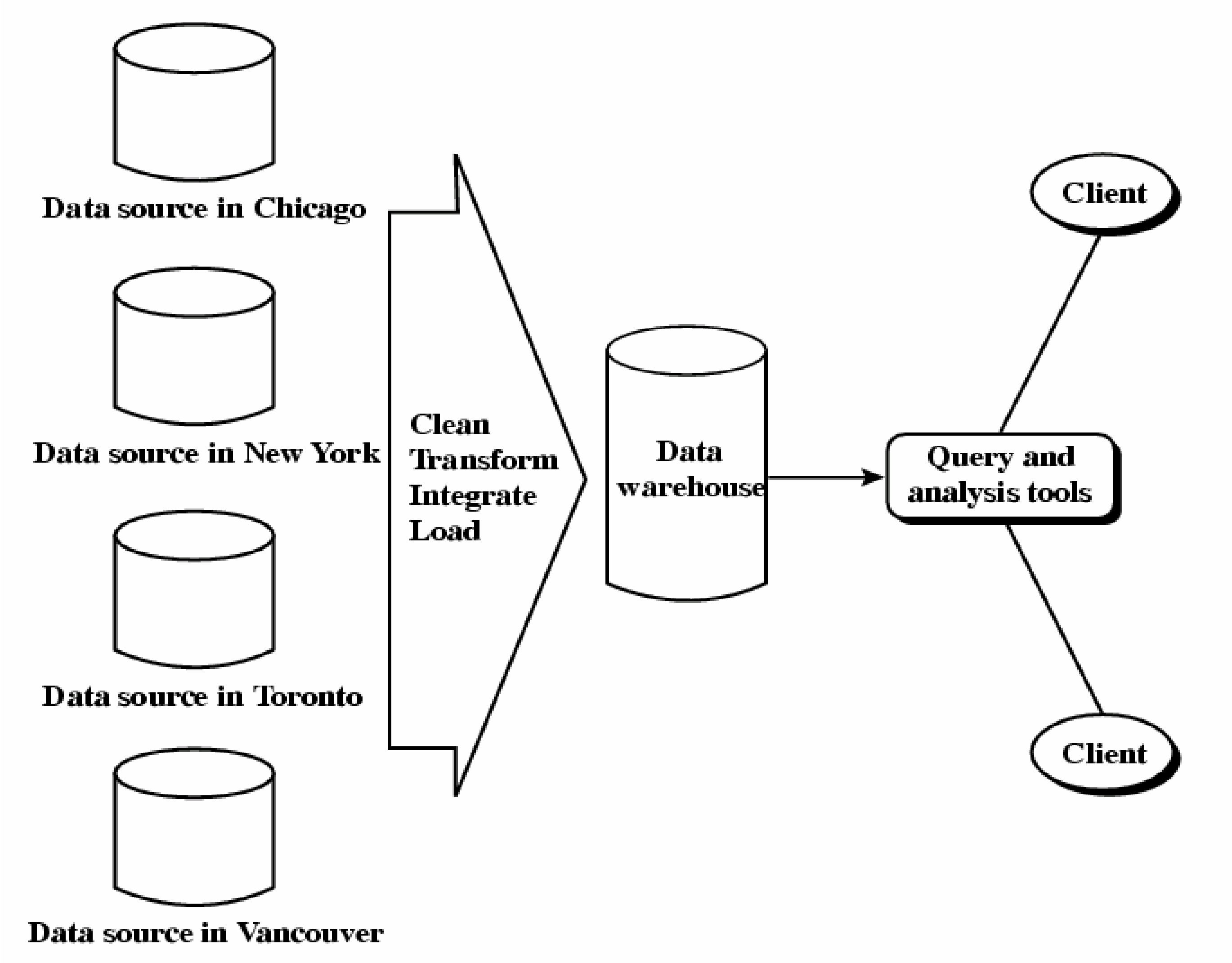


Features = Variables

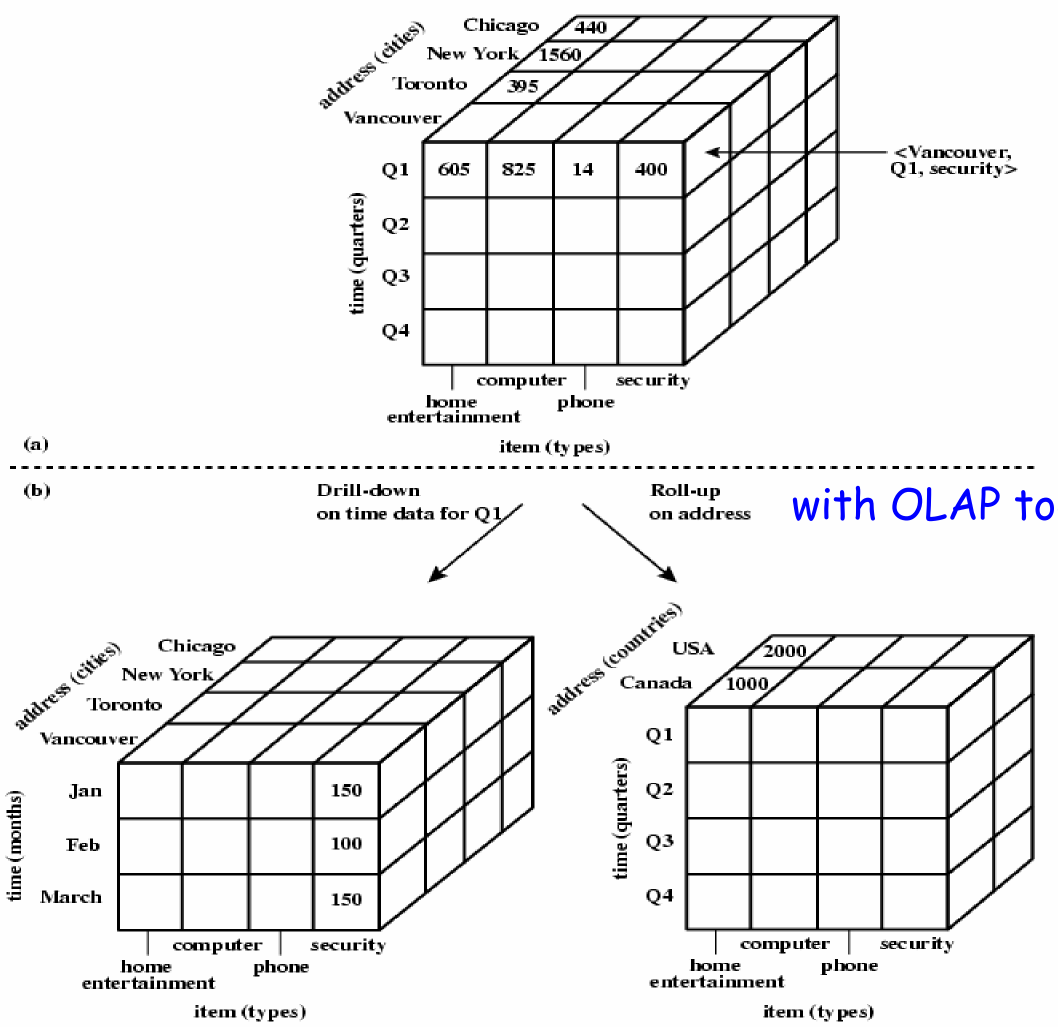
Data Warehouse (1/5)

- Definition
  - A collection of integrated, subject-oriented databases designed to support the decision-support functions (DSF), where each unit of data is relevant to some moment in time
    - Modeled as a multidimensional database structure
  - Or, a repository of multiple heterogeneous data sources, organized under a unified schema usually at a single site in order to facilitate management decision making
- That is, the sole of a data warehouse is to provide information for end users for decision support
- Cf. data mart
  - A department subset of a data warehouse

Data Warehouse (2/5)



Data Warehouse (3/5)



with OLAP tools

Data Warehouse (4/5)

- Applications
  - Data mining
    - Represent one of the major applications for data warehouse
    - Provide end-user with the capability to extract hidden, nontrivial (not obvious) information
      - Act as exploratory queries
  - Structured query languages (SQL)
    - A standard database language
    - Used when we know exactly what we are looking for and we can describe it formally
  - Online Analytical Processing (OLAP)
    - Do not learn from data, nor create new knowledge
    - Let users analyze data by providing multiple views of the data

