Data Mining

Elena Baralis, Paolo Garza

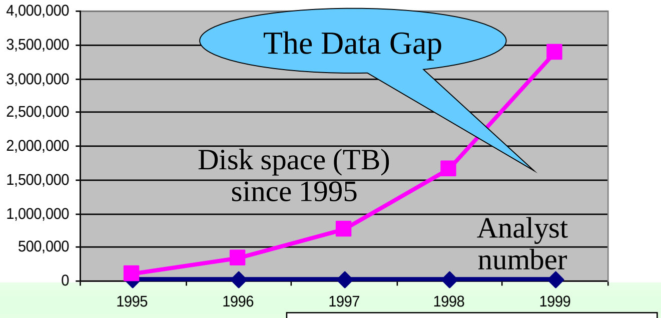
Politecnico di Torino

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

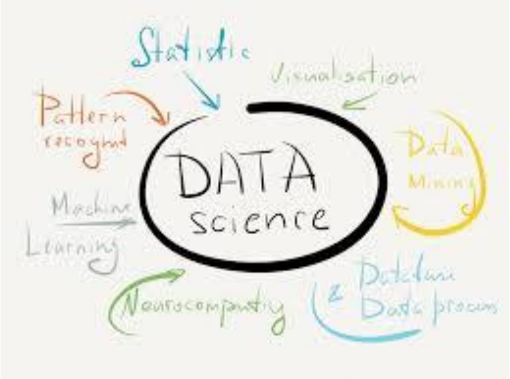
- Information is “hidden” in huge datasets
  - not immediately evident
  - human analysts need a large amount of time for the analysis
  - most data *is never analyzed at all*



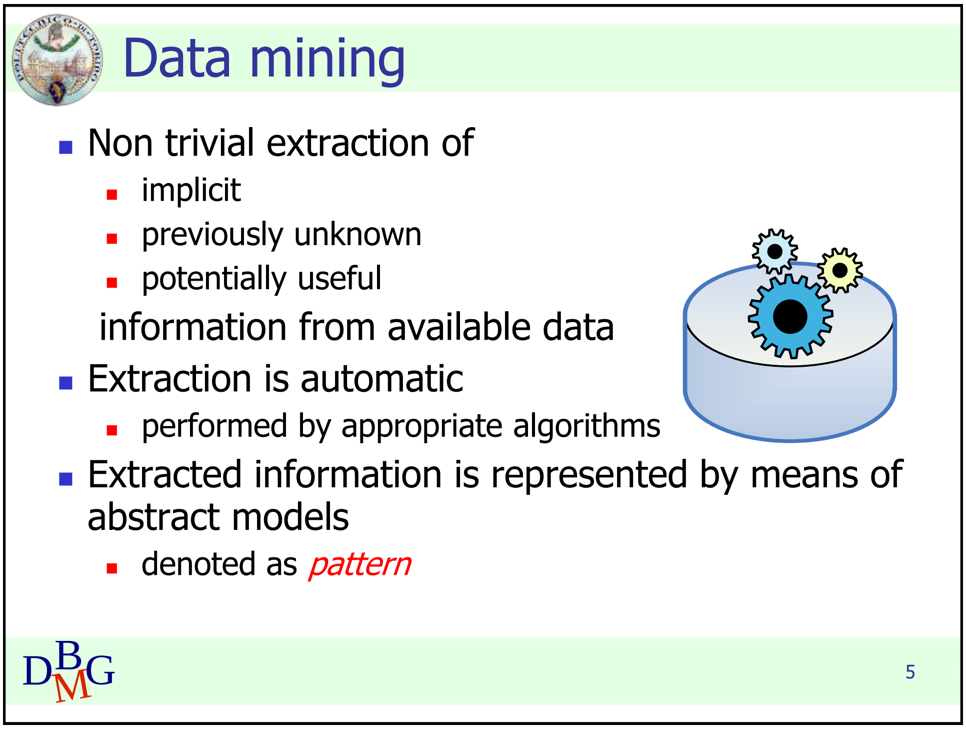
From R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”

Data science

“Extracting meaning from very large quantities of data”



