Data Mining: Introduction

- **Introducing the course**
  - How the course is organized
  - How students are evaluated
  - Deadlines
- **Data Mining** [Chapt. 1 of course book]
  - What is it about?
  - The KDD process
  - Relations to other fields
  - Major techniques
  - Applications

Why Mine Data?

- Lots of data is being collected and warehoused

  - Web data, e-commerce
    - Some of the largest DBs on the web
      - Alexa (www.alexa.com)
      - Internet Archive (www.archive.org)
      - Google (over 4 billion pages!!)
  - Purchases at stores
  - Bank/Credit Card transactions
  - Europe's **V**ery **L**ong **B**aseline **I**nterferometry (VLBI)
    - 16 telescopes, each producing 1Gigabit/second of astronomical data
- Potentially valuable resource

Why Mine Data? Commercial Viewpoint

- Stored data grows very fast
  - Very little can be looked directly by a human
  - Raw data is useless: need techniques to automatically extract information from it
    - Data warehouses provides the enterprise with a memory
    - Data mining provides the enterprise with intelligence, e.g. transforms customer data into customer knowledge
      - Provide better and customized services
      - Competitive advantage

Information is crucial

- Example: *in vitro* fertilization
  - **Problem**: selection of embryos that will survive
  - **Data**: historical records of embryos and outcome
    - Embryos described by 60 features
- Example: customer attrition or churn
  - **Problem**: Find which customers are likely to terminate service
  - **Data**: customer information for the past N months

Example: Marketing

- Bank of America
- 13 million contact Bank of America's call center each month
- In past, customers listened to same marketing message
- Whether relevant to customer or not
- *"...we want to be as relevant as possible to each customer"*
- Customer profiles available to service representatives
- May suggest applicable products or services
- Data mining helps identify marketing approach based on customer's profile

Example: Basketball Strategies

- *Boston Celtics* listed employment position in 12/2003
- Statistics Intern: Work with Basketball Operations
- "Responsibilities include: ...data mining, etc."
- Use IBM's Advanced Scout data mining software
- Software includes NBA's game data
- Each game includes statistics such as shots, passes, points, rebounds, etc.
- Against *Chicago Bulls*, software discovered pattern coaching staff missed
- 16 of 29 NBA teams have turned to Advanced Scout to mine play-by-play data

A brief article at http://findarticles.com/p/articles/mi_m1571/is_n44_v13/ai_20017479/pg_1

Example: Security

- Data mining used in terrorism detection
  - ***T**otal **I**nformation **A**warness* program (**TIA**), US government, 2003
    - http://en.wikipedia.org/wiki/Information_Awareness_Office
  - The project raised some ethic concerns
    - Closed
    - Headlines in ComputerWorld, July, 2003

      "Senate Kills Data Mining Program"
    - http://www.pcworld.com/article/111626/senate_kills_data_mining_program.html

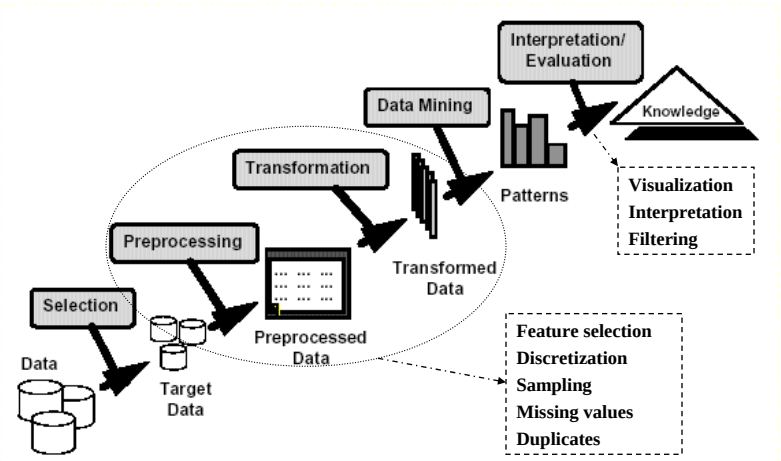
What is Data Mining?

