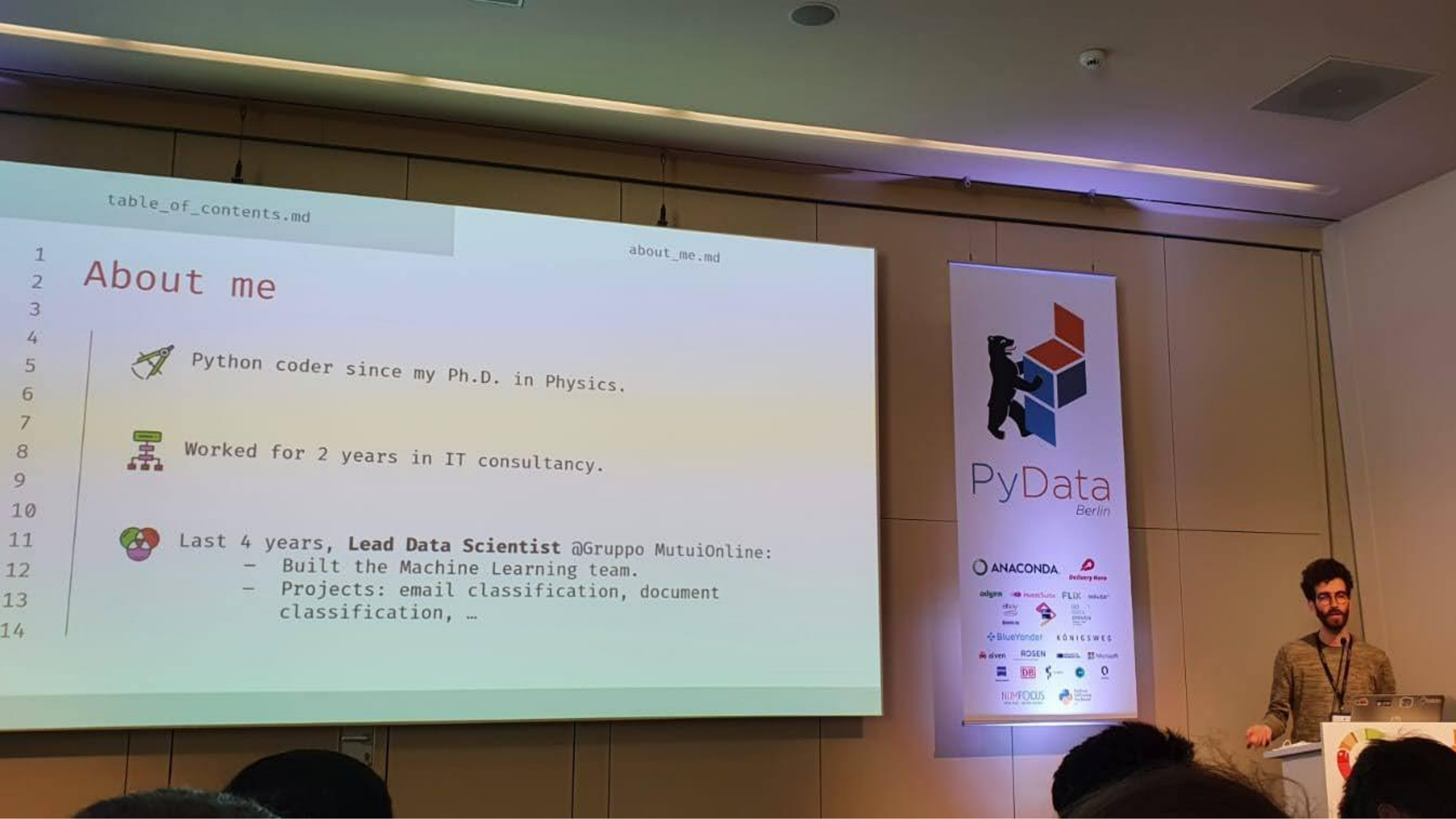
# Machine Learning & Data Science

ImpactDeal 2022



table_of_contents.md

about_me.md

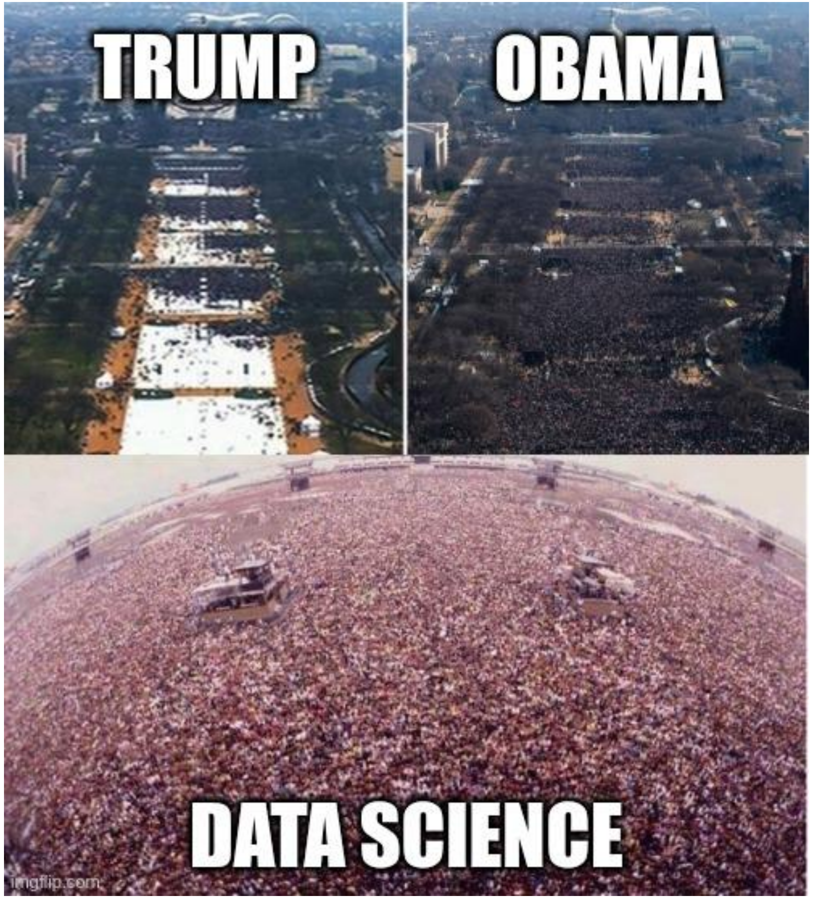
# About me

Python coder since my Ph.D. in Physics.

Worked for 2 years in IT consultancy.

Last 4 years, **Lead Data Scientist** @Gruppo MutuiOnline:

- Built the Machine Learning team.
- Projects: email classification, document classification, …


TRUMP
OBAMA
DATA SCIENCE
imgflip.com

# Yet another online course...?

"How was the online class? What did you get from it?"

Me:

- Books,
- Papers,
- Bootcamps, academies, masters,
- Online courses,
- Online platforms,
- Conferences,
- Videos,
- Blogs,
- More blogs,
- …

# But we can do something different!

- **Hands on**: we will see and write a lot of code.
- **Collaborative**: we will work together and/or in groups.

This course is designed around the idea of participation:

- Ask questions!
- Give feedback!
- Turn your camera on!
- Communicate your ideas!

# Learning Objectives

### Knowledge

- Structure of a data project
- Tools for data science in Python
- Fundamentals of Machine Learning

### Skills

- Techniques for data exploration
- Training ML models
- Ability to deal with complex data

