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

2 Oct.	**8 Oct**	**15 Oct**	**5 Nov.**
8h30 - 11h45 : Vocabulary [0] Data Analysis [1,2]	*8h30 - 11h45* Bayes, Regression and Gaussian processes [5,6]	*8h30 - 11h45 :* Ensemble method Boosting [8,9]	*9h - 11h* Explainability [14]
	9 Oct	**22 & 23 Oct**	**6 Nov.**
13h30 - 16h45 : Supervised learning with SVM [3,4]	*8h30 - 11h45* Surrogate Modelling, Bayesian optim [7]	*8h30 - 11h45* XGboost practice [10] Bagging & Random forest [11,12]	*8h30 - 11h45* : Anomaly detection [13+ evaluation]

- 5 optional home exercices

Links

Courses notebooks :

https://supaerodatascience.github.io/

https://github.com/SupaeroDataScience/machine-learning/

http://scikit-learn.org

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

https://www.kaggle.com/

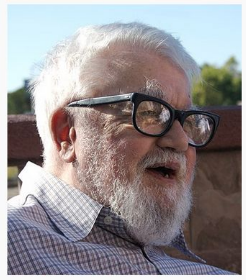
https://www.datascienceweekly.org/

A definition of AI ?

WIKIPÉDIA
L'encyclopédie libre

Artificial intelligence (AI) is a “set of theories and techniques implemented to create machines capable of simulating human intelligence”

“The construction of computer programs that perform tasks that are, for the moment, accomplished more satisfactorily by human beings”



John McCarthy
AI Pioneers with M.L Minsky

Programs that solve complicated tasks: those that are only accomplished today by humans

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.

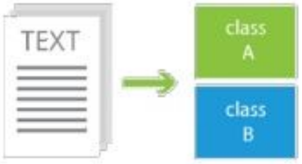
Learn with similitude
Ex. Newton and the apple

Reinforcement Learning

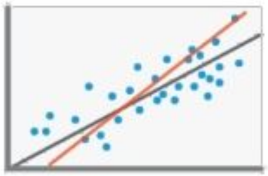
- Learning based on **feedback** or **reward**.

Learn with trial and error
Ex. Ride a bike

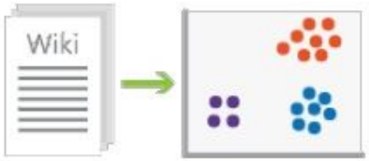
ML to solve different types of problems



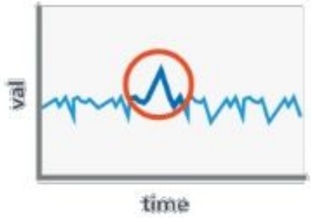
Classification
(supervised – predictive)



Regression
(supervised – predictive)



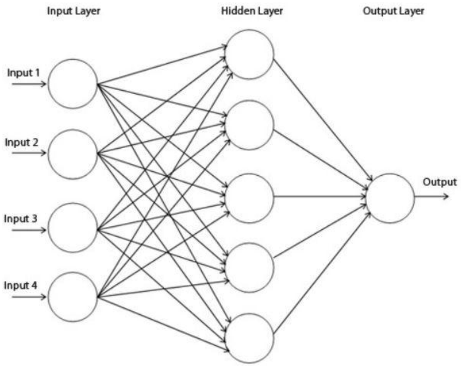
Clustering
(unsupervised – descriptive)



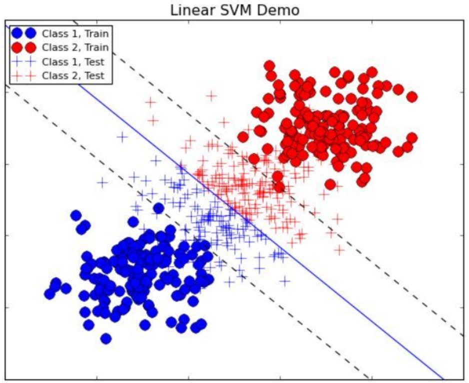
Anomaly Detection
(unsupervised – descriptive)

Classical Machine Learning

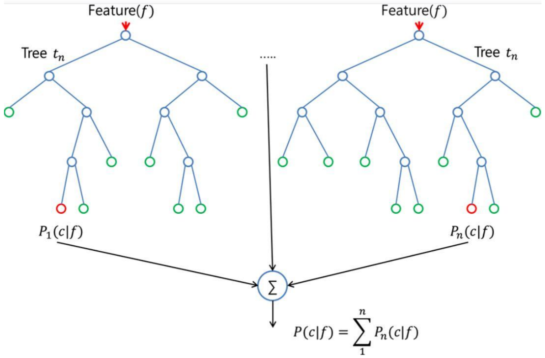
Multi-Layer Perceptron (1986)



