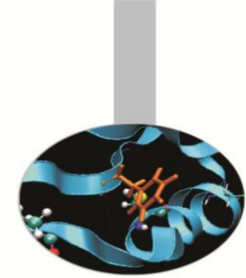
Introduction to Data Analytics

3rd School on Scientific Data Analytics and Visualization

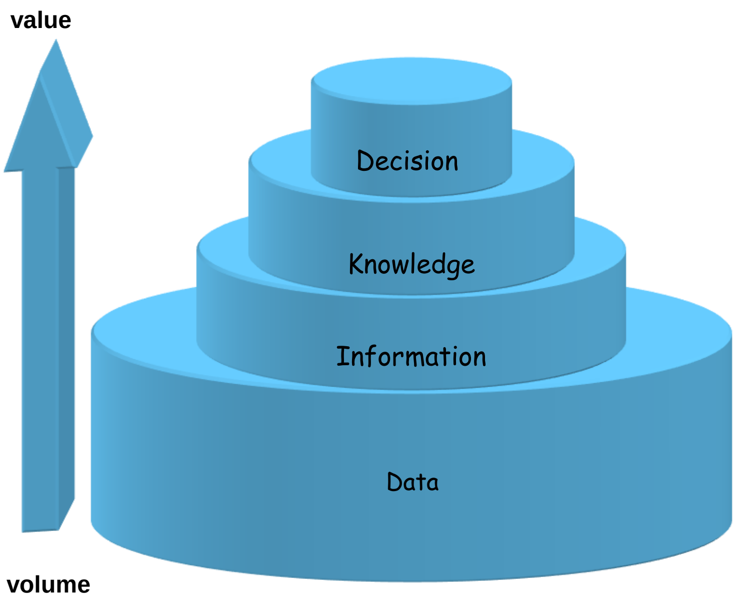
Roberta Turra, *Cineca*

12 June 2017



Data analytics

process of extracting useful insights from raw data



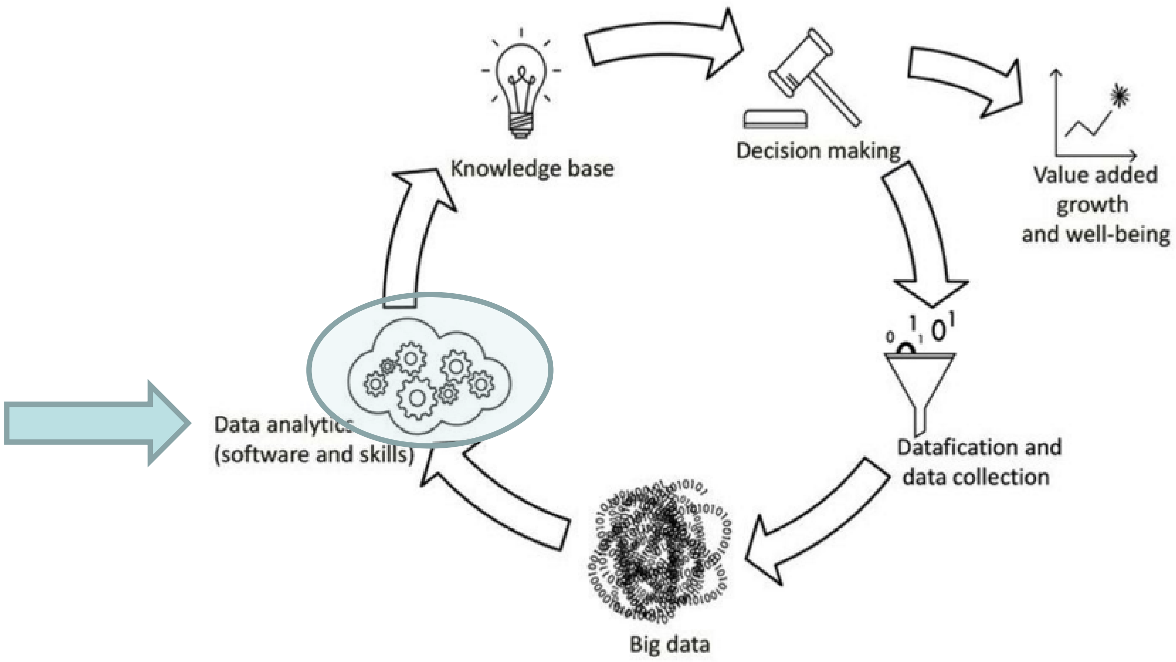
Same as … **Data Mining** (also known as Knowledge Discovery in Databases - KDD):
the process of discovering valuable information from very large databases using algorithms that discover hidden patterns in data (1995)

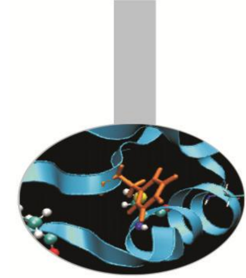
The data value cycle

OECD report on Data-Driven Innovation

(Big Data for Growth and Well-Being)

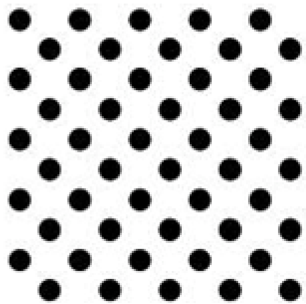
Figure 1.7. **The data value cycle**





Why is it challenging

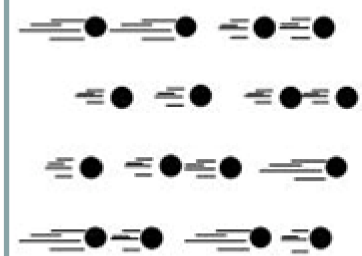
Volume



Data at Rest

Terabytes to exabytes of existing data to process

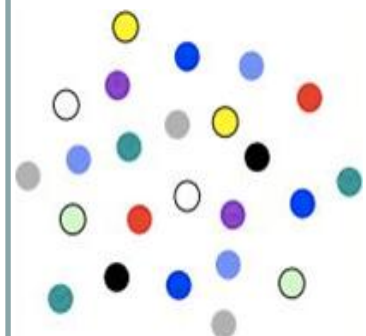
Velocity



Data in Motion

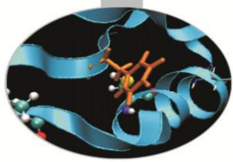
Streaming data, requiring mseconds to respond

