RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE
AND INNOVATION POLICY STUDIES

On-line Course

Optimization and Machine Learning for the Imputation of Missing Interconnected Data

R. Bruni

bruni@dis.uniroma1.it

Università di Roma "Sapienza"
Dip. di Ingegneria Informatica, Automatica e Gestionale (DIAG)

Outline

- Part 1: Introduction to Data Mining and Machine Learning
- Part 2: The Imputation of Interconnected data
  - Higher Educational Institutions Data
  - The Problem of Missing Values
  - Brief Overview of Imputation Techniques
  - Trend Smoothing Imputation
  - Donor Imputation
  - Computational Results

What is Machine Learning ?

Some definitions

1959: Arthur Samuel *“**programming** of a digital computer to **behave** in a way which, if done by human beings or animals, would be described as involving the process of **learning**”*

1997: Tom Mitchell “Machine Learning is the study of computer algorithms that ***improve automatically*** through experience”

or, more precisely

*“We say that a machine learns with respect to a particular task T, performance metric P, and type of experience E, if the system reliably **improves its performance P** at task T, following experience **E**”*

An example: Spam Detection

An e-mail filter able to decide which mail should be classified as «spam» or «not spam» learning from your decisions on past emails

- **T (task)** classify mail as «spam» or «not spam»
- **P (performance measure)** the percentage of correctly classified mails
- **E (experience)** your e-mail classification as «spam» or «not spam»

What is Data Mining ?

- **Information explosion** (a.k.a. data flood) is the rapid increase in the amount of data produced and stored

- **A circle**: technology improvements allow to use more data → using even more data becomes **necessary** → this requires further technological improvements

- We are drowning in data, but starving for **knowledge**!

  Managing those data becomes more and more difficult. We need effective techniques, or we risk an information overload

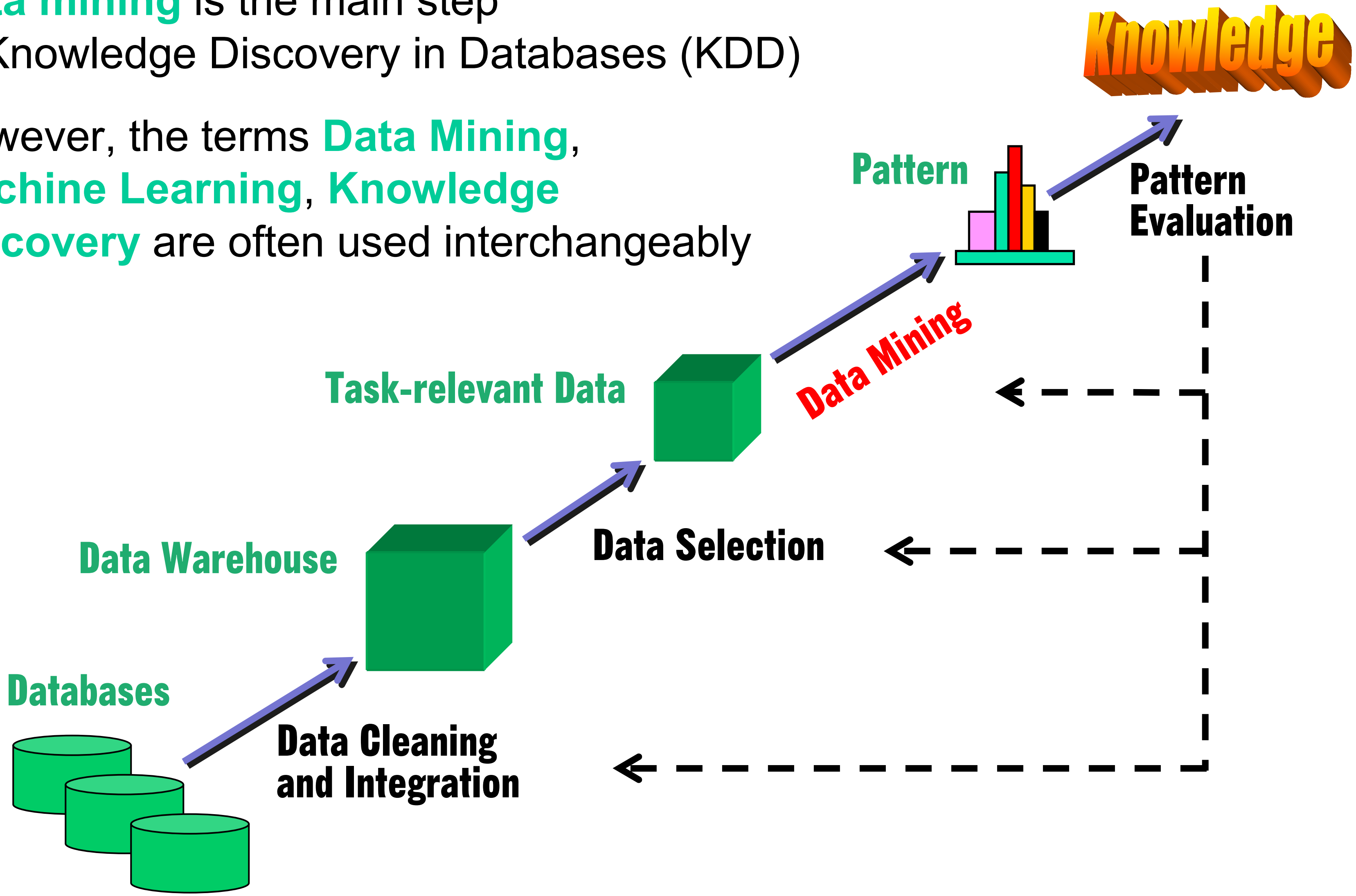
- **Data mining:**

  Extraction from large data sets of information that is **not obvious**, **not immediately available** and **potentially useful** (rules, regularities, patterns, etc. = knowledge) using automatic or semi-automatic methods

Where is Data Mining ?

Data mining is the main step
of Knowledge Discovery in Databases (KDD)

However, the terms **Data Mining**, **Machine Learning**, **Knowledge Discovery** are often used interchangeably

