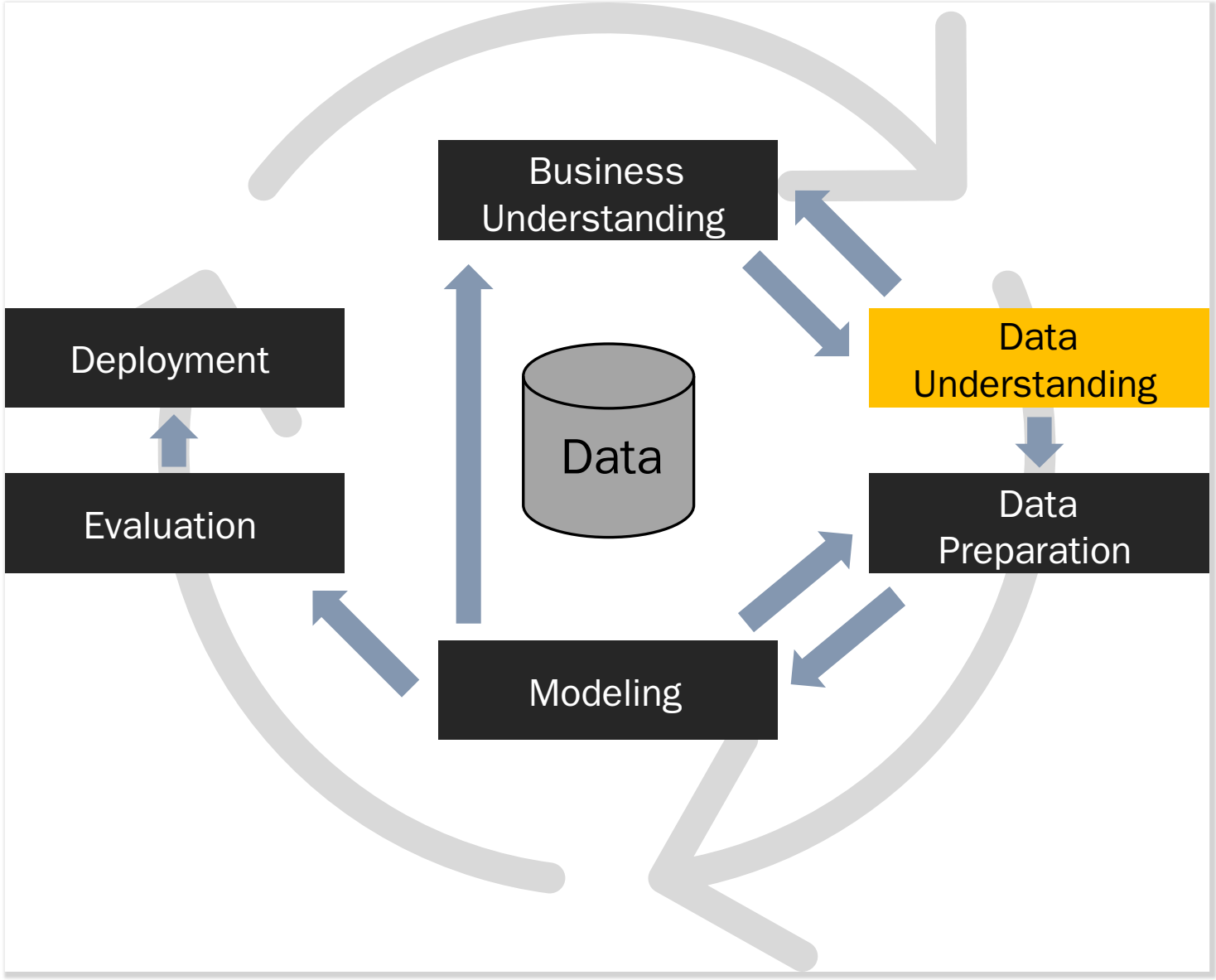
Data Mining

Data Understanding

Matteo Francia
DISI — University of Bologna
m.francia@unibo.it


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

Data understanding

The **data understanding** phase of CRISP-DM involves taking a closer look at the data available for mining

- This step is critical in preventing problems during data preparation, which is typically the longest part of a project
- The data understanding phase involves four steps, including:
    1. *collection* of initial data
    2. *description* of data
    3. *exploration* of data, and
    4. *verification* of data quality

Data collection (or acquisition)

