DATA MINING - 1DL105, 1DL111

Fall 2007

An introductory class in data mining

http://www.it.uu.se/edu/course/homepage/infoutv/ht07

Kjell Orsborn
Uppsala Database Laboratory
Department of Information Technology, Uppsala University,
Uppsala, Sweden

UPPSALA
UNIVERSITET

Personell

- Kjell Orsborn, lecturer, examiner
  - email: kjell.orsborn@it.uu.se, phone: 471 1154, room: 1321
- Per Gustafsson, course assistant
  - email: per.gustafsson@it.uu.se, phone 471 3155, room: 1310
- Tobias Lindahl, course assistant
  - email: tobias.lindahl@it.uu.se, phone: 471 3168, room 1310

Course contents (preliminary)

- Course intro - introduction to data mining
- Overview of data mining techniques
- Data in Data Mining
- Classification
- Clustering
- Association rules
- Mining sequential patterns
- Web content mining
- Search engines
- Data mining and privacy
- Exploring data (if time admits)

Course contents continued …

- Tutorials:
  - Tutorial 1 on MATLAB and assignment 1
  - Tutorial 2 on assignment 2
  - Tutorial 3 on assignment 3
  - Tutorial 4 on assignment 4
- Assignments
  - Assignment 1 - Classification using K Nearest Neighbours (KNN)
  - Assignment 2 - Clustering using K-Means vs. DBSCAN
  - Assignment 3 - Association Rule Mining
  - Assignment 4 - Web Searching using the HITS algorithm

Introduction to Data Mining

(Tan, Steinbach, Kumar ch. 1)

Kjell Orsborn

Department of Information Technology
Uppsala University, Uppsala, Sweden

Data Mining

- The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions, (Simoudis, 1996).

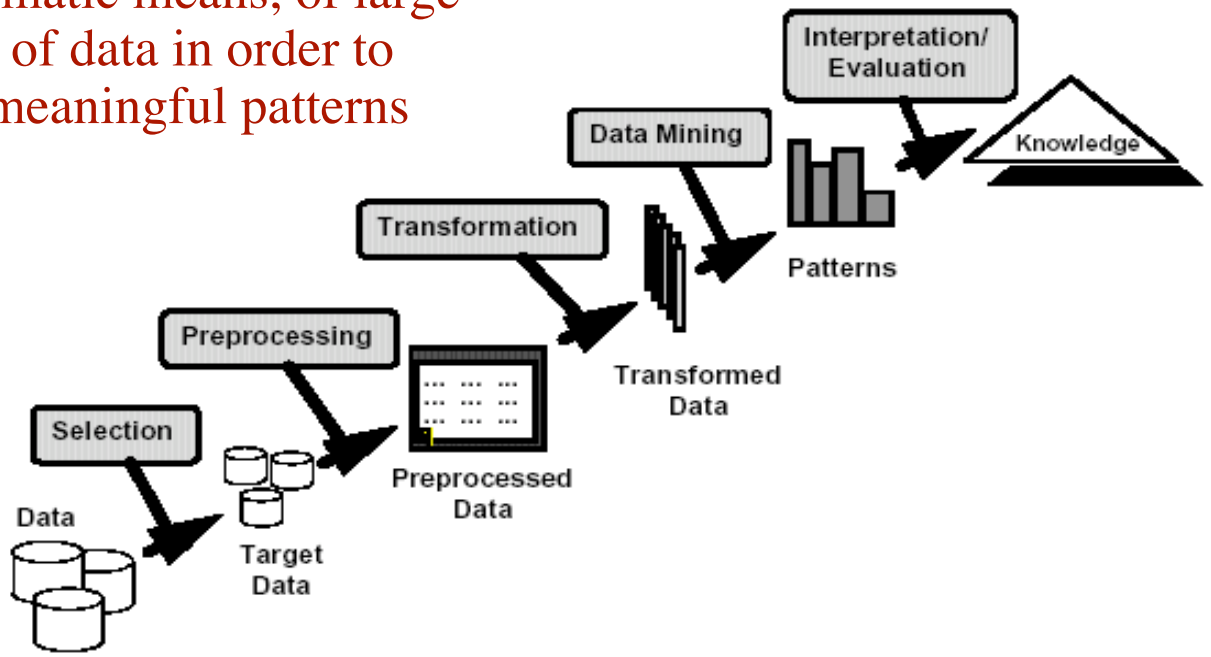
- Involves the analysis of data and the use of software techniques for finding *hidden* and *unexpected* patterns and relationships in sets of data; in contrast to information and knowledge that are already intuitive.
- Patterns and relationships are identified by examining the underlying rules and features in the data.
- Tends to work from the data up and most accurate results normally require large volumes of data to deliver reliable conclusions.
- Data mining can provide huge paybacks for companies who have made a significant investment in data warehousing.
- Relatively new technology, however already used in a number of industries.