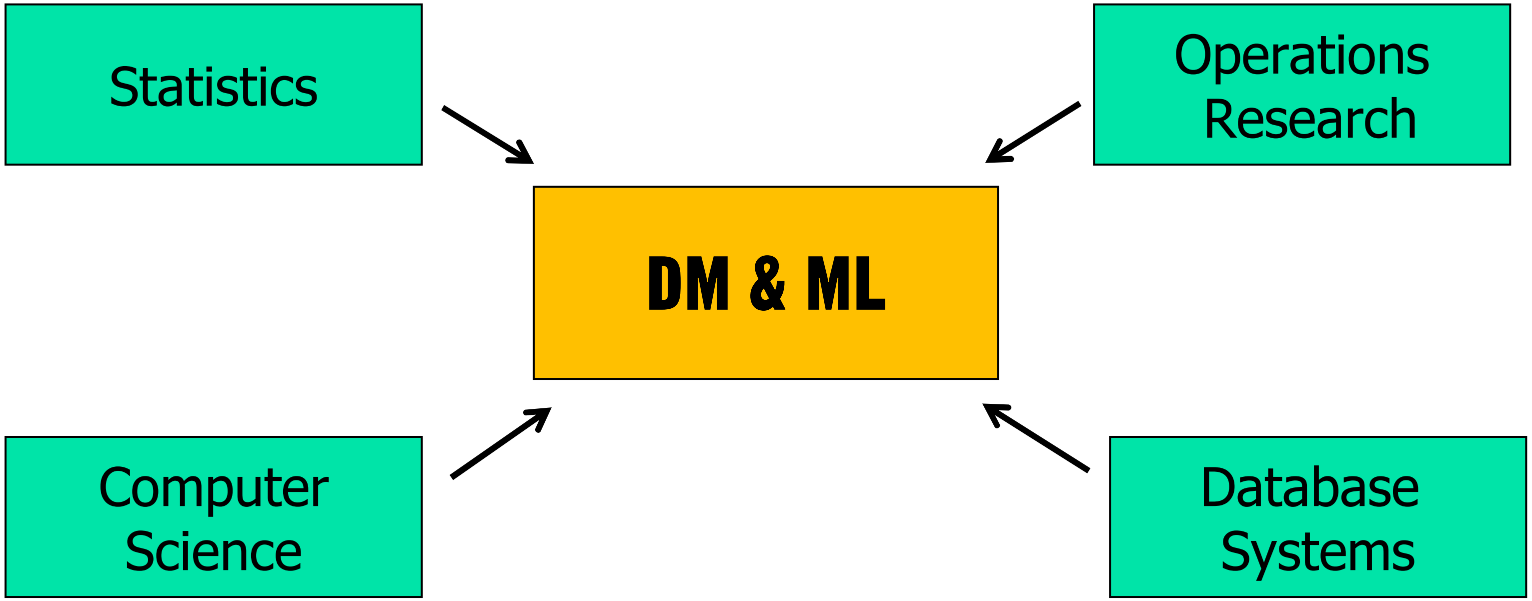


Who needs Data Mining ?

Obtaining **knowledge** and not just **data** is essential in many applications. Some examples:

- Database Analysis (Rules extraction, Associations)
- Market Analysis (Customer profiling, Marketing)
- Risk Analysis (Finance planning, Investments)
- Fraud Detection (Credit cards, Food adulteration)
- Decision Support (Resource management, Allocation)
- Medical Analysis (Diagnosis, Donors management)
- Text mining (Search engines, Anti spam)
- Analysis of Economical or Social Policies (Rule learning)
- …

What do I need for ML and DM ?



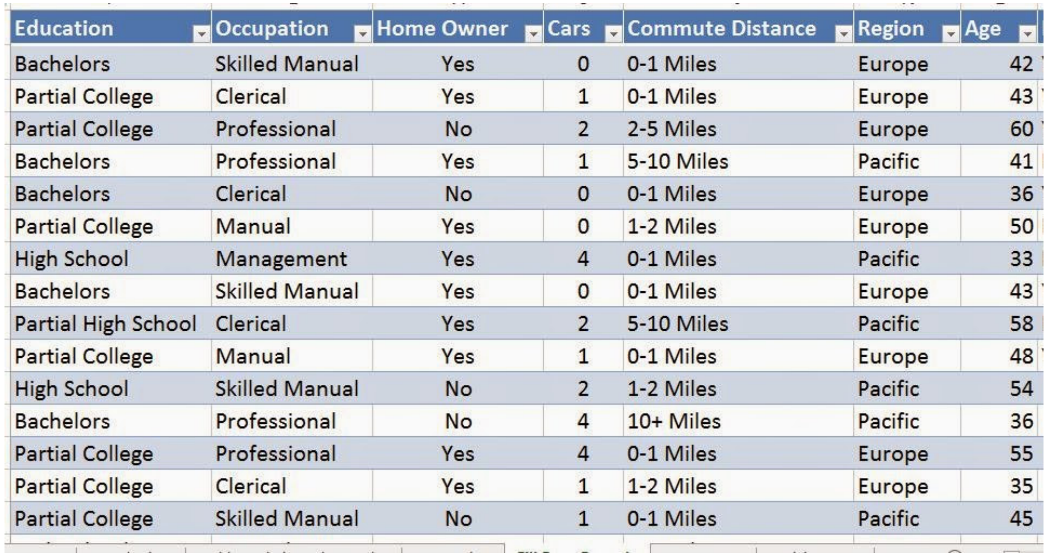
- Several different competences are required to do Machine Learning and Data Mining
- It is a very **interdisciplinary Area**
- For this reason, many things are called with **different names** in the different communities

What exactly is Data ?

- A collection of **objects**, each having some **attributes**
- Each object is usually stored in a record
- An attribute is a property or characteristic of an object
  Examples: name, eye color, income, etc.

Attributes, a.k.a. fields, features, variables, columns, …

Objects, a.k.a. records, tuples, instances, observations, points, samples, rows, …



Education	Occupation	Home Owner	Cars	Commute Distance	Region	Age
Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	42
Partial College	Clerical	Yes	1	0-1 Miles	Europe	43
Partial College	Professional	No	2	2-5 Miles	Europe	60
Bachelors	Professional	Yes	1	5-10 Miles	Pacific	41
Bachelors	Clerical	No	0	0-1 Miles	Europe	36
Partial College	Manual	Yes	0	1-2 Miles	Europe	50
High School	Management	Yes	4	0-1 Miles	Pacific	33
Bachelors	Skilled Manual	Yes	0	0-1 Miles	Europe	43
Partial High School	Clerical	Yes	2	5-10 Miles	Pacific	58
Partial College	Manual	Yes	1	0-1 Miles	Europe	48
High School	Skilled Manual	No	2	1-2 Miles	Pacific	54
Bachelors	Professional	No	4	10+ Miles	Pacific	36
Partial College	Professional	Yes	4	0-1 Miles	Europe	55
Partial College	Clerical	Yes	1	1-2 Miles	Europe	35
Partial College	Skilled Manual	No	1	0-1 Miles	Pacific	45

Data Records

A record scheme is a set of **fields** $R = \{ f_1 \dots f_m \}$

A record instance is a set of **values** $r = \{ v_1 \dots v_m \}$

Each field f_i has its **domain** D_i that is the set of all possible values

Example: fields can be *age, marital status,* corresponding values can be 18, *single*, etc.

Fields can be:

- numerical or quantitative
  - continuous: real-valued
  - discrete: integer or binary
- categorical or qualitative
  - ordered (e.g. first, second)
  - not ordered (e.g. red, blue)

Fields can be re-encoded differently. For example, many procedures convert **each field** f_i into one or more **binary ones**, that we will call **binary attributes** $a_i^j \in \{0,1\}$

Different Tasks in Data Mining

Depending on the application, different **activities** may be required. However boundaries are not sharp at all

- **Classification**: learning a function or a criterion to map objects on a pre-defined set of classes
- **Regression**: learning a function or a criterion to assign each object a real value
- **Clustering**: identification of a partition of the set of objects to group together similar objects
- **Learning of Dependencies and Associations**: identification of significant relationships among data attributes
- **Rule Learning or Summarization**: identification of a compact description of a set or subset of data

Learning Paradigms

Supervised learning: the “correct answer” (**label**) on the instances is available (at least for some of them).

