My 1st Machine Learning model

Atelier « Maker »

D4T4SQU4D

@Booster

Who we are

Nicolas FROT
Data Squad

D4T4SQU4D
Booster

OCTO
Technology

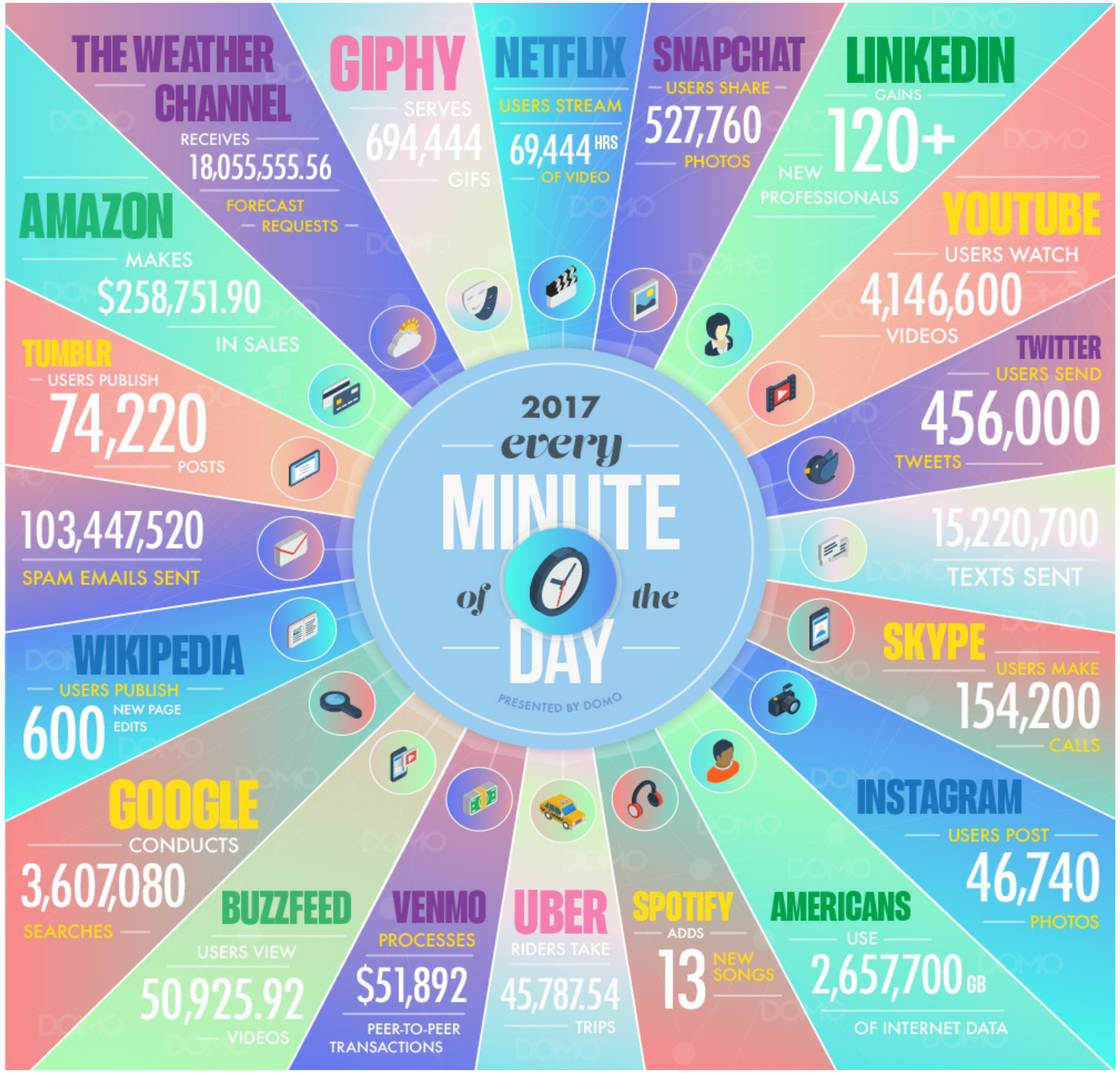
Florian BERGAMASCO
EP/EXPLO/GTS/IGR

DATA LAB
Innovation for data management

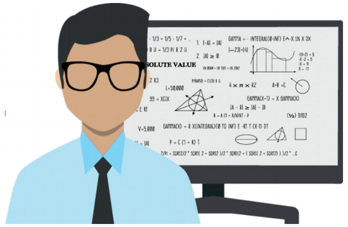
What is Big Data?

