Introduction

CSC 535/635

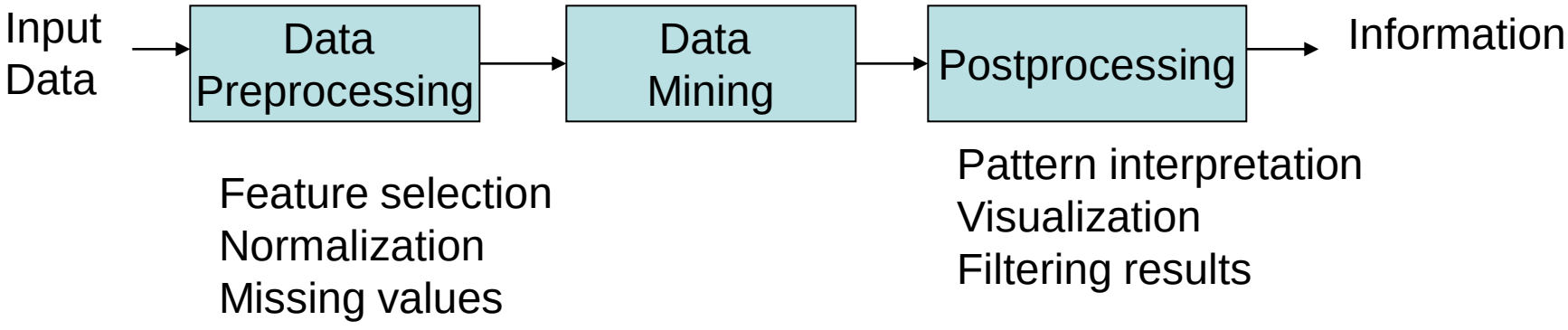
Other Names

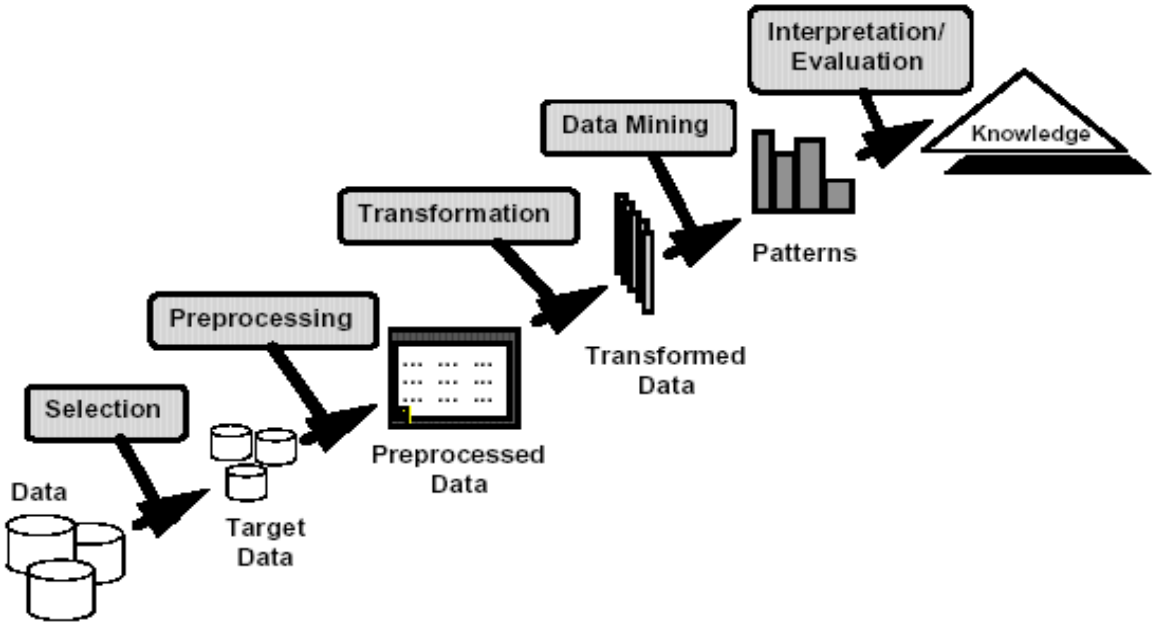
- Knowledge Discovery in Databases (KDD)
- Exploratory data analysis
- Deductive Learning

Data Mining vs. Knowledge Discovery

- Data mining is an important part of knowledge discovery in databases (KDD)
- KDD is the overall process of converting raw data into useful information

The Process of KDD





What is data mining?