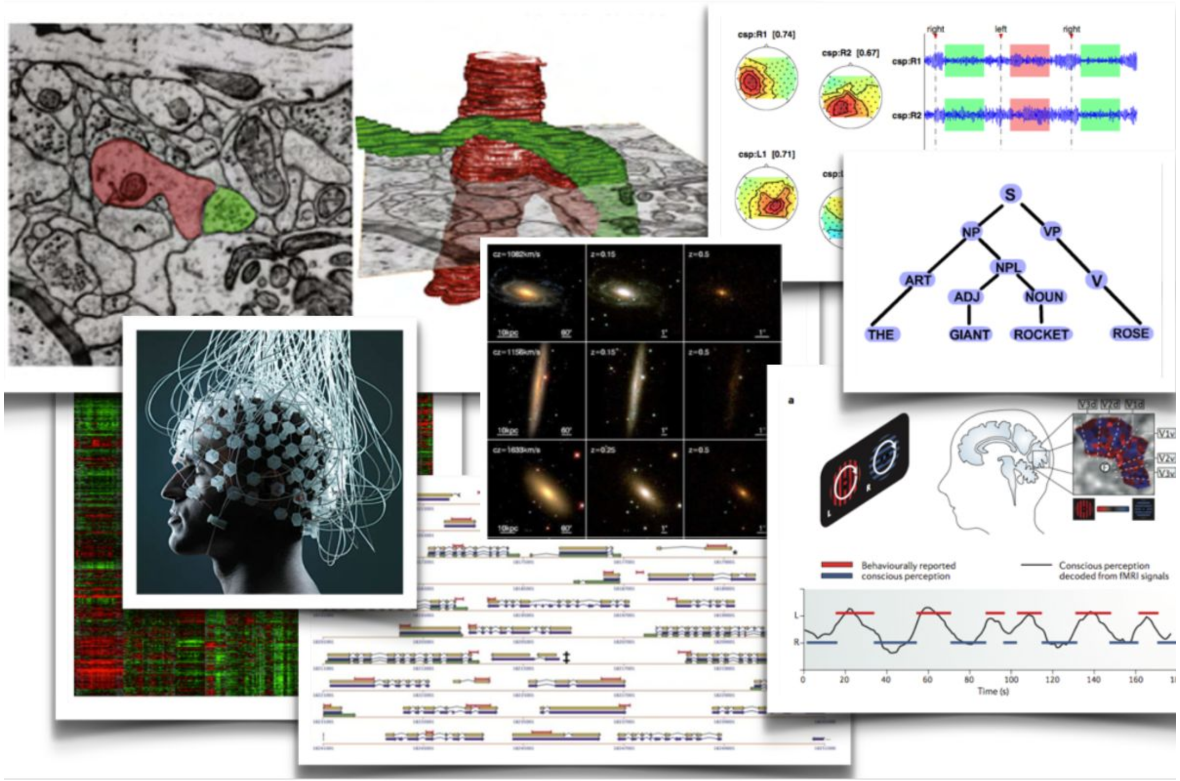
### Attitude

- Critical thinking about data
- Creativity with data analytics

# Learning Tools

## Slides (PDF):

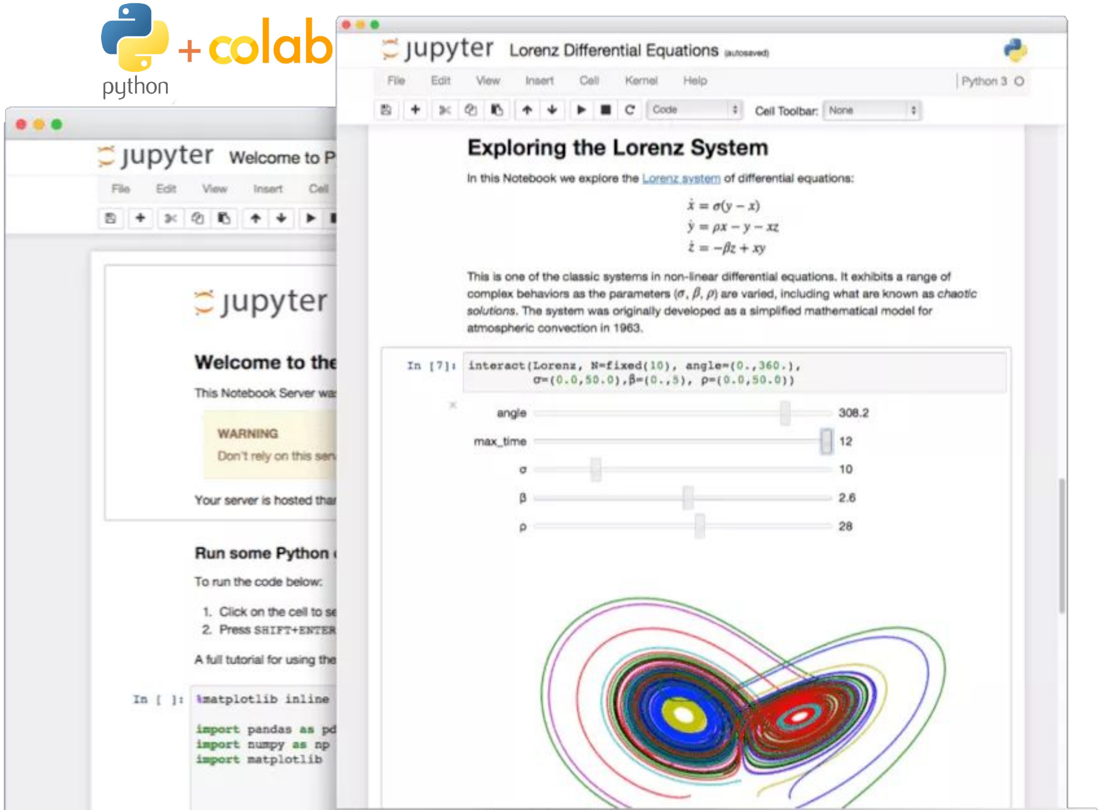
- Introduction to the topics,
- Theoretical concepts,
- No-code examples.



# Learning Tools

## Jupyter notebooks (Colab):

- Example code,
- Exercises,



Welcome to the

This Notebook Server wa

WARNING

Don't rely on this ser

Your server is hosted tha

Run some Python

To run the code below:

1. Click on the cell to se
2. Press SHIFT+ENTER

A full tutorial for using the

```
In [ ]: %matplotlib inline

        import pandas as pd
        import numpy as np
        import matplotlib
```

Exploring the Lorenz System

In this Notebook we explore the Lorenz system of differential equations:

$$\dot{x} = \sigma(y - x)$$
$$\dot{y} = \rho x - y - xz$$
$$\dot{z} = -\beta z + xy$$

This is one of the classic systems in non-linear differential equations. It exhibits a range of complex behaviors as the parameters ($\sigma$, $\beta$, $\rho$) are varied, including what are known as *chaotic solutions*. The system was originally developed as a simplified mathematical model for atmospheric convection in 1963.

```
In [7]: interact(Lorenz, N=fixed(10), angle=(0.,360.),
                 σ=(0.0,50.0),β=(0.,5), ρ=(0.0,50.0))
```

angle 308.2

max_time 12

σ 10

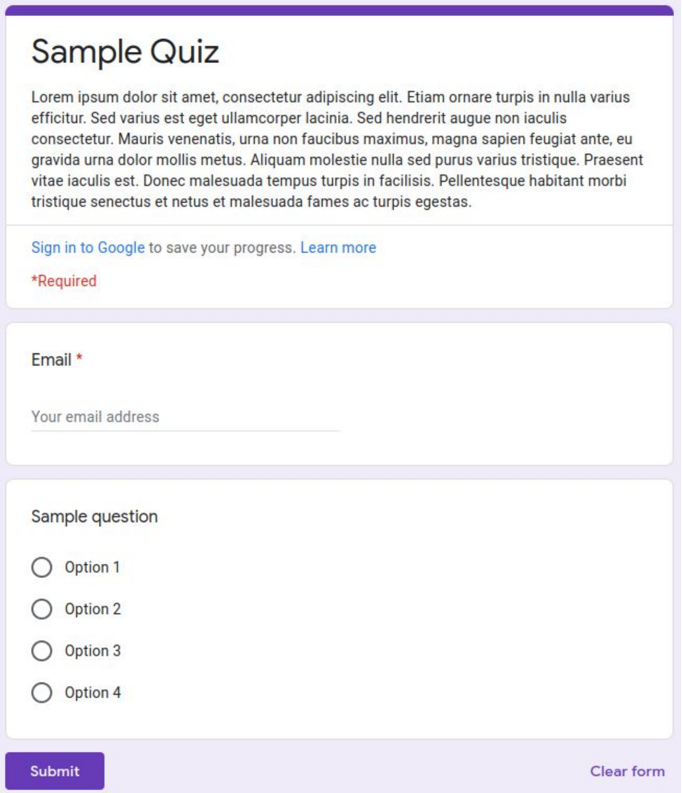
β 2.6

ρ 28

# Learning Tools

## Quizzes (Google Forms):

- Short and simple questions,
- Useful to self-assess learning path,
- Helpful for Q&A session,
- No grades.



Sample Quiz

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam ornare turpis in nulla varius efficitur. Sed varius est eget ullamcorper lacinia. Sed hendrerit augue non iaculis consectetur. Mauris venenatis, urna non faucibus maximus, magna sapien feugiat ante, eu gravida urna dolor mollis metus. Aliquam molestie nulla sed purus varius tristique. Praesent vitae iaculis est. Donec malesuada tempus turpis in facilisis. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Sign in to Google to save your progress. Learn more

*Required

Email *

Your email address

Sample question

○ Option 1

○ Option 2

○ Option 3

○ Option 4

# Learning Tools

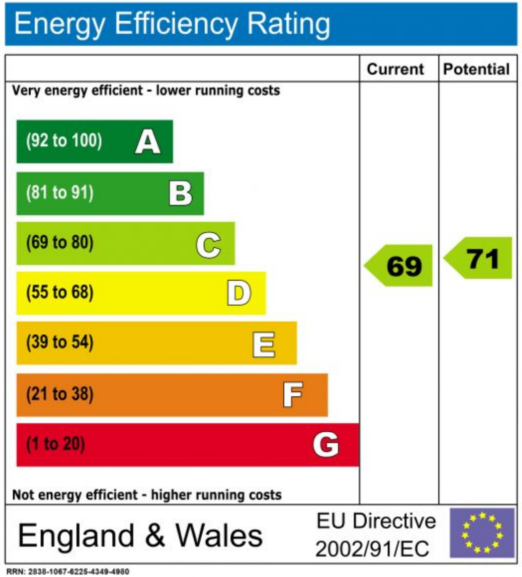
**Collaboration:**

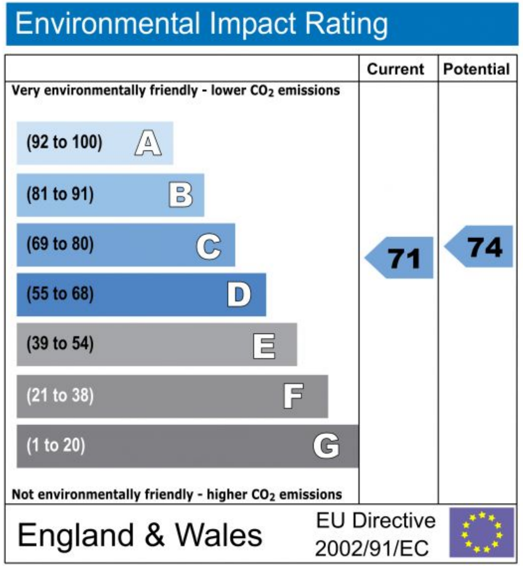
- GitHub (https://github.com/darioka/impactdeal-2022),
- Teams,
- Telegram.

# Project

## Estimation of Building Energy Efficiency

We will try to build a machine learning model able to predict the Energy Efficiency Rating (EER) of a dwelling, using historical data of Energy Performance Certificates of England and Wales.





