Introduction to Data Analytics

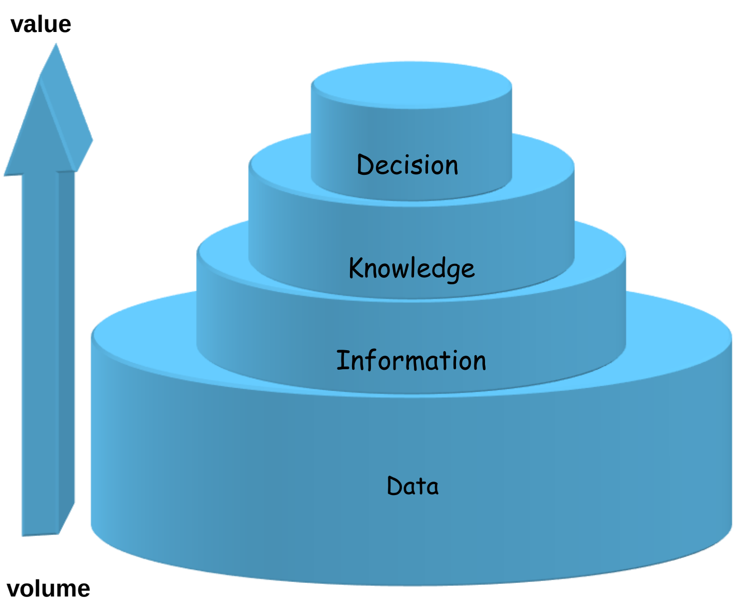
School on Scientific Data Analytics and Visualization

Roberta Turra, *Cineca*

21 June 2016

Data analytics

process of extracting useful insights from raw data



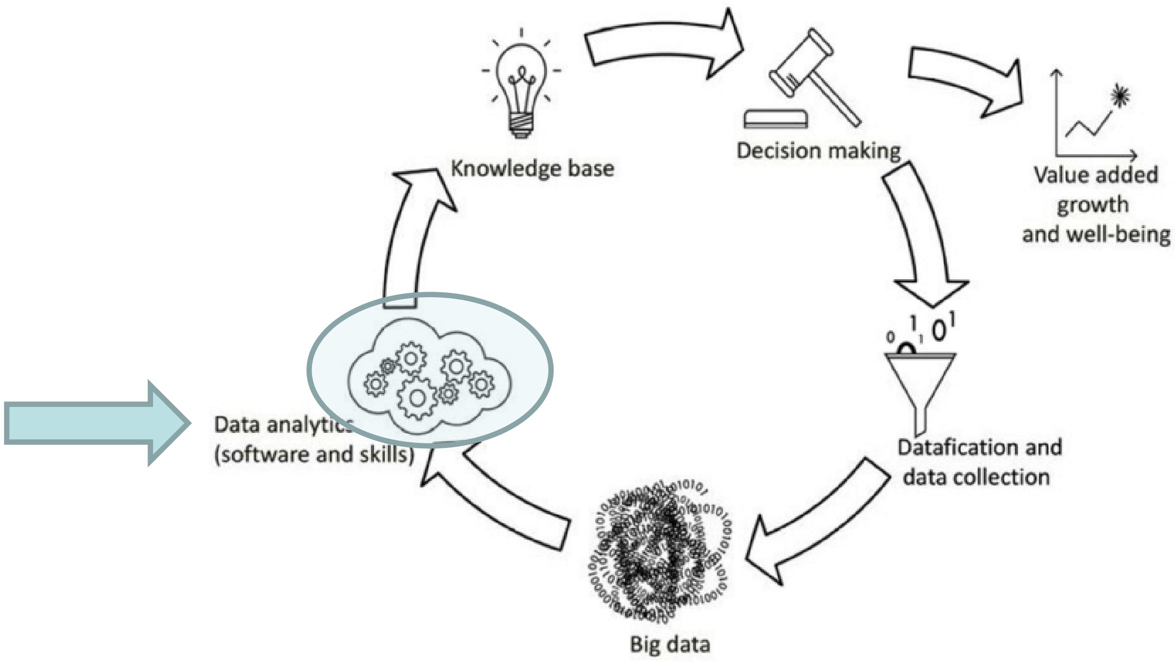
Same as … **Data Mining** (also known as Knowledge Discovery in Databases - KDD):
the process of discovering valuable information from very large databases using algorithms that discover hidden patterns in data (1995)

The data value cycle

OECD report on Data-Driven Innovation

(Big Data for Growth and Well-Being)

