# Introduction to Data Mining

Dr. Sanjay Ranka
Professor
Computer and Information Science and Engineering
University of Florida, Gainesville

ranka@cise.ufl.edu

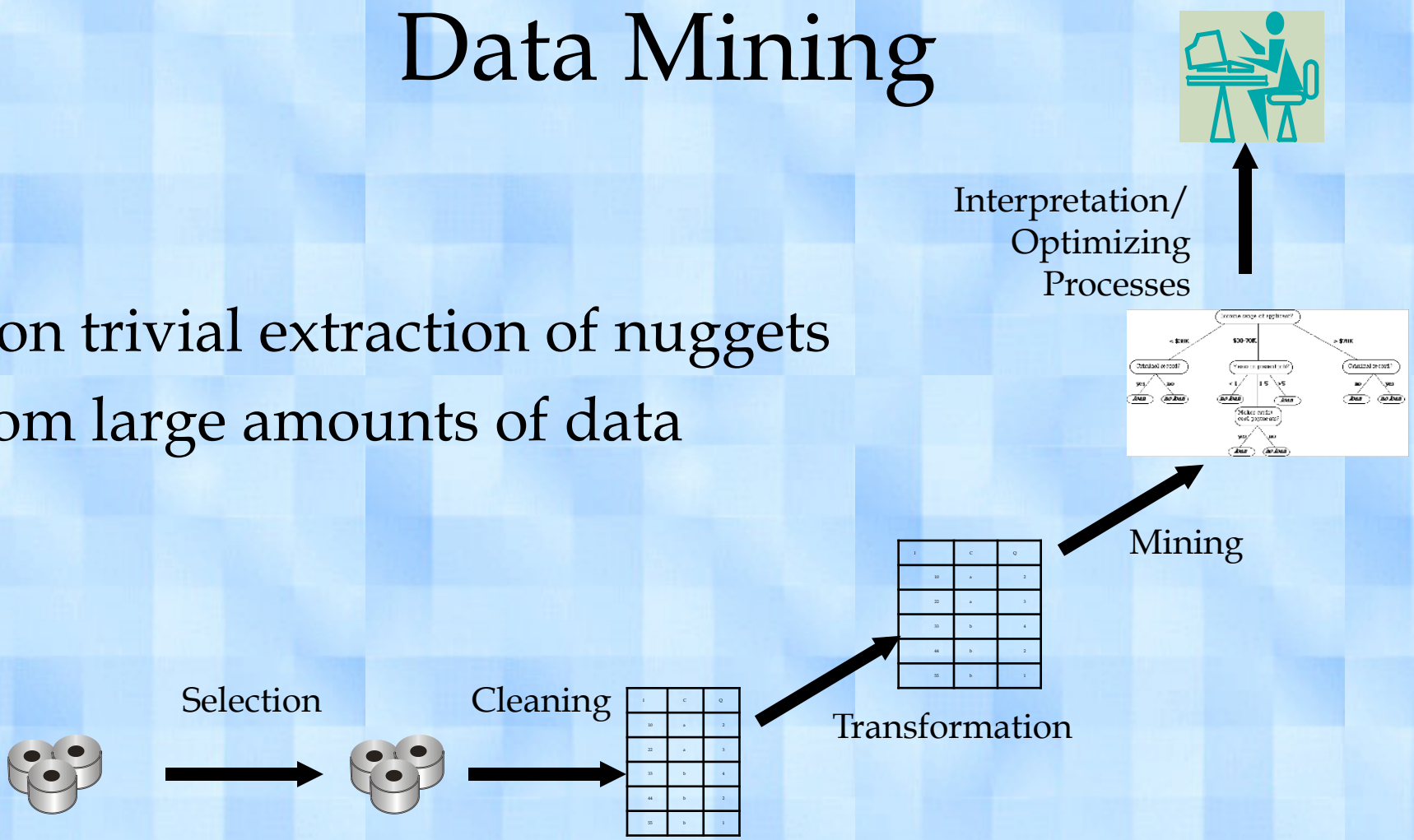
# Course Overview

- Introduction to Data Mining
- Important data mining primitives:
  - Classification
  - Clustering
  - Association Rules
  - Sequential Rules
  - Anomaly Detection
- Commercial and Scientific Applications

- Background required:
  - General background in algorithms and programming
- Grading scheme:
  - 4 to 6 home works (10%)
  - 3 in-class exams ( 30% *each* )
  - Last exam may be replaced by a project
- Textbook:
  - *Introduction to Data Mining* by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, 2003
  - *Data Mining: Concepts and Techniques* by Jiawei Han and Micheline Kamber, 2000

