

# Data mining fundamentals



Elena Baralis  
*Politecnico di Torino*



## Data analysis

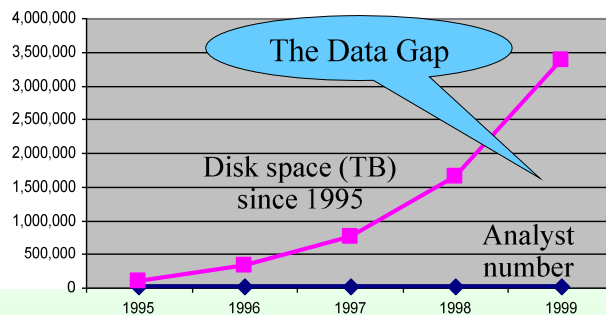
- Most companies own huge databases containing
  - operational data
  - textual documents
  - experiment results
- These databases are a potential source of useful information





## Data analysis

- Information is “hidden” in huge datasets
  - not immediately evident
  - human analysts need a large amount of time for the analysis
  - most data *is never analyzed at all*



DBG  
M

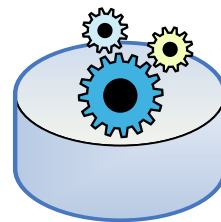
From R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

3



## Data mining

- Non trivial extraction of
  - implicit
  - previously unknown
  - potentially usefulinformation from available data
- Extraction is automatic
  - performed by appropriate algorithms
- Extracted information is represented by means of abstract models
  - denoted as *pattern*



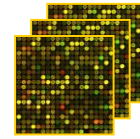
DBG  
M

4



## Example: biological data

- Microarray
  - expression level of genes in a cellular tissue
  - various types (mRNA, DNA)
- Patient clinical records
  - personal and demographic data
  - exam results
- Textual data in public collections
  - heterogeneous formats, different objectives
  - scientific literature (PubMed)
  - ontologies (Gene Ontology)



| CJD               | PATIENT | shv013 | shv060 | shq077 | shv039 | shv014 | shq082 | shq083 | shv008 |
|-------------------|---------|--------|--------|--------|--------|--------|--------|--------|--------|
| ID                | 49A34   | 45A9   | 52A28  | 4434   | 61A31  | 59A6   | 46A15  | 41A31  |        |
| IMAGE:74ISG20   H |         | -1.02  | -2.34  | 1.44   | 0.57   | -0.13  | 0.12   | 0.34   | -0.57  |
| IMAGE:787NFSF13   |         | -0.52  | -4.08  | -0.29  | 0.71   | 1.03   | -0.67  | 0.22   | -0.09  |
| IMAGE:39LCC334    |         | -0.28  | -4.08  | 0.09   | 0.13   | 0.08   | 0.09   | -0.08  | -0.09  |
| IMAGE:23ITGA4   H |         | -1.379 | -1.698 | 0.159  | -0.019 | 0.039  | -0.039 | 0.508  | -0.869 |



## Biological analysis objectives

- Clinical analysis
  - detecting the causes of a pathology
  - monitoring the effect of a therapy
  - ⇒ diagnosis improvement and definition of new specific therapies
- Bio-discovery
  - gene network discovery
  - analysis of multifactorial genetic pathologies
- Pharmacogenesis
  - lab design of new drugs for genic therapies

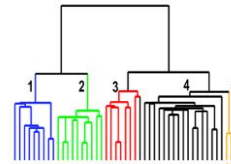
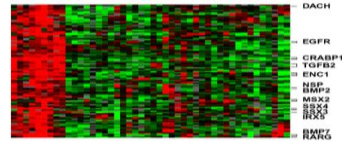


How can data mining contribute?