# Project

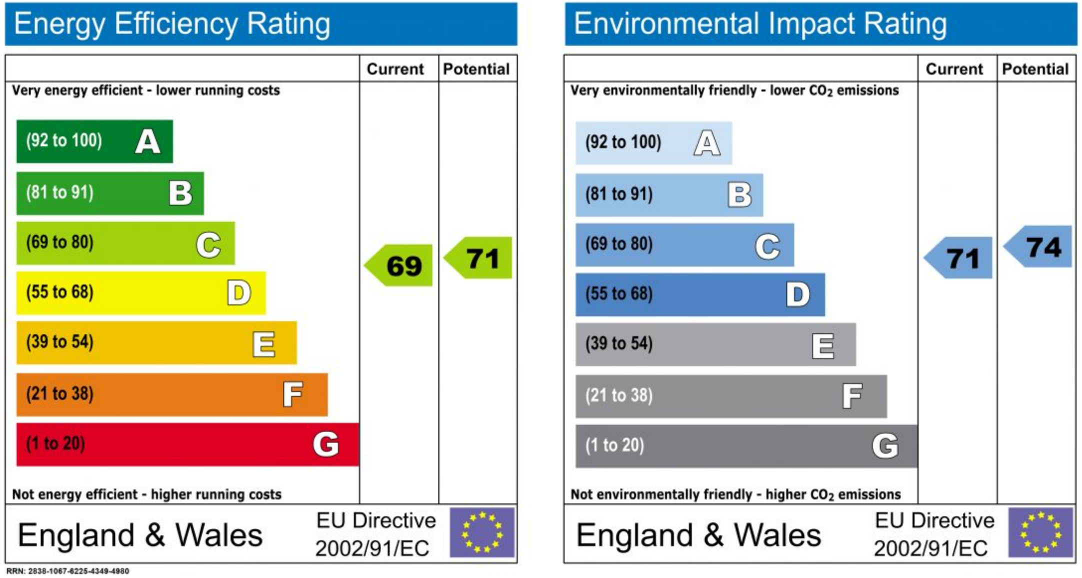
## Estimation of Building Energy Efficiency

**What is a EPC rating?**

EPC is a review of the energy efficiency of a property, which is labelled from A (very efficient) to G (inefficient). EPC are valid for 10 years and are needed whenever a property is sold or rented.

**How is EPC rating calculated?**

A trained professional conducts an inspection of the property and assesses the energy efficiency of walls, windows, heating and water systems, etc.

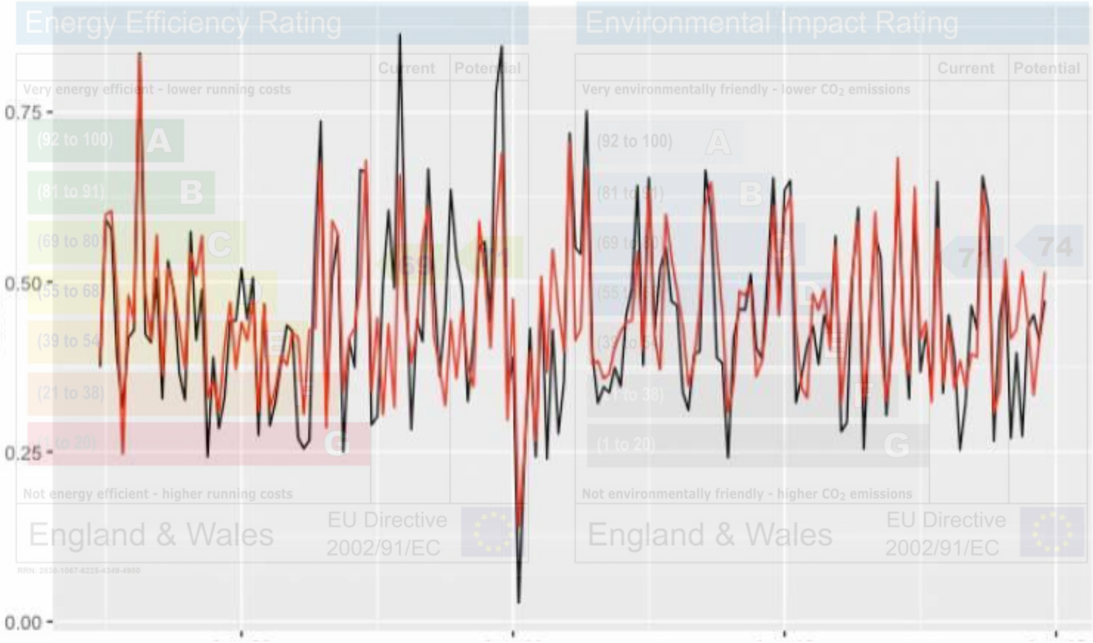


# Project

## Estimation of Building Energy Efficiency

**What is the goal of this project?**
We want to build a machine learning model to predict the EPC band of a property. The model will be trained on historical EPC data, available on opendatacommunities.org.

**How do we do it?**
We will follow the project throughout the course, applying the techniques we will be learning and discussing their implications on the EPC prediction problem. Analyses and training will be performed on EPC data from three major UK cities (*) and can be done entirely on Colab notebooks.



(*) the data has been downloaded, subsampled, pseudonymized (address and postcode) and uploaded to the course's Github repository.

# Data Science Fundamentals

1. **Data Science Fundamentals**
    1. What is Data Science?
    2. What is Machine Learning?
    3. Machine Learning Approaches
2. Machine Learning Projects
    1. Project workflow
    2. Data Exploration
    3. Modeling
    4. Deploy
3. Python for Machine Learning
    1. Why python?

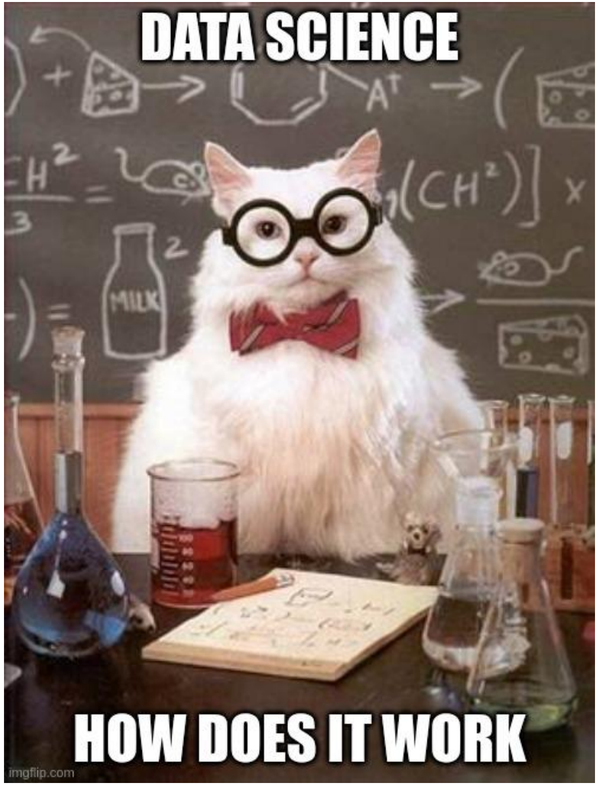
# What is Data Science?

Not very well defined...

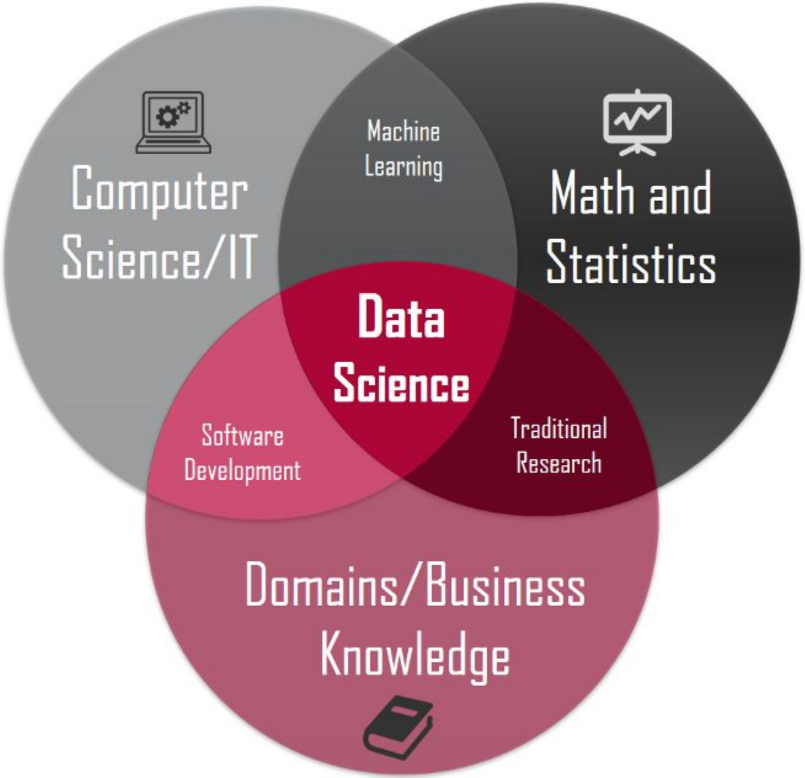
- ***Circular***:
  - *Data science is what data scientists do*
- ***General***:
  - *Data science is the science of learning from data*

Very broad field, but:

- it has to do with **science**,
- it has to do with **data**.



