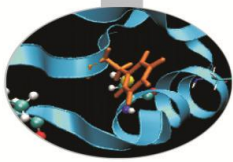
CINECA SCAI
SuperComputing Applications and Innovation

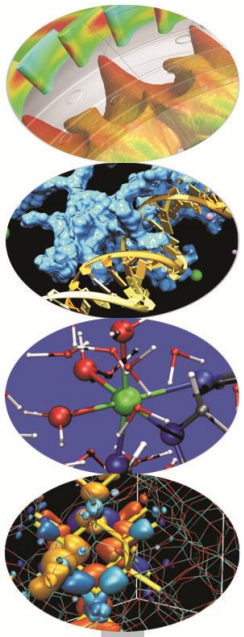


Introduction to Data Analytics

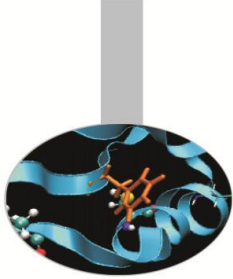
School on Scientific Data Analytics and Visualization

Roberta Turra, *Cineca*

8 June 2015

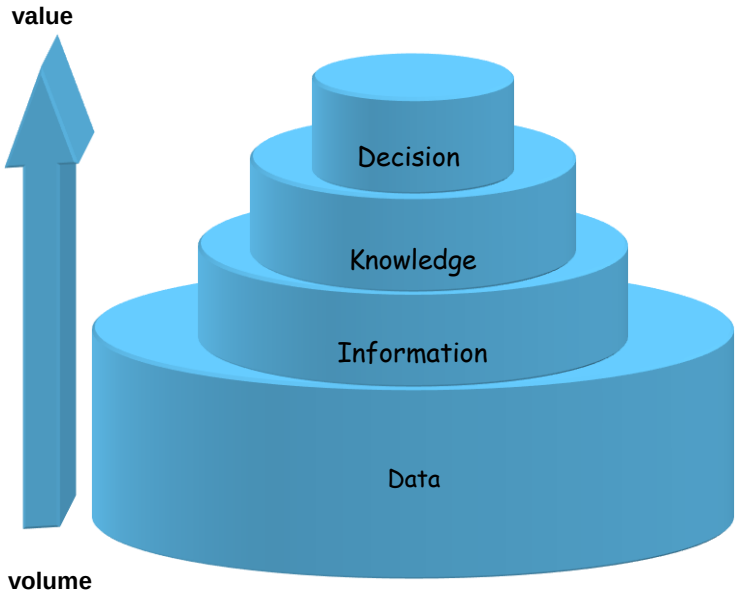


CINECA

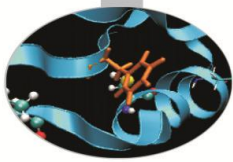


Data analytics

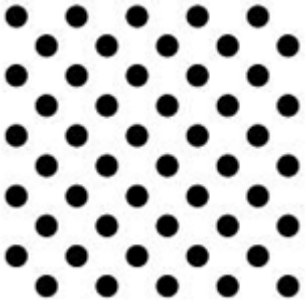
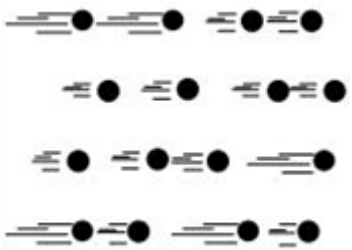
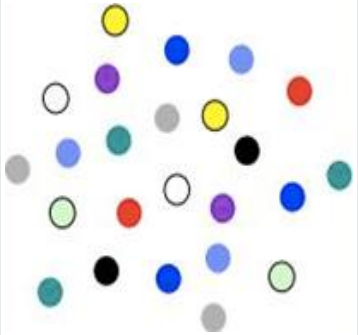
the process of extracting useful insights from raw data

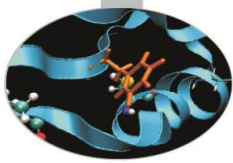


Same as … **Data Mining** (also known as Knowledge Discovery in Databases - KDD):
the process of discovering valuable information from very large databases using algorithms that discover hidden patterns in data
(1995)



Why is it challenging

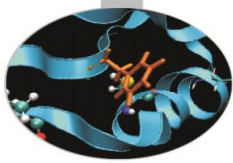
Volume	Velocity	Variety
		
Data at Rest	**Data in Motion**	**Data in Many Forms**
Terabytes to exabytes of existing data to process	**Streaming data, requiring mseconds to respond**	**Structured, unstructured, text, multimedia,…**