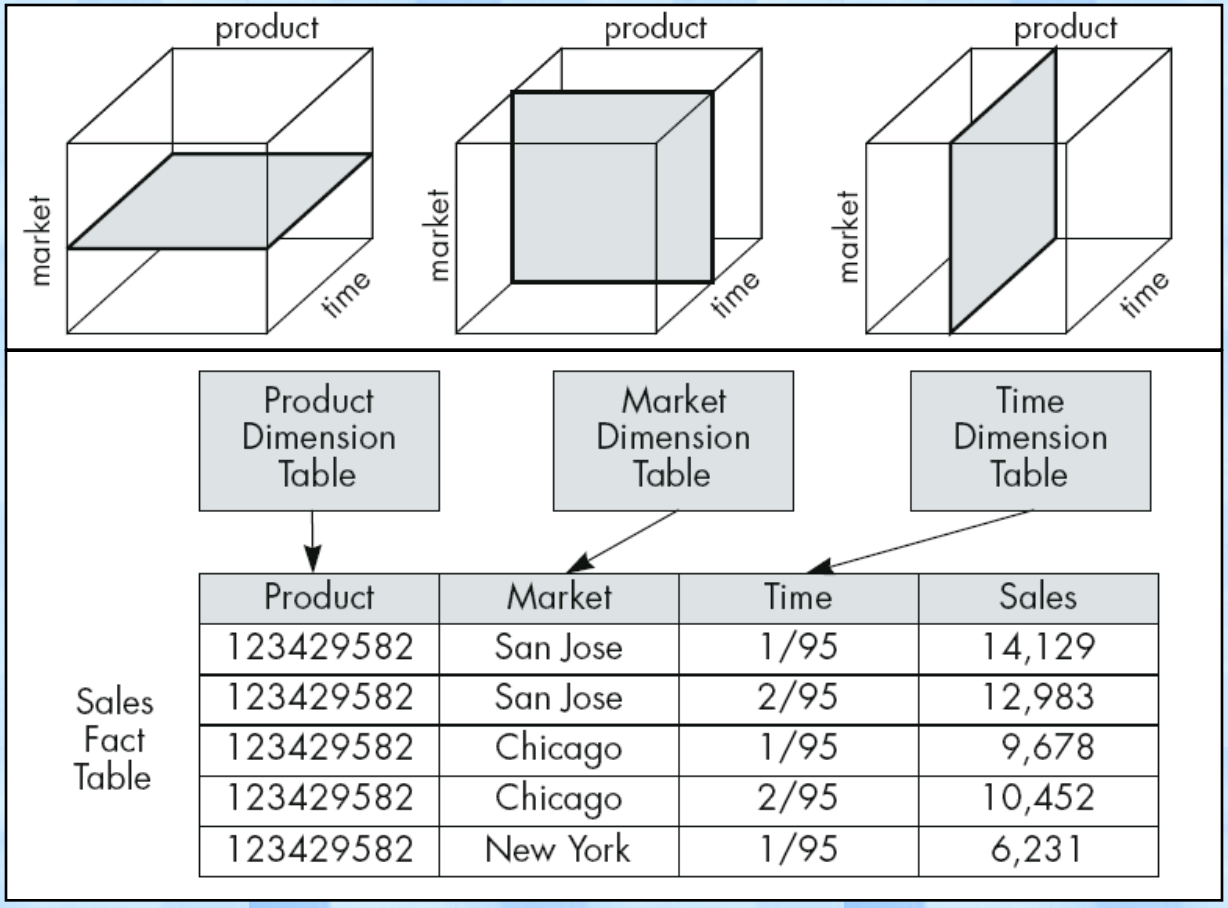
# Data Mining

Non trivial extraction of nuggets from large amounts of data



# Data Mining is not …

- Generating multidimensional cubes of a relational table



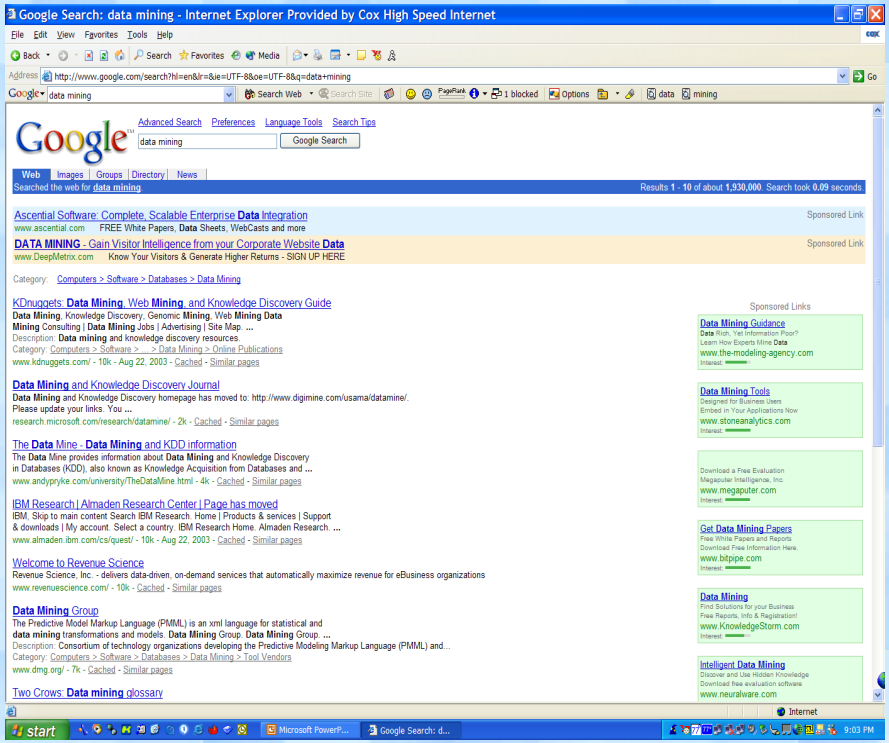

| Product | Market | Time | Sales |
|---|---|---|---|
| 123429582 | San Jose | 1/95 | 14,129 |
| 123429582 | San Jose | 2/95 | 12,983 |
| 123429582 | Chicago | 1/95 | 9,678 |
| 123429582 | Chicago | 2/95 | 10,452 |
| 123429582 | New York | 1/95 | 6,231 |

