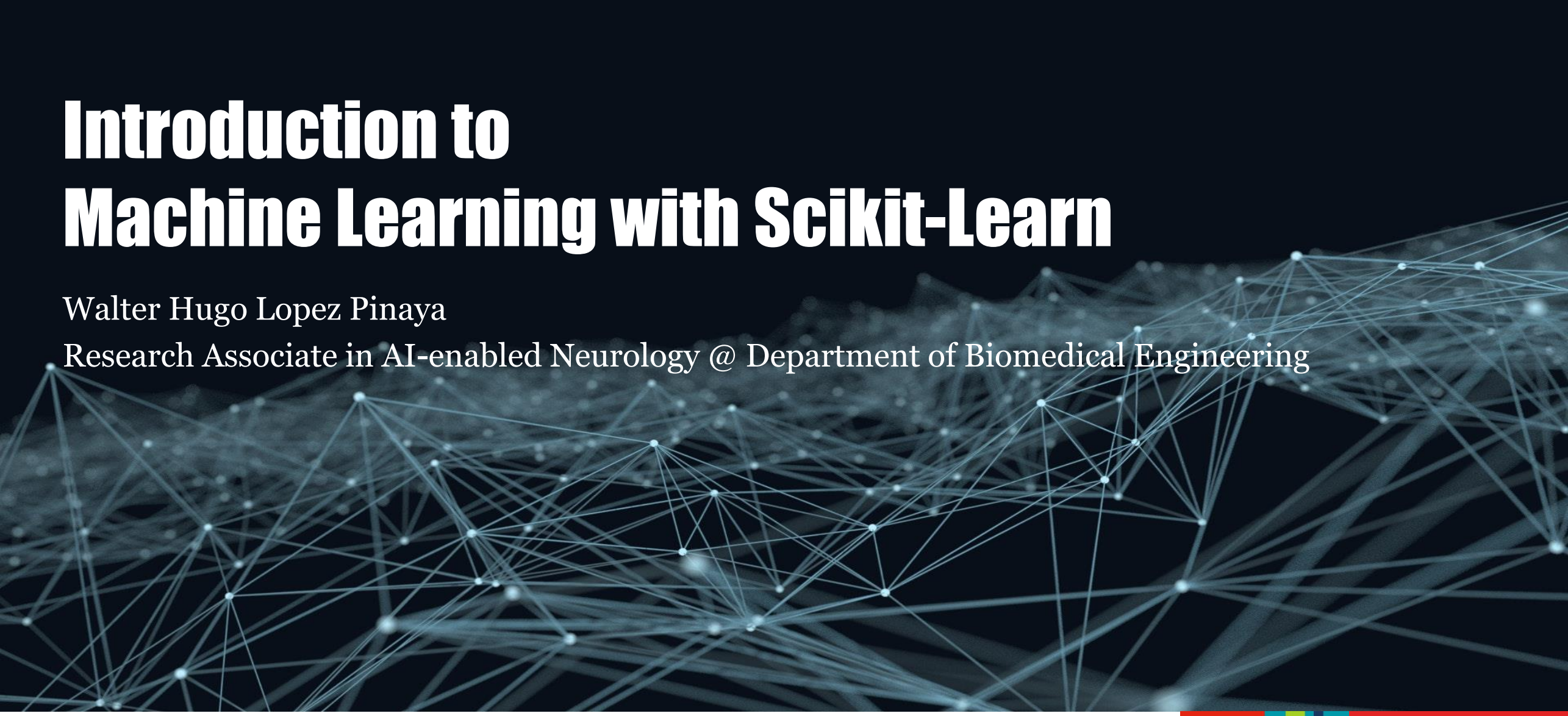
Introduction to Machine Learning with Scikit-Learn

Walter Hugo Lopez Pinaya
Research Associate in AI-enabled Neurology @ Department of Biomedical Engineering



Agenda

- Definition of machine learning
- Types of machine learning
- Workflow for supervised learning
- Practical exercises

WHAT IS MACHINE LEARNING?