Going back to the definition …

the **process** of extracting valuable information from raw **data** using **algorithms** that discover hidden patterns

It's an **explorative approach** or **data driven approach** in contrast with "traditional" data analysis (statistics) that could also be hypothesis driven

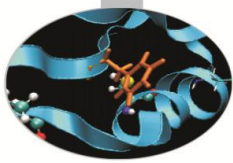


Topics

Data analytics

- data
- process
  - pre-processing
- algorithms / techniques

- *applications*

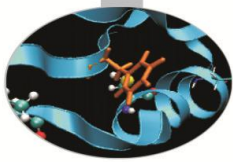


Data

The number and rate of data produced in any particular discipline now exceed our ability to effectively treat and analyse them

Sources:

- digital instruments
- high resolution cameras
- medical scanners
- simulations
- transactional data
- social media
- …

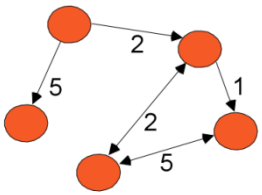


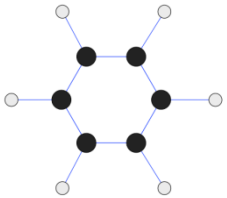
Data typologies

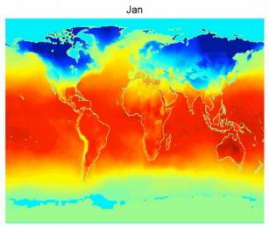
- structured data
  - data matrix
  - transactional data
- graph
  - web and social networks
  - molecular structures
- ordinal data
- spatial data
- time series
- sequences
  - genetic sequences
- unstructured data
  - textual documents
  - images
  - audio and videos (multimodal)

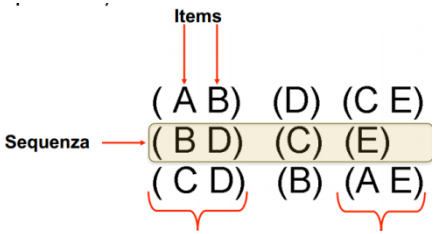
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

