Data Mining: Introduction

Lecture Notes

by Márton Ispány

based on

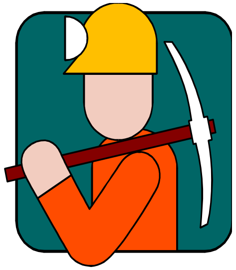
Introduction to Data Mining

by

Tan, Steinbach, Kumar

What is Data Mining?

- After years of data mining there is still no unique answer to this question.

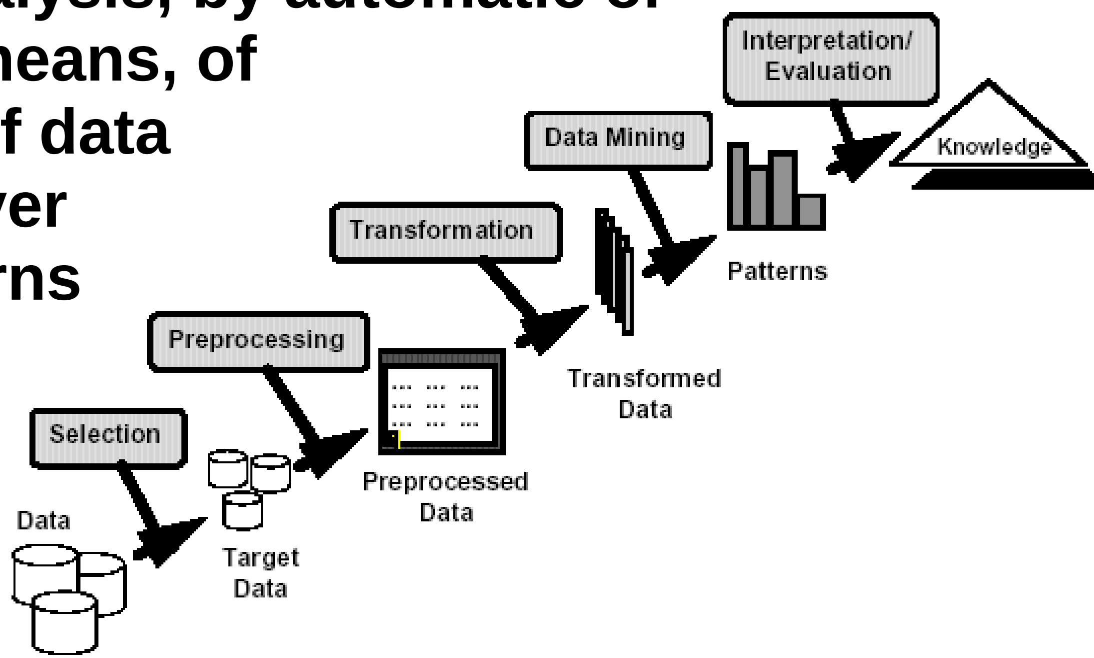


- **Definitions**:

Data mining is the use of efficient techniques for the analysis of very large collections of data and the extraction of useful and possibly unexpected patterns in data.

What is Data Mining?

- **Definitions:**
  - **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**



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

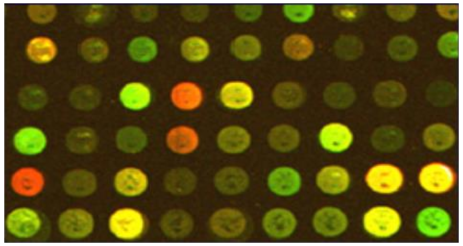
- Computers have become cheaper and more powerful

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

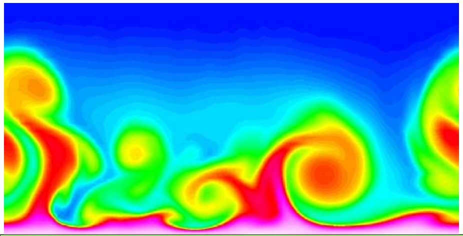
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)

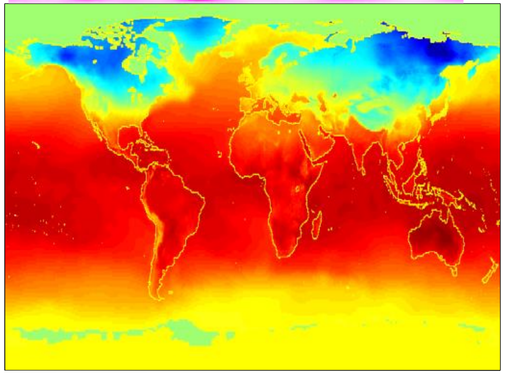
  - remote sensors on a satellite
  - telescopes scanning the skies



  - 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

Why do we need Data Mining?

