What is (not) Data Mining?

- **What is not Data Mining?**
  - data base or Web type quires such as "Who is the president of USA in 2011?"

- **What is Data Mining?**

Data analysis leading to the observations such as

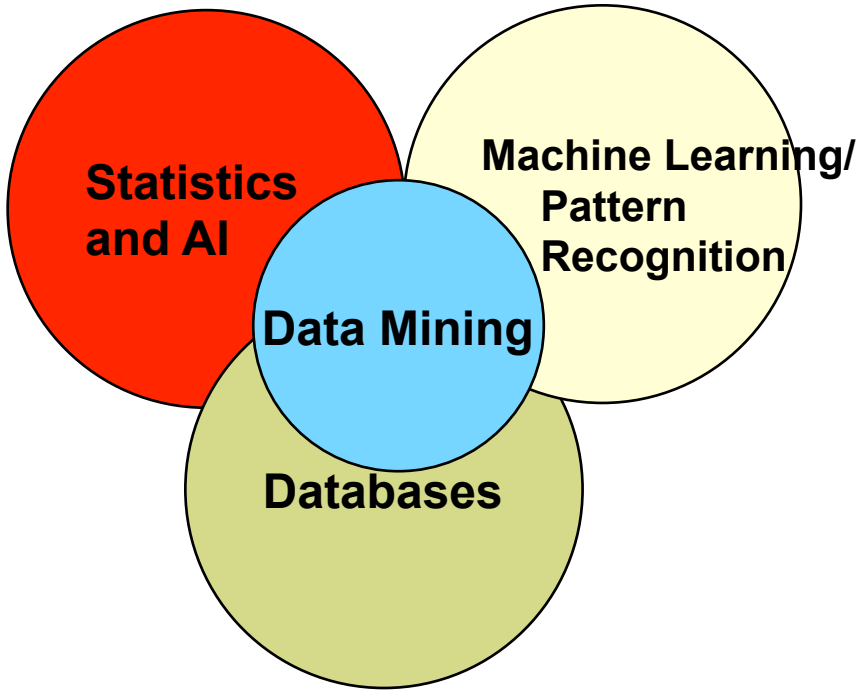
  - "smoking is a primary cause of increase in lung cancer incidents"

Databases, Data Mining, Machine Learning and Statistics and their relationships

Data mining exploits concepts from databases (scalability, semantics), AI (heuristics), Statistics (significance analysis, distributions, etc.), Machine learning (various clustering, classification, regression, etc.) and Pattern mining (image processing).

Some of the goals of Data Mining are:

- to handle large, high dimensional and complex data (relational, time series, graphical, text, etc.)
- to discover interesting patterns that can help understand underlying process
- to discover knowledge that is useful/ actionable



Data Mining Tasks involve

Prediction Methods (supervised learning methods)

- Use some variables to predict unknown or future values of other variables.
  - Classification
  - Regression
  - Deviation Detection (outlier detection, anomaly detection)

Description Methods (unsupervised learning)

- Find human-interpretable patterns that describe the data.
  - Clustering
  - Association Rule Discovery (relationships among features)
  - Sequential Pattern Discovery

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

What is classification?

Given a collection of training data

- Each item of the data contains a set of *attributes (features)*, at least one of the attributes is the *class*.

The goal is to a *model* for class attributes (dependent variables) as a function of the values of other attributes (independent or decision variables).