Variety

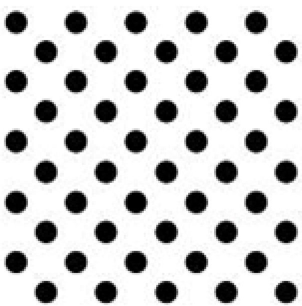
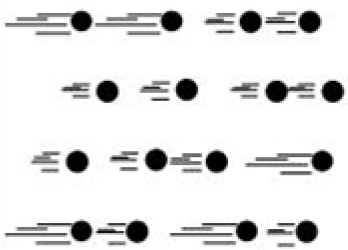
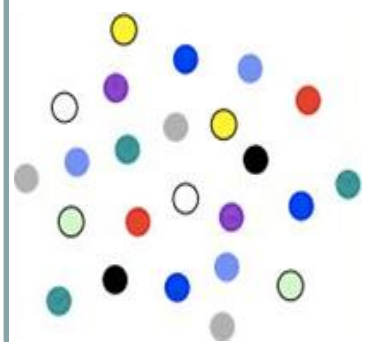
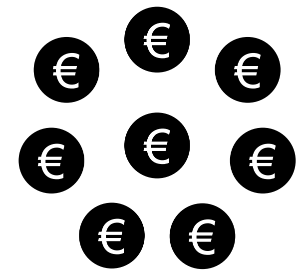


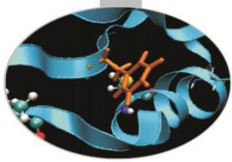
Data in Many Forms

Structured, unstructured, text, multimedia,…



The 5Vs

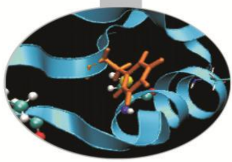
Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	Data in Motion	Data in Many Forms	Data in Doubt	Data into Money
Terabytes to exabytes of existing data to process	**Streaming data, requiring mseconds to respond**	**Structured, unstructured, text, multimedia,…**	**Uncertainty due to data inconsistency & 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

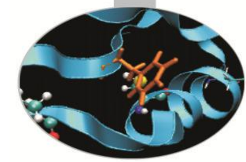
It’s an **explorative approach** or **data driven approach**
in contrast with “traditional” data analysis (in statistics) that could
also be hypothesis driven



Agenda

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

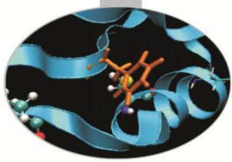
- data
- process
  - pre-processing
- algorithms / techniques



Data

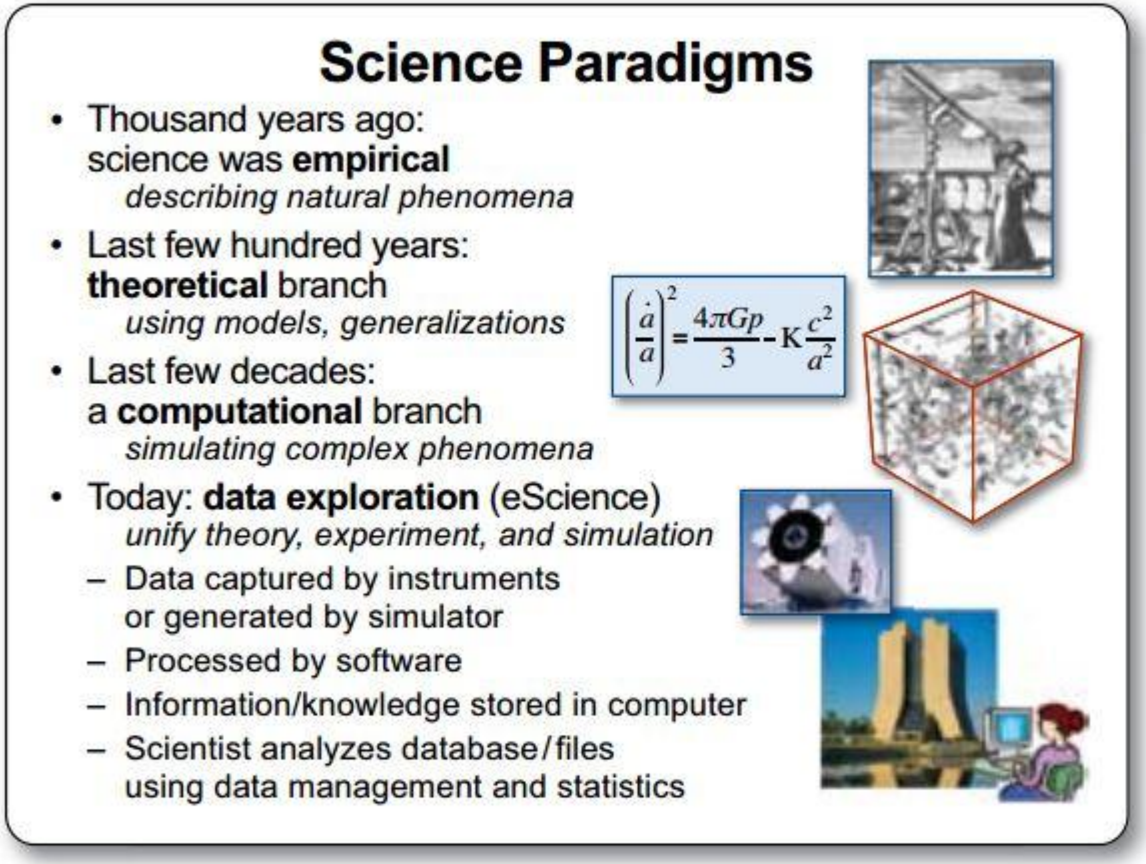
The volume and rate of data produced in any particular discipline now exceed our ability to effectively treat and analyse them

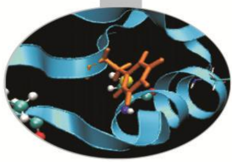
- Internet
  - massive search engines
  - e-commerce
  - social media
  - mobile devices
- Sensor networks
- Scientific data
  - simulations (probing extreme phenomena, e.g. particle physics)
  - digital instruments (exploratory approach to let new phenomena emerge, e.g. genome sequencing, large telescopes,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



The rapid growth in data

The Fourth Paradigm: Data-Intensive Scientific Discovery





The rapid growth in data

Science is about asking questions

traditionally: "*query the world*"

Data acquisition activities coupled to a specific hypothesis

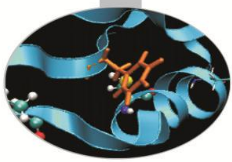
eScience: "*download the world*"

Data acquired massively in support of many hypotheses

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

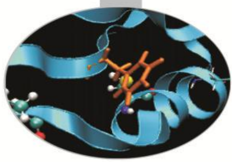
- Astronomy: high-resolution, high-frequency sky surveys
- Life Sciences: lab automation, high-throughput sequencing
- Oceanography: high-resolution models, cheap sensors, satellites

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



Data as an infrastructure

Data has become the key infrastructure for 21st century knowledge economies. Data are not the “new oil”, they are rather an infrastructure and capital good that can be used across society for a theoretically unlimited range of productive purposes, without being depleted.