Write programs that solve the problem

Write programs that

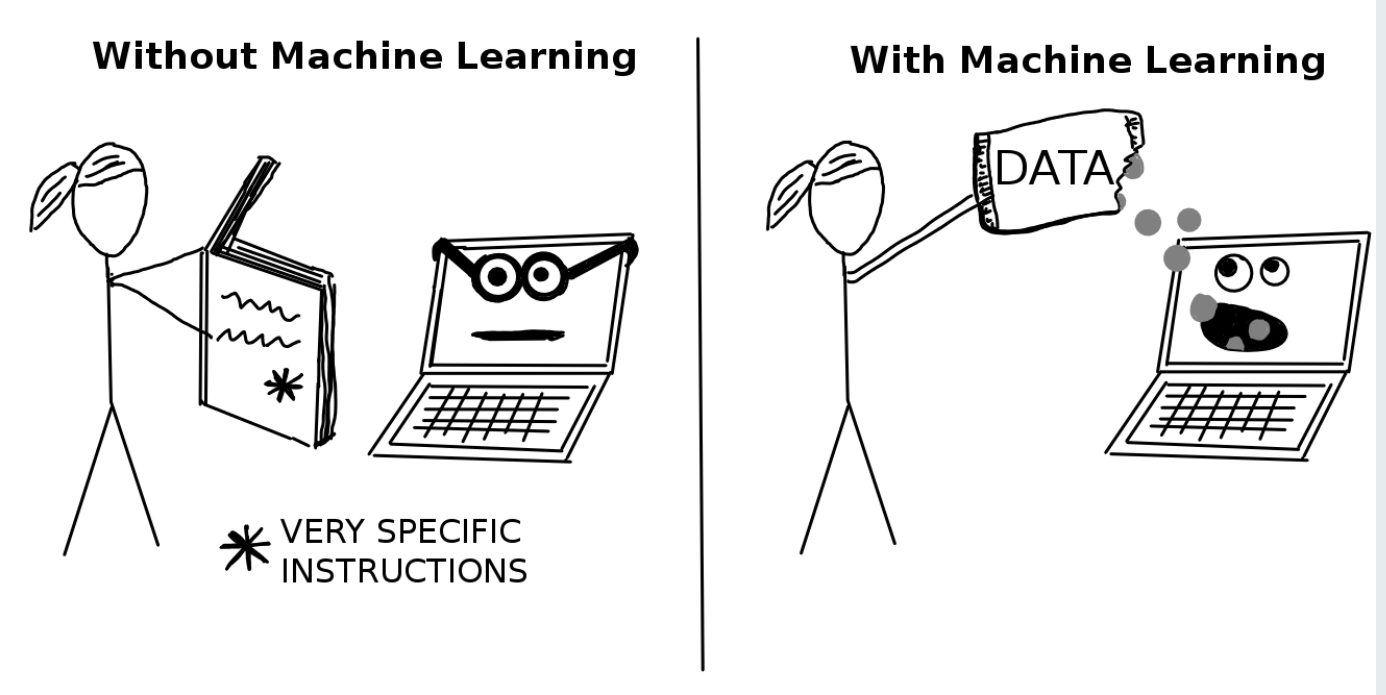
learn to

solve the problem

from examples

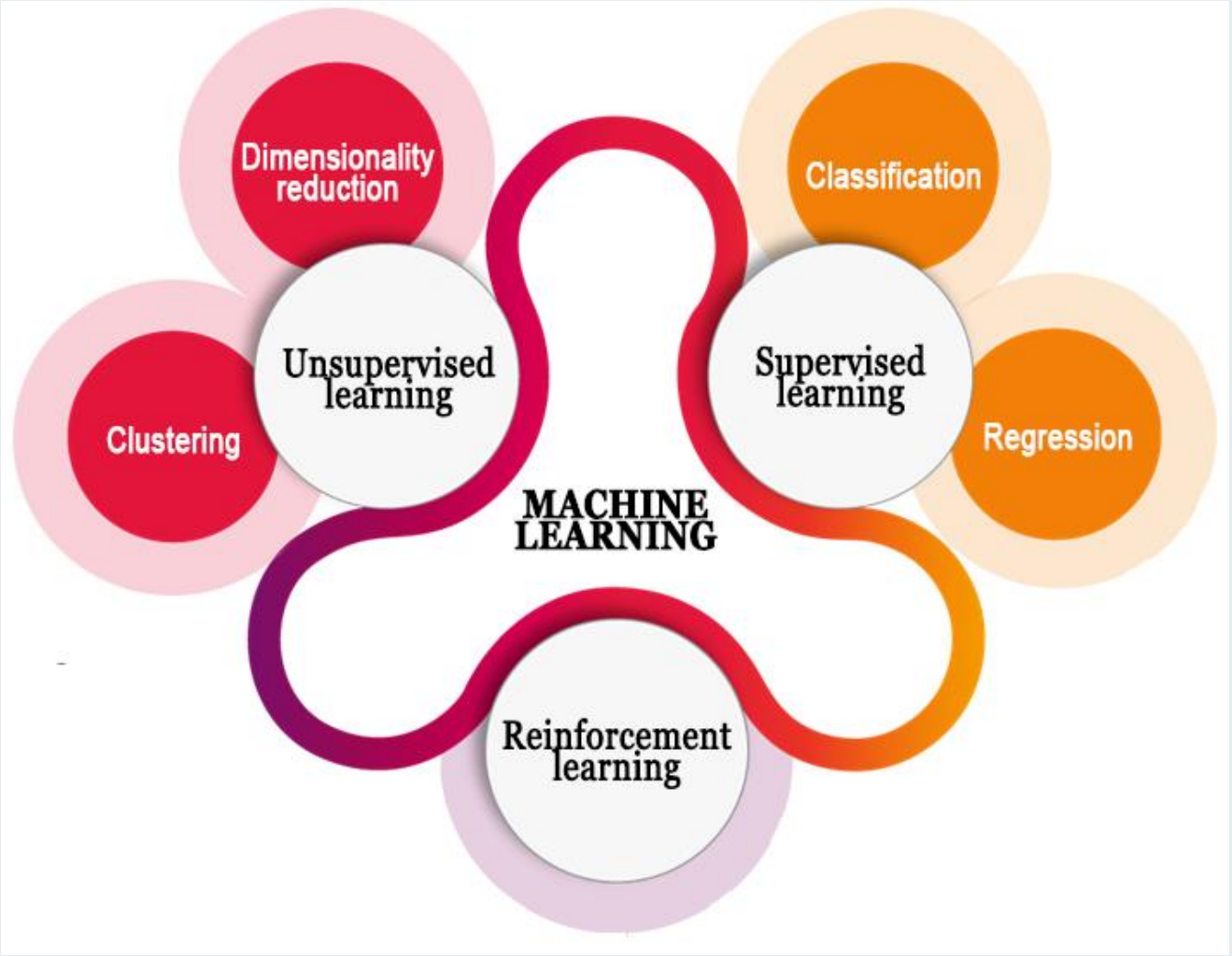
What is Machine Learning?

- Algorithms that can learn from and make predictions on data
- Computational statistics and mathematical optimization to discover trends and patterns
- Helps us discover patterns in the data and use these patterns to make predictions about new data