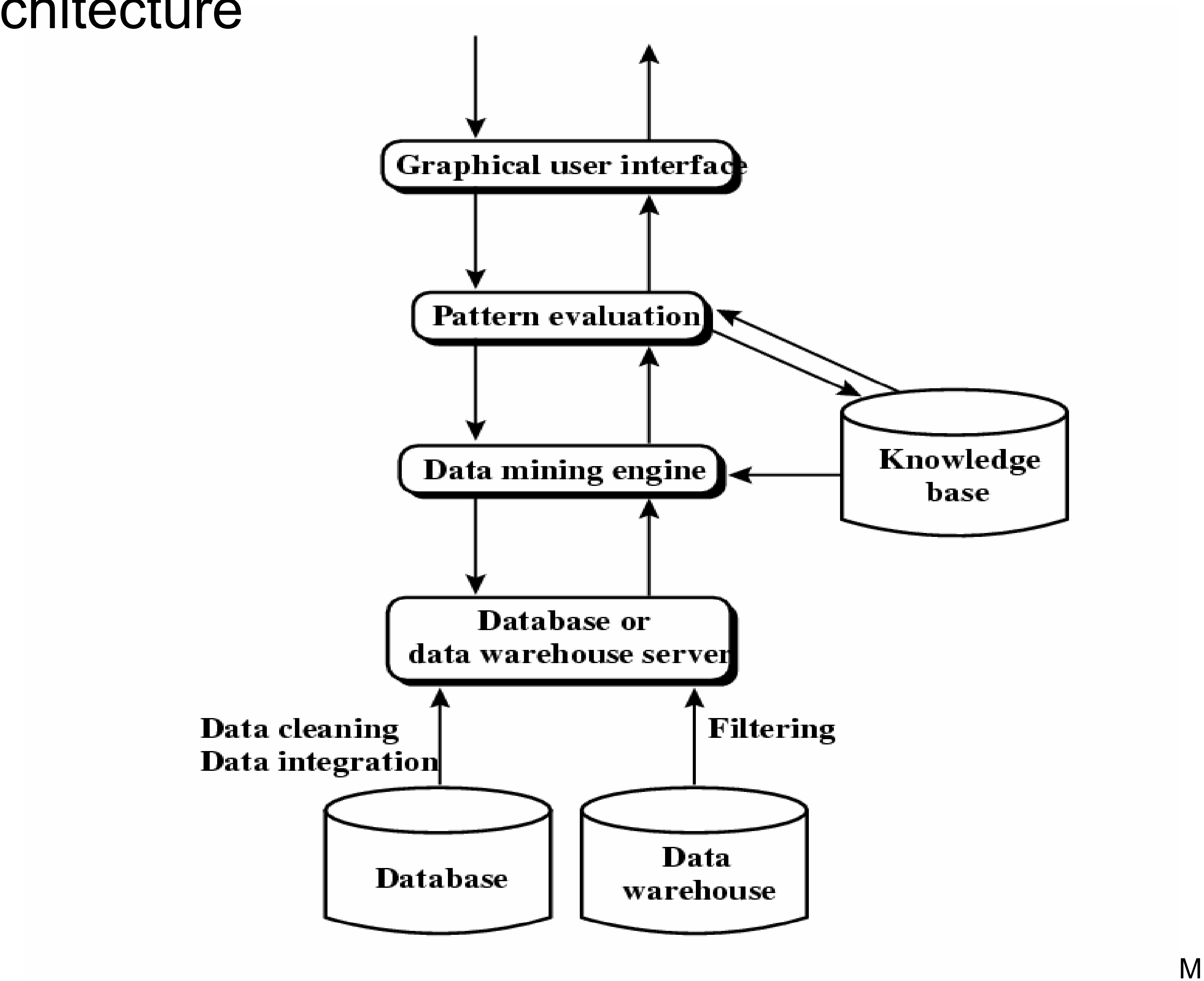
Data Warehouse (5/5)

- Classification of data stored in a data warehouse
  - Old detail data
  - Current (New) detail data
  - Lightly summarized data
  - Highly summarized data
  - Metadata (the data directory or guide)