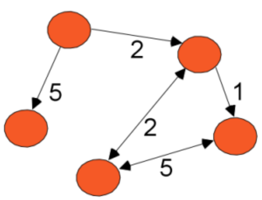
Data typologies

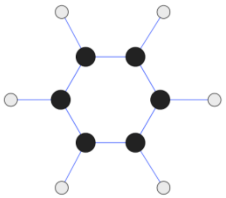
- structured data
  - data matrix
  - transactional data

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

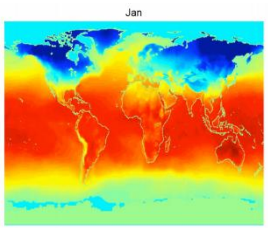
TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

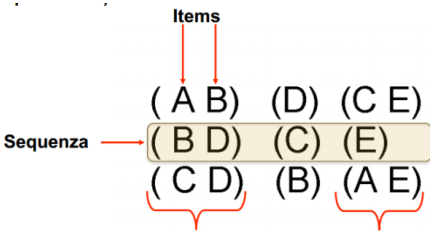
- graph
  - web and social networks
  - molecular structures





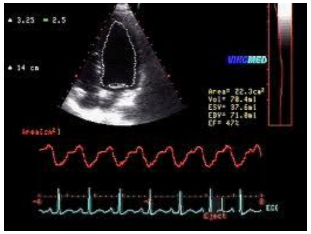
- ordinal data
- spatial data
- time series
- sequences
  - genetic sequences

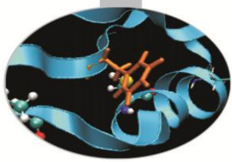




GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG

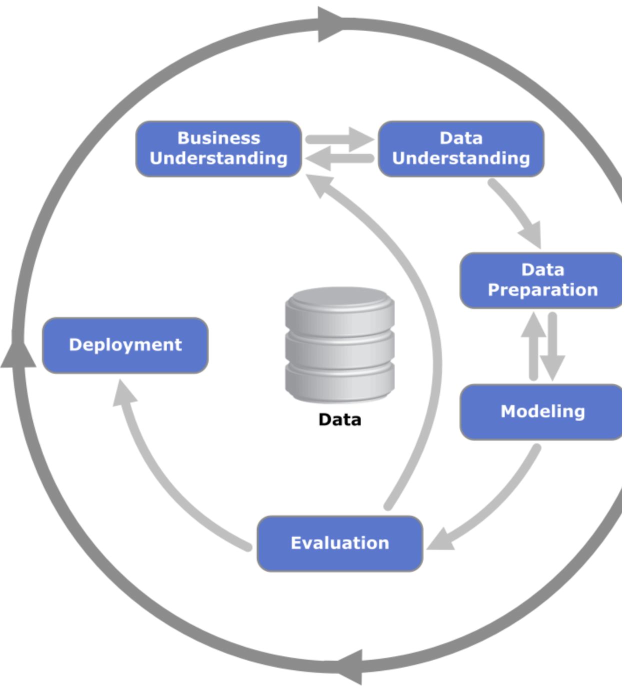
- unstructured data
  - textual documents
  - images
  - audio and videos (multimodal)



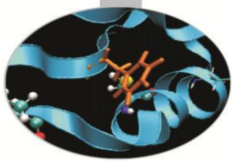


CRISP-DM reference model

Cross Industry Standard Process for Data Mining

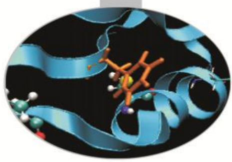


Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *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*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	**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*	**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*	**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*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*



New challenges (1)

- 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)
- 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.

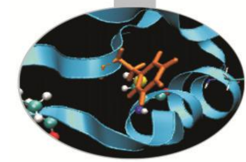


New challenges (2)

- 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.

New trends

- **Re-purposing data** that was collected for a different purpose.
- **Re-purposing algorithms** (e.g. page rank on graphs).
- **Data products**:
  - interactive visualizations, online databases -> not just answering the question once, empower others to use data in new ways
  - data-driven applications (e.g. spell checkers, machine translation, recommendation systems, People You May Know, UPS's route optimization system …) -> turn data into product
- A paradigm shift in knowledge creation (gaining insights) and **decision making** (taking action): analytics obviates the need for decision makers to understand the phenomenon before they act on it (first comes the analytical fact, then the action, and last, if at all, the understanding).



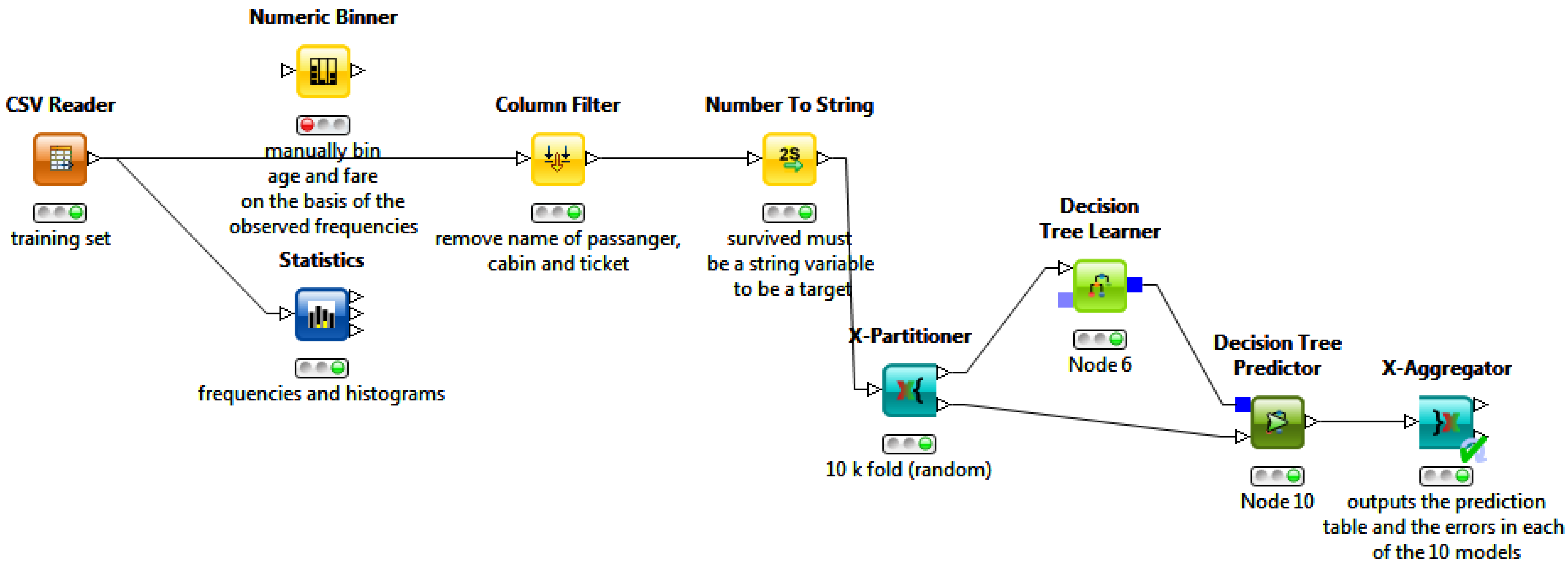
Another way of describing the process (BDVA)