Source: Multidimensional OLAP vs. Relational OLAP by Colin White

# Data Mining is not …

- Searching for a phone number in a phone book

- Searching for keywords on Google



# Data Mining is not …

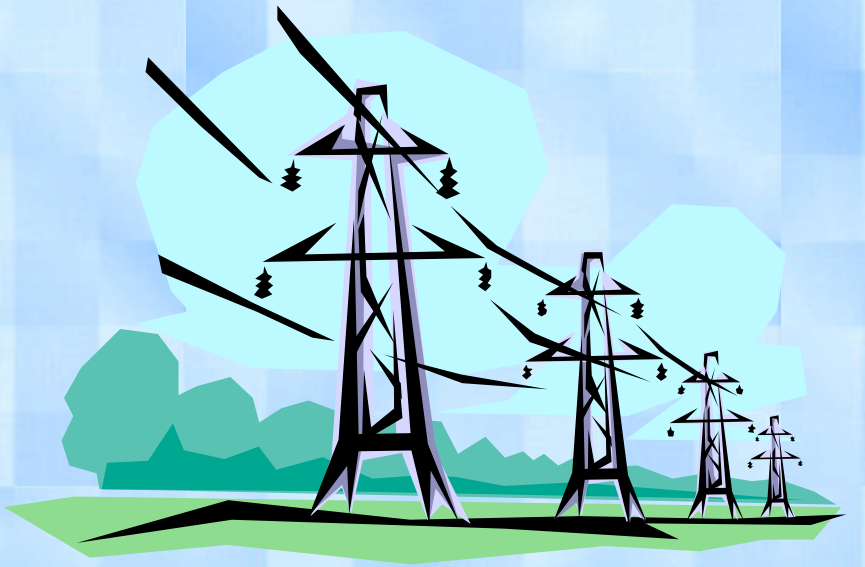
- Generating a histogram of salaries for different age groups

- Issuing SQL query to a database, and reading the reply

# Data Mining is …

- Finding groups of people with similar hobbies

- Are chances of getting cancer higher if you live near a power line?

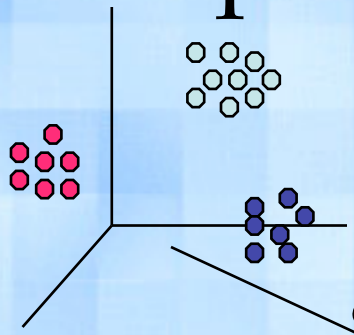


# Data Mining Tasks

- Prediction methods
  - Use some variables to predict unknown or future values of the same or other variables
- Description methods
  - Find human interpretable patterns that describe data

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

# Important Data Mining Primitives

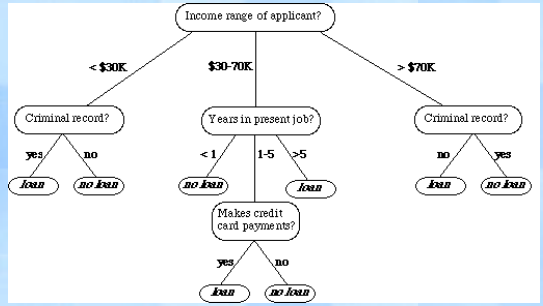


Clustering

Data

| 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 |
| 11 | No | Married | 60K | No |
| 12 | Yes | Divorced | 220K | No |
| 13 | No | Single | 85K | Yes |
| 14 | No | Married | 75K | No |
| 15 | No | Single | 90K | Yes |

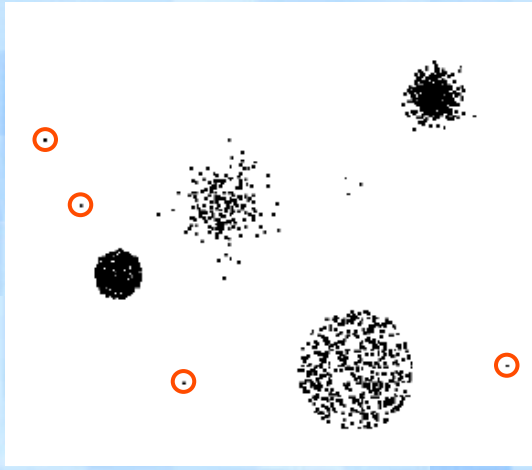
Predictive Modeling




Association Rules

Milk


Anomaly / Deviation Detection



# Data Mining Tasks …

- Classification (predictive)
- Clustering (descriptive)
- Association Rule Discovery (descriptive)
- Sequential Pattern Discovery (descriptive)
- Regression (predictive)
- Deviation Detection (predictive)

# Why is Data Mining prevalent? Lots of data is collected and stored in data warehouses

- Business
  - Wal-Mart logs nearly 20 million transactions per day
- Astronomy
  - Telescope collecting large amounts of data (SDSS)
- Space
  - NASA is collecting peta bytes of data from satellites
- Physics
  - High energy physics experiments are expected to generate 100 to 1000 tera bytes in the next decade