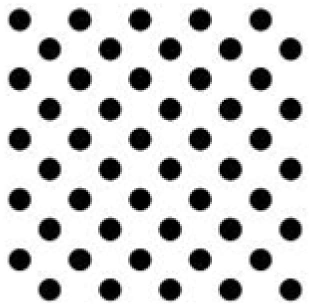
Figure 1.7. **The data value cycle**



SuperComputing Applications and Innovation

Why is it challenging

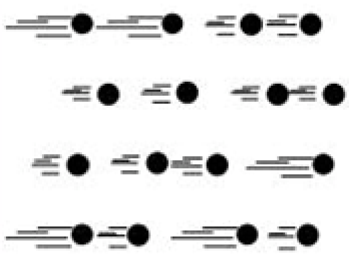
Volume



Data at Rest

Terabytes to exabytes of existing data to process

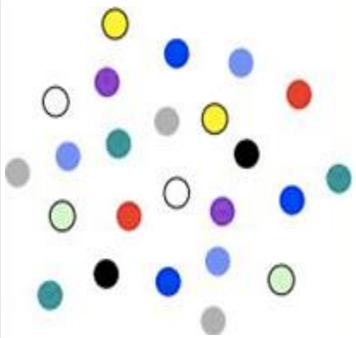
Velocity



Data in Motion

Streaming data, requiring mseconds to respond

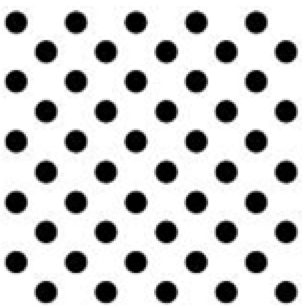
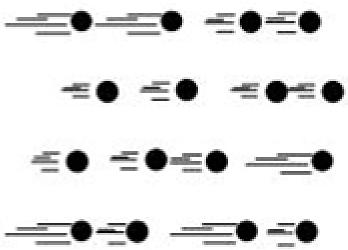
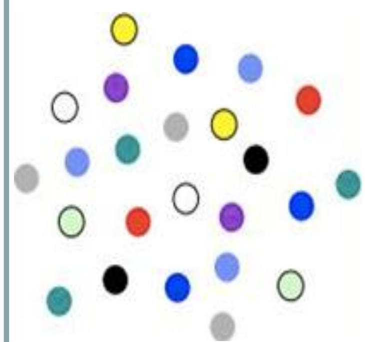
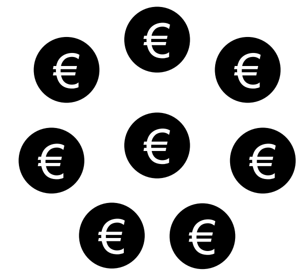
Variety



Data in Many Forms

Structured, unstructured, text, multimedia,…

The 5Vs

Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt	Data into Money
Terabytes to exabytes of existing data to process	**Streaming data, requiring mseconds to respond**	**Structured, unstructured, text, multimedia,…**	**Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception**	**Business models can be associated to the data**

Going back to the definition …

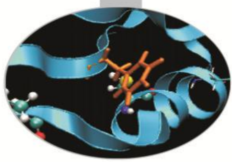
process of extracting valuable information
from raw **data** using **algorithms** that discover
hidden patterns

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

Agenda

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

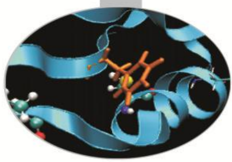
- data
- process
  - pre-processing
- algorithms / techniques



Data

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

- Internet
  - massive search engines
  - e-commerce
  - social media
  - mobile devices
- Sensor networks
- Scientific data
  - simulations (probing extreme phenomena, e.g. particle physics)
  - digital instruments (exploratory approach to let new phenomena emerge, e.g. genome sequencing, large telescopes, …)



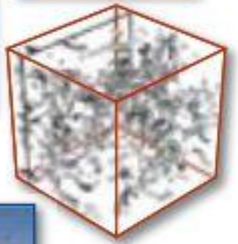
The rapid growth in data

The Fourth Paradigm: Data-Intensive Scientific Discovery

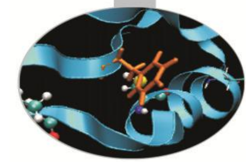
Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$



- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  - Data captured by instruments or generated by simulator
  - Processed by software
  - Information/knowledge stored in computer
  - Scientist analyzes database / files using data management and statistics



The rapid growth in data

Science is about asking questions

traditionally: "*query the world*"

Data acquisition activities coupled to a specific hypothesis

eScience: "*download the world*"

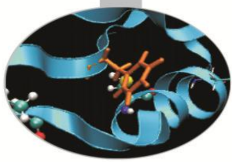
Data acquired massively in support of many hypotheses