- Really, really huge amounts of raw data!!
  - In the digital age, TB of data is generated by the second
    - Mobile devices, digital photographs, web documents.
    - Facebook updates, Tweets, Blogs, User-generated content
    - Transactions, sensor data, surveillance data
    - Queries, clicks, browsing
  - Cheap storage has made possible to maintain this data
- Need to analyze the raw data to extract knowledge

Why do we need Data Mining?

- "The data is the computer"
  - Large amounts of data can be more powerful than complex algorithms and models
    - Google has solved many Natural Language Processing problems, simply by looking at the data
    - Example: misspellings, synonyms
  - Data is power!
    - Today, the collected data is one of the biggest assets of an online company
      - Query logs of Google
      - The friendship and updates of Facebook
      - Tweets and follows of Twitter
      - Amazon transactions
  - We need a way to harness the collective intelligence

The data is also very complex

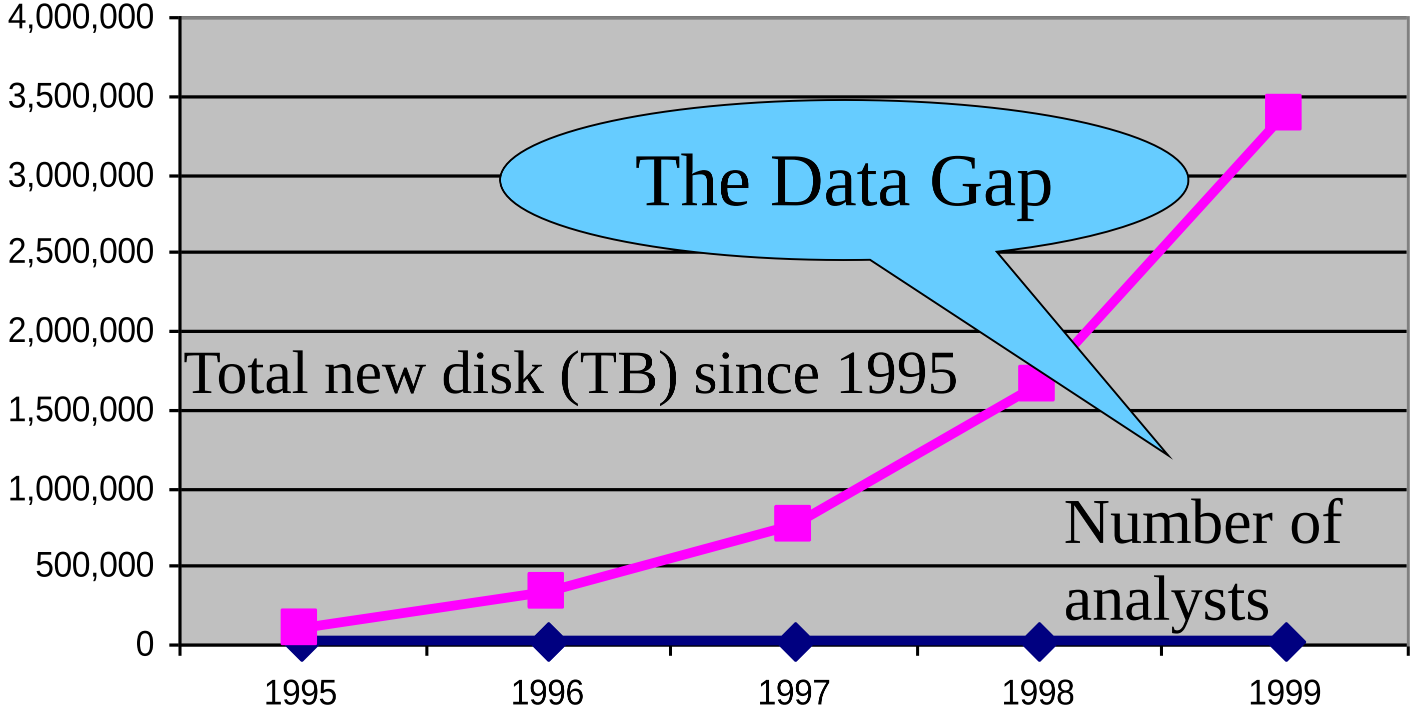
- Multiple types of data: tables, time series, images, graphs, etc