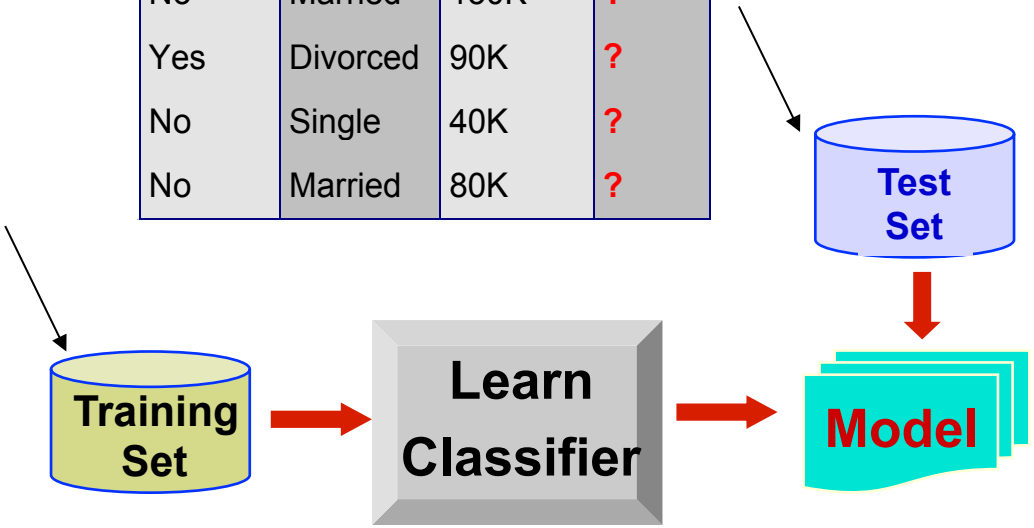
The discovered model (function) is used to predict the label of a previously unseen item.

- For determining how good the discovered model is a *test set* is used. Usually, the given data set is partitioned into two disjoint training and test sets. The training set is used to build the model and the test set is used to validate it by for example the % of the time the class label is correctly discovered.

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: 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

Fraud Detection

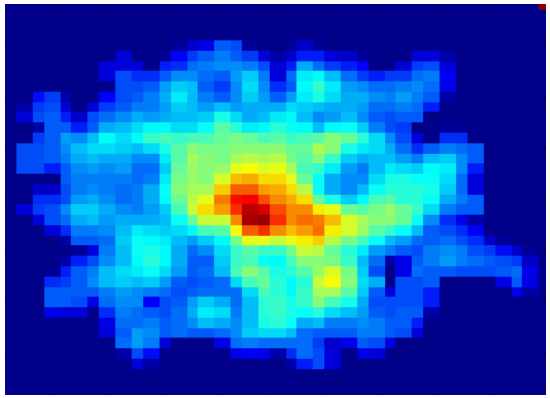
- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  - Use credit card transactions and the information on its account-holder as attributes. When does a customer buy, what does he buy, how often he pays on time, etc.
  - Label past transactions as fraud or fair transactions by an expert. This forms the class attribute.
  - Learn a model for the class of the transactions using a learning technique.
  - Use this model to detect fraud by observing new credit card transactions on an account.