The cost of data acquisition has dropped precipitously thanks to advances in technology

- Astronomy: high-resolution, high-frequency sky surveys
- Life Sciences: lab automation, high-throughput sequencing
- Oceanography: high-resolution models, cheap sensors, satellites

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

CINECA SCAI
SuperComputing Applications and Innovation



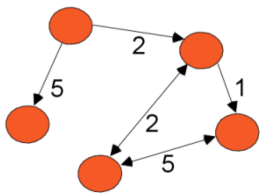
Data typologies

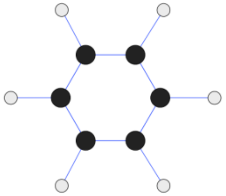
- structured data
  - data matrix
  - transactional data

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

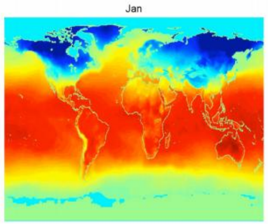
TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

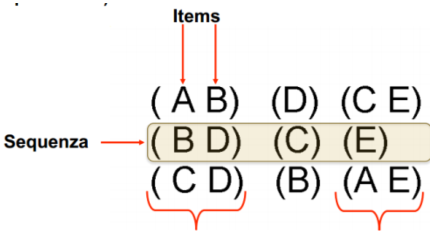
- graph
  - web and social networks
  - molecular structures





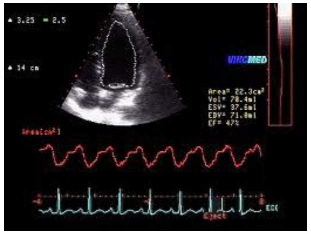
- ordinal data
- spatial data
- time series
- sequences
  - genetic sequences





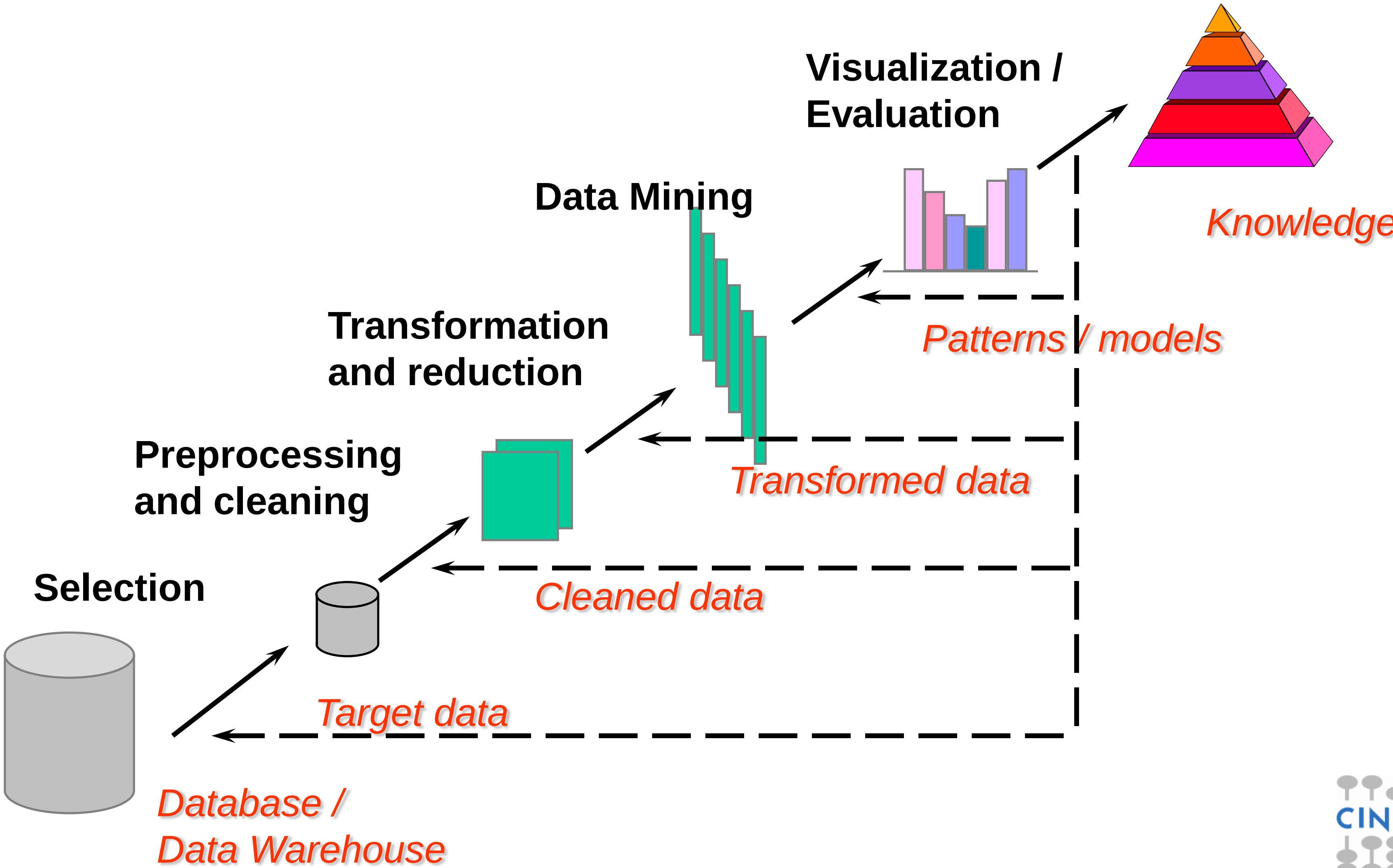
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