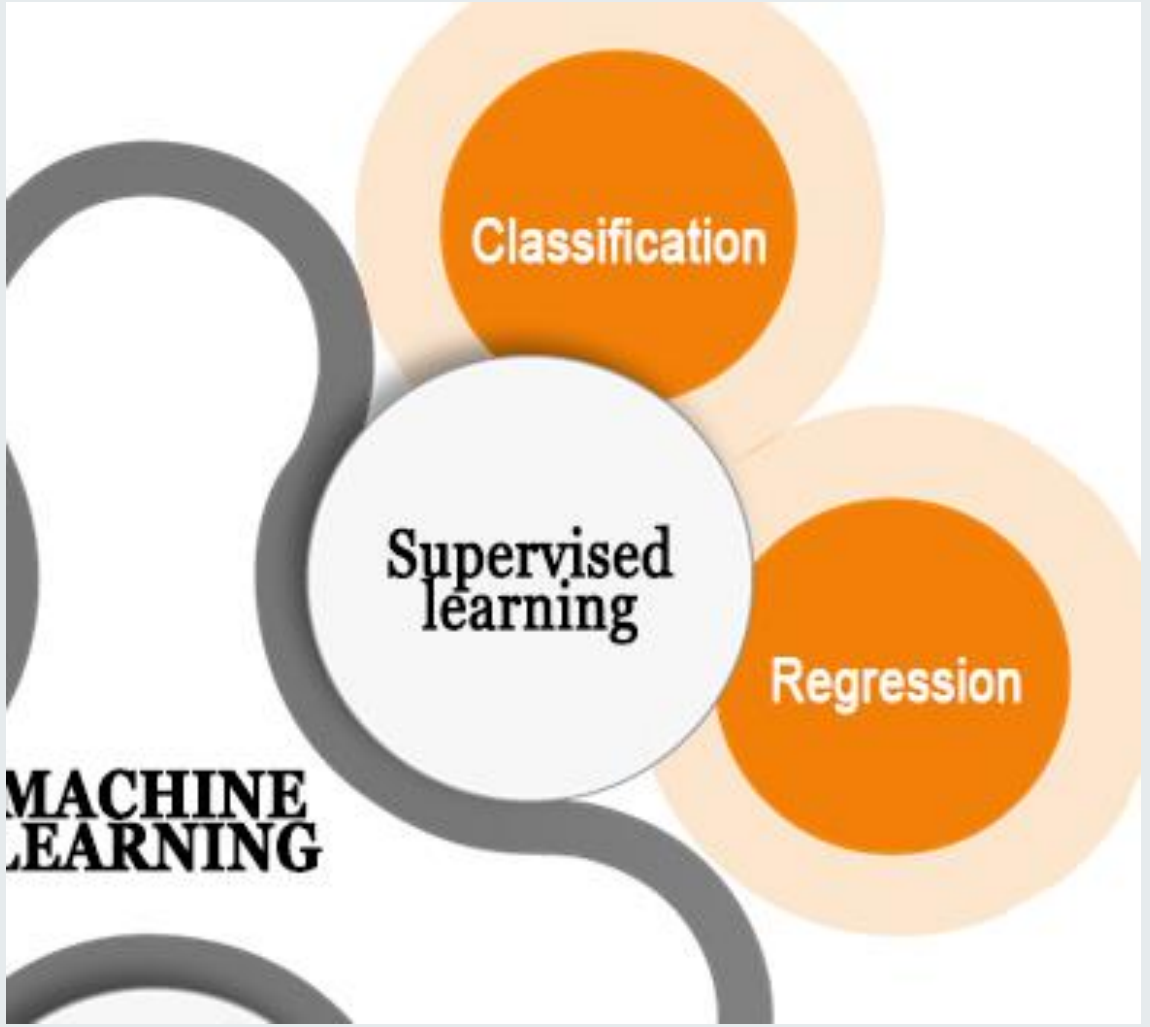


Machine learning types

Machine learning types



Supervised Learning



Done using a **ground truth**

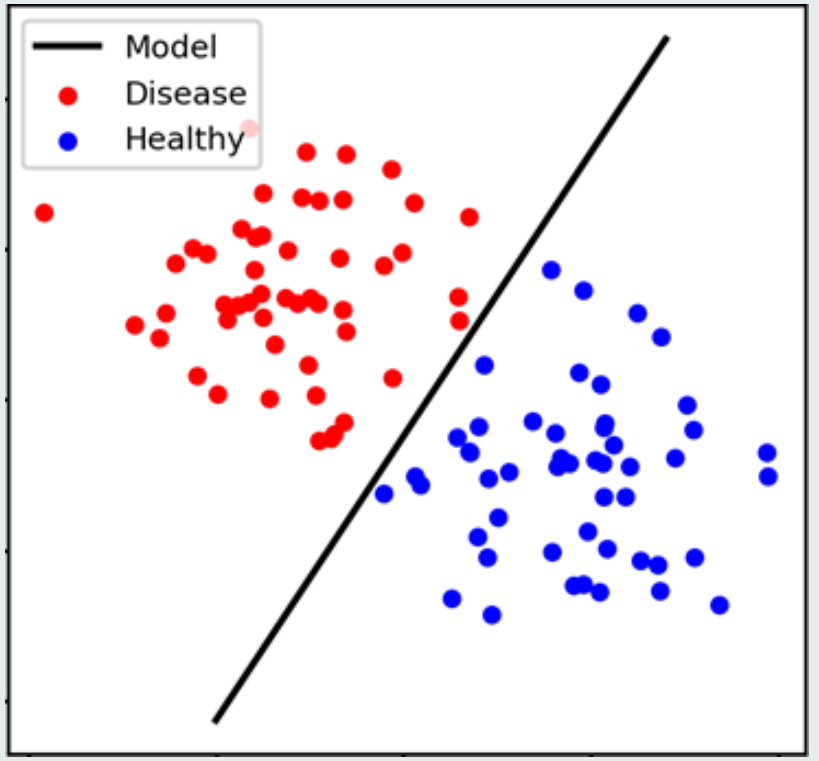
The goal is to learn a **function** that best approximates the relationship between **input** and **output** observable in the data

$$f(x) \rightarrow y$$

Supervised Learning - Examples

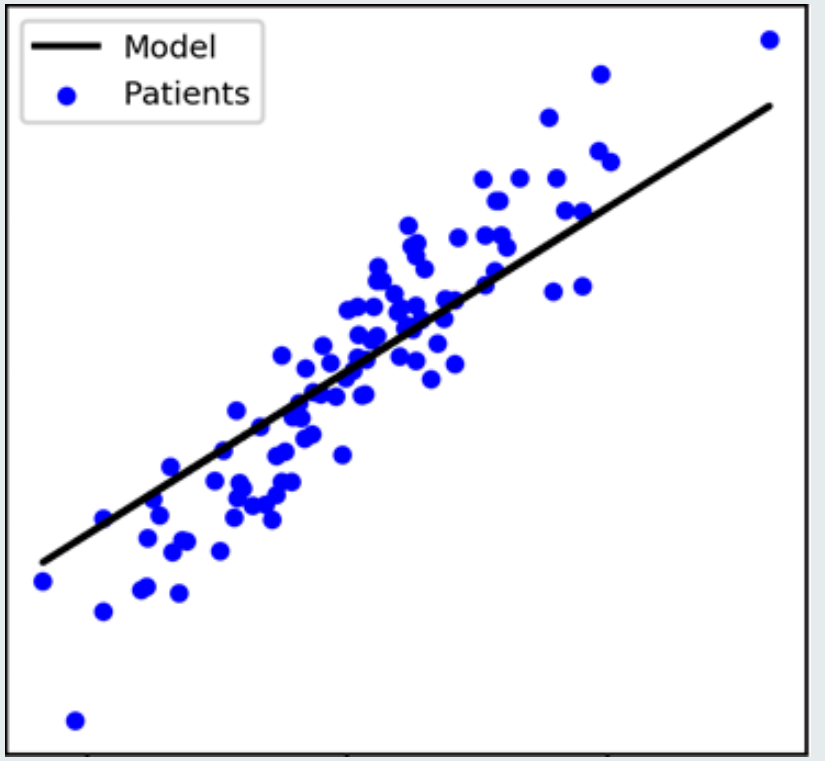
Classification

- Each example is associated with a qualitative target value, which corresponds to a class



Regression

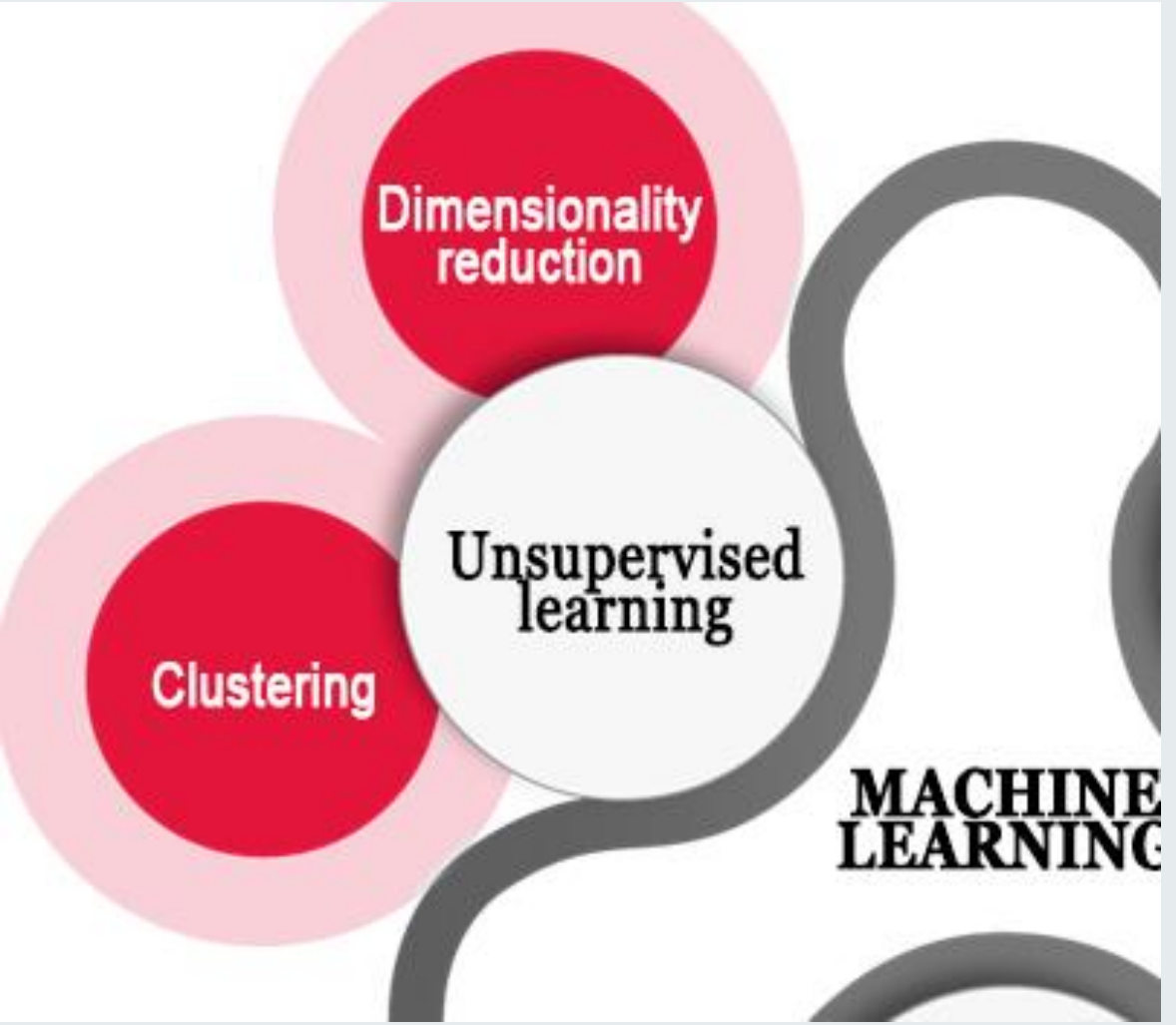
- Each example is associated with a quantitative target value



