SD311 AML-ML

Jonathan Sprauel

What you will be evaluated on (i.e. what you will learn)

Technical skills :

- Hands on practice of all major algorithms (with **sklearn** and **keras**)
- Hands on practice of data analysis tools (**Jupyter**, **bokeh/plotly**, **pandas**)
- Key principles of all major algorithms
- Main bottlenecks of data driven approaches

Methodology skills :

- Use the correct vocabulary from the field
- Choose the correct class of algorithm for each problem
- General Knowledge of the history of the field
- Present the results to aid decision

Planning of the module

5 Oct.	11 Oct	12, 19 & 26 Oct	9 & 15 Nov.
8h30-11h45 : Vocabulary [0] Data Analysis [1,2]	Bayes, Regression and Gaussian processes [4,5,6]	*8h30-11h45 :* Ensemble method Boosting [8,9] Bagging & Random forest [11,12]	*9h30 - 11h30* Explainability [14]
13h15-16h30 : Supervised learning with SVM [3,4]	Surrogate Modelling [7]	XGboost practice [10]	8h30-11h45: Anomaly detection [13+ evaluation]

\+ 5 optional home exercices

Links

Courses notebooks :

https://github.com/erachelson/MLclass

https://supaerodatascience.github.io/

TP : https://github.com/jfabrice/ml-class-anomaly-detection

http://scikit-learn.org

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

https://www.kaggle.com/

https://www.datascienceweekly.org/

The different types of learning

Supervised Learning

- Learning with a **labeled** training set.

Learn with exercises
Ex. Driving license

Unsupervised Learning

- Discovering patterns in **unlabeled** data.

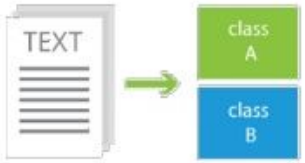
Learn with similitude
Ex. Newton and the apple

Reinforcement Learning

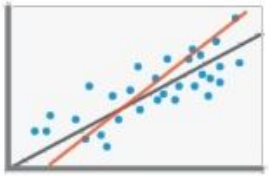
- Learning based on **feedback** or **reward**.

Learn with trial and error
Ex. Ride a bike

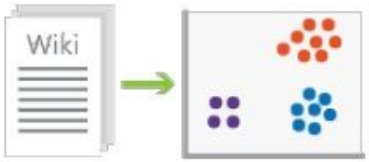
ML to solve different types of problems



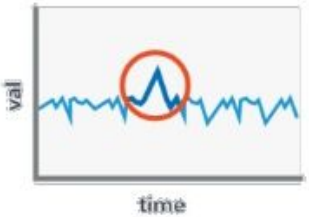
Classification
(supervised – predictive)



Regression
(supervised – predictive)



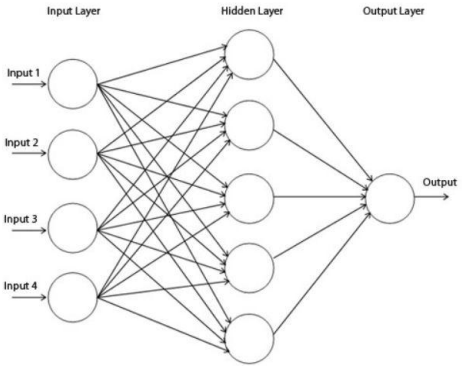
Clustering
(unsupervised – descriptive)



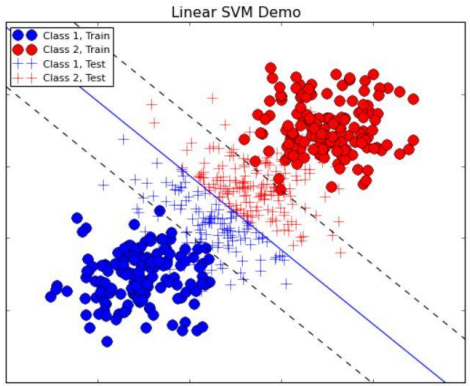
Anomaly Detection
(unsupervised – descriptive)

Classical Machine Learning

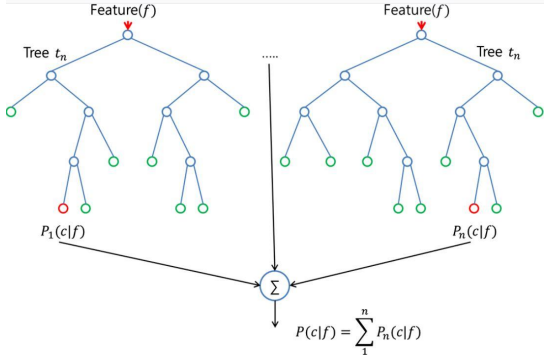
Multi-Layer Perceptron (1986)



