



Data Mining course

Master in Information Technologies
Enginyeria Informàtica

Tomàs Aluja. LIAM – EIO. UPC

Lluís Belanche LSI. UPC



UNIVERSITAT POLITÈCNICA
DE CATALUNYA

Topics

- Introduction to Data Mining
- Preprocess
- Finding profiles
- Visualisation techniques
- Clustering
- Association rules
- Decision trees
- Parametric models
- Non parametric models
- Neurals networks
- Support Vector Machines

Recommended Books

- Aluja T., Morineau A. **Aprender de los datos: El Análisis de Componentes Principales**, EUB, 1999.
- Hand D.J. **Construction and Assessment of Classification Rules.** , John Wiley, 1997.
- Hastie T., Tibshirani R., Friedman J. **The elements of statistical learning. Data mining, inference and prediction.** , Springer, 2001.
- Hernández Orallo J., Ramírez Quintana M.J., Ferri Ramírez C **Introducción a la Minería de Datos**, Prentice Hall, 2004.
- Witten I.H., Frank E **Data Mining**, . Morgan Kaufman Publishers, 2000.
- Berry M.J.A., Linoff G **Data Mining Techniques, for marketing, sales and costumer support**, John Wiley, 1997.
- Hand D., Mannila H., Smyth P. **Principles of Data Mining**, The MIT Press, 2001.
- Ripley B.D. **Pattern Recognition and Neural Networks.** , Cambridge University Press, 1995.
- Bishop C. M. **Neural Networks for Pattern Recognition**, Clarendon Press. Oxford, 1995.
- Cyos, K., Pedyioz, W. I Swiniaski, R. **Data Mining. Methods for Knowledge Discovery**, Kluwer, 1998.

Software Resources

<http://www.cran.r-project.org>

<http://www.kdnuggets.com/>

<http://www.cs.waikato.ac.nz/>

<http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>

<http://ses.telecom-paristech.fr/lebart/>

http://en.wikipedia.org/wiki/Data_mining

<http://www.itl.nist.gov/div898/handbook/pmd/pmd.htm>

Course Grading

- Academic assessment will be based on the grades obtained in the three practical works held during the course, plus a small test.
- Students will write a report on each practical assignment. The report may be jointly written by pairs of students. In addition, the third practical work must be presented orally and publicly.
- The test will take place the last day of the course.
- The relative importance of these three practical works are 15%, 15% and 50%, respectively and the remaining 20% is for the test.

Course Projects

- Form a 2-person group
- Practical work 1 () Write a report
- Practical work 2 () Write a report
- Practical work 3
 - Choose a “real-world” domain and define the problem (Oct 30)
 - Implement and test algorithms
 - Write a report
 - Present it orally

Trends leading to Data Flood

“We are drowning in information but starved for knowledge.
John Naisbitt, “Megatrends” (1982)

- More data is generated:
 - Bank, telecom, other business transactions ...
 - Scientific data: astronomy, genomics, etc
 - Web, text, and e-commerce

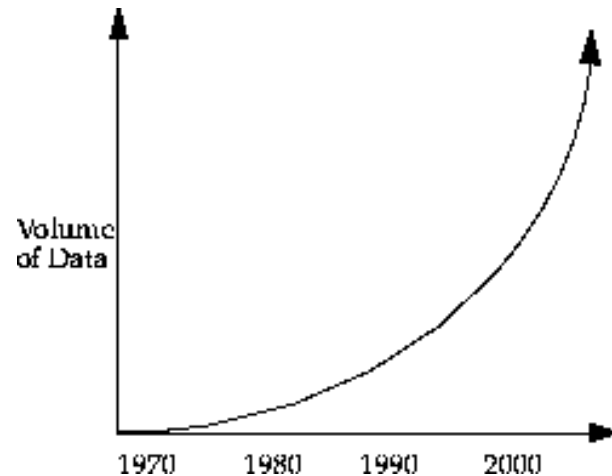
storage and analysis a big problem
so much data cannot be all stored
analysis has to be done “on the fly”,
on streaming data



Paradigm: Data contains information

Data Growth Rate

- The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster.



Very little data will ever be looked at by a human. Knowledge Discovery is **NEEDED** to make sense and use of data.