# Why is Data Mining prevalent? Quality and richness of data collected in improving

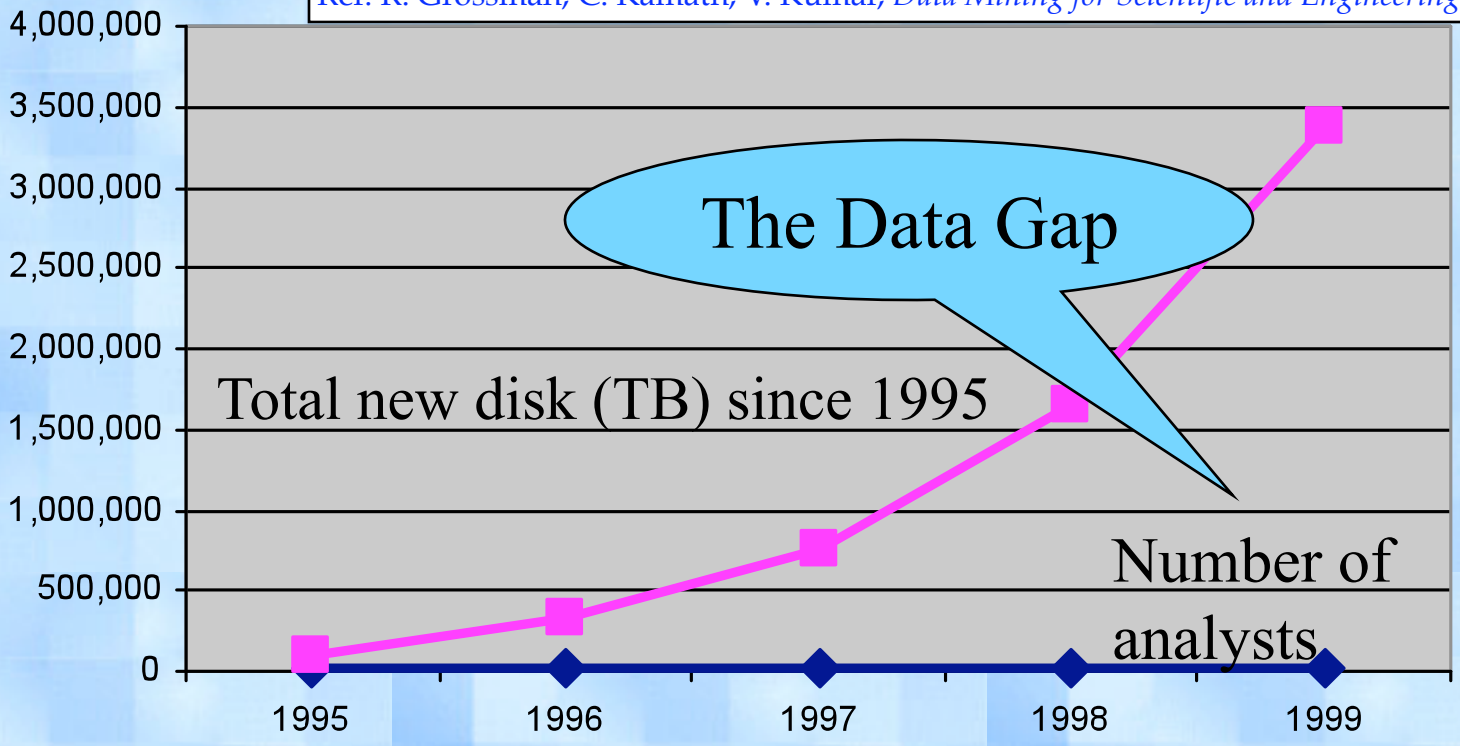
- Retailers
  - Scanner data is much more accurate than other means
- E-Commerce
  - Rich data on consumer browsing
- Science
  - Accuracy of sensors is improving

# Why is Data Mining prevalent?

## The gap between data and analysts is increasing

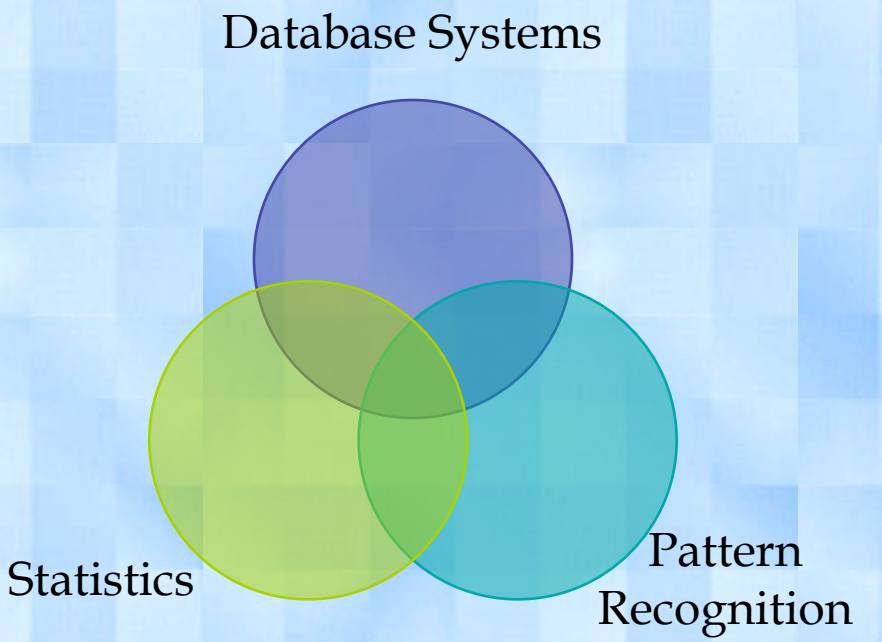
- Hidden information is not always evident
- High cost of human labor
- Much of data is never analyzed at all

Ref: R. Grossman, C. Kamath, V. Kumar, *Data Mining for Scientific and Engineering Applications*



# Origins of Data Mining

- Drawn ideas from Machine Learning, Pattern Recognition, Statistics, and Database systems for applications that have
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous data
  - Unstructured data

