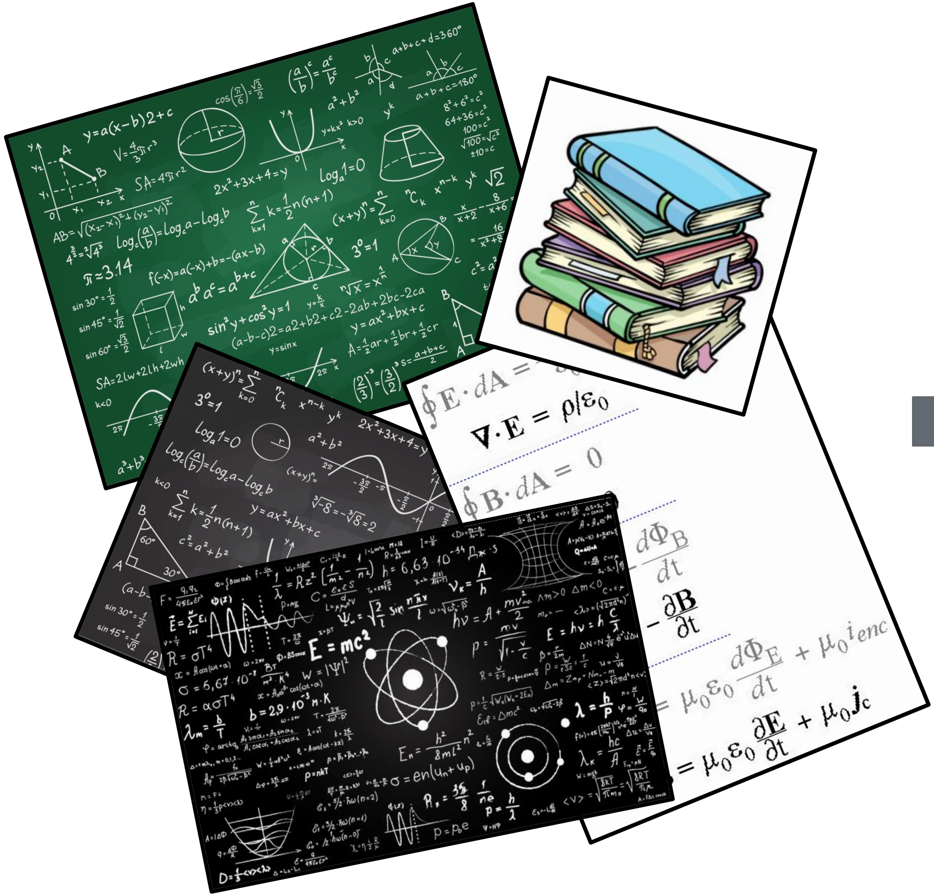
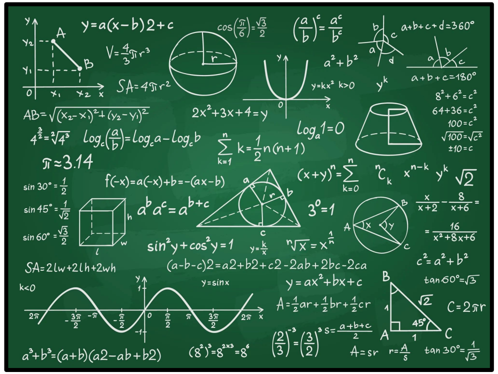
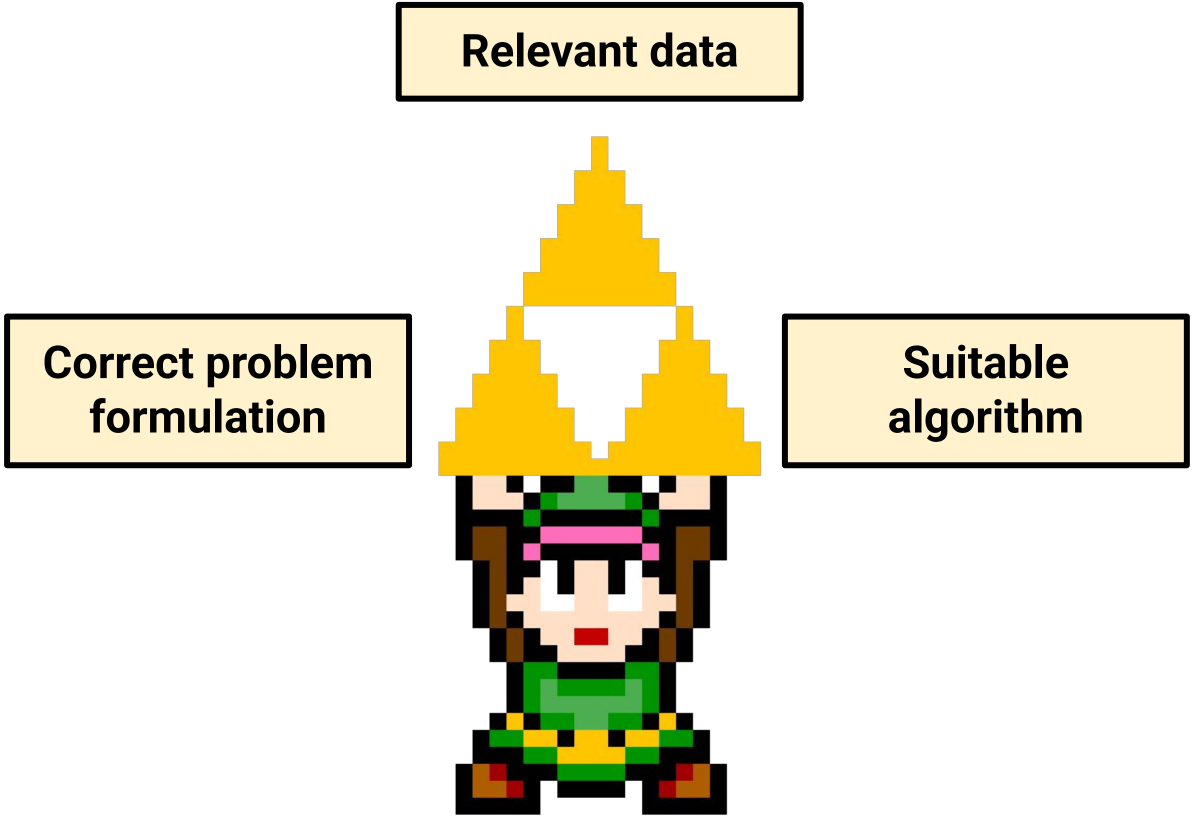