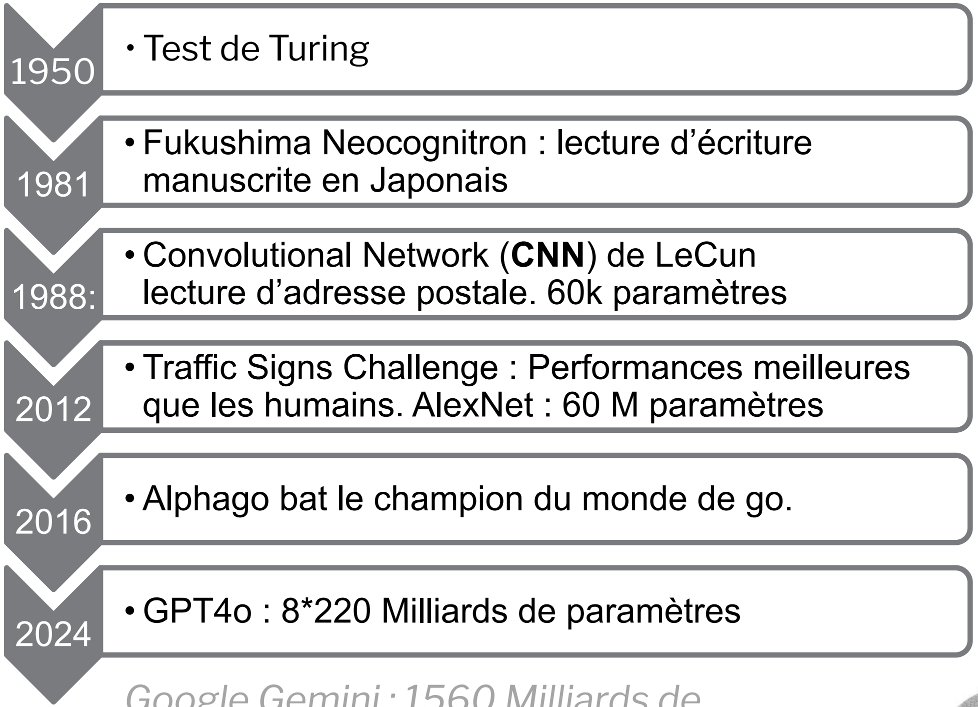
SVM (1995)



Random forest (2001)



A brief history of Deep Learning



1950
- Test de Turing

1981
- Fukushima Neocognitron : lecture d’écriture manuscrite en Japonais

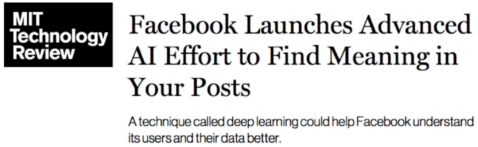
1988:
- Convolutional Network (**CNN**) de LeCun lecture d’adresse postale. 60k paramètres

2012
- Traffic Signs Challenge : Performances meilleures que les humains. AlexNet : 60 M paramètres

2016
- Alphago bat le champion du monde de go.