Unsupervised Learning

Does **not** have desired outputs

Infer the **natural structure** present within a set of data points



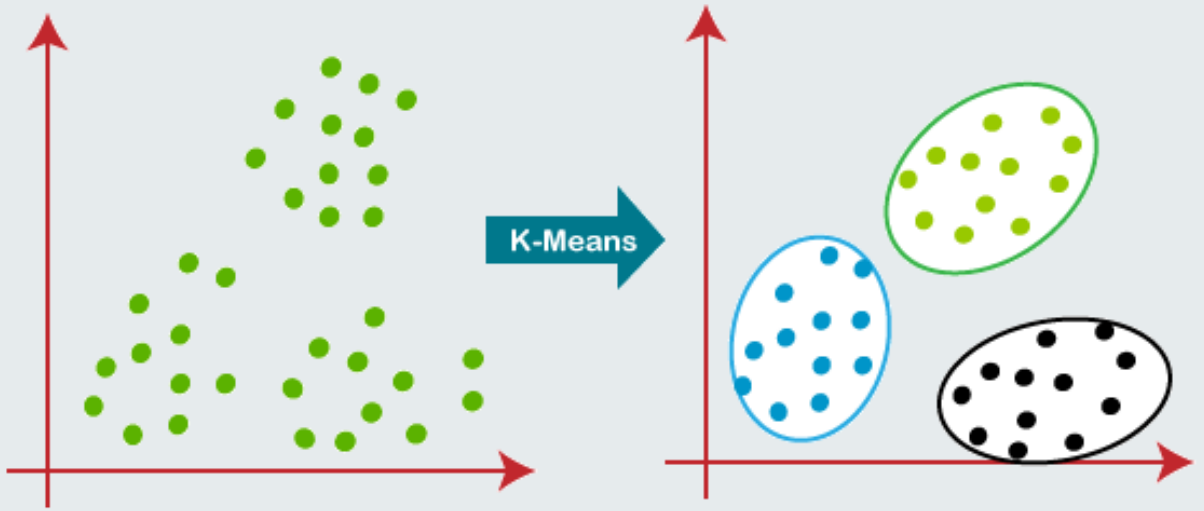
Unsupervised Learning - Examples

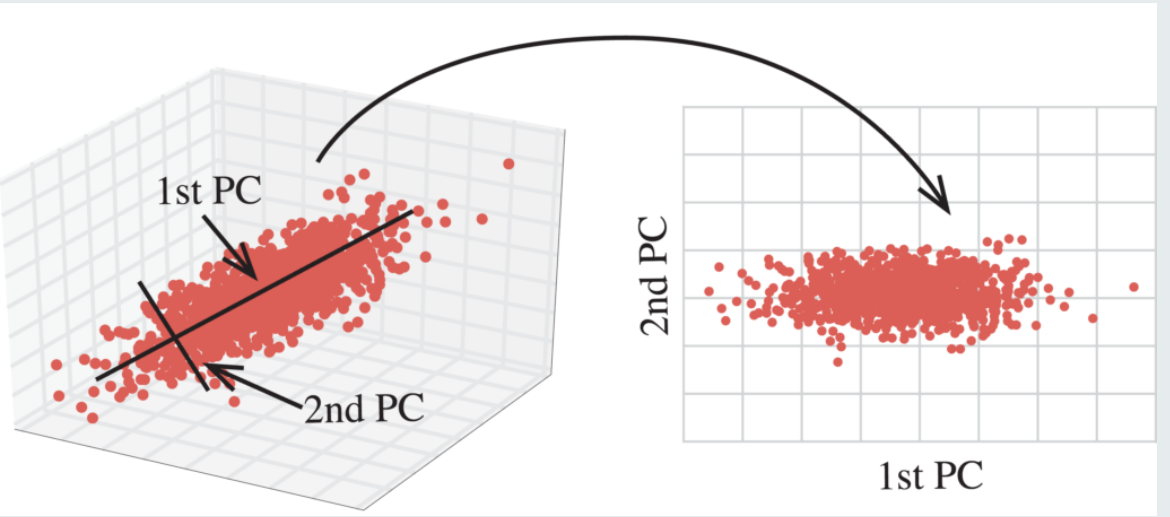
Clustering

- Automatic grouping of similar objects into sets.

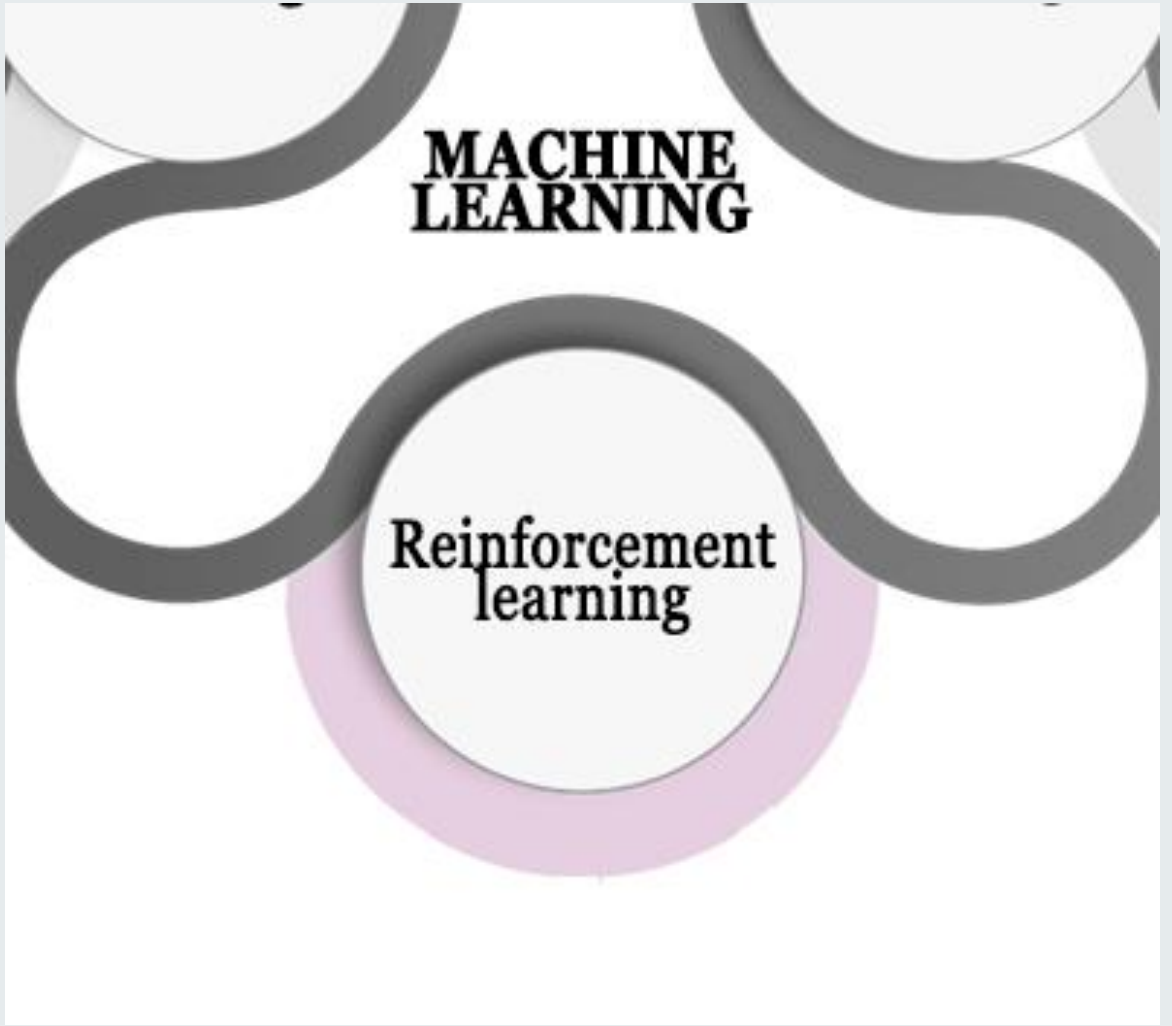
Dimensionality reduction

- Reducing the number of variables to consider.