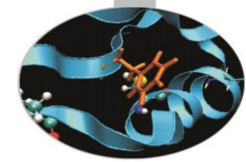
- unstructured data
  - textual documents
  - images
  - audio and videos (multimodal)



CINECA

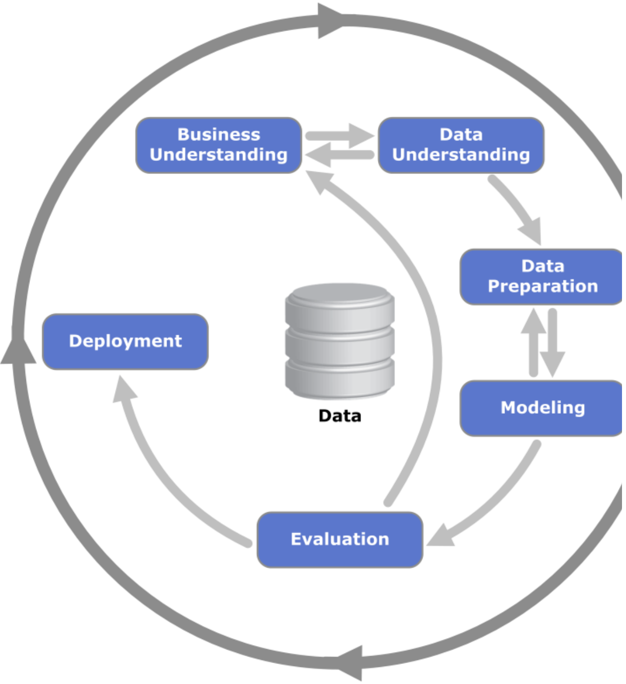
The Data Mining Process





CRISP-DM reference model

Cross Industry Standard Process for Data Mining



Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of Resources Requirements, Assumptions, and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Select Data
Rationale for Inclusion/ Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Dataset
Dataset Description

Modeling

Select Modeling Techniques
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Descriptions

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

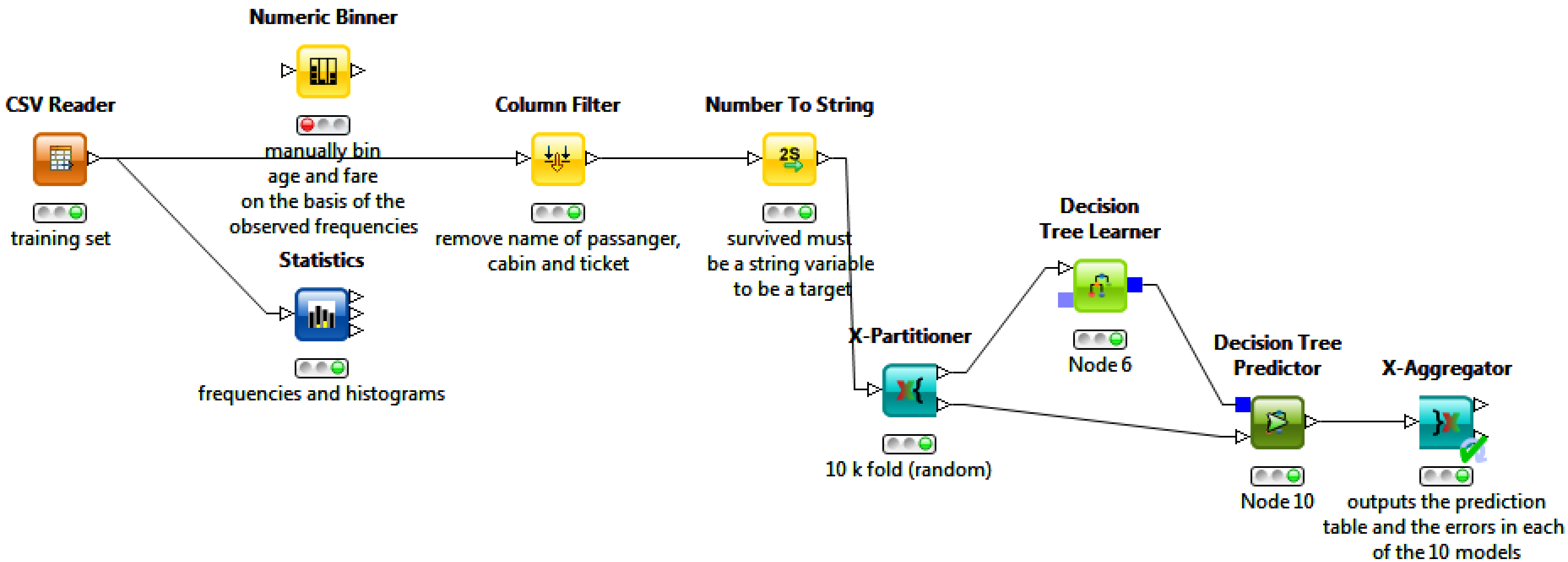
Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