- Data mining is
  - extraction of useful patterns from large volumes of data
- Data mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [Fayyad, Piatetsky-Shapiro, Smyth, 96]

What is (not) Data Mining?

- **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 US locations (O'Brien, O'Rurke, O'Reilly… in Boston area)
  - Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)

Why data mining?

- The data is abundant (big data)
- 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
- Data mining may help scientists

Why is data mining necessary? – Motivation for Mining Large Data Sets

- Make use of your data assets
- There is a big gap from stored data to knowledge; and the transition won't occur automatically
- Many interesting things that one wants to find cannot be found using traditional techniques
  - "Find people likely to buy my products"
  - "Who are likely to respond to my promotion"
  - "Which movies should be recommended to each customer?"
- Solution: Data Mining

Related fields

- Data mining is a multi-disciplinary field:
  - Machine learning, Pattern recognition, AI
  - Statistics
  - Databases
  - Information retrieval
  - Visualization
  - Natural language processing
  - etc.

Data mining (KDD) process

- Understand the application domain
- Identify data sources and select target data
- Pre-processing: cleaning, attribute selection, etc.
- Data mining to extract patterns or models
- Post-processing: identifying interesting or useful patterns/knowledge
- Incorporate patterns/knowledge in real world tasks

Data mining applications

- Marketing
- Customer profiling and retention
- Market segmentation
- Engineering: identify causes of problems in products
- Scientific data analysis, e.g., bioinformatics
- Fraud detection
- Text and web mining
- Any application that involves a large amount of data …

Challenges of Data Mining

- Scalability (huge amounts of data)
- Dimensionality (hundreds or thousands of attributes)
- Complex and Heterogeneous (different domains) Data
- Data Quality (missing values, uncertain data)
- Data Ownership and Distribution
- Privacy Preservation

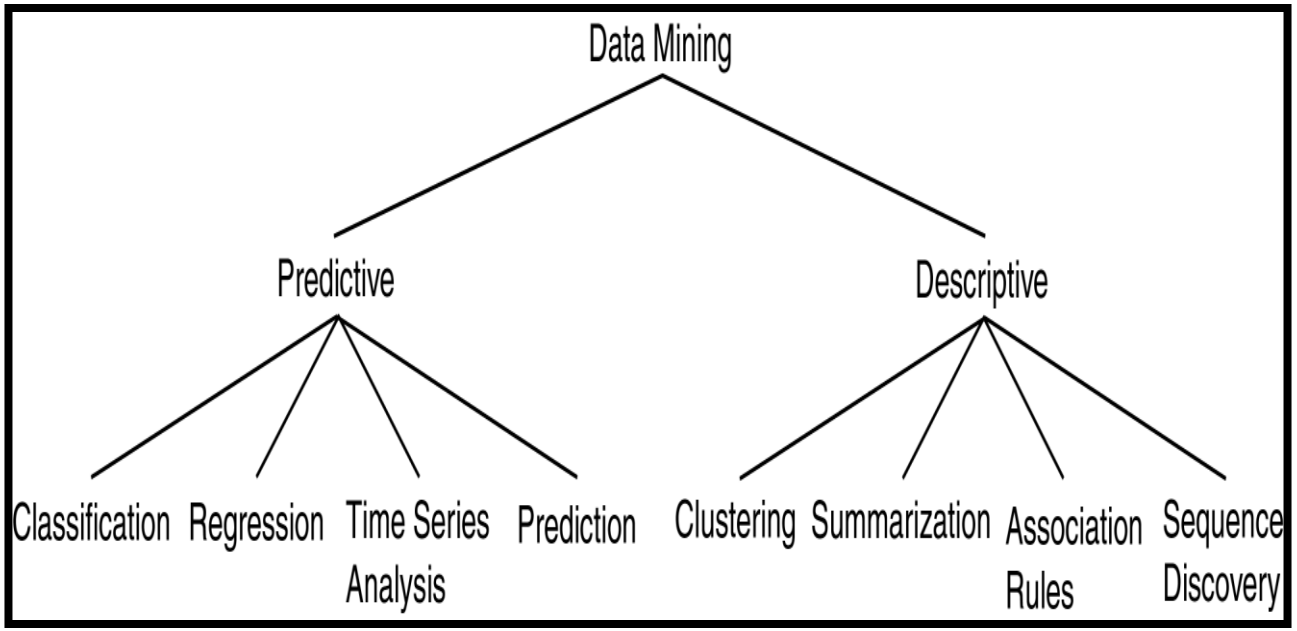
Data Mining Tasks

Data mining tasks are usually divided into two major categories

- Predictive Methods
  - Use some variables to predict unknown or future values of other variables. (classification)