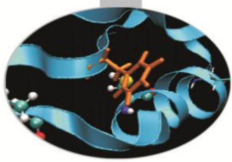
data analysis output can be input for other higher level analysis

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

The process – Knime Workflow





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

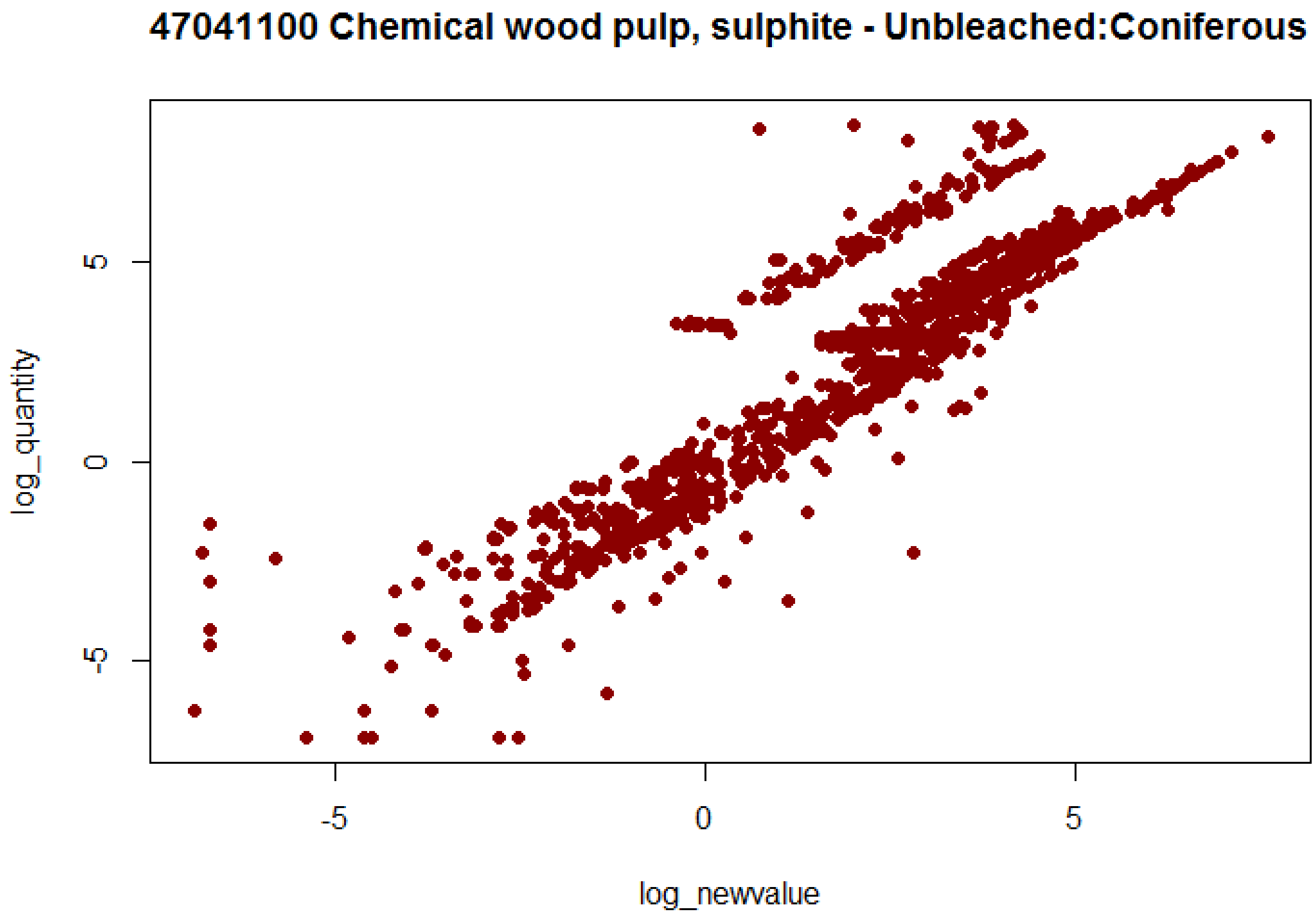
- data preparation
  - Cleaning
  - Transformation (normalization, discretization, aggregation, new variables computation…)
  - Feature extraction
  - Selection / filtering

