Chapter 3

Introduction To Data Mining

Based on CSC603.3:

Students should be able to preprocess the raw data and make it ready for the various data mining tasks

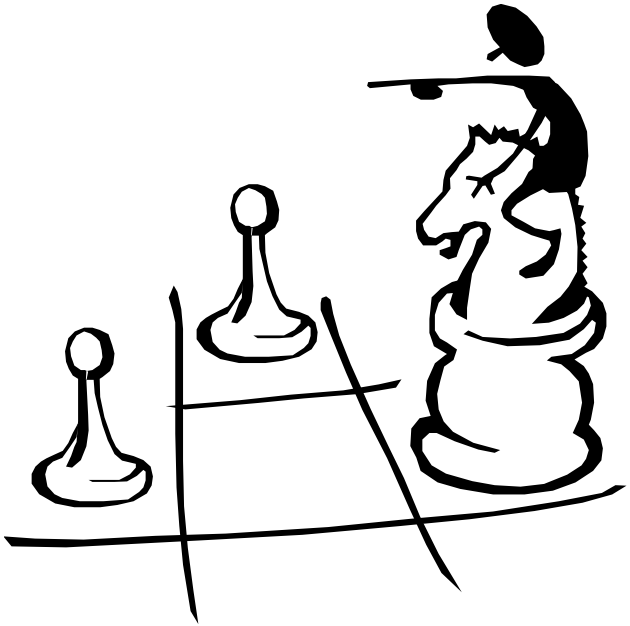
Outline:

- Basics of Data Mining
  - *Motivation & Definition*
- Architecture
- KDD process
- Issues in Data Mining
- Application and Issues in Data Mining
- Data Mining Techniques
  1. *Classification [Predictive]*
  2. *Regression [Predictive]*
  3. *Clustering [Descriptive]*
  4. *Association Rule Discovery [Descriptive]*
  5. *Sequential Pattern Discovery [Descriptive]*

Data Mining works with Warehouse Data

- Data Warehousing provides the Enterprise with a memory

- Data Mining provides the Enterprise with intelligence



Why Data Mining?

- **The Explosive Growth of Data:** from terabytes to petabytes
- Lots of data is being collected and warehoused
  - Major sources of abundant data
    - **Business:** Web, e-commerce, transactions, stocks, …
    - **Science:** Remote sensing, bioinformatics, scientific simulation,
    - **Society and everyone**: news, digital cameras, YouTube
- Data collected and stored at enormous speeds (GB/hour)
- **We are drowning in data, but starving for knowledge!**
- **Traditional techniques infeasible for raw data**
  - ***Data mining may help in classifying and segmenting data***
- Data mining—*Automated analysis of massive data sets*

Why Data Mining

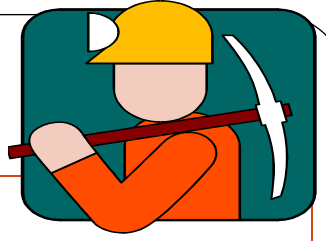
- **Credit ratings/targeted marketing:**
  - Given a database of 100,000 names, which persons are the least likely to default on their credit cards?
  - Identify likely responders to sales promotions
- **Fraud detection**
  - Which types of transactions are likely to be fraudulent, given the demographics and transactional history of a particular customer?
- **Customer relationship management:**
  - Which of my customers are likely to be the most loyal, and which are most likely to leave for a competitor?

Data Mining helps extract such information

Data mining

- Process of semi-automatically analyzing large databases to find patterns that are:
  - valid: hold on new data with some certainty
  - novel: non-obvious to the system
  - useful: should be possible to act on the item
  - understandable: humans should be able to interpret the pattern
- Also known as Knowledge Discovery in Databases (KDD)