We learn from the labeled data (=correct answers) to predict labels (=new correct answers) for unseen instances

Unsupervised learning: no “correct answers” available.

We use the data but the corresponding output values are not known in advance. Example: one wants to find similarity classes and to assign instances to the correct class

Very often, labeled data are scarce, but unlabeled data are easy to collect. **Semi-supervised learning**: techniques that learn from small amount of labeled data and also from large amount of unlabeled data

Learning Process

In many learning tasks, data are partitioned into:

- **Training set** (data+labels, or just data for unsupervised): used to learn
  - incrementally (on-line learning): Data are obtained incrementally during the training process
  - batch (off-line) learning: Data of the training set are available in advance before entering the training process
- **Validation set** (data): after the learning phase, we may need other data to tune parameters etc.
- **Test set** (data): used for doing what we must do (if we know also the labels, we can compute the accuracy)

We deal with **large data sets** and possibly small training sets (e.g. rare events, not controllable events). Labeled data may be costly.

Classification ex.: Fraud detection

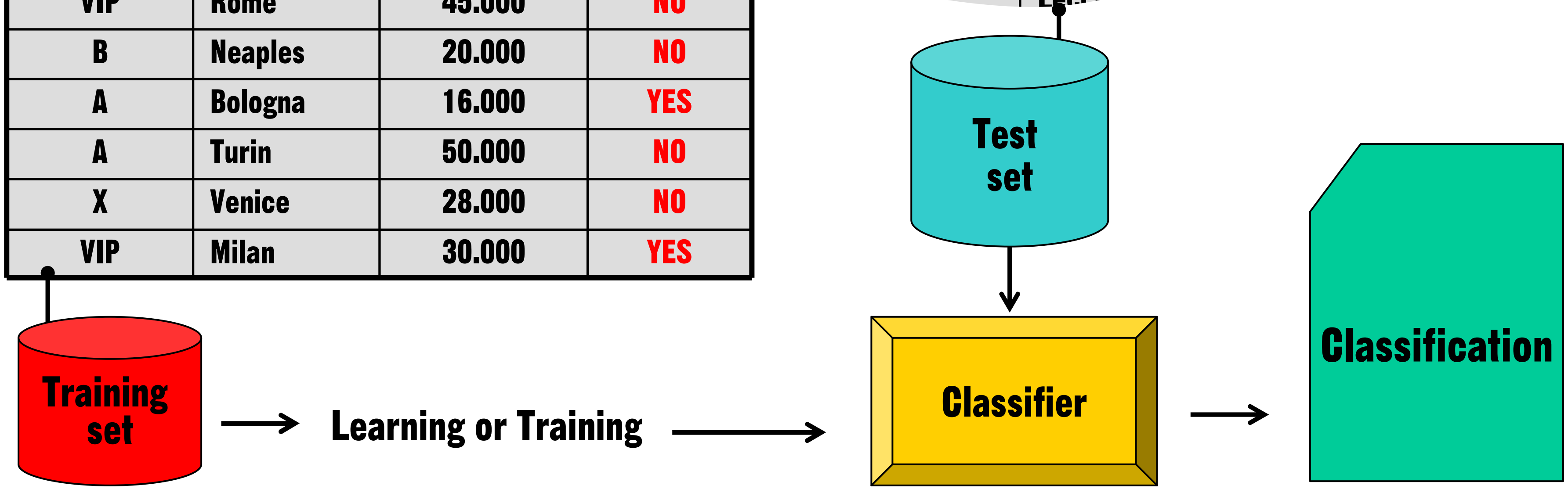
- Given a **training set** partitioned in classes, **predict** the **class** of new data, i.e., learn a classifier

Numerical or categorical

Class

customer type	Town	Income	Fraud
A	Rome	25.000	NO
B	Milan	15.000	YES
X	Florence	18.000	NO
VIP	Rome	45.000	NO
B	Neaples	20.000	NO
A	Bologna	16.000	YES
A	Turin	50.000	NO
X	Venice	28.000	NO
VIP	Milan	30.000	YES

Customer type	Town	Income	Fraud
X	Milan	30.000	?
A	Turin	22.000	?
VIP	Florence	18.000	?
A	Rome	14.000	?
B	Milan	55.000	?
X	Bari	26.000	
A	Lecce		



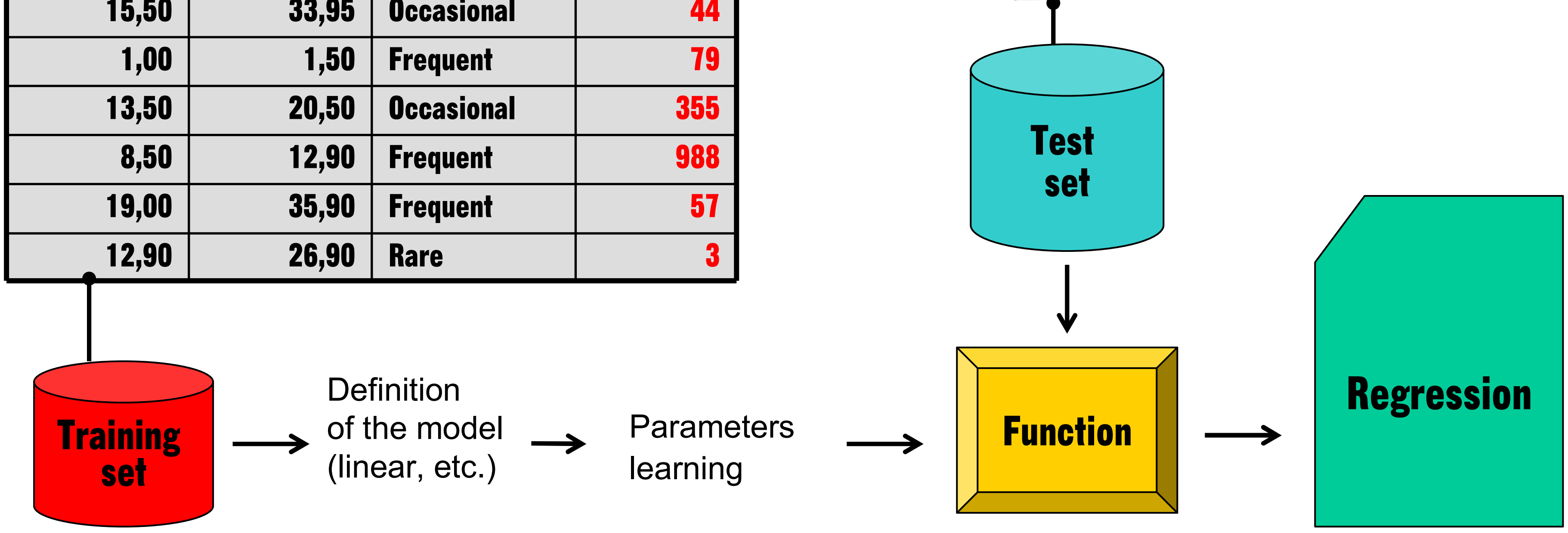
Regression example: Predict Sales

Independent variables (predictors)