- Spatial and temporal aspects

- Interconnected data of different types:
  - From the mobile phone we can collect, location of the user, friendship information, check-ins to venues, opinions through twitter, images though cameras, queries to search engines

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”

Knowledge Discovery from Databases (KDD)

- Data gathering: creating target data sets
- Data cleaning
- Data integration
- Data reduction and projection
- Data transformation
- Data mining
- Interpretation, evaluation
- Knowledge representation

Step 2-4: Data Warehousing

Step 5-6: Data mining

Example: transaction data

- Billions of real-life customers:
  - WALMART: 40M transactions per day, 40Pbyte
  - AT&T 300 M calls per day
  - Credit card companies: billions of transactions per day.

- The point cards allow companies to collect information about specific users

Example: document data

- Web as a document repository: estimated 1.33 billions of web pages

  http://www.internetlivestats.com/total-number-of-websites/

- Wikipedia: 5.5 million articles (and counting)

  https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

- Online news portals: steady stream of 100’s of new articles every day
- Twitter: 500 million tweets every day

  http://www.internetlivestats.com/twitter-statistics/#trend

Example: network data

- Web: 1.335 billion pages linked via hyperlinks
- Facebook: 2.2 billion users
- Twitter: 330 million users
- Instant messenger: 203 million users
- Blogs: 250 million blogs worldwide, presidential candidates run blogs

Example: genomic sequences

- http://www.1000genomes.org/page.php
- Full sequence of 1000 individuals
- $3*10^9$ nucleotides per person → $3*10^{12}$ nucleotides
- Lots more data in fact: medical history of the persons, gene expression data

Behavioral data

- Mobile phones today record a large amount of information about the user behavior
  - GPS records position
  - Camera produces images
  - Communication via phone and SMS
  - Text via facebook updates
  - Association with entities via check-ins

- Amazon collects all the items that you browsed, placed into your basket, read reviews about, purchased.

- Google and Bing record all your browsing activity via toolbar plugins. They also record the queries you asked, the pages you saw and the clicks you did.

- Data collected for millions of users on a daily basis

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,)

Why data mining again?

- Commercial point of view
  - Data has become the key competitive advantage of companies
    - Examples: Facebook, Google, Amazon
  - Being able to extract useful information out of the data is key for exploiting them commercially.
- Scientific point of view
  - Scientists are at an unprecedented position where they can collect TB of information
    - Examples: Sensor data, astronomy data, social network data, gene data
  - We need the tools to analyze such data to get a better understanding of the world and advance science
- Scale (in data size and feature dimension)
  - Why not use traditional analytic methods?
  - Enormity of data, curse of dimensionality
  - The amount and the complexity of data does not allow for manual processing of the data. We need automated techniques.

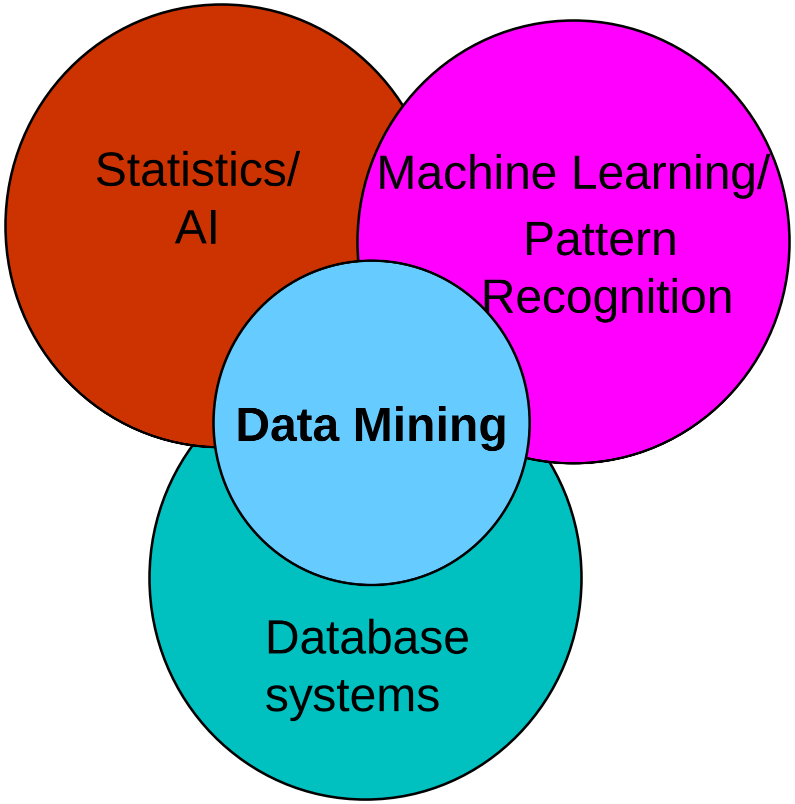
What is Data Mining again?