2024
- GPT4o : 8*220 Milliards de paramètres

Google Gemini : 1560 Milliards de paramètres



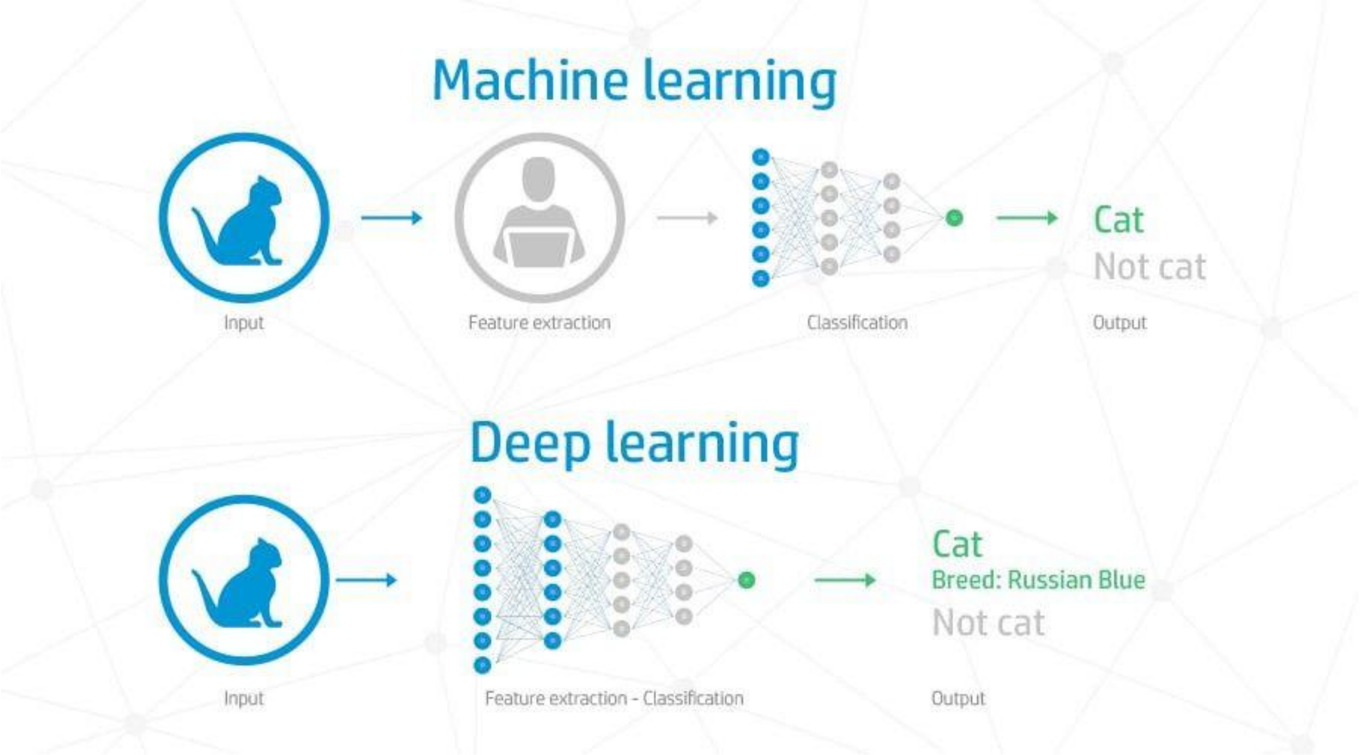
MIT Technology Review

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

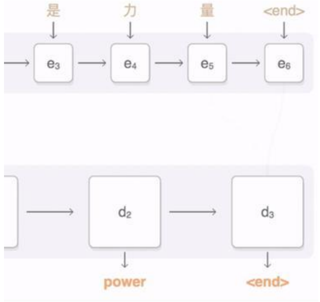


Solved Applications

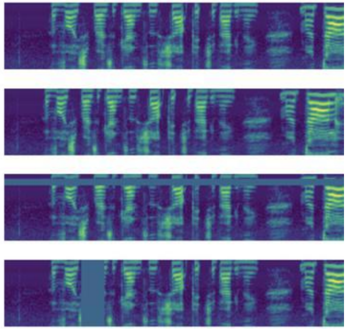
Image Classification :
92% on Image Net

Object Detection

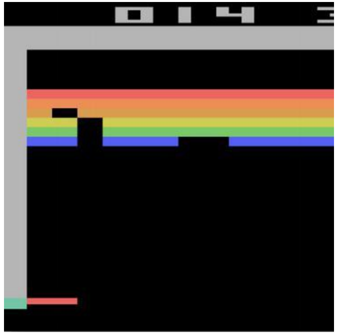
Sentiment analysis
(amazon, twitter, ...)
96% on IMDB



Machine Translation
BLEU score 40
(34 human pro)



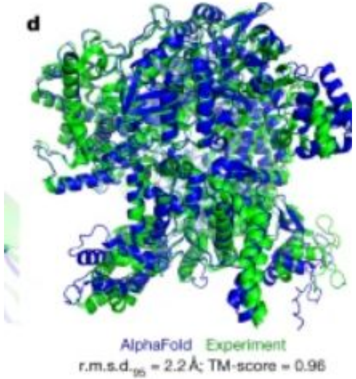
Speech Recognition
97% on Noisy



Atari, Chess, Go

Applications still under research

Image & Video Generation
Diffusion Models



Protein Prediction
>90% AlphaFold 2

Conversation agents (LLM)
67,6% Accuracy on US Medical exam

Multi agents games :
Starcraft, Diplomacy...

Exercice 1 : Regression

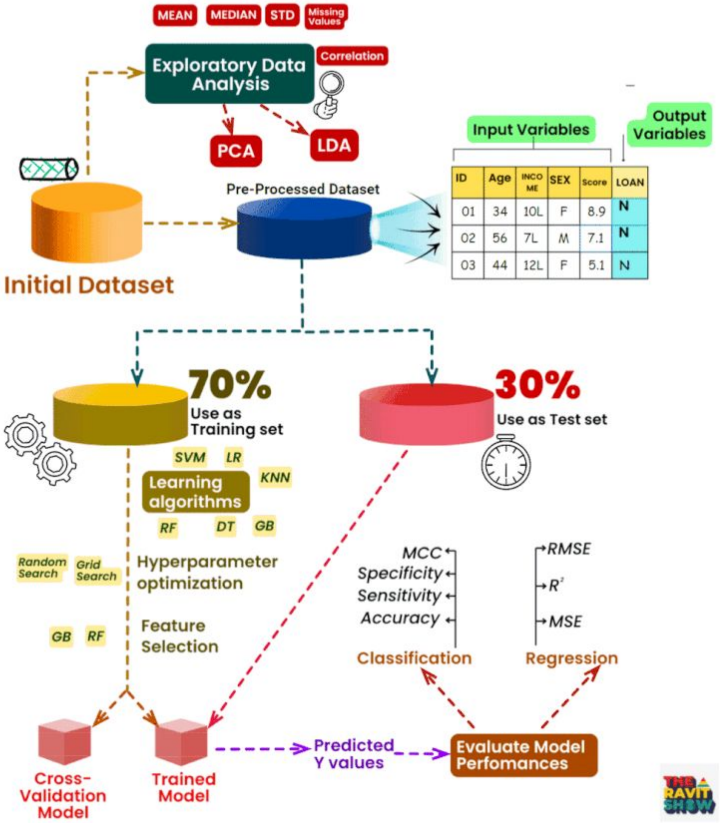
Objectives :

- Understand the difference between Regression and Classification

- Understand the definition of a Label

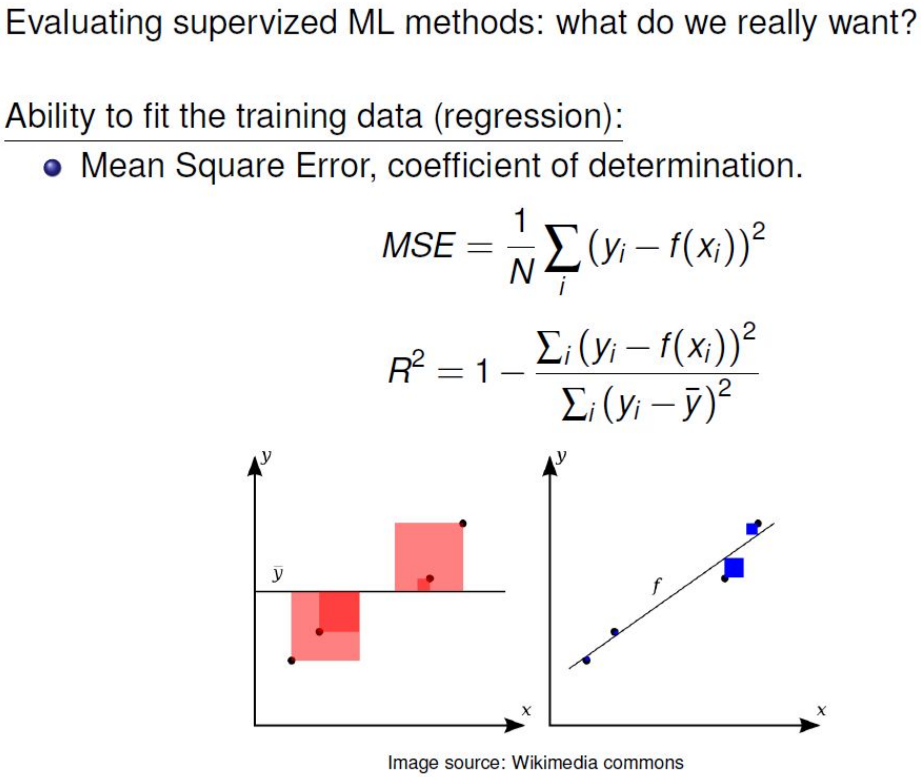
Working of Machine Learning Model

Don't Forget to Save For Later



ID	Age	INCOME	SEX	Score	LOAN
01	34	10L	F	8.9	N
02	56	7L	M	7.1	N
03	44	12L	F	5.1	N

Evaluation criterias



Evaluating supervized ML methods: what do we really want?