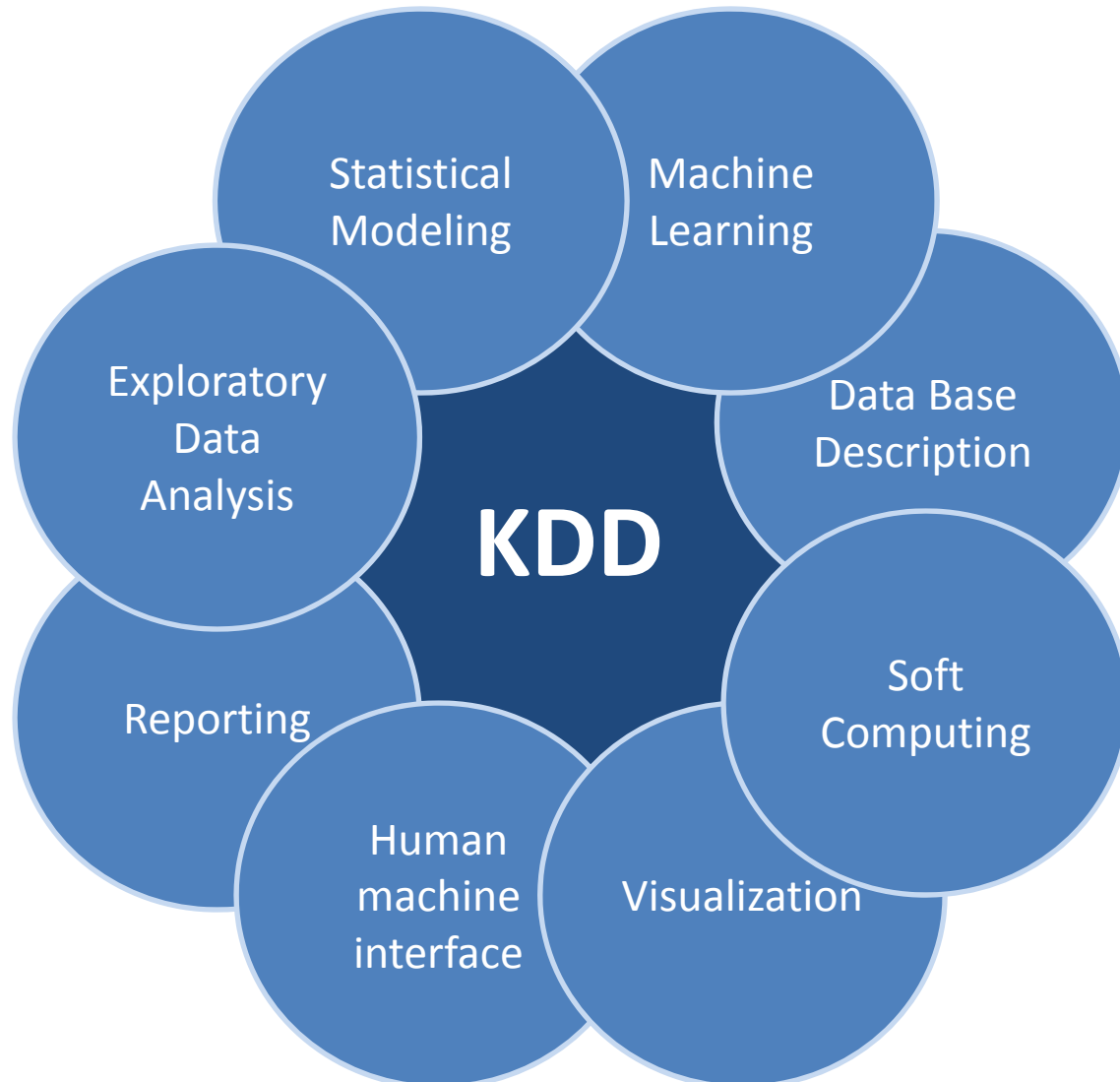
Knowledge Discovery Definition

Knowledge Discovery in Data is the *non-trivial* process of identifying

- *valid*
- *novel*
- *potentially useful*
- and ultimately *understandable patterns* in data.