# What is Data Science?

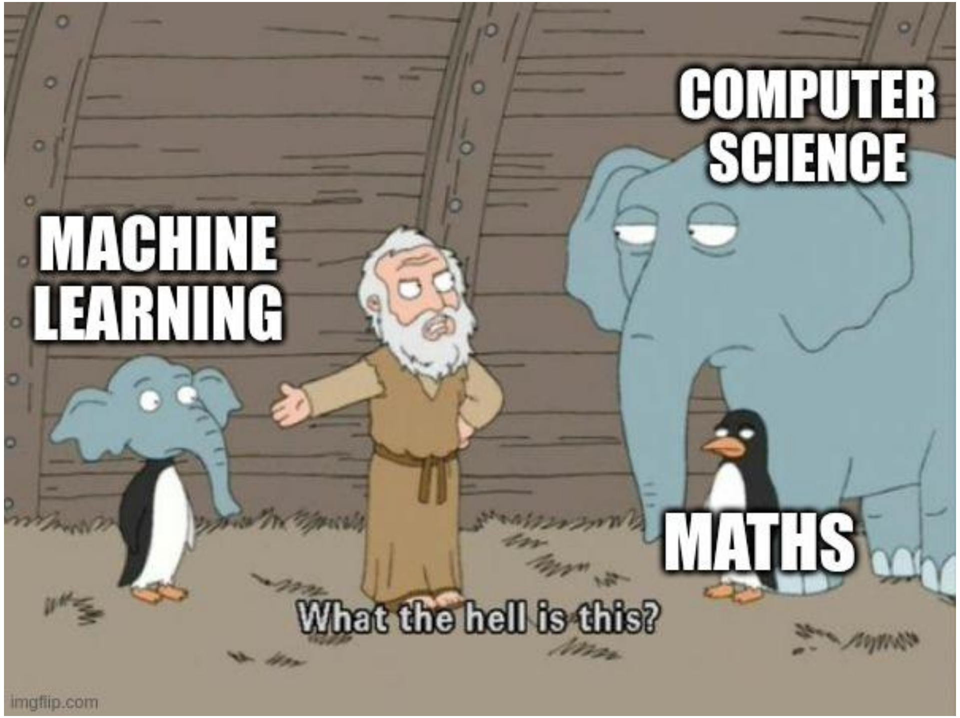


# What is Machine Learning?

Better definitions:

***The study of computer algorithms that can "learn" from data to solve tasks, without being explicitly programmed to do so.***

Machine learning algorithms are based on **training data** and are often formulated as **minimization** of some **loss function** i.e. as **optimization** problems.



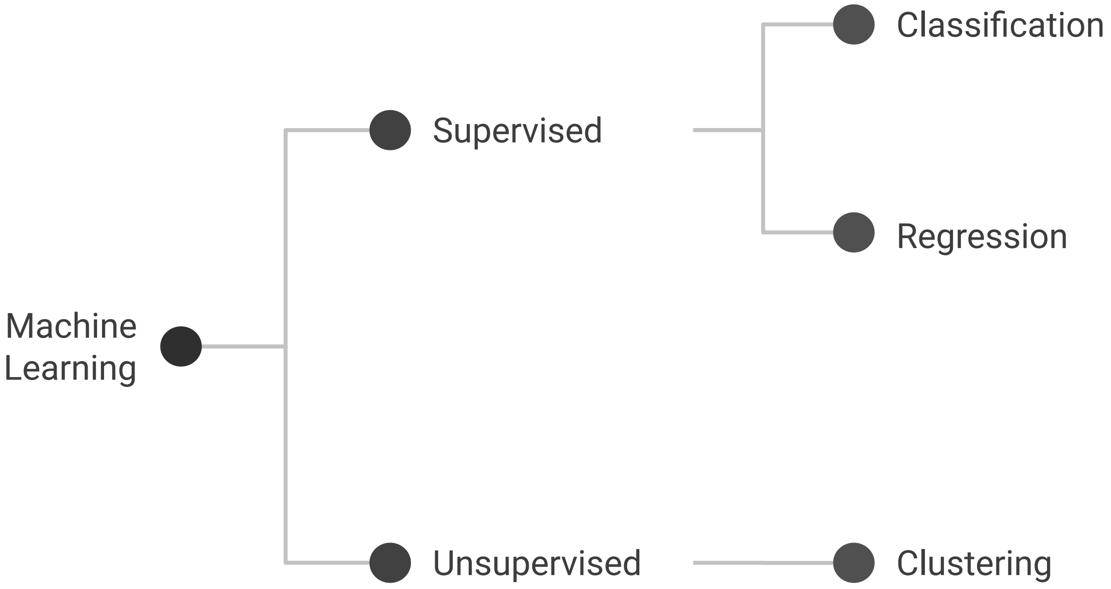
# What is Machine Learning?

Examples:

- **Recommenders**:
  - Not explicitly programmed to suggest object A if object B has been chosen. Not a set of rules.
  - Try to predict users’ preferences, based on historical data.

- **Image recognition**:
  - Does not require manual feature extraction or expert computer vision knowledge,
  - Learn properties and relationships of pixels in images.

# Machine Learning Approaches



Machine Learning

Supervised

Classification

Regression

Unsupervised

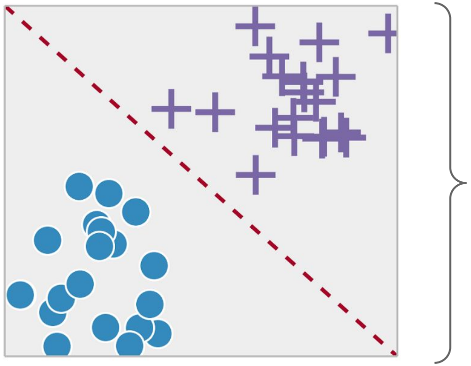
Clustering

# Machine Learning Approaches

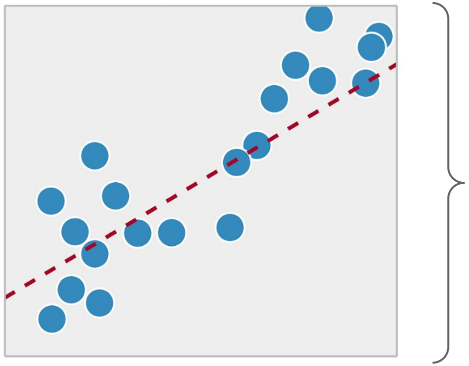
## Supervised

- **Predict the value of the target** for each sample,
- Models requires both **features** and **target** in the training data (i.e. labeled data).

- **Classification**:
  - Target is a discrete variable,
  - Model example: logisitc regression,
  - Application example: fraud detection.
- **Regression**:
  - Target is a continuous variable,
  - Model example: linear regression,
  - Application example: demand forecasting



*Classification*



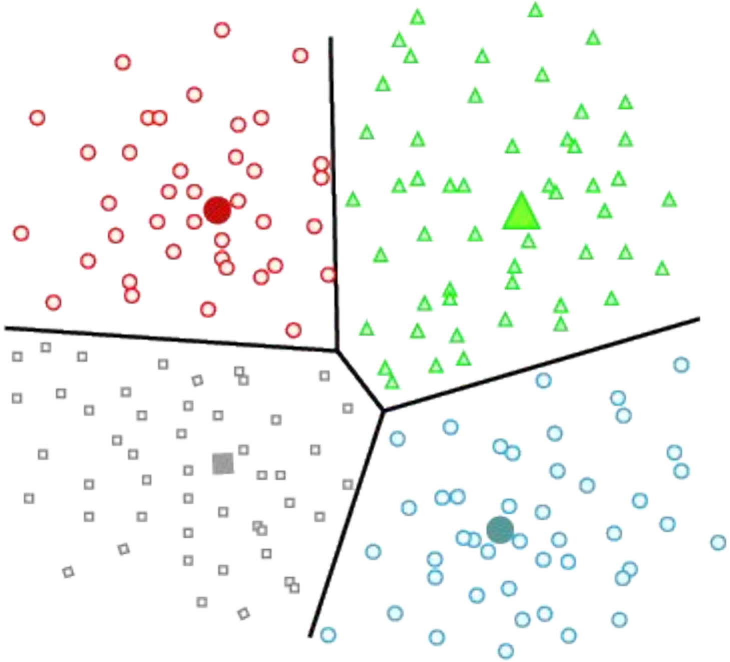
*Regression*

# Machine Learning Approaches

## Unsupervised

- **Find patterns in the data**,
- There is not “target” value to predict or the target is absent from training data.

- **Clustering**:
  - Divide input samples into groups,
  - Model example: K-means,
  - Application example: customer segmentation



# Machine Learning Approaches

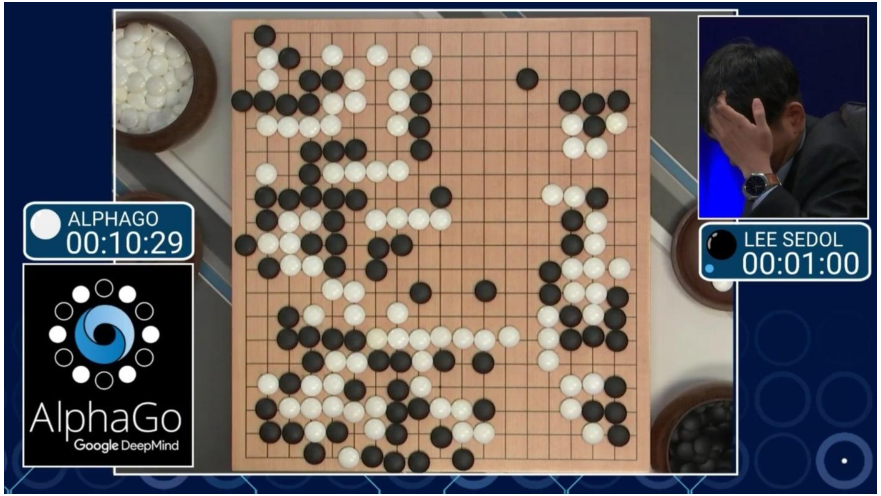
There are techniques and/or tasks that may be either supervised, unsupervised or both:

- **Semi-supervised learning**: training dataset has both labeled and unlabeled samples.
- **Dimensionality reduction** and **manifold learning**: transformations of data from a high-dimensional space to a low-dimensional representation.
- **Association rule learning**: methods for discovering relationships and strong rules between variables in large databases.