Reinforcement Learning



Learn to **react** to an **environment** on their own

Imagine you want to teach a machine to play a very basic video game and never lose. You set up the model with the game, and you tell the model not to get a "game over" screen.

During training, the agent receives a **reward** when it performs this task.

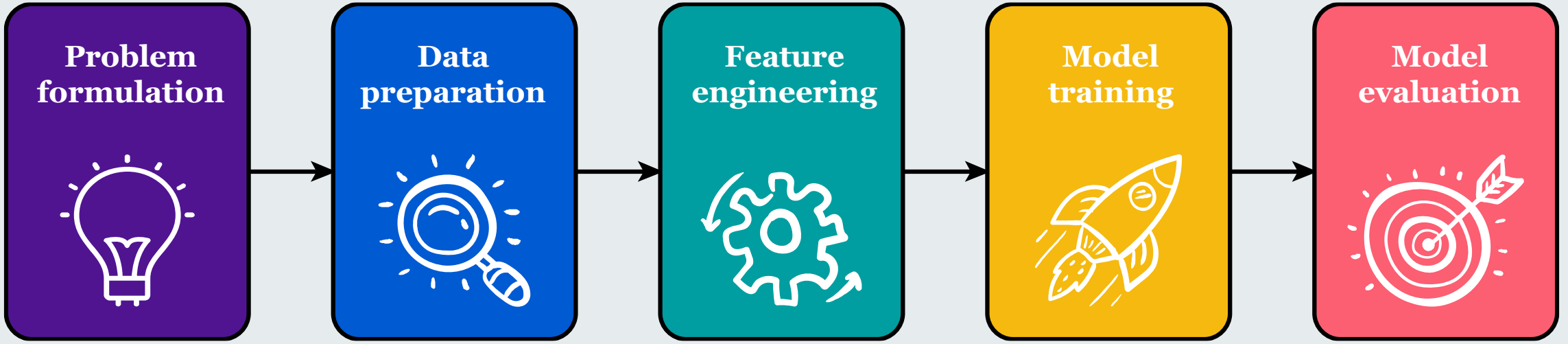
Scikit-learn

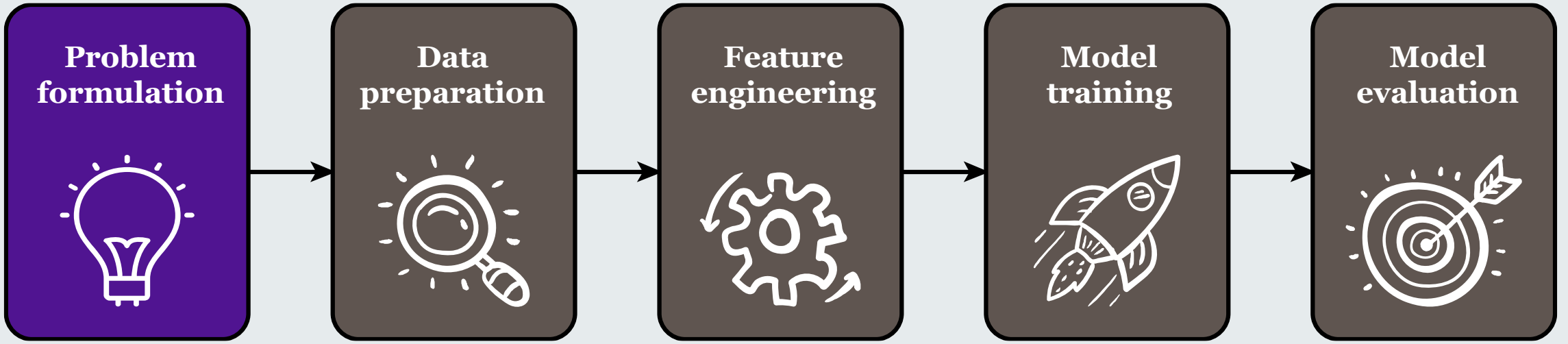
- Scikit-learn is an open-source machine learning library
- Supports **supervised** and **unsupervised** learning
- Provides various tools for model fitting, data pre-processing, model selection and evaluation, and many other utilities

Today, we will implement a solution for a task using machine learning (supervised learning). For this, we will follow an organized manner.

Machine Learning Workflow

Machine Learning workflow




Problem formulation
Data preparation
Feature engineering
Model training
Model evaluation

Problem formulation

Well-defined research question requires clearly defined **feature set**, **target variable**, and **task**.

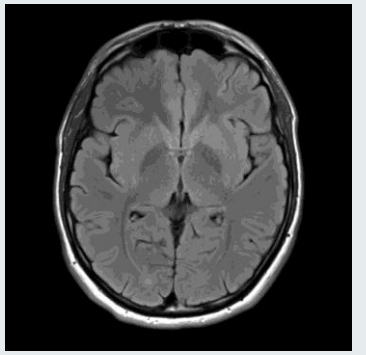
Problem formulation

Feature set

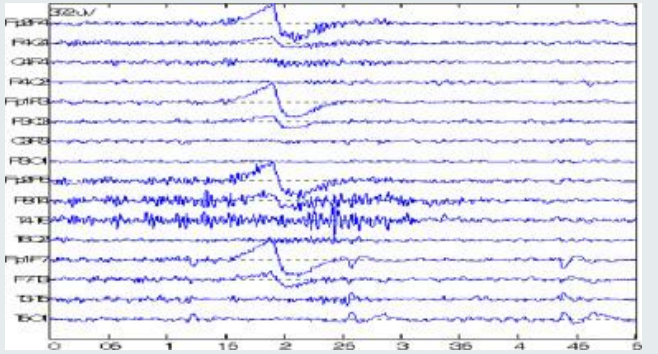
Comprises the data that will serve as input to the machine learning algorithm

Examples

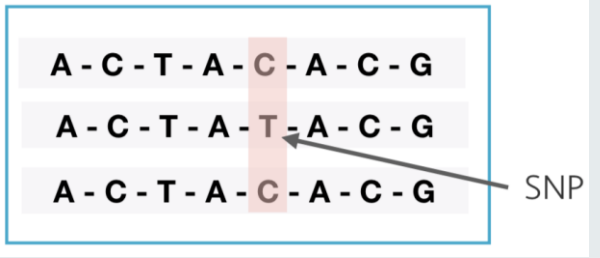
Brain image



EEG signals



Genetic data



Clinical scores

Problem formulation

Target variable

What the machine learning algorithm will learn to predict

Examples

- Diagnosis
- Treatment response
- Recovery
- Mortality risk

Problem formulation

Task

Defines what the machine learning algorithm should do with the features and target variable

Examples

- Classification
- Segmentation
- Regression
- Object detection

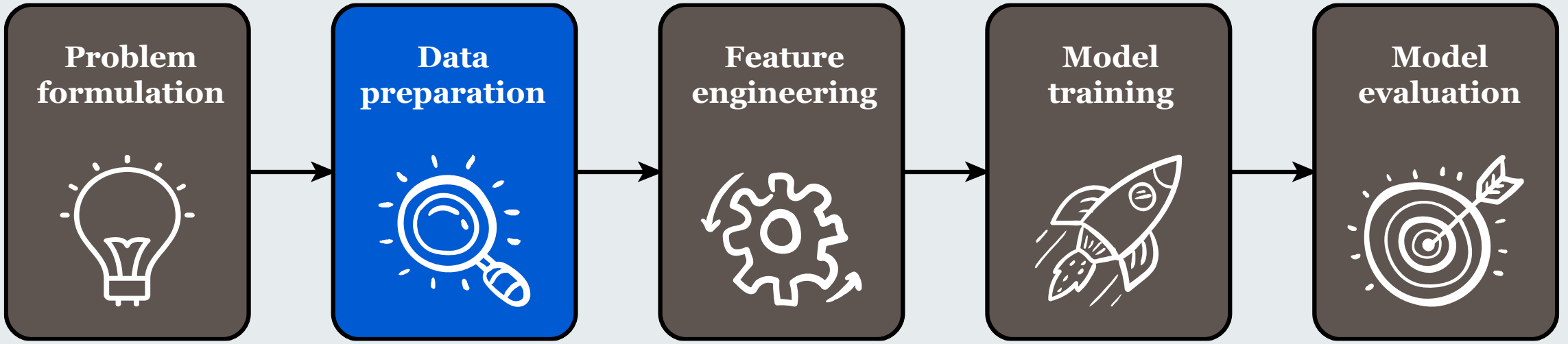
Problem formulation

Putting all these elements together enables one to formulate a succinct problem statement