Pre–processing

Why is it useful - a few examples

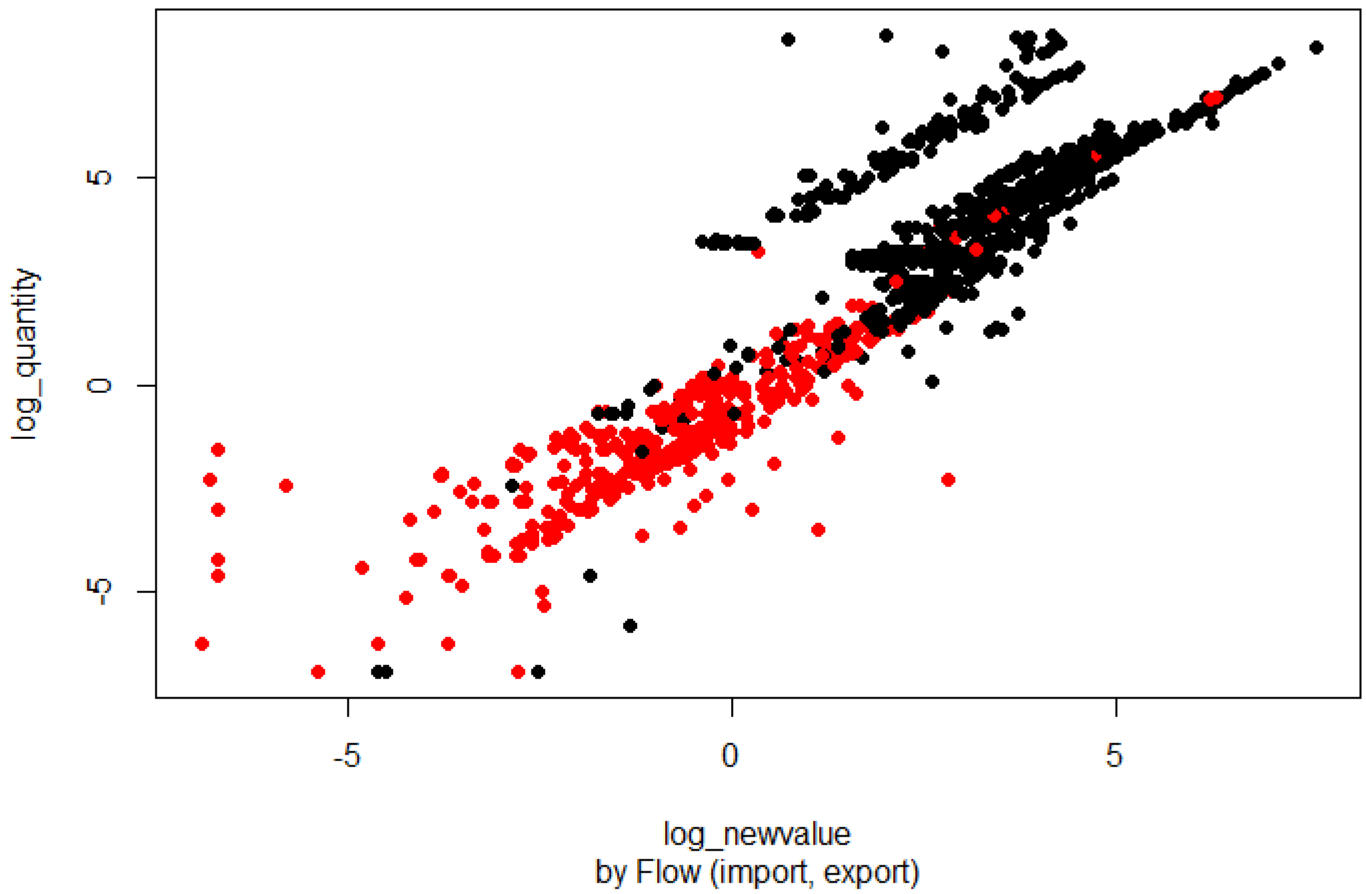
- L'Equité: high peak of 96 years old insured
  - missing birth dates had been codified 1/1/1900
- Trento University: a high number of students with very low grades in the high school diplomas
  - grades in the high school diplomas have undergone a scale change (from 60 as a maximum to 100)
- Local Health 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 declarant Country was a key variable in international trade outliers identification

