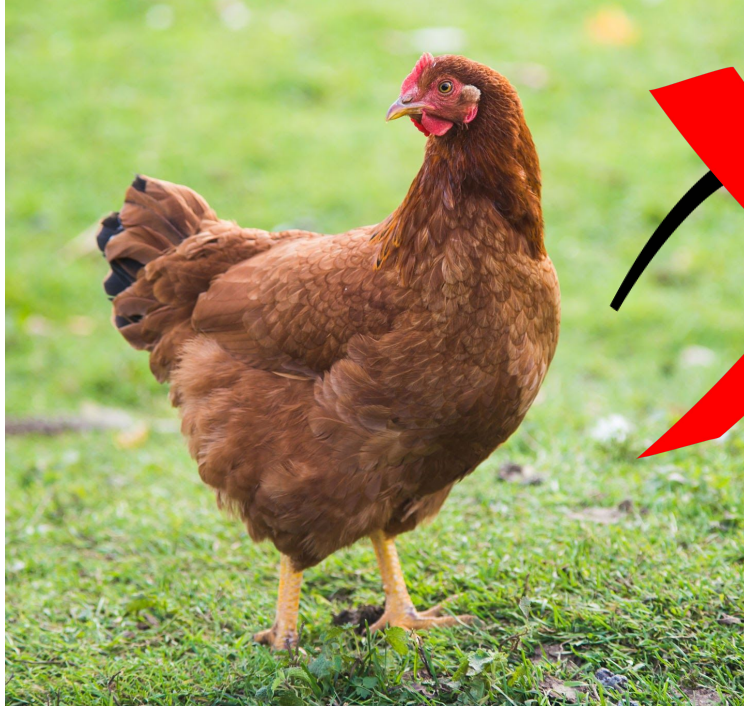




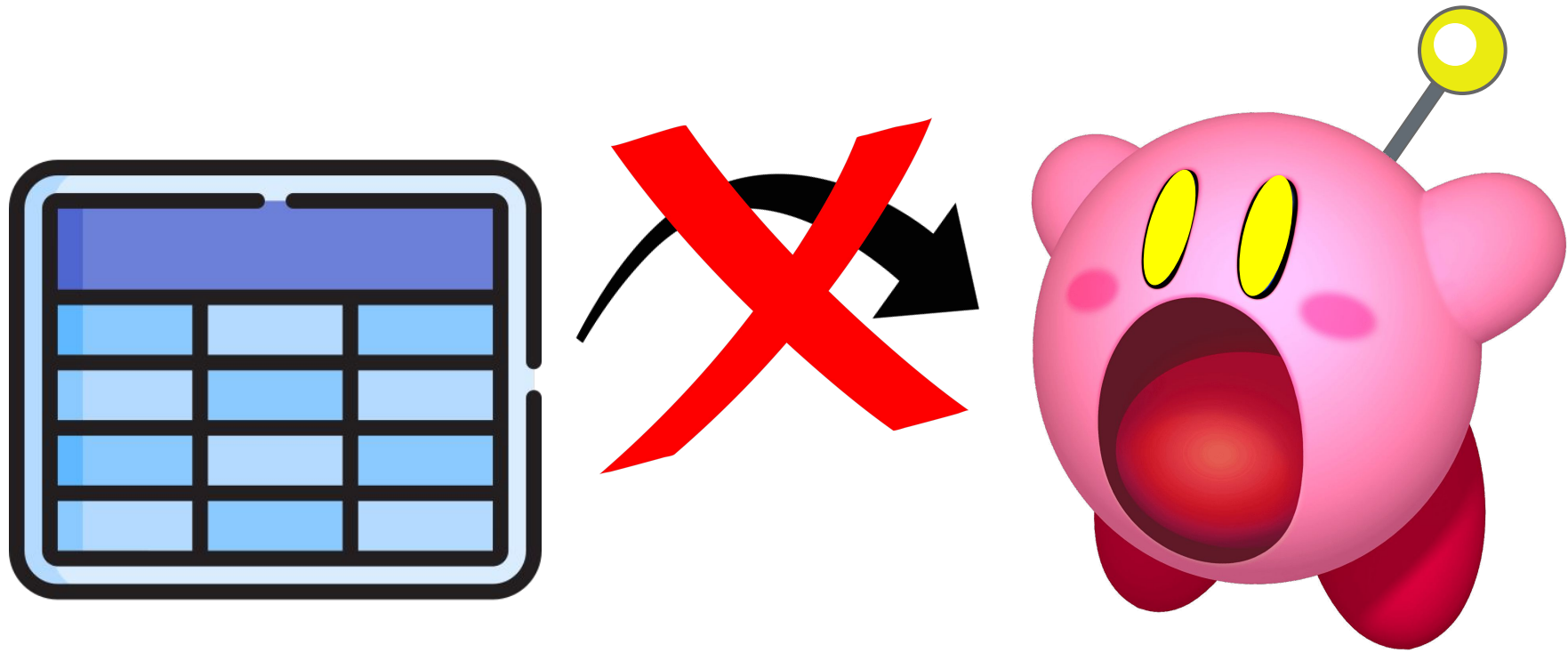
Would you eat raw chicken ?



You don't eat raw chicken...

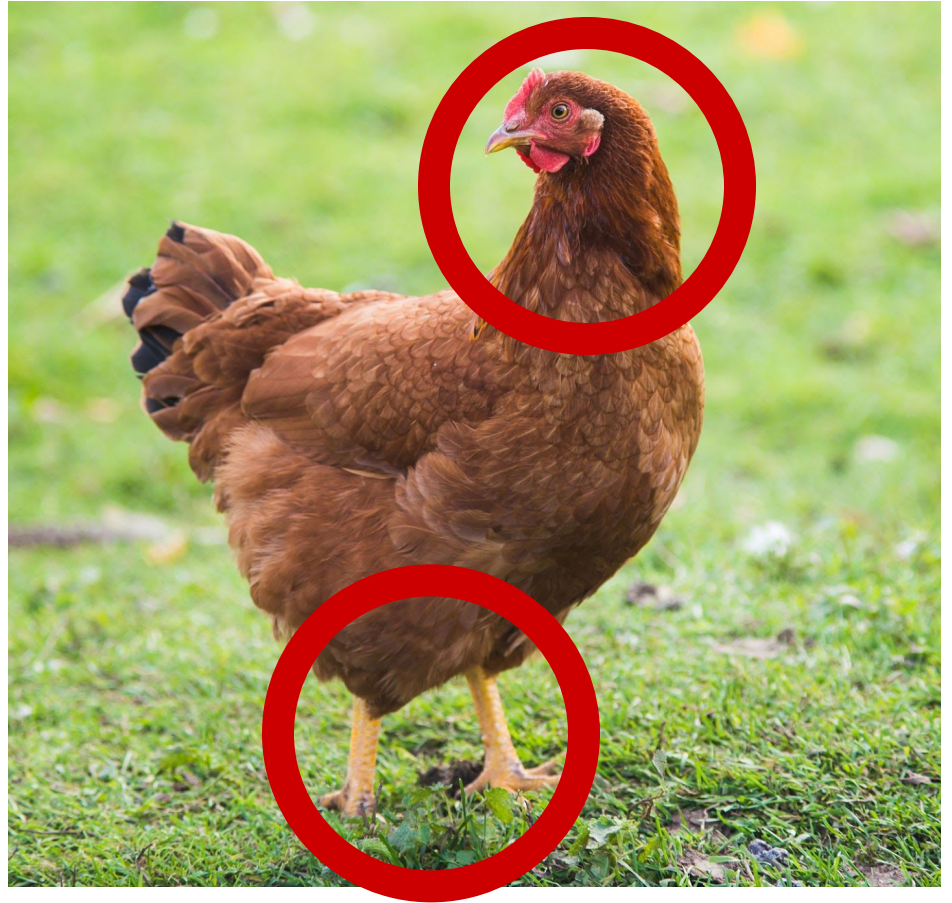


You eat chicken nuggets!