Query examples

- Database query
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk
- Data mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (clustering)
  - Find all items which are frequently purchased with milk. (association rules)

What is 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
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

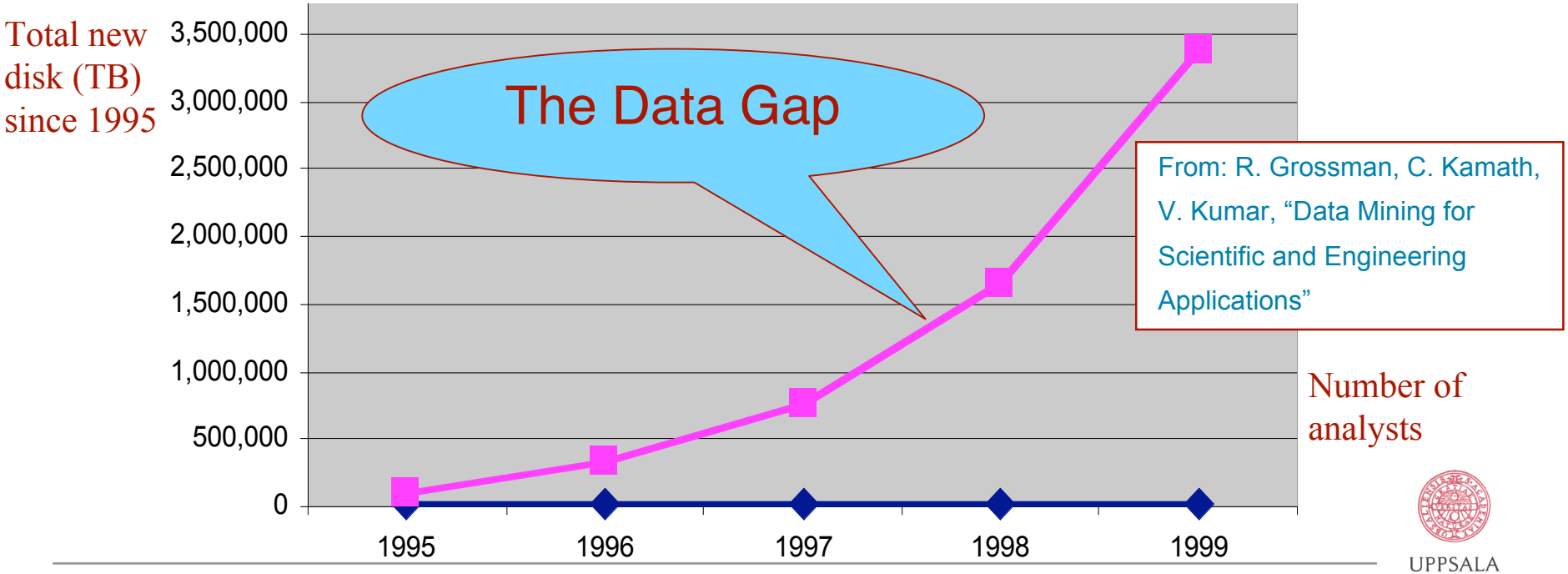


What is data mining - continued …

- Data mining - a misnomer?
  - Alternative names: knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

Why data mining?