- Descriptive Methods
  - Find human-interpretable patterns that describe the data. (clustering, association rules)

Data Mining Tasks



Classification: Definition

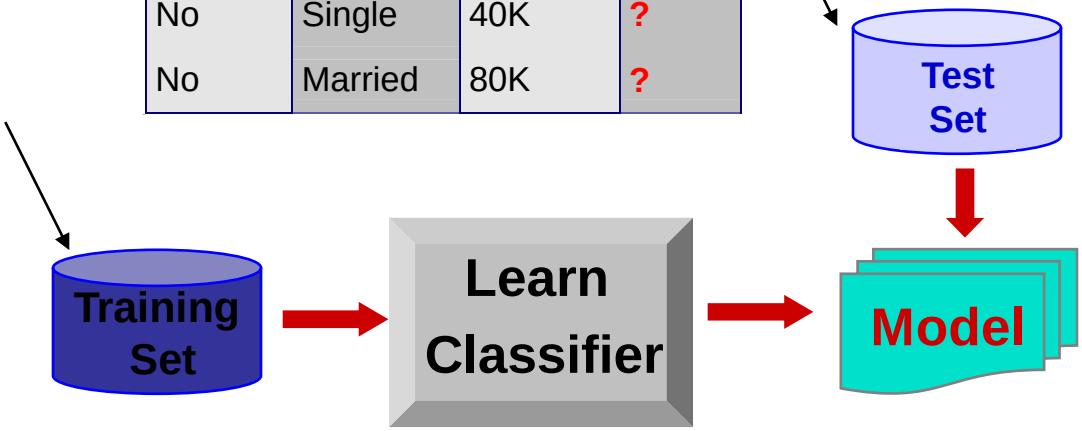
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.
  - A *test set* is used to determine the accuracy of the model.

Classification Example

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
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	?



Clustering Definition

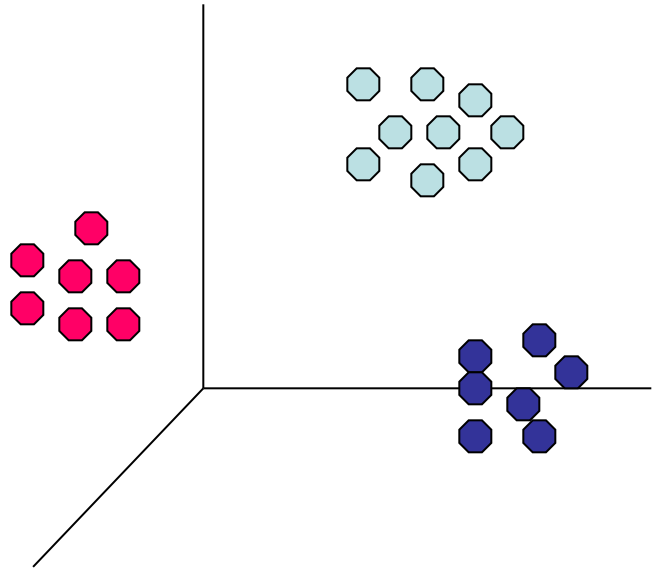
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean Distance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

Illustrating Clustering

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized



Association Rule Discovery: Definition

- 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

- 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!

Supervised and Unsupervised Learning

- Supervised: classification
- Unsupervised: clustering
- Semi-supervised

Data Mining Resources

- ACM SIGKDD: http://www.kdd.org
  - ACM Special Interest Group on Knowledge Discovery and Data Mining
- Kdnuggets: http://www.kdnuggets.com/
  - News and resources.
- Data mining related conferences
  - Data mining: KDD, ICDM, SDM, …
  - AI: ICML, NIPS, AAAI, IJCAI, ACL, …
  - Databases: SIGMOD, VLDB, ICDE, …
  - Web: WWW, WSDM, …
  - Information retrieval: SIGIR, CIKM, …