You, when trying to learn everything at once

Having more data to learn does not mean the learning will be more efficient



It is much healthier to focus on learning to do one thing very well, and to learn it the right way



Solving a problem requires the right problem formulation, and the right learning method

Machine Learning

Session 1 - Introduction to ML & Regression

hadrien.salem@centralelille.fr

[introduction-to-data-science]

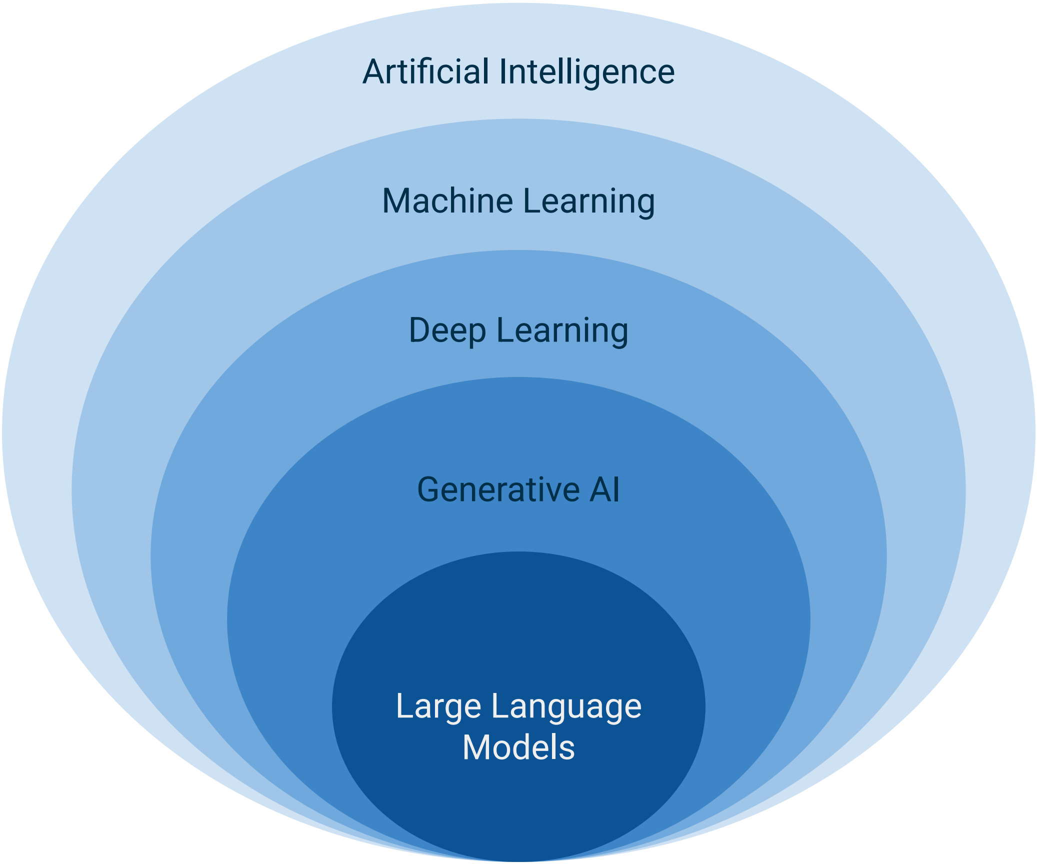
Introduction

What is Machine Learning and why should I care?

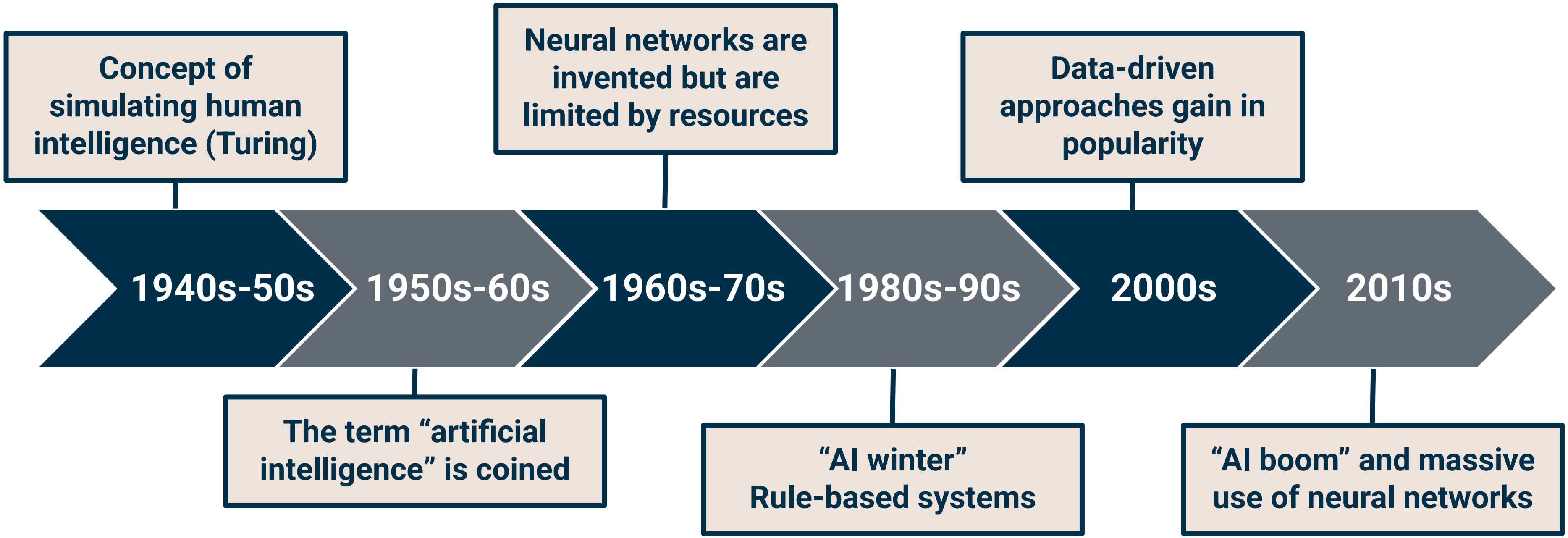
LEARNING

The process of acquiring new **understanding**, **knowledge**, **behaviors**, **skills**, **values**, **attitudes** and **preferences** through study, experience or being taught