SVM (1995)



Random forest (2001)



A brief history of Deep Learning

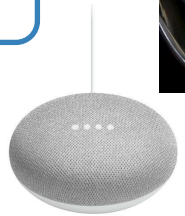
- **1981**
  - Fukushima Neocognitron
- **1988:**
  - Convolutional Network (**CNN**) de LeCun.
- **2011**
  - Traffic Signs Challenge : Performances above humans
- **2016**
  - Alphago wins against the best human champion at Go

MIT Technology Review

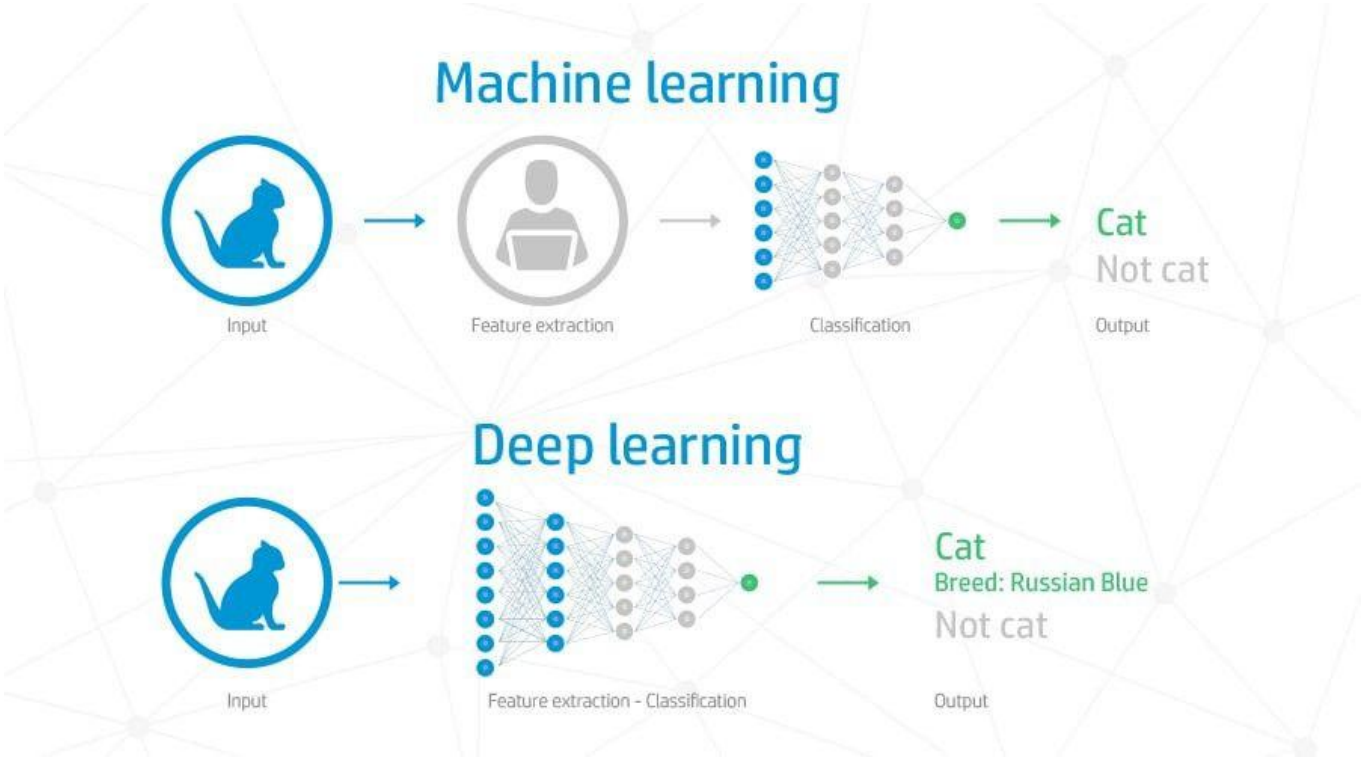
Facebook Launches Advanced AI Effort to Find Meaning in Your Posts

A technique called deep learning could help Facebook understand its users and their data better.