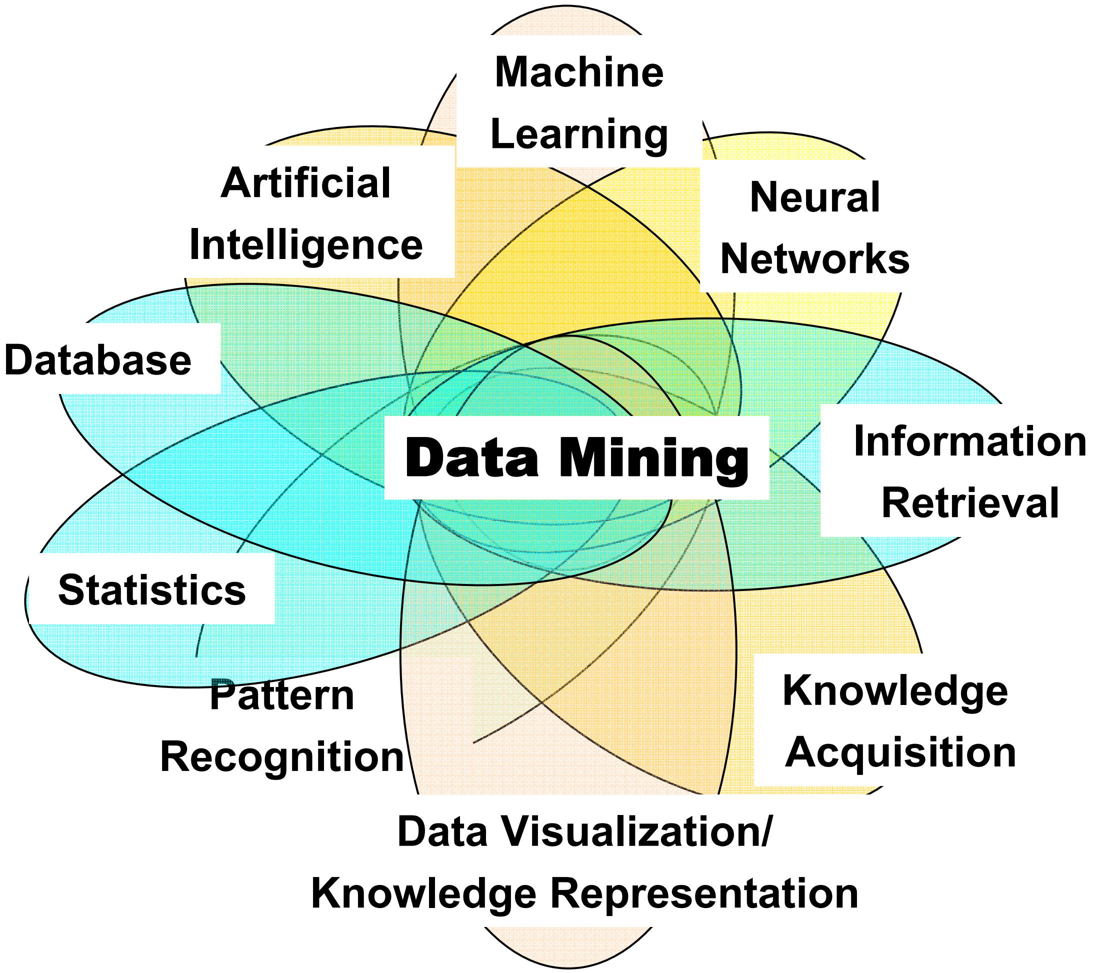
- Fundamental types of data transformation
  - Simple transformations (encoding/decoding)
  - Cleansing and scrubbing
  - Integration
  - Aggregation and summarization

