Data Mining

Introduction to Data Mining

Slides by Tan, Steinbach, Kumar adapted by Pimprapai Thainiam

Data vs Information

Difference between data and information?

Data vs Information

Data is raw facts that is to be processed for further information

Data is useless unless it is processed or has been made into something.

Data has no meaning when it has not been interpreted.

while

Information is the processed data.

Information is meaningful when data is gathered and meaning is generated.

Information is a useful context for decision making.

Information depends upon data. It cannot be generated without the help of data.

Data vs Information

Examples of Data

- **Student Data on Admission Forms:** When students get admission in a college. They fill admission form. This form contains raw facts (data of student) like name, father's name, address of student etc.
- **Data of Citizens:** During census, data of all citizens is collected.
- **Survey Data:** Different companies collect data by survey to know the opinion of people about their product.
- **Students Examination data:** In examination data about obtained marks of different subjects for all students is collected.

Data vs Information

Examples of Information

- **Census Report:** Census data is used to get report/information about total population of a country and literacy rate etc.
- **Survey Reports and Results:** Survey data is summarized into reports/information to present to management of the company.
- **Result Cards of Individual Students:** In examination system collected data (obtained marks in each subject) is processed to get total obtained marks of a student. Total obtained marks are Information. It is also used to prepare result card of a student.
- **Merit List:** After collecting admission forms from candidates, merit is calculated on the basis of obtained marks of each candidate. Normally, percentage of marks obtained is calculated for each candidate. Now all the candidates names are arranged in descending order by percentage. This makes a merit list. Merit list is used to decide whether a candidate will get admission in the college or not.

Why Mine Data? Commercial Viewpoint

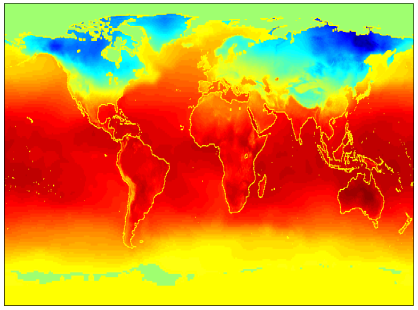
- Lots of data is being collected and warehoused
  - Web data, e-commerce (e.g.lazada, amazon, ebay)
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful

- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

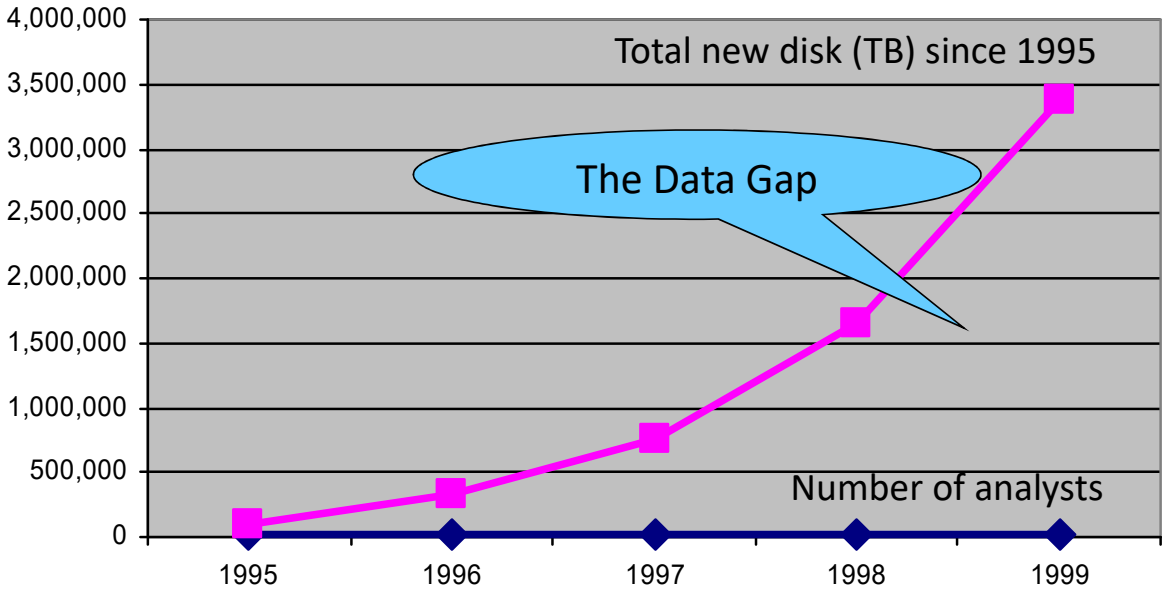
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - Remote sensors on a satellite
  - Telescopes scanning the skies
  - Microarrays 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



Mining Large Data Sets - 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"

Definitions of Data Mining

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
- Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Process of automatically discover useful information in large data repositories
- A business process for exploring large amounts of data to discover meaningful patterns and rules