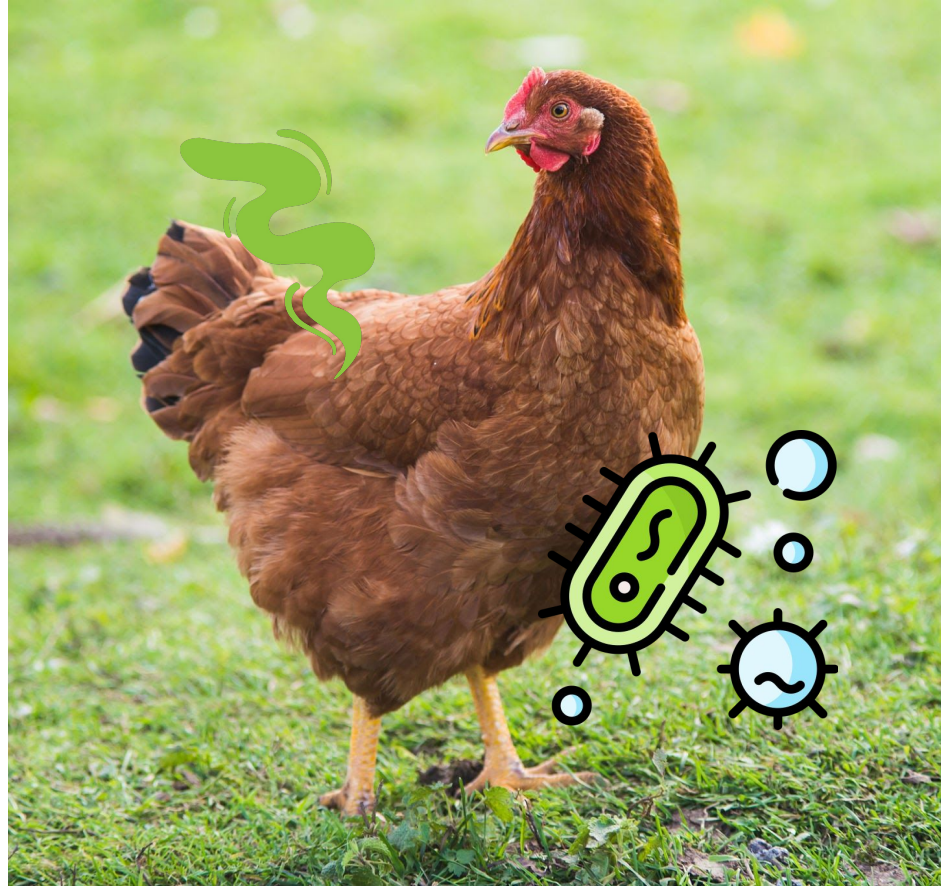


Similarly, you don't feed raw data to an algorithm

There are things
you don't eat in the
chicken



The chicken itself
is not clean



Your data is the
exact same!



Data Science

Session 1 - Understanding data



hadrien.salem@centralelille.fr



[introduction-to-data-science](#)

Introduction

The importance of data

DATA

Value carrying information

Literal, numerical,
boolean, etc.

Amounts, facts, statistics, etc.

⇒ Using data is using information to your advantage



Each and every activity generates data – [What is Data Science? on hackr.io](https://www.hackr.io/what-is-data-science)

