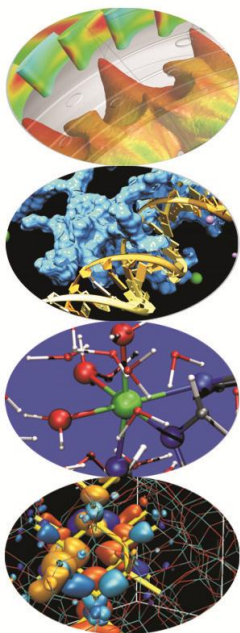
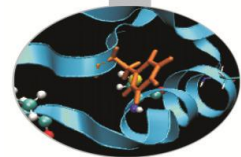


Introduction to Data Analytics

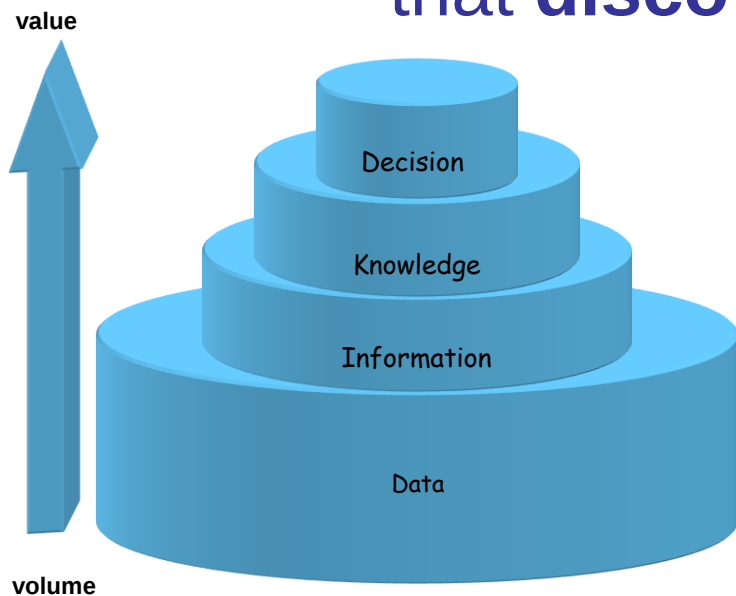
Roberta Turra, *Cineca*



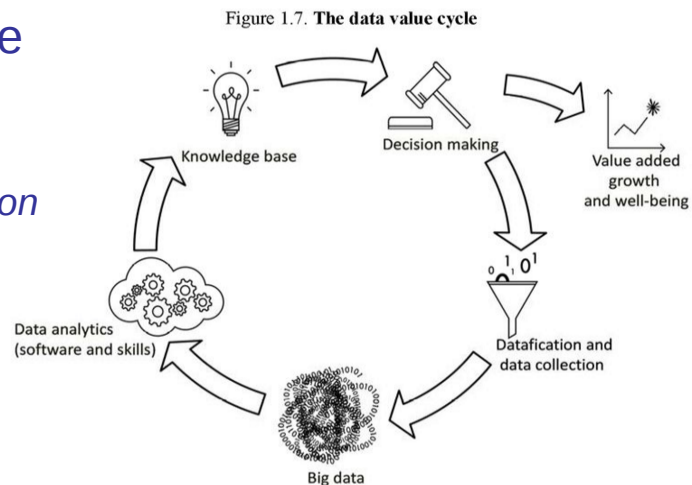


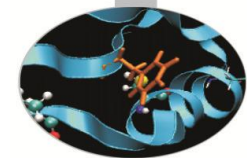
Data analytics

The **process** of extracting useful insights from raw data using algorithms that **discover** hidden patterns