Dependent variable (numerical)

Cost	Price	Usage	Sales
5,00	11,50	Frequent	154
6,00	12,80	Rare	21
15,50	25,50	Frequent	234
15,50	33,95	Occasional	44
1,00	1,50	Frequent	79
13,50	20,50	Occasional	355
8,50	12,90	Frequent	988
19,00	35,90	Frequent	57
12,90	26,90	Rare	3

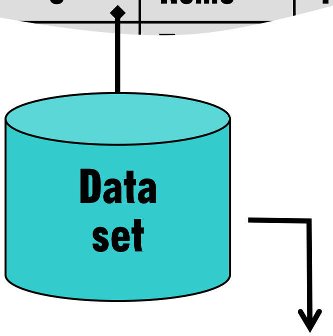
Cost	Price	Usage	Sales
10,00	19,90	Frequent	?
5,50	11,00	Occasional	?
14,50	25,90	Occasional	?
63,00	128.00	Rare	?
2,50	4,90	Frequent	?
24,00	49,90	Occasi	
12,00			



Clustering ex.: Market Segmentation

ID Cust.	Town	Income	Marital status	Revenue
1	Milan	21.470	unmarried	2.500
2	Rome	12.500	unmarried	400
3	Turin	63.600	Divorced	250
4	Neaples	21.900	married	12.000
5	Milan	20.300	married	645
6	Rome	40.500		

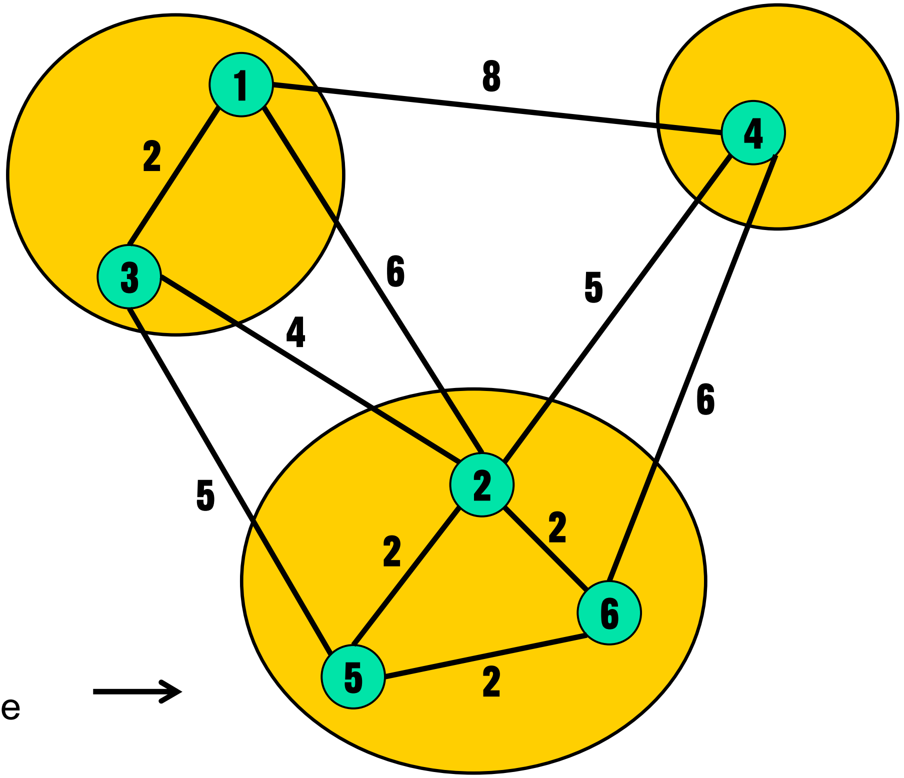
Given the data, partition all customers in k=3 groups that should be treated differently



Definition of a distance criterion and computation of distances →

↓

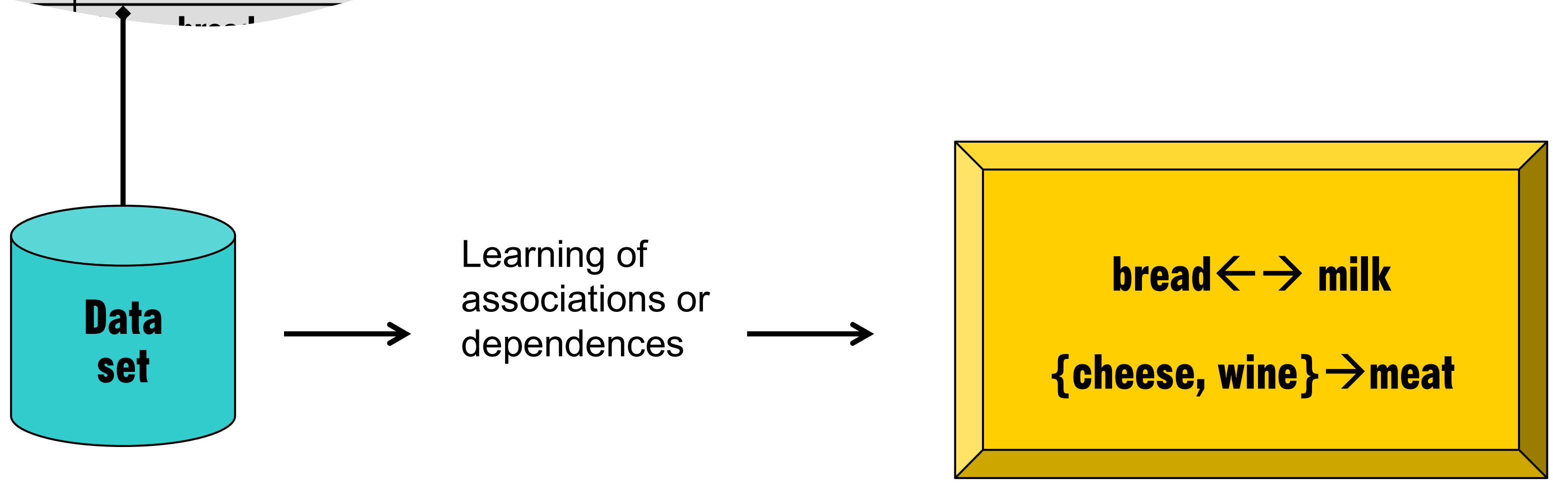
Partition in k groups minimizing intra-group distance or maximizing group-to-group distance →