⇒ Machines learn by observing patterns in data



Vocabulary reminders



A brief history of machine learning

Important dates in recent history

Since the 2000s, and especially since the 2010s, AI has been a very active field of research and is becoming very present in the industry

2006 - Introduction of “deep learning” and use of CNNs for image recognition

2012 - AlexNet (deep CNN using GPU acceleration) wins the ImageNet competition

2014 - GANs are introduced

2015 - AlphaGo defeats the world’s Go champion

2018 - NLP models based on deep learning (Transformers, BERT, etc.)

2022 - ChatGPT and large language models

What CAN machine learning do?

What machine learning CAN do

With recent advances, the range of tasks that can be performed by algorithms has increased

In general

- Image recognition and classification
- Natural Language Processing
- Recommandation systems
- Fraud detection
- Financial forecasting
- Language / Image generation
- Optimization (logistics, supply chain, etc.)

In healthcare

- Medical imagery analysis
- Disease diagnosis
- Disease prediction
- Personalized treatment plans
- Chatbots (e.g. symptom checkers)
- EHR analysis and processing

What CAN machine learning NOT do?

What machine learning CANNOT do

Algorithms cannot do everything, especially when human reasoning is required

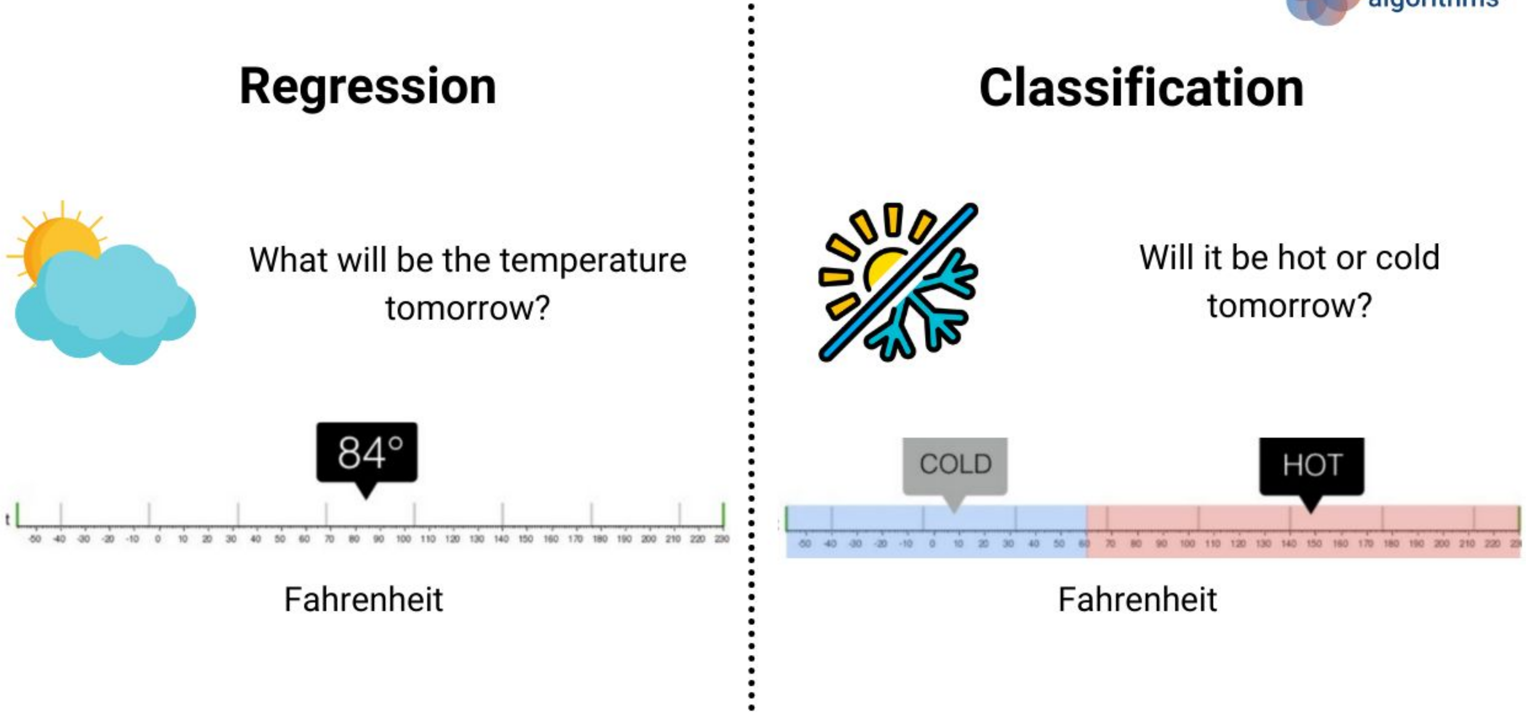
Algorithms typically struggle with…

- Common sense reasoning
- Creativity and abstraction
- Ethical and moral reasoning
- Understanding emotions
- Unstructured problem solving

Overview
What is there to study in order to train algorithms?