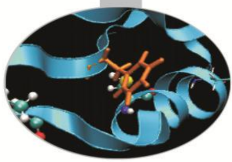
The process – Knime Workflow



Is it still the reference model? (1)

New challenges

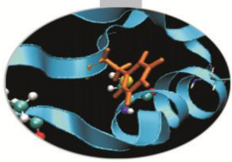
- The CRISP model reflects a data management perspective where all relevant information can be stored and cleaned before any further manipulation. Often the data flow is too massive to allow an **exhaustive storage** (filtering/compressing data on the fly to allow that would require some awareness of 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 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 -> Turning data into product

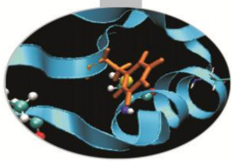
Not just answering the question once, empower others to use data in new ways

Another way of describing the process (BDVA)

data analysis output can be input for other higher level analysis

Data Generation Acquisition	Data Analysis Processing	Data Storage Curation	Data Visualisation & Services
Structured Data	Data pre-processing	In-Memory Storage	Decision Support
Unstructured Data	In-memory processing	Data Augmentation	Modeling
Event Processing	Semantic Analysis	Data Annotation	Simulation
Streams	Sentiment Analysis	Data Validation	Prediction
Sensor Networks	Data Correlation	Data redundancy	Exploration
Multimodality	Pattern Recognition	Cloud	Domain Usage
	Real time Analysis	No / NewSQL	Control
	Machine Learning	Consistency	
		Revision & Update	

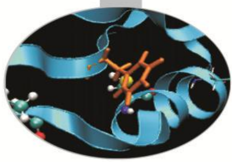
Security, Data protection, Privacy, Trust