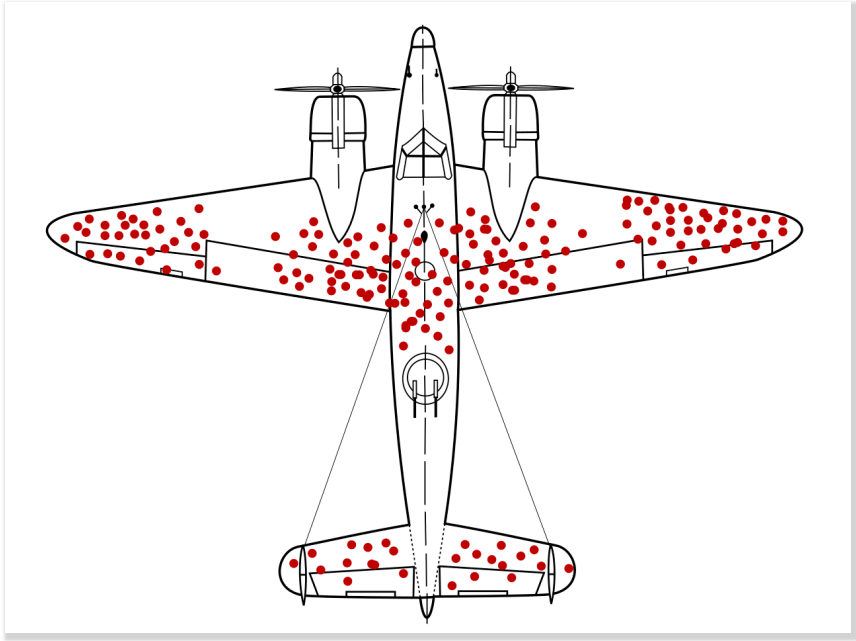
Data collection is the process of *gathering information on targeted variables* in an established system

- Capture evidence that allows data analysis to formulate credible answers to the questions that have been posed
- The analyst should make sure to report problems and solutions to aid with future replications of the project.
- Data may have to be collected from several different sources, and some of these sources may have a lag time.

The analyst then proceeds to

- increase familiarity with the data,
- identify data quality problems,
- discover initial insights into the data,
- detect interesting subsets to form hypotheses about hidden information

Problem: what about biases?



Red dots stand for places where surviving planes were shot. How would you reinforce the planes?

Biases

> "If you torture the data long enough, it will confess to anything"
>
> Ronald H. Coase

During data collection and analysis, several **biases** can occur

- *Selection*: sample used for data collection is not representative of the population being studied
- *Sampling*: certain segments of the population are more likely to be included or excluded from the sample
- *Response*: participants in a survey or study provide inaccurate or misleading responses
- *Confirmation*: refers to the tendency to favor information that confirms pre-existing beliefs or hypotheses while ignoring or discounting contradictory evidence
- *Cultural*: data collection methods, survey questions, or study designs are culturally insensitive or fail to account for cultural differences
- *Time-Interval*: the timing of data collection influences the results
- *Publication*: tendency for researchers or journals to publish studies with positive or significant results while neglecting to publish studies with null or negative results
- … and many others

Survivorship bias

The “survivors” get studied, while the failures are excluded, leading to potentially flawed conclusions.

Start-up Success Stories

> People often hear stories of wildly successful companies like Apple, Amazon, or Tesla and assume that hard work and a good idea are enough to succeed. This overlooks the countless failed start-ups that had hard-working teams and great ideas but didn’t survive due to market conditions, competition, or other factors.

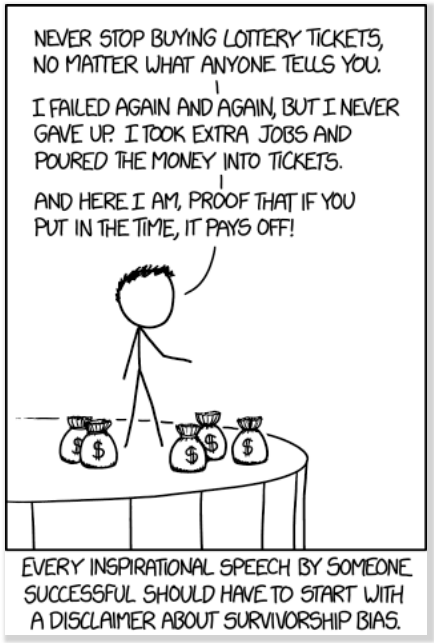
Fitness and Weight Loss Programs

> Testimonials for fitness programs often highlight people who achieved dramatic results. These “success stories” ignore the many individuals who followed the same program but didn’t achieve noticeable results, either due to differences in genetics, lifestyle, or other factors.

Investment Portfolios

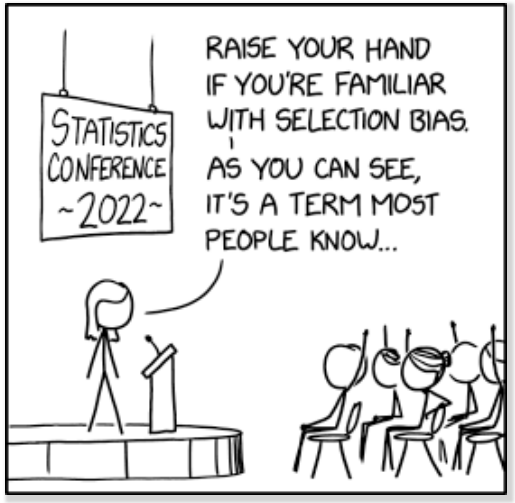
> Financial advice often highlights top-performing stocks or mutual funds as examples of great investments. These examples focus on the “survivors” in the market, ignoring the many investments that failed or underperformed, which can lead to overestimating the likelihood of similar success in the future.