- "Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data analyst" (Hand, Mannila, Smyth)

- "Data mining is the discovery of models for data" (Rajaraman, Ullman)
  - We can have the following types of models
    - Models that explain the data (e.g., a single function)
    - Models that predict the future data instances.
    - Models that summarize the data
    - Models the extract the most prominent features of the data.

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



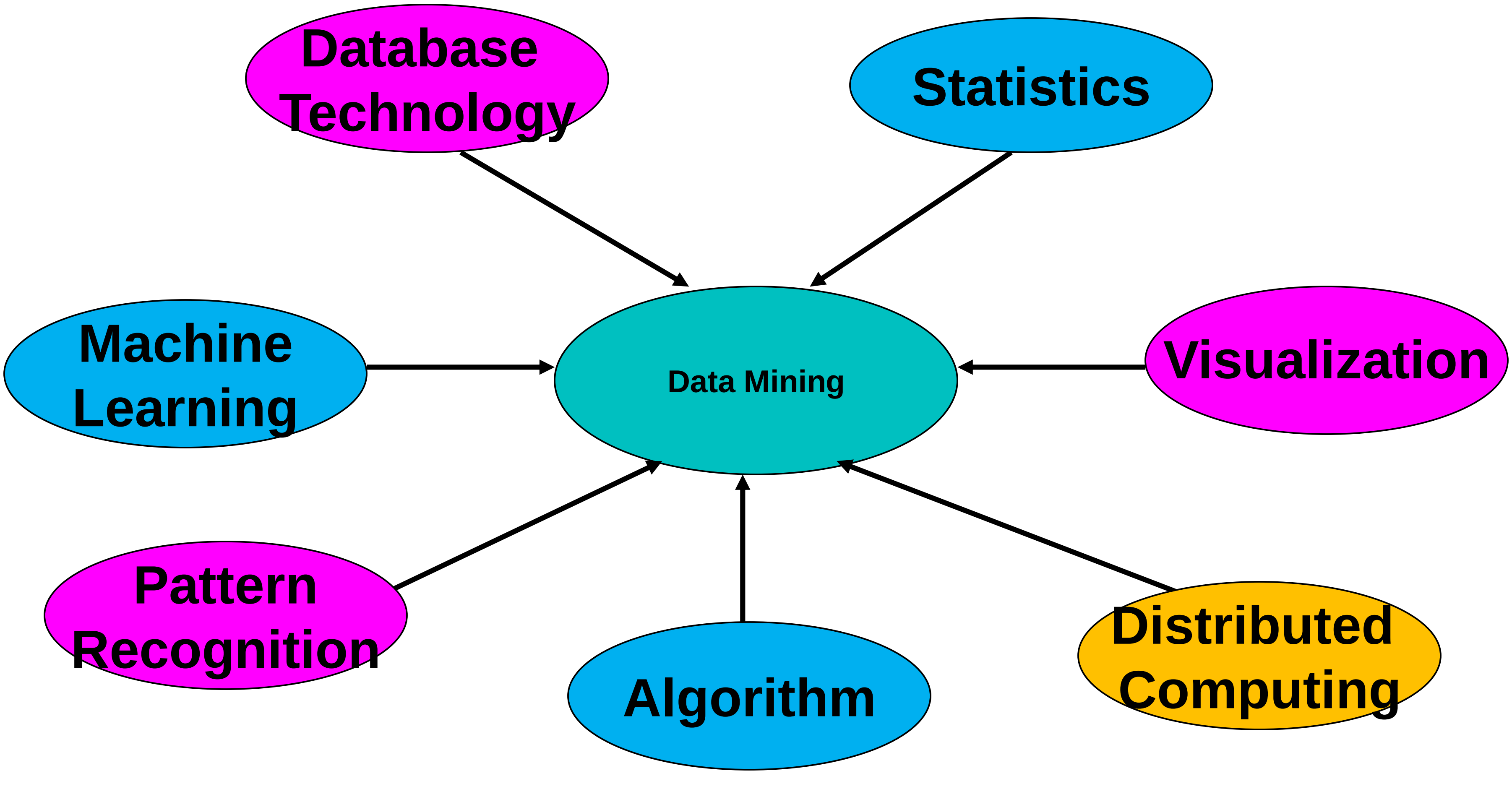
Cultures

- Databases: concentrate on large-scale (non-main-memory) data.
- AI (machine-learning): concentrate on complex methods, small data.
  - In today's world data is more important than algorithms
- Statistics: concentrate on models.

Models vs. Analytic Processing

- To a database person, data-mining is an extreme form of analytic processing – queries that examine large amounts of data.
  – Result is the query answer.