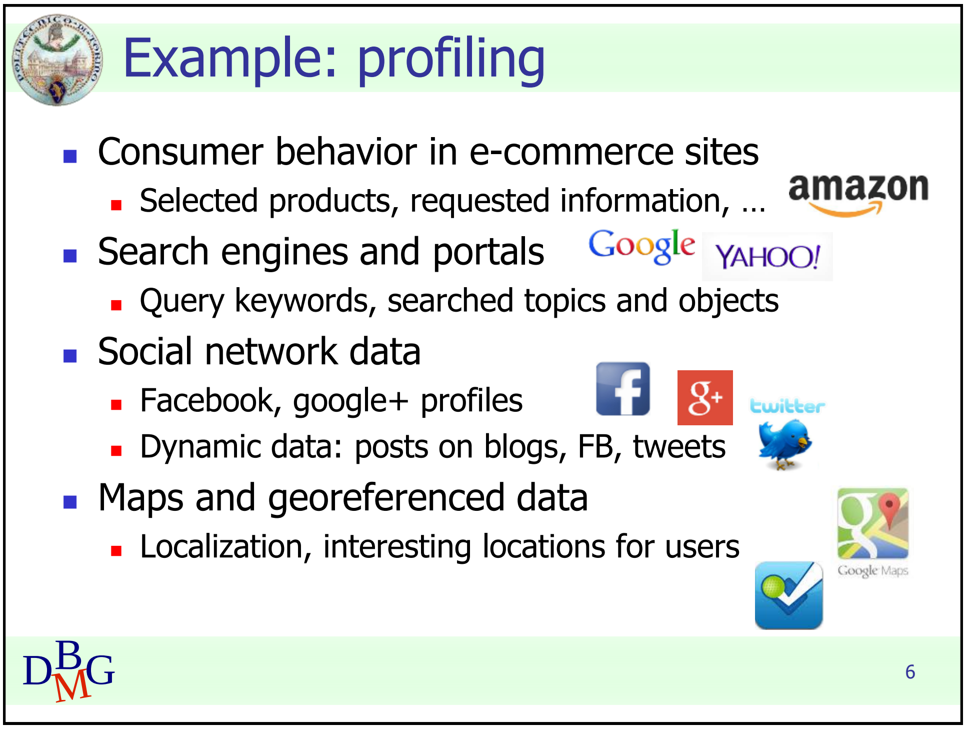
D.J. Patil coined the word *data scientist*



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: profiling

- Consumer behavior in e-commerce sites
  - Selected products, requested information, …
- Search engines and portals
  - Query keywords, searched topics and objects
- Social network data
  - Facebook, google+ profiles
  - Dynamic data: posts on blogs, FB, tweets
- Maps and georeferenced data
  - Localization, interesting locations for users

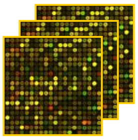
amazon
Google
YAHOO!
twitter
Google Maps

Example: profiling

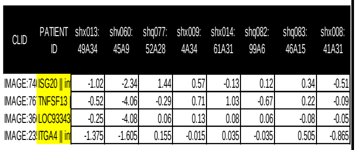
- User/service profiling
  - Recommendation systems
  - Advertisements
- Market basket analysis
  - Correlated objects for cross selling
    - User registration, fidelity cards
- Context-aware data analysis
  - Integration of different dimensions
    - E.g., location, time of the day, user interest
- Text mining
  - Brand reputation, sentiment analysis, topic trends

Example: biological data

- Microarray
  - expression level of genes in a cellular tissue
  - various types (mRNA, DNA)