Association example: Food Shopping

ID	Oggetti Acquistati
1	bread, milk, eggs
2	vegetables, cookies, juice, pasta
3	meat, cheese, cookies, wine
4	bread, cheese, milk
5	bread, wine, meat, vegetables, milk, cheese
6	pasta, juice, eggs

Each record contains a variable number of objects from a list of foods. Find dependences or associations in the records, so as to predict what a customer still has to buy and help him\her (the more we sell, the better)



Data set

Learning of associations or dependences

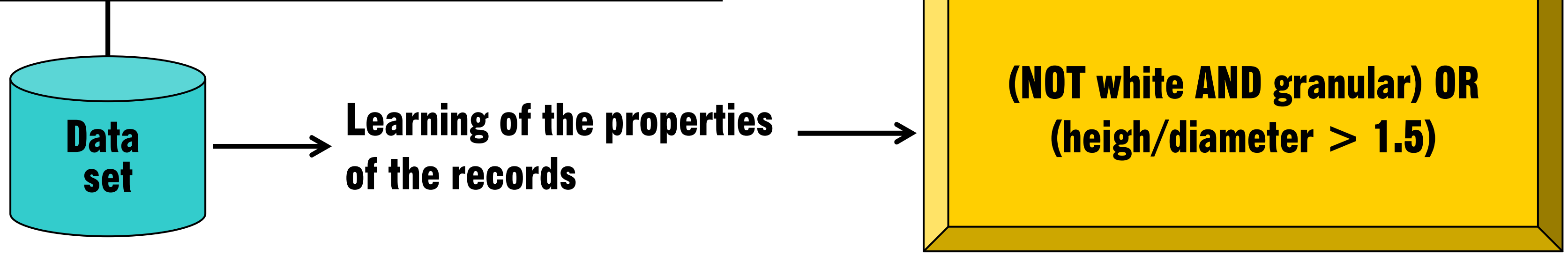
bread ←→ milk

{cheese, wine} → meat

Rule extraction: pois. mushrooms

Categorical	Categorical	Numerical	Numerical
Colour	**Skin**	**Diameter**	**Heigh**
red	granular	13	5
white	smooth	4	7
grey	granular	10	8
grey	smooth	6	12
red	granular	10	10
white	granular	5	9
grey	smooth	6	10
white	granular	3	6
red	smooth	10	16

Given the description of many poisonous mushrooms, find a compact description (an intensive description) of this set



Data set

Learning of the properties of the records

(NOT white AND granular) OR (heigh/diameter > 1.5)

“If you torture the data long enough, it will always confess”

How to obtain the results?

- There exist many approaches, each approach has several variants, and algorithms can also be designed by mixing approaches
- The background of researchers often will make the choice
- In general, **there is not** a “best technique”: no single algorithm is currently able to provide the best performance on all datasets
- This seems to be **inevitable**: if you chose a “best technique”, one can make a dataset composed of the records wrongly labeled by this “best technique” and make it the “worst technique” (no free lunch theorem)
- Therefore, **Ensemble techniques**: use many weak learners and combine their outputs to obtain both accuracy and robustness

Part 2: Interconnected Data

Higher Educational Institutions

Organizations providing higher level education are called Higher Educational Institutions (HEIs)

- Their data constitute the basis for many important **analyses on the educational systems**
- Key data are for example number of students, number of graduates, etc.
- The **European Tertiary Education Register** (ETER) is the database of European HEIs
- It has been developed during the course of 2 European research projects
- Covers **37 European Countries**, at present most of the data go **from 2011 until 2016**

HEI Data

- Each Unit is described by a data record: generally there is a **value** for each **variable**

- Here, a unit is a University, and a variable is for example “name of the University”, or “number of students”; the latter is a **time series** of values: one value for each of the covered years

- **Example**: The students of University AAA over the years 2011,…,2016 are denoted as

$$V_{students\ 2011}^{AAA}\ ,\ldots,\ V_{students\ 2016}^{AAA}$$

- To make it easier, when there is no ambiguity, we define our temporal horizon $\{1,2,\ \ldots,\ t\ \} = T$, and we may simply call them

$$s_1\ ,\ s_2\ ,\ \ldots,\ s_t$$

What is a Missing Value ?

- Sometimes we do not have “numbers” for the values

- For example, we do not have the number of students in 2011 for a University founded in 2015. To specify that this is correct, we call that value “not applicable”

- In other cases, we do not have a number, but **it should have been present**. For example, the number was cancelled by an error. We call that value “missing value”, denoted by m

- Missing values can be due to **many reasons**, and unfortunately they are very common in practice

- For ETER, due to the vast size of the project, they are very **frequent**

Why missing values are a problem ?

- Missing values do not allow the **micro analysis** of the institutions containing them
- Also, they prevent the **macro analysis** (at the aggregate level) of categories of institutions when they include the incomplete ones
- We need to mathematically “guess” the numerical value which is missing, by imputing a value **as similar as possible** to the original **unknown** value
- Giving a value instead of missing is called **Imputation** (or information reconstruction, …). Since this is a very common problem, many imputation techniques have been developed, in statistics, in database theory, etc.

How to deal with missing values ?

- This is a **difficult** problem, and easy solutions do not work. Example: the number of students in 2015 for University AAA is missing.

- If we just impute missing students with a predefined value, say 100, or the average value, we change the frequency distributions of the data: the imputed value will become **too frequent** (+ additional problems…)

- If we insert a random value taken from the distribution of students in the Universities, the value may still be **unsuitable** for AAA (if the random value is 55 and AAA is Sapienza, that is way too small!)

- Moreover, it is unclear **which** Universities must be taken for computing the distribution, and **which** distribution use. Also, the generated value may have never appeared in real data, and we generate a kind of “Frankenstein record”.

Techniques 1/3: from same unit

- Derive the missing info from the other values **available** in the unit
- Good in some cases, bad in others; must be **carefully calibrated** on the specific case
- If we have a series of values, and some are missing, the problem can be seen as **Regression**. There are Statistical or Machine Learning approaches (e.g. neural networks)
- Most of these approaches require setting some **parameters**, a.k.a. hyperparameters. For example, neural networks require to define the topology of the network, number of layers, neurons, etc.
- Different parameter choices may provide **very different results**, and usually there are no arguments to support a univocal choice
- **Non-parametric** techniques exist, but they require many data and, again, choosing the technique can be seen as a parameter choice…

Techniques 2/3: from other units

- Use the values of similar units, which are called **donors**. This has the advantage of **respecting** the frequency distributions. No need to guess a distribution or a regression model, and the value is guaranteed to have appeared in the data
- Donor imputation is very popular, of course the problem is **selecting the right donor** for each unit (we need a large set of donors)
- We need to set up **similarity criteria** for the specific application. For example, a university is similar to another univ. if: same country, same size, same type of univ., etc.
- In other words, we need a **distance function** between couples of units
- Hot deck means taking a donor from the same dataset, Cold deck means take it from another dataset (e.g. a previous edition of the same survey)