© reuters/ Kim Hong Ji

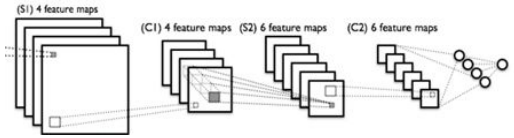


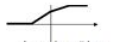
Machine Learning != Deep Learning != Artificial Intelligence



A brief history of what happened

New models:
convolutions,
subsampling, etc



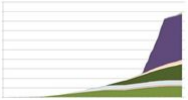


MLP 1986

Tricks on model architectures,
Learning methods

High performance computing,
GPUs for training models

Big Data : a lot of data available for training models



Open source community very active

Buzz cleverly orchestrated
(Google, Facebook, etc.)

Deep Learning

Exercice 1 : Regression

Objectives :

- Understand the difference between Regression and Classification

- Understand the definition of a Label

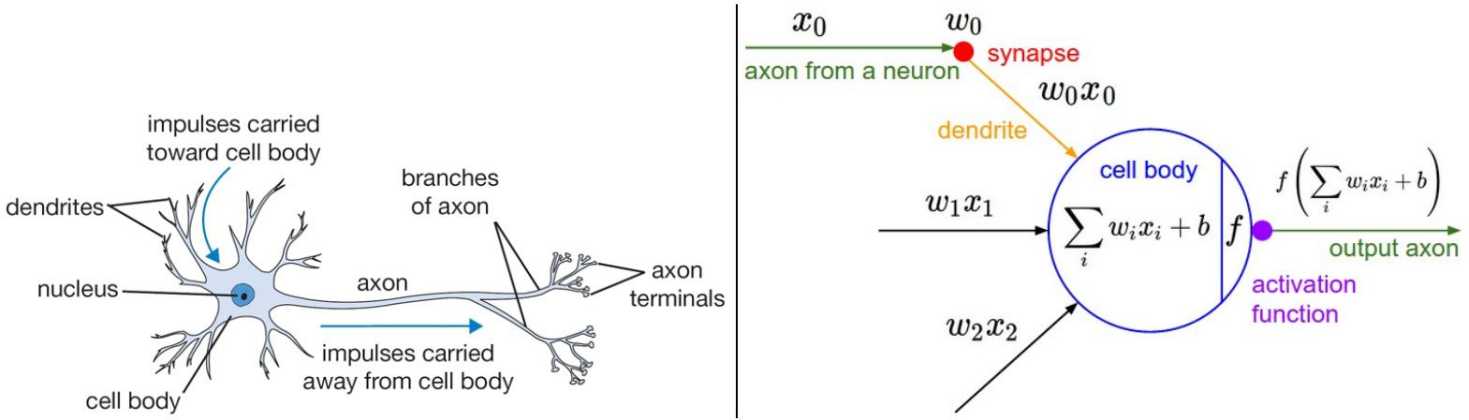
Exercice 2 : Features

Objectives :

- Understand the notion of Feature
- Understand the importance of Feature selection
- Understand how Deep learning changes the computation of Features

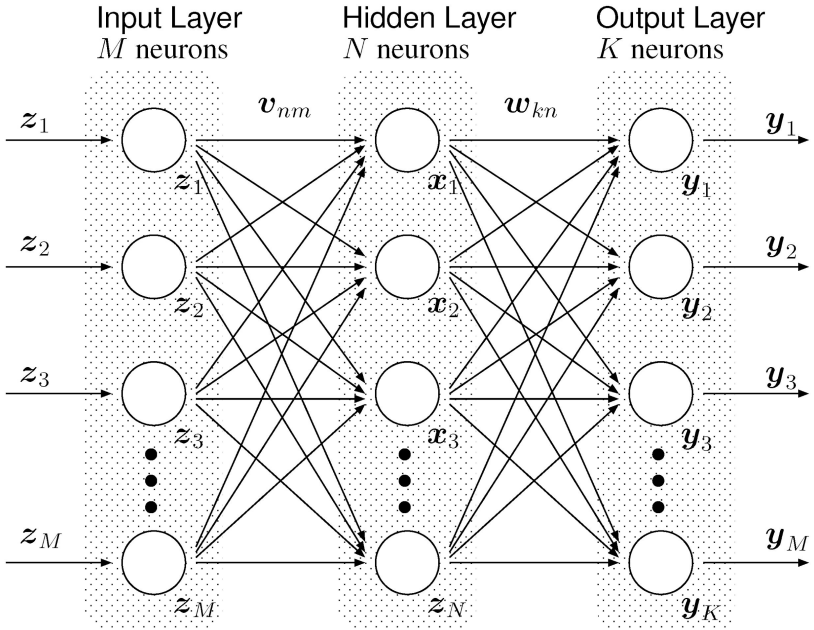
Neurons

- Neurons are trained to filter and detect features such as edges, shapes, textures, by receiving weighted inputs from the previous neurons, transforming it with an activation function and passing it to the outgoing connections.