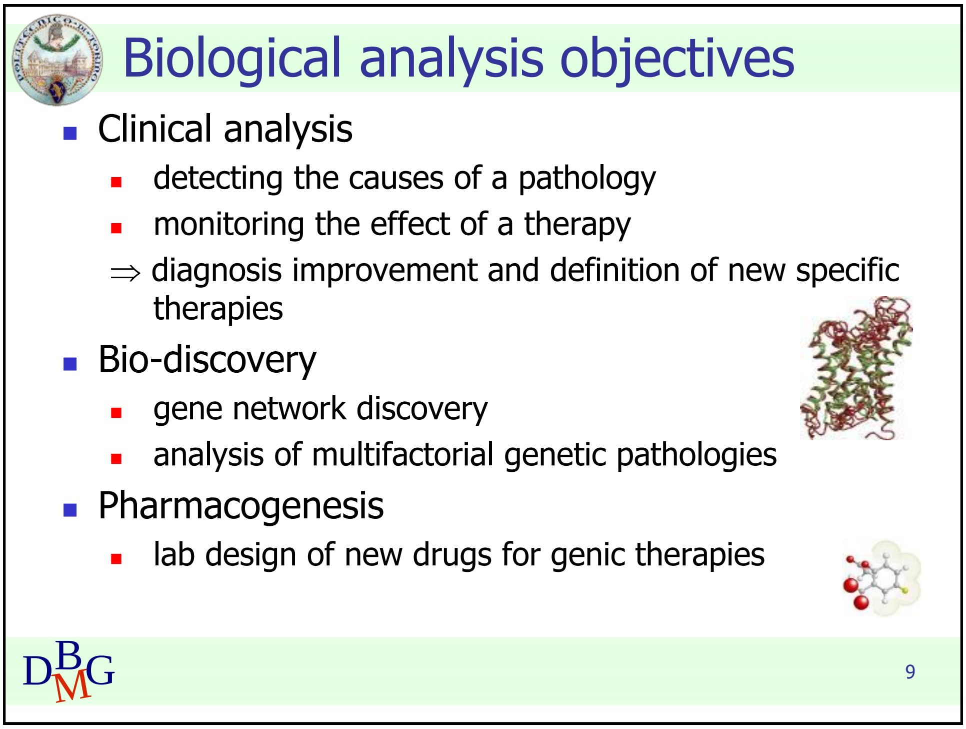


- Patient clinical records
  - personal and demographic data
  - exam results



CLID	PATIENT ID	shx013: 49A34	shv060: 45A9	shq077: 52A28	shx009: 4A34	shx014: 61A31	shq082: 99A6	shq083: 46A15	shx008: 41A31
IMAGE:74	ISG20 || in	-1.02	-2.34	1.44	0.57	-0.13	0.12	0.34	-0.51
IMAGE:76	TNFSF13 |	-0.52	-4.06	-0.29	0.71	1.03	-0.67	0.22	-0.09
IMAGE:36	LOC93343	-0.25	-4.08	0.06	0.13	0.08	0.06	-0.08	-0.05
IMAGE:23	ITGA4 || in	-1.375	-1.605	0.155	-0.015	0.035	-0.035	0.505	-0.865

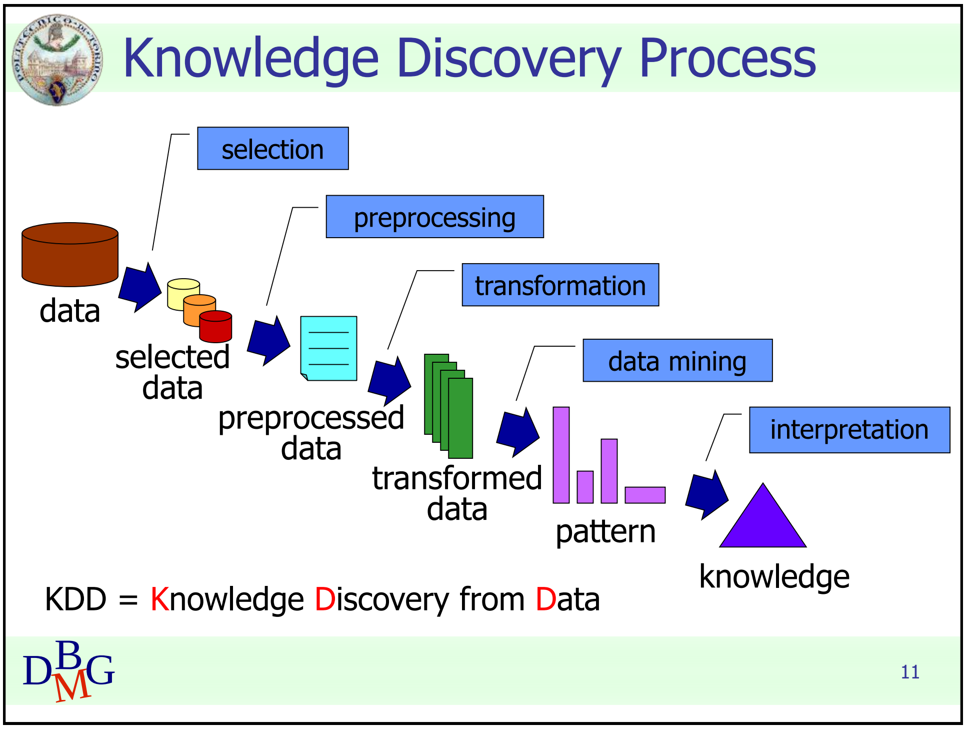
- Textual data in public collections
  - heterogeneous formats, different objectives
  - scientific literature (PUBMed)
  - ontologies (Gene Ontology)