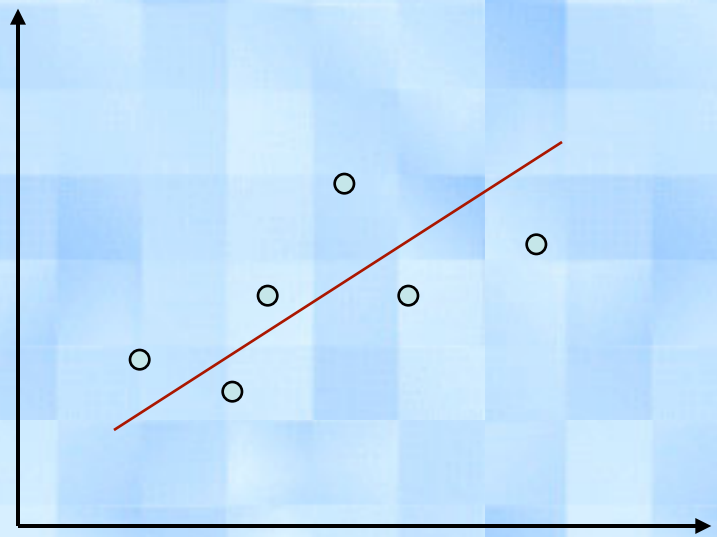


# Regression

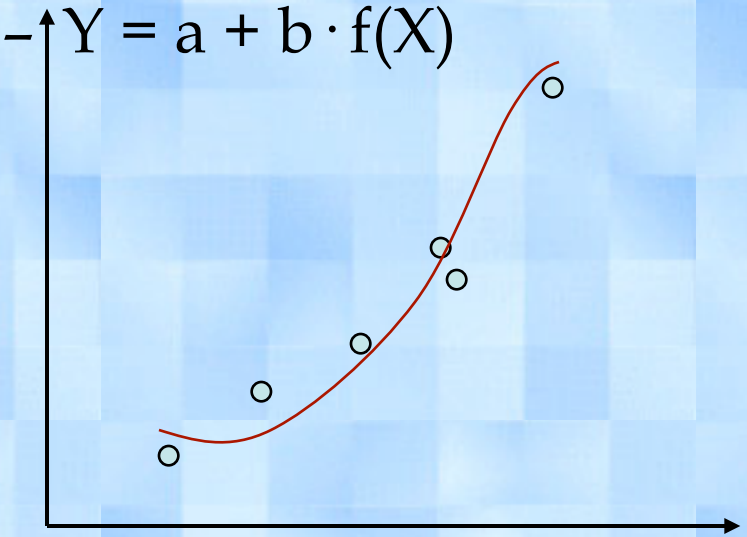
- Predict the value of a given continuous valued variable based on the values of other variables, assuming a linear or non-linear model of dependency
- Extensively studied in the fields of Statistics and Neural Networks
- Examples
  - Predicting sales numbers of a new product based on advertising expenditure
  - Predicting wind velocities based on temperature, humidity, air pressure, etc
  - Time series prediction of stock market indices

# Regression

- Linear regression
  - Data is modeled using a straight line
  - $Y = a + bX$



- Non-linear regression
  - Data is more accurately correctly modeled using a nonlinear function
  - $Y = a + b \cdot f(X)$



# Association Rule Discovery

- Given a set of transactions, each of which is a set of items, find all rules (X → Y ) that satisfy user specified minimum support and confidence constraints
- Support = (#T containing X and Y) / (#T)
- Confidence = (#T containing X and Y ) / (#T containing X)
- Applications
  - Cross selling and up selling
  - Supermarket shelf management

Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

| Transaction | Items |
|---|---|
| T1 | Bread, Jelly, Peanut Butter |
| T2 | Bread, Peanut Butter |
| T3 | Bread, Milk, Peanut Butter |
| T4 | Beer, Bread |
| T5 | Beer, Milk |

- Some rules discovered
  - Bread → Peanut Butter
    - support=60%, confidence=75%
  - Peanut Butter → Bread
    - support=60%, confidence=100%
  - Jelly → Peanut Butter
    - support=20%, confidence=100%
  - Jelly → Milk
    - support=0%

# 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
- Example:
  - {Bread} → {Peanut Butter}
  - {Jelly} → {Peanut Butter}

# Association Rule Discovery: Marketing and sales promotion

- Say the rule discovered is

  *{Bread, …. } → {Peanut Butter}*

- *Peanut Butter as a consequent*: can be used to determine what products will boost its sales
- *Bread as an antecedent*: can be used to see which products will be impacted if the store stops selling bread (e.g. cheap soda is a "loss leader" for many grocery stores.)
- *Bread as an antecedent and Peanut Butter as a consequent*: can be used to see what products should be stocked along with Bread to promote the sale of Peanut Butter

# Association Rule Discovery: Super market shelf management

- *Goal*: To identify items that are bought concomitantly by a reasonable fraction of customers so that they can be shelved appropriately based on business goals.
- *Data Used*: Point-of-sale data collected with barcode scanners to find dependencies among products
- *Example*
  - If a customer buys Jelly, then he is very likely to buy Peanut Butter.
  - So don't be surprised if you find Peanut Butter next to Jelly on an aisle in the super market. Also, salsa next to tortilla chips.