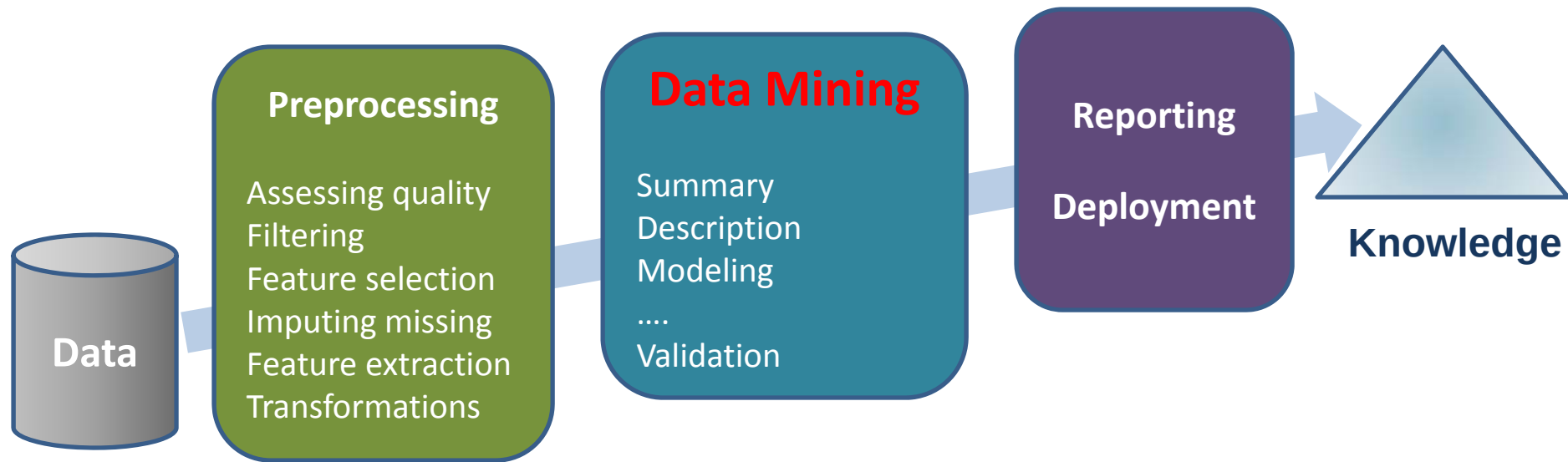
from *Advances in Knowledge Discovery and Data Mining*, Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, (Chapter 1), AAAI/MIT Press 1996

KDD and related fields



KDDDB versus DM

KDDDB encompasses from Data (DB) to Knowledge, whereas DM refers to the technical phase of applying statistical modeling or learning algorithms, but in practice DM is used indistinguishable of the whole process.

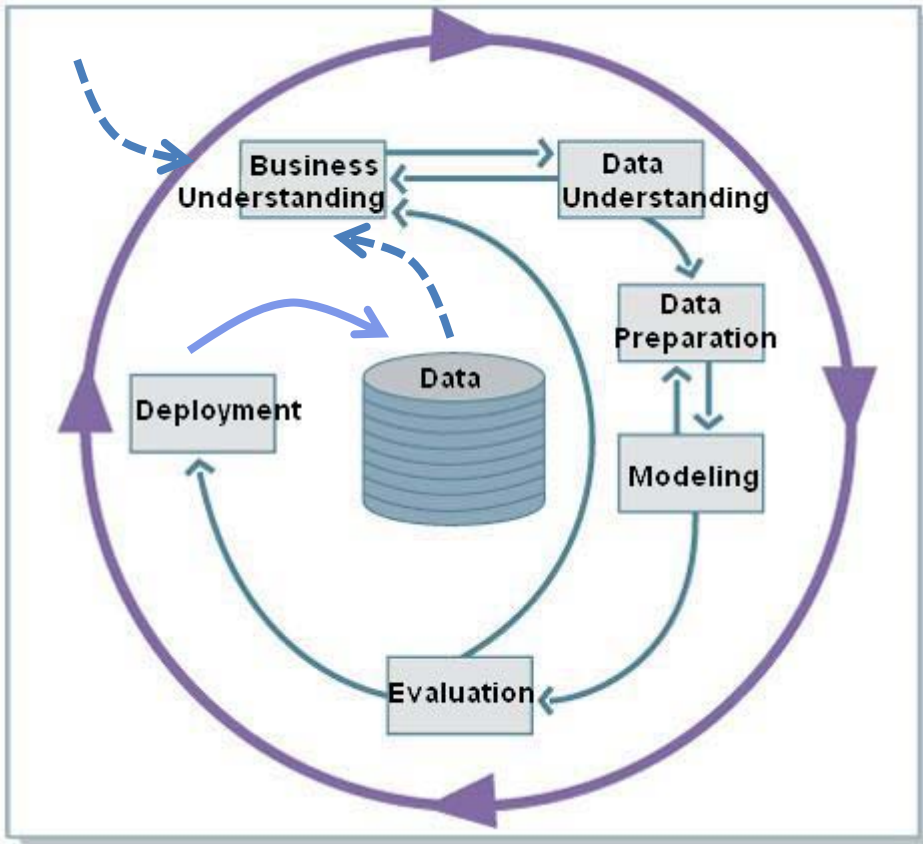


- Numeric
- Web
- Text
- Sound & Images

Data Mining
Web Mining
Text Mining
Multimedia Mining

The DM cycle

There is a problem



1. Data collection
2. Data preparation
 1. Cleansing
 2. Feature selection
 3. Feature extraction
3. Modeling
 1. Select modeling techniques
 2. Select validation
 3. Find optimal model
4. Evaluation
5. Deployment (Decision making)
(confidence on visualization of data and results)

What is Data Mining about ...