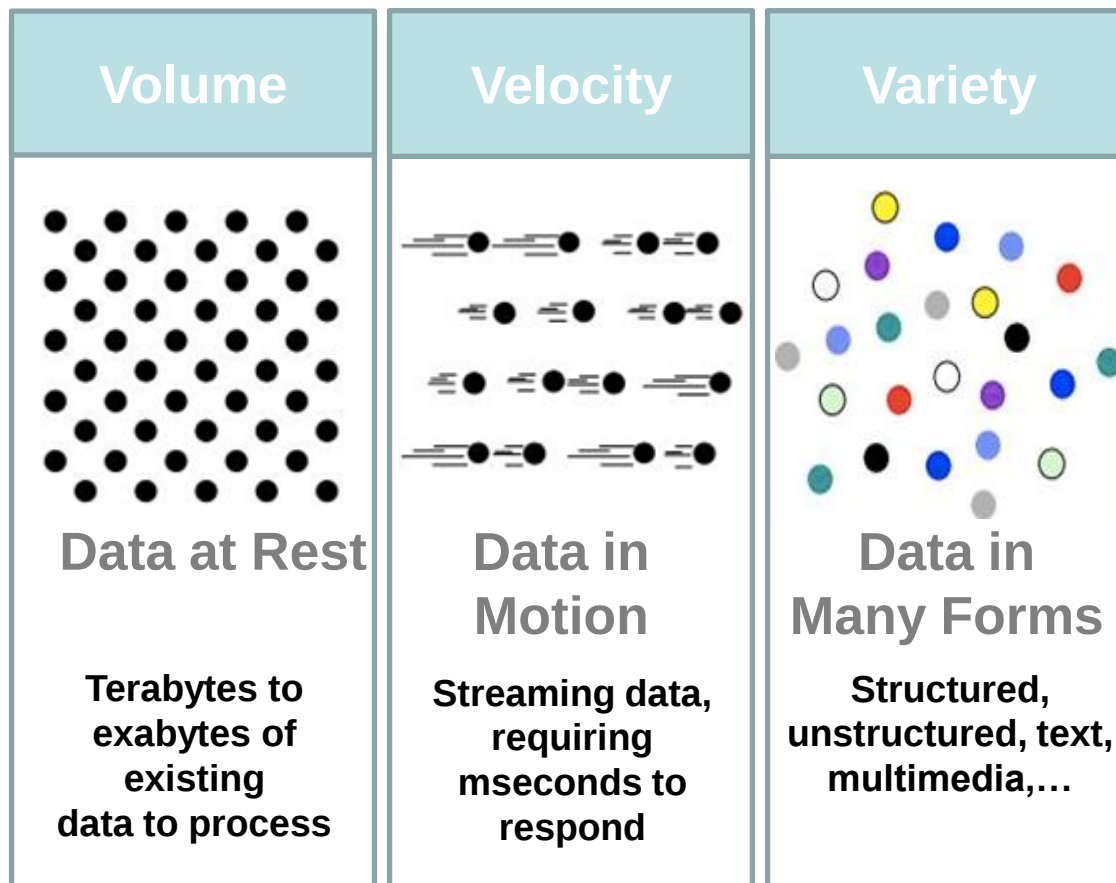


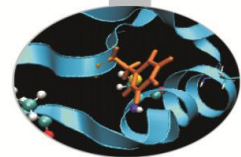
A step in the data value cycle
(OECD report on Data Driven Innovation)





Why is it challenging





Data typologies

structured data

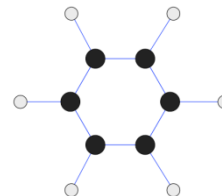
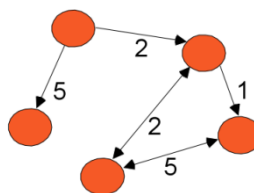
- data matrix
- transactional data

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

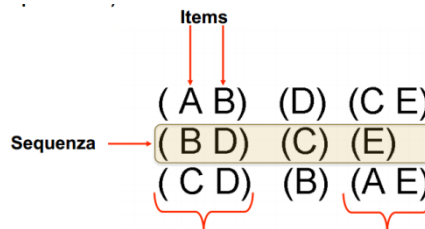
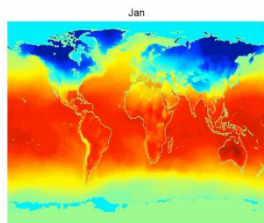
time series

sequences

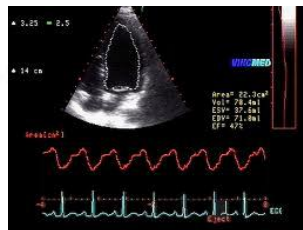
- genetic sequences

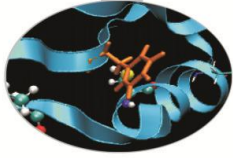
unstructured data

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



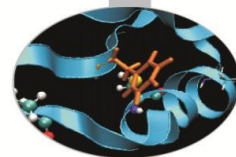
GGTTCGCGCTTCAGCCCCGCGCC
 CGCAGGGCCCCCGCCGCGCGTC
 GAGAAGGGCCCCGCTGGCGGGCG
 GGGGGAGGCGGGCGCCCGAGC
 CCAACCGAGTCCGACCAGGTGCC
 CCCTCTGCTCGGCCCTAGACCTGA
 GCTCATTAGGCGGCAGCGGACAG
 GCCAAGTAGAACACGCGAAGCGC
 TGGGCTGCCTGCTGCGACCAGGG