Survivorship bias



https://xkcd.com/1827/

Selection bias



https://xkcd.com/2618/

… and more!

Confirmation Bias

> A person who believes in a specific political ideology might seek out news articles or social media posts that reinforce their beliefs, while ignoring or dismissing information that contradicts their views. This selective exposure to information strengthens their pre-existing opinions.

Anchoring Bias

> When shopping for a new car, a person may see a car priced at $30,000, and then a second car priced at $25,000. Even if the second car is not objectively better or a great deal, the first price “anchors” their perception of the value of the second car, leading them to think it’s a better deal simply because it’s cheaper than the first option.

Availability Bias

> After watching several news reports about airplane crashes, a person might overestimate the risk of flying. They might avoid flying, despite it being statistically safer than driving, because the images and stories of crashes are more readily available in their memory.

Problem: is data the new oil?

The more data we have, the more analysis we can do (however, more data != smarter data)

- There are several disciplines focusing on data (e.g., Data Science, Data Mining, Big Data, Business Intelligence)
- In Europe (but now in many areas of the world), there can be problems related to privacy
  - When is it “right” to protect privacy?
  - When does it become a limit?

Acquiring data is a time-consuming, investment, and knowledge-intensive process

- How much data is *enough*?

Rule of thumb: *one in ten/twenty* (Chowdhury and Turin 2020)

> There is no set rule as to the number of variables to include in a prediction model as it often depends on several factors. The ‘one in ten rule’, a rule that stipulates how many variables/parameters can be estimated from a data set, is quite popular in traditional clinical prediction modeling strategies (e.g., logistic regression and survival models). According to this rule, one variable can be considered in a model for every 10 events

Problem: how can we collect data?

Public Datasets

We can get data mainly in two ways:

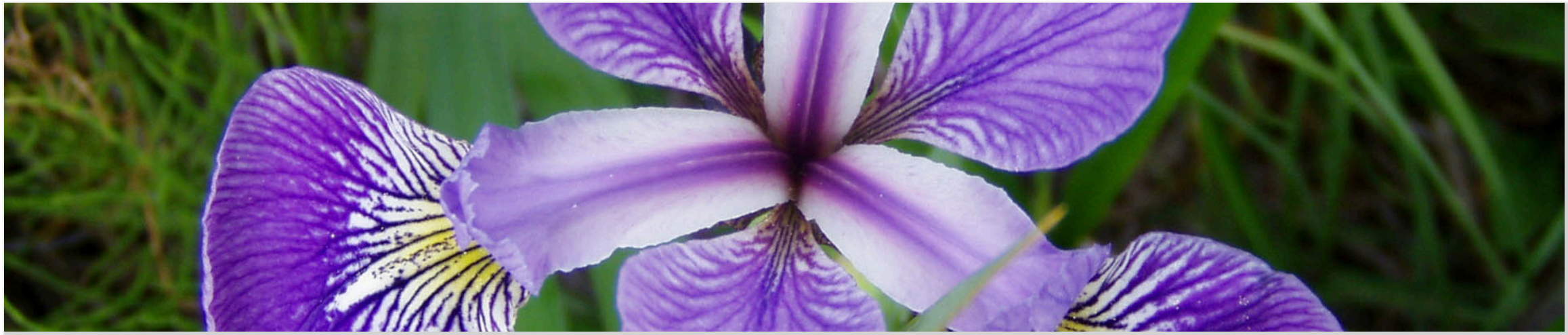
1. By using *publicly available data* (datasets or databases) → someone collected them for us!
    - They can be free or for a fee
    - The quality of the data made available must be checked
2. By *acquiring a new set of data*, but why?
    - It is not certain that public data well represent the problem we want to solve
    - We want to acquire specific data and thus generate specific expertise for the company (know-how)
    - We are forced to acquire data that due to their sensitive nature would not otherwise be available (privacy issues)
    - The company we work for already has a data collection process that we can use

Many universities publicly release their datasets:

- There are no requirements related to profit or non-disclosure agreement (NDA)
    - It is the basis of the scientific method, in particular for the reproducibility of the results obtained
    - I release my data so that others can conduct my own experiments and verify my results
    - Examples: https://www.image-net.org/

Some platforms make datasets available for competitions, such as Kaggle and others