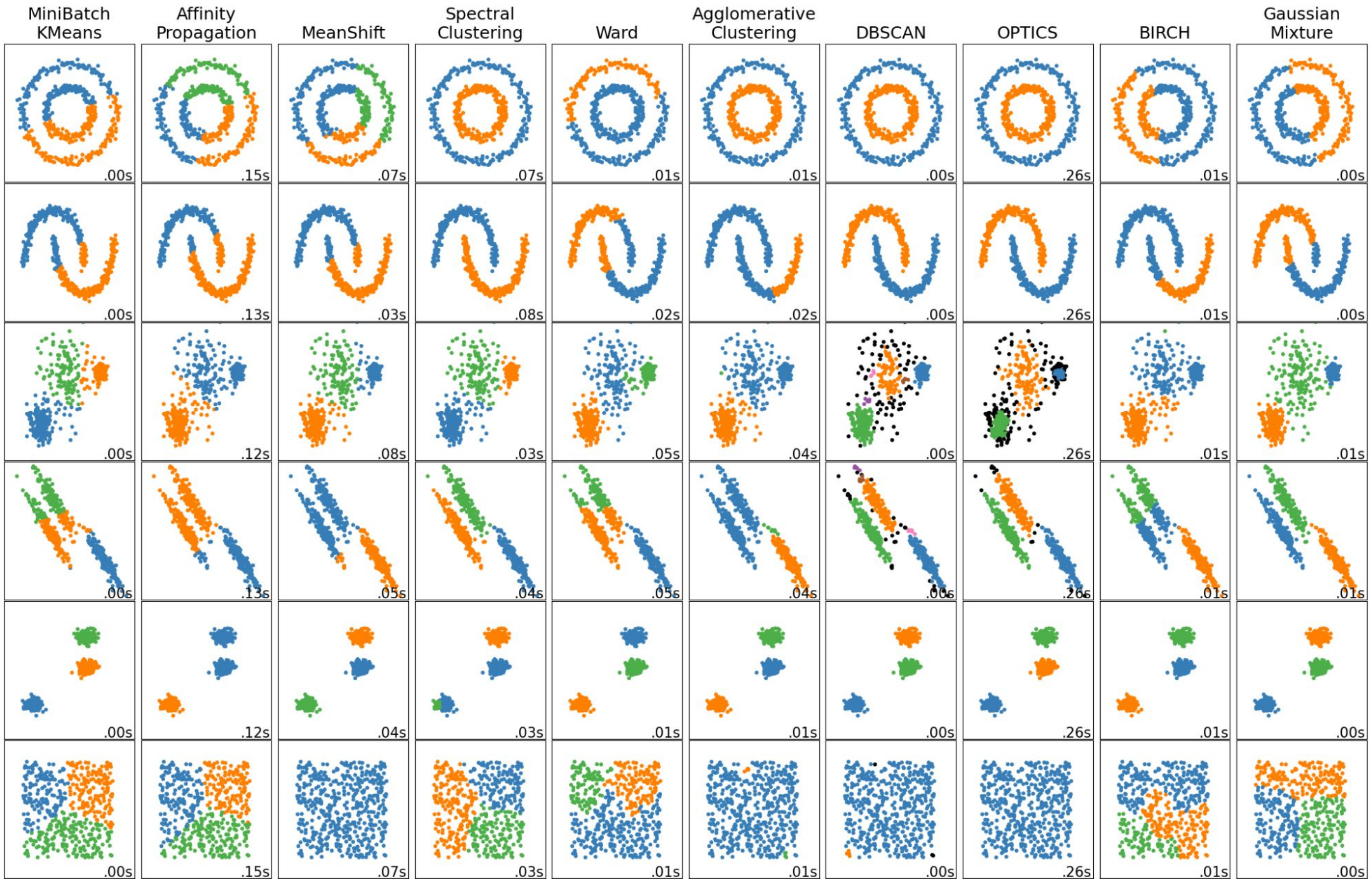
There are three main classes of machine learning problems

Regression, (Supervised) **Classification** and **Clustering**



Regression vs. Classification

Image source

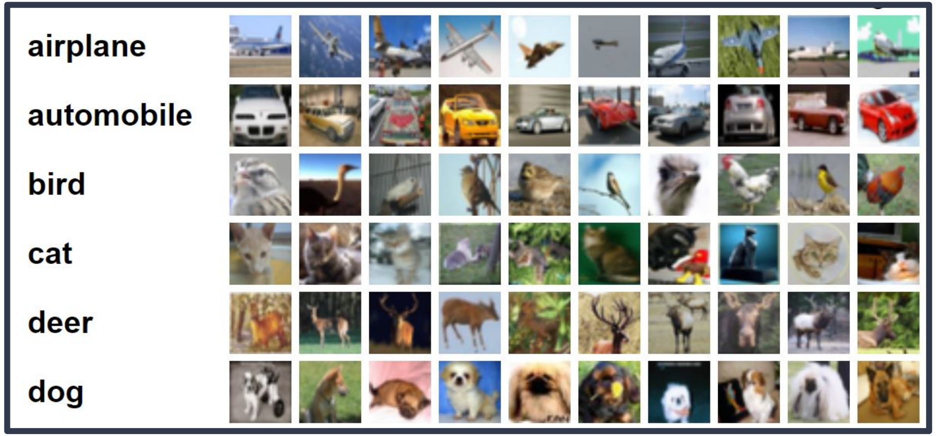


Clustering

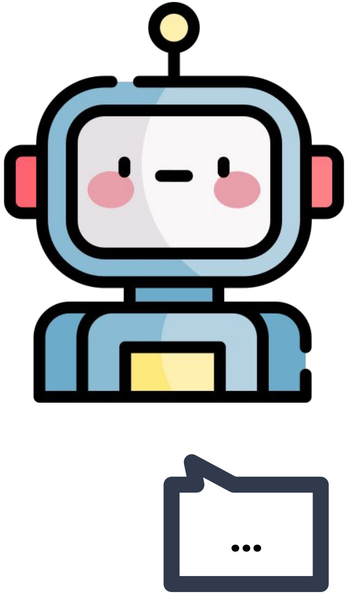
[Image source]()