- Extracting
  - **implicit,**
  - **previously unknown,**
  - **potentially useful**

  information from data
- Needed: programs that detect patterns and regularities in the data
- Strong patterns ⇒ good predictions
  - **Problem 1**: data may be garbled or missing
  - **Problem 2**: most patterns are not interesting
  - **Problem 3**: patterns may be inexact (or spurious)

KDD and Data Mining

Knowledge Discovery in Data



Fallacies of Data Mining

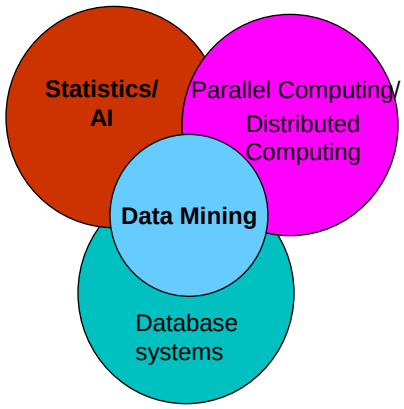
- Fallacy 1
  - Set of tools can be turned loose on data repositories
  - Finds answers to all business problems
- Reality 1
  - No automatic data mining tools solve problems
  - Rather, data mining is a process
- Fallacy 2
  - Data mining process is autonomous
  - Requires little oversight
- Reality 2
  - Requires significant intervention during every phase
  - Every phase may require several iterations

Fallacies of Data Mining

- Fallacy 3
  - Data mining quickly pays for itself
- Reality 3
  - Return rates vary
  - Costs with personnel, equipment/software, data preparation costs, etc.
- Fallacy 4
  - Data mining software easy to use
- Reality 4
  - Ease of use varies across projects
  - Analysts must combine subject matter knowledge with specific problem domain

Origins of Data Mining

- **Draws ideas from AI, visualization, statistics, data structures, and database systems**
- **Traditional Techniques may be unsuitable due to**
  - **Enormity of data**
  - **High dimensionality of data**
  - **Heterogeneous, distributed nature of data**



Statistics versus Data Mining (oversimplified)

- Standard statistical techniques assume that
  - Data sets are small, clean, contain only numerical data, and are random samples
- Statistics is most concerned with *primary data analysis*
  - Data are collected with a particular question in mind
  - Use the data to test hypothesis
    - E.g. *Average hourly wage of construction workers in Norrkoping is different from the national average.*

- Data mining techniques handle
  - Large data sets, with noise and missing values
  - Image data, audio data, text data, geographical data
  - Some regions may be sampled more heavily than others
    - E.g. Credit card data is likely to have much less fraudulent transactions
- Data mining is most concerned with *secondary data analysis*
  - To find the right hypothesis
    - *Which products can boost the sales of potato chips?*
    - *Characterize the workers in the different Swedish regions?*

Statistics versus Data Mining (oversimplified)

- Both are concerned with data analysis
- Statistical techniques can profitable be used in data mining
  - Sampling: questions about data can be answered with good accuracy with less than the entire data set
    - More powerful and computationally intensive algorithms can be used
  - Interval estimation and hypothesis testing
    - To predict the true performance of a model
    - To compare performance of different models
    - Verifying quality of extracted patterns

Potential Role of AI in Data Mining

- Natural language processing
  - Can be used in text mining
    - Parsing the text can be important to decide what an article is about
- Search strategies
  - Heuristic search (e.g. based on information gain) to build the best decision tree for a given data set
- Knowledge representation
  - Knowledge representation techniques can be used to express patterns
  - Use of ontologies
- See the paper *"Data Mining: An AI Perspective"*
  - Available from the course web page

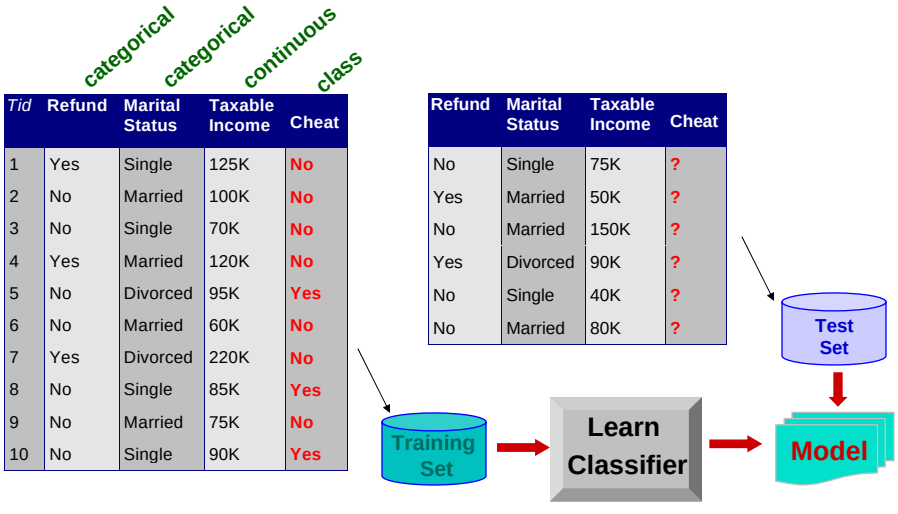
Data Mining Tasks

- Classification
- Estimation
- Database Segmentation (clustering)
- Associations Discovery
- Anomaly Detection