- To a statistician, data-mining is the inference of models.
  – Result is the parameters of the model.

Data Mining: Confluence of Multiple Disciplines



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.

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

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

Classification: Definition

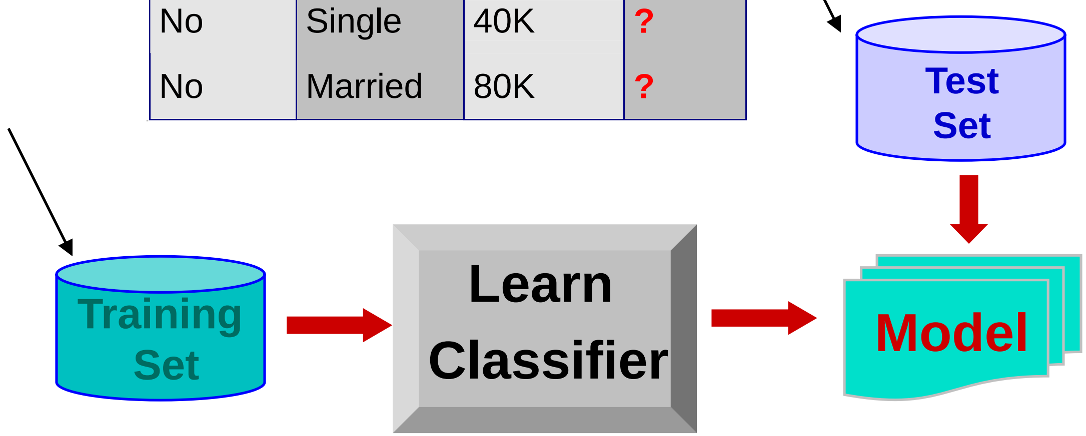
- Given a collection of records (*training set*)
  - Each record contains a set of *attributes*, one of the attributes is the *class*.
- Find a *model* for class attribute as a function of the values of other attributes.
- Goal: previously unseen records should be assigned a class as accurately as possible.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

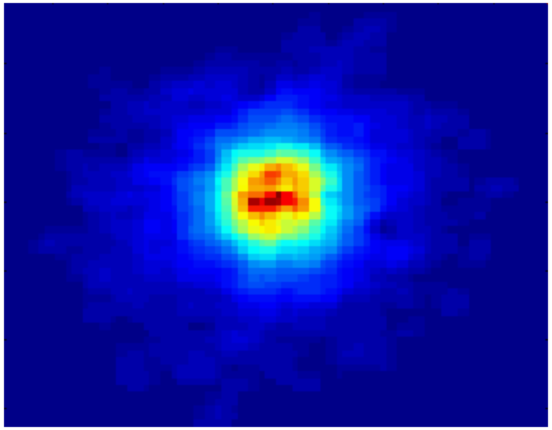
- Sky Survey Cataloging
  - Goal: To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
    - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes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