Ability to fit the training data (regression):

- Mean Square Error, coefficient of determination.

$$MSE = \frac{1}{N}\sum_i (y_i - f(x_i))^2$$

$$R^2 = 1 - \frac{\sum_i (y_i - f(x_i))^2}{\sum_i (y_i - \bar{y})^2}$$

Image source: Wikimedia commons

Evaluation criterias

Ability to fit the training data (classification):

- Accuracy, TP, FP, confusion matrix [link]

$$Acc = \frac{TP + TN}{TP + FP + TN + FN}$$

- ROC, AUC, [link]

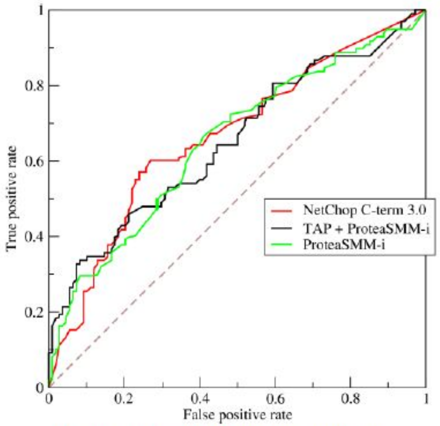


Image source: Wikimedia commons

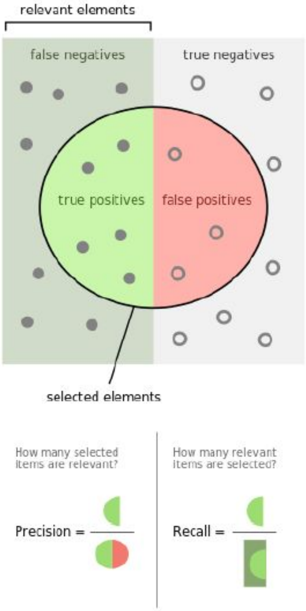
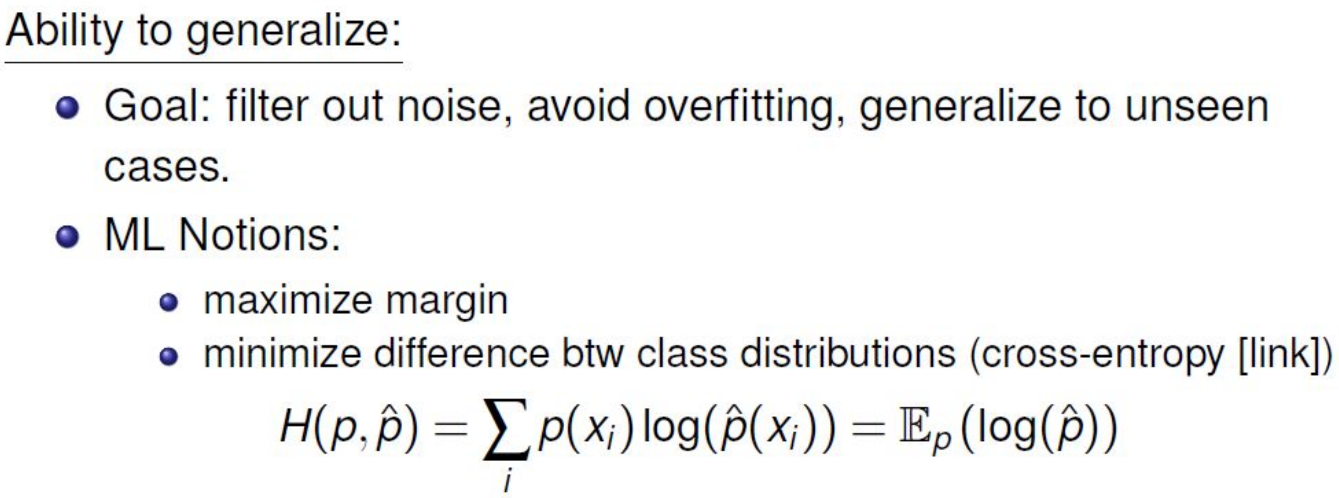


Image source: Wikimedia commons

Evaluation criterias



Ability to generalize:

- Goal: filter out noise, avoid overfitting, generalize to unseen cases.
- ML Notions:
  - maximize margin
  - minimize difference btw class distributions (cross-entropy [link])

$$H(p, \hat{p}) = \sum_i p(x_i)\log(\hat{p}(x_i)) = \mathbb{E}_p\left(\log(\hat{p})\right)$$