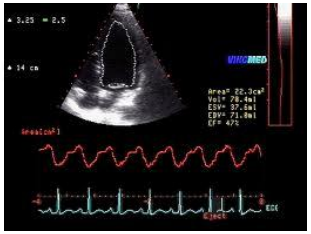


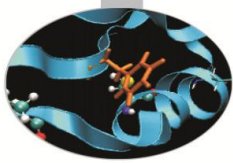




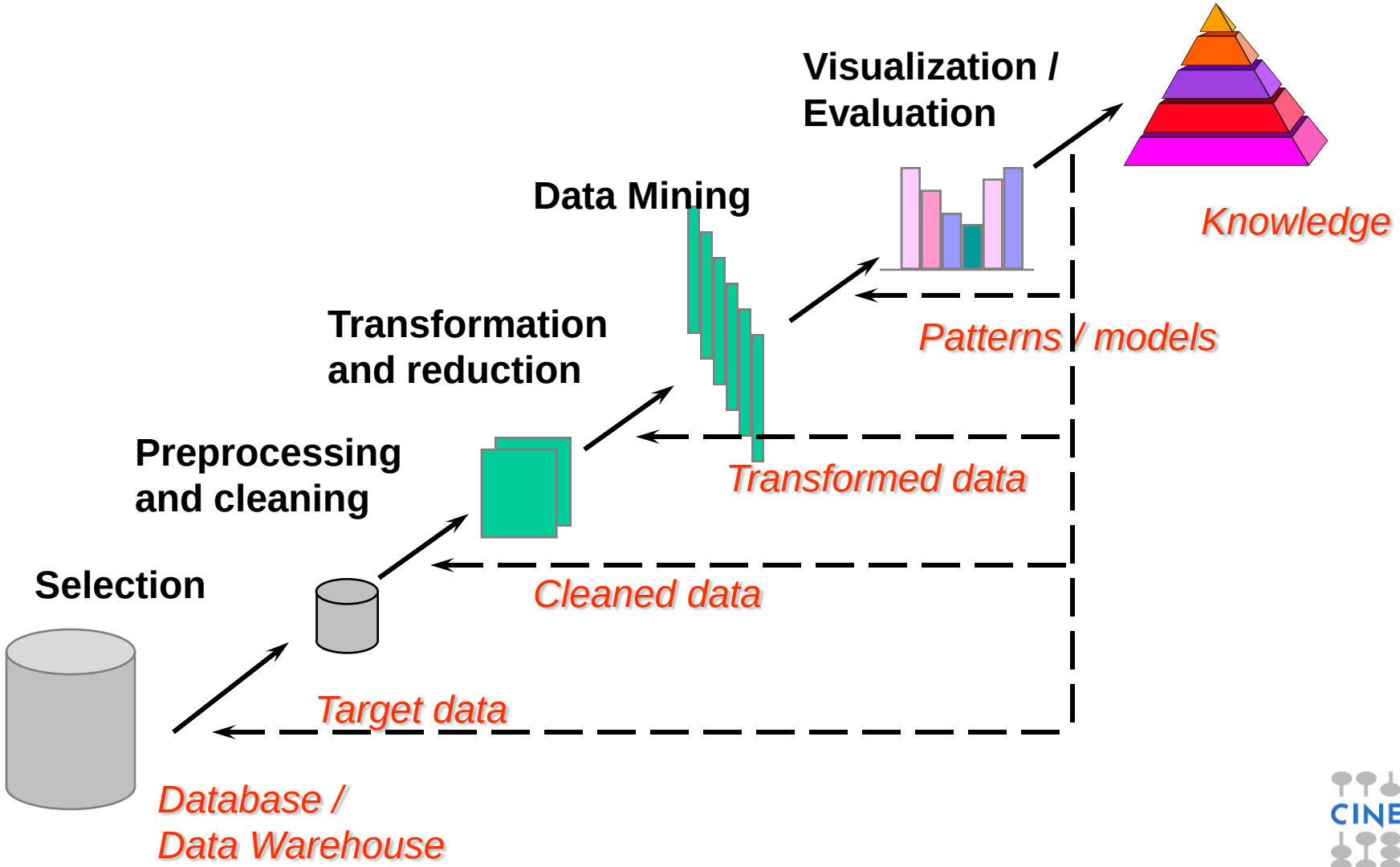


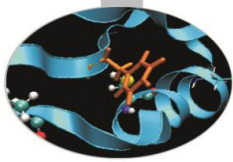

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```





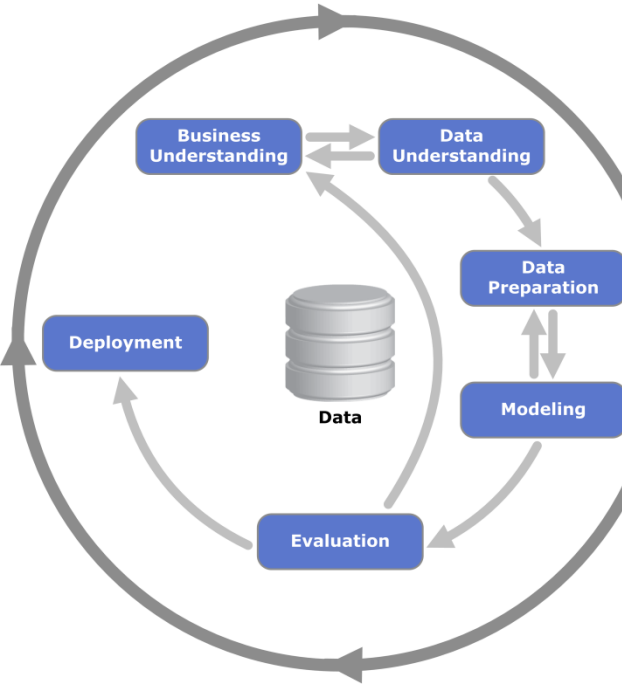
The Data Mining Process



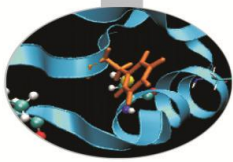


CRISP-DM reference model

Cross Industry Standard Process for Data Mining



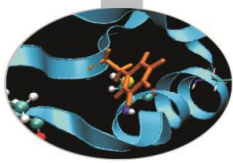
Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria*	**Collect Initial Data** *Initial Data Collection Report*	**Select Data** *Rationale for Inclusion/ Exclusion*	**Select Modeling Techniques** *Modeling Technique* *Modeling Assumptions*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models*	**Plan Deployment** *Deployment Plan*
Assess Situation *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits*	**Describe Data** *Data Description Report*	**Clean Data** *Data Cleaning Report*	**Generate Test Design** *Test Design*	**Review Process** *Review of Process*	**Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan*
Determine Data Mining Goals *Data Mining Goals* *Data Mining Success Criteria*	**Explore Data** *Data Exploration Report*	**Construct Data** *Derived Attributes* *Generated Records*	**Build Model** *Parameter Settings* *Models* *Model Descriptions*	**Determine Next Steps** *List of Possible Actions* *Decision*	**Produce Final Report** *Final Report* *Final Presentation*
Produce Project Plan *Project Plan* *Initial Assessment of Tools and Techniques*	**Verify Data Quality** *Data Quality Report*	**Integrate Data** *Merged Data*	**Assess Model** *Model Assessment* *Revised Parameter Settings*		**Review Project** *Experience Documentation*
		Format Data *Reformatted Data* *Dataset* *Dataset Description*			