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



Knowledge Discovery Process

selection

preprocessing

transformation

data mining

interpretation

data

selected data

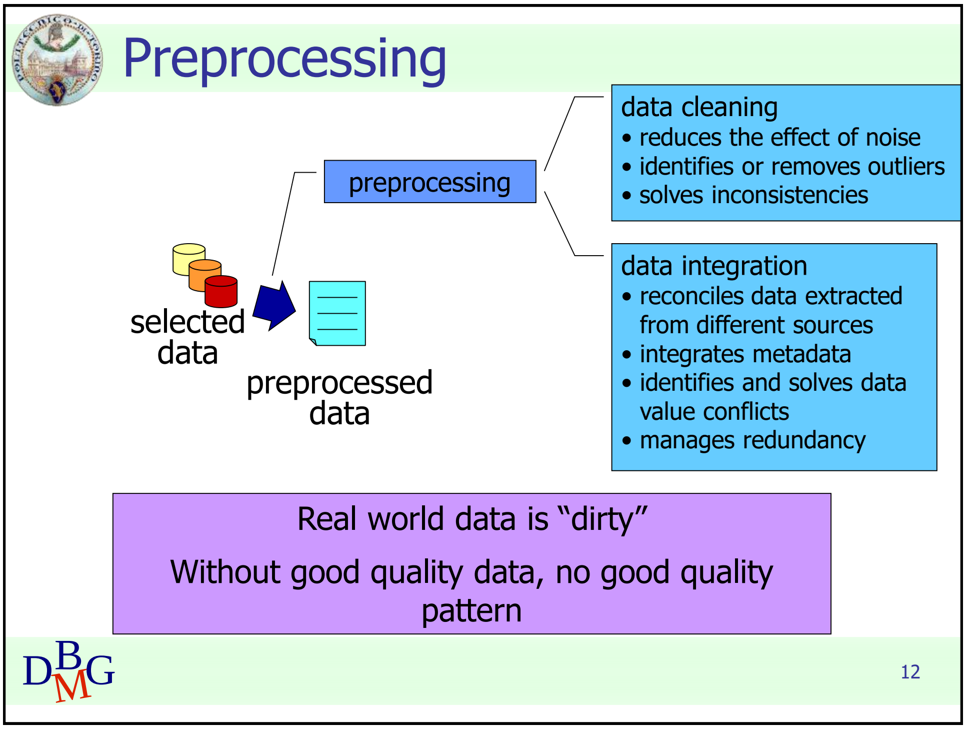
preprocessed data

transformed data

pattern

knowledge

KDD = Knowledge Discovery from Data



Preprocessing

selected data

preprocessing

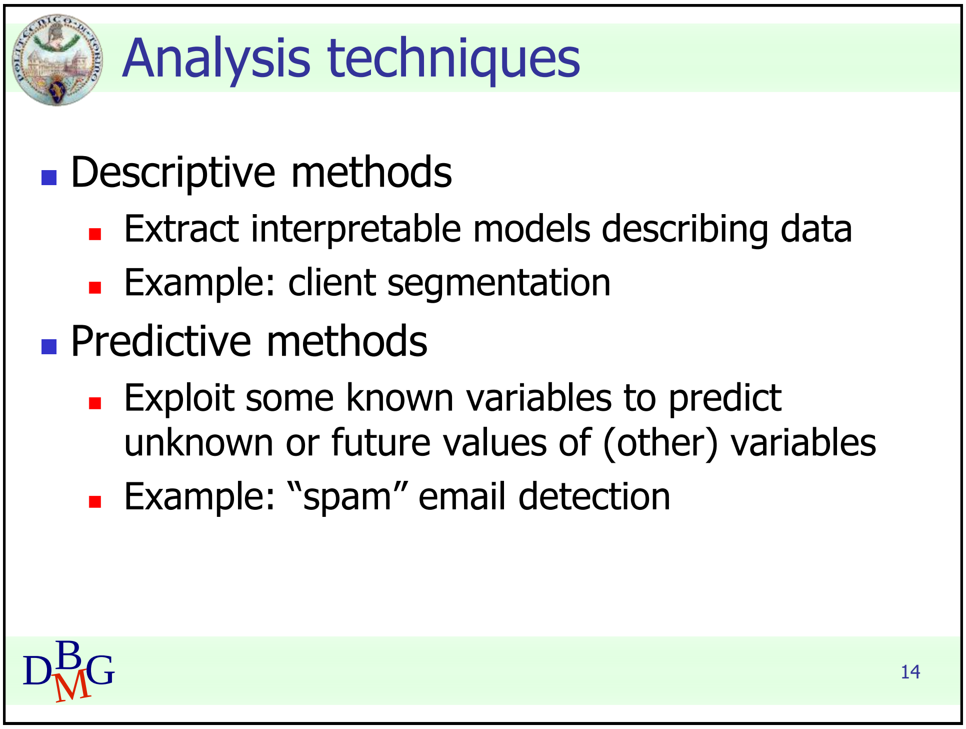
preprocessed data

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

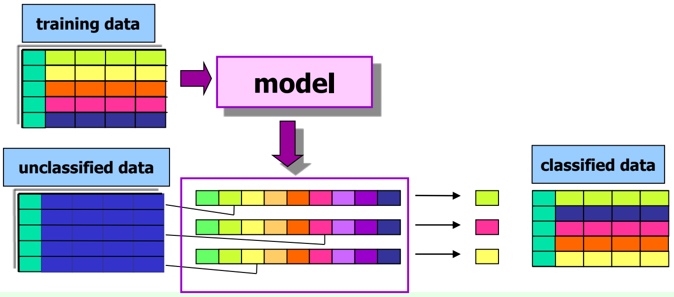