Pre–processing

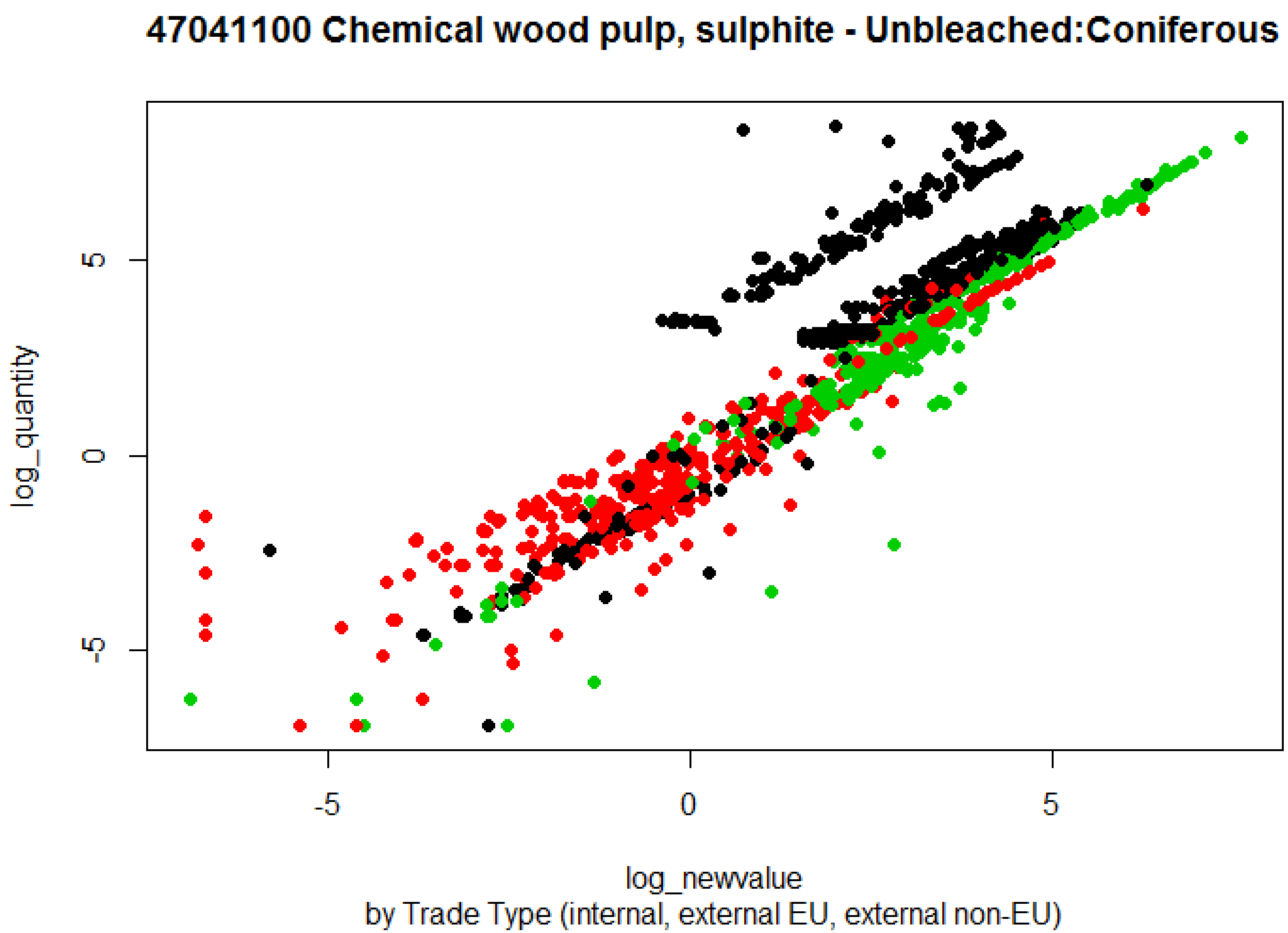


Pre–processing

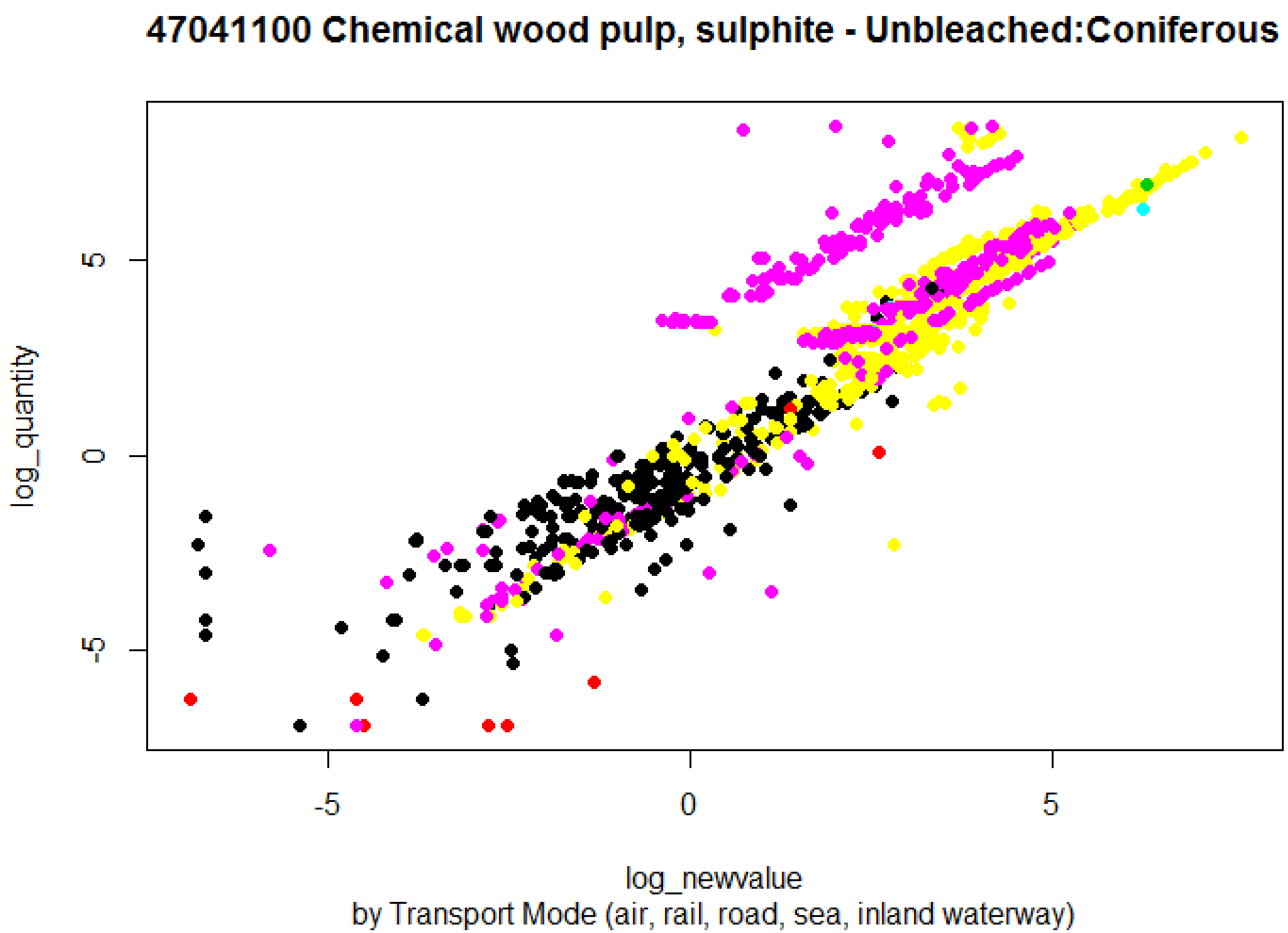
47041100 Chemical wood pulp, sulphite - Unbleached:Coniferous