SuperComputing Applications and Innovation

Pre–processing

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



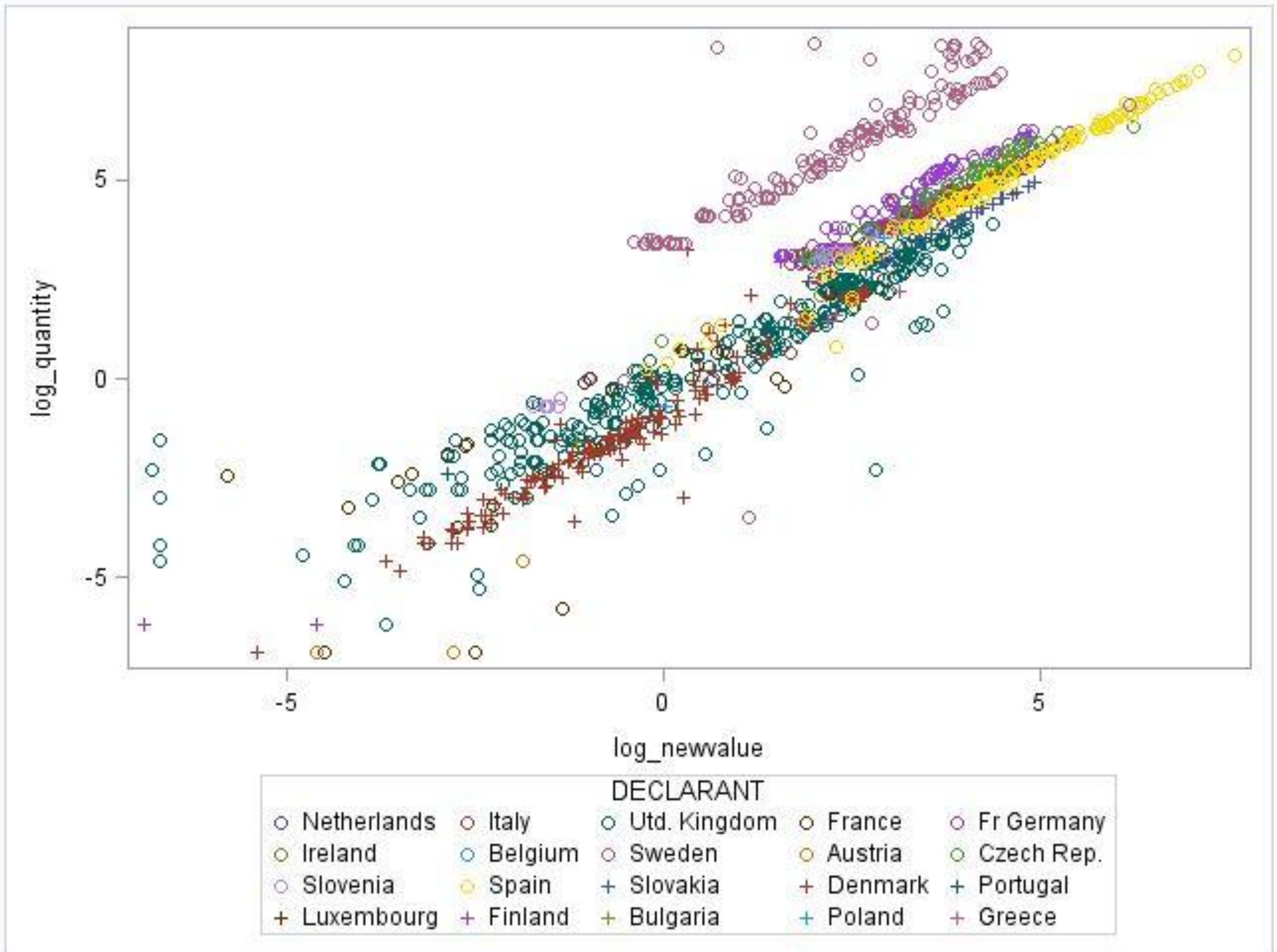
Pre–processing

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

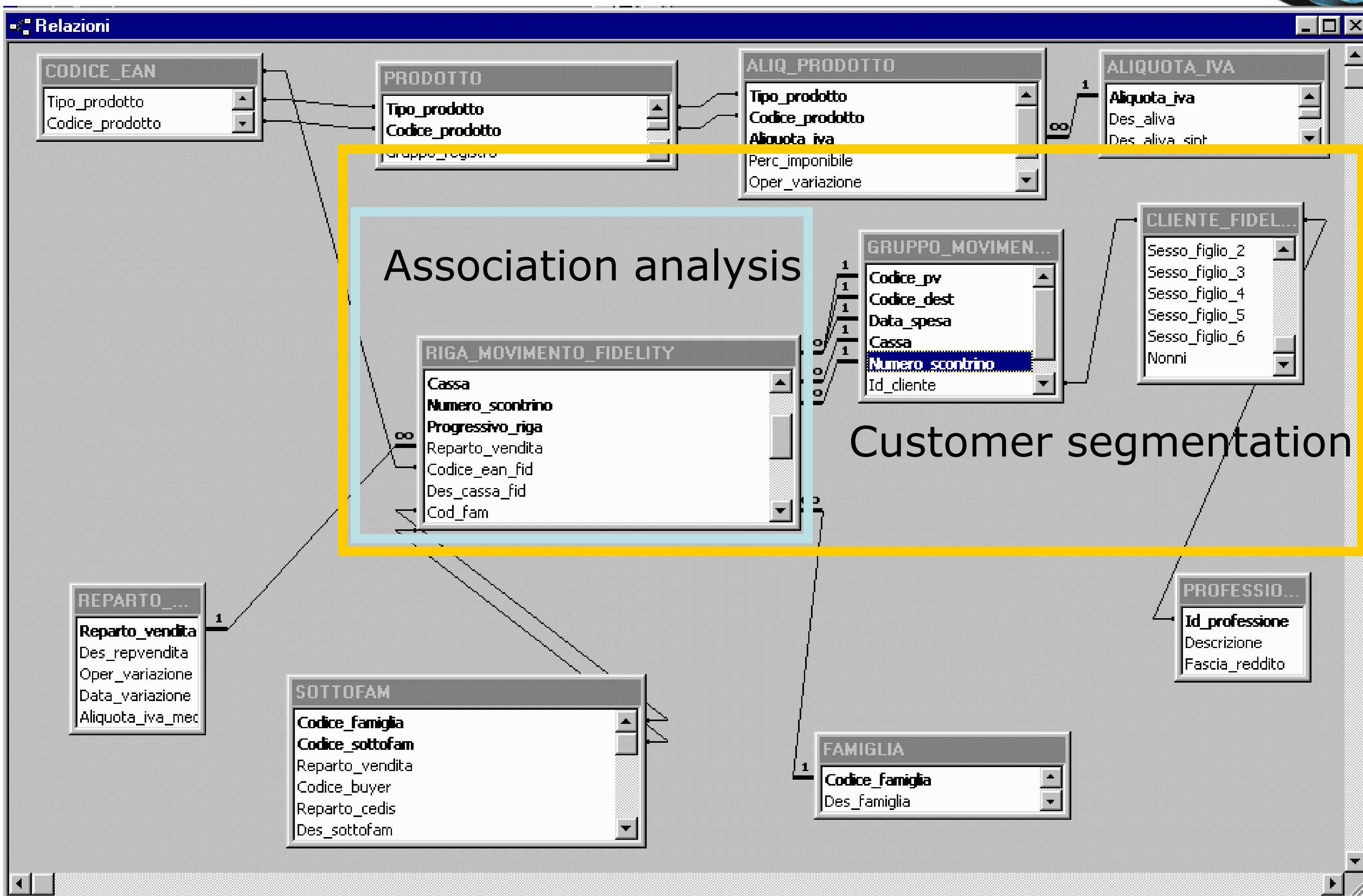
Analysis matrix

variable

$$
\begin{array}{ccccc}
X_{11} & X_{12} & X_{13} & \dots & X_{1d} \\
X_{21} & X_{22} & X_{23} & \dots & X_{2d} \\
\dots & & & & \\
X_{n1} & X_{n2} & X_{n3} & \dots & x_{nd}
\end{array}
$$

observation

Coal: data structure



SuperComputing Applications and Innovation

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

The process in text mining

- collecting
- indexing
- mining
- evaluation

Collecting

- document selection
  - Document collection from multiple sources
    - retreiving from DBs (query)
    - downloading (through API)
    - web crawling / web scraping
- pre – processing
  - parsing
  - integration
  - transformation to a common format

Indexing

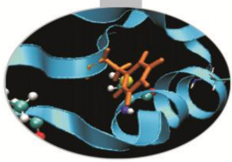
- document preparation (**indexing**)
  - tokenization
  - Part Of Speech tagging
  - selection of terms (nouns, verbs, adjectives, …)
  - stemming / lemmatization