Vocabulary

Dataset

Big Data

Data Analysis

Data Engineering

Data Science

Vocabulary

Dataset

An organised structure
containing data

Big Data

A lot of data

Data Analysis

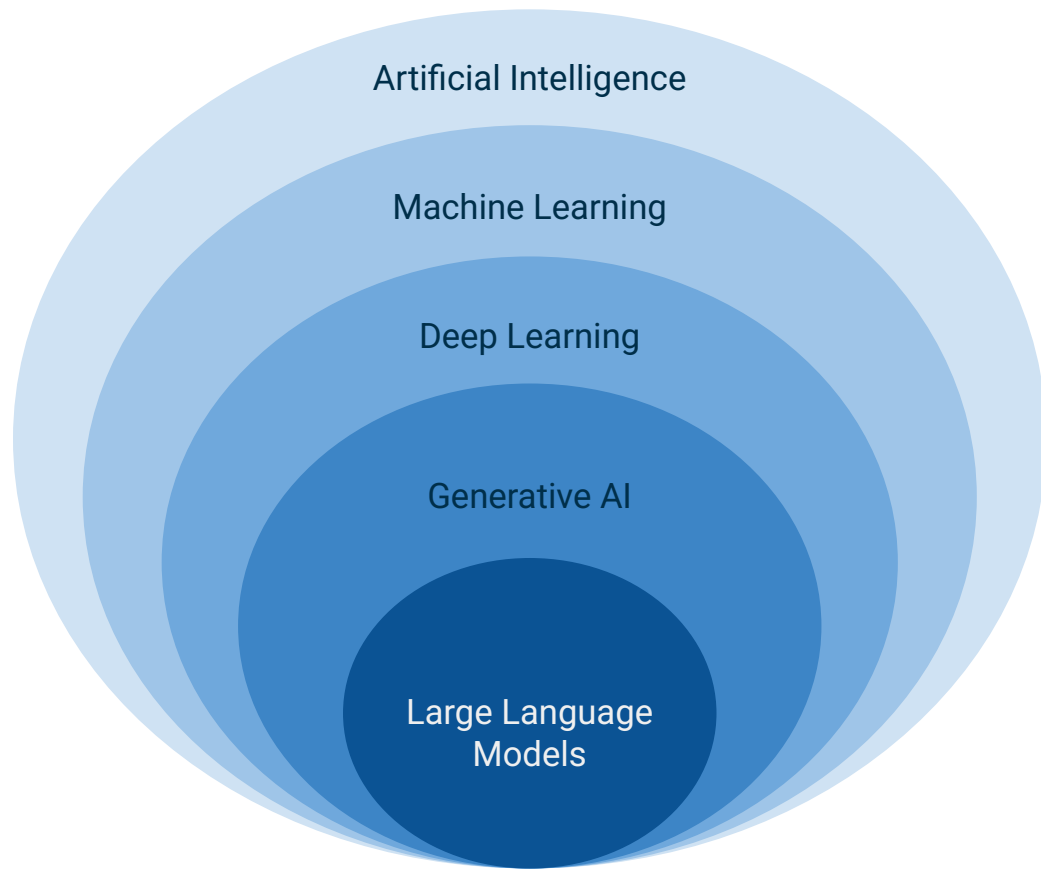
Analyse data to understand it

Data Engineering

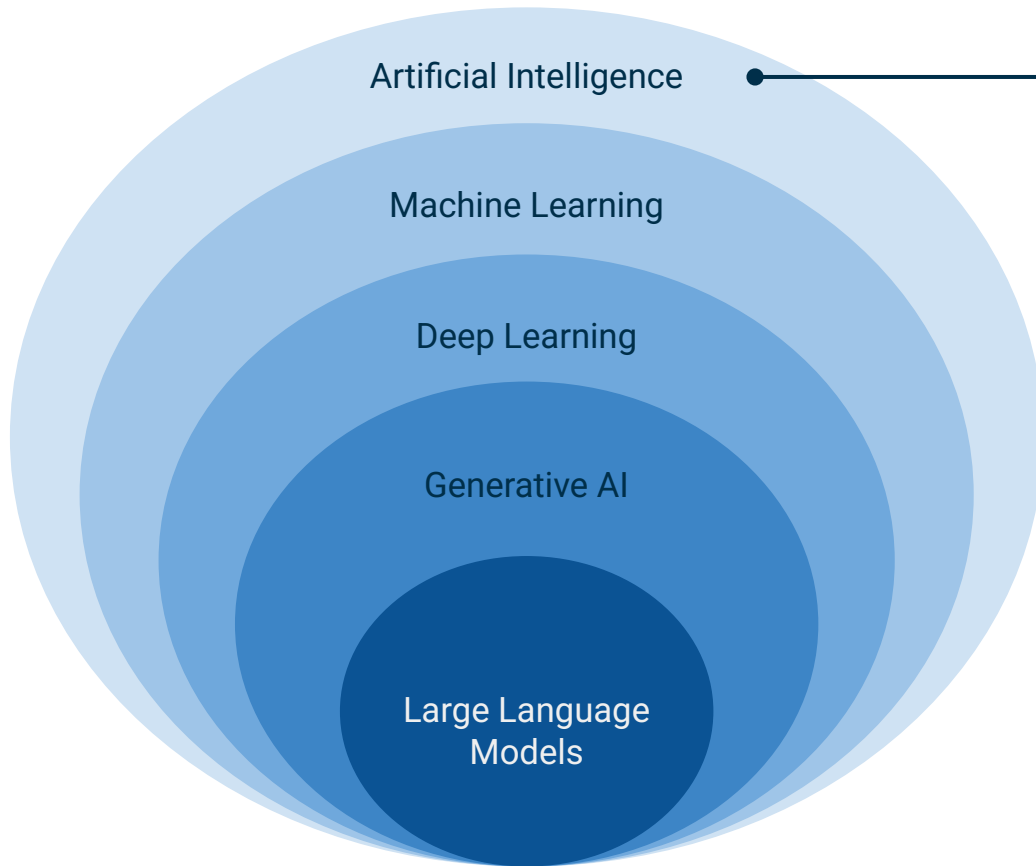
Prepare data for future use

Data Science

Modelling data

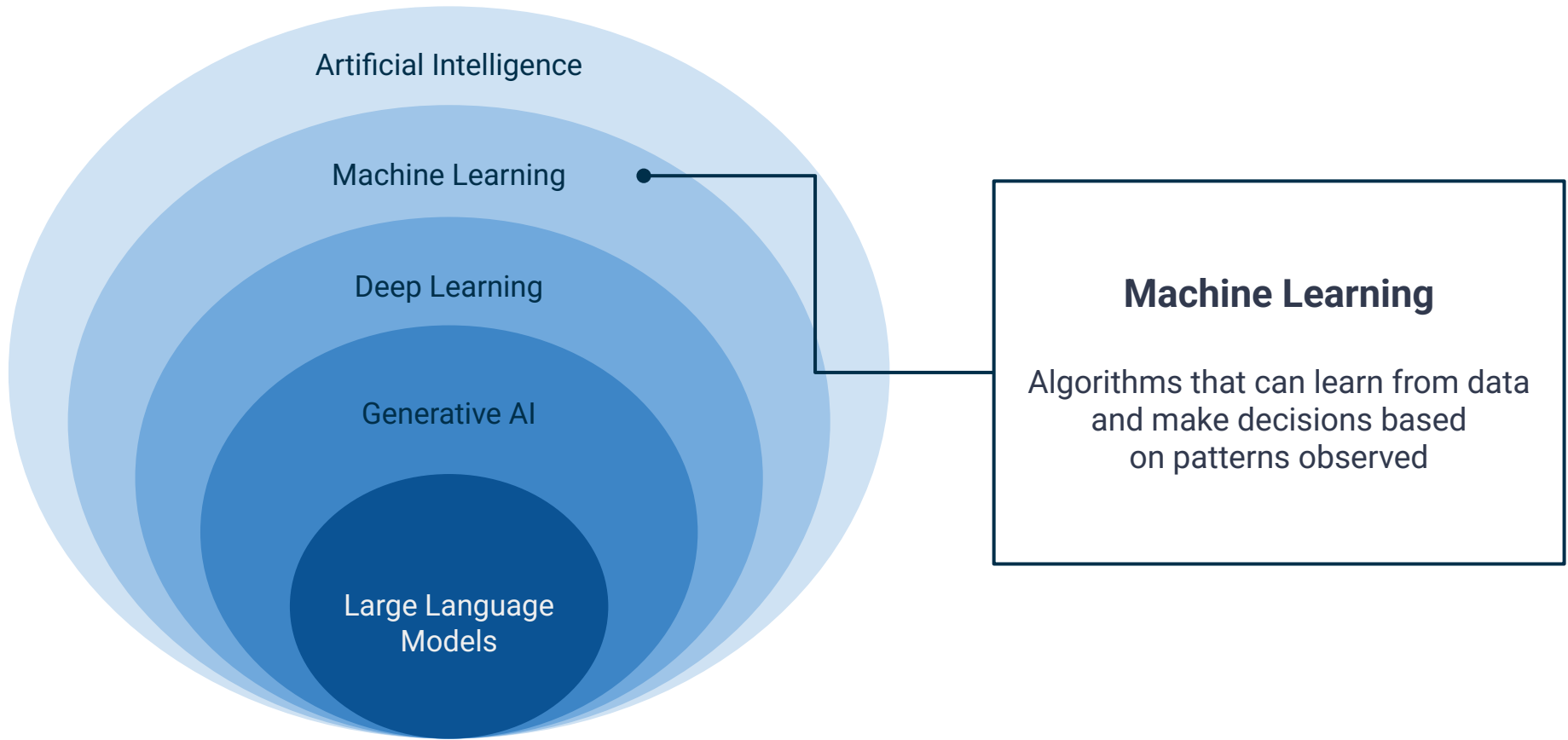


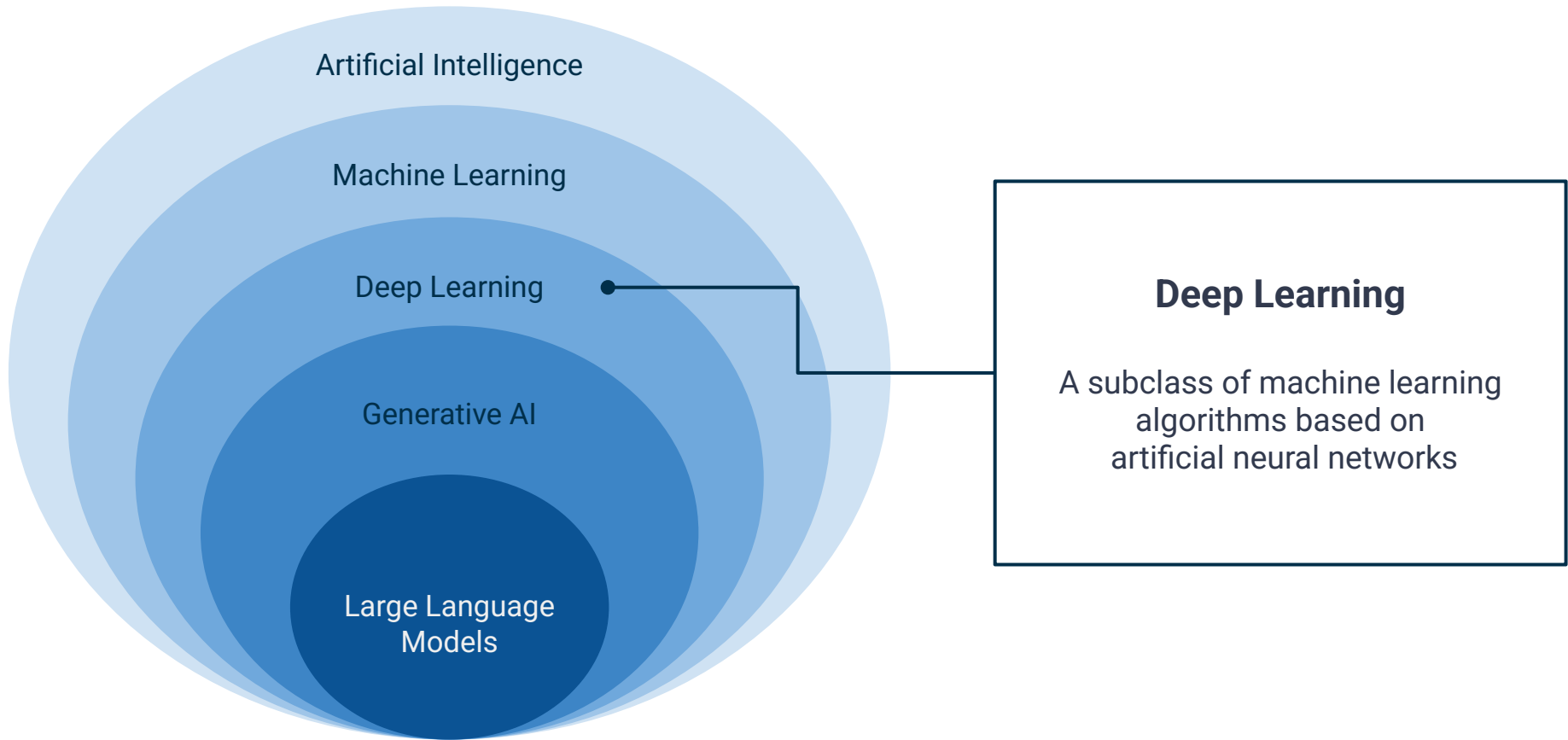
What “Artificial Intelligence” *really* means