1956: 5 Mo,
$50k

2018: 256 Go,
$30



What are the roles in a Big Data organization?



Data Scientist

Machine Learning Engineer

Data Manager

Data engineer

Data Architect

Data Analyst

What is Machine Learning?

Machine learning algorithms build a mathematical model of sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to perform the task.

Machine learning is used in:

- Email filtering
- Image classification
- Fraud detection
- Etc…

Which use cases we saw at TOTAL?

- Use Case 1: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

- Use Case 2: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

- Use Case 3: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

- Use Case 4: XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

Some basic vocabulary

Pb: predict the quantity of apples sold in a supermarket for a given day

Variable / feature

Target variable

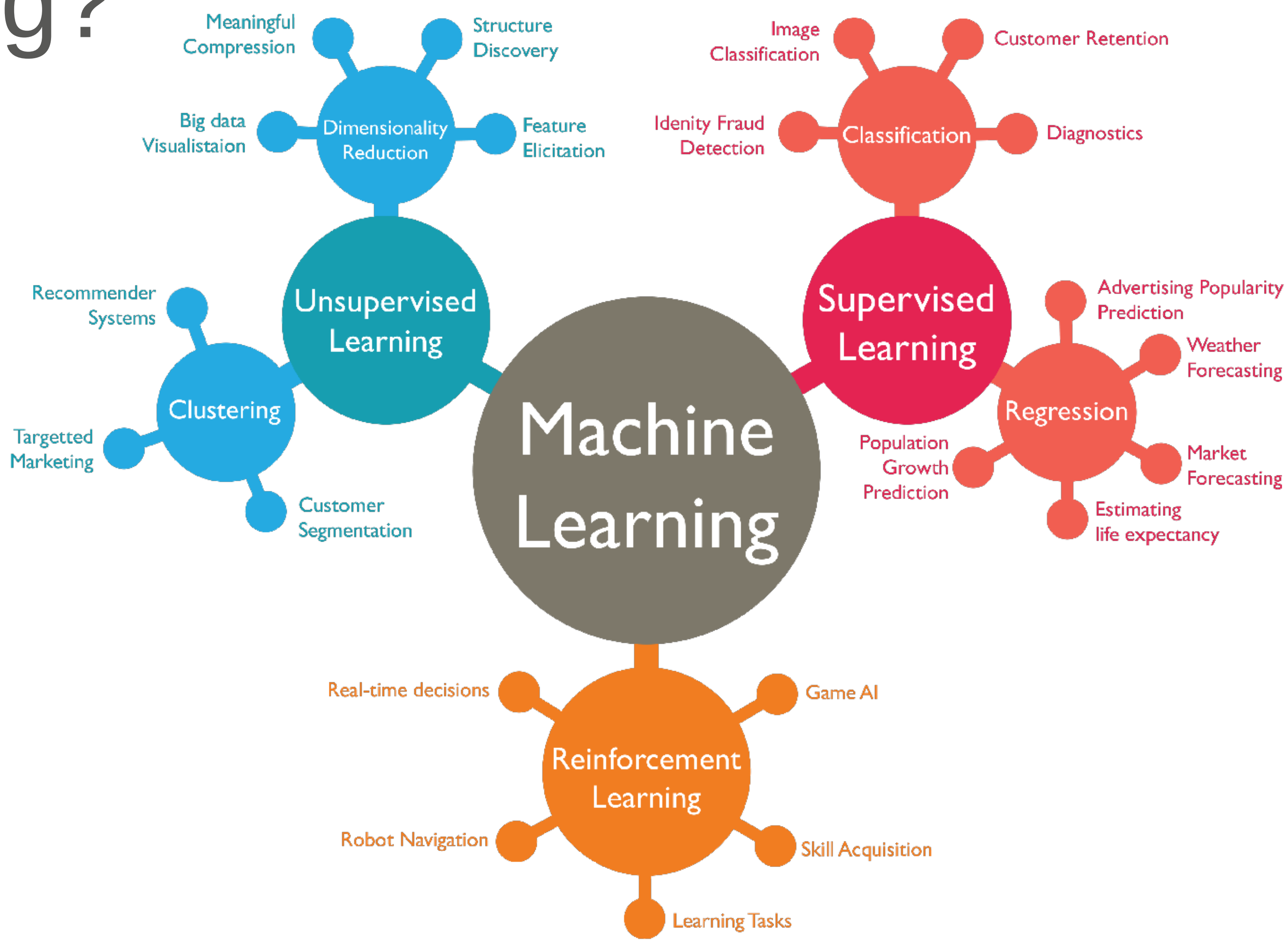
observation

	Temp ext (°C)	Day of week	…	Price/ kg for apples
01/01/2017	-10	3	…	2
02/01/2017	-8	4	…	2,03
03/01/2017	5	5	…	2,04
04/01/2017	6	6	…	2,50
05/01/2017	2	7	…	2,50

Apples sold (kg)
34
37
67
64
33
…
87

Dataset

What are the branches of machine learning?



Classification

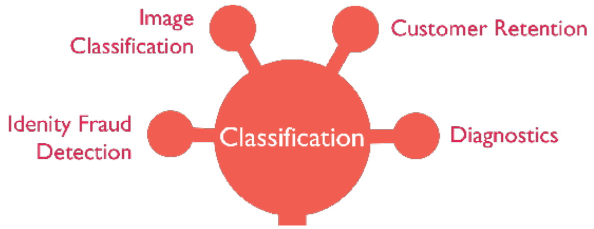


Image Classification

Customer Retention

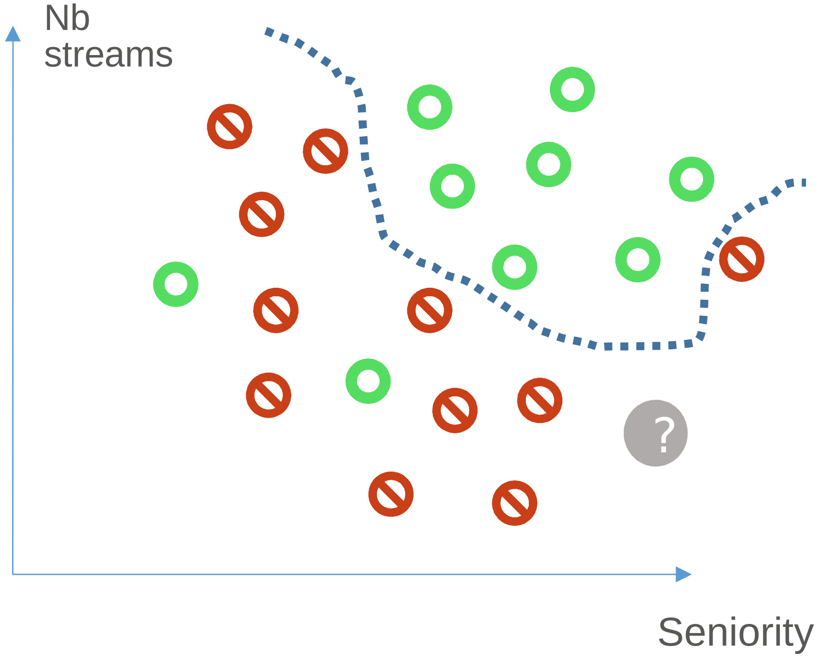
Idenity Fraud Detection

Classification

Diagnostics

Pb: (Spotify) Will the users buy our premium offer?

	Nb streams per day	Seniority	Buy after trial?
User 1	12	1	🚫
User 2	56	24	○
User 3	467	13	🚫
…	…	…	…
User n	32	4	🚫

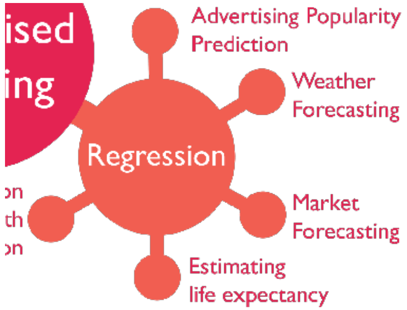


After the model has been fitted to the training set, let’s apply the prediction on a new observation:

User n+1			?