  - chunking (n-grams, nominal phrases)
  - weighting (binary, frequencies, tfidf, …)
  - stop-words filtering
  - dimensionality reduction
  - meta-information tagging

tn.5.26.35 SOURCE Reuters
tn.5.26.35 DATE 6/21/2000
tn.5.26.35 MONTHYEAR 2000_06
tn.5.26.35 SUBJECTS Japan
tn.5.26.35 SUBJECTS Passenger_Vehicles
tn.5.26.35 SUBJECTS Safety
tn.5.26.35 STATE Japan
tn.5.26.35 LANGUAGE English
tn.5.26.35 ORG2 TOYOTA
tn.5.26.35 NN area
tn.5.26.35 NN automobile
tn.5.26.35 NN average
tn.5.26.35 NN barrier
tn.5.26.35 NN car
tn.5.26.35 NN chest
tn.5.26.35 NN compartment
tn.5.26.35 NN crash
tn.5.26.35 NN driver
tn.5.26.35 NN dummy
tn.5.26.35 NN foot
tn.5.26.35 NN force
tn.5.26.35 NN group
tn.5.26.35 NN head

tn.5.26.35 NN hour
tn.5.26.35 NN impact
tn.5.26.35 NN injury
tn.5.26.35 NN insurer
tn.5.26.35 NN intrusion
tn.5.26.35 NN likelihood
tn.5.26.35 NN luxury
tn.5.26.35 NN mark
tn.5.26.35 NN mile
tn.5.26.35 NN neck
tn.5.26.35 NN offset
tn.5.26.35 NN passenger
tn.5.26.35 NN potential
tn.5.26.35 NN rating
tn.5.26.35 NN risk
tn.5.26.35 NN safety
tn.5.26.35 NN score
tn.5.26.35 NN sedan
tn.5.26.35 NN side
tn.5.26.35 NN sport
tn.5.26.35 NN test
tn.5.26.35 NN utility
tn.5.26.35 NN vehicle