# Association Rule Discovery: Inventory Management

- *Goal*: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products, and wants to keep the service vehicles equipped with frequently used parts to reduce the number of visits to consumer household.
- *Approach*: Process the data on tools and parts required in repairs at different consumer locations and discover the co-occurrence patterns

# Association Rules: Apparel

Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

| Transaction | Items | Transaction | Items |
|---|---|---|---|
| T1 | Blouse | T11 | T-Shirt |
| T2 | Shoes, Skirt, T-Shirt | T12 | Blouse, Jeans, Shoes, Skirt, T-Shirt |
| T3 | Jeans, T-Shirt | T13 | Jeans, Shoes, Shorts, T-Shirt |
| T4 | Jeans, Shoes, T-Shirt | T14 | Shoes, Skirt, T-Shirt |
| T5 | Jeans, Shorts | T15 | Jeans, T-Shirt |
| T6 | Shoes, T-Shirt | T16 | Skirt, T-Shirt |
| T7 | Jeans, Skirt | T17 | Blouse, Jeans, Skirt |
| T8 | Jeans, Shoes, Shorts, T-Shirt | T18 | Jeans, Shoes, Shorts, T-Shirt |
| T9 | Jeans | T19 | Jeans |
| T10 | Jeans, Shoes, T-Shirt | T20 | Jeans, Shoes, Shorts, T-Shirt |

{Jeans, T-Shirt, Shoes} → { Shorts}
Support: 20% Confidence: 100%

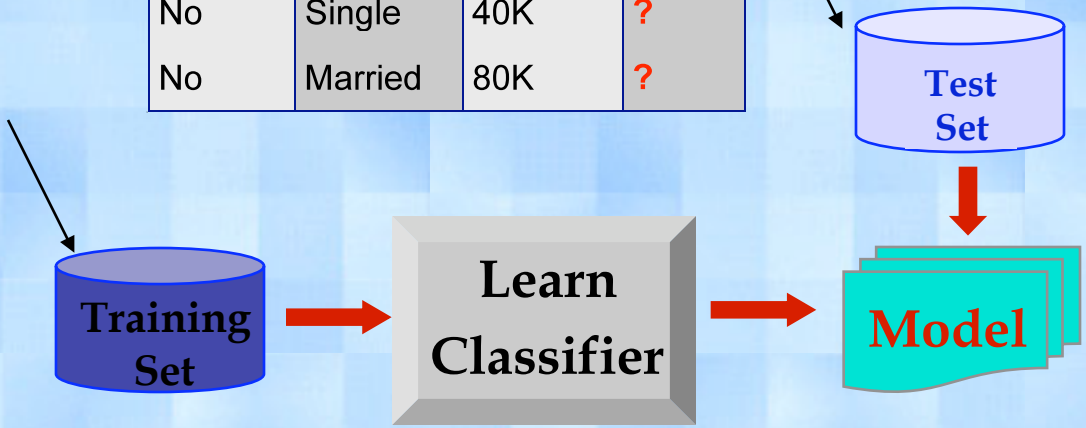
# Classification: Definition

- Given a set of records (called the training set)
  - Each record contains a set of attributes. One of the attributes is the class
- Find a model for the class attribute as a function of the values of other attributes
- *Goal*: Previously *unseen* records should be assigned to a class as accurately as possible
  - Usually, the given data set is divided into training and test set, with training set used to build the model and test set used to validate it. The accuracy of the model is determined on the test set.

# Classification Example

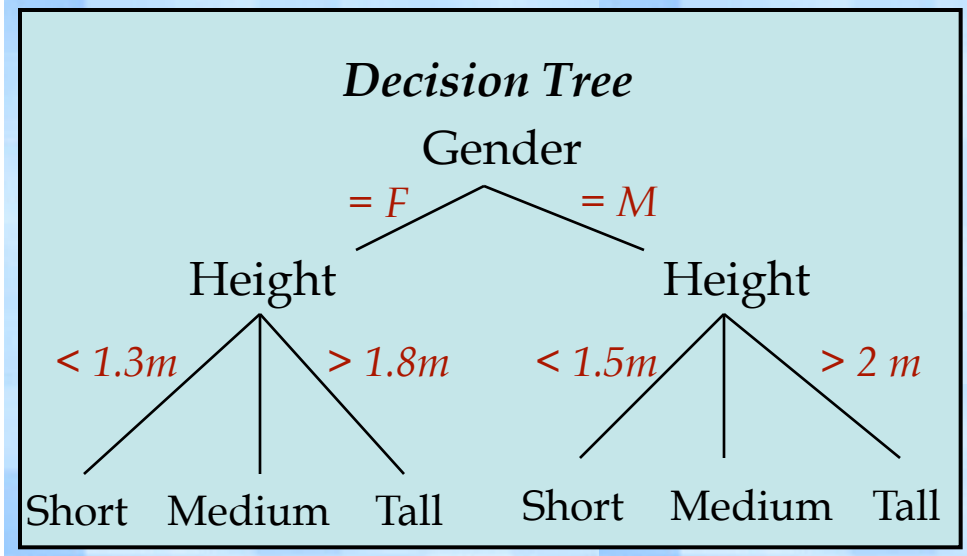
| Tid | Refund (categorical) | Marital Status (categorical) | Taxable Income (continuous) | Cheat (class) |
|---|---|---|---|---|
| 1 | Yes | Single | 125K | No |
| 2 | No | Married | 100K | No |
| 3 | No | Single | 70K | No |
| 4 | Yes | Married | 120K | No |
| 5 | No | Divorced | 95K | Yes |
| 6 | 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 | ? |



# Classification

- Modeling a class attribute, using other attributes
- Applications
  - Targeted marketing
  - Customer attrition