- The explosive growth of data: from terabytes to petabytes
  - Data collection from automated data collection tools, database systems, web, e-commerce, transactions, stocks, remote sensing, bioinformatics, scientific simulation, computerized society, news, digital cameras, …
  - Human analysts may take weeks to discover useful information
  - Much of the data is never analyzed at all



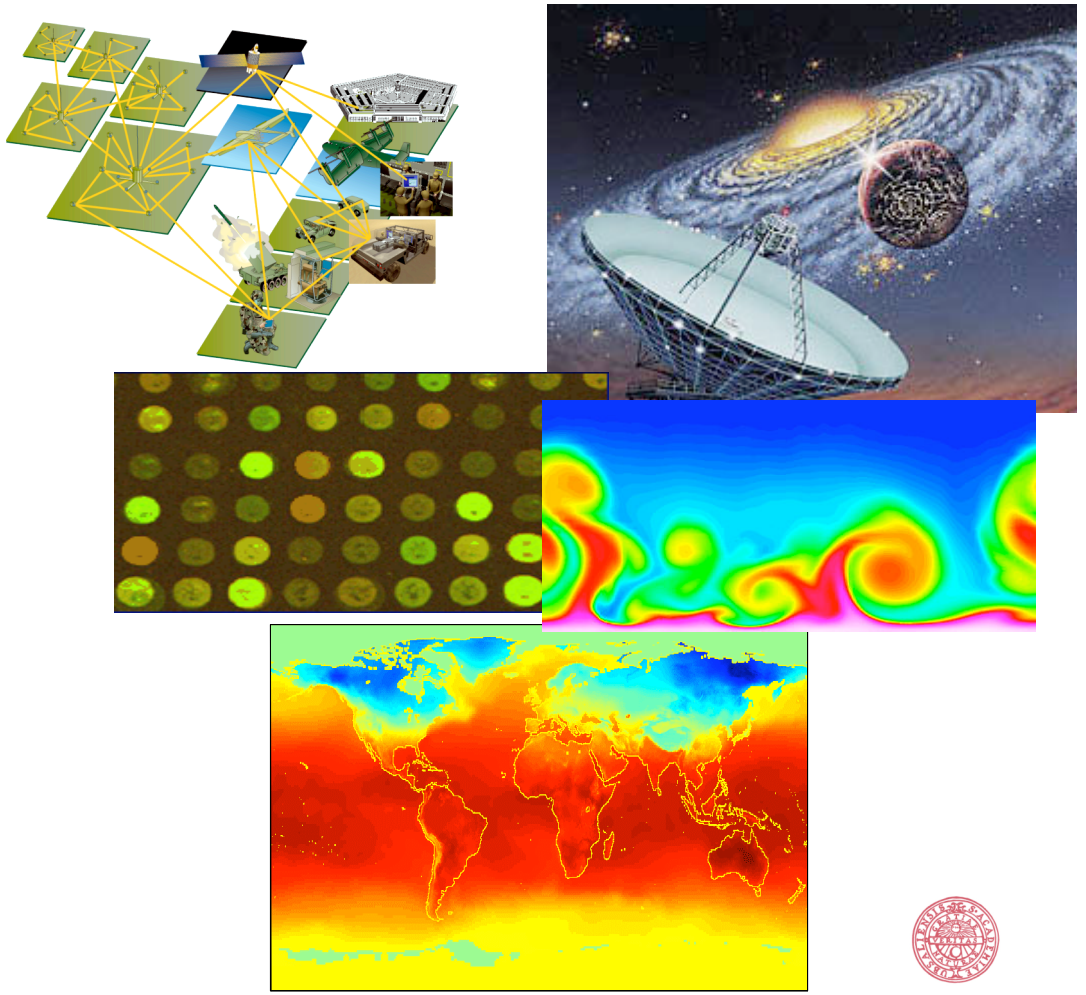
Why mine data (commercial viewpoint)?

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank & credit card transactions