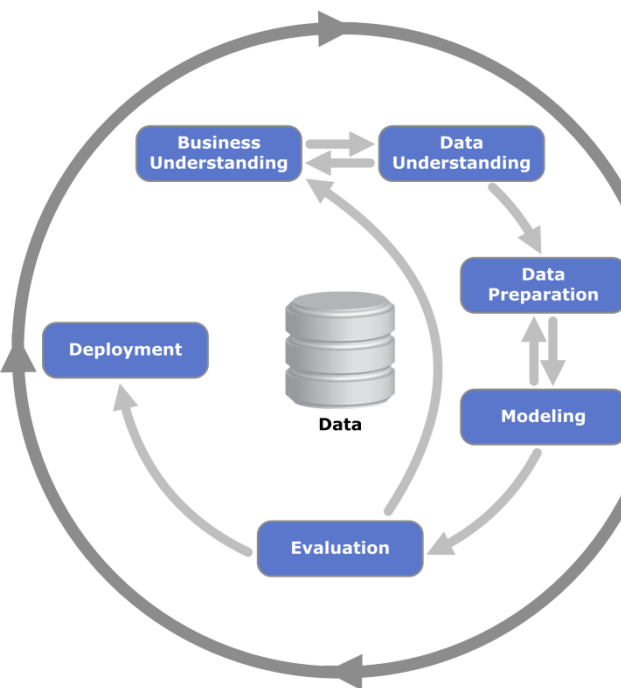
Data as an infrastructure

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.

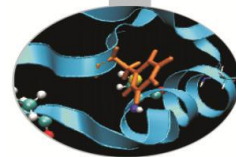


CRISP-DM reference model

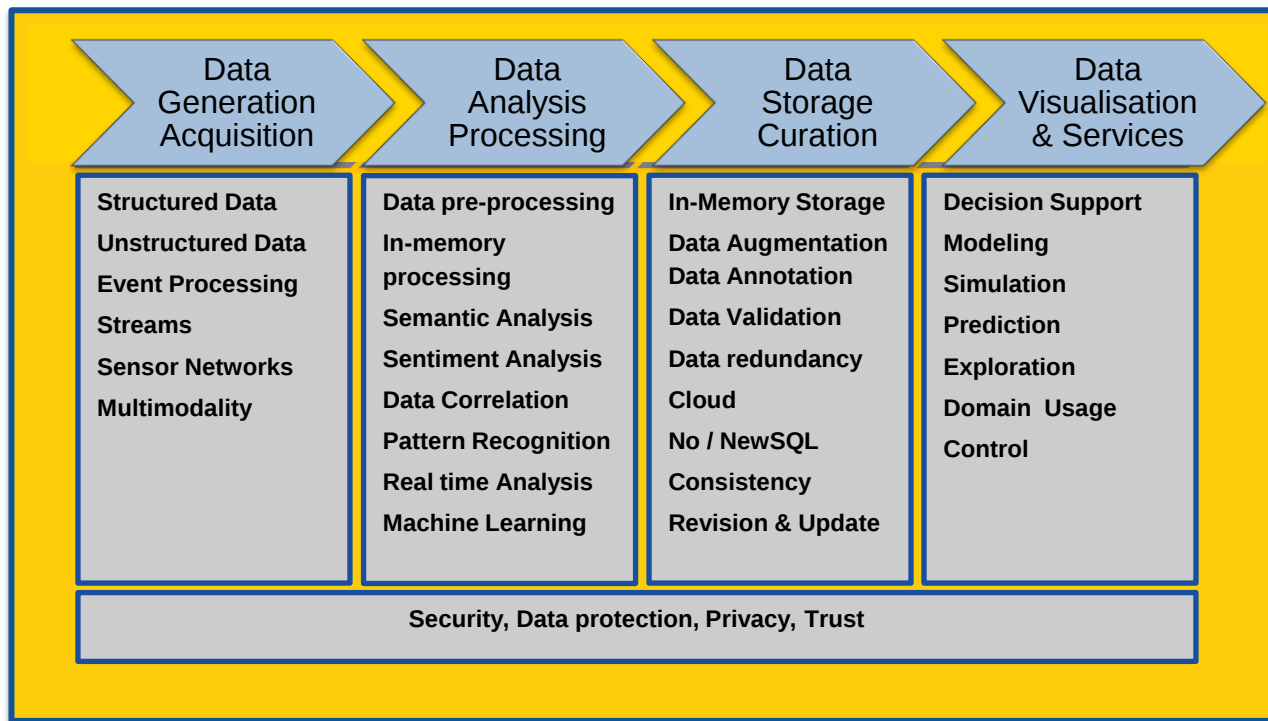
Cross Industry Standard Process for Data Mining