Source: Data Mining – Introductory and Advanced topics by Margaret Dunham

| Name | Gender | Height | Output |
|---|---|---|---|
| Kristina | F | 1.6 m | Medium |
| Jim | M | 2 m | Medium |
| Maggie | F | 1.9 m | Tall |
| Martha | F | 1.88 m | Tall |
| Stephanie | F | 1.7 m | Medium |
| Bob | M | 1.85 m | Medium |
| Kathy | F | 1.6 m | Medium |
| Dave | M | 1.7 m | Medium |
| Worth | M | 2.2 m | Tall |
| Steven | M | 2.1 m | Tall |
| Debbie | F | 1.8 m | Medium |
| Todd | M | 1.95 m | Medium |
| Kim | F | 1.9 m | Tall |
| Amy | F | 1.8 m | Medium |
| Lynette | F | 1.75 m | Medium |

# Classification: 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 collected for a similar product introduced in the recent past.
  - Use the profiles of customers along with their {buy, didn't buy} decision. The latter becomes the class attribute.
  - The profile of the information may consist of demographic, lifestyle and company interaction.
    - Demographic – Age, Gender, Geography, Salary
    - Psychographic – Hobbies
    - Company Interaction –Recentness, Frequency, Monetary
  - Use these information as input attributes to learn a classifier model

Source: Data Mining Techniques, Berry and Linoff, 1997

# Classification: Fraud Detection

- *Goal:* Predict fraudulent cases in credit card transactions
- *Approach*:
  - Use credit card transactions and the information on its account holders as attributes (important information: when and where the card was used)
  - Label past transactions as {fraud, fair} transactions. This forms the class attribute
  - Learn a model for the class of transactions
  - Use this model to detect fraud by observing credit card transactions on an account

# Classification: Customer Churn

- *Goal:* To predict whether a customer is likely to be lost to a competitor
- *Approach*:
  - Use detailed record of transaction with each of the past and current customers, to find attributes
    - How often does the customer call, Where does he call, What time of the day does he call most, His financial status, His marital status, etc. (Important Information: Expiration of the current contract).
  - Label the customers as {churn, not churn}
  - Find a model for Churn

Source: Data Mining Techniques, Berry and Linoff, 1997

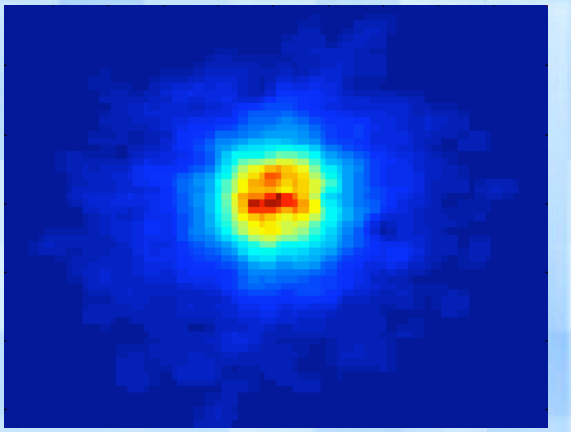
# Classification: Sky survey cataloging

- *Goal*: To predict class {star, 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 (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) !!!

Source: Advances in Knowledge Discovery and Data Mining, Fayyad et al., 1996

# Classification: Classifying Galaxies

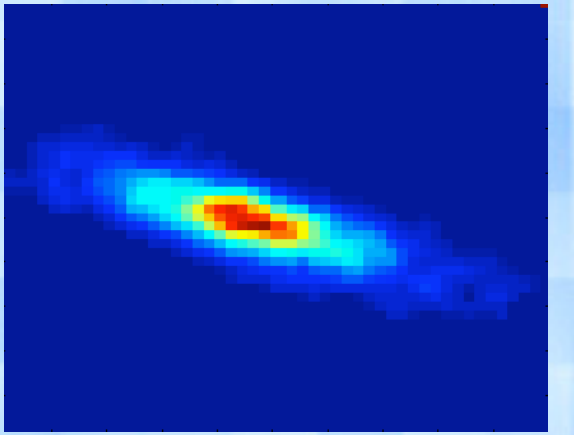
*Early*



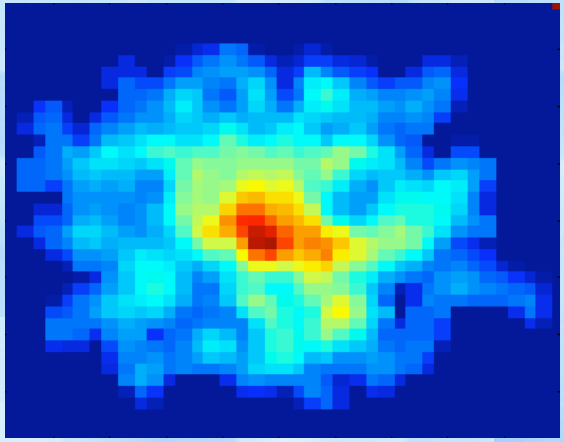
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