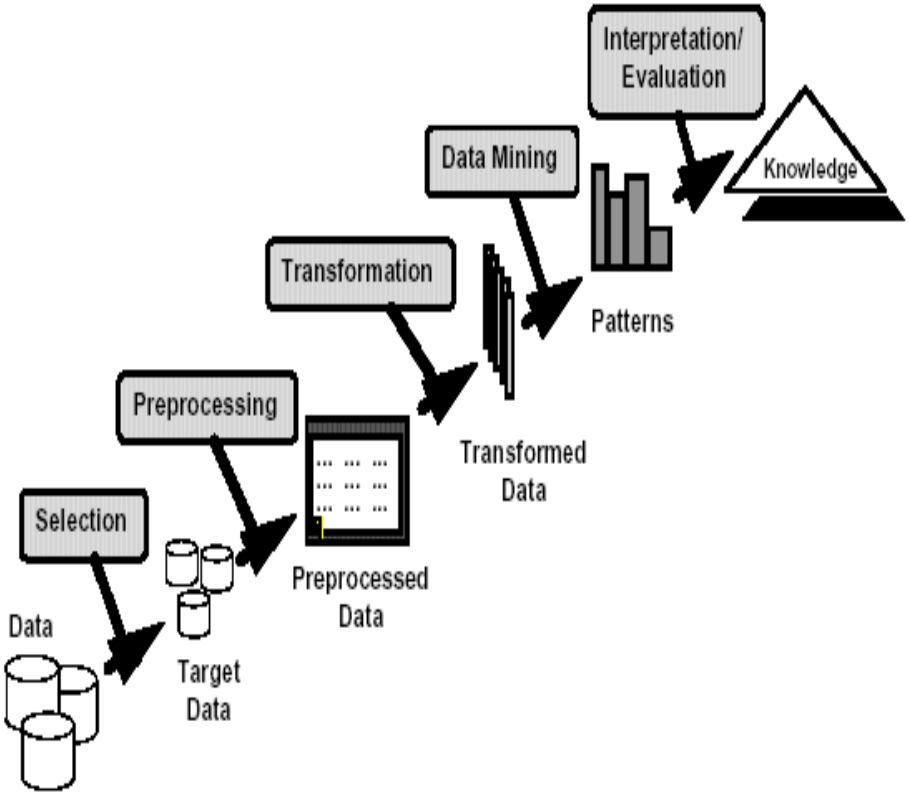
Definition of Data Mining



- **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***

Alternative names

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



Concern: Is everything "data mining"?

Examples: 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,)

Decisions in 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.
  - Multiple/integrated functions and mining at multiple levels
- **Techniques utilized**
  - Database-oriented, data warehouse (OLAP), machine learning, statistics, visualization, neural network, etc.
- **Applications adapted**
  - Retail, telecommunication, banking, fraud analysis, DNA mining, stock market analysis, Web mining, Weblog analysis, etc.

Major Issues in Data Mining- (Dec 17,May 16, May 17)

1. **Mining methodology**
   - Mining different kinds of knowledge from diverse data types, e.g., bio, stream, Web
   - Performance: efficiency, effectiveness, and scalability
   - Pattern evaluation: the interestingness problem
   - Incorporation of background knowledge
   - Handling noise and incomplete data
   - Parallel, distributed and incremental mining methods
   - Integration of the discovered knowledge with existing one: knowledge fusion

Major Issues in Data Mining (Dec 17,May 16, May 17)

2. **User interaction**
   - Data mining query languages and ad-hoc mining
   - Expression and visualization of data mining results
   - Interactive mining of knowledge at multiple levels of abstraction
3. **Applications and social impacts**
   - Domain-specific data mining & invisible data mining
   - Protection of data security, integrity, and privacy

Applications (Dec 16, Dec 17, May 16, May 18)

Application of Data mining to Financial Analysis? (Dec 16)

- Banking: loan/credit card approval
  - predict good customers based on old customers
- Customer relationship management:
  - identify those who are likely to leave for a competitor.
- Targeted marketing:
  - identify likely responders to promotions
- Fraud detection: telecommunications, financial transactions
  - from an online stream of event identify fraudulent events
- Manufacturing and production:
  - automatically adjust knobs when process parameter changes

Applications (Dec 16, Dec 17, May 16, May 18)