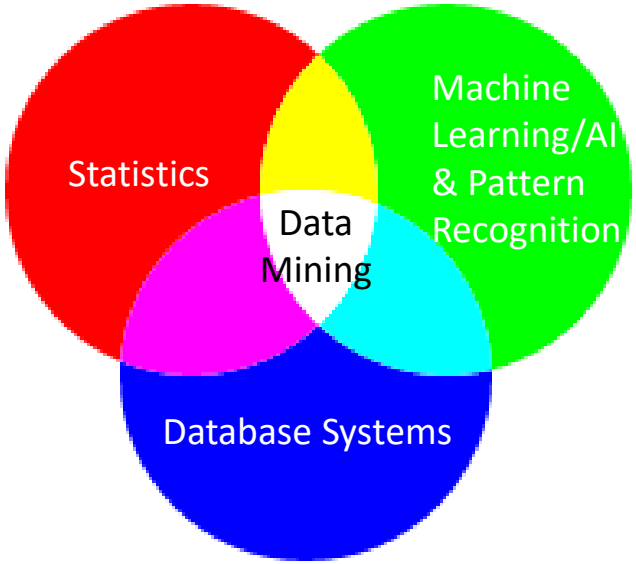
What is Data Mining? & What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”
- 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

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



Data Mining Tasks

- **Predictive Methods:** Use some variables (attributes) to predict unknown or future values of other variables (attributes).
  - Classification – discrete (categorical) target attribute
  - Regression – continuous target attribute

- **Descriptive Methods:** Find human-interpretable patterns that describe the data.
  - Clustering
  - Association Rules Mining

Data Mining Tasks

Task	Type	Goal
Classification	Predictive	To accurately predict the target class for each object.
Clustering	Descriptive	To group a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).
Association Rules Mining	Descriptive	To identify strong rules discovered in databases using some measures of interestingness.
Sequential Pattern Discovery	Descriptive	To find statistically relevant patterns between data examples where the values are delivered in a sequence.
Regression	Predictive	To find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
Deviation/Anomaly Detection	Predictive	To identify unusual data records, that might be interesting or data errors that require further investigation.

Classification: Definition

- **Given:** a collection of records (training set) where each record contains a set of attributes, one of the attributes is the class or target attribute.
  column
- **Find:** a model for target class attribute as a function of the values of other attributes.
- **Goal:** previously unseen records should be assigned a class as accurately as possible.
  Items
- 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.

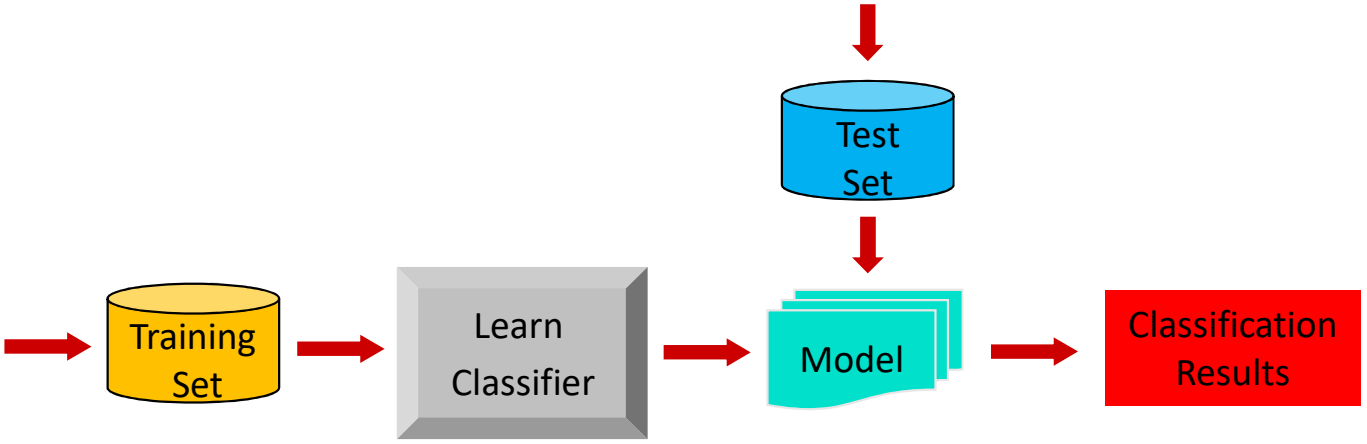
Classification Example

Will a person cheat on income his/her tax?