*Intermediate*



*Late*



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

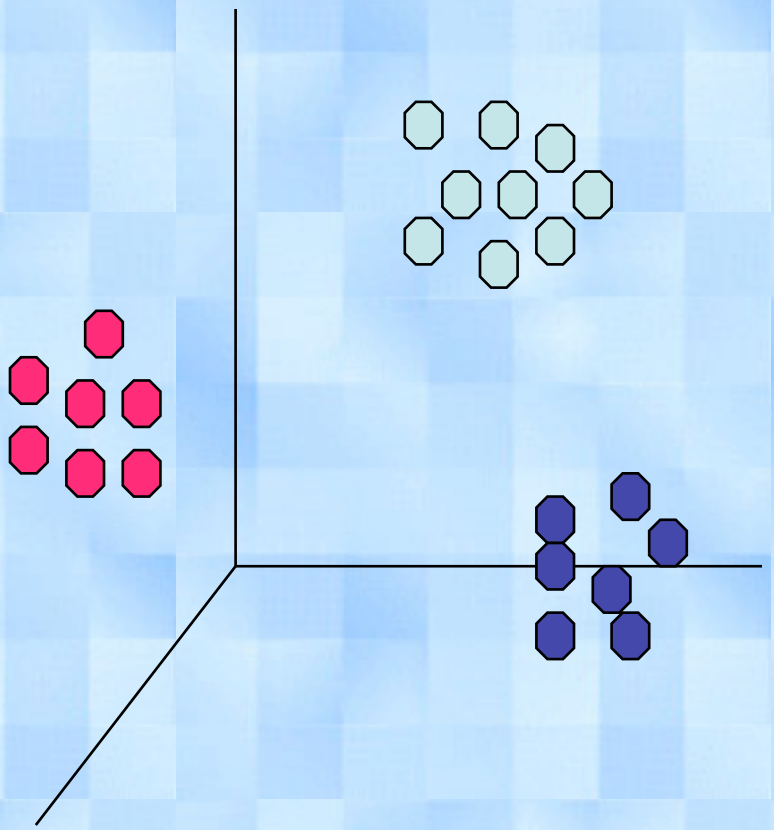
Source: Minnesota Automated Plate Scanner Catalog, http://aps.umn.edu

# Clustering

- Determine object groupings such that objects within the same cluster are similar to each other, while objects in different groups are not
- Typically objects are represented by data points in a multidimensional space with each dimension corresponding to one or more attributes. Clustering problem in this case reduces to the following:
  - Given a set of data points, each having a set of attributes, and a similarity measure, 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

# Clustering Example

- Euclidean distance based clustering in 3D space
  - Intra cluster distances are minimized
  - Inter cluster distances are maximized



# Clustering: Market Segmentation

- *Goal*: To subdivide a market into distinct subset of customers where each subset can be targeted 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 the buying patterns of customers in the same cluster vs. those from different clusters

# Clustering: Document Clustering

- *Goal*: To find groups of documents that are similar to each other based on important terms appearing in them
- *Approach*: To identify frequently occurring terms in each document. Form a similarity measure based on frequencies of different terms. Use it to generate clusters
- *Gain*: Information Retrieval can utilize the clusters to relate a new document or search term to clustered documents

# Clustering: Document Clustering Example

- Clustering points: 3024 articles of Los Angeles Times
- Similarity measure: Number of common words in documents (after some word filtering)

| Category | Total articles | Correctly placed articles |
|---|---|---|
| Financial | 555 | 364 |
| Foreign | 341 | 260 |
| National | 273 | 36 |
| Metro | 943 | 746 |
| Sports | 738 | 573 |
| Entertainment | 354 | 278 |

# Clustering: S&P 500 stock data

- Observe stock movements everyday
- Clustering points: Stock – {UP / DOWN}
- Similarity measure: Two points are more similar if the events described by them frequently happen together on the same day

| | Discovered Clusters | Industry Group |
|---|---|---|
| 1 | Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN | Technology1-DOWN |
| 2 | Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device- DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN | Technology2-DOWN |
| 3 | Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN | Financial-DOWN |
| 4 | Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP | Oil-UP |

# Deviation / Anomaly Detection

- Some data objects do not comply with the general behavior or model of the data. Data objects that are different from or inconsistent with the remaining set are called outliers
- Outliers can be caused by measurement or execution error. Or they represent some kind of fraudulent activity.
- Goal of Deviation / Anomaly Detection is to detect significant deviations from normal behavior

# Deviation / Anomaly Detection: Definition

- Given a set of *n* data points or objects, and *k*, the expected number of outliers, find the top *k* objects that considerably dissimilar, exceptional or inconsistent with the remaining data
- This can be viewed as two sub problems
  - Define what data can be considered as inconsistent in a given data set
  - Find an efficient method to mine the outliers so defined

# Deviation: Credit Card Fraud Detection

- *Goal*: To detect fraudulent credit card transactions
- *Approach*:
  - Based on past usage patterns, develop model for authorized credit card transactions
  - Check for deviation form model, before authenticating new credit card transactions
  - Hold payment and verify authenticity of "doubtful" transactions by other means (phone call, etc.)

# Anomaly Detection: Network Intrusion Detection

- *Goal*: To detect intrusion of a computer network
- *Approach*:
  - Define and develop a model for normal user behavior on the computer network
  - Continuously monitor behavior of users to check if it deviates from the defined normal behavior
  - Raise an alarm, if such deviation is found

# 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

(A B) (C) ⟶ (D E)

# Sequential Pattern Discovery: Telecommunication Alarm Logs

- Telecommunication alarm logs
  - (Inverter_Problem Excessive_Line_Current) (Rectifier_Alarm) → (Fire_Alarm)

# Sequential Pattern Discovery: Point of Sale Up Sell / Cross Sell

Point of sale transaction sequences

- Computer bookstore
  - (Intro_to_Visual_C) (C++ Primer) → (Perl_For_Dummies, Tcl_Tk)
- Athletic apparel store
  - (Shoes) (Racket, Racket ball) → (Sports_Jacket)