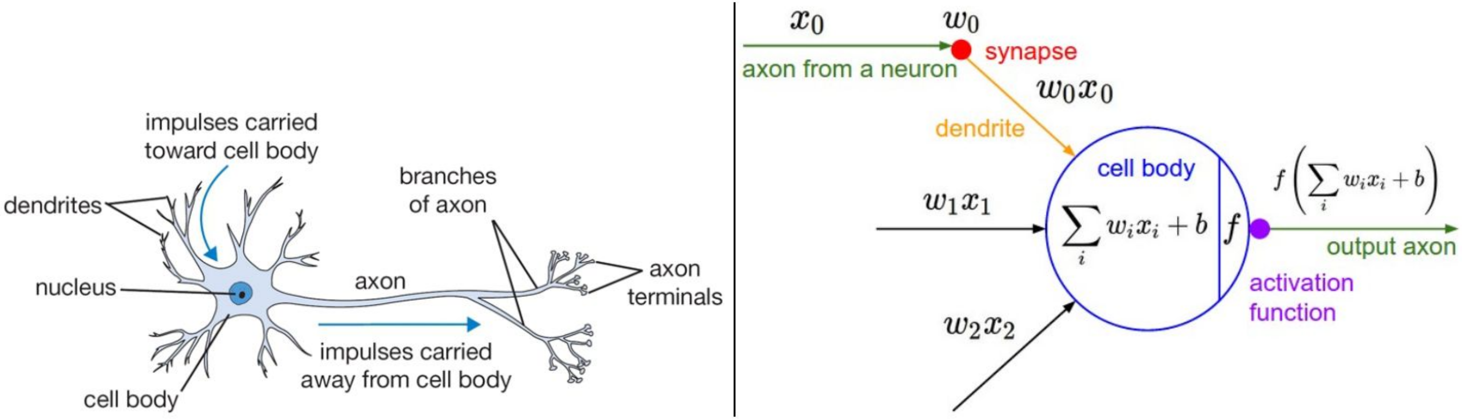
Exercice 2 : Features

Objectives :

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

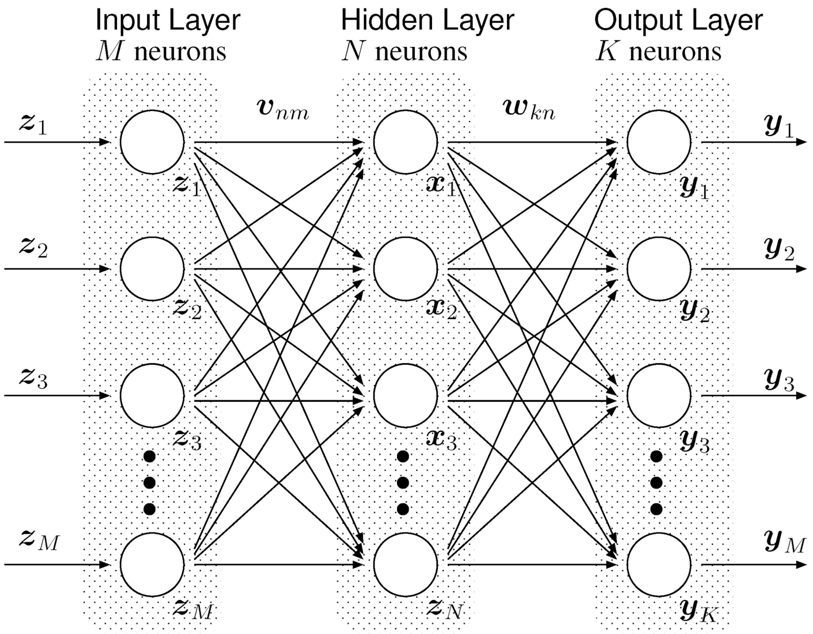
Neurons

- **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.

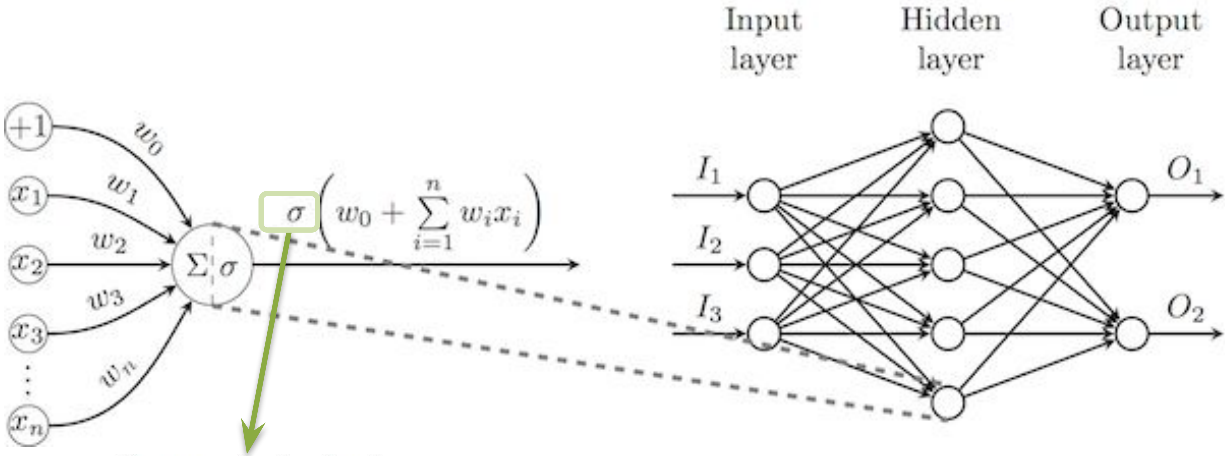


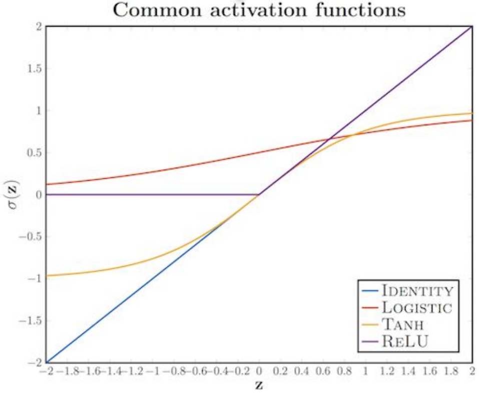
- Parameters estimation:

- Quadratic error is known (estimated – known)^2 => we can estimate the gradient for the last layer

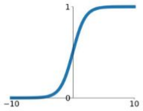
- We don’t know the quadratic error associated to each hidden layer.