Example of a public dataset: the Iris dataset



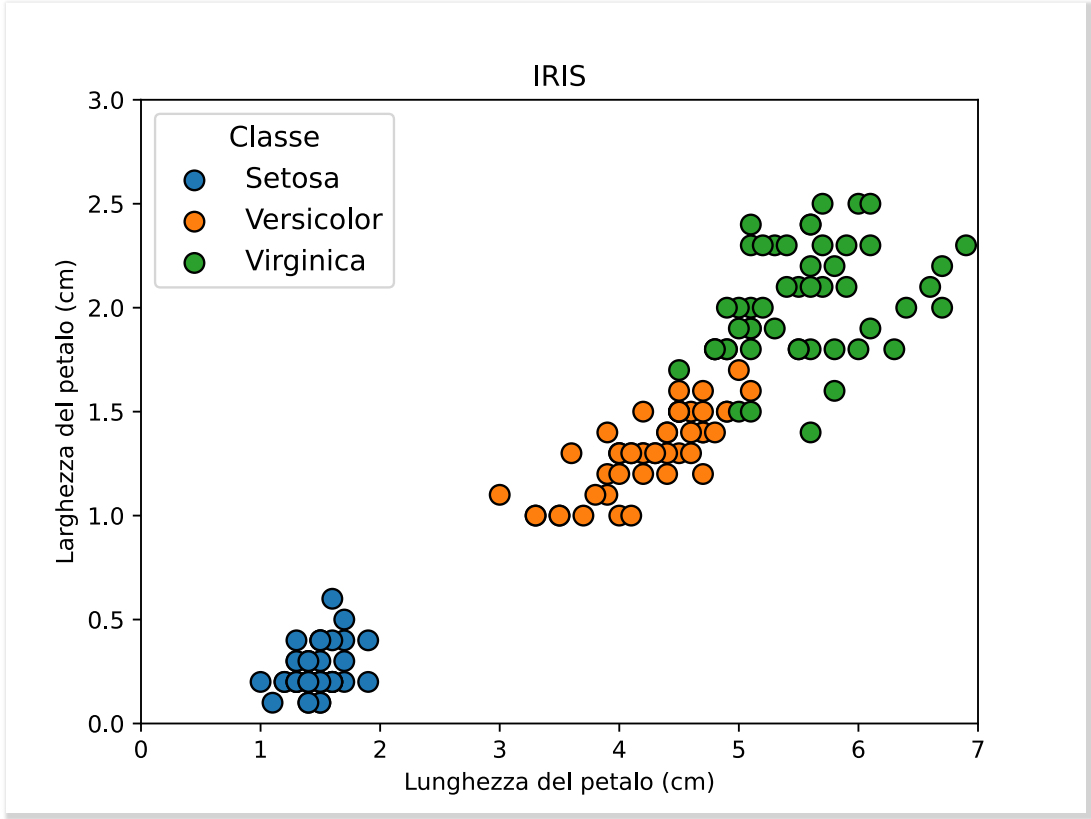
The Iris dataset is public data that was used in R.A. Fisher's classic 1936 paper (Fisher 1936)

- It can also be found on the UCI Machine Learning Repository.
- It includes 3 iris species (Setosa, Virginica, and Versicolor) with 50 samples each
- It characterizes flowers with some properties about each flower
    1. `SepalLengthCm`
    2. `SepalWidthCm`
    3. `PetalLengthCm`
    4. `PetalWidthCm`
    5. `Species`

Example of a public dataset: the Iris dataset

Used in classification problems where the goal is to predict the species of Iris flowers based on their features

- It is a simple dataset, it is easy to distinguish the different flowers and does not need data preparation



Acquisition of a new dataset

Acquiring a new dataset is usually a costly process!

- *Investment of time and money* for:
  - Programming or learning to use an acquisition tool
  - Handling of *large amounts of data*
  - Testing to find any bugs that could compromise the success of the acquisition
    - Unfortunately, we often notice them at the end of the process
  - *Acquire new hardware* for data collection and storage

It is necessary to carefully consider whether it is appropriate to acquire a new dataset

- Considerations not only in engineering but also in management and economics aspects
- Future needs must be foreseen in advance

Data Annotation

Acquiring a new dataset does not mean acquiring only new data!

Indeed, one of the most relevant aspects is the **annotation of the data**

> What would we do with the Iris dataset if we do not have the labels of each flower?

The specific annotation is usually called a “label” and is the (semantic) content of the data.

- A single data is therefore defined as *annotated* if it is associated with a label
- The *label* depends on the problem we want to solve and *can be numerical or categorical*
- Examples:
  - A person’s height prediction → data: joint lengths, label: height (cm)
  - Pedestrian Detection → data: images, label: presence of a pedestrian (yes/no)
  - Pedestrian Localization → data: images, label: position of the pedestrian (x, y, w, h)
  - Audio classification numbers → data: audio sequences, label: number (‘five’)

Data collected without correct and timely annotation is often useless

- However, it is also possible to “extract knowledge” from un-annotated data through, for instance, clustering

Data Annotation Process

The data annotation process can take place in several ways:

- *Manual*: each data is manually annotated
  - A long and expensive process
  - The quality of the annotations is usually controllable and high
  - This is not always an applicable process (for example, is it possible to annotate a dataset with 1M of images?)
- *Automatic*: each data is automatically annotated, using specific tools
  - It is based on particular a priori knowledge (for example, all images acquired in a dog shelter depict dogs).
  - The quality of the annotations is not always easily controlled
- *Third parties*: all data is noted by a third party
  - *Free of charge*: this is the case, for example, in which users barter the free use of some platform with the transfer of their annotated data (for example, photos uploaded - to Facebook accompanied by information regarding the content, the position of the face, or scene acquired).
  - *Paid*: there are platforms where it is possible to purchase annotation time from third parties (often from “developing countries”). Example: Amazon Mechanical Turk

Amazon Mechanical Turk

Pricing

Pay only for what you use.