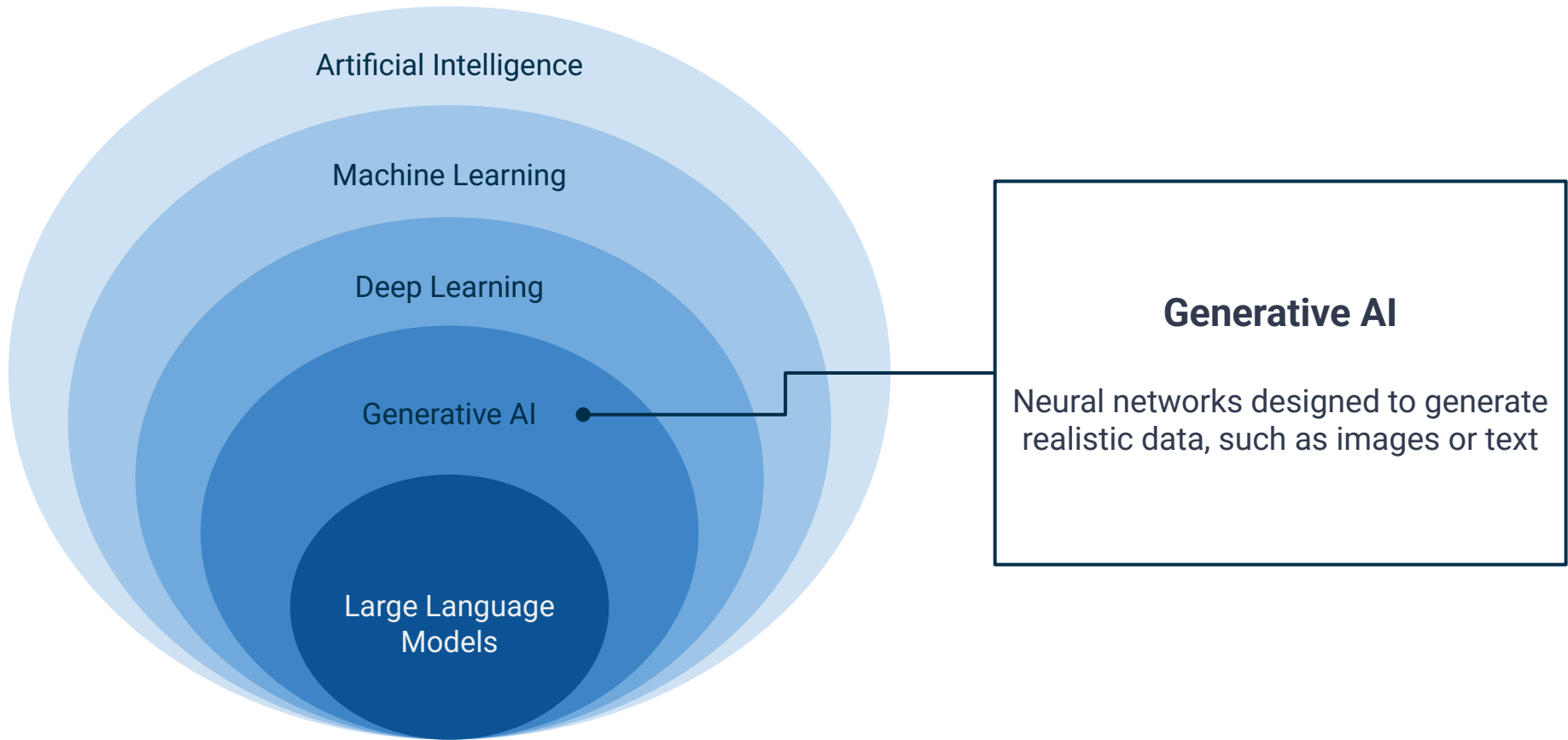


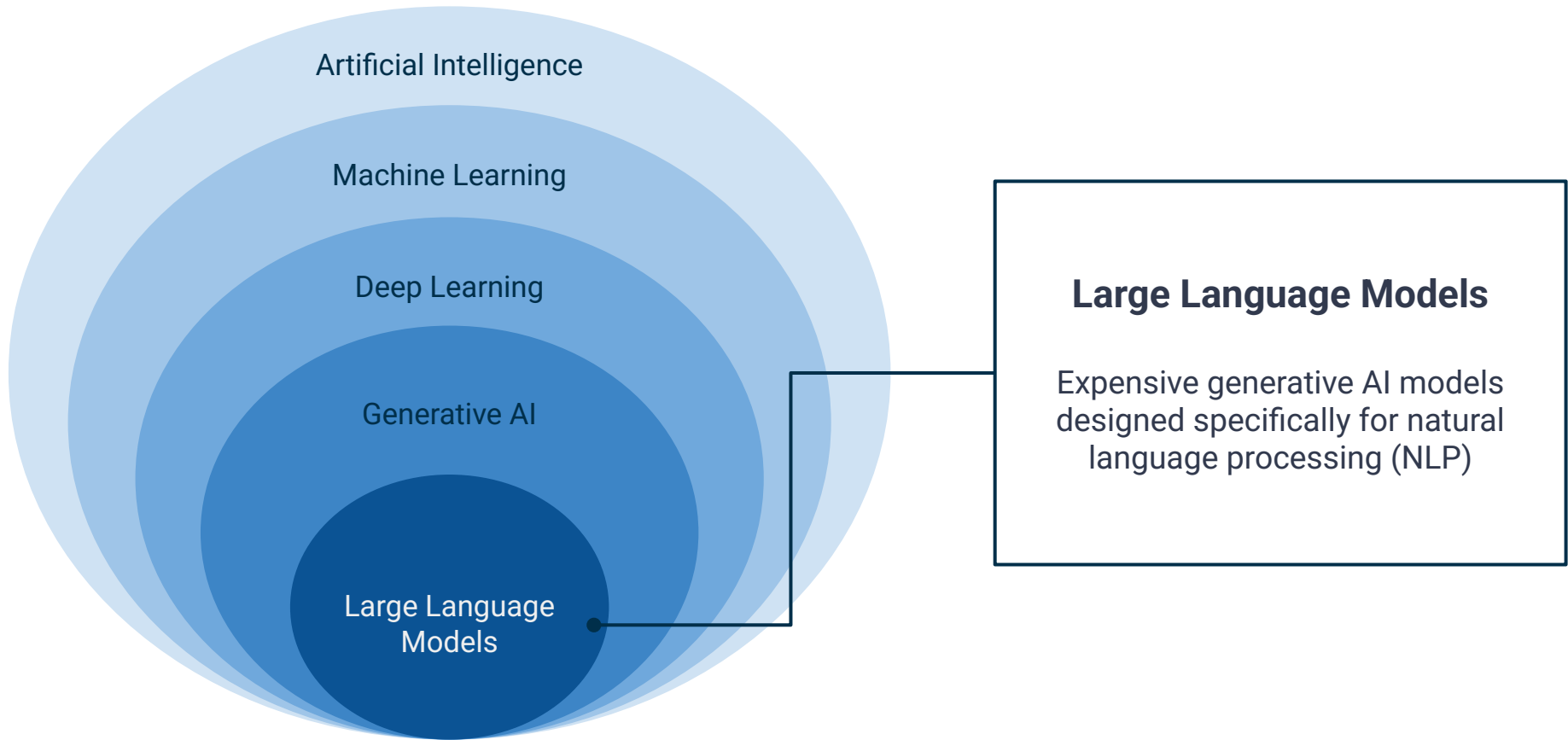
Artificial Intelligence

Development of smart systems and machines that can carry out tasks that typically require human intelligence









Examples in healthcare

There are many applications for data exploitation in healthcare, both in research and in the industry.

Disease prediction

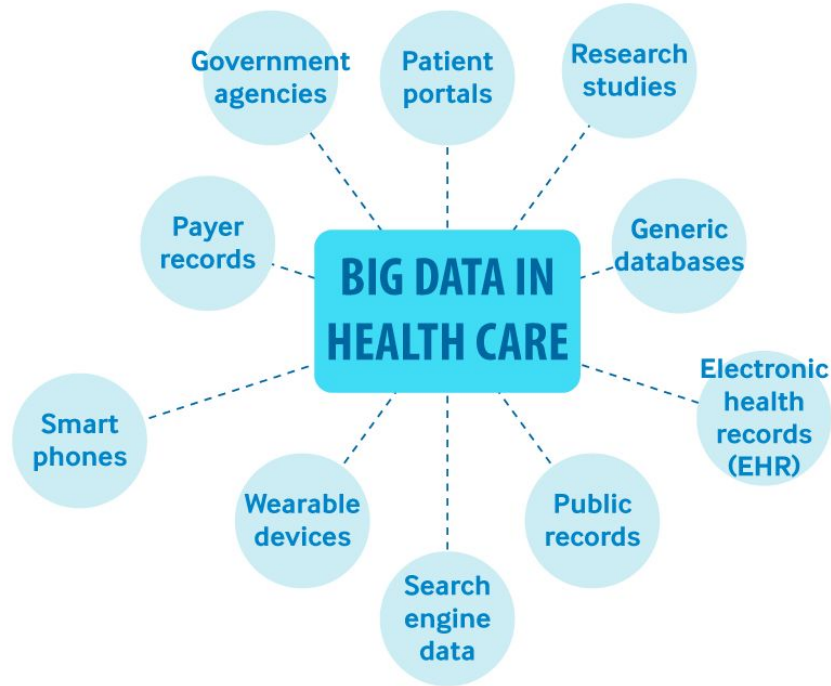
Chat bots

Appointments management

Alerting patients

... etc.