tn.5.26.35 UTERM crash_test
tn.5.26.35 UTERM top_score
tn.5.26.35 ORG honda_motor_co
tn.5.26.35 ORG insurance_institute for …
tn.5.26.35 ORG isuzu_motors
tn.5.26.35 ORG mazda_motor
tn.5.26.35 ORG nissan_motor
tn.5.26.35 ORG toyota_motor
tn.5.26.35 UNAME avalon
tn.5.26.35 UNAME honda_passport
tn.5.26.35 UNAME infiniti_i30
tn.5.26.35 UNAME maxima
tn.5.26.35 UNAME mazda_mpv
tn.5.26.35 UNAME rodeo

SuperComputing Applications and Innovation



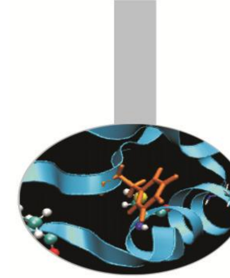
Data representation

The result of the indexing phase is a document vector (a sequence of terms and tags).

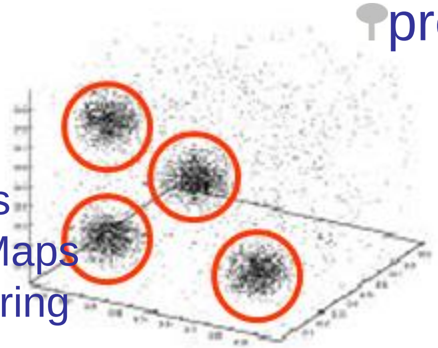
All document vectors are then converted to a common format: the analysis matrix.

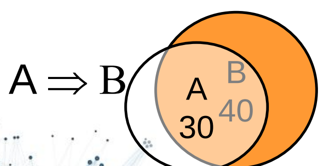
	team	coach	play	ball	score
Document 1	3	0	5	0	2
Document 2	0	7	0	2	1
Document 3	0	1	0	0	1

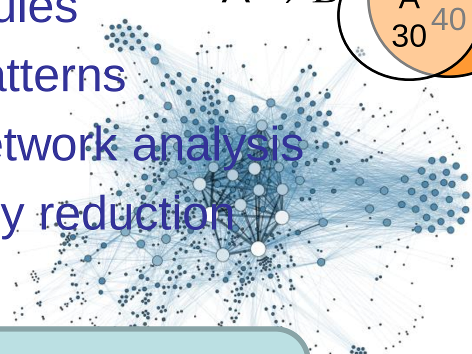
Tasks and techniques



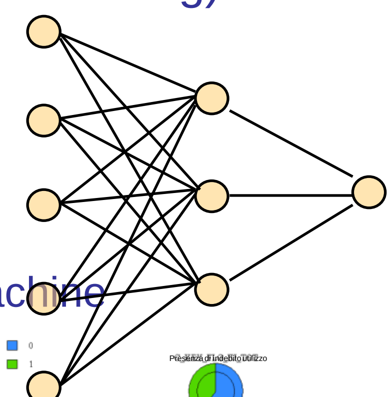
- descriptive
  - clustering
    - k-means
    - relational analysis
    - Self Organizing Maps
    - hierachical clustering
    - mixture model
    - …
  - association rules
  - sequential patterns
  - graph and network analysis
  - dimensionality reduction
  - …

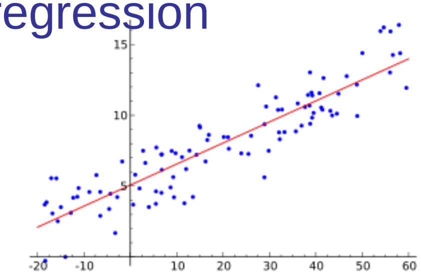


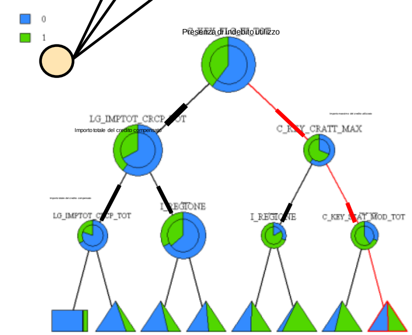




- predictive
  - classification (machine learning)
    - Naive Bayes
    - Decision Trees
    - Neural Networks
    - KNN
    - Rocchio
    - Support Vectors Machine
    - …
  - regression







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)