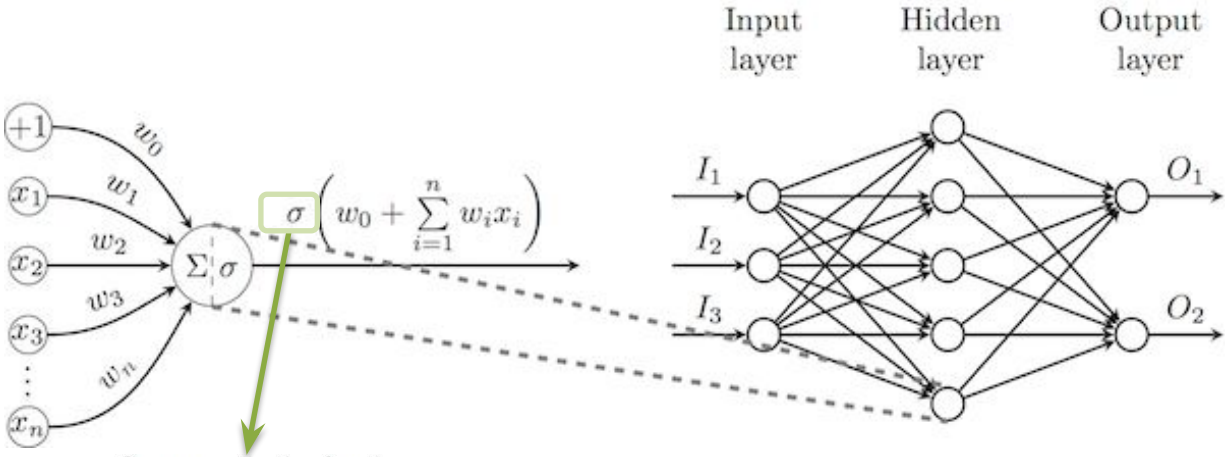
Multi-layer Perceptron (MLP)

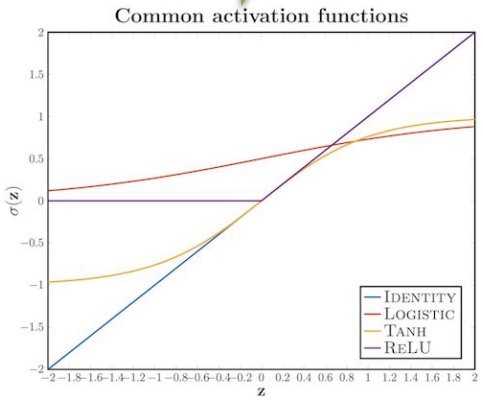
- MLP interest is in the association of neurons in multi layers : it results in a composition of non linear functions that can represent complex problematics.



- Parameters estimation:
- Quadratic error is known (estimated – known)² => we can estimate the gradient for the last layer
- We don't know the quadratic error associated to each hidden layer.

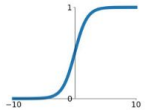
Activation Functions





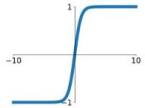
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



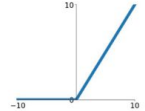
tanh

$\tanh(x)$



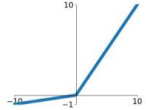
ReLU

$\max(0, x)$



Leaky ReLU

$\max(0.1x, x)$

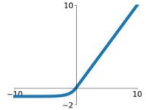


Maxout

$\max(w_1^T x + b_1, w_2^T x + b_2)$

ELU

$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



Exercice 3 : Neurones

Objectives :

- Understand the influence of hyper-parameters
- Reinforce the notion of Feature and the distinction between ML and DL

playground.tensorflow.org/

Quizz time : Fill in the definitions

Level 1	Machine Learning	Deep Learning	Artificial Intelligence	Big Data
Level 2	Supervised vs Unsupervised learning	Classification vs Regression	Correlation	Feature vs target
Level 3 [you are here]	Overfitting	Hyper parameter	Training vs Testing Dataset	Feature engineering
Level 4	ROC curve	Cross validation	Gradient descent	Bias vs Variance