Sources of Big Data in Health Care



NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

The healthcare sector involves many actors who generate data

[Image source](#)

The healthcare sector can be difficult to work with

Healthcare is a high-impact subject involving many actors with conflictual interests.

Heavy legal constraints

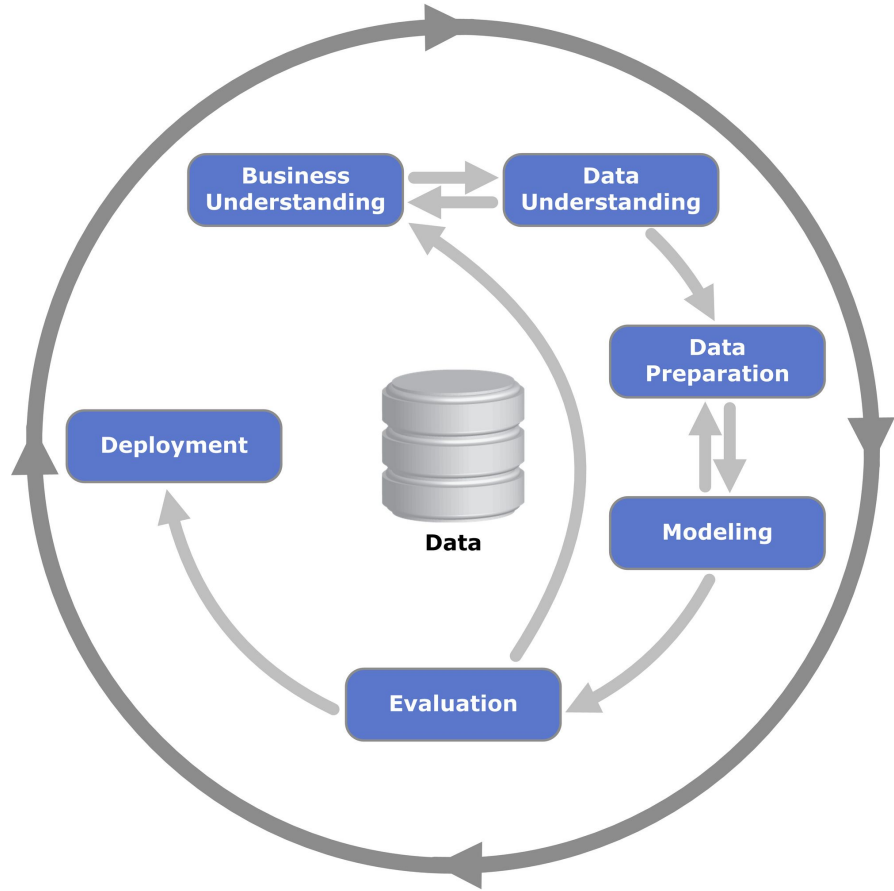
Political issues

Reluctance of certain actors

Abundant but unclean data

... etc.

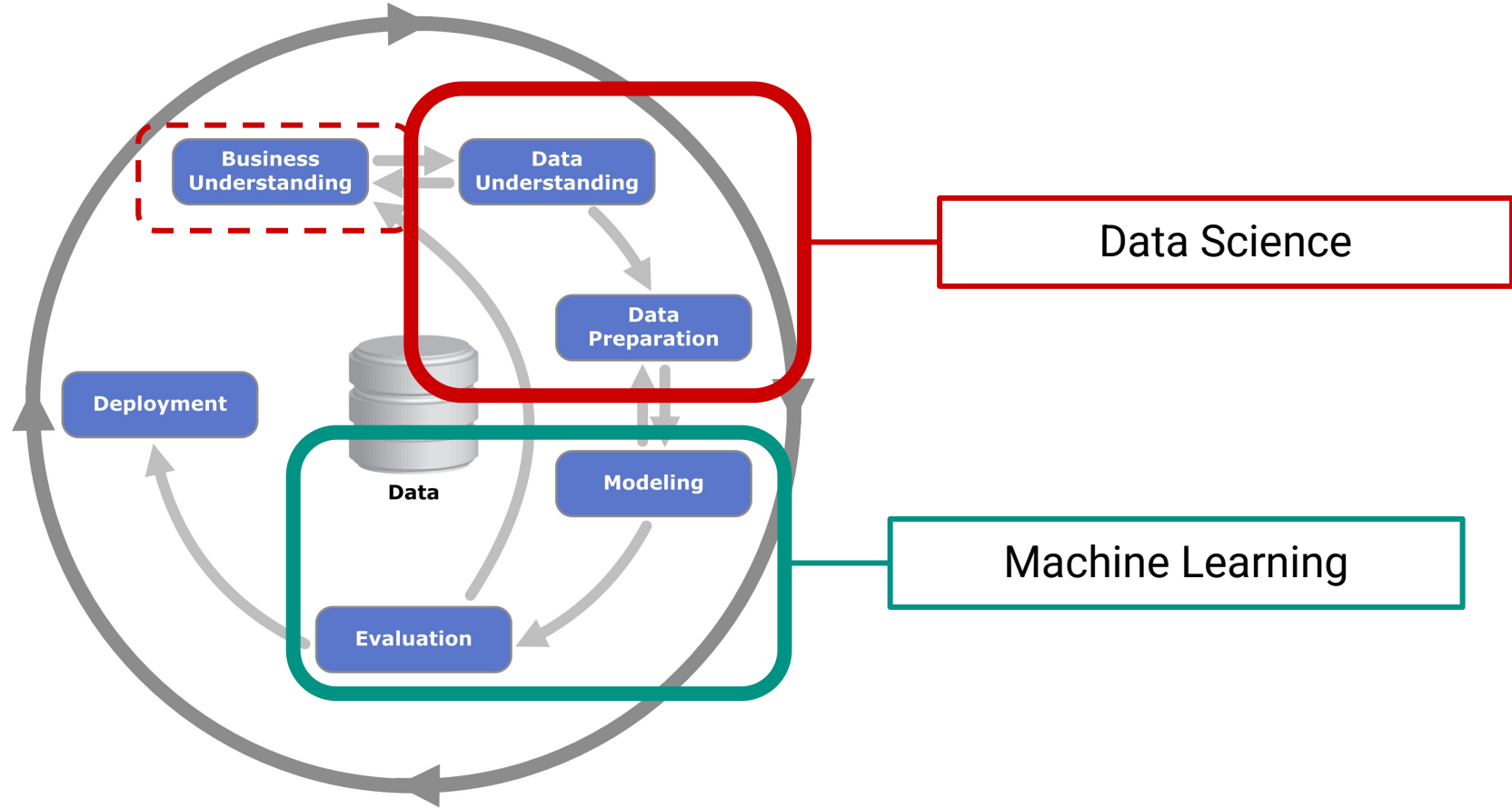
How does one
leverage data?