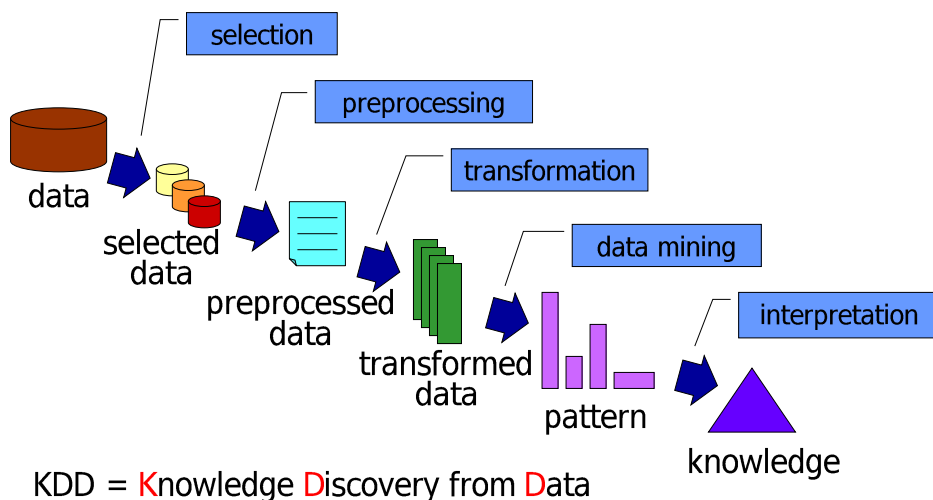



## Data mining contributions

- Pathology diagnosis
  - classification
- Selecting genes involved in a specific pathology
  - feature selection
  - clustering
- Grouping genes with similar functional behavior
  - clustering
- Multifactorial pathologies analysis
  - association rules
- Detecting chemical components appropriate for specific therapies
  - classification

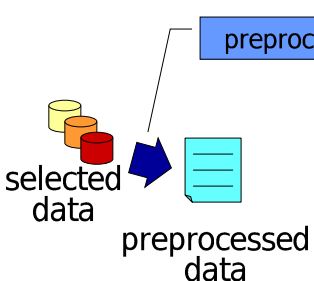


## Knowledge Discovery Process





## Preprocessing




selected data

preprocessed data


preprocessing

- data cleaning
  - reduces the effect of noise
  - identifies or removes outliers
  - solves inconsistencies
- data integration
  - reconciles data extracted from different sources
  - integrates metadata
  - identifies and solves data value conflicts
  - manages redundancy

Real world data is "dirty"  
Without good quality data, no good quality pattern

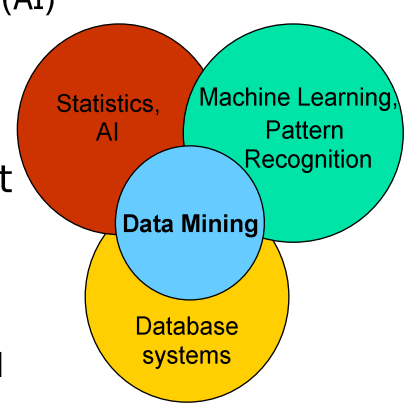


9



## Data mining origins

- Draws from
  - statistics, artificial intelligence (AI)
  - pattern recognition, machine learning
  - database systems
- Traditional techniques are not appropriate because of
  - significant data volume
  - large data dimensionality
  - heterogeneous and distributed nature of data




Statistics, AI

Machine Learning, Pattern Recognition

Data Mining

Database systems

From: P. Tan, M. Steinbach, V. Kumar, "Introduction to Data Mining"



10



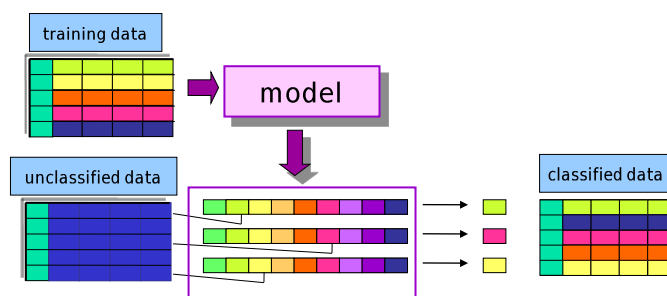
## Analysis techniques

- Descriptive methods
  - Extract interpretable models describing data
  - Example: client segmentation
- Predictive methods
  - Exploit some known variables to predict unknown or future values of (other) variables
  - Example: "spam" email detection



## Classification

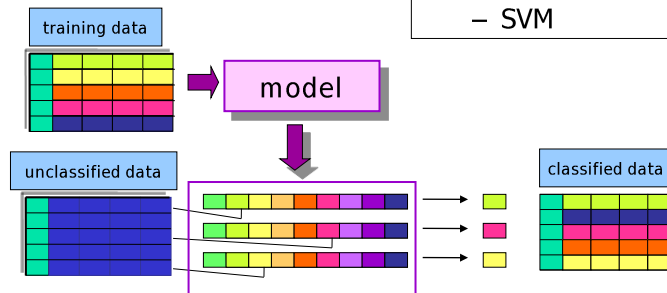
- Objectives
  - prediction of a class label
  - definition of an interpretable model of a given phenomenon