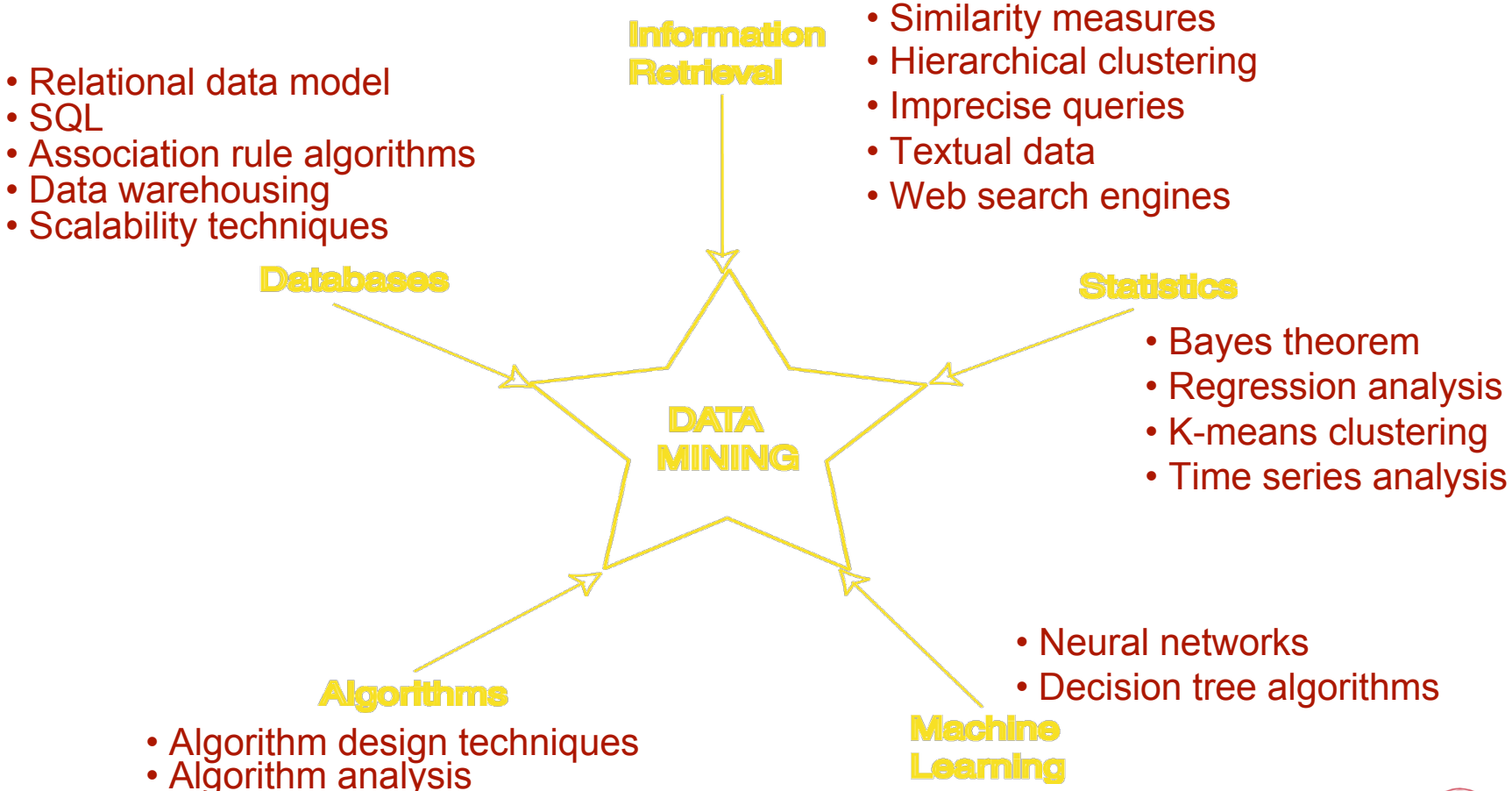
- Computers have become cheaper and more powerful
- Competitive pressure is strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

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



Data mining and related areas (Dunham, 2003)