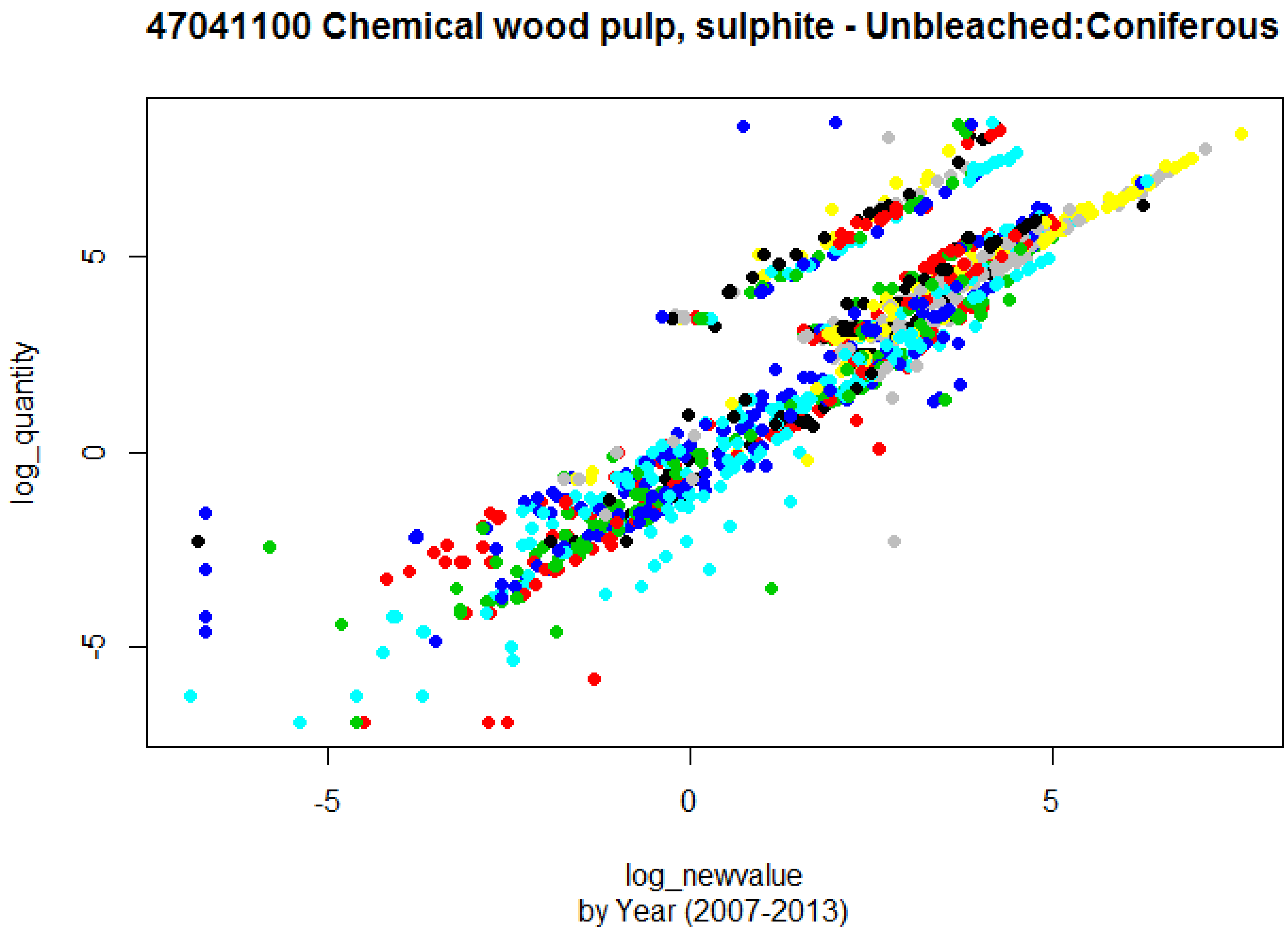
Pre–processing



Pre–processing

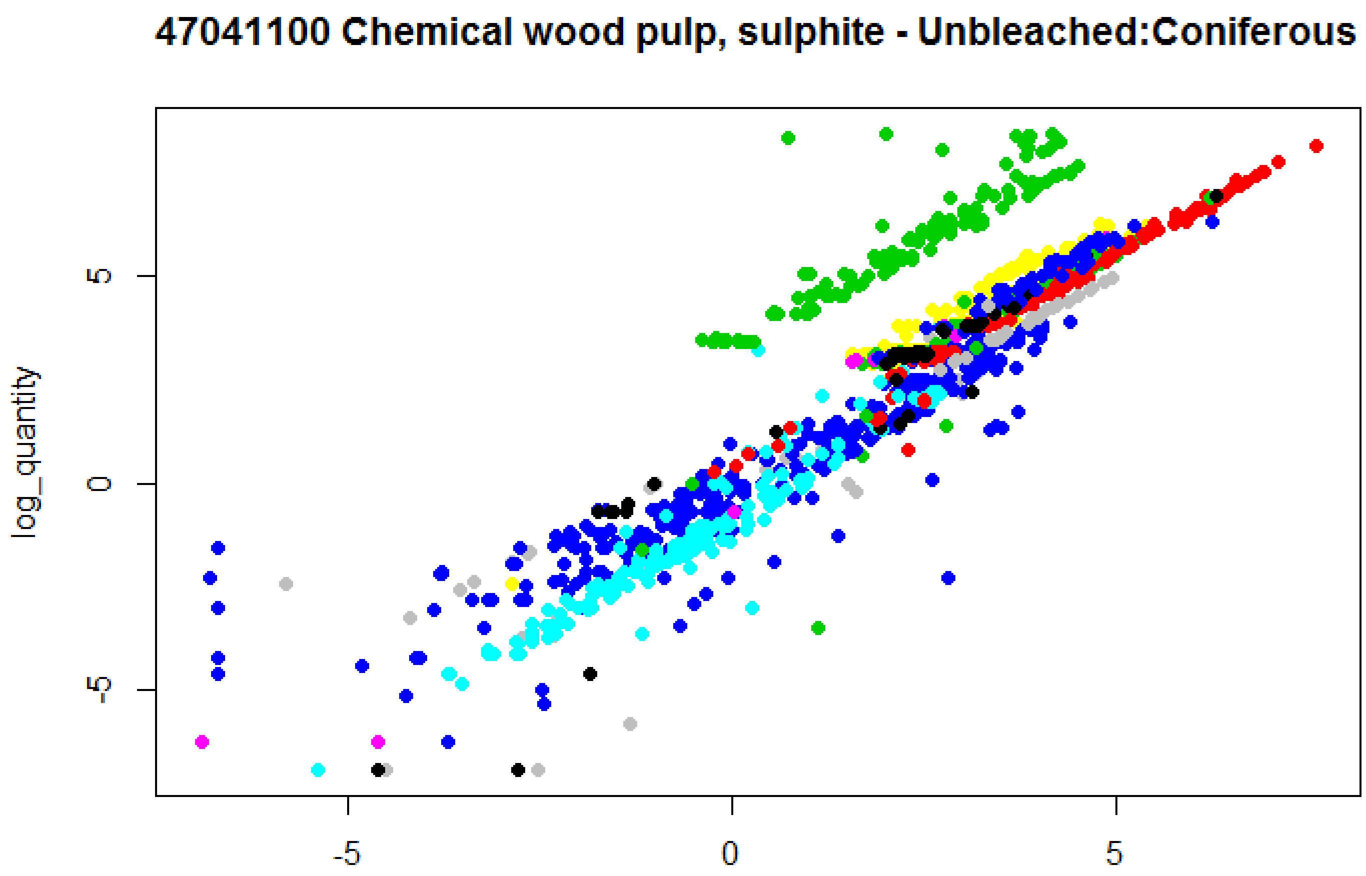


Pre–processing



SuperComputing Applications and Innovation

Pre–processing



A general model is not possible: the declarant country must be accounted for, otherwise the cloud of points referring to Sweden exports would be labelled as outlier and evened out