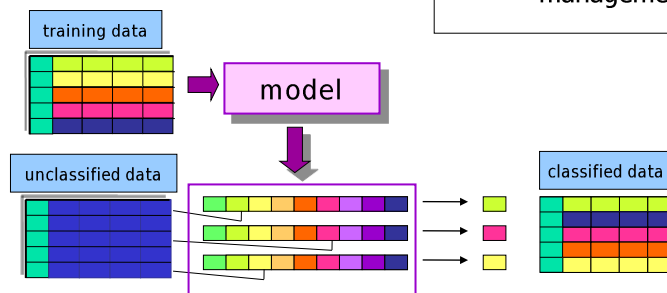
# Classification

- Approaches
  - decision trees
  - bayesian classification
  - classification rules
  - neural networks
  - k-nearest neighbours
  - SVM



# Classification

- Requirements
  - accuracy
  - interpretability
  - scalability
  - noise and outlier management

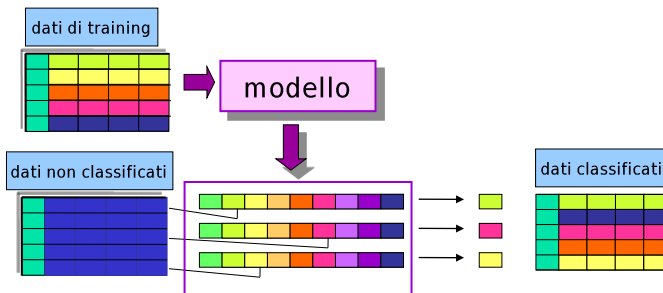




# Classification

## ■ Applications

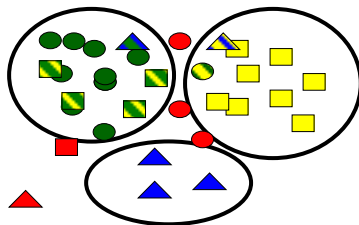
- detection of customer propension to leave a company (churn or attrition)
- fraud detection
- classification of different pathology types
- ...



# Clustering

## ■ Objectives

- detecting groups of similar data objects
- identifying exceptions and outliers



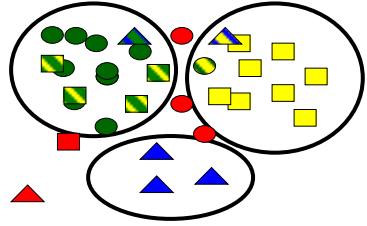




# Clustering

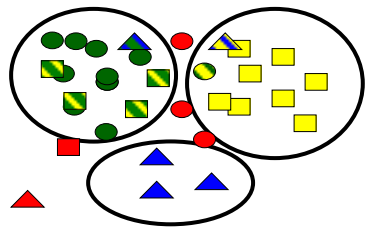
- Approaches
  - partitional (K-means)
  - hierarchical
  - density-based (DBSCAN)
  - SOM

- Requirements
  - scalability
  - management of
    - noise and outliers
    - large dimensionality
  - interpretability



# Clustering

- Applications
  - customer segmentation
  - clustering of documents containing similar information
  - grouping genes with similar expression pattern
  - ...





## Association rules

- Objective
  - extraction of frequent correlations or pattern from a transactional database

Tickets at a supermarket counter

| TID | Items                      |
|-----|----------------------------|
| 1   | Bread, Coke, Milk          |
| 2   | Beer, Bread                |
| 3   | Beer, Coke, Diapers, Milk  |
| 4   | Beer, Bread, Diapers, Milk |
| 5   | Coke, Diapers, Milk        |
| ... | ...                        |

- Association rule  
diapers  $\Rightarrow$  beer
  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contain beer



## Association rules

- Applications
  - market basket analysis
  - cross-selling
  - shop layout or catalogue design

Tickets at a supermarket counter

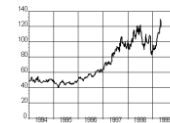
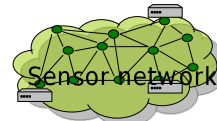
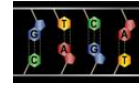
| TID | Items                          |
|-----|--------------------------------|
| 1   | Bread, Coca Cola, Milk         |
| 2   | Beer, Bread                    |
| 3   | Beer, Coca Cola, Diapers, Milk |
| 4   | Beer, Bread, Diapers, Milk     |
| 5   | Coca Cola, Diapers, Milk       |
| ... | ...                            |

- Association rule  
diapers  $\Rightarrow$  beer
  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contain beer



## Other data mining techniques

- Sequence mining
  - ordering criteria on analyzed data are taken into account
  - example: motif detection in proteins
- Time series and geospatial data
  - temporal and spatial information are considered
  - example: sensor network data
- Regression
  - prediction of a continuous value
  - example: prediction of stock quotes
- Outlier detection
  - example: intrusion detection in network traffic analysis



## Open issues

- Scalability to *huge* data volumes
- Data dimensionality
- Complex data structures, heterogeneous data formats
- Data quality
- Privacy preservation
- Streaming data