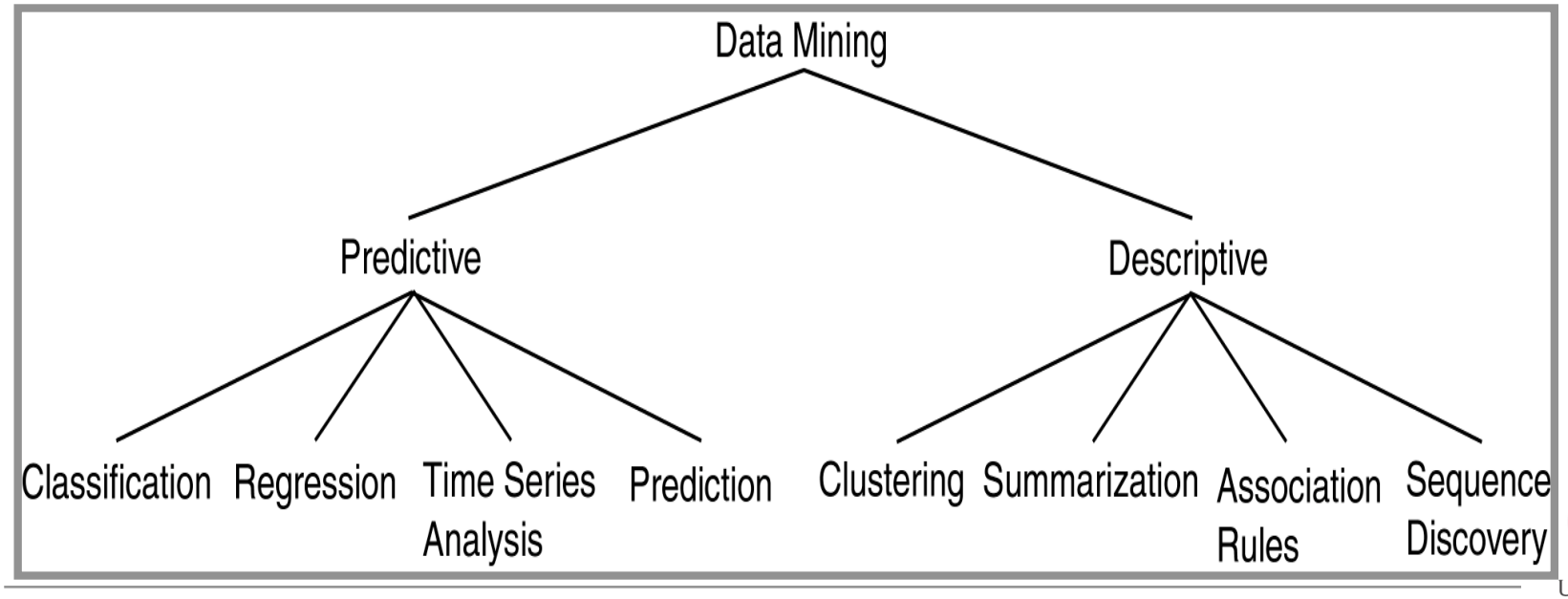


Why not traditional data analysis?

- Tremendous amount of data
  - Algorithms must be highly scalable to handle such as tera-bytes of data
- High-dimensionality of data
  - Micro-array may have tens of thousands of dimensions
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Heterogeneous databases and legacy databases
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Data mining tasks

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



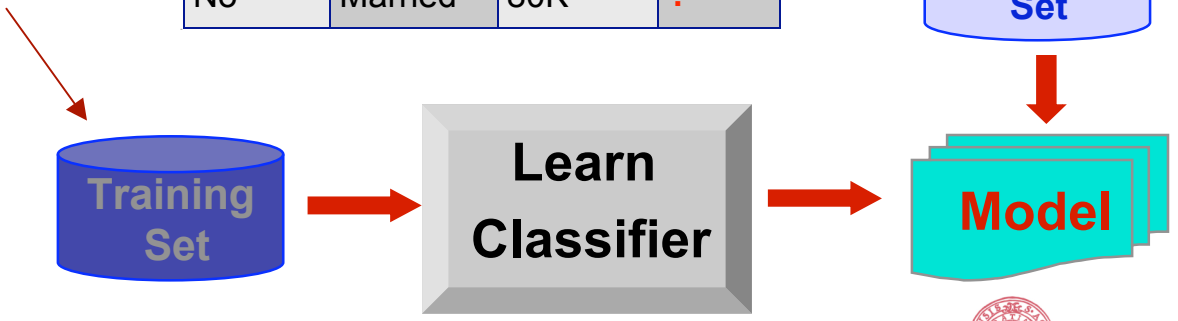
Classification - definition

- 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	?



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. 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 and 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.

From [Berry & Linoff] Data Mining Techniques, 1997

Classification - application 4

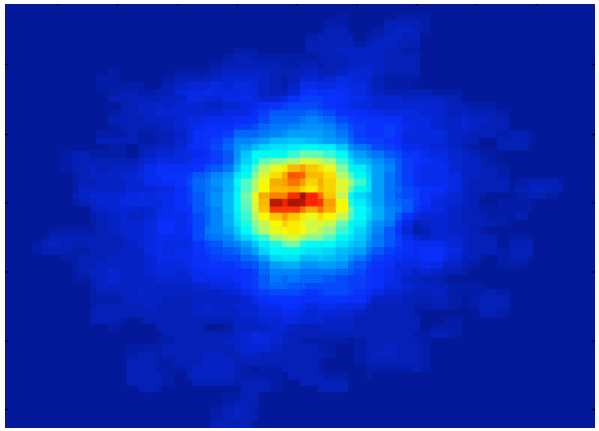
- Sky Survey Cataloging
  - Goal: to predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (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