The price you (the Requester) pay for a Human Intelligence Task ("HIT") is comprised of two components: the amount you pay Workers, plus a fee you pay Amazon Mechanical Turk (MTurk) which is based on the amount you pay Workers. Additional details are as follows:

Worker Reward

- You decide how much to pay Workers for each assignment.

MTurk Fee

- 20% fee on the reward and bonus amount (if any) you pay Workers.
- HITs with 10 or more assignments will be charged an additional 20% fee on the reward you pay Workers.
- The minimum fee is $0.01 per assignment or bonus payment.

Masters Qualification

- There is an additional fee for using the Masters Qualification (What is the Masters Qualification?)
- The fee is 5% of the reward you pay Workers.

Premium Qualifications

- There is an additional fee for using Premium Qualifications (How do I use Premium Qualifications?)
- The additional fee per assignment starts at $0.05 and varies depending on the qualification. View the full list of Premium Qualifications and their associated fees.

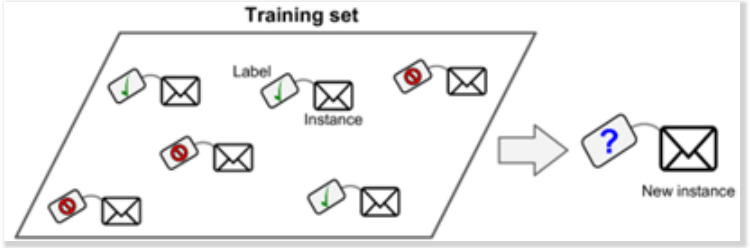
Different Ways of Learning

We define different types of learning depending on data annotation:

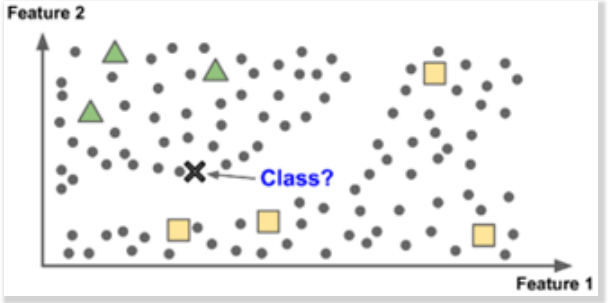
- *Annotated data* → Supervised Learning
  - One of the most studied types that allows to obtain the best results
- *Not annotated data* → Unsupervised Learning
  - Results that can be obtained are usually worse than the previous case
- *Partially annotated data* → Semi-Supervised Learning

Specific algorithms correspond to each of these areas

- Best performances are usually obtained with supervised trained algorithms
- We will mainly work on fully annotated data → Supervised Learning



Fully annotated



Partially annotated

Open and Closed Sets

The last aspect to be defined relating to data annotation: **do we know all annotations**?

Closed Set: it is assumed that the pattern to be classified belongs to one of the known classes.

- The most common case in machine learning benchmarks
- Ideal condition, but not always suitable for real-world systems

Open Set: the patterns to be classified can belong to none of the known classes.

- More realistic condition, but more challenging
- Example: classify all fruits into {pears, bananas}

Two possible solutions to the open set problem:

- An additional fictitious class is added to the classes (“the rest of the world”, “unknown”)
  - The so-called “negative examples” are added to the training set
- You allow the system not to assign the pattern
  - A threshold is defined and the pattern is assigned to the most likely class only when the probability is higher than the threshold

Common Problems in Data Collection

Companies usually face common problems:

- The business process produces huge amounts of data
  - It is almost impossible to acquire all the data
  - Also, physical limitations when the data stream is bigger than the storing capacity
  - Usually, it is necessary to choose which ones to store
- Sometimes companies have a lot of “old” data in their databases or information systems:
  - They don’t know what to do with it
  - Data re-collection on existing data (since data must be clean or something similar)
- In many business processes it is unclear understanding:
  - Which data is possible to collect (also due to privacy issues)
  - Which data is (really) useful for the business

Problem: now that we have the data, what do we do?

Describe the Data

The key question to ask is: does the data acquired satisfy the relevant requirements?

- This step also provides a basic understanding of the data on which subsequent steps will be built.
- For instance, if age is important and the data does not reflect the entire age range, it may be wise to collect a different dataset

The data analyst examines the “surface” properties of the acquired data, examining issues such as:

- the *format* of the data,
- the *quantity* of the data,
- the *number of records and fields* in each table,
- the *identities* of the fields,
- and any other *surface features of the data*.