Data Mining consist on transforming Data into Information (usable) (=Knowledge)

- Data Mining is the exploration and analysis, automatic or semiautomatic, of huge quantities of secondary information, using statistics or machine learning tools, to discover relevant information useful for the decision making process.

Getting relevant information is a competitive factor for companies. Those which are able to learn more quickly and efficiently from their processes are able to take better decisions to assure their profitability.

- Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful". Clementine User Guide.

The roots of Data Mining

- *Machine learning*: “a branch of AI that deals with the design and application of learning algorithms” (Mena, 1999)
 - Algorithmic solution (complexity, scalability, ...)
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics – areas not part of data mining
- *Statistics*: “methodology for extracting information from data and expressing the amount of uncertainty in decisions we make” (Rao, 1989).
 - Inferential aspects (p.value, ...)
 - more theory-based
 - more focused on testing hypotheses
- *Data Bases*: Treatment of very large data bases

Roseta stone of DM (Lebart, 1995)

STATISTICS	MACHINE LEARNING
VARIABLES	ATRIBUTS (DB: FIELDS)
INDIVIDUALS	INSTANCES (DB: REGISTRES)
EXPLANATORY VARIABLES , PREDICTORS,	INPUT
RESPONSE VARIABLES	OUTPUT (TARGET)
MODEL	NETWORK, TREE, ...
COEFFICIENTS	WEIGHTS
FIT CRITERIUM (OLS, WLS, ML)	COST FUNCTION
ESTIMATION	LEARNING (TRAINING)
CLASSIFICATION (“CLUSTERING”)	UNSUPERVISED CLASSIFICATION
DISCRIMINATION	SUPERVISED CLASSIFICATION

Some DM problems, ...

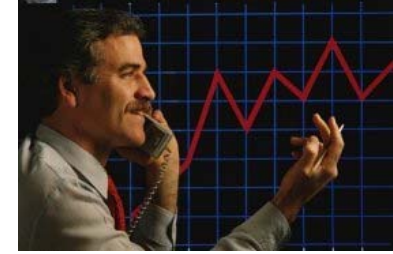
- Science
 - astronomy, bioinformatics, drug discovery, genomics, ...
- Business
 - CRM (Customer Relationship management)
 - BI (strategic integration of KDD in the decision taking process)
 - Telecom profiling
 - Credit risk
 - Fraud detection, ...
- Web:
 - Rank pages according their importance (authorities, hubs).
Classification of pages, advertising in the web. ...
- Government
 - law enforcement, profiling tax cheaters, anti-terror.

Data Mining on line: Stream data

Data Mining in the future: the Service Society

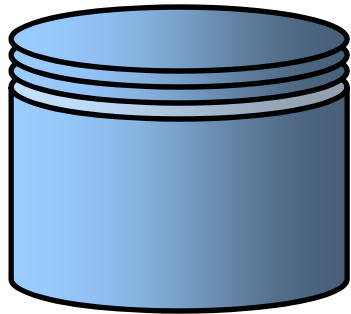
- Customer Tasks:
 - attrition prediction (identify loyal costumers)
 - Value of customers
 - targeted marketing:
 - cross-selling, advertising , ...
- Industries
 - banking, insurance, telecom, retail sales, ...

DM problem: Marketing campaigns



Required information:

Customer Data Base

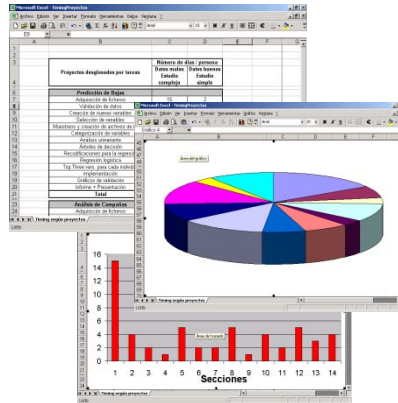


Socio-demographic data

Previous acquisitions

Enrichment with external
DB (be aware to comply
with privacy regulations)