There are three main classes of machine learning algorithms

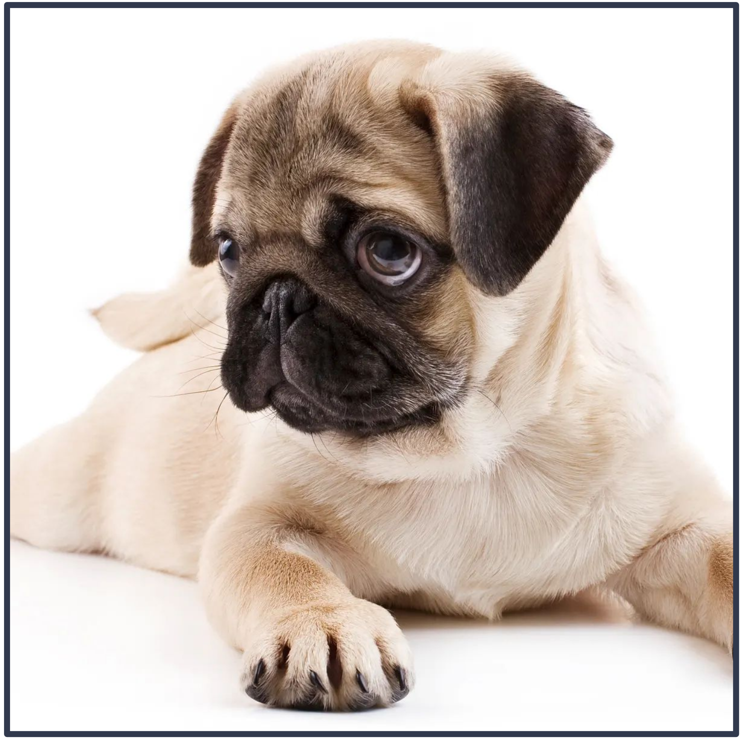
Supervised, **Unsupervised** and **Reinforcement** learning