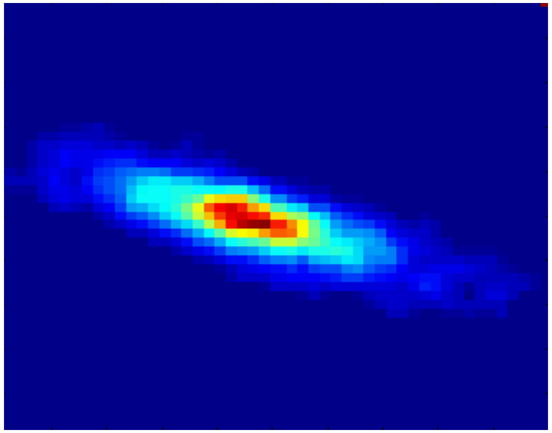
Early



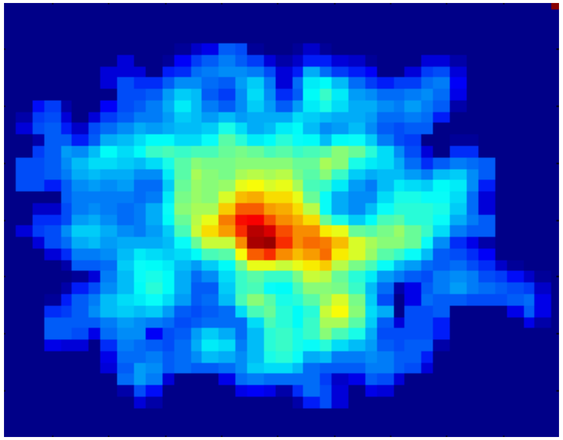
Class:
- **Stages of Formation**

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

Intermediate



Late



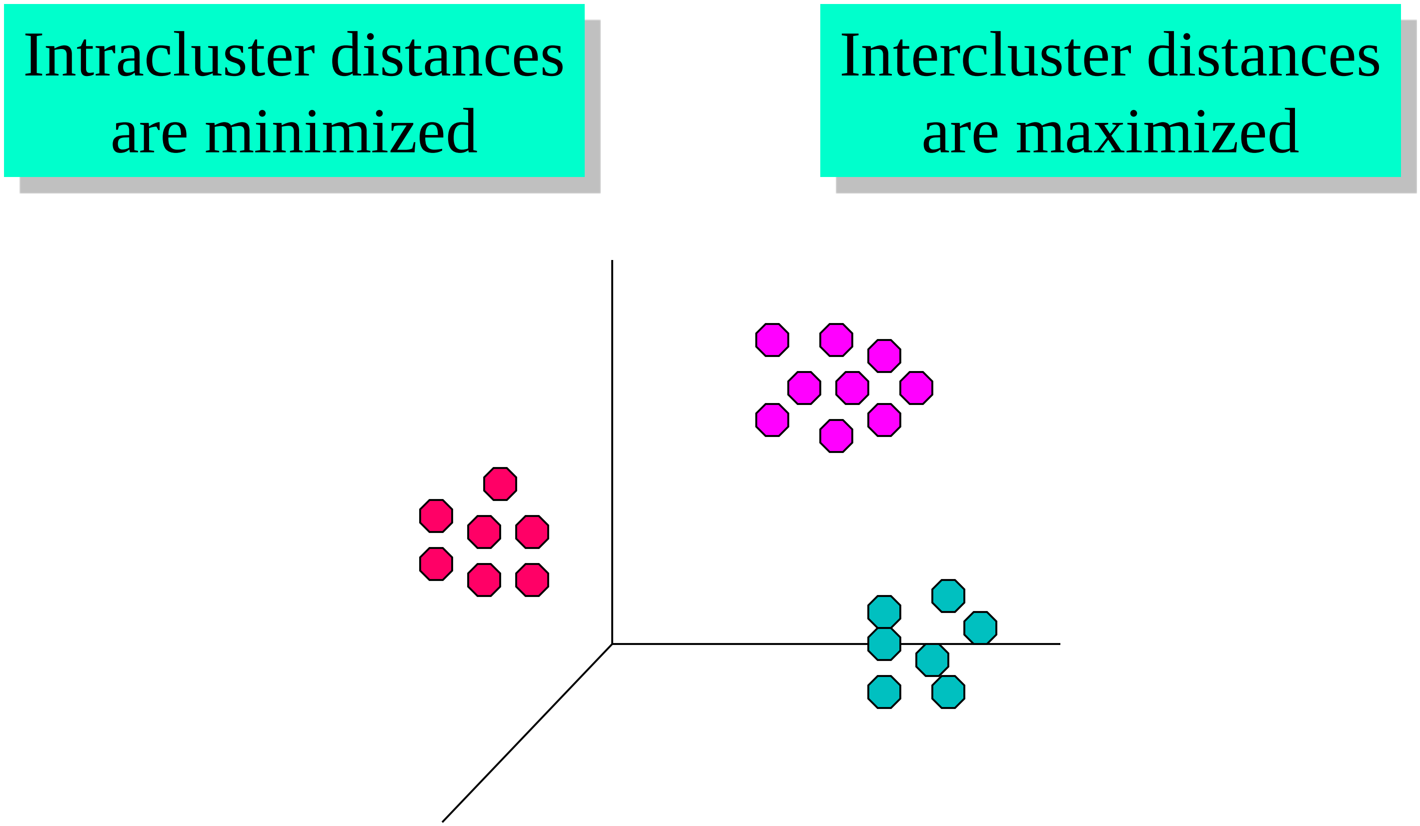
Data Size:
- **72 million stars, 20 million galaxies**
- **Object Catalog: 9 GB**
- **Image Database: 150 GB**