Data from last campaign

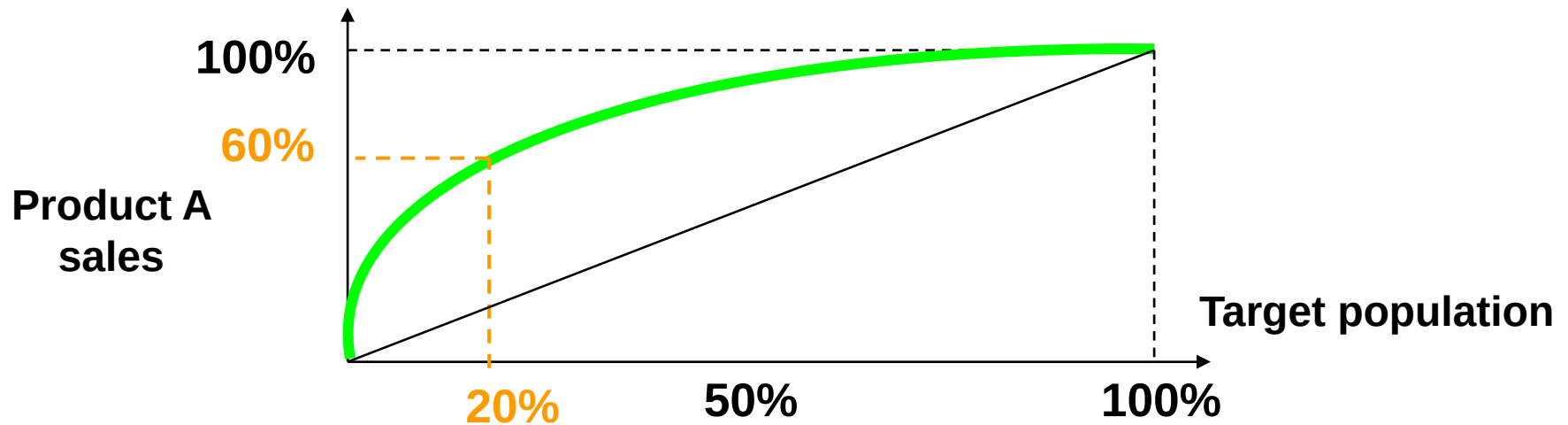


Product A?

Preprocess of data

Description of data.
Filtering for outliers
Feature selection
Feature extraction

Marketing campaigns. Results



El 60% of sales is done by 20% of potential buyers

⇒ **New campaign target**

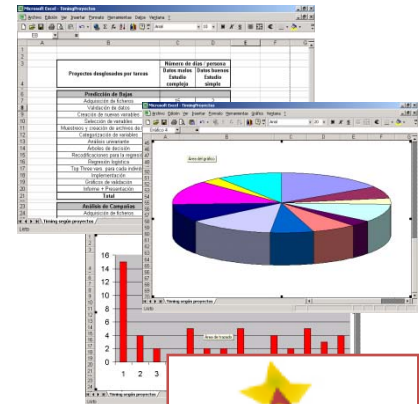
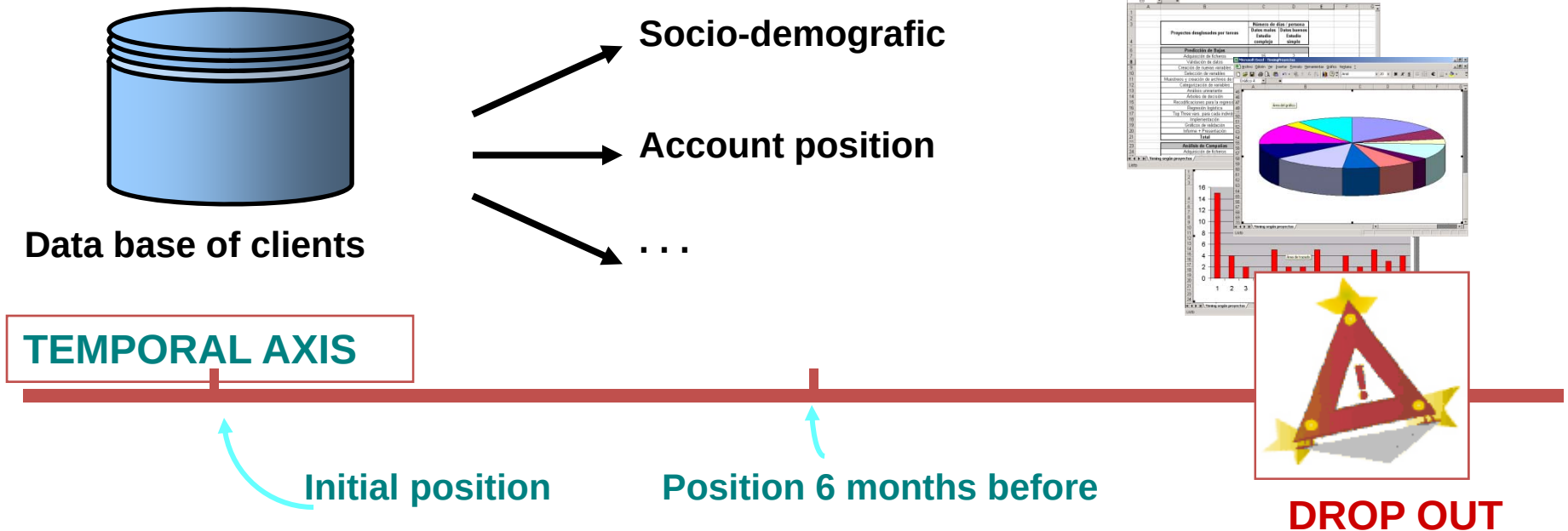
DM problem: Attrition modeling