"Using EEG image data (features) to categorize (task) which patients will and will not benefit from a certain treatment (target)"

"Using clinical scores (features) to categorize (task) which patients will and will not recover after 6 months (target)"

"Using MRI voxels data (features) to categorize (task) which voxels are part of a tumour or part of the health tissue (target)"


Problem formulation
Data preparation
Feature engineering
Model training
Model evaluation

Data Preparation

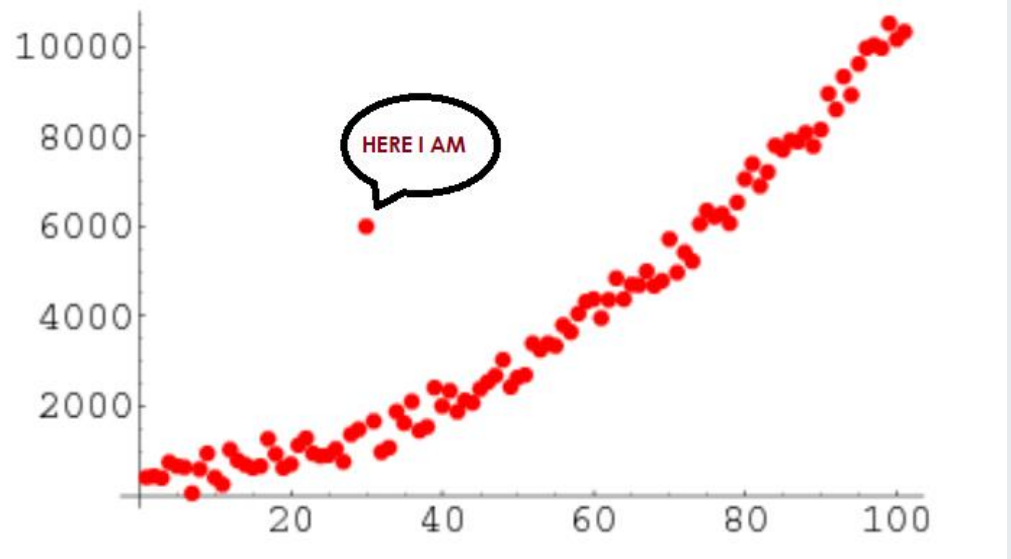
Machine learning helps us discover patterns in the data and use these patterns to make predictions about new data

To achieve this, however, it is important to clean, explore, and prepare the data to improve the overall quality of the dataset

Data Preparation

Histograms and scatter plots to explore the data

Range of strategies for minimizing the impact of **outliers** and **missing data**



	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

"A machine learning model is only as good as the data used to develop it"

Data Preparation - Splitting your data

Learning the parameters of a prediction function and testing it on the same data is a **methodological mistake**

A model that would repeat the labels of the samples that it has just seen would have a perfect score but would fail to predict anything useful on yet-unseen data

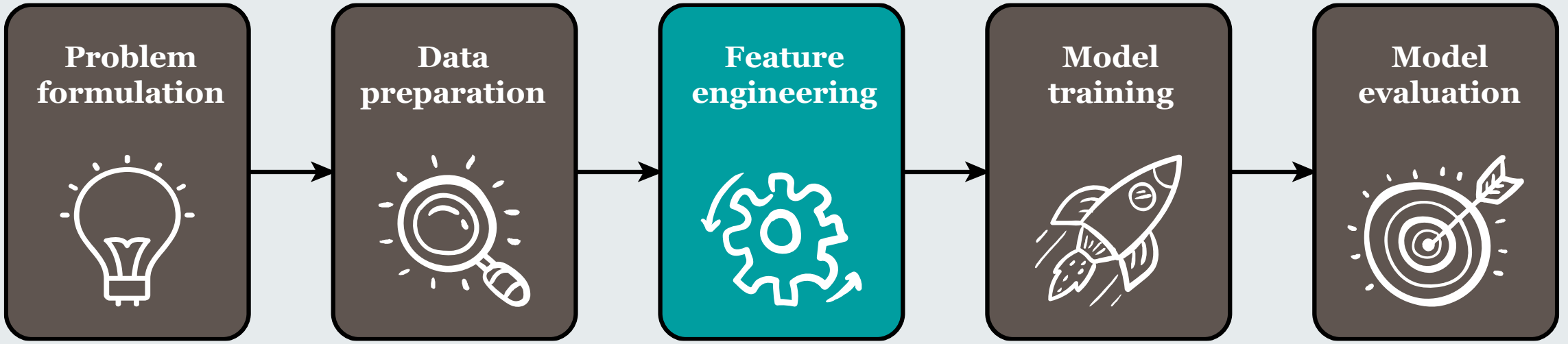
To avoid it, it is common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set

Data Preparation - Splitting your data

To train and evaluate our machine learning solution, we split the data into subsets

- **Training set** — a subset to train a model
- **Test set** — a subset to test the trained model

If we are exploring different configurations for our pre-processing and models, we should have a **Validation set** – a subset of the training set where we evaluate different strategies

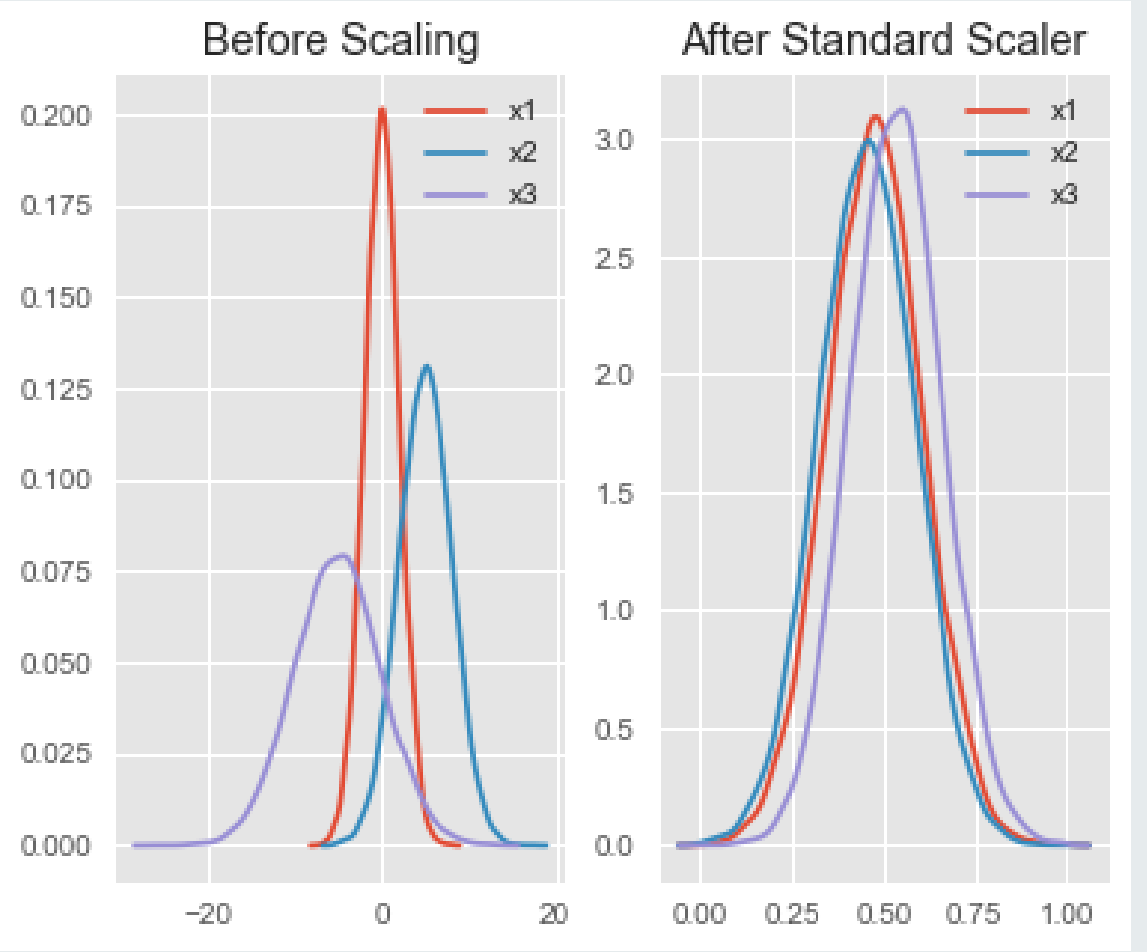

Problem formulation
Data preparation
Feature engineering
Model training
Model evaluation