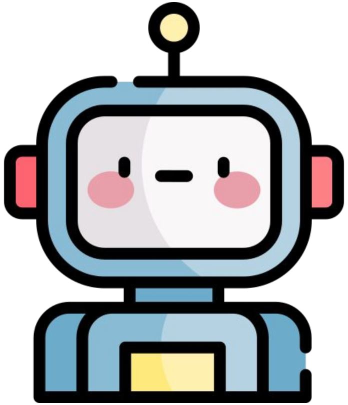
Learning



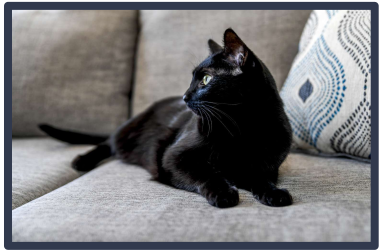
Supervised learning

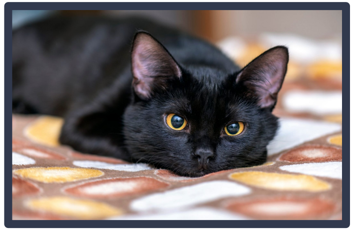


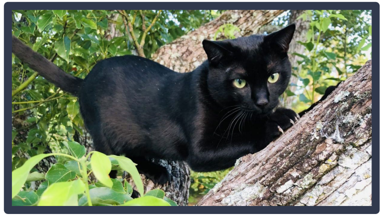
Unknown example



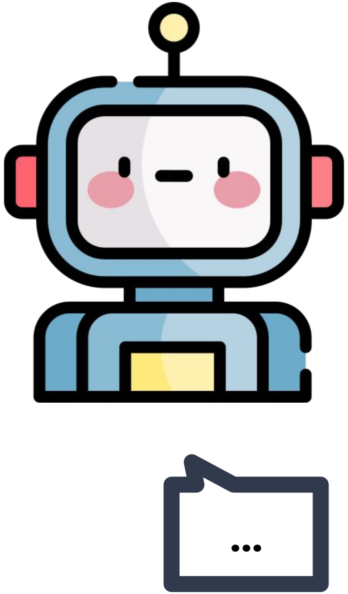
Dog



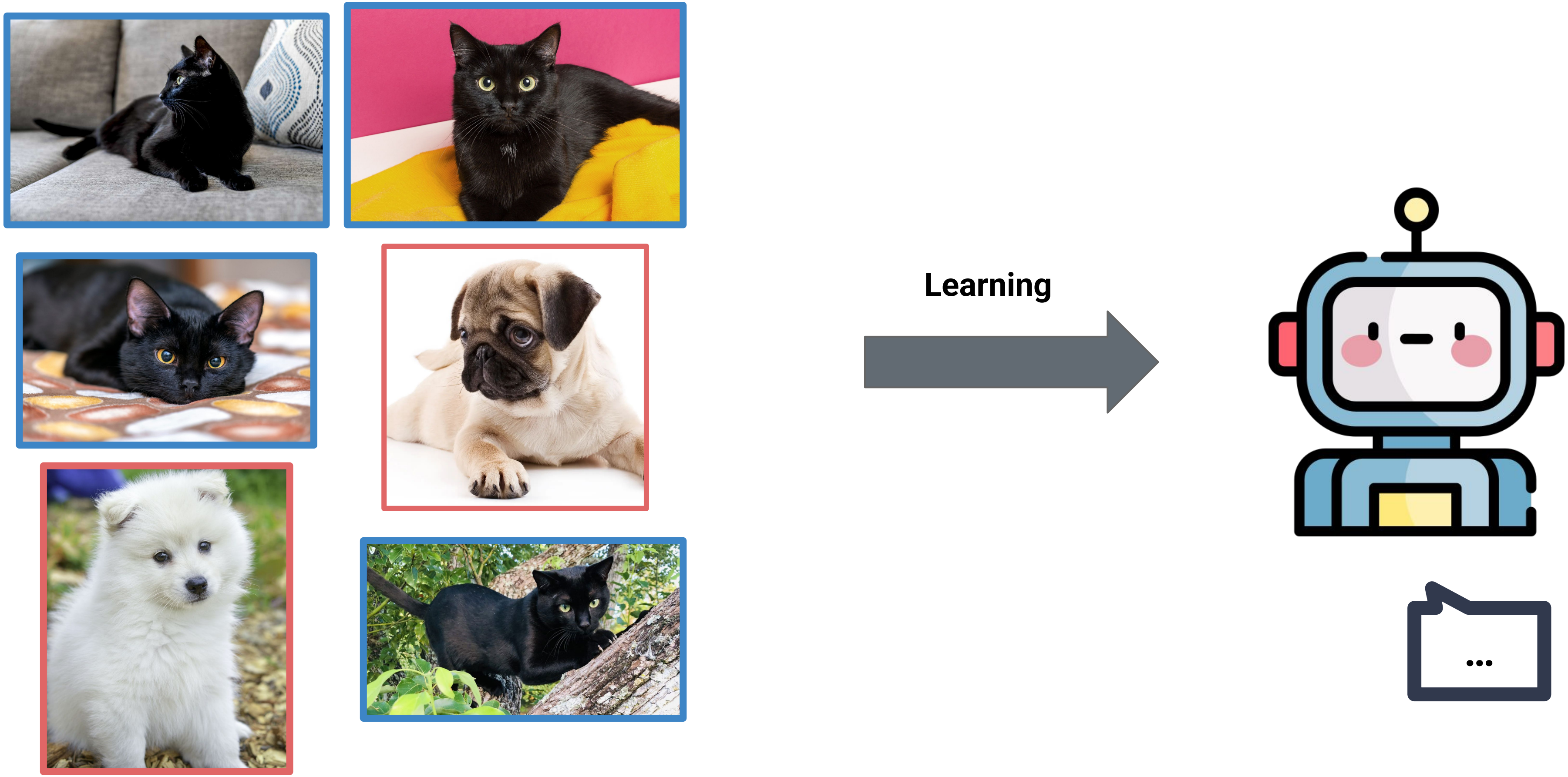




Learning

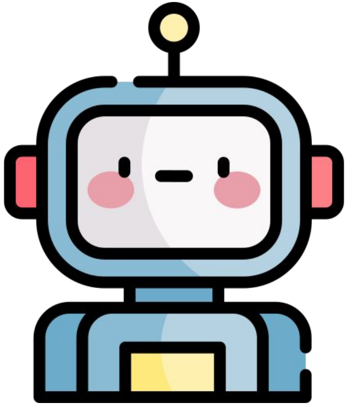


Unsupervised learning

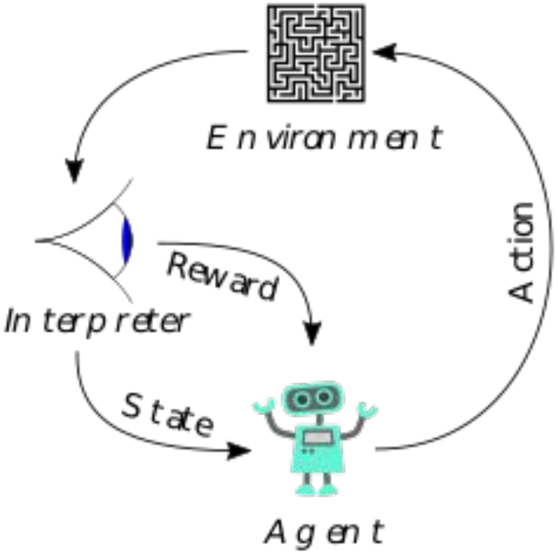


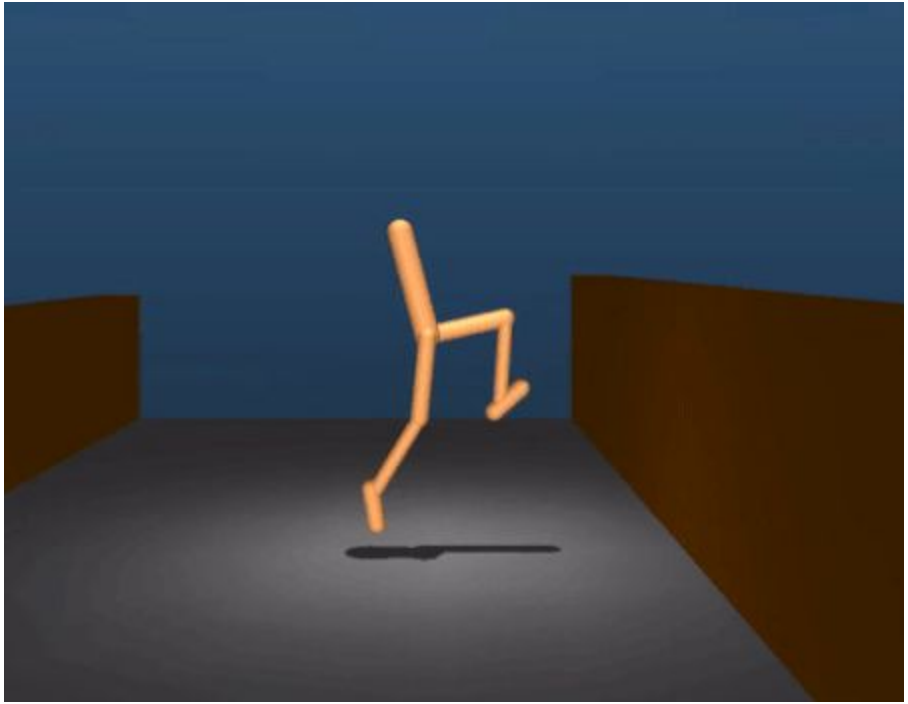
Unsupervised learning

Unknown example



Thing of group “red”





Reinforcement learning

In this course, we will learn what algorithms are fit for different types of problems

We will be presenting and using several algorithms, their strengths, weaknesses and use cases

Course outline

Intro to ML course

Session 1: Introduction to ML & Regression

Session 2: Supervised classification

Session 3: Neural networks

Workflow

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

Philosophy

In this course, the goal is to have an overview of the main categories of machine learning problems and methods.