# Machine Learning Approaches

## Reinforcement Learning

- **Agents** interacting and taking **actions** in an environment.
- Solve a task trying to find an optimal strategy that maximizes a **reward** for the agent.
- Different approach: no need of large training data.
- Usually much harder than supervised/unsupervised. Active research area.
- Examples: AlphaGo

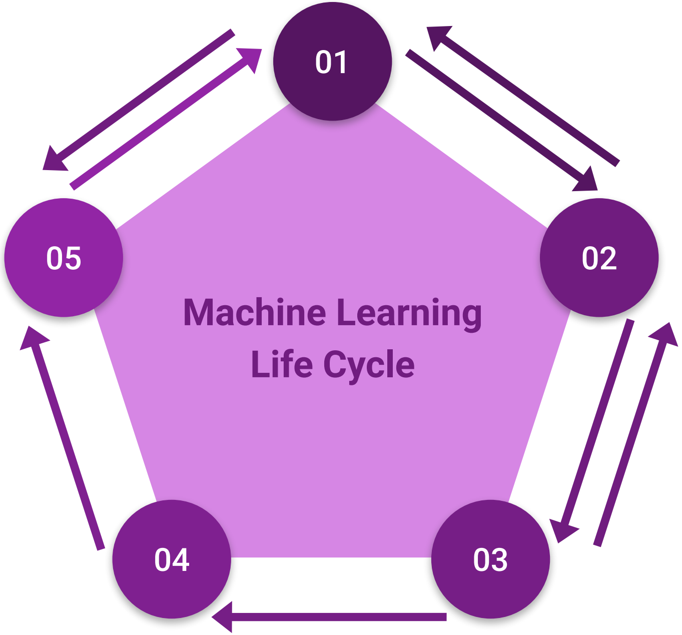


# Machine Learning Projects

1. Data Science Fundamentals
   a. What is Data Science?
   b. What is Machine Learning?
   c. Machine Learning Approaches
2. **Machine Learning Projects**
   a. Project workflow
   b. Data Exploration
   c. Modeling
   d. Deploy
3. Python for Machine Learning
   a. Why python?

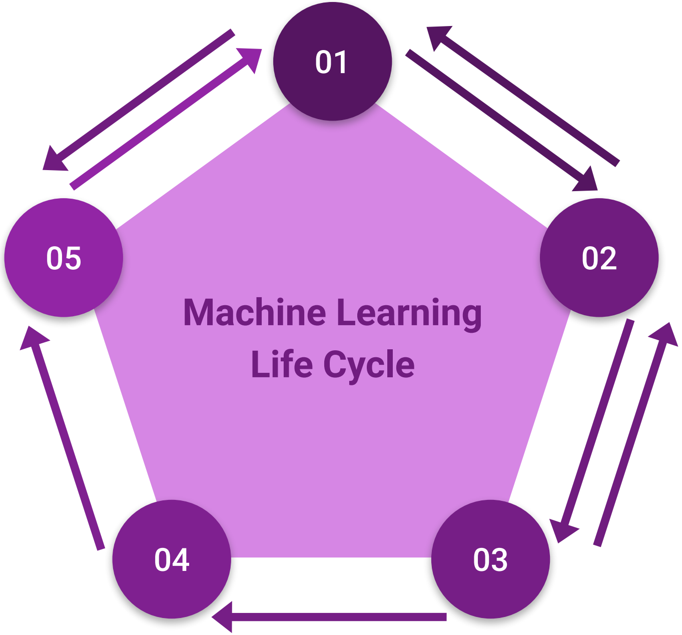
# Project Workflow

1. Business Understanding
2. Data Preparation
3. Modeling
4. Deployment
5. Management



# Project Workflow

1. **Business Understanding**
   - ○ Problem definition
   - ○ Objectives
   - ○ Expected outcomes
   - ○ Assessments (risk, infrastructure, ...)
   - ○ Understand integration/deploy
2. Data Preparation
3. Modeling
4. Deployment
5. Management



# Project Workflow

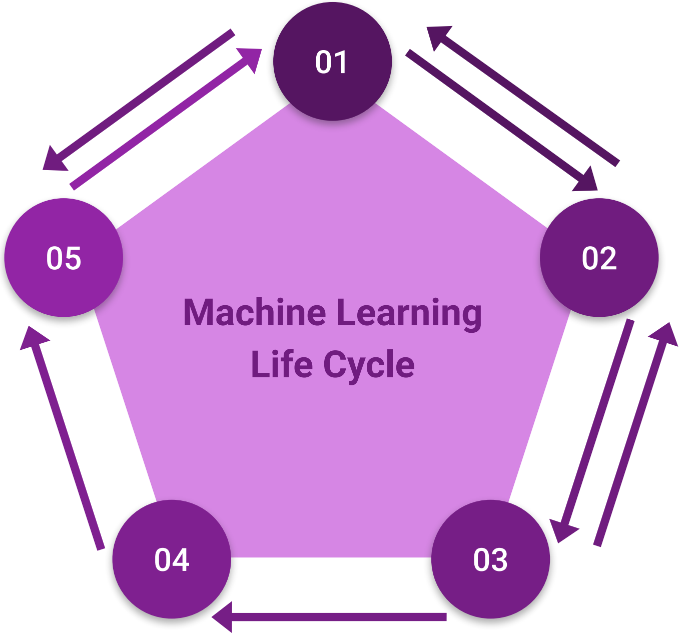
1. Business

2. **Data**
    - Data gathering
    - Data exploration
    - Data cleaning/preprocessing
    - Feature engineering

3. Modeling

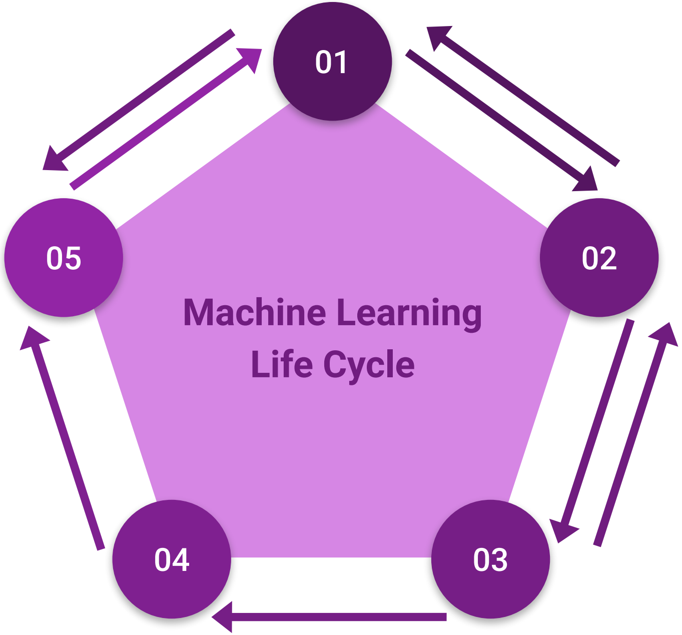
4. Deployment

5. Management



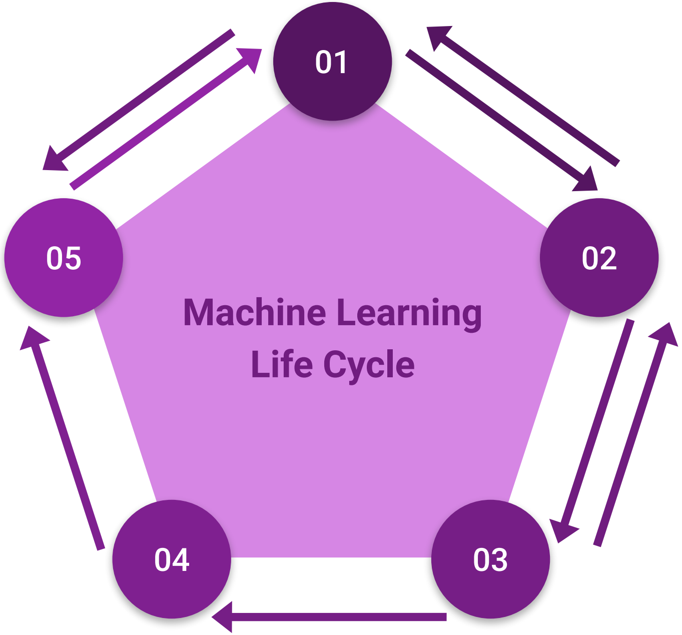
# Project Workflow

1. Business Understanding
2. Data Preparation
3. **Modeling**
   - ○ Design experiments
   - ○ Train models
   - ○ Evaluate and test models
   - ○ Interpret results
   - ○ Review and cross-check
4. Deployment
5. Management



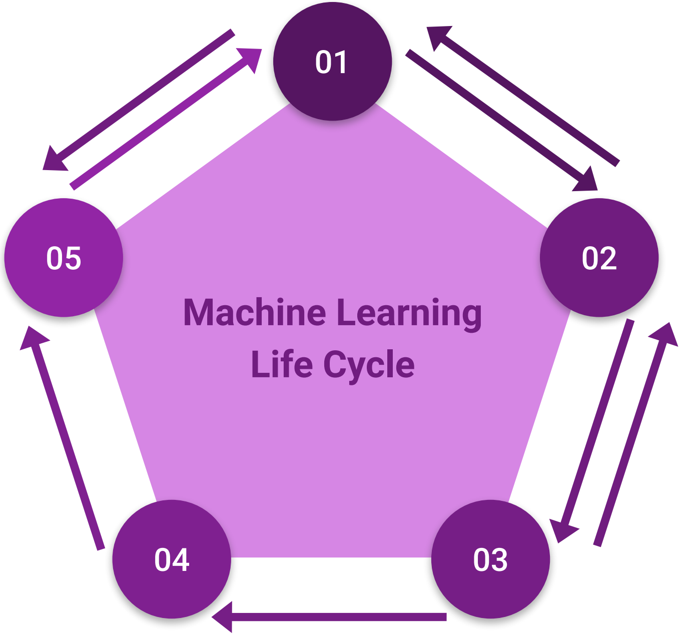
# Project Workflow

1. Business Understanding
2. Data Preparation
3. Modeling
4. **Deployment**
    - ○ Create model artifacts
    - ○ Write production-ready code
    - ○ Integration
    - ○ Plan maintenance
5. Management