Activation Functions

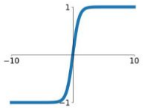




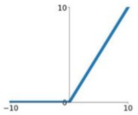
Sigmoid
$\sigma(x) = \frac{1}{1+e^{-x}}$



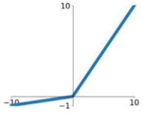
tanh
$\tanh(x)$



ReLU
$\max(0, x)$

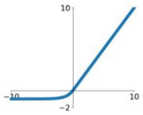


Leaky ReLU
$\max(0.1x, x)$



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

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



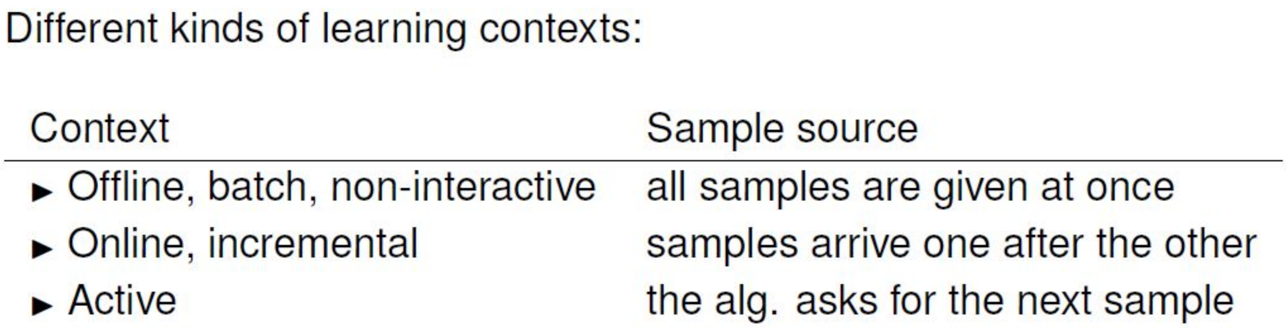
Exercice 3 : Neurones

Objectives :

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

playground.tensorflow.org/

Learning contexts



Different kinds of learning contexts:

Context	Sample source
► Offline, batch, non-interactive	all samples are given at once
► Online, incremental	samples arrive one after the other
► Active	the alg. asks for the next sample

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

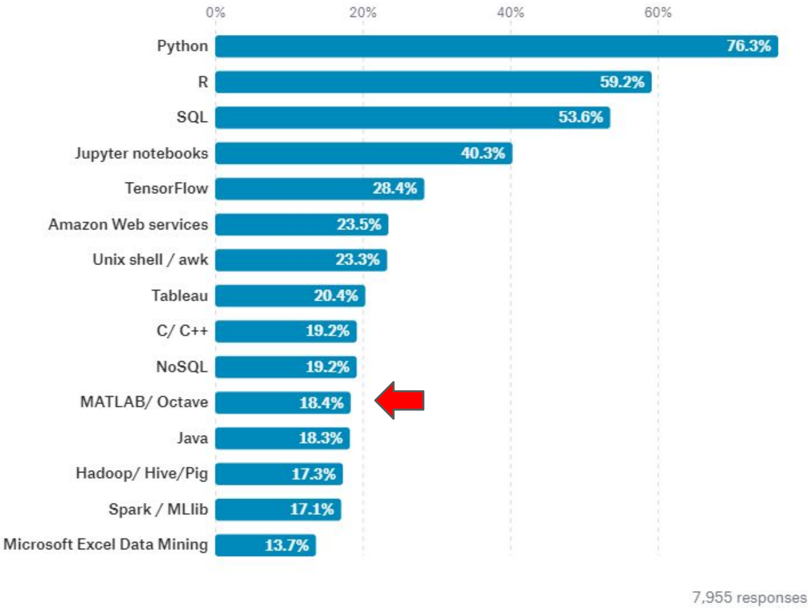
Machine learning is a field of computer science that gives computer systems the ability to “learn” (i.e. progressively improve performance on a specific task) with data, without being explicitly programmed. (Wikipedia)

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

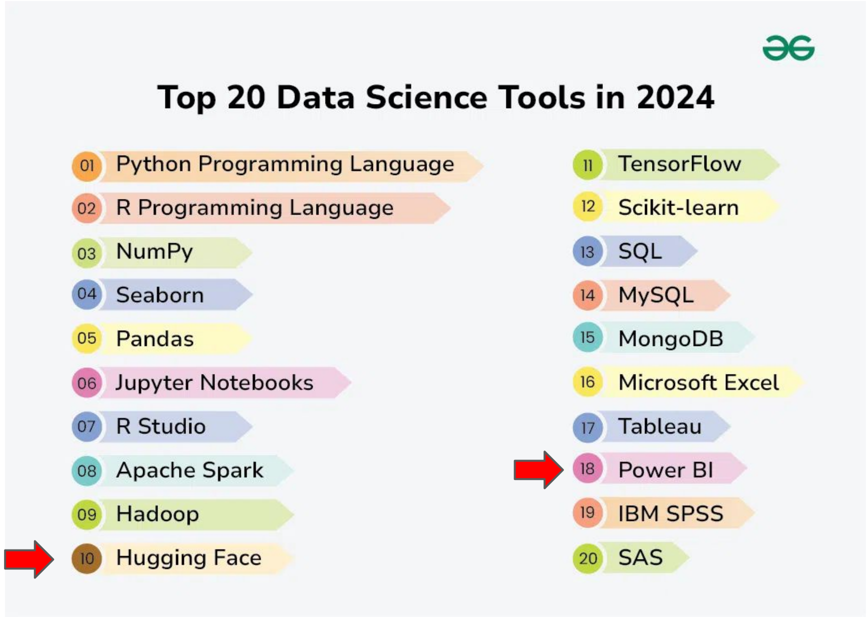
Big Data refers to working with datasets that have large Volume,Variety, Velocity (, Veracity, and Value).