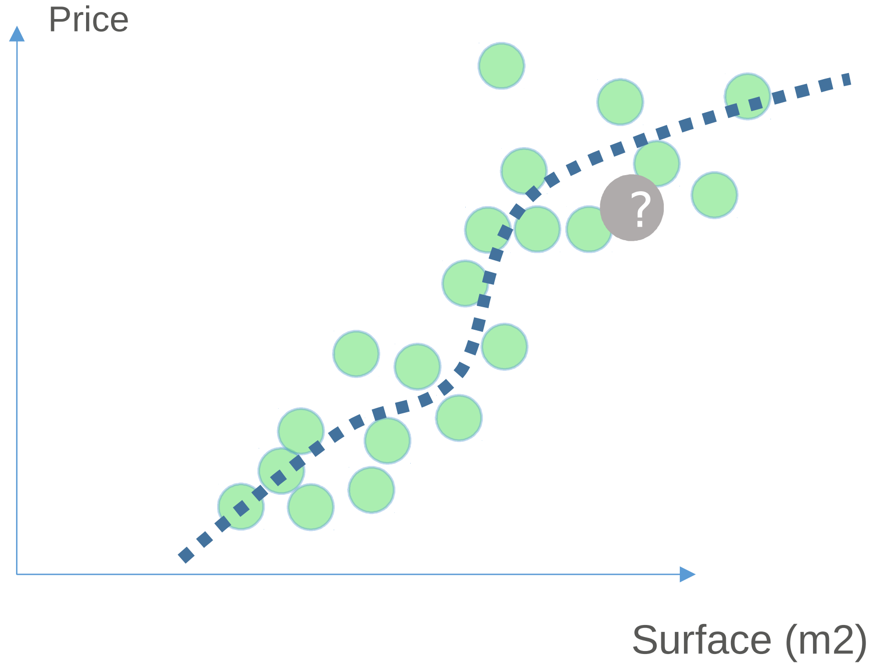
The model tries to find the best border that splits the positive and negative observations

Regression



Pb: (MeilleursAgents) How to estimate the price of an appartment?

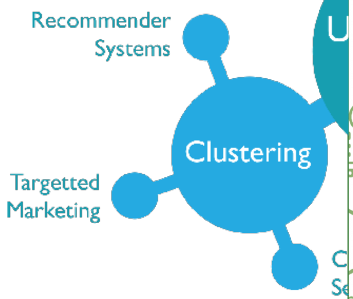
	Surface (m2)	Price (k€)
Apt 1	12	200 k
Apt 2	56	450 k
Apt 3	130	1200 k
…	…	…
Apt n	32	300 k



After the model has been fitted to the training set, let's apply the prediction on a new observation:

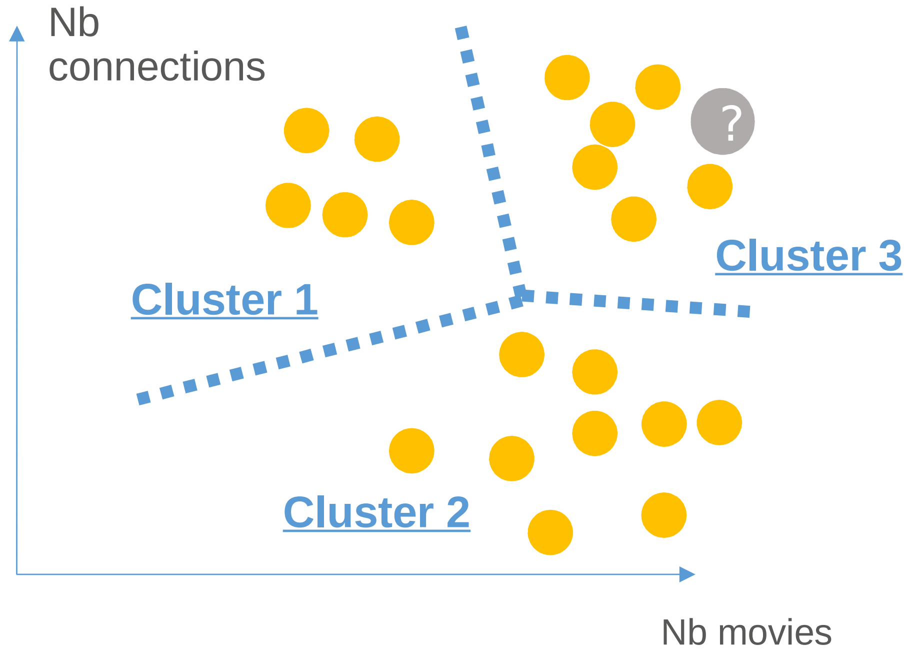
Apt n+1		?