Iris

Plain Iris dataset

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
…	…	…	…	…	…

Iris

Example of profiling the schema of the data in Iris

```
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
#   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
0   sepal length (cm)  150 non-null    float64
1   sepal width (cm)   150 non-null    float64
2   petal length (cm)  150 non-null    float64
3   petal width (cm)   150 non-null    float64
4   species            150 non-null    object 
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

Iris

Example of profiling the distribution of the data in Iris

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm
count	150	150	150	150	150
mean	75.5	5.84333	3.054	3.75867	1.19867
std	43.4454	0.828066	0.433594	1.76442	0.763161
min	1	4.3	2	1	0.1
25%	38.25	5.1	2.8	1.6	0.3
50%	75.5	5.8	3	4.35	1.3
75%	112.75	6.4	3.3	5.1	1.8
max	150	7.9	4.4	6.9	2.5

Iris

In descriptive statistics, a **box plot** shows graphically the locality, spread, and skewness groups of numerical data

A boxplot is a standardized way of displaying the dataset based on the five-number summary:

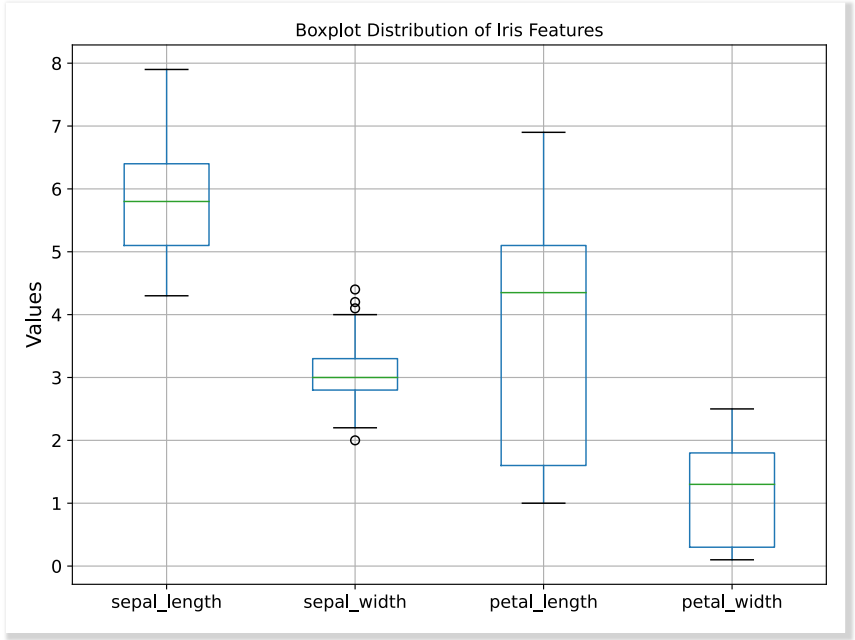
1. *Minimum* (Q_0 or 0th percentile): the lowest data point
2. *First quartile* (Q_1 or 25th percentile)
3. *Median* (Q_2 or 50th percentile): the middle value
4. *Third quartile* (Q_3 or 75th percentile)
5. *Maximum* (Q4 or 100th percentile): the highest data point

Interquartile range: $\mathrm{IQR} = Q_3 - Q_1$

Graphical elements

- The *box* is drawn from Q_1 to Q_3
- *Whiskers* are based on the $1.5 \cdot IQR$ value
  - A whisker is drawn up to the largest/lowest observed data point from the dataset that falls within this distance
  - The whisker lengths can look unequal
- All other points outside the whiskers are plotted as *outliers*

Iris data



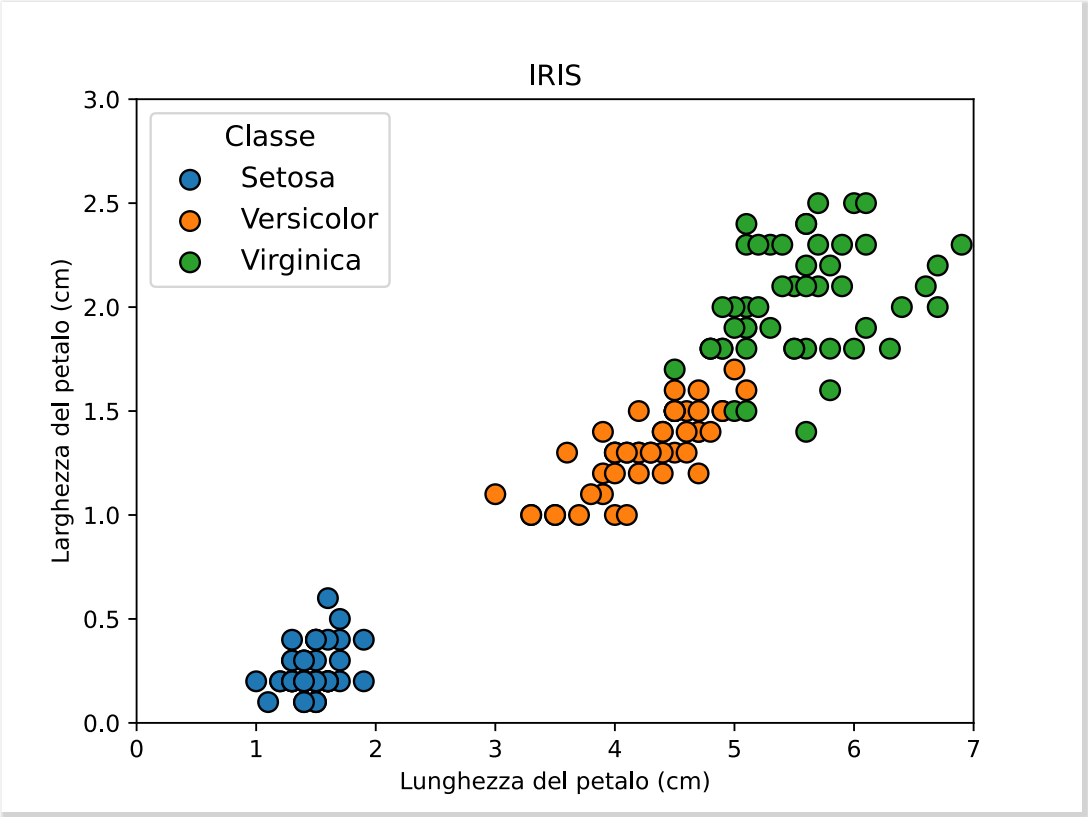
Box plot

Explore the Data

This task tackles the data mining questions, which can be addressed using querying, visualization, and reporting.

- Create a data exploration report that outlines the first findings, or an initial hypothesis
- For instance, query the data to discover the types of products that purchasers in a particular income group usually buy

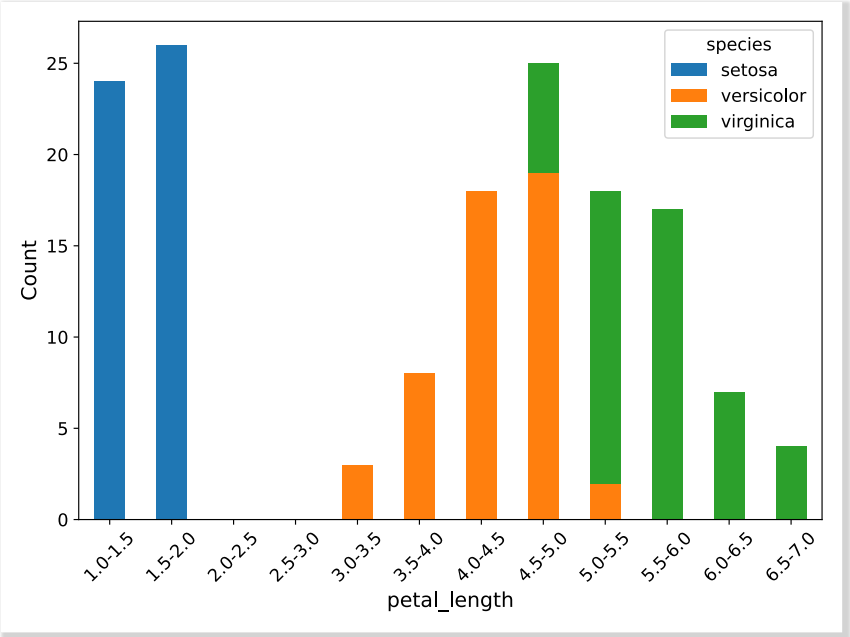
> 2D visualization of the Iris dataset



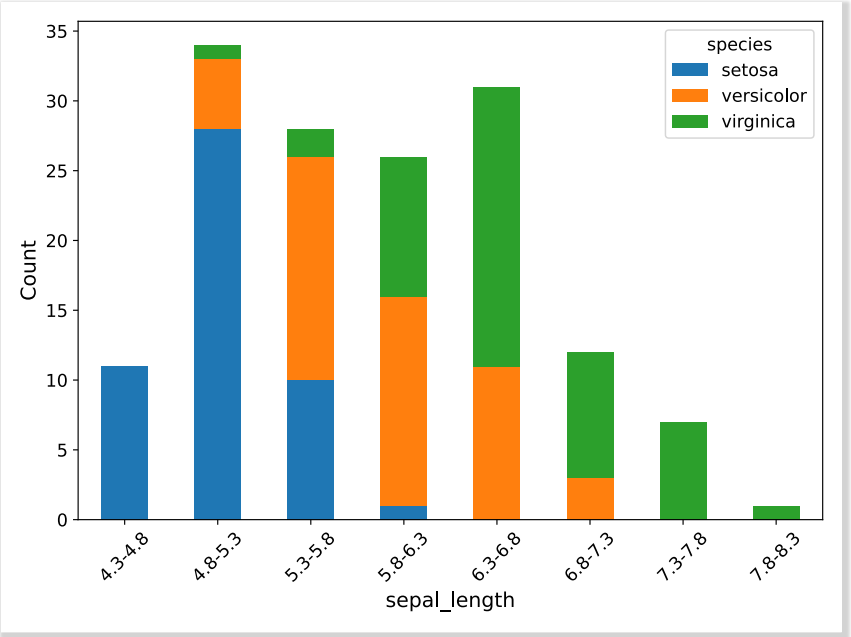
Iris

Value distribution vs `species`

`petal_length` vs `species`



`sepal_length` vs `species`



Verify Data Quality

At this point, the analyst examines the quality of the data, addressing questions such as:

Some common data quality issues to check include:

- *missing values* or attributes;
  - E.g., unknown date of death (Is it missing? Is the person alive?)
- whether *all possible values are represented*;
  - E.g., all age groups are contained in the dataset
- the *plausibility of values*, review any attributes that may give answers that conflict with common sense;
  - E.g., teenagers with high income
- whether attributes with different values have *similar meanings*;
  - E.g., `low fat` and `diet`
- the *spelling* of values.
  - E.g., `law fat` or `low fat`?

> Nothing to worry about in Iris

Dimensions of Data Quality (Sidi et al. 2012)

Types of Data	Definition
Consistency	The extent to which data is presented in the same format and compatible with previous data [15]. Refer to the violation of semantic rules defined over the set of data [2].
Accuracy	Data are accurate when data values stored in the database correspond to real-world values [2, 19]. The extent which data is correct, reliable and certified [15]. Accuracy is a measure of the proximity of a data value, v, to some other value, v', that is considered correct [2, 17]. A measure of the correction of the data (which requires an authoritative source of reference to be identified and accessible [14].
Completeness	The ability of an information system to represent every meaningful state of the represented real world system [2, 11]. The extent to which data are of sufficient breadth, depth and scope for the task at hand [15]. The degree to which values are present in a data collection [2, 17]. Percentage of the real-world information entered in the sources and/or the data warehouse [2, 18]. Information having all having all required parts of an entity's information present [2, 16]. Ratio between the number of non-null values in a source and the size of the universal relation [2, 20]. All values that are supposed to be collected as per a collection theory [2, 21].

Dimension	*Definition*
Timeliness	The extent to which age of the data is appropriated for the task at hand [15]. Timeliness refers only to the delay between a change of a real world state and the resulting modification of the information system state [2, 11]. Timeliness has two components: age and volatility. Age or currency is a measure of how old the information is, based on how long age it was recorded. Volatility is a measure of information instability the frequency of change of the value for an entity attribute [2, 16].