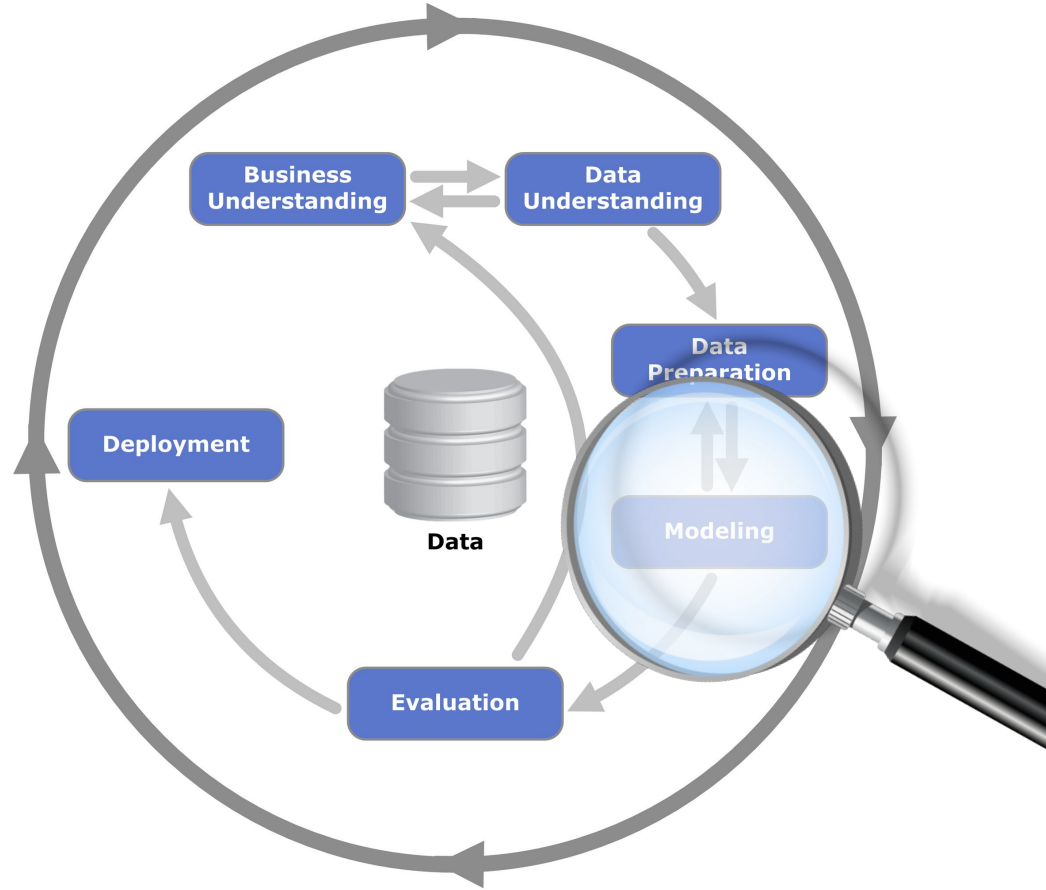


The CRISP-DM method

Cross-Industry Standard Process for Data Mining

- Published in 1999
- Common in the industry
- Still relevant today





The CRISP-DM method to carry out data-driven projects

(Image source: Wikipedia)

Leveraging data is a
complex subject that goes
beyond using algorithms

Course outline

Data science course

Session 1: Understanding data

Session 2: Clean code & Git

Session 3: Preparing data - Cleaning & Missingness

Session 4: Preparing data - Dimensionality reduction

Session 5: Preparing data - Data imbalance



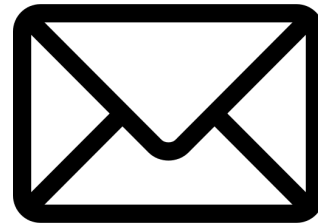
Machine learning course

Workflow

1. Introduction - Reminders - Questions
2. Theoretical elements for the day's subject
3. Practical application
4. Correction
5. Debrief

Philosophy

In this first course, we focus **only** on the preparation of data. Machine learning algorithms may be used, but will be explained in the dedicated course.



hadrien.salem@centralelille.fr



[introduction-to-data-science](#)

Exploratory data analysis

Introduction

Exploratory data analysis

**Learning to know your data
is always the first step**



What are we
trying to learn ?



What are we trying to learn ?

General questions (observe and count)

- What data is contained in the dataset?
- How is this data represented?
- What is the type of each feature?
- Are there “holes” in the data?
- Are there duplicates in the data?
- Is there imbalance in the data?

Advanced questions (understand)

- What is the statistical distribution of this data?
- Are some features correlated?
- If there are, which ones and why?

⇒ The more you explore, the more questions you will find, and the more specific the questions will be

Exploratory data analysis

Practical application

What languages for data analysis?

Python and R are the most common, but there are many more (e.g. Kotlin, Java, etc.).

These languages offer many packages to analyse and model your data.

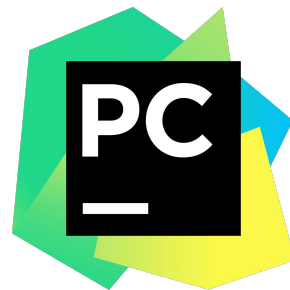


We will be using Python

What software for data analysis?

We will be using jupyter notebooks to run code and visualize results.

colab
kaggle



Which packages for data analysis?

Different libraries cover different aspects of data science.



Mathématiques



Manipulation de datasets



**Machine Learning
(hors Deep learning)**



Affichages



Course material

The screenshot shows the GitHub interface for the repository 'SnowHawkeye / introduction-to-data-science'. The repository is public and has 1 branch and 0 tags. A warning icon is visible on the left. The 'main' branch is selected in the dropdown menu. The 'Clone' button is highlighted with a red box, and the 'Download ZIP' button is also highlighted with a red box. A red arrow points to the 'Code' button. The repository name is 'introduction-to-data-science' and the owner is 'SnowHawkeye'. The repository description is 'Introduction course to data science'. The repository has 0 stars, 1 watching, and 0 forks. The repository has no releases published and no packages published.

<https://github.com/SnowHawkeye/introduction-to-data-science>

Opening the notebook

It can be imported in any IDE.

Google colab is an easy way to get started if you do not have a local IDE with Python set up.



Pros

- Easy installation, does not depend on your machine
- Use online resources (including free GPU time)
- Can share with other team members
- Integrations with Drive and GitHub

Cons

- Not as much control / customization as local install
- Managing extra files is cumbersome
- No concurrent writing with team members

Practical work

The notebook contains all the necessary instructions

Data visualization

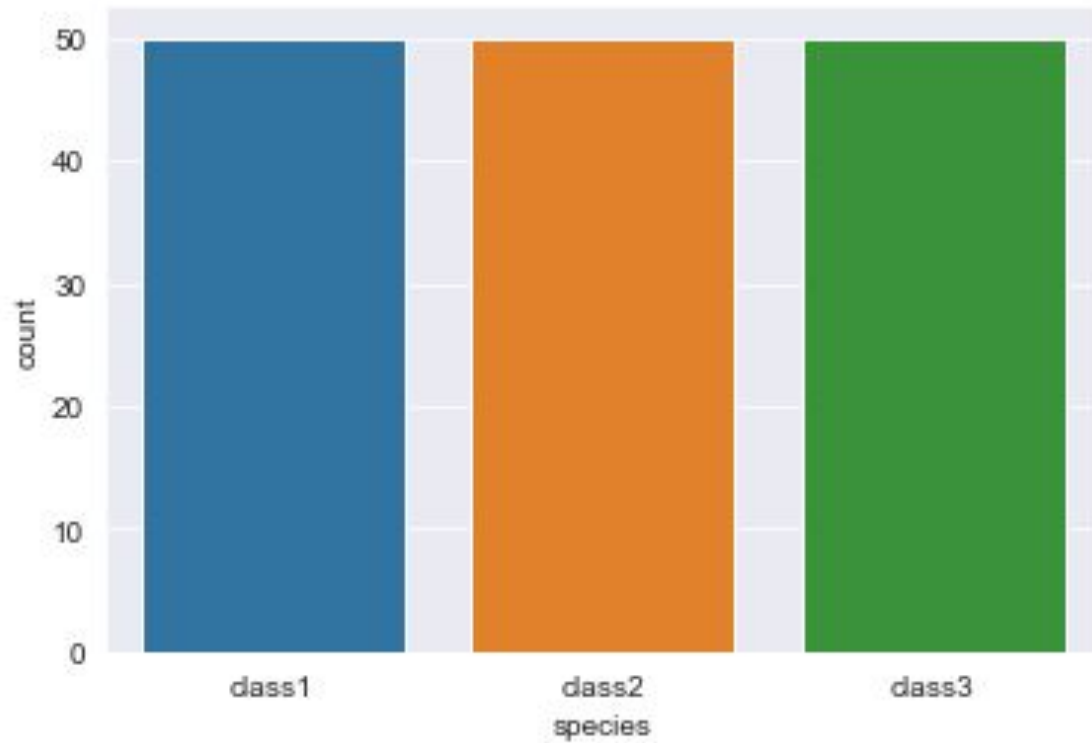
Why do we want to
visualize our data?



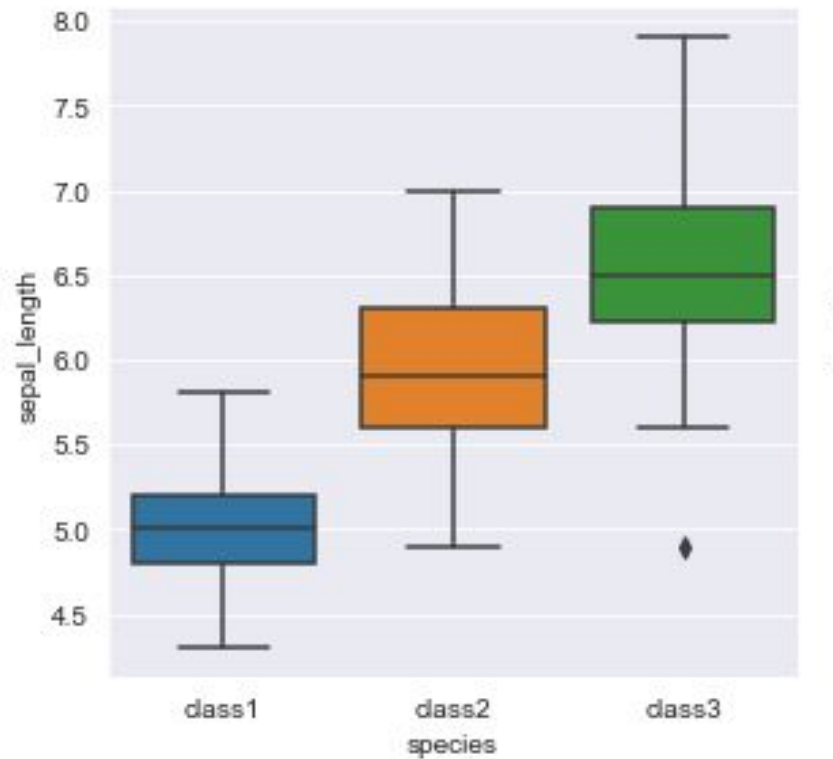
Why do we want to visualize our data?