Clustering



Pb: (Netflix) Can I group similar users by behaviour on the app?

	Nb movies (/month)	Nb connections
User 1	12	1
User 2	32	24
User 3	46	13
…	…	…
User n	32	44

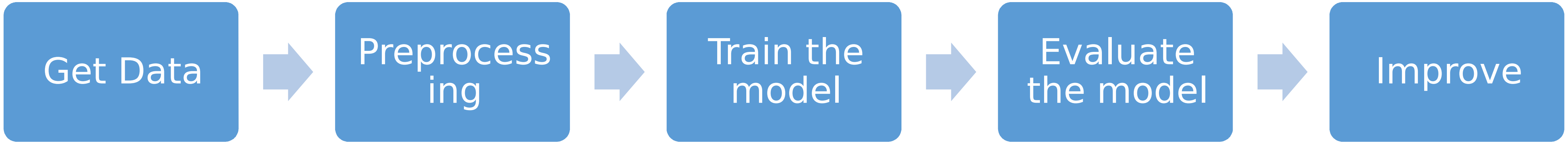


After the model has been fitted to the training set, let's apply the prediction on a new observation:

User n+1			?

The model tries to find the best border that splits the positive and negative observations

Workflow of a ML project



Get Data

Preprocessing

Train the model

Evaluate the model

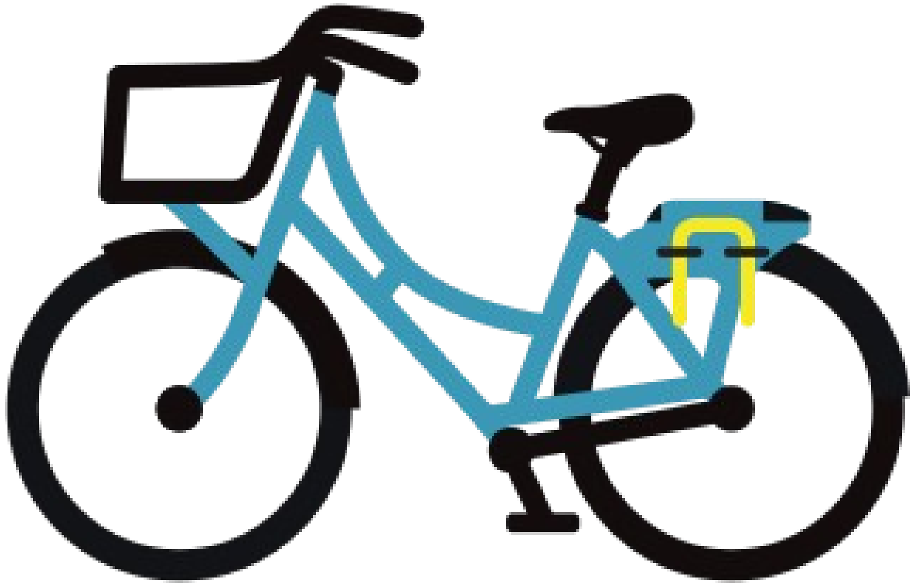
Improve

What are we going to do today?

The Challenge

Predict the number of shared bikes rented every hour in San Diego given meteorological information

- **Features**:
  - Temperature
  - Time
  - Humidity
  - Wheather
  - Weekday
  - Is_holiday
  - Etc…