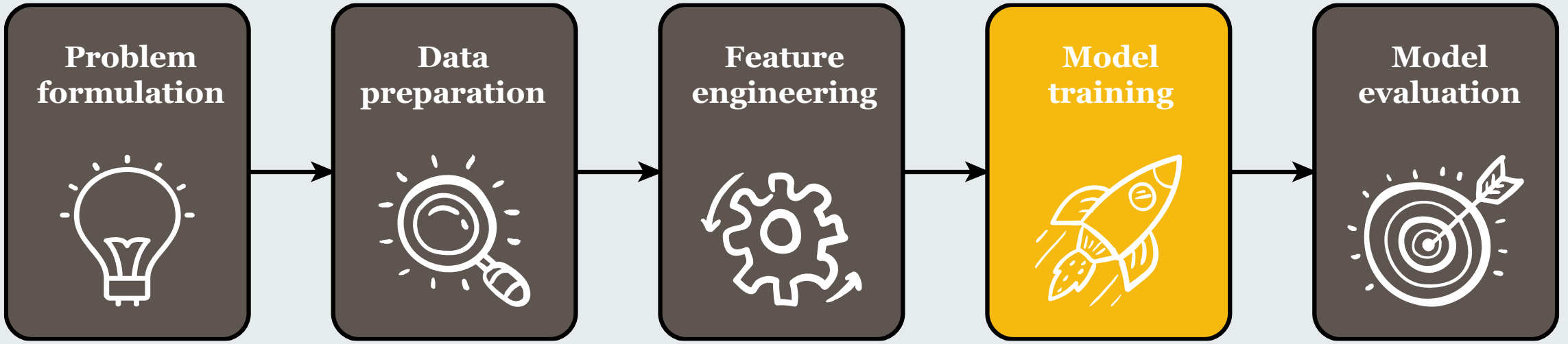
Feature Engineering

Feature engineering usually refers to **transformations** performed on the **raw data** to generate **features**

Some examples:

- Dimensionality reduction
- Feature selection
- Feature scaling/normalizing



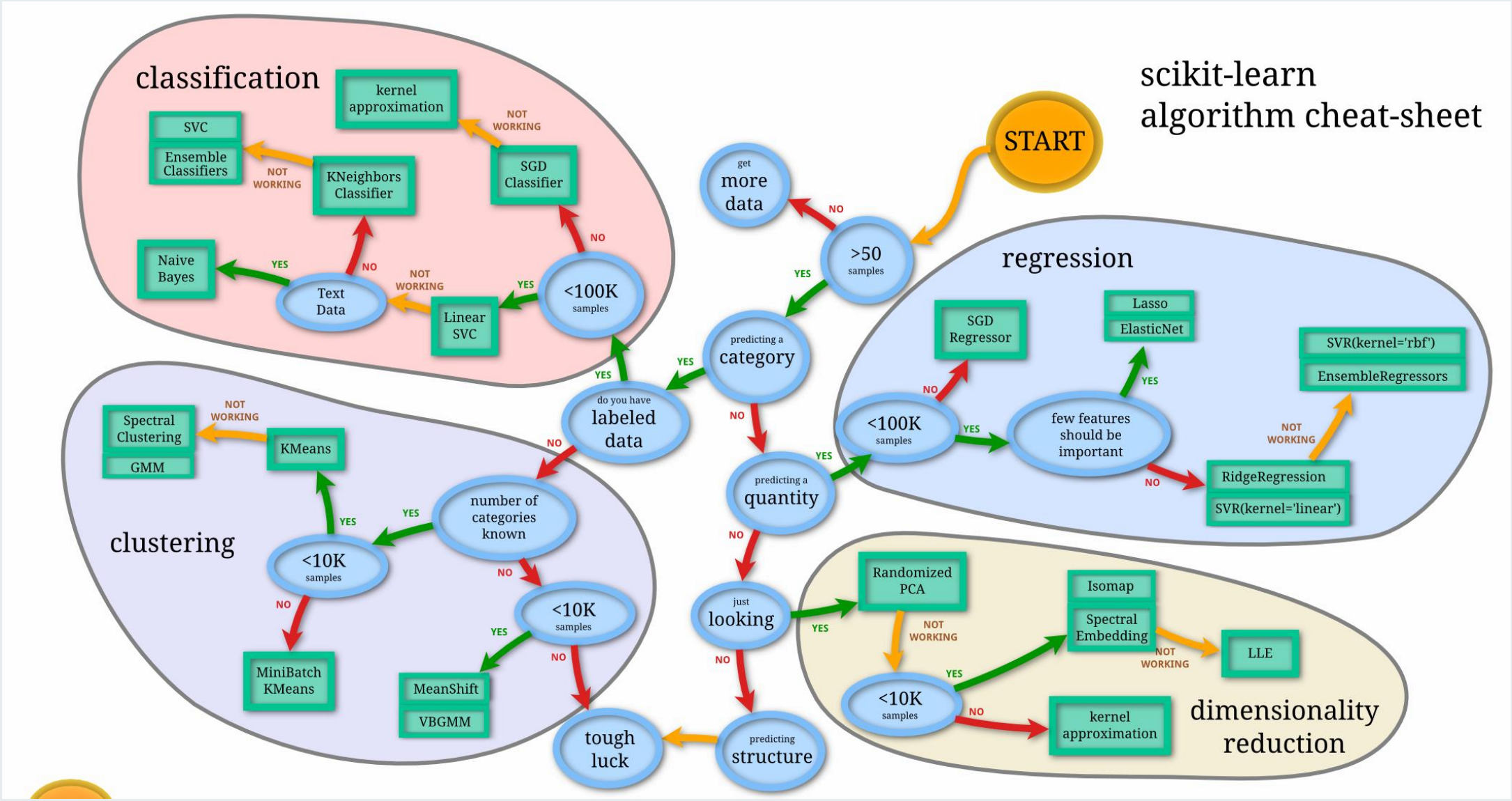

Problem formulation
Data preparation
Feature engineering
Model training
Model evaluation

Model Training

Model training refers to the process in which a machine learning algorithm finds a function f that best maps some given input features X and target variable y

Scikit-learn provides dozens of built-in machine learning algorithms and models, called **estimators**. Each estimator can be fitted to some data using its fit method.