The benefits of data visualization

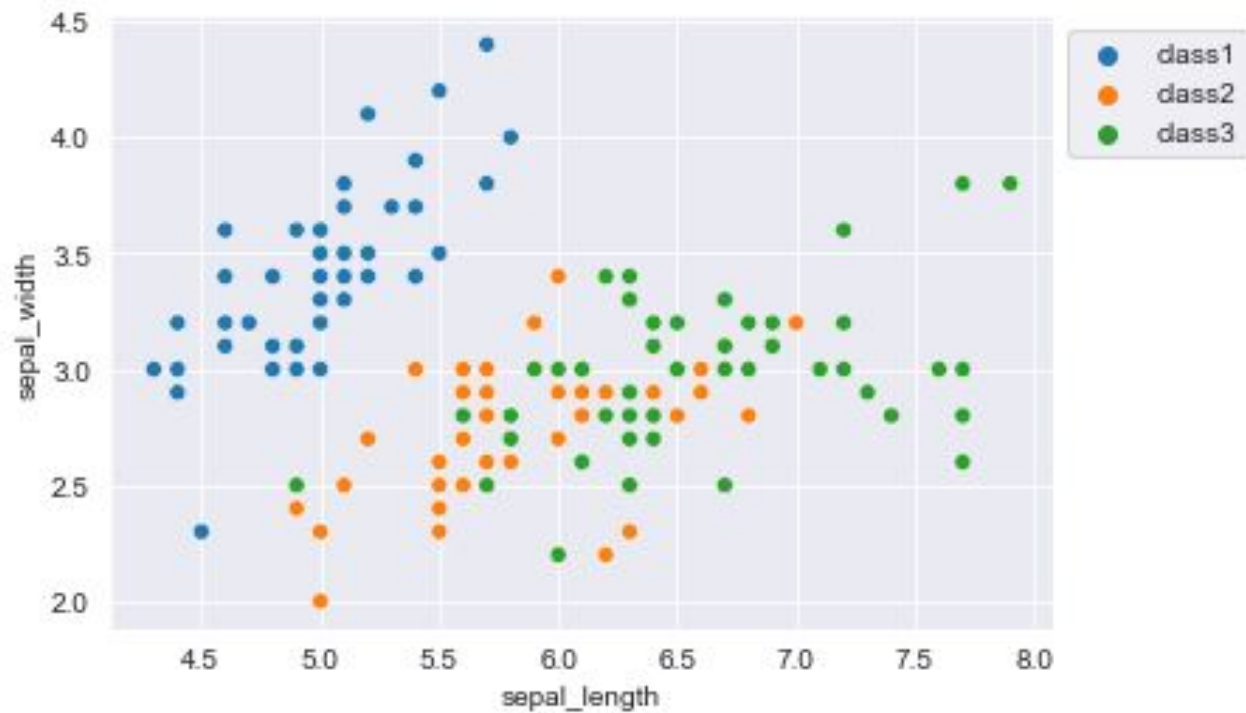
- Visualization helps **understanding the data**: detect outliers, understand the distribution of a variable, the number of elements in a class, feature correlation, feature importance, etc.
- It can help you **choose an algorithm** (in particular if your data is linearly separable)
- Graphs are essential for **communication**, in particular with non-technicians



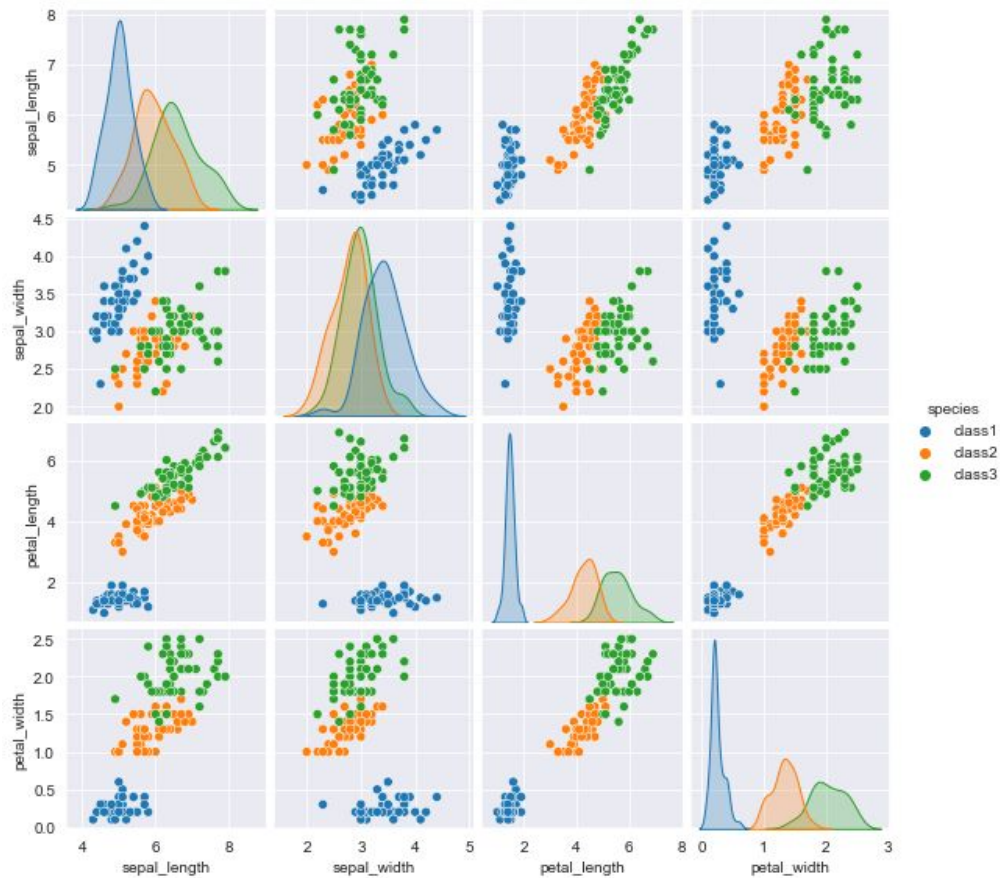
Graphs to count your data: `countplots`



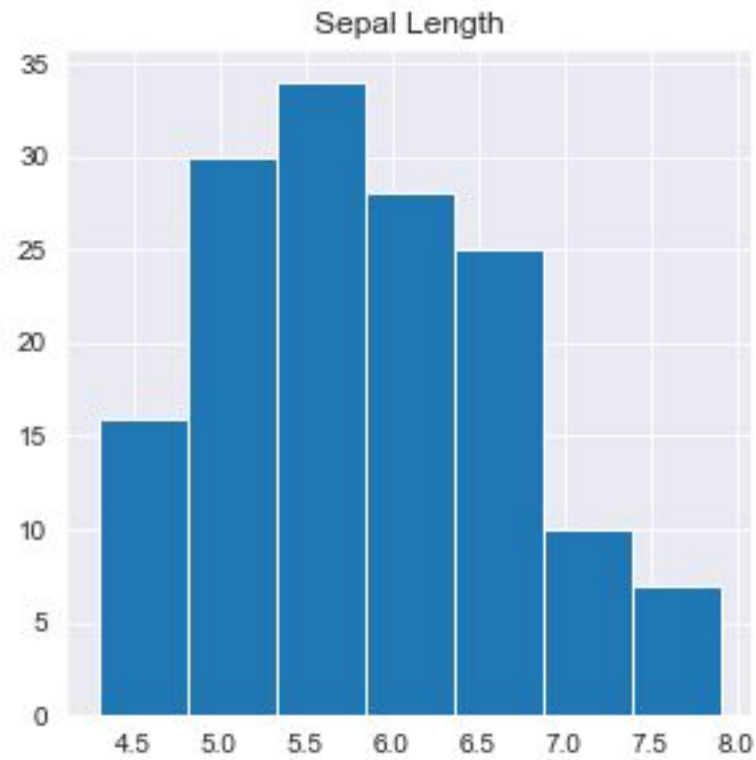
Graphs to understand the data distribution: `boxplots`