hadrien.salem@centralelille.fr

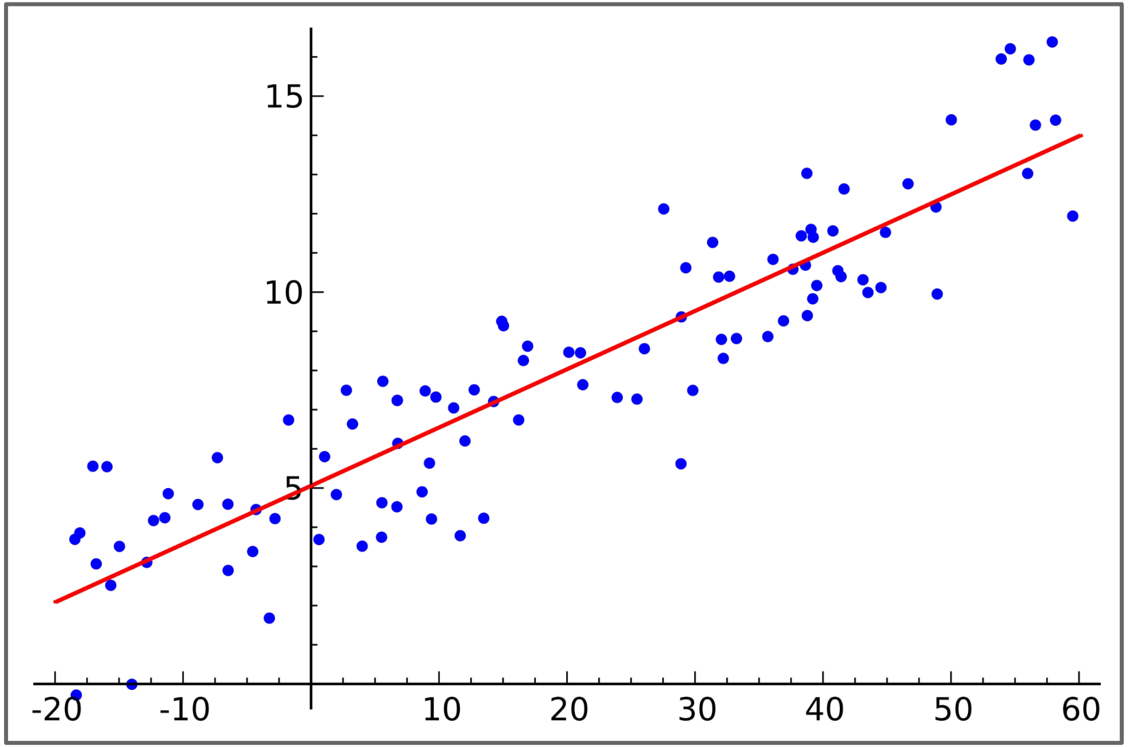
introduction-to-data-science

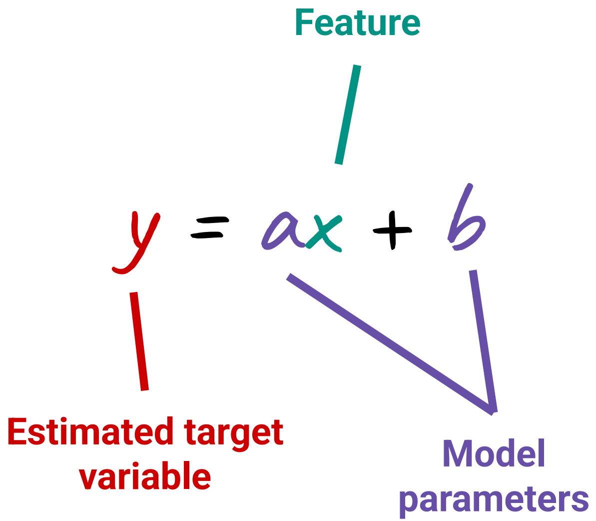
Regression

Introduction

Regression is a task consisting in **modeling** the relationship between input variables (**features**) and an output variable (**target**)

In Machine Learning, regression is the learning of the underlying relationship (function) between a number of features and a target variable





The goal is to find **a** and **b** such that the **estimation of the target variable** is **closest to reality**

Use cases for regression

Regression is the learning of a function modeling the relationship between features and a target value.

As such, it is used to study this relationship, or make inferences based on this model.