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class, target
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

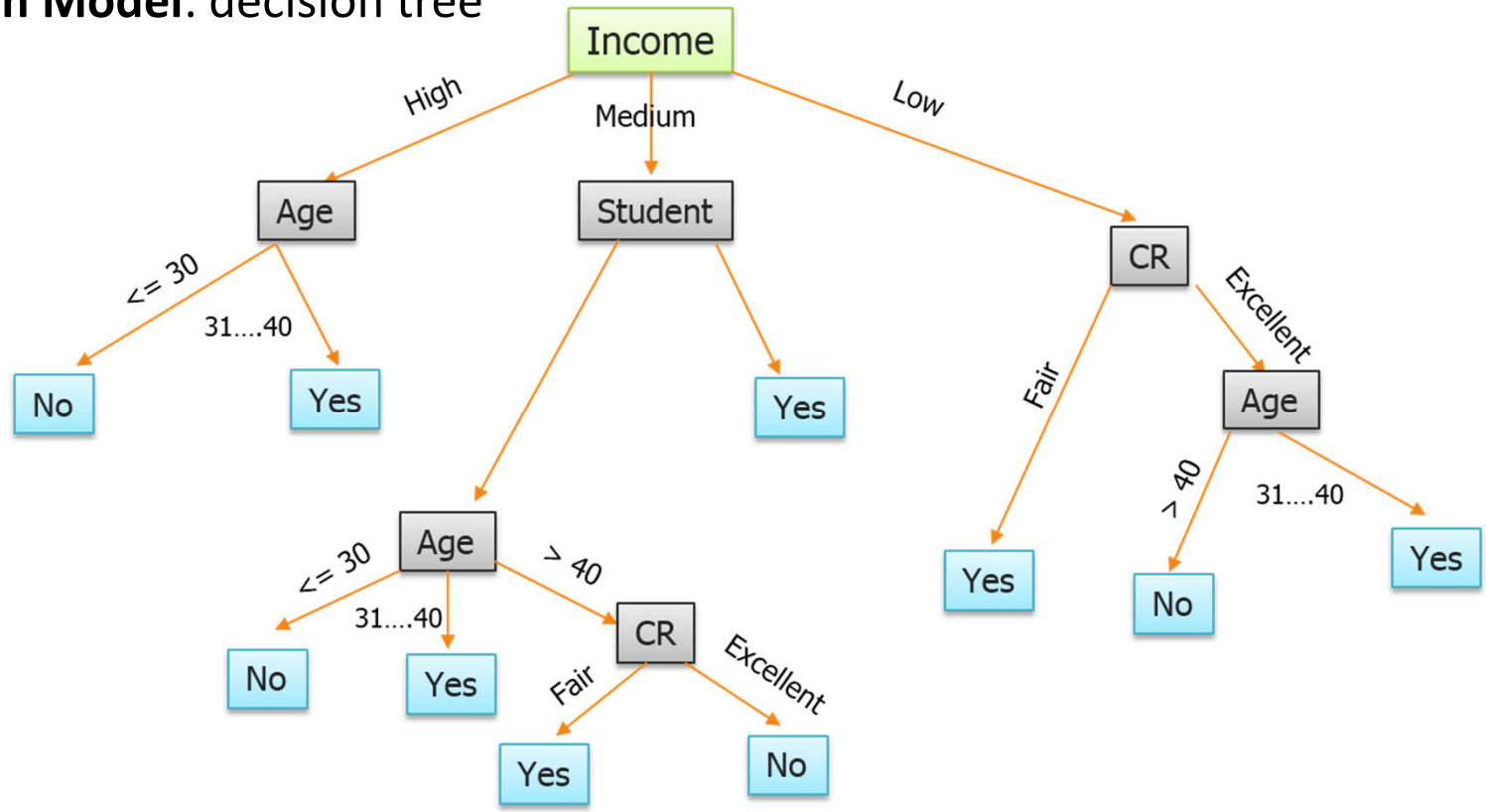
← Attributes

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 Example

Classification Model: decision tree



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 such as type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model.

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 such as 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.

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.

Association Rules Mining: Definition

- **Given:** a set of records each of which contain some number of items from a given collection

- **Find:** dependency rules which will predict occurrence of an item based on occurrences of other items.

Transaction ID

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 Rules Mining: 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 Rules Mining: 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!

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 Rules Mining: Application 3

Inventory Management

- **Goal:** A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.

- **Approach:** Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.

Clustering: Definition

- **Given:** a set of data points, each having a set of attributes, and a similarity measure among them

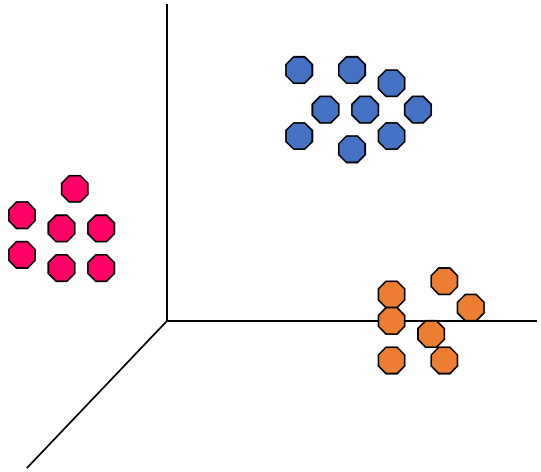
- **Find:** clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.

- 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



Clustering: Application 1

Market Segmentation

- **Goal:** To 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.

Illustrating Document Clustering

- **Clustering Points:** 3,195 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	276
Sports	738	573
Entertainment	345	278

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.

Regression

Simple linear regression (1 independent attribute)