Classification

- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- **Goal**:
  - Find a *model* for class attribute as a function of the values of other attributes.
  - Previously unseen records should be assigned a class 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



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

From [Berry & Linoff] Data Mining Techniques, 1997

Classification: Application 2

- **Sky Survey Cataloging**
  - **Goal**: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory)
    - 3000 images with 23,040 x 23,040 pixels per image.
  - **Approach**:
    - Segment the image
    - Measure image attributes (features) - 40 of them per object
    - Model the class based on these features
    - **Success Story**: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

Estimation

- **Regression**: predicting a continuous value based on the values of other variables, assuming a *linear or nonlinear model* of dependency
- Greatly studied in statistics
- **Examples**:
  - Predicting sales amounts of new product based on advertising expenditure
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc

Predicting CPU performance

- Example: 209 different computer configurations

	Cycle time (ns)	Main memory (Kb)		Cache (Kb)	Channels		Performance
	MYCT	MMIN	MMAX	CACH	CHMIN	CHMAX	PRP
1	125	256	6000	256	16	128	198
2	29	8000	32000	32	8	32	269
…							
208	480	512	8000	32	0	0	67
209	480	1000	4000	0	0	0	45

- Linear regression function

```
PRP =  -55.9 + 0.0489 MYCT + 0.0153 MMIN + 0.0056 MMAX
       + 0.6410 CACH - 0.2700 CHMIN + 1.480 CHMAX
```

Database Segmentation

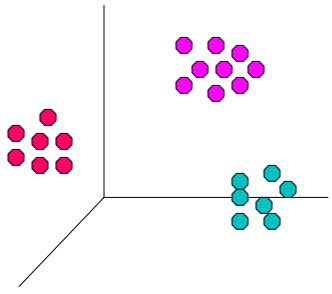
- **Clustering**: given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another
  - Data points in separate clusters are less similar to one another
  - Target variable (class) not specified
- Similarity Measures:
  - Euclidean Distance, if attributes are continuous
  - Other Problem-specific Measures

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Clustering: Application 1

- **Goal:** Create demographic profiles for different geographic areas according to zip code
- **Approach**:
  - Collect different attributes based on lifestyle related information.
  - Distinct "lifestyle" types (clusters) where identified

E.g. Clusters for Beverly Hills, CA 90210 include:

Cluster 1: Blue Blood Estates
Cluster 2: Bohemian Mix
Cluster 3: Winner's Circle
Cluster 4: Young Literati

  - Description of Cluster 01, *"...'old money' heirs that live in America's wealthiest suburbs...accustomed to privilege and live luxuriously..."*

Clustering: Other applications

- Subdivide a market into distinct subsets of customers where each can be targeted with a distinct marketing strategy
- Use for data summarization
- Clustering often used as preliminary step in data mining
  - Resulting clusters used as input to different DM technique

Associations Discovery

- 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}

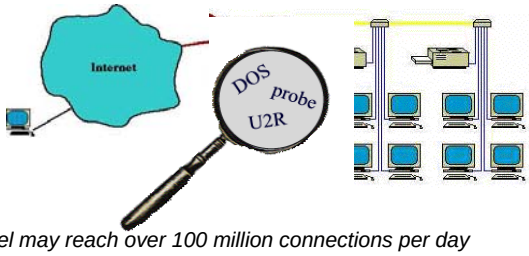
Association Rule Discovery: Application 1

- **Marketing and Sales Promotion**:
  - Let the rule discovered be
    {Bagels, Eggs, Bacon} --> {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!
- **Supermarket shelf management**.

Anomaly Detection

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

Clustering techniques can be used.



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

Summary

- Technology nowadays leads to data flood
  - Data mining helps to make sense of data
- Data mining has many applications
  - Who is likely to remain a loyal customer and who is likely to quit?
  - What products should be marketed for a particular customer?
  - Which patients with a heart condition should undergo the "expensive" medical procedure "X" to determine whether they are likely to have a heart failure?
  - Which web pages are likely to be visited next?

Summary (cont.)

- **K**nowledge **D**iscovery in **D**ata (KDD) process
- Data mining tasks
  - Classification, clustering, …
- An interesting web site
  - http://www.kdnuggets.com