Prediction & Forecast

Understanding relationships

Time series modeling

$$y = w_0 + w_1 x_1 + \dots + w_p x_p = X.w$$

General formula for linear regression

Features

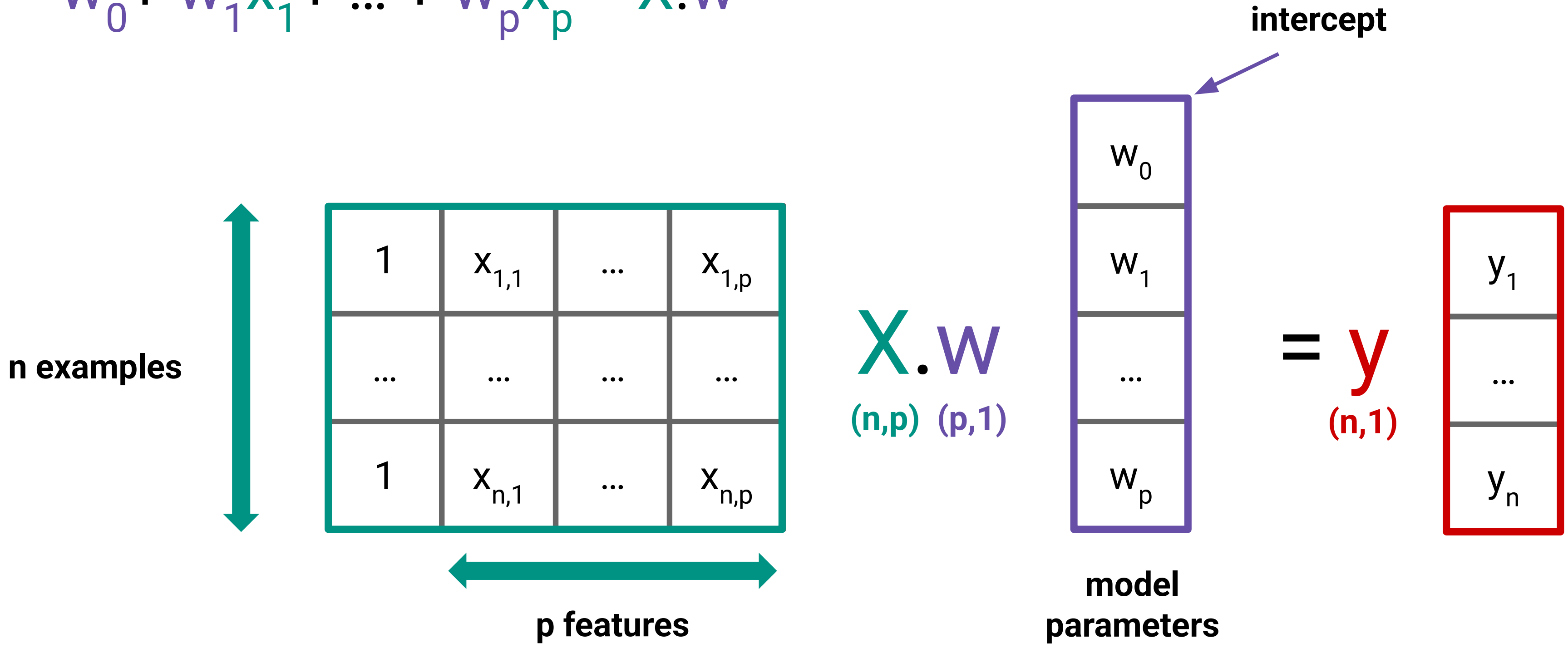
Estimated target variable

$$y = w_0 + w_1 x_1 + \dots + w_p x_p = X.w$$

Intercept
“Baseline effect” that is present even all features are at zero

Model parameters

$$y = w_0 + w_1 x_1 + \dots + w_p x_p = X.w$$



Linear regression with *p* features

$$y = w_0 + w_1 x_1 + \dots + w_p x_p = X.w$$

$$y \quad \begin{bmatrix} y_1 \\ \dots \\ y_n \end{bmatrix} \overset{?}{=} \begin{bmatrix} t_1 \\ \dots \\ t_n \end{bmatrix} \quad t$$

estimated values **target values**

The estimated values will never be exactly equal to the target values
We need a way to quantify how "close" they are

Fitting a regression model is finding the parameters such that the **sum of squared errors** is minimized

$$\boxed{w = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} |t_i - y_i(x_i, w)|^2}$$

Many algorithms exist to find the optimal parameters :
gradient descent, Newton's method, genetic algorithms, etc.

Once a model is fit, its performance can be evaluated by computing the **Mean Squared Error**

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |t_i - y_i|^2$$

Using regression in practice

Select a model
linear, polynomial, etc.

↓

Fit the model to part of the data
"training" set

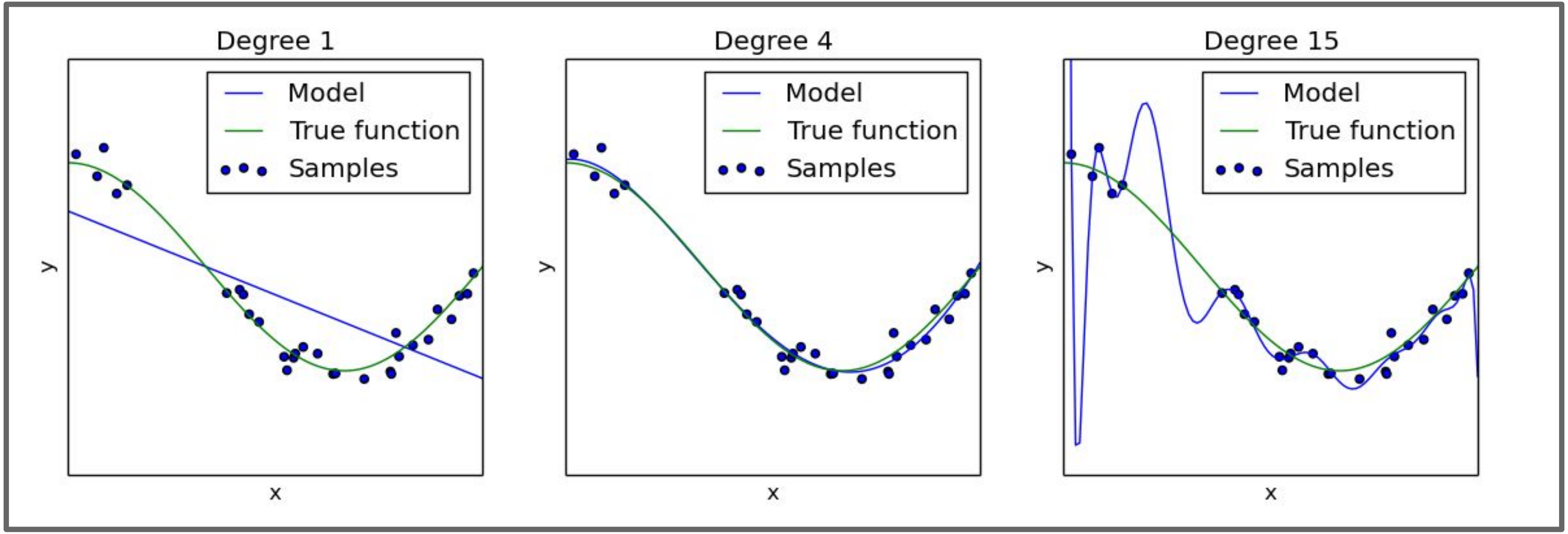
↓

Evaluate the model on unseen data
"test" set

Overfitting

Overfitting is “learning by heart”

It means your algorithm is **unable to generalize**



If the model is not complex enough, it cannot reproduce the patterns in the data
If the model is too complex, it will only reproduce the noise in the data

This is known as the **bias-variance trade-off**

Image source

$$w = \underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} |t_i - y_i(x_i, w)|^2 + \lambda ||w||$$

By adding an extra term to the cost function, we can penalize large values for w

This helps prevent overfitting

λ controls the strength of the regularization

Different norms have different properties

Practical work

The notebook contains all the necessary instructions

Debrief

Debrief

What did we learn today?

What could we have done better?

What are we doing next time?

Machine Learning

Session 1 - Introduction to ML & Regression

hadrien.salem@centralelille.fr
introduction-to-data-science