How to retain our clients (some of them are quitting and we don't know why?)

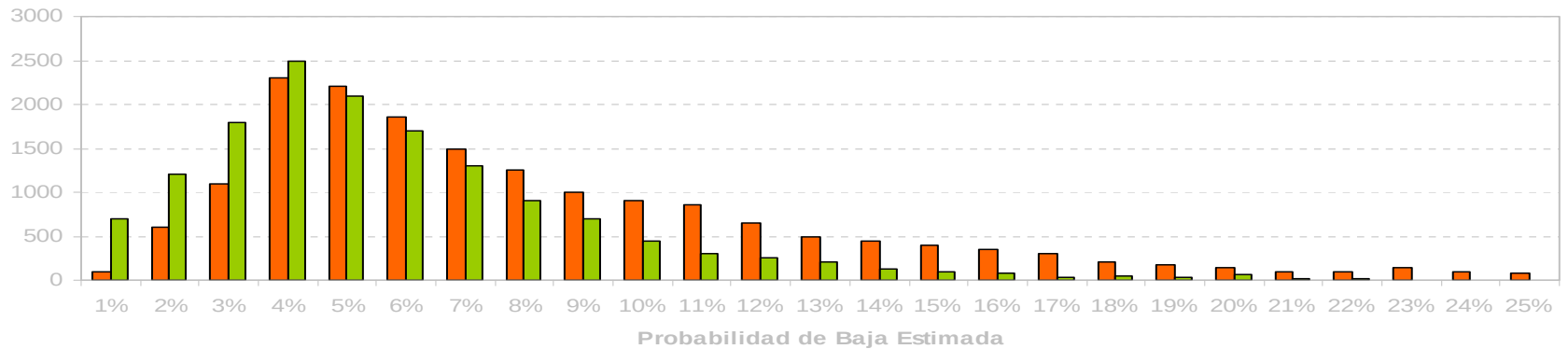
We need a model to calculate the probability of attrition for each one.

And we need to do so with enough anticipation to take effective measures.



Validation of the model

Contrast of the resigning and non resigning according the estimated probability of attrition.



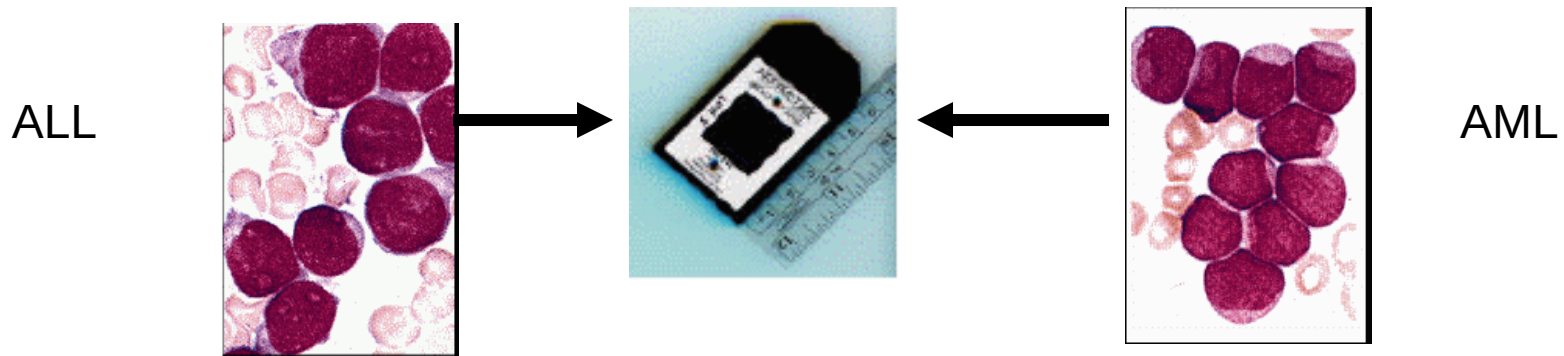
Genomic Microarrays – Case Study

Given microarray data for a number of samples (patients),
can we

- Accurately diagnose the disease?
- Predict outcome for given treatment?
- Recommend best treatment?