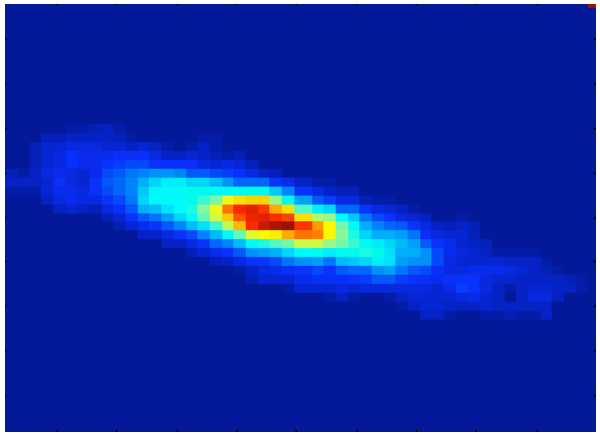
Early



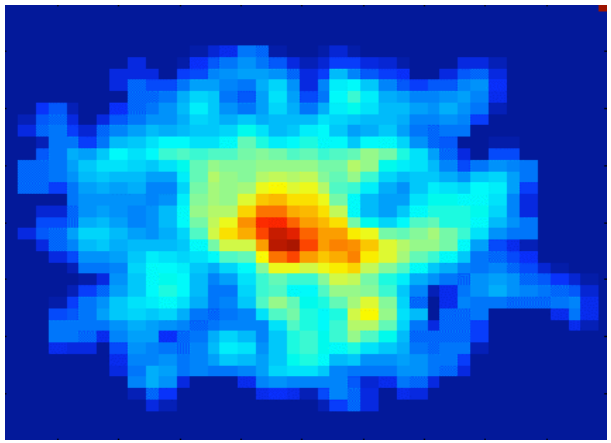
Class:
- **Stages of Formation**

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**

Clustering - definition

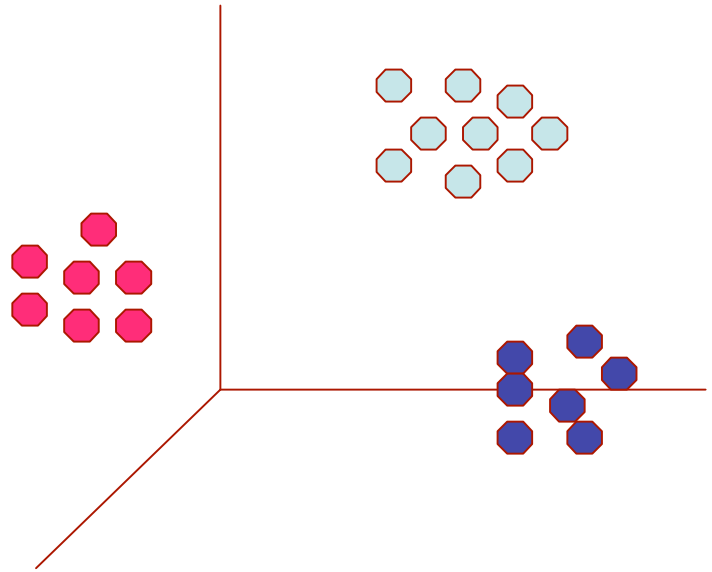
- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - Data points in one cluster are more similar to one another.
  - Data points in separate clusters are less similar to one another.
- Similarity Measures:
  - Euclidean distance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

UPPSALA UNIVERSITET

Clustering of S&P 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.
  - We used association rules to quantify a similarity measure.

	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

UPPSALA UNIVERSITET

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}

UPPSALA UNIVERSITET

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!

Association rule discovery: application 2

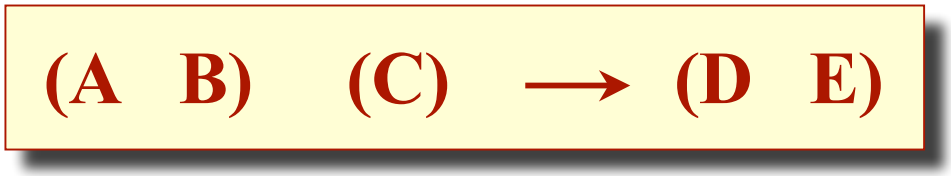
- Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule --
    - If a customer buys diaper and milk, then he is very likely to buy beer.
    - So, don't be surprised if you find six-packs stacked next to diapers!