Analysis techniques

- 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

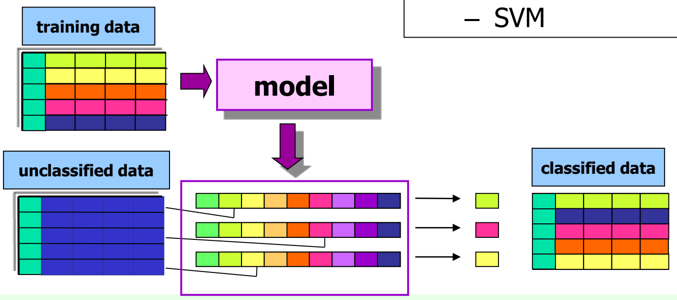
Classification

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



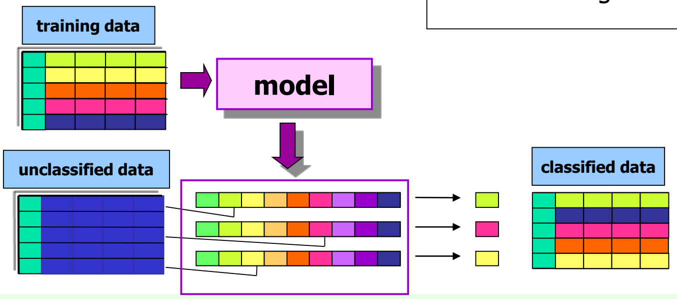
Classification

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



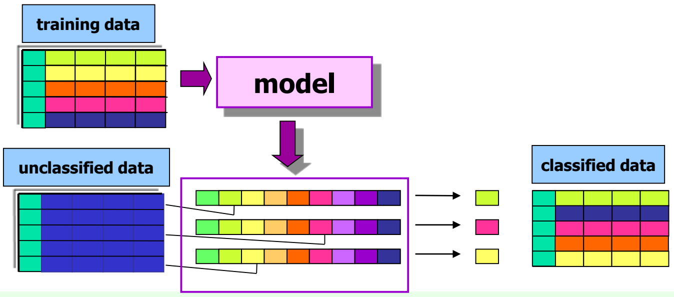
Classification

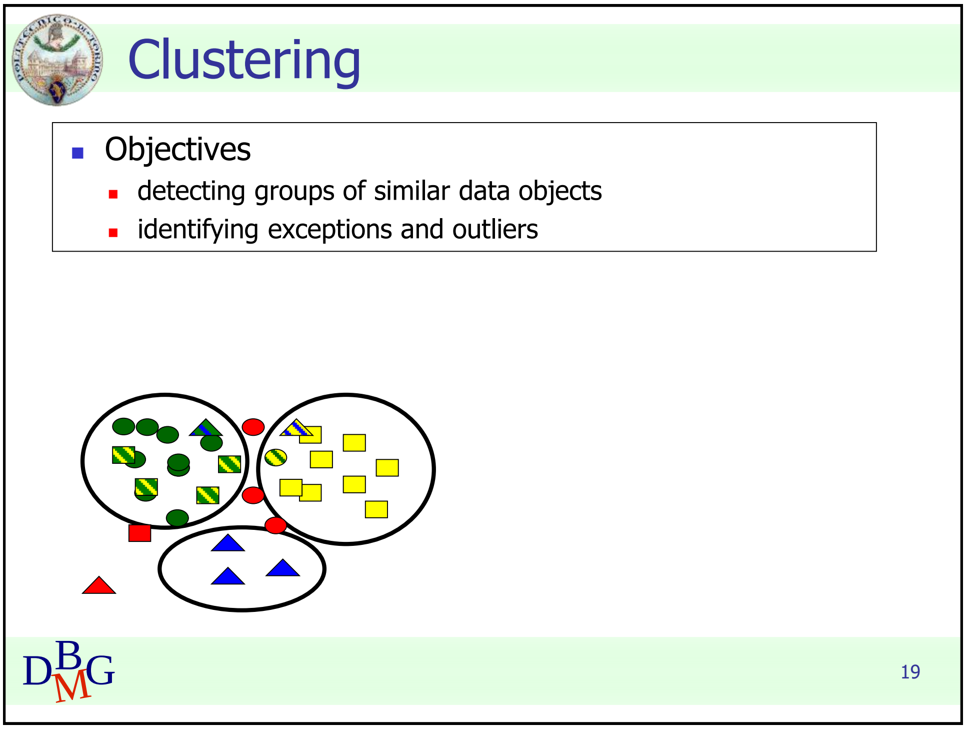
- Requirements
  - accuracy
  - interpretability
  - scalability
  - noise and outlier management