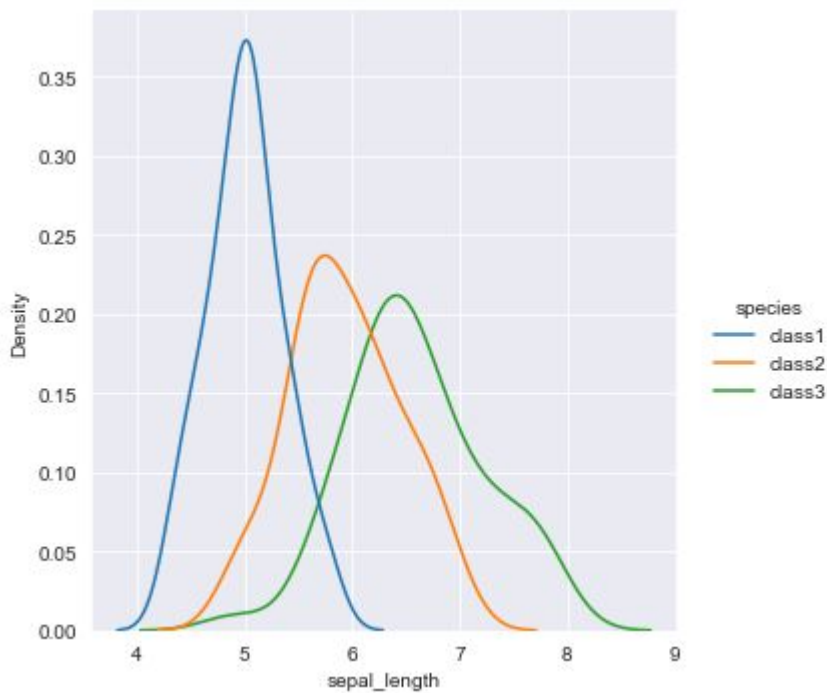
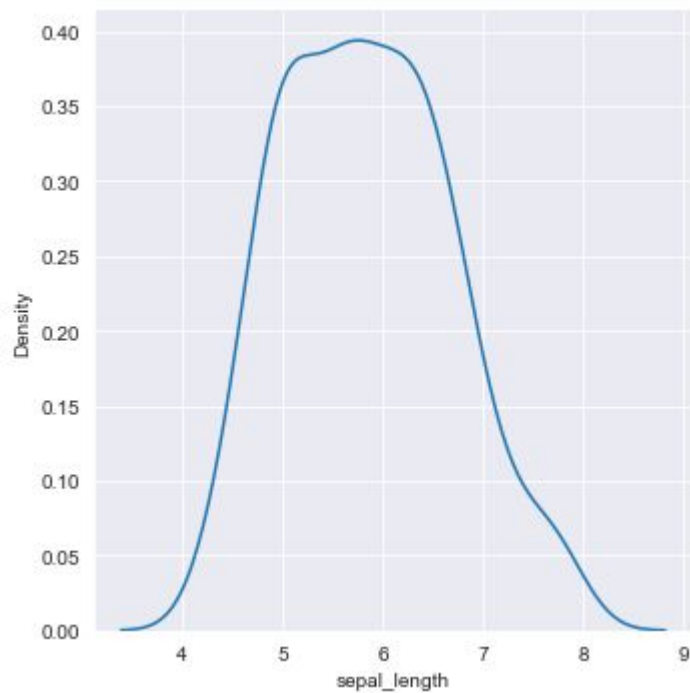
Graphs to visualize your features: `scatterplots`



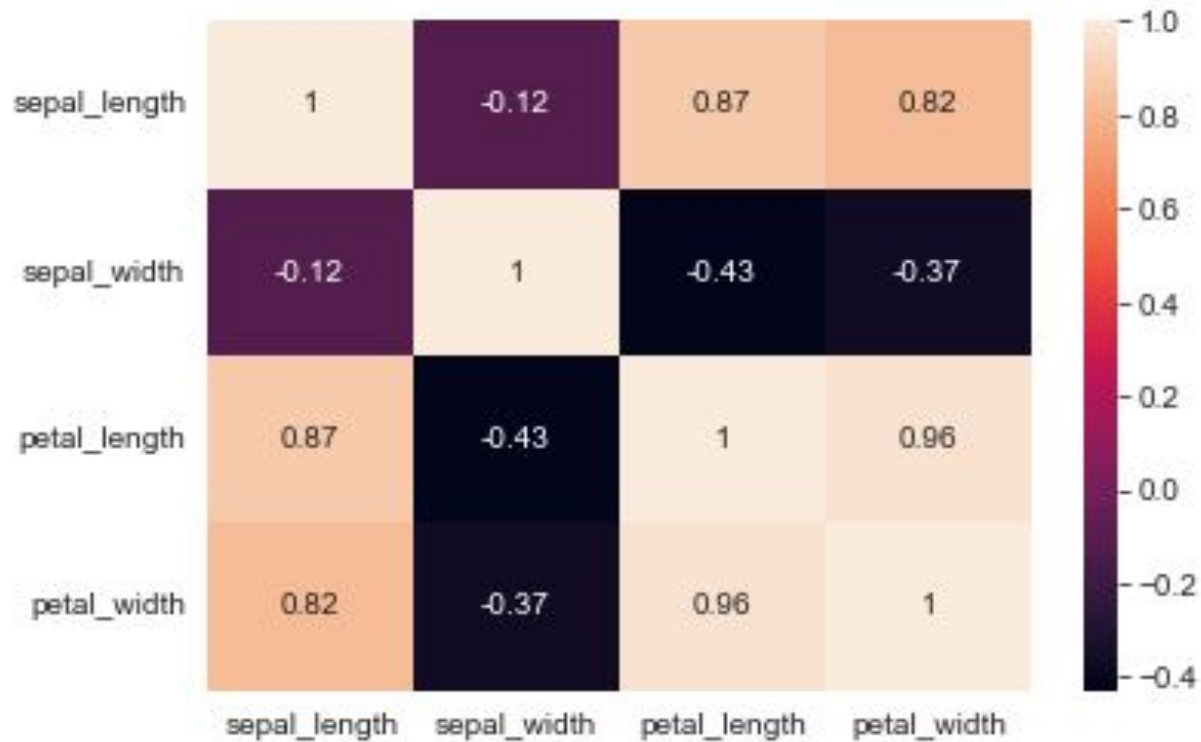
Graphs to visualize your features: `pairplot`



Graphs to understand the distribution of your features: `histogram`



Graphs to understand the distribution of your features: `displot`



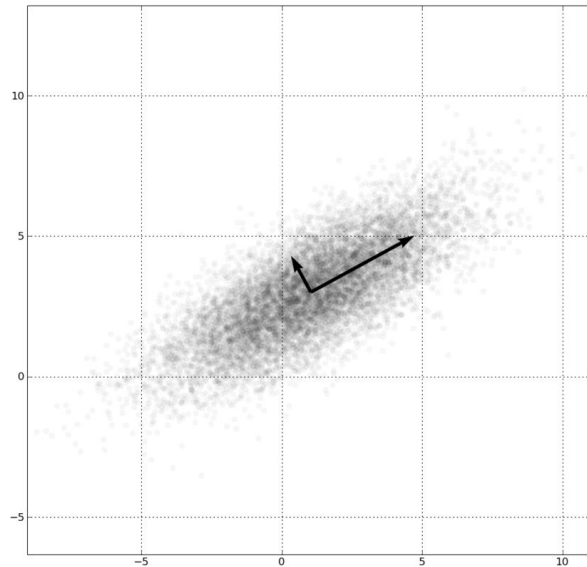
Graphs to understand relations within your features: `heatmap`

$$\text{Cov}(X, Y) \equiv \mathbb{E}[(X - \mathbb{E}[X]) (Y - \mathbb{E}[Y])]$$

Covariance of two random variables

Quantifies to what extent a change in one variable implies a change in the other variable.

In machine learning, we tend to like high (co)variance (high amount of information)



To represent with a heatmap: covariance

$$r = \frac{\sum (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Pearson's correlation coefficient

Quantifies to what extent the variables evolve similarly

To represent with a `heatmap`: correlation

Debrief

Debrief

What did we learn today?

What could we have done better?

What are we doing next time?

Data Science

Session 1 - Understanding data



hadrien.salem@centralelille.fr



[introduction-to-data-science](#)