Association rule discovery: application 3

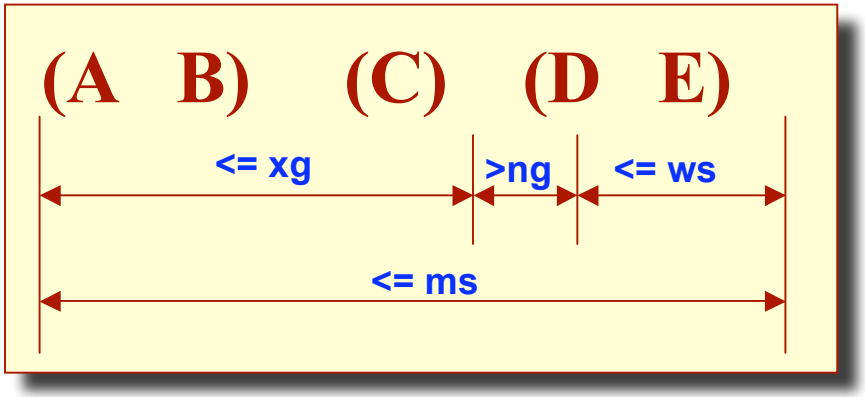
- Inventory management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required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

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 disovering patterns. Event occurrences in the patterns are governed by timing constraints.



Sequential pattern discovery examples

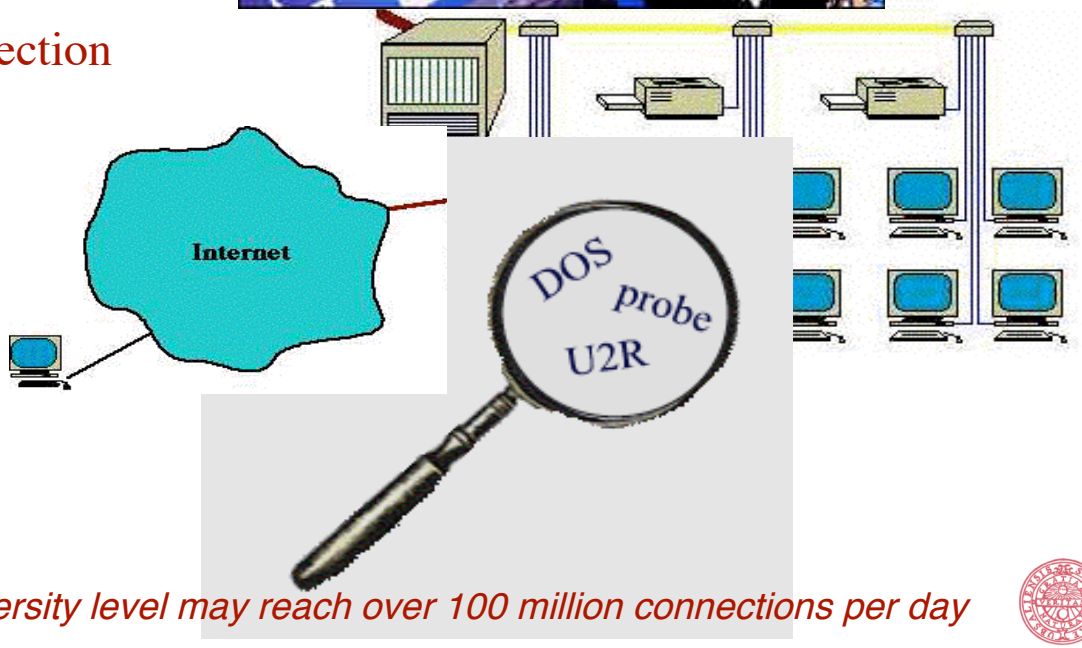
- In telecommunications alarm logs,
  - (Inverter_Problem Excessive_Line_Current)
  - (Rectifier_Alarm) --> (Fire_Alarm)
- In point-of-sale transaction sequences,
  - Computer Bookstore:
    - (Intro_To_Visual_C) (C++_Primer) --> (Perl_for_dummies,Tcl_Tk)
  - Athletic Apparel Store:
    - (Shoes) (Racket, 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 advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

Deviation or 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