Deep Learning is Machine Learning with Deep Neural Networks.

Which tools are used



2017

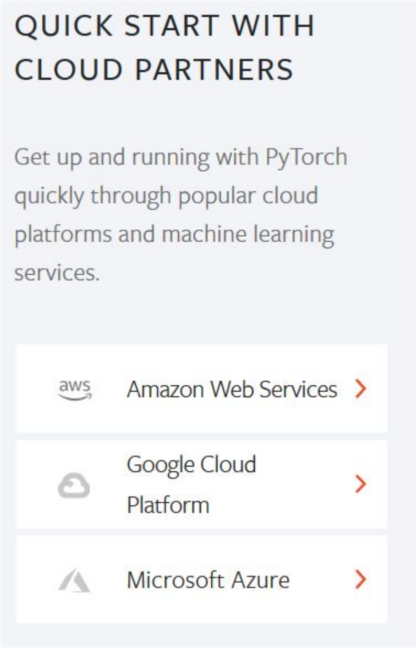


Top 20 Data Science Tools in 2024

01 Python Programming Language
02 R Programming Language
03 NumPy
04 Seaborn
05 Pandas
06 Jupyter Notebooks
07 R Studio
08 Apache Spark
09 Hadoop
10 Hugging Face
11 TensorFlow
12 Scikit-learn
13 SQL
14 MySQL
15 MongoDB
16 Microsoft Excel
17 Tableau
18 Power BI
19 IBM SPSS
20 SAS

2024

… and a lot of tools to put ML/DL in production



QUICK START WITH CLOUD PARTNERS

Get up and running with PyTorch quickly through popular cloud platforms and machine learning services.

https://pytorch.org/

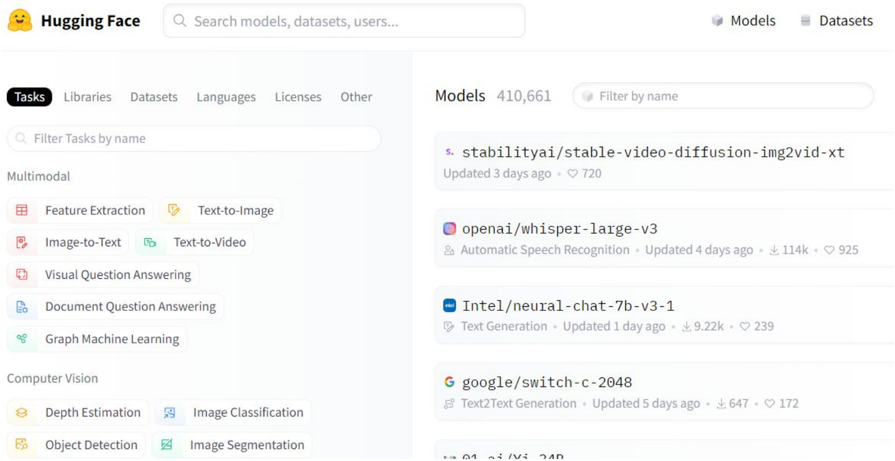
Cloud is the easiest option (by far…)
Optimized stacks for training and inference

But …
Exploding costs depending on the number of parameters

	Training cost
DeepMind AlphaGO	35 Million $
GPT3	12 Million $
CoAtNet (top 1 ImageNet)	250 000 $
BERT	7000 $
Yolo V5	100 $
ResNet 50	10 $

Start from a model already trained

Hugging Face / Pytorch hub : state of the art models with weight already tuned
=> we add images, continue training and voilà !



Hugging Face

Search models, datasets, users...

Models Datasets

Tasks Libraries Datasets Languages Licenses Other

Filter Tasks by name

Multimodal

- Feature Extraction
- Text-to-Image
- Image-to-Text
- Text-to-Video
- Visual Question Answering
- Document Question Answering
- Graph Machine Learning

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation

Models 410,661

Filter by name

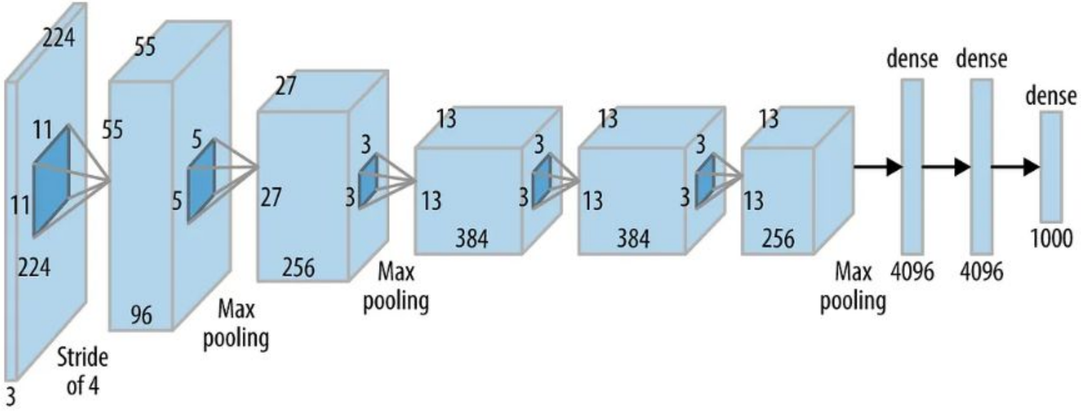
stabilityai/stable-video-diffusion-img2vid-xt
Updated 3 days ago • ♡ 720

openai/whisper-large-v3
Automatic Speech Recognition • Updated 4 days ago • ⤓ 114k • ♡ 925