Classification

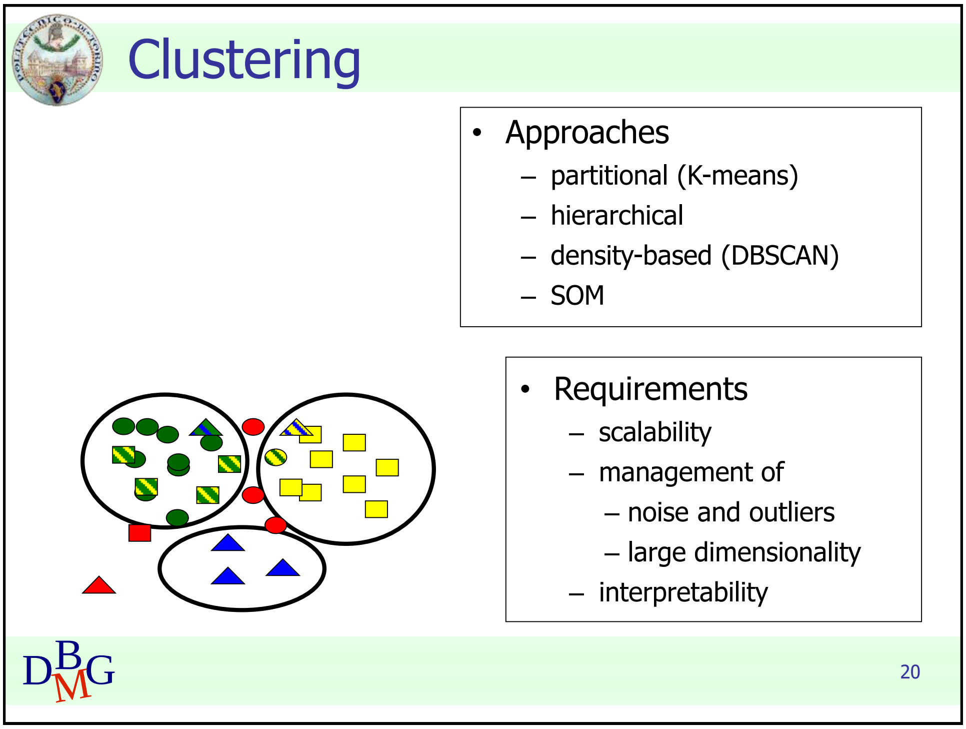
- Applications
  - detection of customer propension to leave a company (churn or attrition)
  - fraud detection
  - classification of different pathology types
  - …





Clustering

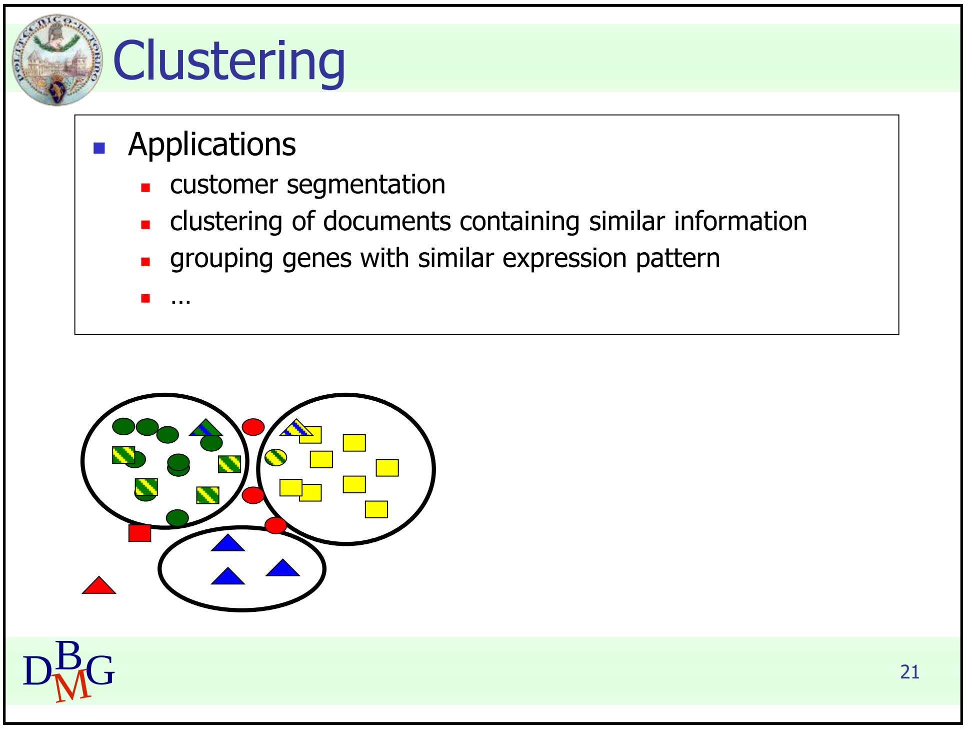
- Objectives
  - detecting groups of similar data objects
  - identifying exceptions and outliers