A Typical Data Mining System

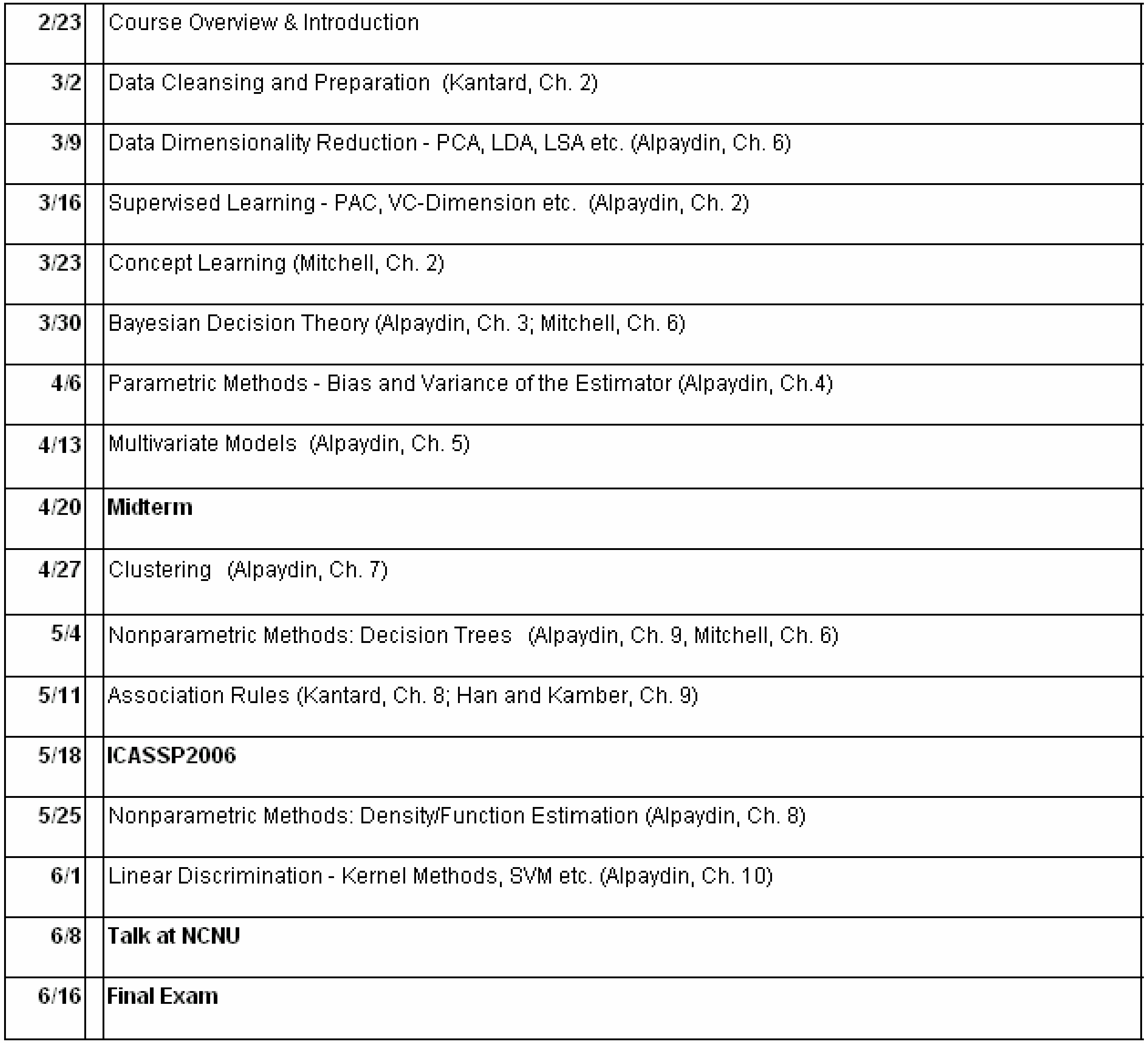
- Architecture



Confluence of Multiple Disciplines



Topic List and Tentative Schedule



2/23	Course Overview & Introduction
3/2	Data Cleansing and Preparation (Kantard, Ch. 2)
3/9	Data Dimensionality Reduction - PCA, LDA, LSA etc. (Alpaydin, Ch. 6)
3/16	Supervised Learning - PAC, VC-Dimension etc. (Alpaydin, Ch. 2)
3/23	Concept Learning (Mitchell, Ch. 2)
3/30	Bayesian Decision Theory (Alpaydin, Ch. 3; Mitchell, Ch. 6)
4/6	Parametric Methods - Bias and Variance of the Estimator (Alpaydin, Ch.4)
4/13	Multivariate Models (Alpaydin, Ch. 5)
4/20	**Midterm**
4/27	Clustering (Alpaydin, Ch. 7)
5/4	Nonparametric Methods: Decision Trees (Alpaydin, Ch. 9, Mitchell, Ch. 6)
5/11	Association Rules (Kantard, Ch. 8; Han and Kamber, Ch. 9)
5/18	**ICASSP2006**
5/25	Nonparametric Methods: Density/Function Estimation (Alpaydin, Ch. 8)
6/1	Linear Discrimination - Kernel Methods, SVM etc. (Alpaydin, Ch. 10)
6/8	**Talk at NCNU**
6/16	**Final Exam**

Resources: Journals

- Journal of Machine Learning Research
- Machine Learning
- Neural Computation
- Neural Networks
- IEEE Transactions on Neural Networks
- IEEE Transactions on Pattern Analysis and Machine Intelligence
- Annals of Statistics
- Journal of the American Statistical Association
- ...

Resources: Conferences

- International Conference on Machine Learning (ICML)
- European Conference on Machine Learning (ECML)
- Neural Information Processing Systems (NIPS)
- Uncertainty in Artificial Intelligence (UAI)
- Computational Learning Theory (COLT)
- International Joint Conference on Artificial Intelligence (IJCAI)
- International Conference on Neural Networks