Intel/neural-chat-7b-v3-1
Text Generation • Updated 1 day ago • ⤓ 9.22k • ♡ 239

google/switch-c-2048
Text2Text Generation • Updated 5 days ago • ⤓ 647 • ♡ 172

Articles and blogs describe architectures (how many layers, which types), which are known to work well on a given problem



Alexnet Block Diagram (source:oreilly.com)

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

Must have :

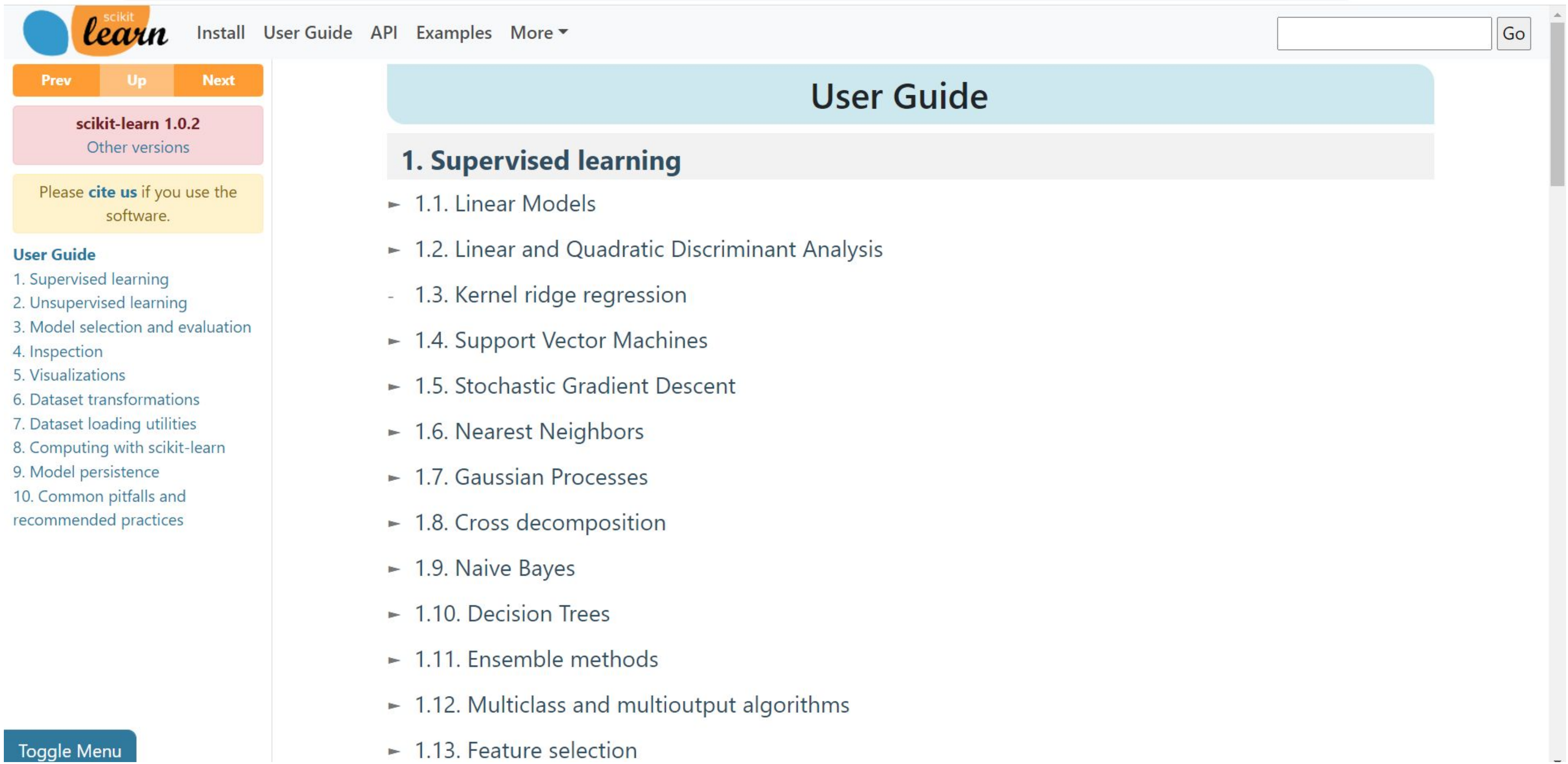
- Technology names (most of them) : sklearn, python / R, pytorch / keras / tensorflow, jupyter, numpy, pandas, spark
- Experiences with datasets outside of a MOOC
- Likes understanding people’s problems

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 : let’s have a look

http://scikit-learn.org



scikit-learn 1.0.2
Other versions

Please **cite us** if you use the software.

User Guide

1. Supervised learning
2. Unsupervised learning
3. Model selection and evaluation
4. Inspection
5. Visualizations
6. Dataset transformations
7. Dataset loading utilities
8. Computing with scikit-learn
9. Model persistence
10. Common pitfalls and recommended practices

User Guide

1. Supervised learning

- ► 1.1. Linear Models
- ► 1.2. Linear and Quadratic Discriminant Analysis
- - 1.3. Kernel ridge regression
- ► 1.4. Support Vector Machines
- ► 1.5. Stochastic Gradient Descent
- ► 1.6. Nearest Neighbors
- ► 1.7. Gaussian Processes
- ► 1.8. Cross decomposition
- ► 1.9. Naive Bayes
- ► 1.10. Decision Trees
- ► 1.11. Ensemble methods
- ► 1.12. Multiclass and multioutput algorithms
- ► 1.13. Feature selection