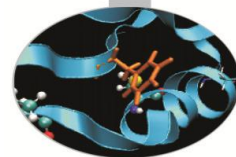


Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
Determine Business Objectives <i>Background Business Objectives Business Success Criteria</i>	Collect Initial Data <i>Initial Data Collection Report</i>	Select Data <i>Rationale for Inclusion/ Exclusion</i>	Select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



Another way of describing the process (BDVA)



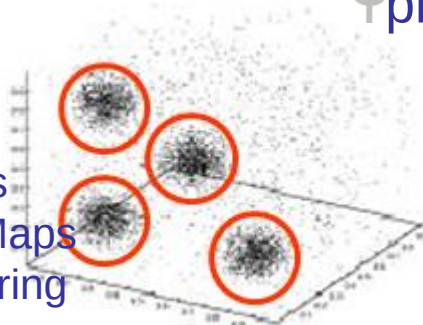


Tasks and techniques

🌳 descriptive

🌳 clustering

- 🌳 k-means
- 🌳 relational analysis
- 🌳 Self Organizing Maps
- 🌳 hierarchical clustering
- 🌳 mixture model
- 🌳 ...



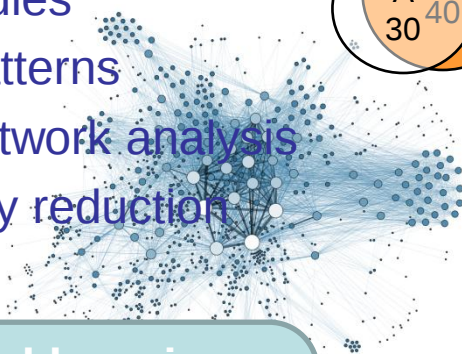
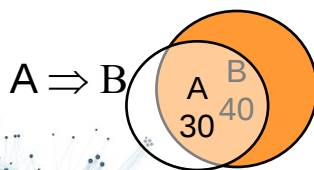
🌳 association rules

🌳 sequential patterns

🌳 graph and network analysis

🌳 dimensionality reduction

🌳 ...



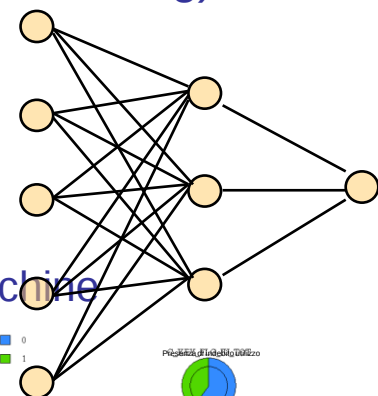
Unsupervised learning

training samples have no class information
 guess classes or clusters in the data
 we are given inputs but no outputs
 (unlabeled data)
 we learn the "latent" labels

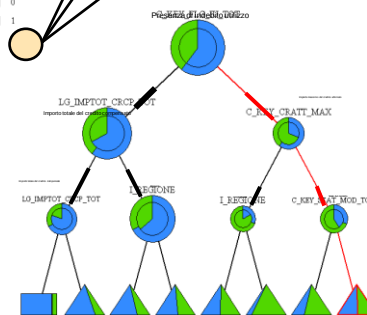
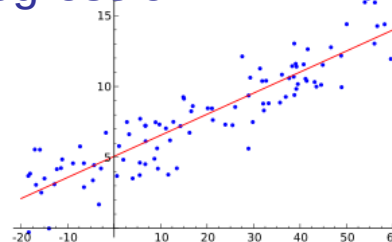
🌳 predictive

🌳 classification (machine learning)

- 🌳 Naive Bayes
- 🌳 Decision Trees
- 🌳 Neural Networks
- 🌳 KNN
- 🌳 Rocchio
- 🌳 Support Vectors Machine
- 🌳 ...

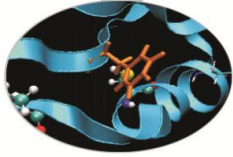


🌳 regression



Supervised learning

use training samples with known classes
 to classify new data
 we are given examples of inputs and
 associated outputs
 we learn the relationship between them



Different approaches to the predictive task

predictive

- classification (the learned attribute is categorical ,“nominal”)

- Naive Bayes
- Decision Trees
- Neural Networks
- KNN
- Rocchio
- Support Vectors Machine
- ...

- regression (the learned attribute is numeric)

infer how to map input to output

Statisticians: model the process that gave rise to data

ML: make an accurate prediction, given the data