Classification: Application 3

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

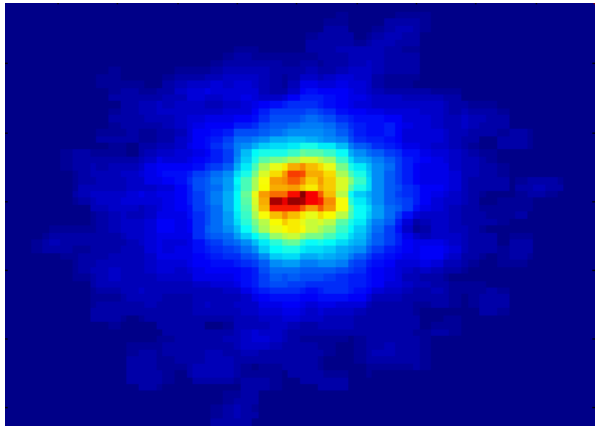
Classifying Galaxies

Courtesy: http://aps.umn.edu

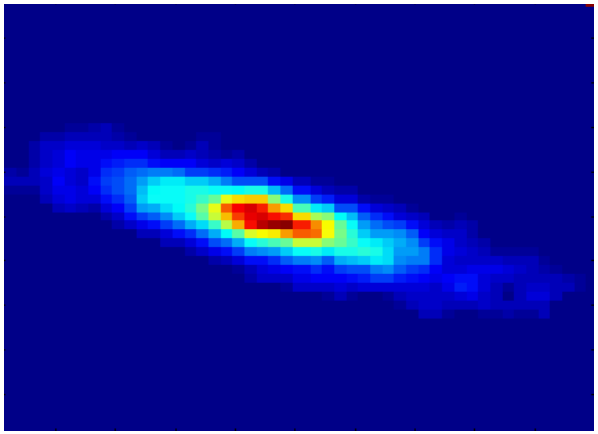
Class:
- **Stages of Formation**

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

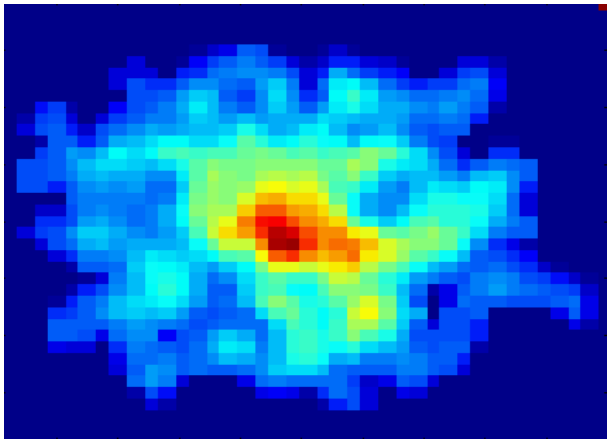
Early



Intermediate



Late



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

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. We need define appropriate similarity function suitable for the application!
- Data points in separate clusters are less similar to one another.

Similarity Measures:

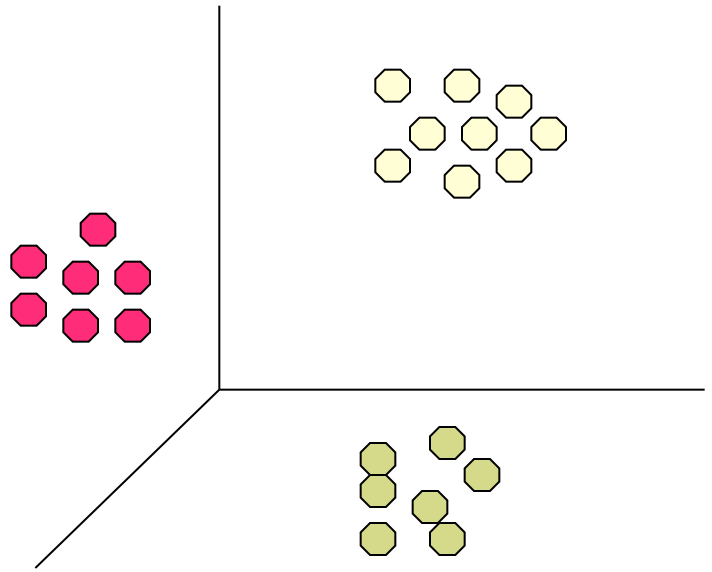
- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures, e.g., Cosine measure for documents as we studied in Information Retrieval.

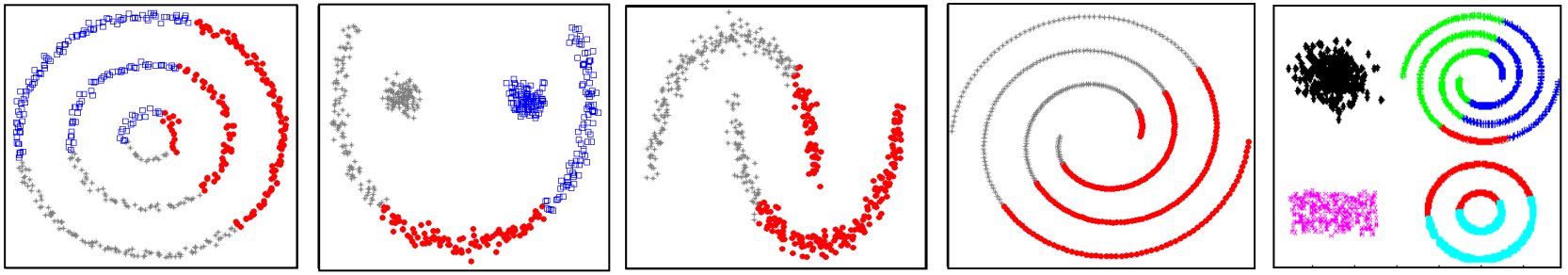
Illustrating Clustering

- Euclidean Distance Based Clustering in 3-D space.