Techniques 3/3: from other sources

- We impute the value by obtain the missing information from **another source** (other databases, statistical ledgers, etc.)
- When another source is not available, it is clearly **not applicable**. This happens frequently, for instance it happens for ETER
- It requires to solve **data linkage** problems: recognize the same unit in another dataset, probably with a different structure. This may again be difficult (no free lunch!)
- Also, the other source may be outdated

In conclusion: the problem is difficult: we are guessing something we don’t know. Good news: over large numbers, positive and negative errors should tend to balance each other, so good for the **aggregate level**

Interconnected data

- One additional difficulty of HEI data: data are **interconnected**, both vertically and horizontally
- **Over the same year**. Examples: number of graduates is not independent from number of students, the expenditure is not independent from the staff, etc.
- **Over the time series of the same variable**. Example: number of students in one year is not independent from that in another year, etc.
- So, each imputed value may impact on many other values of the unit
- Many known imputation techniques have problems with this case
- For **interconnected** data we identify 3 important types of features in a unit, which must be respected by the imputed values: **Size**, **Ratios**, **Trends**

Size, Ratios and Trends

- **Size**: if we have some values for an institution, we can evaluate its size, and we must impute values with comparable size
- **Ratios**: due to the existing connections, the graduates are a portion of the students, the PhD students are portion of the graduates, the staff is proportional to the number of students, and so on. We identify a number of important ratios: Graduates/Students, PhD Stud/Graduates, PhD Grad/PhD Stud, Academic Staff/All Students Non-Academic Staff/All Students, Expenditure/All Students, Revenues/All Students, Expenditure/Revenues. When we impute, we must respect not only size but also these ratios
- **Trends**: if the number of students is increasing over the years, the same should happen for the number of graduates, and so on. When we impute the values, we must respect also these trends

Target variables

In ETER, the **target** of our imputation is constituted of

- Total number of students enrolled (Total Stud. Enr. at ISCED 5-7)
- Total number of graduates (Tot. Stud. Grad. at ISCED 5-7)
- Total number of PhD students (Tot. Stud. Enr. at ISCED 8)
- Total number of PhD graduates (Tot. Stud. Gr. at ISCED 8)
- Total academic staff (researchers and professors, measured either in Full Time Equivalent - FTE or in Head Count - HC)
- Total non-academic staff (technical and administrative staff, measured either in Full Time Equivalent - FTE or in Head Count - HC)
- Total current expenditure (in Euro)
- Total current revenues (in Euro)

Types of missing in ETER

- **Isolated Internal Missing**: occurs when there is one missing between two numerical values
- Example: 165, 193, 220, m, 205, 288
- **Isolated Extreme Missing**: when there is one missing at one extreme of the time series
- Example: m, 193, 220, 250, 205, 288
- **Missing Sequence of length L**: occurs when there are L consecutive missing values, but not the whole sequence
- Example: 165, 193, m, m, m, m
- **Full Sequence Missing**: when all the time series is m

Current Situation

	Total # of miss	HEI without missing
Students	3040	69%
Graduates	3852	61%
PhD students	3332	71%
PhD graduates	3697	70%
Academic staff FTE	7915	51%
Academic staff HC	7284	50%
Non-academic staff FTE	8807	45%
Non-academic staff HC	9145	42%
Expenditure	10041	31%
Revenues	10252	32%

Imputation from the same unit

- We use two basic techniques as **building blocks**: weighted average and linear regression

- **Weighted Average**: suitable for isolated internal missing

$$v_i = \sum_{\substack{h \in T \\ h \neq i}} w_h \, v_h$$

weights w_h progressively decreasing with
the distance from i and such that $\sum w_h = 1$
the decrease rate is learned from the data

- **Linear regression**: suitable for isolated extreme missing, possible for partial missing sequences with at least two numerical values

v_i = the value given by the straight line interpolating the available v_h

- Can follow a **trend** in the values. **Avoid negative** values by taking a power of the nearest value $(v_{i+1})^c$ with $c < 1$

Example: m,100,250,410,550,690 would give m = -44, impossible!
Instead, we could take $(v_2)^{0.5}$ (square root) and obtain m = 10

Imputation from the same unit

- We combine the weighted average value v_i^{WA} and the linear regression value v_i^{LR}
- **Trend smoothing imputation**: we do follow the trend, but don't want to be **excessively** (mis)lead by it
- Suitable for missing sequences up to $\mathrm{L}= t\text{-}2$ (at least two numerical values are available, otherwise no trend is possible)

$$v_i = (a^2 / a^2 +1)\, v_i^{\mathrm{WA}} + (1 / a^2 +1)\, v_i^{\mathrm{LR}}$$

- Coefficient $a \geq 0$ is the **key** for passing without discontinuities from LR when the trend is flat ($a = 0$) to WA when the trend is excessive ($a \to\infty$)

$$a = \frac{2|m|}{\min\limits_{h \neq i} \{w_h\}}$$

m angular coefficient of the interpolating straight line of LR normalized by using the minimum available value

Example: average case

- University of Western Brittany (France, ETER_id FR0026)

	Students	graduates	PhD stud	PhD grad
2011	16143	7337	540	135
2012	16618	8065	556	133
2013	16808	7447	575	152
2014	17203	7729	594	157
2015	17776 (smooth)	7784 (smooth)	611 (smooth)	165 (smooth)
2016	18199 (smooth)	7839 (smooth)	629 (smooth)	174 (smooth)
2017	18728	7895 (smooth)	647 (smooth)	181 (smooth)

- The **last years** were missing for many variables, imputed with trend smoothing
- Values appear very **plausible**, and follow a mild increasing trend. Ratios are comparable to similar available institutions; Trends are comparable to each other
- Of course, there is no **guarantee** of having the original values, which would be almost impossible. We are working on big data, not just few institutions
- But we can expect they are **not too far** from the unknown original values, and positive and negative errors should tend to **balance** each other

Example: difficult case

- Jožef Stefan International Postgraduate School (Slovenia, SI0022)

	Students	graduates	PhD stud	PhD grad
2011	36 (smooth)	8 (smooth)	216 (smooth)	61 (smooth)
2012	32 (smooth)	8 (smooth)	204 (smooth)	55 (smooth)
2013	29 (smooth)	8 (smooth)	190 (smooth)	49 (smooth)
2014	26	8	173	43
2015	24	8	175	40
2016	19	8	146	27

- The **first 3 years** were missing for many variables, and are reconstructed with the described trend smoothing imputation

- **Peculiar** institution because it has more PhD than masters (probably no bachelors programs in this institution)

- The available data are a bit **strange**: graduates are constant while other variables are mildly decreasing. There could be several causes for this, but we ignore them. However, the imputed values appear **in line with the available values** without creating too distortions in the relations among variables

Imputation from donor

- The cases of **full sequence missing** (e.g. 6 in 6 are m) or of **full sequence except one** (e.g. 5 in 6 are m) contains the majority of the missing values
- Trend smoothing cannot be used, **no trend** is available
- We impute from a **donor**. Unit under imputation is called **recipient**
- We take the donor(s) at minimum distance and use their values
- We need to define our **distance** function
- If there is plenty of possible donors, we can be very selective. If they are not so abundant (as it is in ETER, especially for some categories) we need to **compromise**
- The distance value of the donor is a measure of the **confidence** in the imputed values
- A donor cannot donate more than tot times, for instance 2