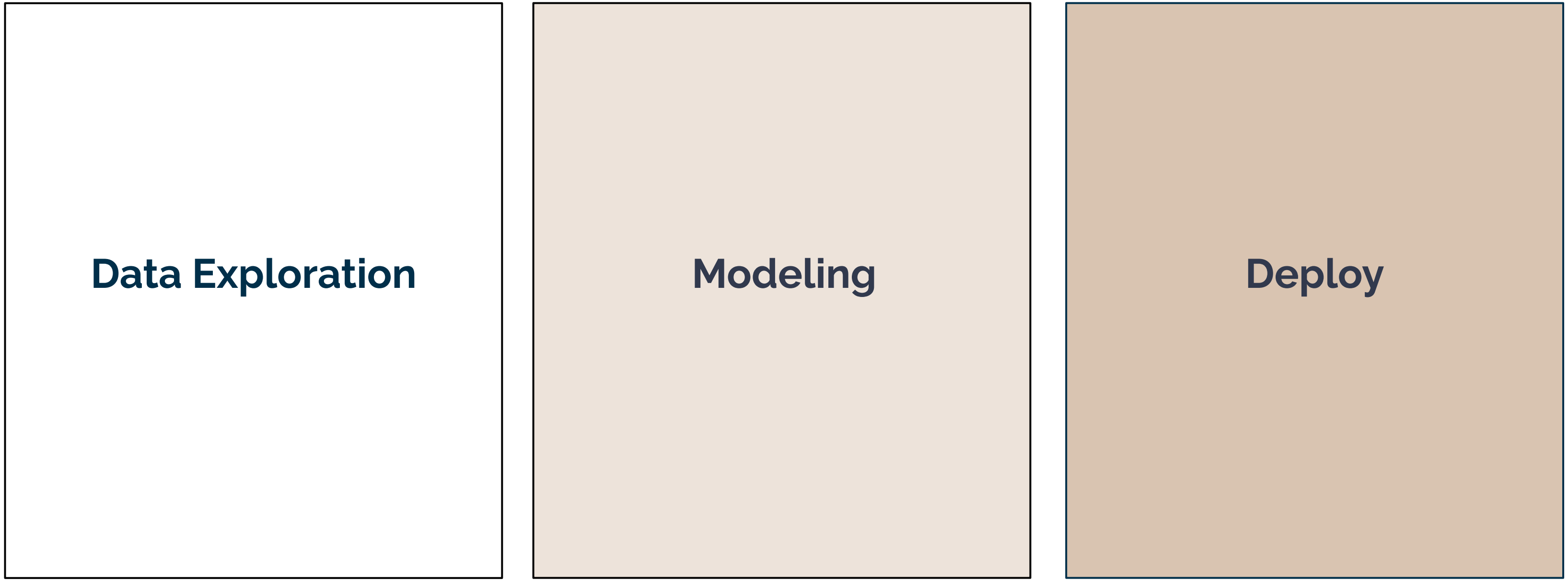


# Project Workflow

1. Business Understanding
2. Data Preparation
3. Modeling
4. Deployment
5. **Management**
   - ○ Monitoring input/output
   - ○ Monitoring performances
   - ○ Model updates



# Project Workflow



Data Exploration

Modeling

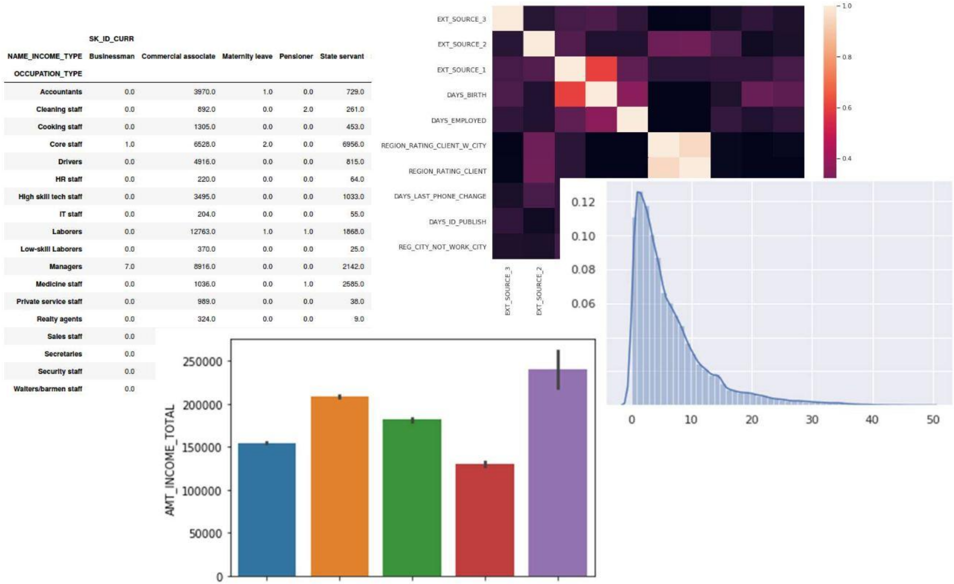
Deploy

# Data Exploration

It’s the process of analyzing a dataset to understand its characteristics.

It is a mixture of:

- summary statistics,
- data visualization,
- hypothesis testing,
- manual drill-down.

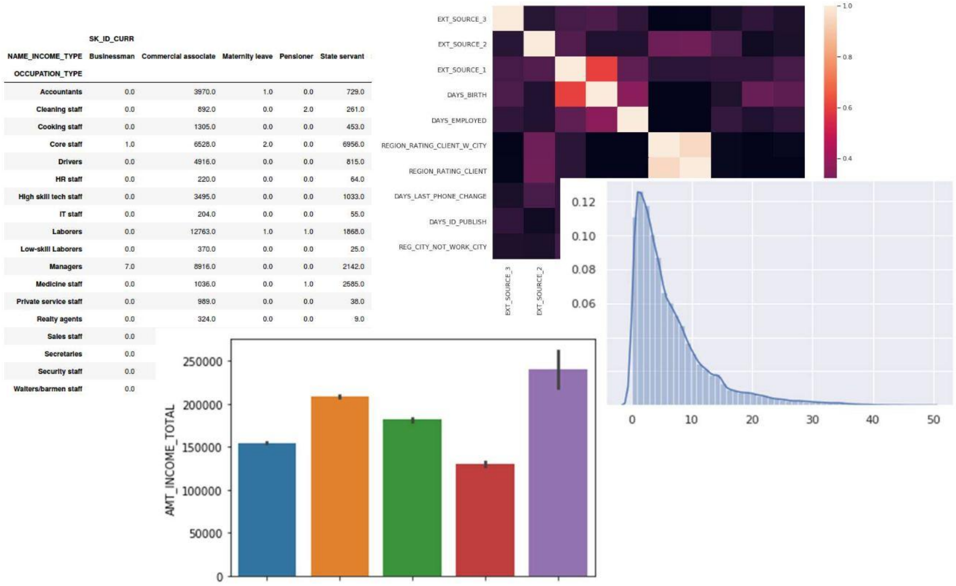


|  | SK_ID_CURR |  |  |  |  |
|---|---|---|---|---|---|
| NAME_INCOME_TYPE | Businessman | Commercial associate | Maternity leave | Pensioner | State servant |
| OCCUPATION_TYPE |  |  |  |  |  |
| Accountants | 0.0 | 3970.0 | 1.0 | 0.0 | 729.0 |
| Cleaning staff | 0.0 | 892.0 | 0.0 | 2.0 | 261.0 |
| Cooking staff | 0.0 | 1305.0 | 0.0 | 0.0 | 453.0 |
| Core staff | 1.0 | 6528.0 | 2.0 | 0.0 | 6956.0 |
| Drivers | 0.0 | 4916.0 | 0.0 | 0.0 | 815.0 |
| HR staff | 0.0 | 220.0 | 0.0 | 0.0 | 64.0 |
| High skill tech staff | 0.0 | 3495.0 | 0.0 | 0.0 | 1033.0 |
| IT staff | 0.0 | 204.0 | 0.0 | 0.0 | 55.0 |
| Laborers | 0.0 | 12763.0 | 1.0 | 1.0 | 1868.0 |
| Low-skill Laborers | 0.0 | 370.0 | 0.0 | 0.0 | 25.0 |
| Managers | 7.0 | 8916.0 | 0.0 | 0.0 | 2142.0 |
| Medicine staff | 0.0 | 1036.0 | 0.0 | 1.0 | 2585.0 |
| Private service staff | 0.0 | 989.0 | 0.0 | 0.0 | 38.0 |
| Realty agents | 0.0 | 324.0 | 0.0 | 0.0 | 9.0 |
| Sales staff | 0.0 |  |  |  |  |
| Secretaries | 0.0 |  |  |  |  |
| Security staff | 0.0 |  |  |  |  |
| Waiters/barmen staff | 0.0 |  |  |  |  |

# Data Exploration

Typical steps in data exploration are:

- Variable identification,
- Univariate analysis,
- Bivariate analysis,
- Missing data analysis.



