

# Data mining fundamentals



Elena Baralis  
Politecnico di Torino

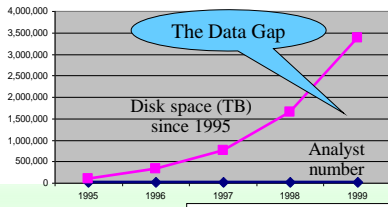
## Data analysis

- 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*

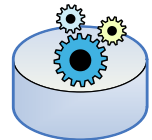


From R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"



## Data mining

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



## 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)



CD	PATENT	PHOS	PHAS	PHOT	PHOS	PHOT	PHOS	PHOT	PHOS	PHOT
D	634	631	6282	634	632	634	631	632	631	632
RANGE:1482(14)	-0.03	0.24	1.44	1.03	0.12	0.12	0.24	0.24	0.24	0.24
RANGE:1679(14)	-0.03	-4.08	-0.28	0.17	1.14	-0.97	0.14	0.14	0.14	0.14
RANGE:1670(14)	-0.03	-1.08	0.08	0.18	1.08	0.18	0.18	0.18	0.18	0.18
RANGE:1725(14)	-0.03	-1.08	0.18	-0.03	0.18	0.18	0.18	0.18	0.18	0.18



## 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

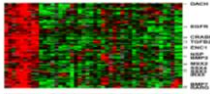
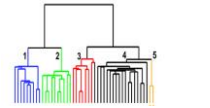


How can data mining contribute?



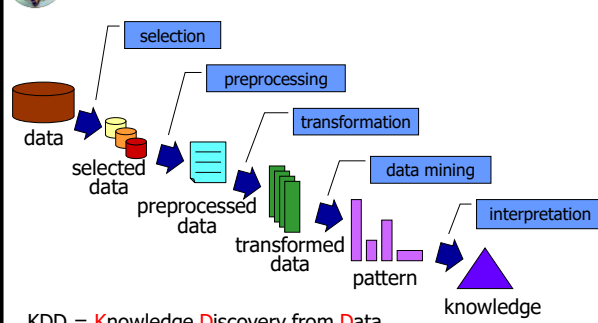
## Data mining contributions

- 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

DBGM 7

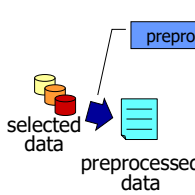
## Knowledge Discovery Process



KDD = Knowledge Discovery from Data

DBGM 8

## Preprocessing



**data cleaning**

- reduces the effect of noise
- identifies or removes outliers
- solves inconsistencies

**data integration**

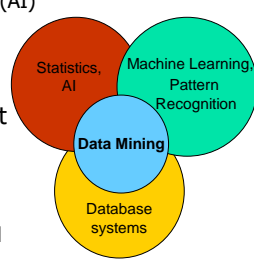
- reconciles data extracted from different sources
- integrates metadata
- identifies and solves data value conflicts
- manages redundancy

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

DBGM 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



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

DBGM 10

## Analysis techniques

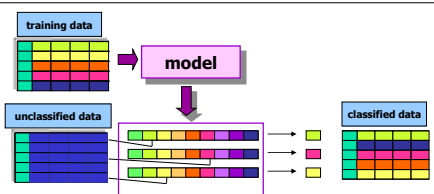
- Descriptive methods
  - Extract interpretable models describing data
  - Example: client segmentation
- Predictive methods
  - Exploit some known variables to predict unknown or future values of (other) variables
  - Example: "spam" email detection

DBGM 11

## Classification

**Objectives**

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



DBGM 12

## Classification

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

The diagram shows a flow from 'training data' (a grid of colored squares) to a 'model' box. Below this, 'unclassified data' (another grid) is processed by the model to produce 'classified data' (a grid where each square is a different color, representing different classes).

DBG

13

## Classification

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

The diagram shows a flow from 'training data' (a grid of colored squares) to a 'model' box. Below this, 'unclassified data' (another grid) is processed by the model to produce 'classified data' (a grid where each square is a different color, representing different classes).

DBG

14

## Classification

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

The diagram shows a flow from 'dati di training' (a grid of colored squares) to a 'modello' box. Below this, 'dati non classificati' (another grid) is processed by the model to produce 'dati classificati' (a grid where each square is a different color, representing different classes).

DBG

15

## Clustering

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

The diagram shows several groups of data points (circles, squares, triangles) clustered together. Some points are circled, indicating they are part of a cluster, while others are not, indicating they are outliers.

DBG

16

## Clustering

- Approaches
  - partitional (K-means)
  - hierarchical
  - density-based (DBSCAN)
  - SOM
- Requirements
  - scalability
  - management of
    - noise and outliers
    - large dimensionality
  - interpretability

The diagram shows several groups of data points (circles, squares, triangles) clustered together. Some points are circled, indicating they are part of a cluster, while others are not, indicating they are outliers.

DBG

17

## Clustering

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

The diagram shows several groups of data points (circles, squares, triangles) clustered together. Some points are circled, indicating they are part of a cluster, while others are not, indicating they are outliers.

DBG

18



## 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



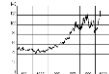
### Regression

- 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