Distance function 1/2

- **Institution Category standardized**: 1 = University, 2 = University of Applied Science, 0= Other. Accept only donors from same category
- **Distance education institution**: if the institution is telematic or traditional. If different, gives a contribution p_1 (strong penalty) to the distance
- **Institution Category**: more granular. If different, adds p_1
- **Legal status**: public or private. If different, adds p_2 (weak penalty)
- **Expenditure, Revenues, Staff**. They describe the size of the institution. Each difference between two values v_a and v_b gives a contribution

$$p_1 \frac{| v_a - v_b |}{max \{v_a , v_b \}}$$

If a value is not available, its contribution becomes p_1

So, the contribution of each variable is between p_1 and 0

Distance function 2/2

- **Country**: we define geographical **areas**. Same country adds 0 to the distance. Different countries but same area adds p_2 (weak); different areas adds p_1 (strong)

1: Belgium, Liechten., Luxembourg, Netherlands, Switzerland

2: Austria, Germany

3: Greece, Italy, Portugal, Spain

4: Czech Republic, Slovakia, Estonia, Lithuania, Latvia, Hungary, Poland

5: Albania, Bulgaria, Croatia, North Macedonia, Romania, Serbia, Slovenia, Montenegro

6: Finland, Norway, Denmark, Iceland, Sweden

7: Ireland, Malta, United Kingdom

8: France

9: Cyprus, Turkey

Example: heavily incomplete

- EBC Hochschule Berlin (Germany, DE0256)

	Students	graduates	Acad HC	Nonacad HC
2014	2248	510	337	81
2015	2339	501	372	87
2016	2409	533	365	77

- The institution has only 3 years, and **all treated variables** are missing! Each time series is imputed from donor
- Donor: Katholische Stiftungsfachhochschule München (Germany) at distance 2. They are very similar, both universities of applied sciences
- Of course, in this situation we cannot **guarantee** values near to the unknown original values. However, given the similarity of the units, they should be **not too far**, and at the aggregate level positive and negative errors should tend to **balance** each other
- In any case, having values and a measure of the confidence (the distance) is **better than** having nothing!

Additional hints for donors

- We set also some exclusion criteria (filters). They guarantee the **quality** of the imputation by forbidding some donors. Clearly, they may leave some units without donors, so they are **not** imputed (the **price** of quality)
- In a second phase we may **relax** some requirements to impute as much as possible the remaining units, knowing that we sacrifice quality
- **Exclusion for size**: We require at least one size match between donor and recipient, i.e. at least one among students, graduates, staff, etc. should differ less than 30%. Note that the recipient is often heavily incomplete, so the size match may be not computable and the recipient can be imputed only in the relaxed phase
- **Exclusion for Ratios or Trends**: similarly, we require a match between donor and recipient on all ratios and trends involved in the imputation
- We also exclude form the donor set all units with **extreme values of ratios and trends** (e.g. top and bottom 2% - they represent oddities we don't want to replicate)

Donor Scaling

- If we impute a **full-except-one** sequence of missing values, for example: v_1 , m, m, m, m, m, and the donor sequence is $w_1, w_2, w_3, w_4, w_5, w_6$
- then we can use the available value v_1 to learn a **size ratio** $r = v_1 / w_1$ between donor and recipient
- Now, we use that size ratio to scale all the donor values and obtain for the recipient $v_1, rw_2, rw_3, rw_4, rw_5, rw_6$ with 2 **advantages**: we better respect the recipient features, and the sequence connects more smoothly to the only available value v_1
- This technique can be used also to impute a **full sequence** missing of a variable when the recipient has the values of another correlated variable
- For example, the recipient has $s_1, s_2, s_3, s_4, s_5, s_6$ for students and the full sequence of graduates is missing. If the donor has $t_1, t_2, t_3, t_4, t_5, t_6$ for students and $g_1, g_2, g_3, g_4, g_5, g_6$ for graduates, we learn the sequence of size ratios $r_1 = s_1 / t_1 , \ldots , r_6 = s_6 / t_6$ and impute the graduates of the recipient with $r_1 g_1, \ldots, r_6 g_6$

Example: donor scaling

- Finnish National Defence University (FI0024, recipient)
- The best available donor is the University of Lapland (FI0021)

	Donor Students	Donor Acad Staff FTE	Recipient Students	Acad Staff FTE **rescaled** from donor
2011	4517	284	659	(659 / 4517) x 284 = 41
2012	4338	290	774	(774 / 4338) x 290 = 52
2013	4240	304	788	(788 / 4240) x 304 = 56
2014	4002	296	861	(861 / 4002) x 296 = 64
2015	4032	297	840	(840 / 4032) x 297 = 62
2016	4020	305	824	(824 / 4020) x 305 = 63

- We are using the values of the donor to impute new values respecting the **size** and the **trend** of the recipient
- It is a **self-calibrated mechanism** that allow to use donors of different size and trend with respect to the recipient
- Particularly **useful** when the donors are not so abundant, as in our case

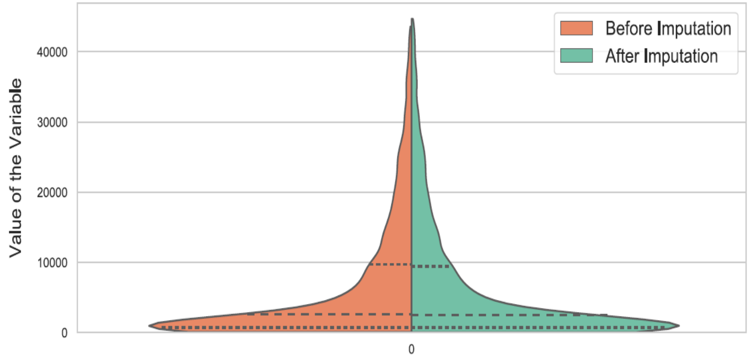
Results

- We use the described approach to impute all units in **ETER**
- Clearly the procedure has been designed **simple enough** to tackle this large dataset. Many additional improvements and subcases **could be added**, but they would greatly increase the complexity of the code and the computational burden
- The percentage of units without missing goes from **30%** for some variables before imputation to about **90%** or more for all variables after imputation
- Some units remain without an **acceptable** donor due to our filters. If we remove them, we could impute all units but the **quality** of the imputation would worsen