Is it still the reference model? (1)

New challenges

- The CRISP model reflects a data management perspective where all relevant information can be stored and cleaned before any further manipulation. This assumption might be easily violated in all those cases where the data flow is too massive to allow an **exhaustive storage** (filtering/compressing data on the fly to allow that would require some awareness of the analyses expected afterward) or when there are timeliness constraints.
- The CRISP model suggests a flat approach. Mastering the data variety and complexity requires several **levels of analysis**, combining the results of various processing tools to obtain complex patterns or models, to form hierarchical dependencies among the steps performed.

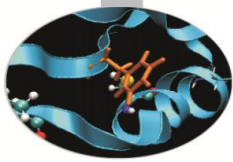


Is it still the reference model? (2)

New challenges

- In complex applications, the design of an analytical process is actually a **multi-disciplinary** effort that involves actors with different backgrounds.
- The **computational complexity** requires new scalable algorithms and the distribution of workloads on clusters (eg MapReduce) or on cloud.
- Big Data Analytics often involve the use of personal data, ranging from medical records to location information, activity records on social networks, web navigation and searching history, etc. All this calls for mechanism that ensure that the information flow employed in the analyses does not harm the **privacy** of individuals.

Is it still the reference model? (3)



New challenges

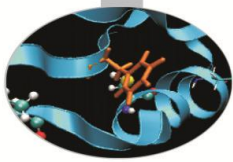
- **Data integration** from multiple and heterogeneous sources.
- **Data quality**.
- Models **fast adapting** to temporal changes.

New enphasis on

- **Re-purposing data** that was collected for a different purpose.
- **Re-purposing algorithms** (e.g. page rank on graphs).
- **Data products**: data driven applications (e.g. spell checkers, machine translation, recommendation systems, …) interactive visualizations, online databases.

⇨ Not just answering the question once, empower others to use data in new ways

Is it still the reference model? (4)



e-Science

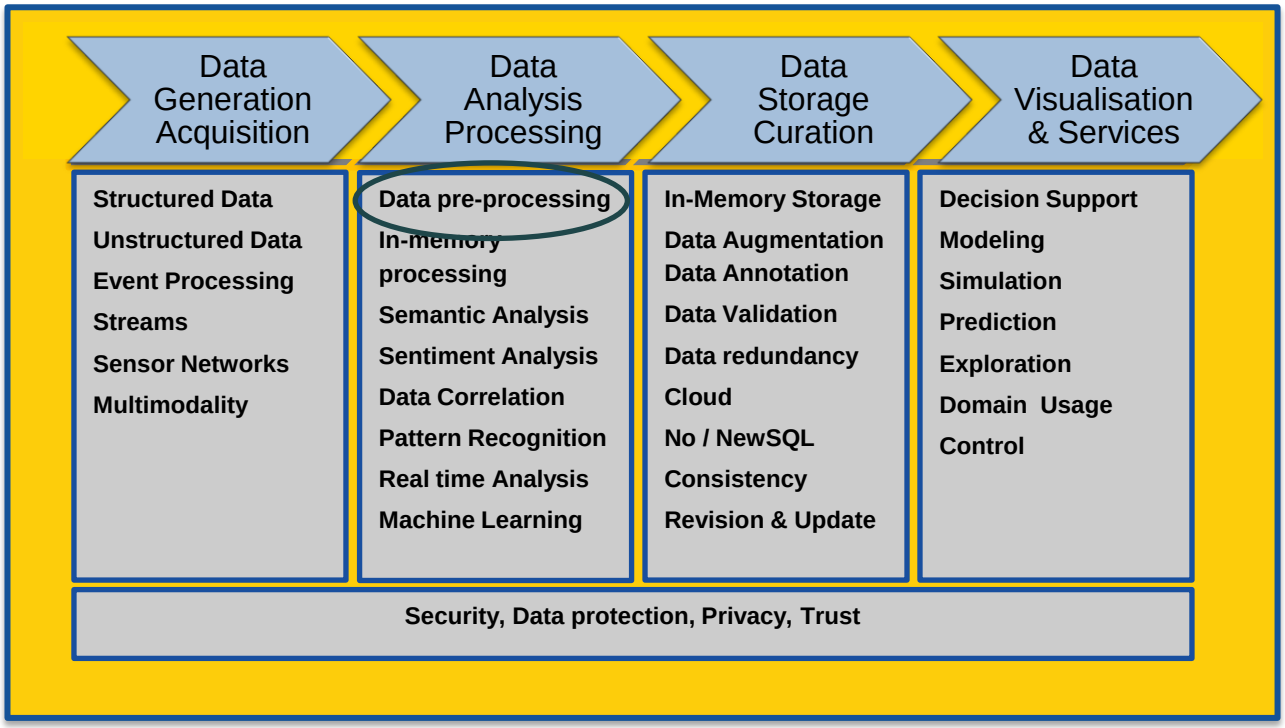
4th paradigm of scientific inquiry:

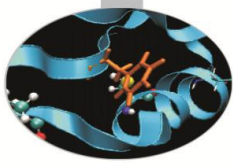
to acquire massive data sets from instruments or from simulations

- e-Science is **driven by data** more than by the computation
- **data analysis** has replaced data acquisition as the new bottleneck to discovery

Another way of describing the process (BDVA)

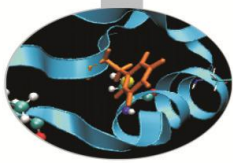
data analysis output can be input for other higher level analysis





Pre-processing

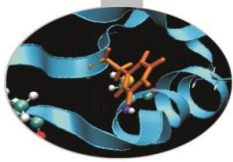
- data understanding and data quality assessment (evaluation of data accuracy and reliability, completeness, consistence, … correlation)
  - Presence of missing values, outliers, inconsitencies
  - Level of noise
  - Redundance
- data preparation
  - Cleaning
  - Transformation (normalization, discretization, aggregation, new variables computation…)
  - Feature extraction
  - Selection / filtering



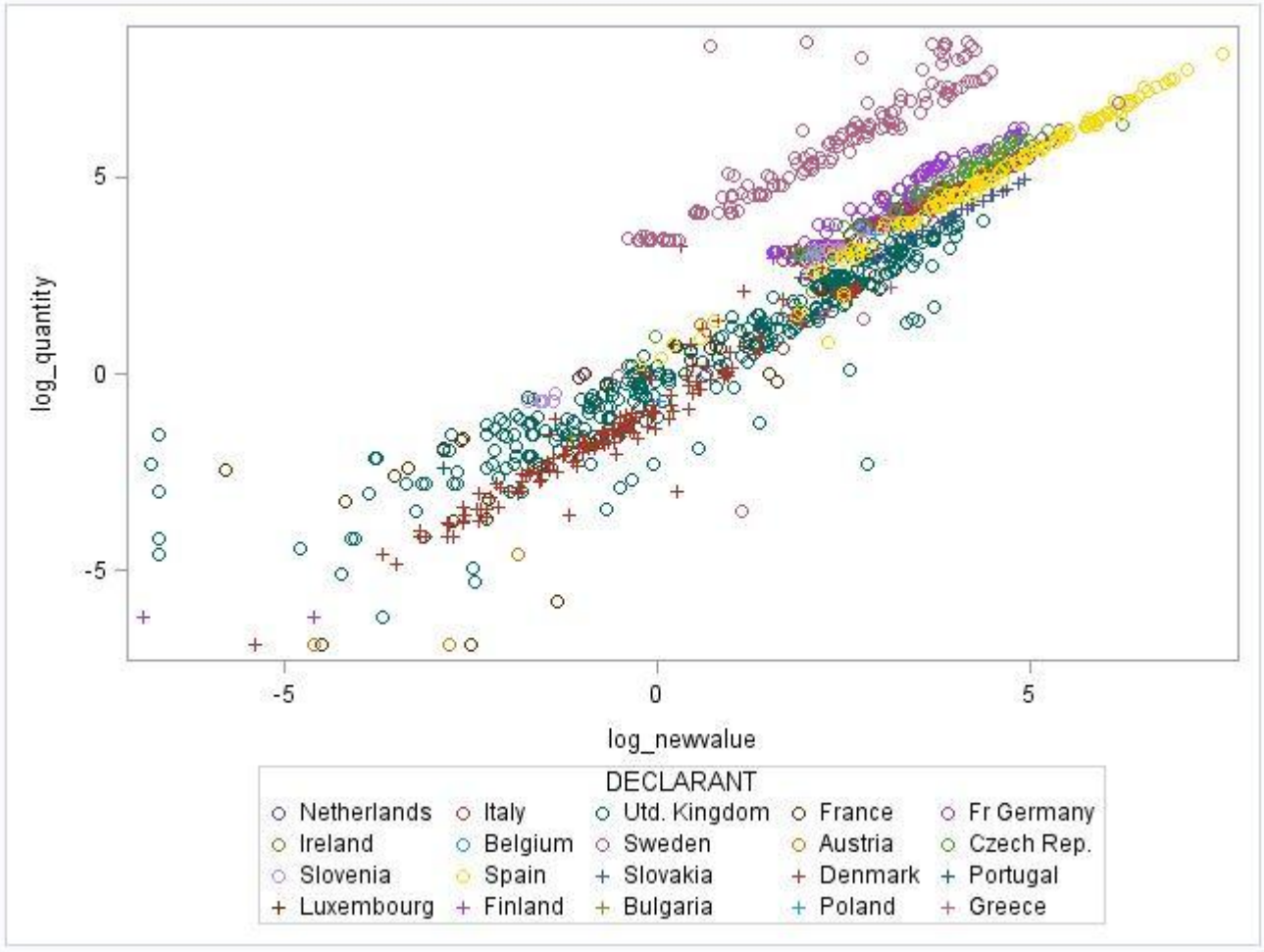
Pre–processing

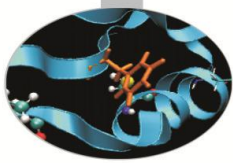
Why is it useful - a few examples

- L'Equité: high peak of 96 years old insured
  - missing birth dates had been codified 1/1/1900
- Trento University: a high number of students with very low grades in the high school diplomas
  - grades in the high school diplomas have undergone a scale change (from 60 as a maximum to 100)
- Local Health Service: high consumpion of cardiovascular drugs in diabetics
  - the quantity of active ingredient for cardiovascular drugs was in milligrams (instead of grams)
- Eurostat: visual patterns of outliers
  - the Country was a key variable in international trade outliers identification