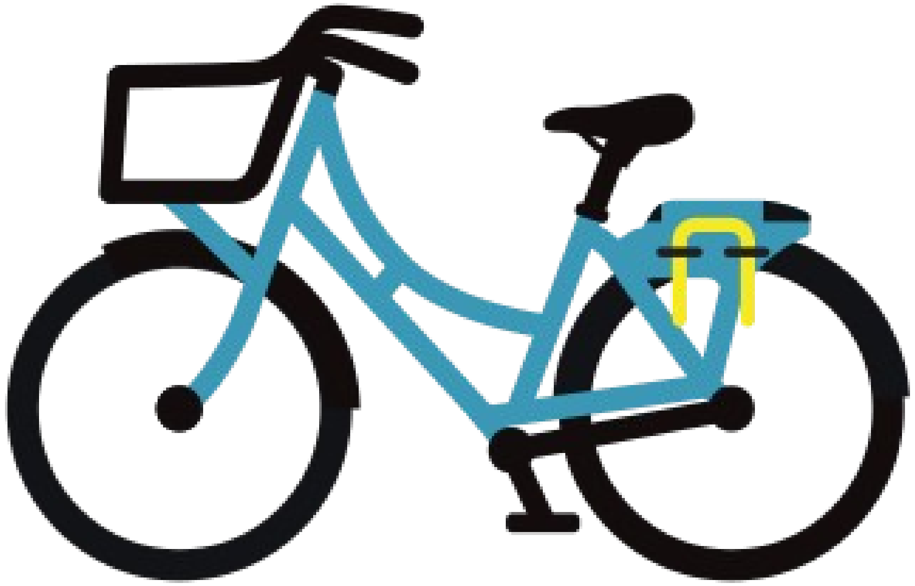
- **Data**: records from 2016/2017
- **Tools**: using Python (via Jupyter Notebook)

What are we going to do today?

The steps

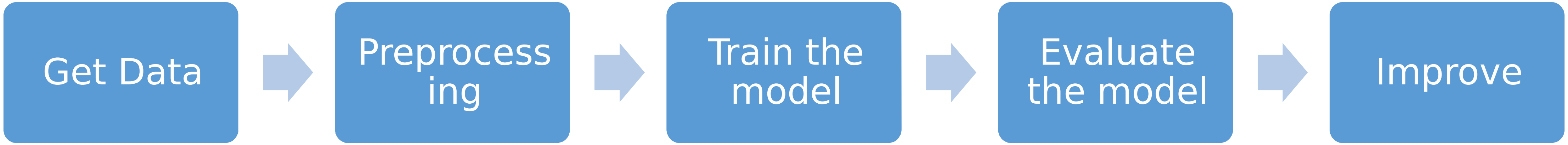
Theory & Hands on:

- Step 0:
  Introduction to Python and Jupyter
- Step 1:
  Build a first basic model
- Step 2:
  Improve your model: preprocessing
- Step 3:
  Improve your model: model choice and optimise the hyperparameters



Step 0: Introduction to Python and Jupyter notebooks

Step 1: Build a 1st basic model



Get Data

Preprocessing

Train the model

Evaluate the model

Improve

- Load data

Do the minimum:

- Remove rows with NA values
- Dummify categorical values
- Split Dataframe in train and test

Do the minimum:

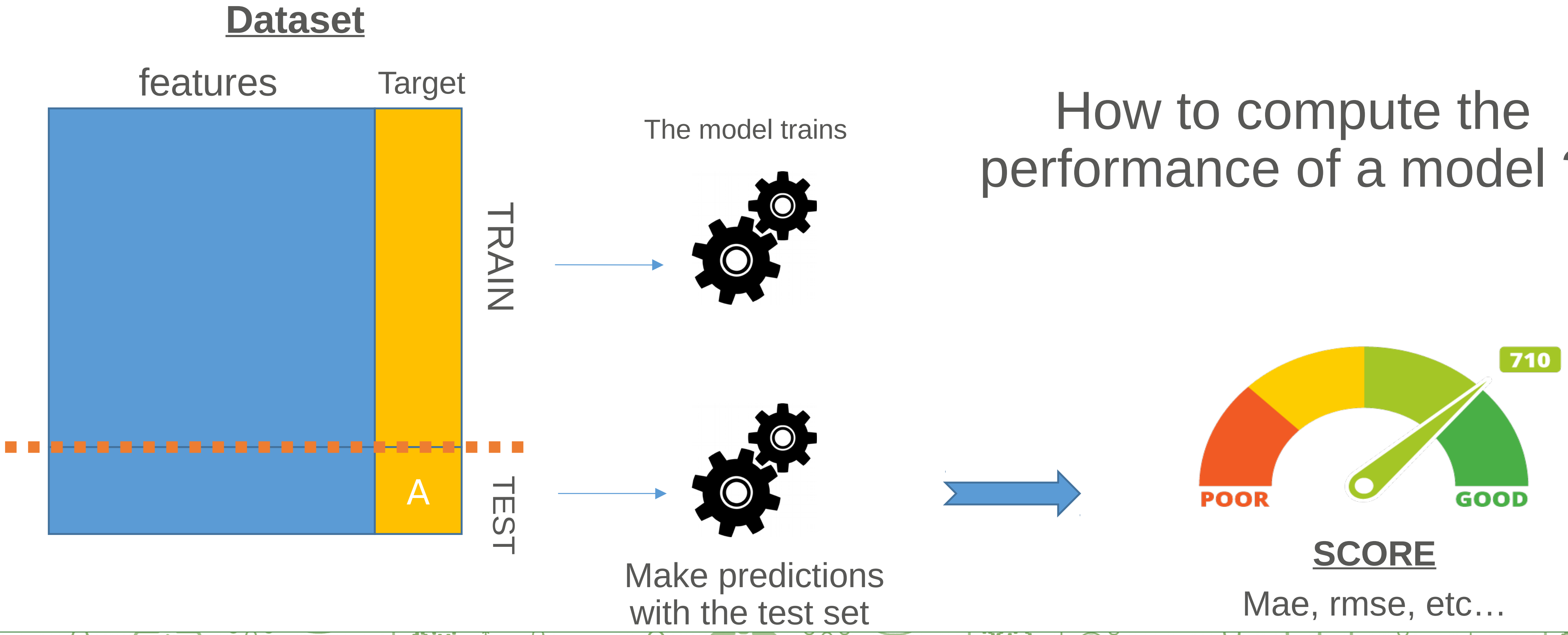
- Choose 1 model only

Do the minimum:

- Choose an evaluation metric
- Compute the score of your model

Step 1: Build a 1st basic model

Split the Dataset in order to evaluate your model



Dataset

features

Target

TRAIN

A

TEST

The model trains

Make predictions with the test set

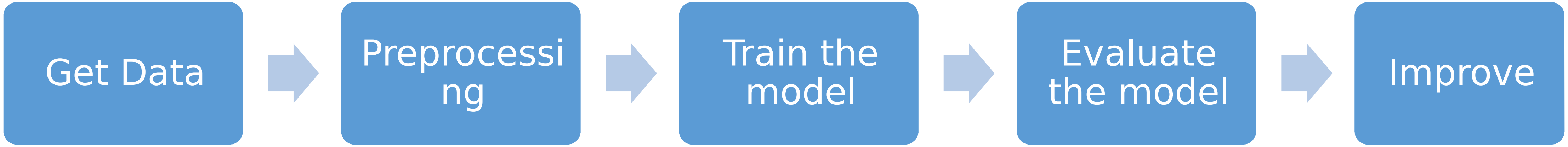
How to compute the performance of a model ?

SCORE

Mae, rmse, etc…

Step 1: Build a 1st basic model

Step 2: Improve your model : Preprocessing



Get Data → Preprocessing → Train the model → Evaluate the model → Improve

- Load data

- Dummify categorical values
- Split Dataframe in train and test
- Impute missing values
- Add some feature engineering

Do the minimum:

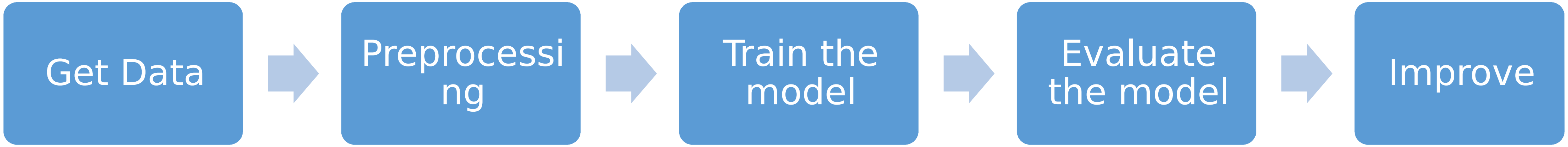
- Choose 1 model only

Do the minimum:

- Choose an evaluation metric
- Compute the score of your model

Step 2: Improve your model : Preprocessing

Step 3: Improve your model : Models and hyperparameters optimisation



- Load data

- Dummify categorical values
- Split Dataframe in train and test
- Impute missing values
- Add some feature engineering

- Try several models (NNs, RFs, Gradient Boosting etc…)
- Use several sets of hyperparameters

- Choose an evaluation metric
- Compute the score of your model

Step 3: Improve your model : Models and hyperparameters optimisation