Example: ALL/AML data

- 38 training cases, 34 test, ~ 7,000 genes
- 2 Classes: Acute Lymphoblastic Leukemia (ALL) vs Acute Myeloid Leukemia (AML)
- Use train data to build diagnostic model



Results on test data:

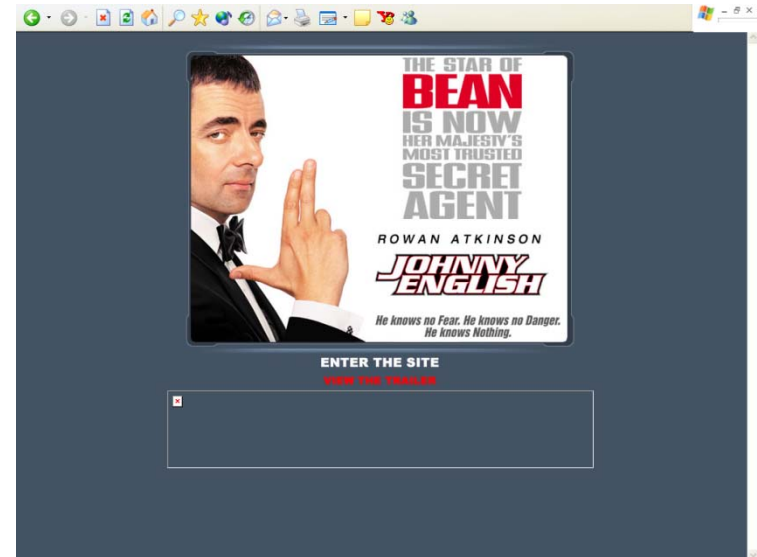
33/34 correct, 1 error may be mislabeled

Web document classification

Task:

Assign one or more class labels to a text document

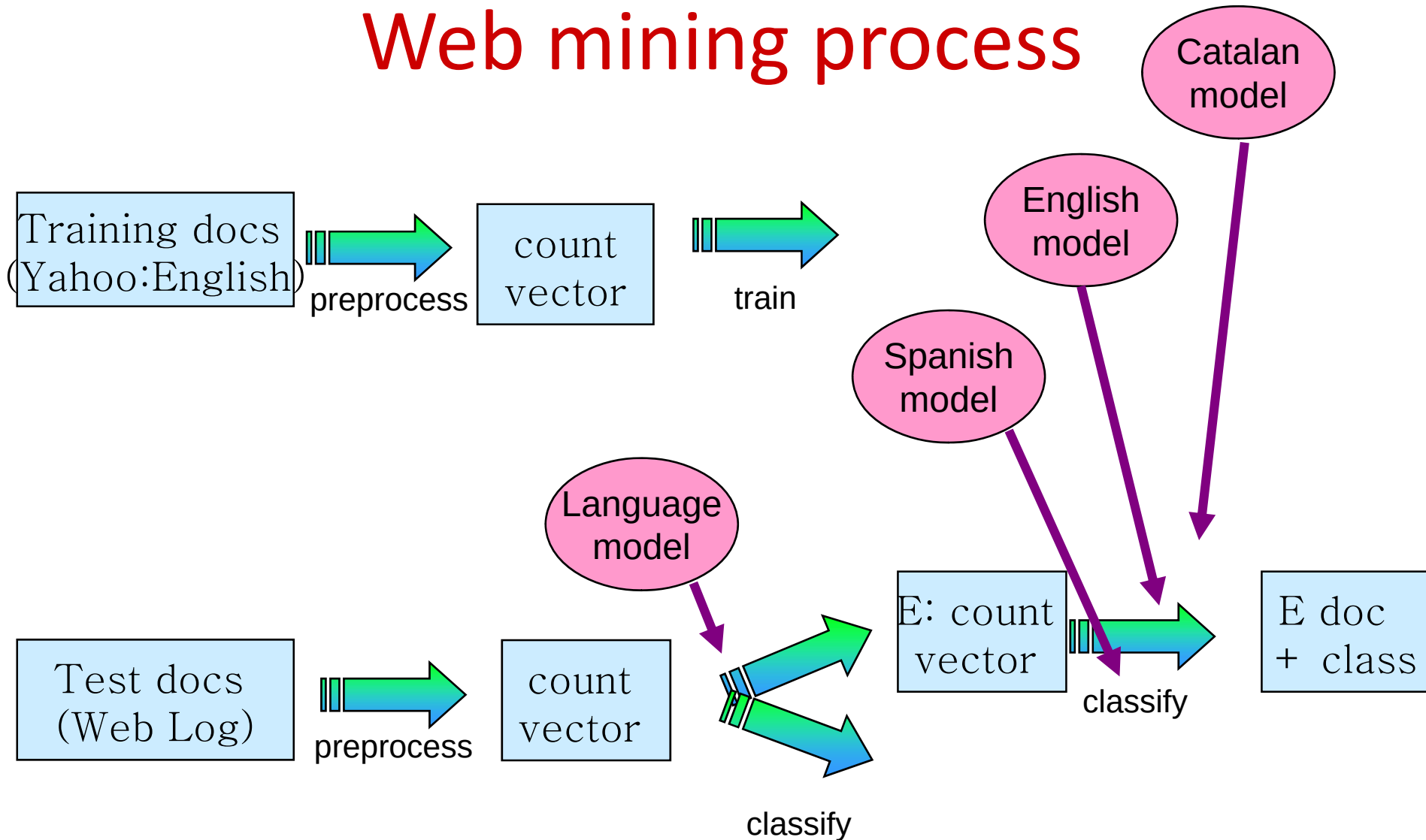
- Define classes by example:
each sample document belongs to one or more classes
→ implicit definition of class
- Several classes may hold
- Learn from set of training examples
Yahoo directory