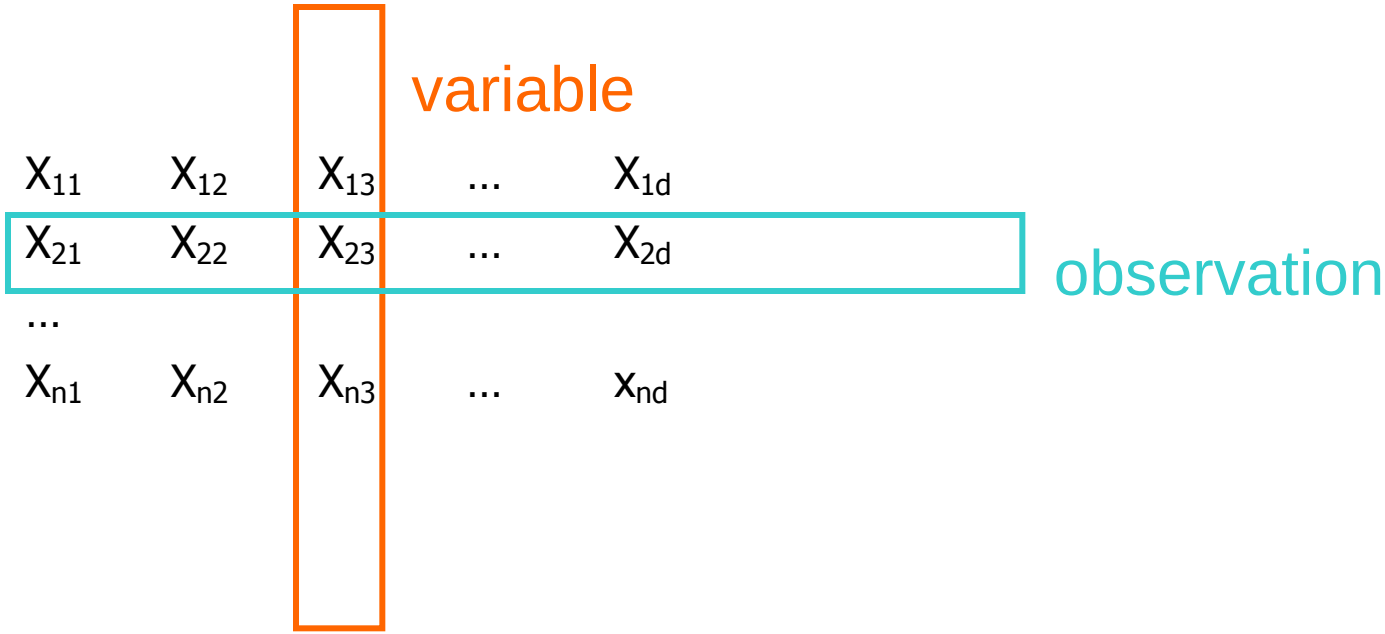
Pre–processing
Ask the right question



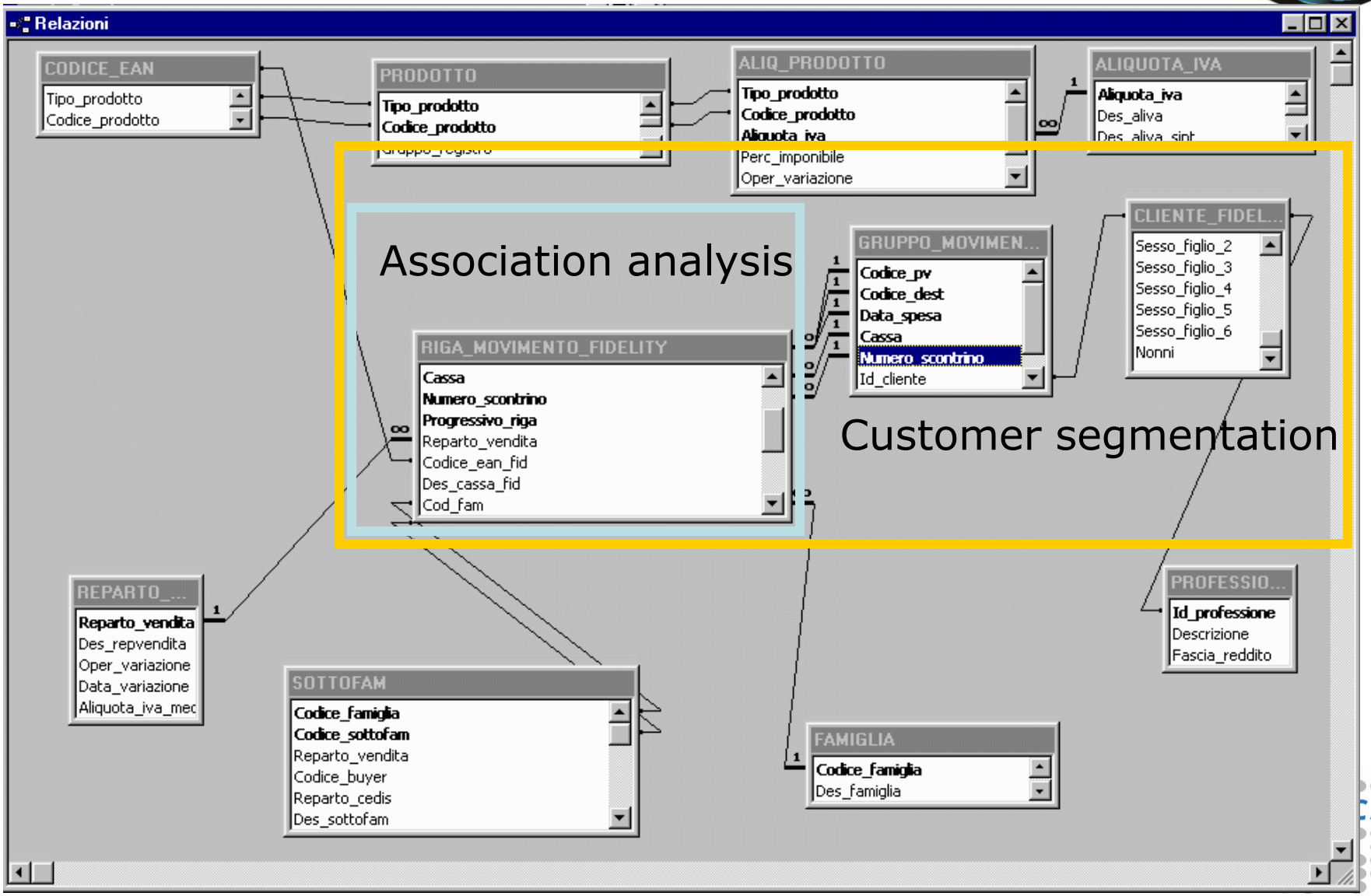


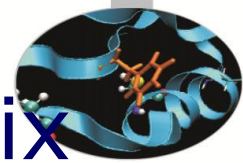
Data representation

Analysis matrix



Coal: data structure



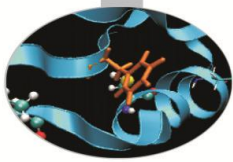


Coal: customer segmentation matrix

- variables describing the buyer behavior:
  - items list (only the characterizing, distinguishing items)
  - number of receipts
  - average number of items per receipt
  - average expense
  - percentage of items having a promotion
- socio-demographic variables:
  - genre
  - age
  - job
  - marital status
  - number of sons
  - number of children
  - cats
  - dogs