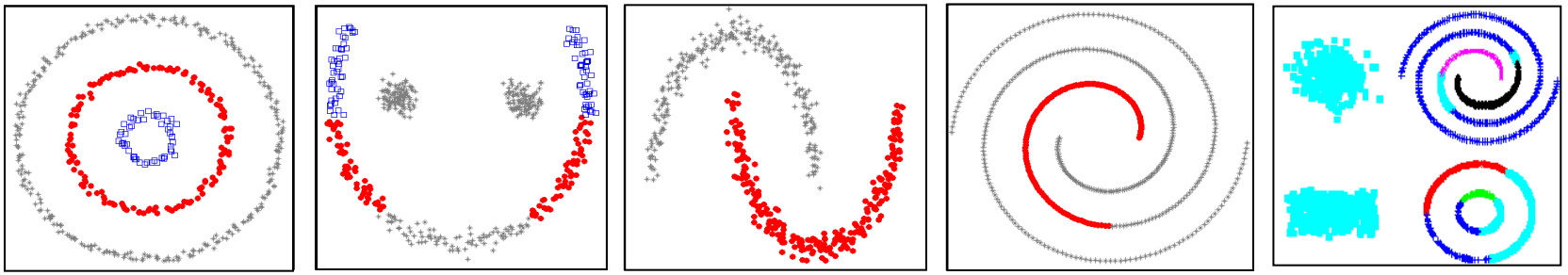
Intracluster distances are minimized

Intercluster distances are maximized

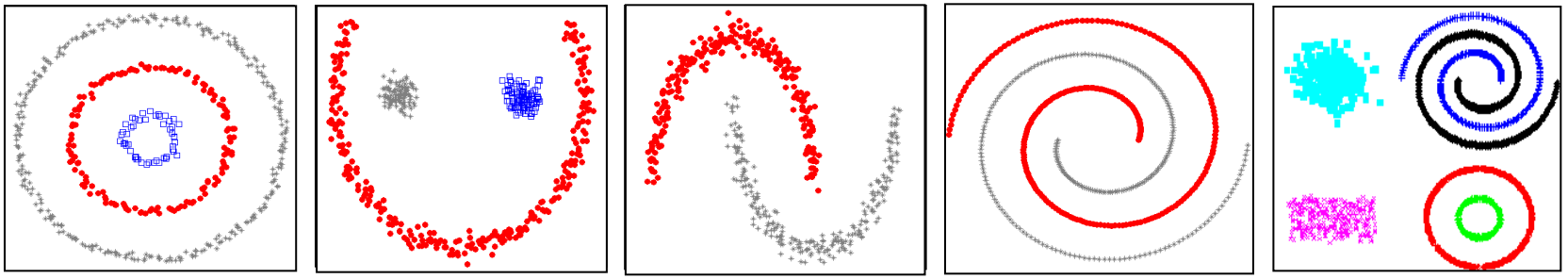




K-means



Spectral clustering



Path based method

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 strategy.
- 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: 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

Clustering of S&P (Standard & Poor's) 500 Stock Data

- Observe Stock Movements every day.
- 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

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}
Confidence for the rule = 3/4
Support for the rule = 3 / 5
{Diaper, Milk} --> {Beer}
Confidence = 2 / 3 confidence
Support for the rule = 2 / 5

Association Rule Discovery: 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!
- Bagels in antecedent *and* Potato chips in consequent => Can be used to co-locate Bagels and Potato Chips to further boost the sales of both products.
  Bagels in antecedent *and* Potato chips in consequent => The store may reduce the price of Bagels to actually increase the profit!

Association Rule Discovery: Application 2

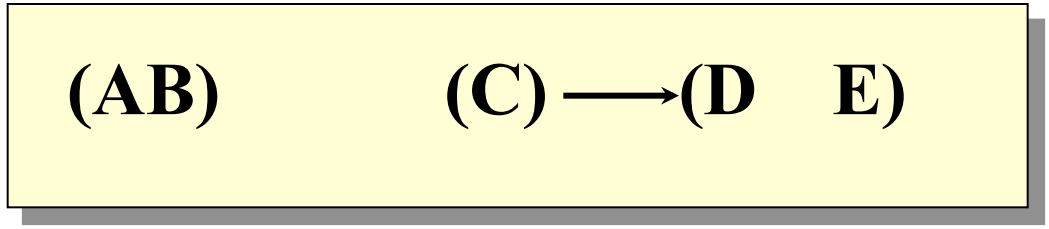
Inventory Management:

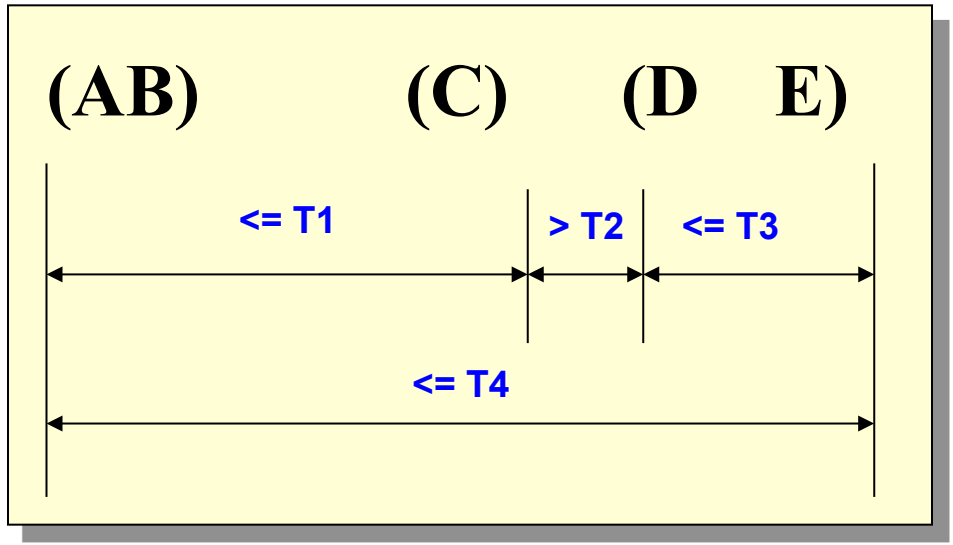
- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
- Approach: Process the data on tools and parts required in previous repairs at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

Sequential Pattern Discovery: Definition

Given 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.





Sequential Pattern Discovery: Examples

In telecommunications alarm logs,

(Inverter_Problem) (Excessive_Line_Current) (Rectifier_Alarm)→(Fire_Alarm)

In point-of-sale transaction sequences,

Computer Bookstore:

(Intro_To_Visual_C) (C++_Primer) → (Perl_for_dummies,Tcl_Tk)

Athletic Apparel Store:

(Shoes) (Racket, Racketball) → (Sports_Jacket)

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:

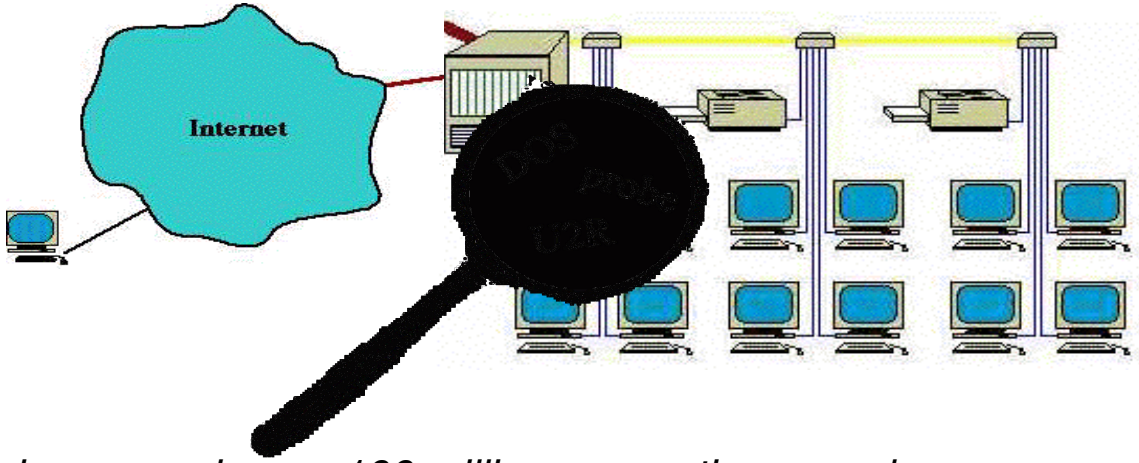
- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.

Deviation/Anomaly Detection

Detect significant deviations from normal behavior

Applications:

- Credit Card Fraud Detection
- Network Intrusion Detection



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

Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data
- Incremental learning
- How to combine domain knowledge with Data Mining
- Understandability of the models discovered