Text Classifier

Society
& culture

Web mining process



Market Basket Analysis



Analysis of retail sales (El Corte Ingles, web, ...)

Detect frequent itemsets, what items are bought together (e.g. milk+cereal, chips+salsa)

Trivial statistical concepts, but very complex computational implementation.

IF Friday AND Diapers

THEN

Beer



IF Red wine without Denomination THEN

Fizzy soda



Data mining problems/issues 1

Limited Information

A database is often designed for purposes different from data mining (normally they are produced routinely in a process, so the data entry hasn't been designed taking into account the Data Mining goals) and sometimes the properties or attributes that would simplify the learning task are not present nor can they be requested from the real world. Inconclusive data causes problems because if some attributes essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain.

This leads to build biased models

Data mining problems/issues 2

Noise, missing values and outliers

- Noise (as random fluctuation) is always present to some extent in every real phenomenon. It conveys the complexity of reality (a philosophical dispute), two individuals with the same characteristics will come out with different outputs due to unknown causes. But noise is not an error.
- Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Obviously where possible it is desirable to minimize errors from the classification information as this affects the overall accuracy of the generated rules.
- Missing data and outliers must be detected and treated: simply disregard missing values; omit the corresponding records; infer missing values from known values; treat missing data as a special value to be included additionally in the attribute domain; or average over the missing values using Bayesian techniques.

Missing values and outliers gives us a measure of the quality of the data collection, whereas noise is inherent to the phenomenon being studied, but both give raise to uncertainty in the results.

Data mining problems/issues

Size, updates, and irrelevant fields

- Databases tend to be large and dynamic in that their contents are ever-changing as information is added, modified or removed. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date and consistent with the most current information. Also the learning system has to be time-sensitive as some data values vary over time and the discovery system is affected by the 'timeliness' of the data. Be aware of false predictors (i.e. the increase in expenses when the client has got a loan can't be use to predict the concession of the loan, strong predictors are highly suspect).
- Another issue is the relevance or irrelevance of the fields in the database to the current focus of discovery.

Major Data Mining Tasks

- **Visualization:** to facilitate human discovery
- **Summarization:** describing a group
- **Deviation Detection:** finding changes
- **Profiling:** finding the significative characteristics of a group of individuals
- **Associations:** e.g. A & B & C occur frequently
- **Clustering:** finding clusters in data
- **Prediction:**
 - **Classification:** predicting an item class
 - **Regression:** predicting a continuous value
- **Link Analysis:** finding relationships
- ...

Major statistical learning problems

- **Density estimation: “Unsupervised” learning**
 - Determine (joint) distribution of the data $P(X)$
 - Robot inferring a map of a building
 - Distribution of words in text documents
 - Amino acid distribution in the human genome
 - Pixel intensity distribution over images
- **Classification: “Supervised” learning**
 - Determine conditional distribution $P(Y/X)$
 - Text classification, visual recognition, Credit card screening, medical diagnosis, gene finding, biometrics, optical character recognition, stock analysis, ...
- **Regression: Function approximation**
 - Determine conditional mean $E(Y/X)$
 - Applications: Reinforcement learning, scientific models