Quizz time : Some answers

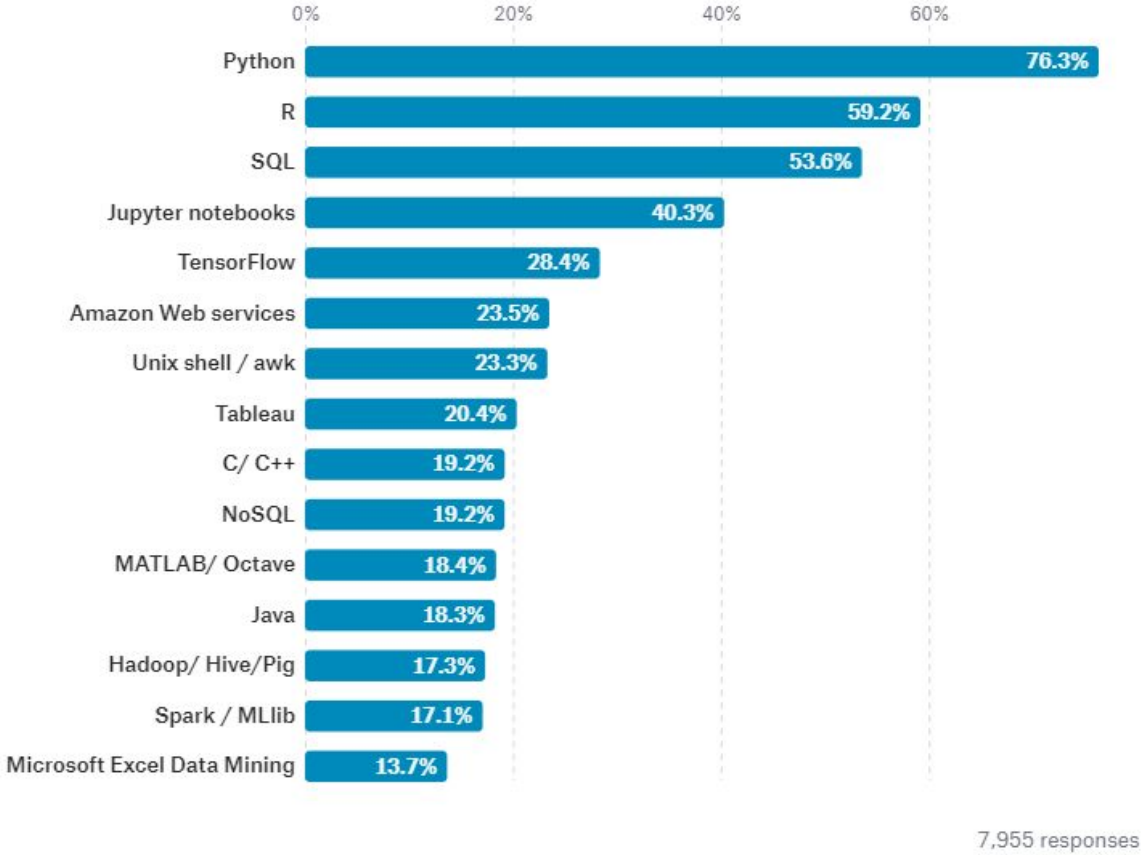
Machine learning is a field of computer science that gives computer systems the ability to “learn” (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed. (Wikipedia)

Artificial intelligence (AI) is the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings. (Brittanica)

Big Data refers to working with datasets that have large Volume,Variety, Velocity (, Veracity, and Value).

Deep Learning is Machine Learning with Deep Neural Networks.

Which tools are used



What should I look for in a data scientist's CV?

Must have :

- Technology names (most of them) : sklearn, python / R, keras / tensorflow, jupyter, numpy, pandas, spark
- Experiences with datasets outside of a MOOC
- Likes understanding people's problems

Nice to have :

- PhD (in computer science, applied math or physics)
- kaggle competition/score
- publications (Arxiv, JMLR, MLJ, IEEE PAMI, NIPS, ICML, ICLR…)
- cloud experience (AWS,GCP, Azure) or deployment experience (docker, terraform, kubernetes,...)

Sklearn : lets have a look

http://scikit-learn.org