| | SK_ID_CURR | | | | |
|---|---|---|---|---|---|
| NAME_INCOME_TYPE | Businessman | Commercial associate | Maternity leave | Pensioner | State servant |
| OCCUPATION_TYPE | | | | | |
| Accountants | 0.0 | 3970.0 | 1.0 | 0.0 | 729.0 |
| Cleaning staff | 0.0 | 892.0 | 0.0 | 2.0 | 261.0 |
| Cooking staff | 0.0 | 1305.0 | 0.0 | 0.0 | 453.0 |
| Core staff | 1.0 | 6528.0 | 2.0 | 0.0 | 6956.0 |
| Drivers | 0.0 | 4916.0 | 0.0 | 0.0 | 815.0 |
| HR staff | 0.0 | 220.0 | 0.0 | 0.0 | 64.0 |
| High skill tech staff | 0.0 | 3495.0 | 0.0 | 0.0 | 1033.0 |
| IT staff | 0.0 | 204.0 | 0.0 | 0.0 | 55.0 |
| Laborers | 0.0 | 12763.0 | 1.0 | 1.0 | 1868.0 |
| Low-skill Laborers | 0.0 | 370.0 | 0.0 | 0.0 | 25.0 |
| Managers | 7.0 | 8916.0 | 0.0 | 0.0 | 2142.0 |
| Medicine staff | 0.0 | 1036.0 | 0.0 | 1.0 | 2585.0 |
| Private service staff | 0.0 | 989.0 | 0.0 | 0.0 | 38.0 |
| Realty agents | 0.0 | 324.0 | 0.0 | 0.0 | 9.0 |
| Sales staff | 0.0 | | | | |
| Secretaries | 0.0 | | | | |
| Security staff | 0.0 | | | | |
| Waiters/barmen staff | 0.0 | | | | |

# Data Exploration

## Variable identification

*What kind of information does the dataset contain?*

- Identify the target (if any),
- Understand the meaning of the variables,
- Identify data types.

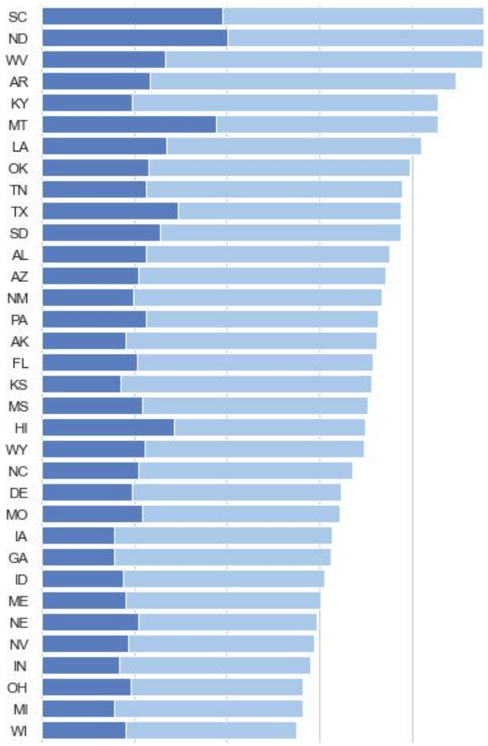
# Data Exploration

## Univariate analysis

*How the data is distributed?*

For categorical variables:

| Methods | Visualization |
|---|---|
| Counts<br>Frequencies | Bar plots |



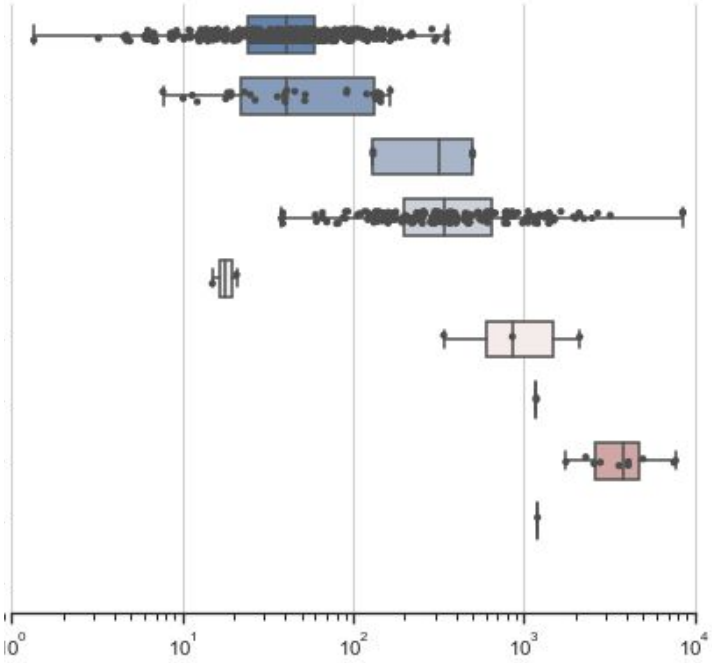
# Data Exploration

## Univariate analysis

*How the data is distributed?*

For numerical variables:

| Methods | Visualizations |
|---|---|
| Central tendency<br>(mean, median, …)<br><br>Dispersion<br>(range, quartiles, variance, …) | Histograms<br>Box Plots<br>Violin plots |

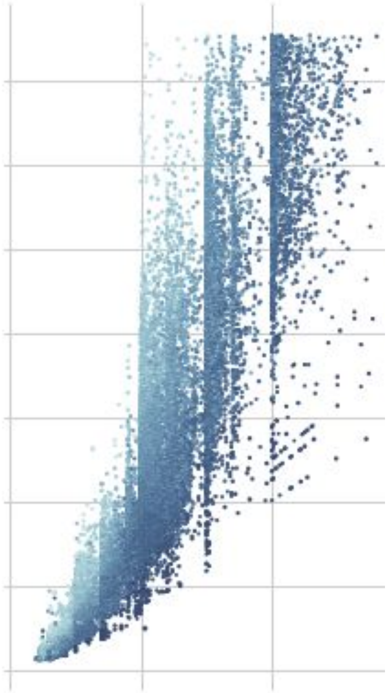


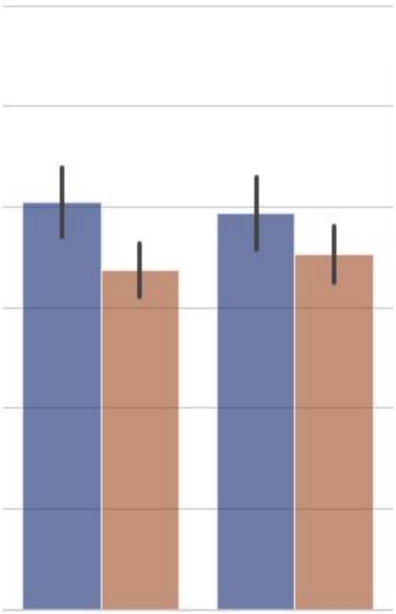
# Data Exploration

## Bivariate analysis

*What are the relationships between the variables?*

| Methods | Visualizations |
|---|---|
| Correlation | Scatter plots<br>Heatmap |
| Joint frequencies<br>Distribution difference between groups | Histograms<br>Box Plots<br>Violin plots |





# Data Exploration

## Missing Values

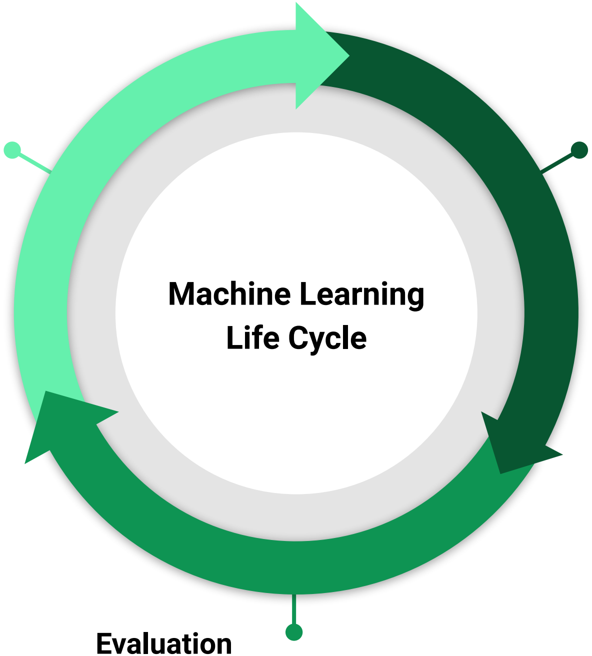
*Is the data complete?*

Common situation with real-world data sets, where some variables contains no information, for example because of errors during the data collection process.

Missing data can be a problem for any analysis:

- Loss of information and statistical power,
- If systematic, introduce bias and distortions.