Clustering

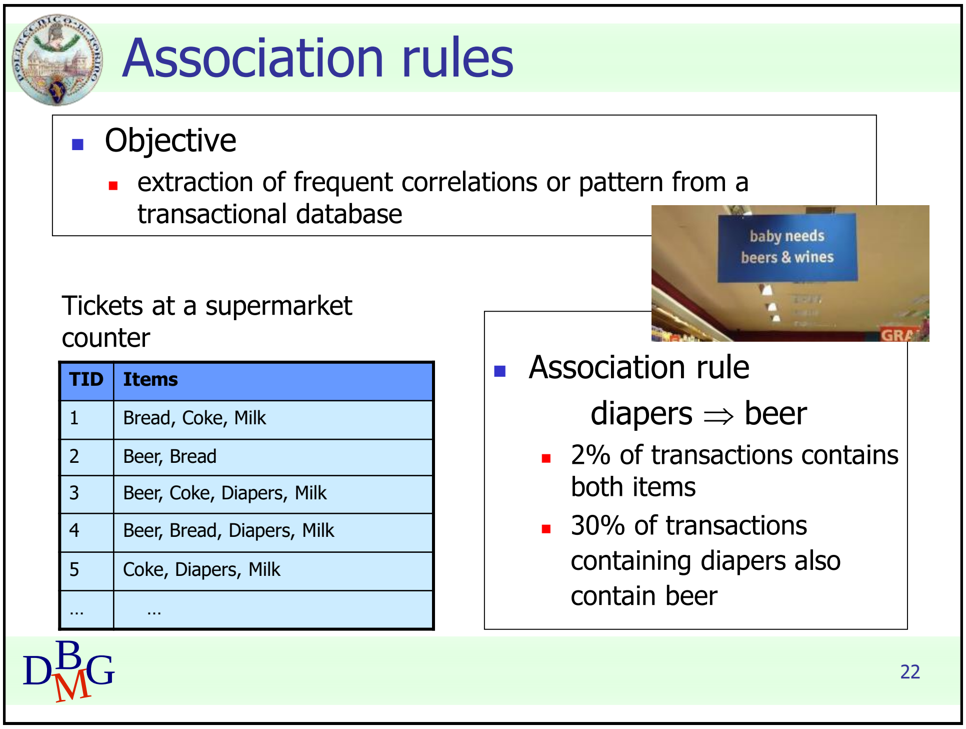
- Approaches
  - partitional (K-means)
  - hierarchical
  - density-based (DBSCAN)
  - SOM

- 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

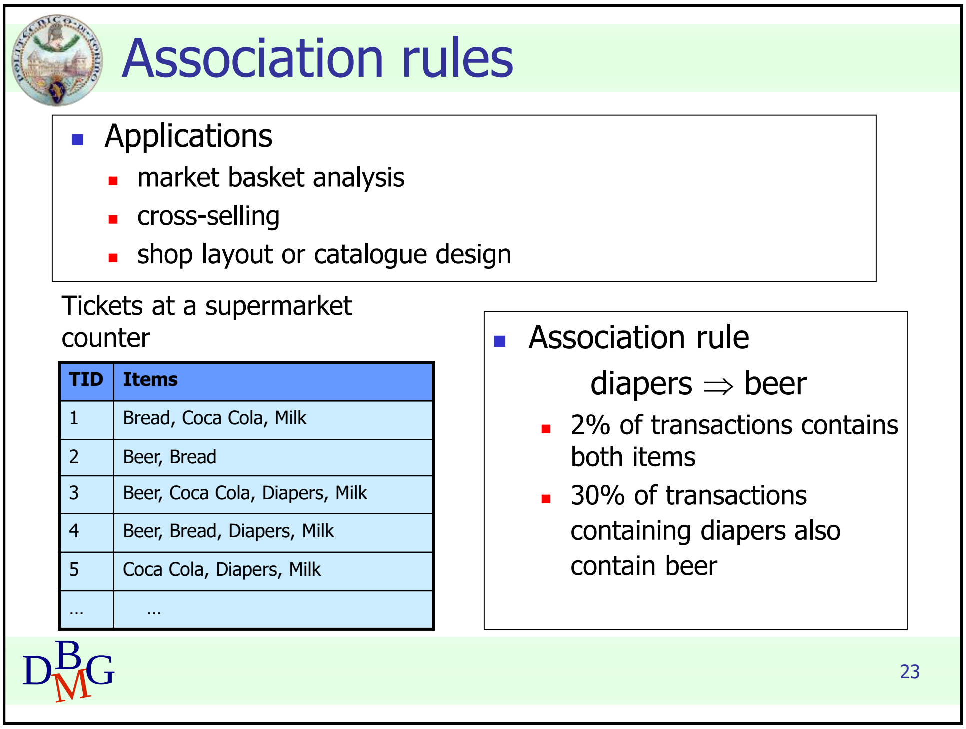
baby needs
beers & wines
GRA

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 ⇒ beer
  - 2% of transactions contains both items
  - 30% of transactions containing diapers also contain beer