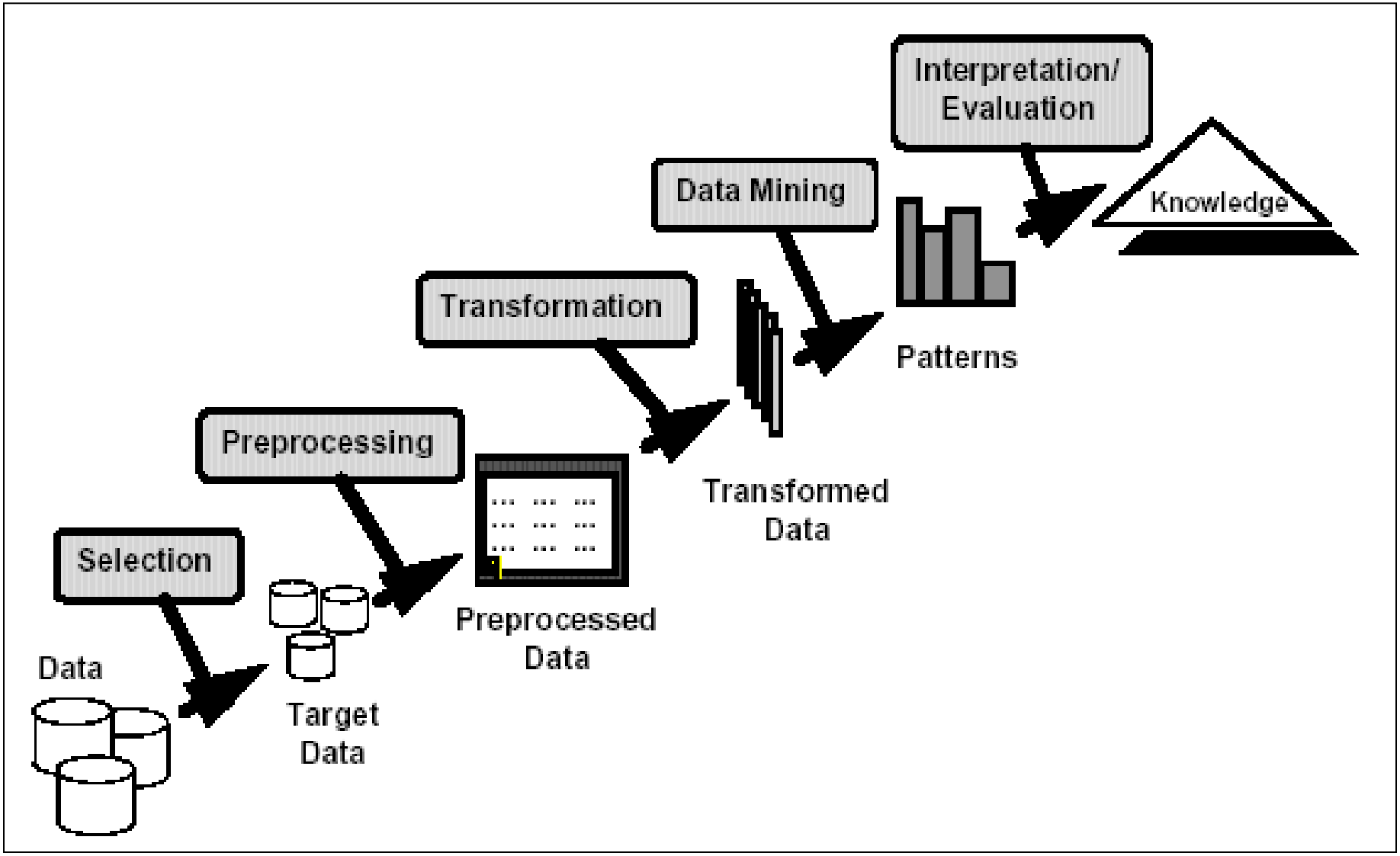
- Medicine: disease outcome, effectiveness of treatments
  - analyze patient disease history: find relationship between diseases
- Molecular/Pharmaceutical: identify new drugs
- Scientific data analysis:
  - identify new galaxies by searching for sub clusters
- Web site/store design and promotion:
  - find affinity of visitor to pages and modify layout

KDD Process and Architecture of Data mining

1. Describe Steps in the KDD Process with a suitable block Diagram. (May 2011, Dec 2012)
2. Explain data mining as a step in KDD. Give architecture of typical DM Systems. (May 2010)
3. Describe the various functionalities of data mining as a step in the process of Knowledge discovery. Dec 16
4. Architecture of a typical data mining system. Dec 17, May 16
5. Explain data mining as a steps in KDD. Give the architecture of typical data mining System. May 18

Data Mining: Knowledge Discovery (KDD) Process

Data mining: the core of knowledge discovery process.