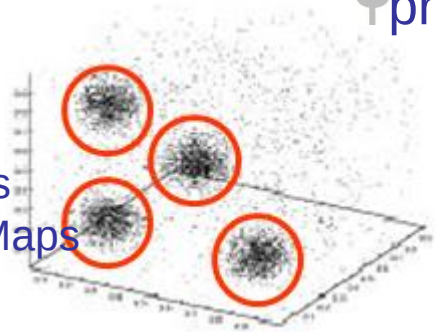
"active" variables

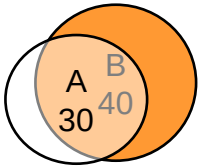
"descriptive" variables

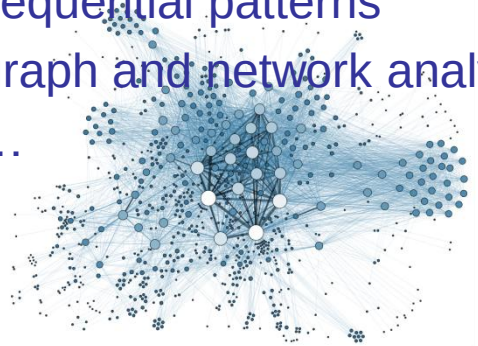


Tasks and techniques

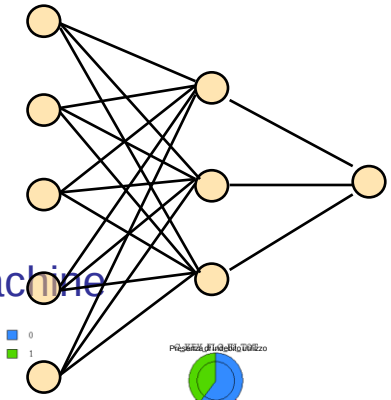
- descriptive
  - clustering
    - k-means
    - relational analysis
    - Self Organizing Maps
    - …
  - association rules
  - sequential patterns
  - graph and network analysis
  - …

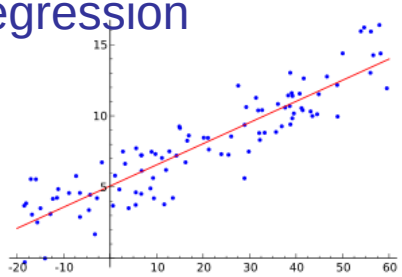


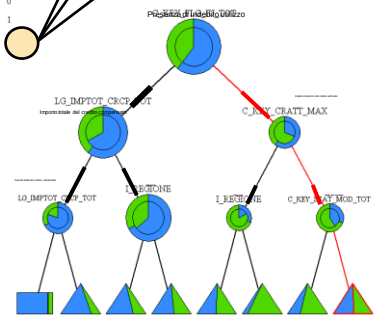




- predictive
  - classification (machine learning)
    - Naive Bayes
    - Decision Trees
    - Neural Networks
    - KNN
    - Rocchio
    - Support Vectors Machine
    - …
  - regression

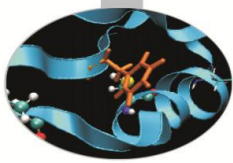






Unsupervised learning
training samples have no class information
guess classes or clusters in the data

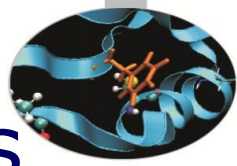
Supervised learning
use training samples with known classes
to classify new data



Terminology

- Supervised learning (“Training”)
  - we are given examples of inputs an associated outputs
  - we learn the relationship between them
- Unsupervised learning (sometimes “Mining”)
  - we are given inputs but no outputs
    - unlabeled data
  - we learn the “latent” labels

  (e.g. clustering, dimensionality reduction)



Tasks, techniques and applications

- descriptive
  - clustering
    - k-means
    - relational analysis
    - Self Organizing Maps
    - …
  - association rules
  - sequential patterns
  - graph and network analysis

Customer segmentation
Thematic grouping
Market Basket Analysis
Social Network Analysis
….

- predictive
  - classification (machine learning)
    - Naive Bayes
    - Decision Trees
    - Neural Networks
    - KNN
    - Rocchio
    - Support Vectors Machine
    - …
  - regression

Churn analysis
Fraud detection
Prospect identification
Recommendation systems
Document classification
…