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.



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

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

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}

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

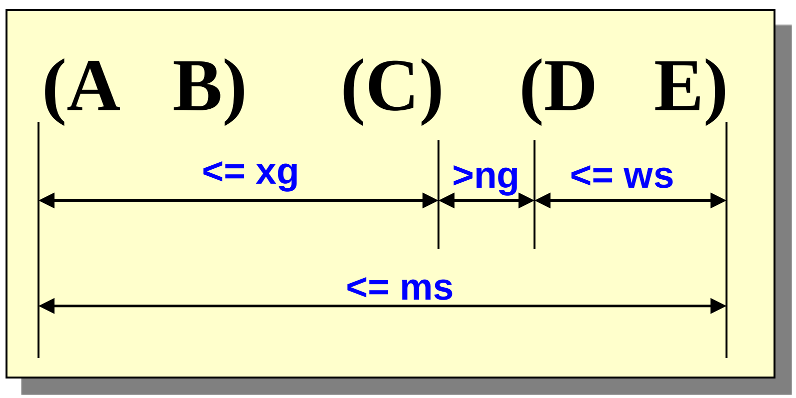
- Inventory Management:
  - Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.
  - Approach: Process the data on tools and parts required 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.

(A B) (C) → (D E)

- 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)

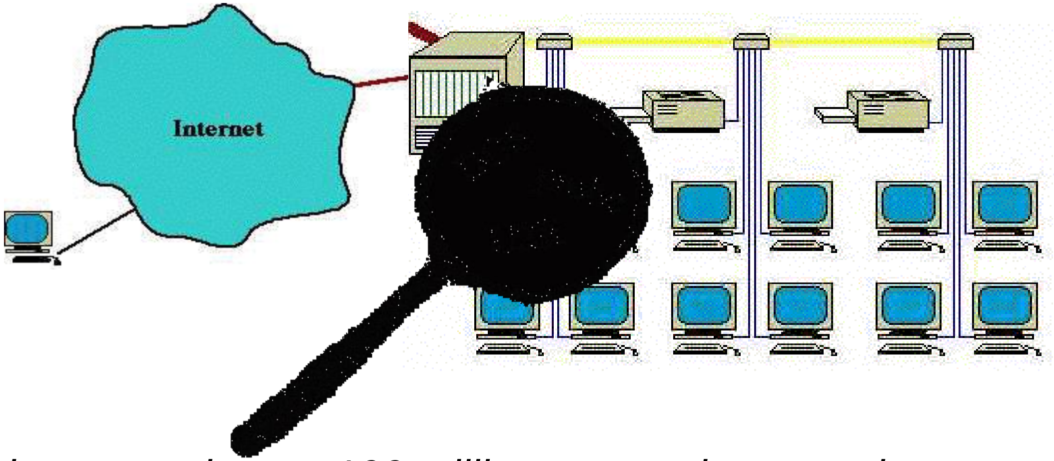
Regression

- 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/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