# Modeling



**Machine Learning Life Cycle**

**Data Preparation**

- Cleaning
- Preprocessing
- Feature Engineering

**Model Training**

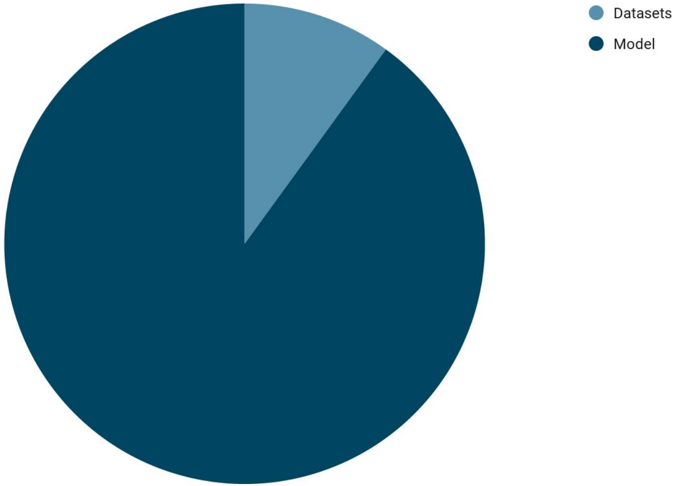
- Model fitting
- Hyperparameter tuning

**Evaluation**

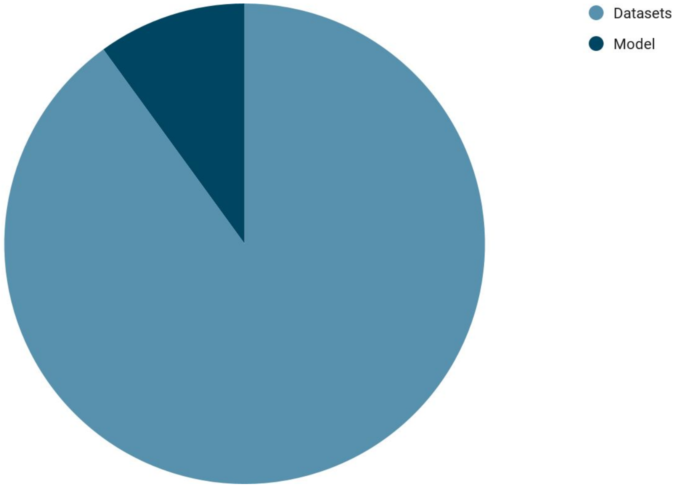
- Test performance
- Interpretation

# Modeling

*You may think this is how data scientists spend their time…*



*…but it is the other way around.*



# Deploy

**Deployment** is the integration of a machine learning model into a production environment.

Examples:

- A program that targets customer at risk of churning and send them personalized offers every month.
- An online bank website that recognizes uploaded documents.
- A virtual assistant device that answers to questions.

# Deploy

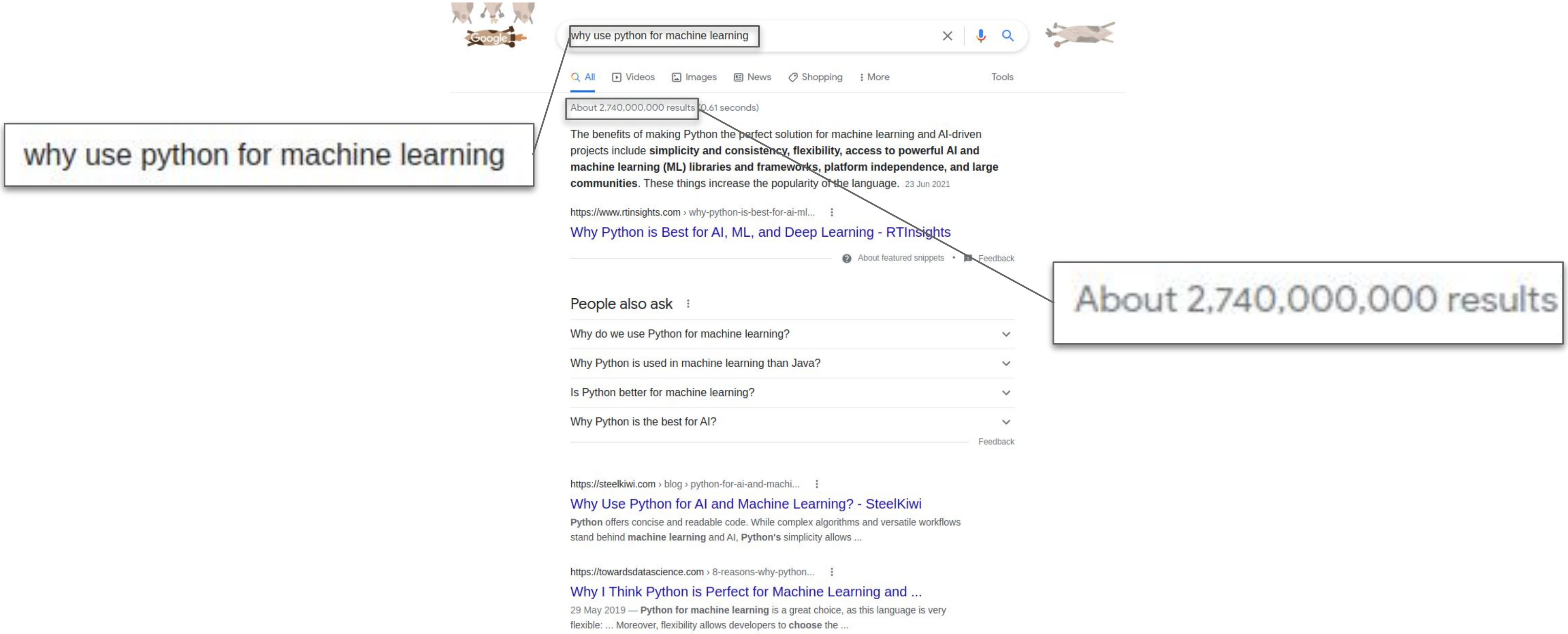
Challenges:

- Integration,
- Reproducibility,
- Scalability,
- Drift.

# Python for Machine Learning

1. Data Science Fundamentals
   a. What is Data Science?
   b. What is Machine Learning?
   c. Machine Learning Approaches
2. Machine Learning Projects
   a. Project workflow
   b. Data Exploration
   c. Modeling
   d. Deploy
3. **Python for Machine Learning**
   a. Why python?

# Python for Machine Learning



why use python for machine learning

About 2,740,000,000 results (0.61 seconds)

The benefits of making Python the perfect solution for machine learning and AI-driven projects include **simplicity and consistency, flexibility, access to powerful AI and machine learning (ML) libraries and frameworks, platform independence, and large communities**. These things increase the popularity of the language. 23 Jun 2021

https://www.rtinsights.com › why-python-is-best-for-ai-ml...

Why Python is Best for AI, ML, and Deep Learning - RTInsights

People also ask

- Why do we use Python for machine learning?
- Why Python is used in machine learning than Java?
- Is Python better for machine learning?
- Why Python is the best for AI?

https://steelkiwi.com › blog › python-for-ai-and-machi...

Why Use Python for AI and Machine Learning? - SteelKiwi

**Python** offers concise and readable code. While complex algorithms and versatile workflows stand behind **machine learning** and AI, **Python's** simplicity allows ...

https://towardsdatascience.com › 8-reasons-why-python...

Why I Think Python is Perfect for Machine Learning and ...

29 May 2019 — **Python for machine learning** is a great choice, as this language is very flexible: ... Moreover, flexibility allows developers to **choose** the ...

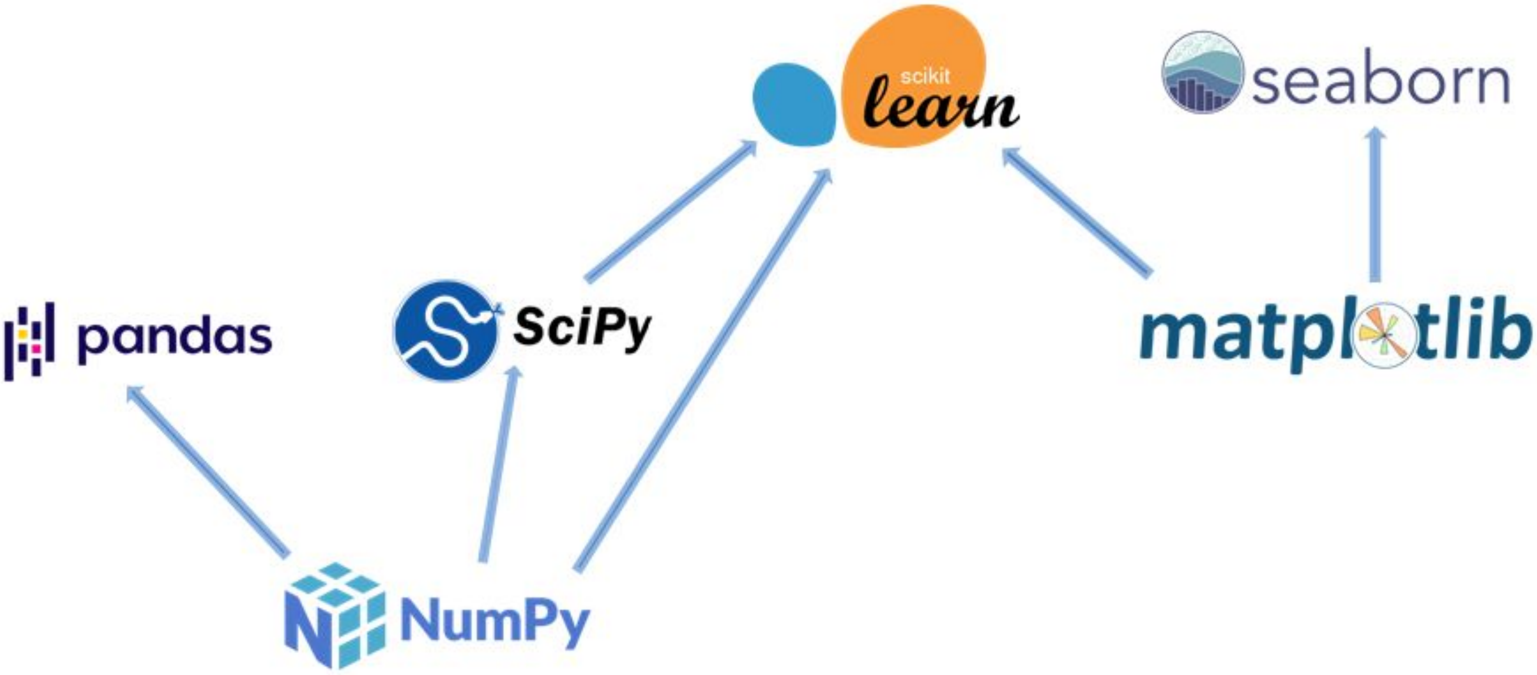
why use python for machine learning

About 2,740,000,000 results

# Python for Machine Learning

*What do **you** think?*

# Python for Machine Learning


scikit learn
seaborn
pandas
SciPy
matplotlib
NumPy