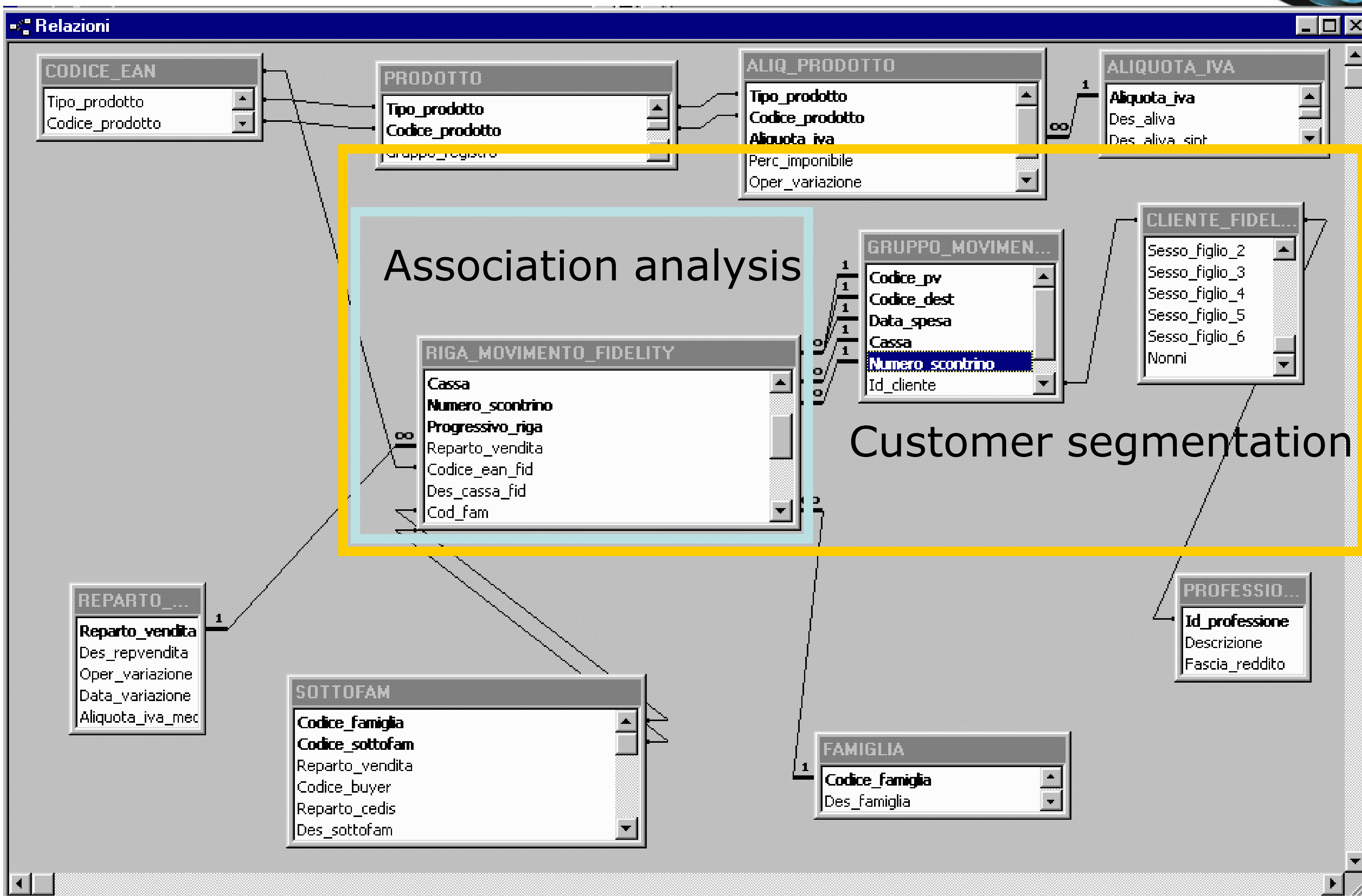
SuperComputing Applications and Innovation

Data representation

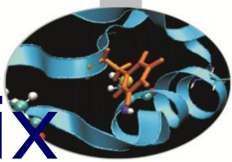
Analysis matrix

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

Coal: data structure



SuperComputing Applications and Innovation



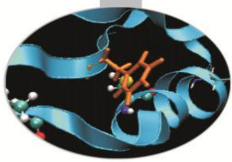
Coal: customer segmentation matrix

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

“active” variables

“descriptive” variables

SOGEI: target variable definition



Two information were available:

• the " tax credit accrued during the year, which is not due " means the credit , as calculated by the taxpayer , in the absence of the conditions for entitlement .

• the " tax credit used during the year without being entitled " indicates the amount used in excess of the amount due , as estimated by the auditor.

Four possible outcomes:

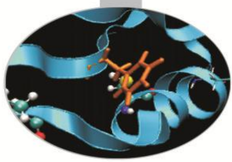
Group	Description	N.	%	Audit Outcome	Target Variable
1	Undue tax credit declared = 0 AND Undue tax credit benefited = 0	26.484	48,58	No remarks	0
2	Undue tax credit declared = 0 AND Undue tax credit benefited > 0	12.647	23,20	Substantial remarks	1
3	Undue tax credit declared > 0 AND Undue tax credit benefited = 0	6.514	11,95	Formal remarks	0
4	Undue tax credit declared > 0 AND Undue tax credit benefited > 0	8.864	16,26	Formal and Substantial Remarks	1
	TOTAL	**54.517**	**100**		

The process in text mining

- collecting
- indexing
- mining
- evaluation

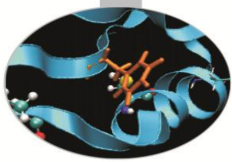
Collecting

- 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

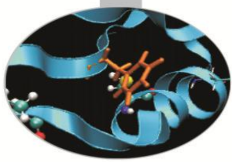
Data representation

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.

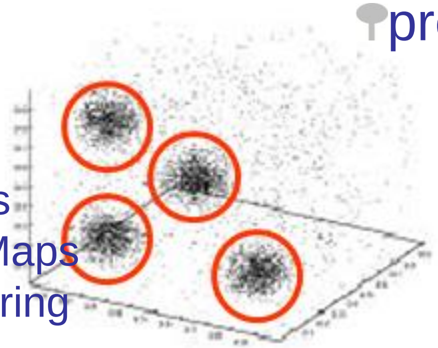
	team	coach	play	ball	score
Document 1	3	0	5	0	2
Document 2	0	7	0	2	1
Document 3	0	1	0	0	1

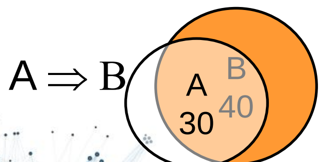
SuperComputing Applications and Innovation



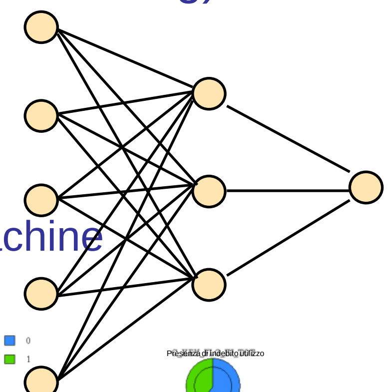
Tasks and techniques

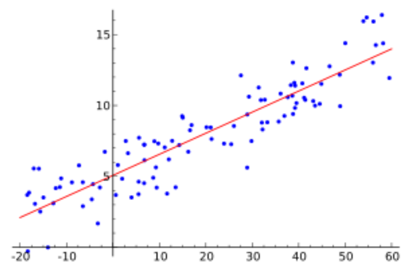
- descriptive
  - clustering
    - k-means
    - relational analysis
    - Self Organizing Maps
    - hierachical clustering
    - mixture model
    - …
  - association rules
  - sequential patterns
  - graph and network analysis
  - dimensionality reduction
  - …

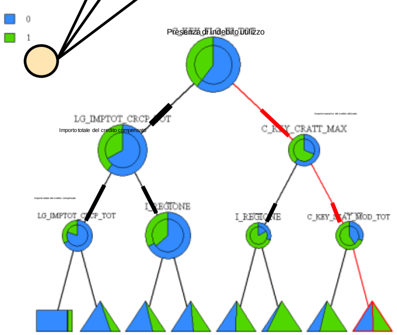




- predictive
  - classification (machine learning)
    - Naive Bayes
    - Decision Trees
    - Neural Networks
    - KNN
    - Support Vectors Machine
    - …
  - regression







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)