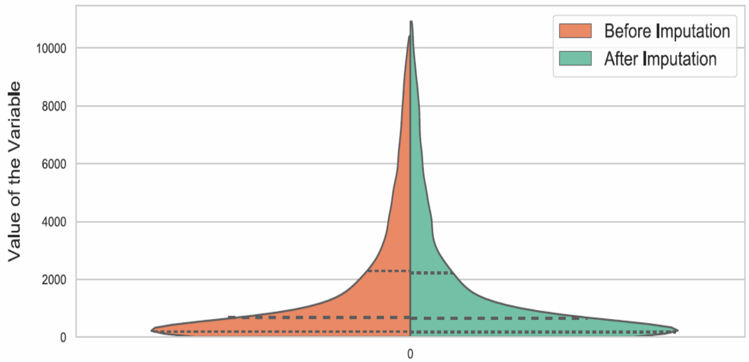
Before and After Imputation

- The analysis of the **frequency distributions** of the values of each variable **before and after imputation** shows that the frequency destitutions are generally respected
- We visualize it by means of the so-called **violin plots**. Examples:

Students enrolled



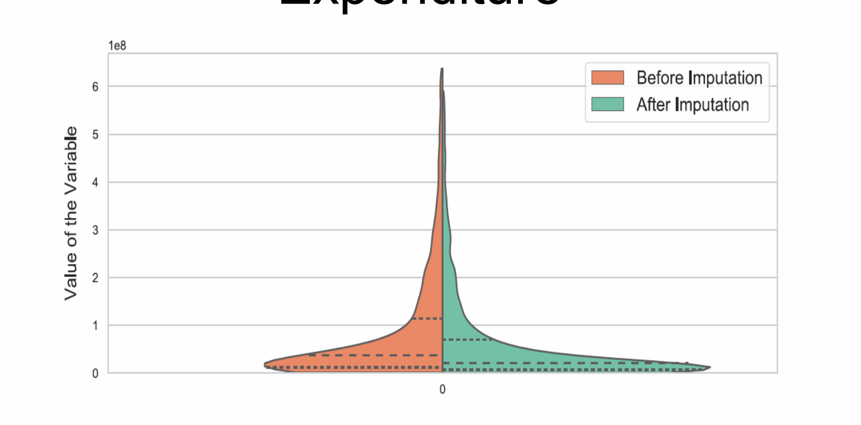
Graduates



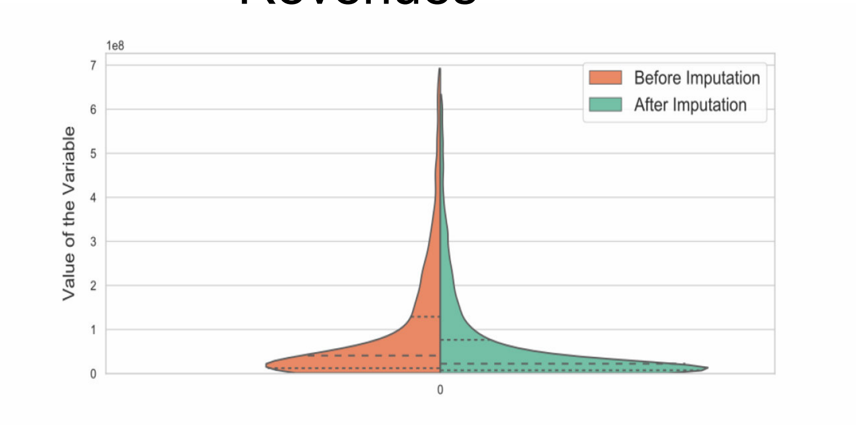
More In-Depth Analysis

- Sometimes they are **not** fully respected, in particular for Expenditure and Revenues. However, there is a **good reason**
- Missing values are not equally distributed over the institutions, but more concentrated on **small** institutions, especially for these 2 variables
- When the small institutions are imputed, **small values** would appear **more frequently** in the distribution, and this is correct
- On the contrary, imputing the small institutions with values similar to those of the larger institutions would not be correct

Expenditure



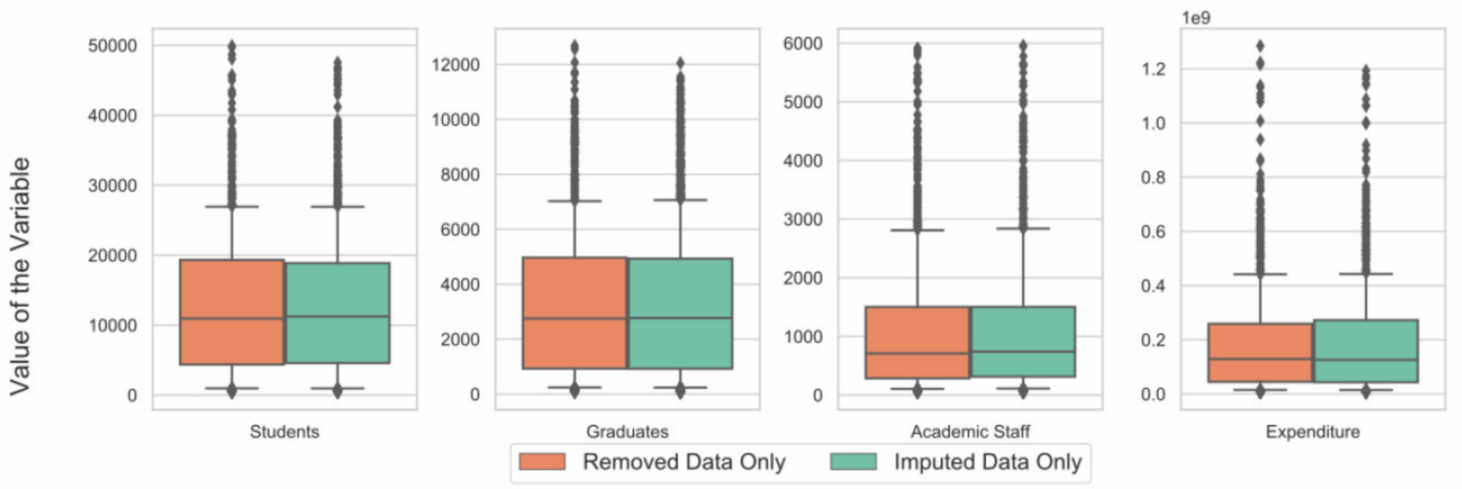
Revenues



How far are the imputed values?

- When we impute real missing values, we do not know the original values, so we don’t know how far we are from them
- To evaluate the **difference** between imputed and original value, we make another experiment
- We introduce **artificial** missing value in complete units: in this case **we know the original values**
- Now, we impute them and we compare **imputed and original values**
- We report the box plots. Note that the analysis is limited to removed data on the left and imputed data on the right, **to avoid diluting differences**!

Box Plot Analysis



- At the **aggregate** level, the imputed values appear **almost equivalent** to the original values
- More results in Bruni, Daraio, Aureli: Imputation Techniques for the Reconstruction of Missing Interconnected Data from Higher Educational Institutions, submitted

Conclusions

- HEI data are very important but they structurally contain non-negligible shares or missing values
- Experiments on the large real-world dataset ETER confirm that the imputation process is practically **feasible** and **useful**
- Experiments on the imputation of artificial missing values show that the reconstructed data are **satisfactory similar** to the original data
- The described procedure works at the **formal** level, with a **data driven** approach. It could be **adapted** to impute different datasets containing educational data from other origins, or even other interconnected datasets with completely different meaning