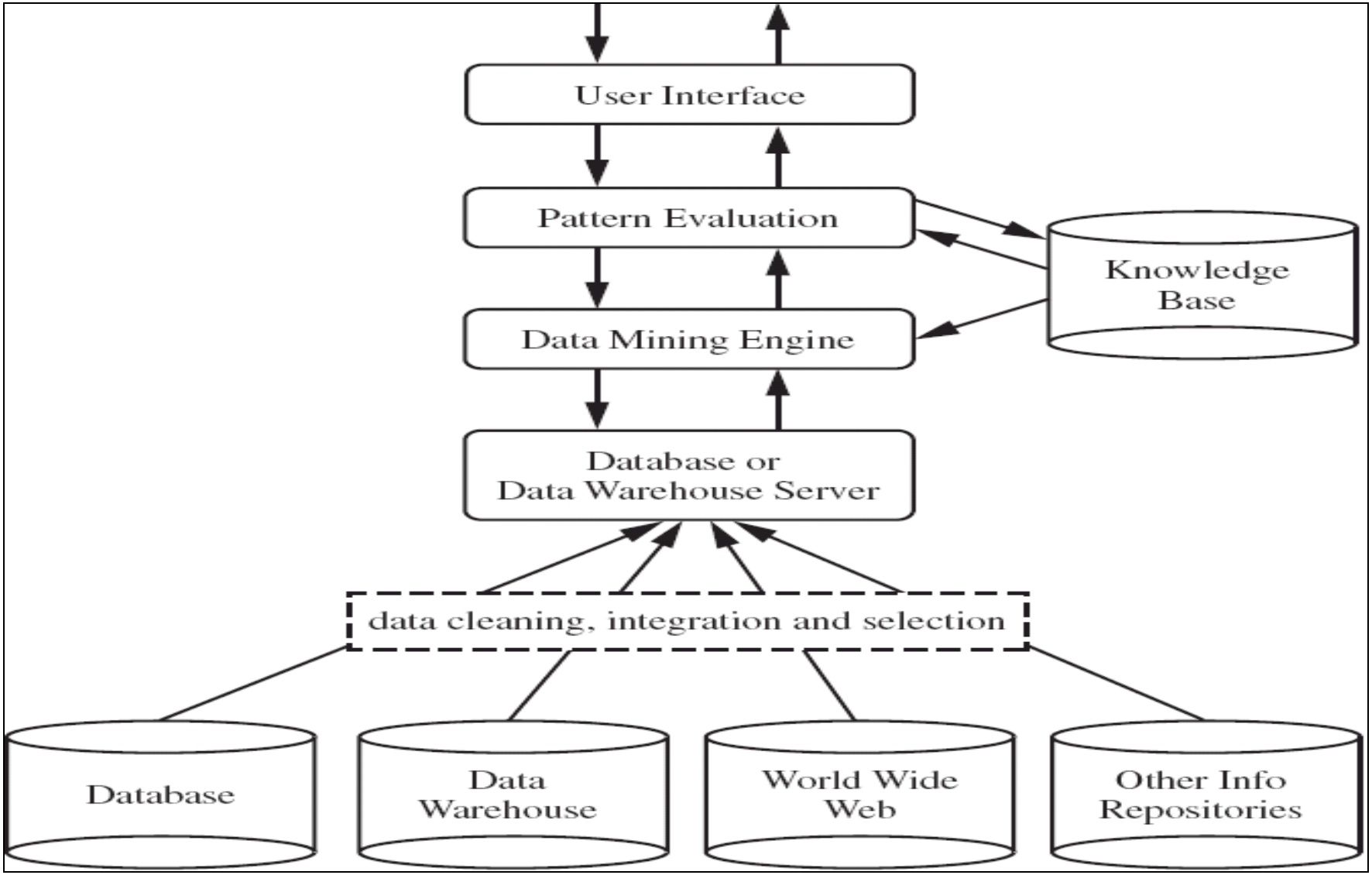


KDD Process: Several Key Steps

Key steps in the Knowledge Discovery cycle:

1. Data Cleaning: remove noise and incosistent data
2. Data Integration: combine multiple data sources
3. Data Selection: select the part of the data that are relevant for the problem
4. Data Transformation: transform the data into a suitable format (e.g., a single table, by summary or aggregation operations)
5. Data Mining: apply machine learning and machine discovery techniques
6. Pattern Evaluation: evaluate whether the found patterns meet the requirements (e.g., interestingness)
7. Knowledge Presentation: present the mined knowledge to the user (e.g., visualization)

Architecture: Typical Data Mining System



Data Mining: Classification Schemes

Data Mining Methods:

- **Prediction Methods**
  - Use some variables to predict unknown or future values of other variables.
  - Based on this methods these are two data mining Techniques:
    1. **Classification**
    2. **Regression**
- **Description Methods**
  - Find human-interpretable patterns that describe the data.
  - Based on this methods these are three data mining Techniques:
    1. **Clustering**
    2. **Association Rule Discovery**
    3. **Sequential Pattern Discovery**

1. Classification: Definition

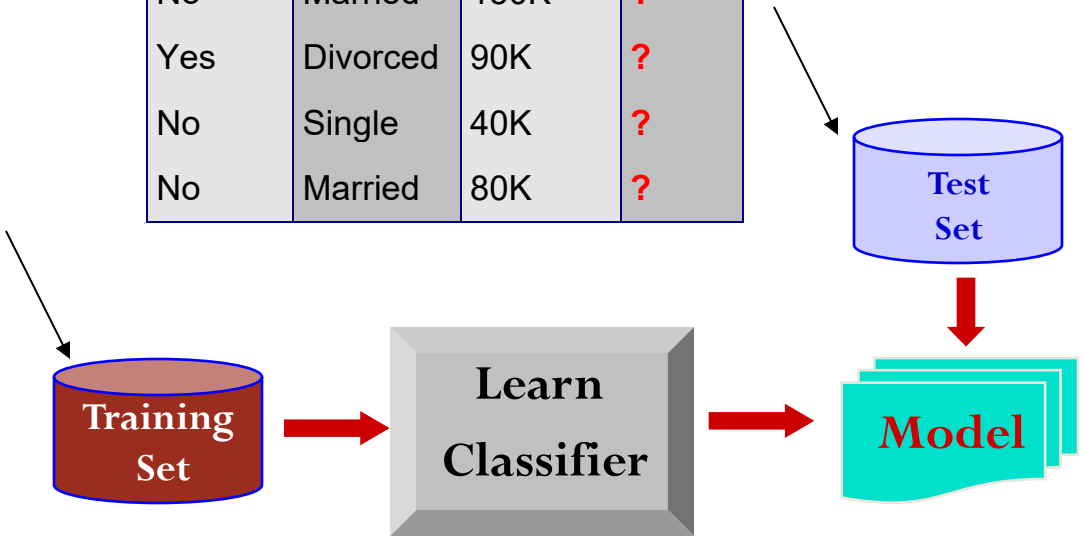
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.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

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	?



1. 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.

1. 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.

1. Classification: Application 3

- **Customer Attrition/Churn:**
  - ***Goal:*** To predict whether a customer is likely to be lost to a competitor.
  - ***Approach:***
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
    - Label the customers as loyal or disloyal.
    - Find a model for loyalty.

2. 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 advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

3. Clustering Definition

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

3. 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.

3. 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.

Illustrating Document Clustering

- 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

4. 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.
- Initially used for Market Basket Analysis to find how items purchased by customers are related.

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}

4. 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 beer:

$$Diapers \rightarrow Beer,\ support = 20\%,\ confidence = 85\%$$

4. Association Rule Discovery: Application 2

- 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!

5. Sequential Pattern Discovery: Definition

- Given is a set of *objects*, with each object associated with its own *timeline of events*, 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.

(A B) (C) ⟶ (D E)

Difference between Association Rule & Sequential Patterns:

- **Terminology:**
  - **Association Rules** refer to what items are bought together (at the same time)
    - Intra-transaction patterns
  - **Sequential Patterns** refer to what items are bought at different times
    - Inter-transaction patterns

Applications of sequential pattern mining

1. **Customer shopping sequences:**
   - First buy computer, then CD-ROM, and then Web camera, within 3 months.
2. **Medical treatments, natural disasters (e.g., earthquakes), science & eng. processes, stocks and markets, etc.**
3. **Telephone calling patterns, Weblog click streams**
4. **DNA sequences and gene structures**

Thank You