Association rules

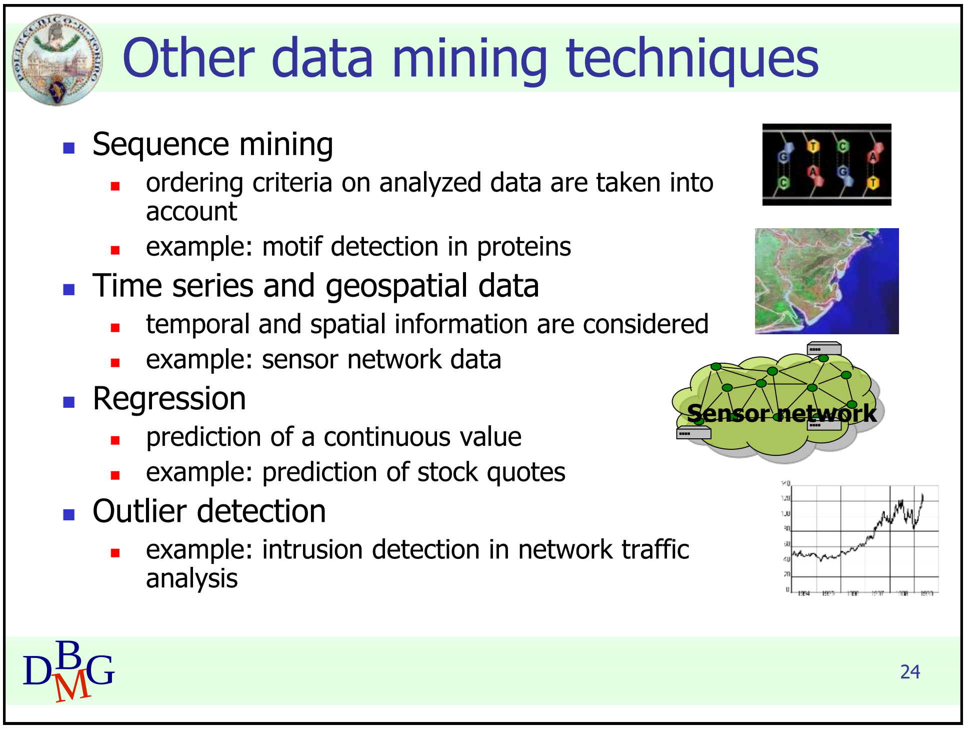
- Applications
  - market basket analysis
  - cross-selling
  - shop layout or catalogue design

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
…	…

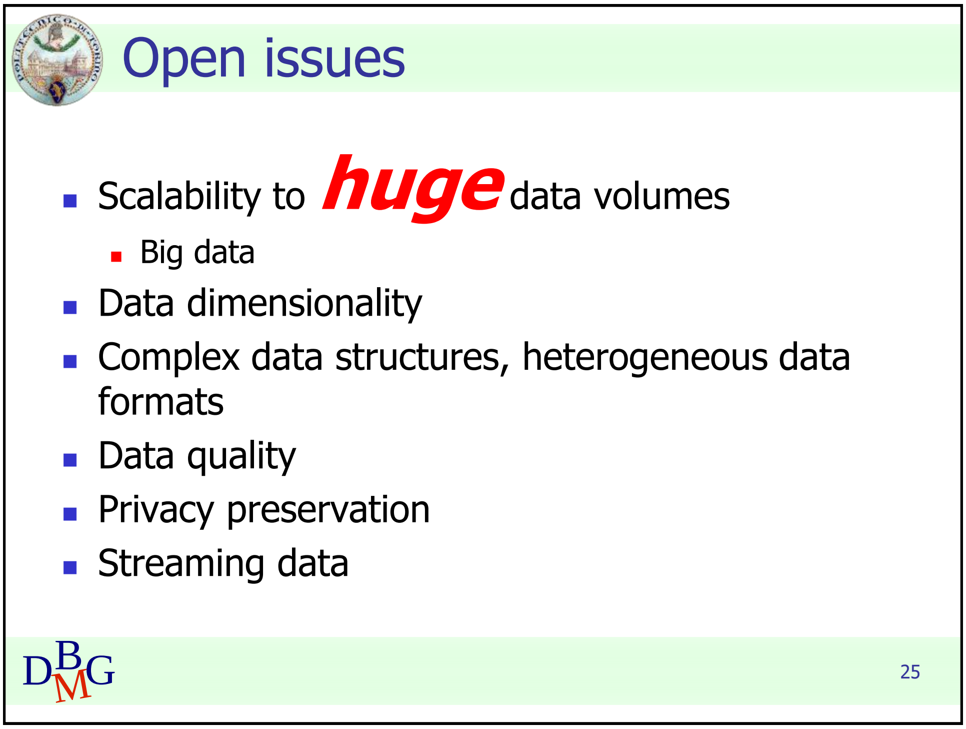
- Association rule

  diapers ⇒ 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
  - Big data
- Data dimensionality
- Complex data structures, heterogeneous data formats
- Data quality
- Privacy preservation
- Streaming data