Currency	Currency is the degree to which a datum is up-to-date. A datum value is up-to-date if it is correct is spite of possible discrepancies caused by time-related changes to the correct value [2, 17]. Currency describes when the information was entered in the sources and/or the data warehouse. Volatility describes the time period for which information is valid in the real world [2, 18].

Metrics of Data Quality (Batini et al. 2009)

Table IX. Dimensions and Metrics

Dimensions	Name	Metrics Definition
Accuracy	Acc1	Syntactic accuracy: it is measured as the distance between the value stored in the database and the correct one Syntactic Accuracy=Number of correct values/number of total values
	Acc2	Number of delivered accurate tuples
	Acc3	User Survey - Questionnaire
Completeness	Compl1	Completeness = Number of not null values/total number of values
	Compl2	Completeness = Number of tuples delivered/Expected number
	Compl3	Completeness of Web data = $(T_{max} - T_{current})^{*}$ $(Completeness_{Max} - Completeness_{Current})/2$
	Compl4	User Survey - Questionnaire
Consistency	Cons1	Consistency = Number of consistent values/number of total values
	Cons2	Number of tuples violating constraints, number of coding differences
	Cons3	Number of pages with style guide deviation
	Cons4	User Survey - Questionnaire
Timeliness	Time1	Timeliness = $(\max(0; 1\text{-Currency/Volatility}))^{s}$
	Time2	Percentage of process executions able to be performed within the required time frame
	Time3	User Survey - Questionnaire
Currency	Curr1	Currency = Time in which data are stored in the system - time in which data are updated in the real world
	Curr2	Time of last update
	Curr3	Currency = Request time- last update
	Curr4	Currency = Age + (Delivery time- Input time)
	Curr5	User Survey - Questionnaire
Volatility	Vol1	Time length for which data remain valid
Uniqueness	Uni1	Number of duplicates
Appropriate amount of data	Appr1	Appropriate Amount of data = Min ((Number of data units provided/Number of data units needed); (Number of data units needed/Number of data units provided))

References

Batini, Carlo, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. 2009. “Methodologies for Data Quality Assessment and Improvement.” *ACM Computing Surveys (CSUR)* 41 (3): 1–52.

Chowdhury, Mohammad Ziaul Islam, and Tanvir C Turin. 2020. “Variable Selection Strategies and Its Importance in Clinical Prediction Modelling.” *Family Medicine and Community Health* 8 (1).

Fisher, Ronald A. 1936. “The Use of Multiple Measurements in Taxonomic Problems.” *Annals of Eugenics* 7 (2): 179–88.

Sidi, Fatimah, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. “Data Quality: A Survey of Data Quality Dimensions.” In *2012 International Conference on Information Retrieval & Knowledge Management*, 300–304. IEEE.