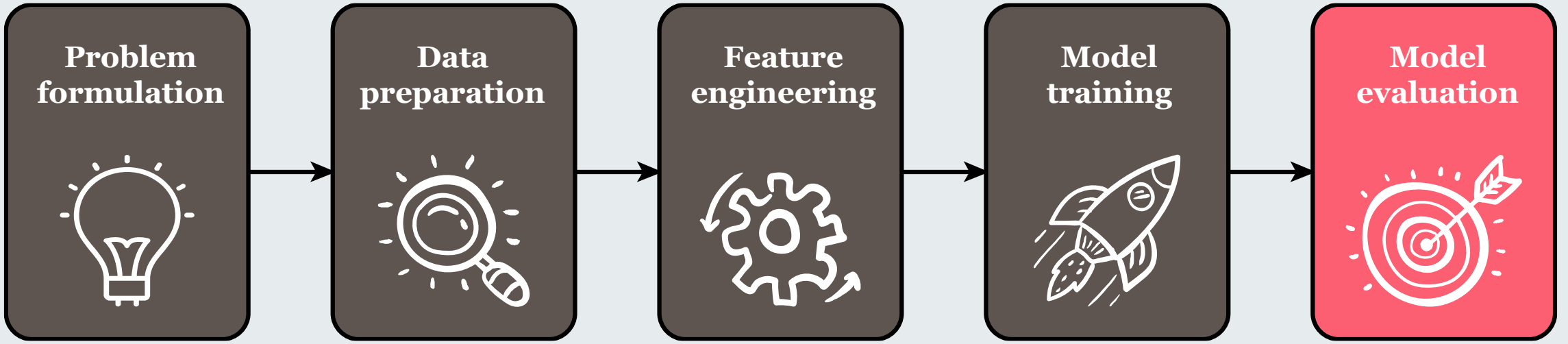
Model Training



Model Training

Different models can have very different results

https://ml-playground.com/


Problem formulation
Data preparation
Feature engineering
Model training
Model evaluation

Model Evaluation

- Fitting a model to some data does not entail that it will predict well on unseen data

- Performance is calculated by comparing the predicted labels against the true labels in the test set

Some examples:

- Accuracy
- Mean absolute error
- Mean squared error
- Sensitivity
- Specificity
- AUC ROC
- R Squared

Practical exercises

30~40 minutes to go through the Google's Colab notebook available at the Github page

Next steps

Online material

- https://scikit-learn.org/stable/getting_started.html
- https://machinelearningmastery.com/

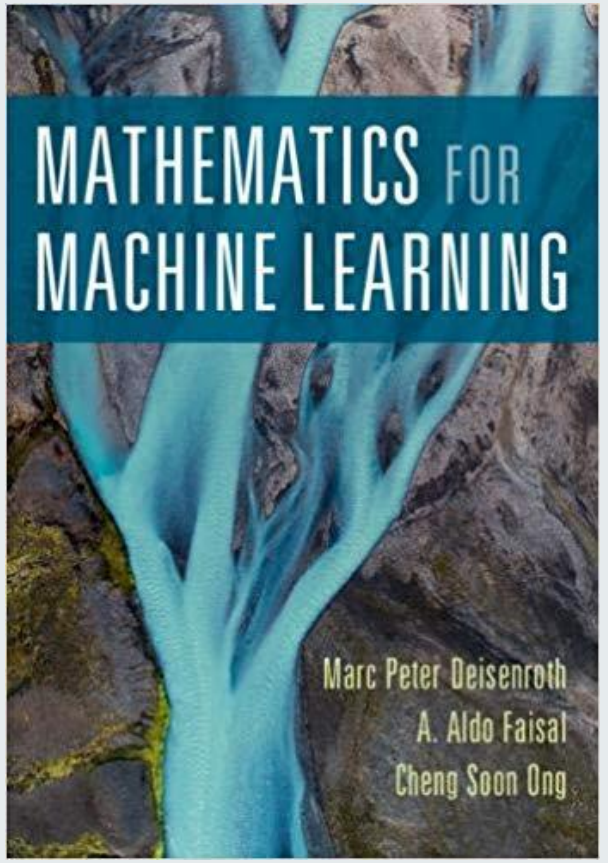
Online courses

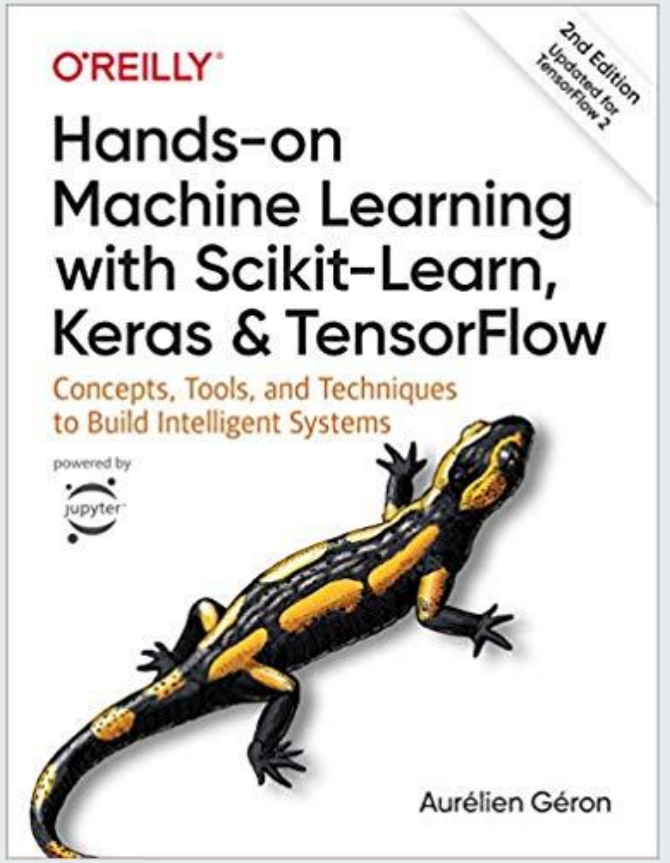
- https://www.coursera.org/learn/machine-learning *
- http://course18.fast.ai/ml
- https://developers.google.com/machine-learning/crash-course
- https://www.kaggle.com/learn/intro-to-machine-learning

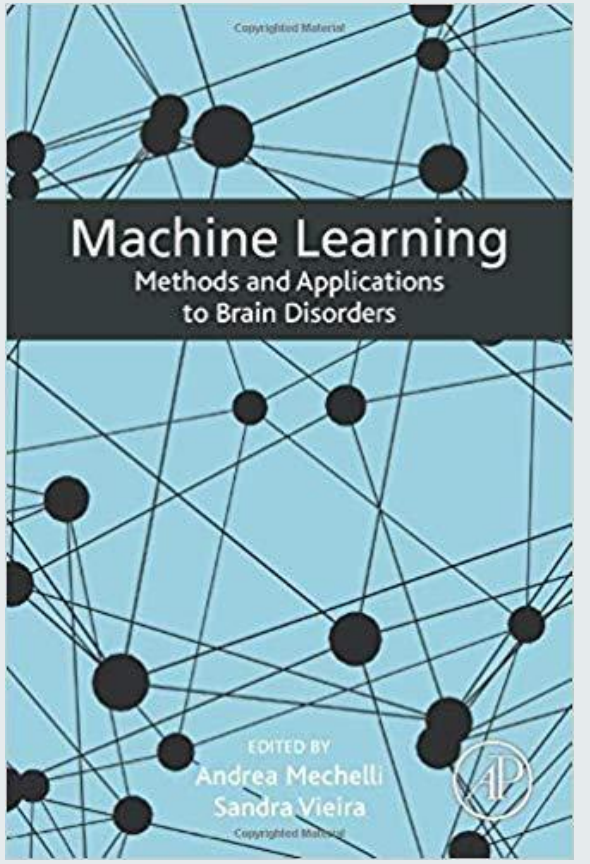
Next steps

Books

- https://mml-book.github.io/ *
- https://www.sciencedirect.com/book/9780128157398/machine-learning
- https://www.amazon.co.uk/Hands-Machine-Learning-Scikit-Learn-TensorFlow/dp/1492032646







References

https://machinelearningmastery.com/types-of-learning-in-machine-learning/

https://scikit-learn.org/stable/datasets/index.html

https://datasetsearch.research.google.com/

Thank you

Twitter: @Warvito
Linkedin: /walter-lopez-pinaya/