Machine Learning for Data Science

Philippe Besse, Sébastien Gerchinovitz, Béatrice Laurent

Machine Learning for Data Science
CERFACS – May 2019

Credits: Aurélien Garivier

Outline

Artificial Intelligence (AI): Definition

Intelligence exhibited by machines

- emulate cognitive capabilities of humans (big data: humans learn from abundant and diverse sources of data).
- a machine mimics "cognitive" functions that humans associate with other human minds, such as "learning" and "problem solving".

Ideal "intelligent" machine =

flexible rational agent that perceives its environment and takes actions that maximize its chance of success at some goal.

Founded on the claim that human intelligence

"can be so precisely described that a machine can be made to simulate it."

Artificial Intelligence: Tension

Operational goals

- Autonomous robots for not-too-specialized tasks
- In particular, vision + understand and produce language

Tension between operational and philosophical goals

- As machines become increasingly capable, facilities once thought to require intelligence are removed from the definition. For example, optical character recognition is no longer perceived as an exemplar of "artificial intelligence"; having become a routine technology.
- Capabilities still classified as AI include advanced Chess and Go systems and self-driving cars.

Machine Learning (ML): Definition

Arthur Samuel (1959)

Field of study that gives computers the ability to learn without being explicitly programmed

Tom M. Mitchell (1997)

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E.

ML: Learn from and make predictions on **data**

- Algorithms operate by building a model from **example** inputs in order to make data-driven **predictions or decisions**...
- ...rather than following strictly static program instructions: useful when designing and programming explicit algorithms is unfeasible or poorly efficient.

Within Data Analytics

- Machine Learning used to devise complex models and algorithms that lend themselves to **prediction** - in commercial use, this is known as *predictive analytics*.
- www.sas.com: "Produce reliable, repeatable decisions and results" and uncover "hidden insights" through learning from historical **relationships and trends** in the data.
- evolved from the study of pattern recognition and computational learning theory in artificial intelligence.

Machine Learning: Typical Problems

- spam filtering, text classification
- optical character recognition (OCR)
- search engines
- recommendation platforms
- speach recognition software
- computer vision
- bio-informatics, DNA analysis, medicine
- etc.

For each of these tasks, it is possible but very inefficient to write an explicit program reaching the prescribed goal.
It proves much more succesful to have a machine infer what the good decision rules are.

Related Fields

- **Computational Statistics**: focuses in prediction-making through the use of computers together with statistical models (ex: Bayesian methods).
- **Statistical Learning**: ML by statistical methods, with statistical point of view (probabilistic guarantees: consistency, oracle inequalities, minimax)
  $\rightarrow$ more focused on *correlation*, less on *causality*
- **Data Mining** (unsupervised learning) focuses more on exploratory data analysis: discovery of (previously) unknown properties in the data. This is the analysis step of Knowledge Discovery in Databases.
- Importance of **probability**- and **statistics**-based methods $\rightarrow$ **Data Science** (Michael Jordan)
- Strong ties to **Mathematical Optimization**, which delivers methods, theory and application domains to the field

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]

Qu'est-ce qu'une (très grande) masse de données ?

Big Data

VLDB

XLDB

Very Big Data

Data Deluge

Massive Data

Data Masses

Data inflation 2

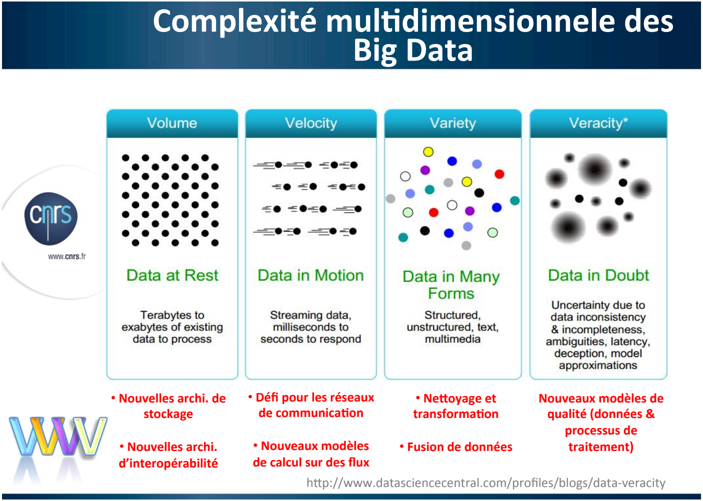
Unit	Size	What it means
Bit (b)	1 or 0	Short for "binary digit", after the binary code (1 or 0) computers use to store and process data
Byte (B)	8 bits	Enough information to create an English letter or number in computer code. It is the basic unit of computing
Kilobyte (KB)	1,000, or 2^{10}, bytes	From "thousand" in Greek. One page of typed text is 2KB
Megabyte (MB)	1,000KB; 2^{20} bytes	From "large" in Greek. The complete works of Shakespeare total 5MB. A typical pop song is about 4MB
Gigabyte (GB)	1,000MB; 2^{30} bytes	From "giant" in Greek. A two-hour film can be compressed into 1-2GB
Terabyte (TB)	1,000GB; 2^{40} bytes	From "monster" in Greek. All the catalogued books in America's Library of Congress total 15TB
Petabyte (PB)	1,000TB; 2^{50} bytes	All letters delivered by America's postal service this year will amount to around 5PB. Google processes around 1PB every hour
Exabyte (EB)	1,000PB; 2^{60} bytes	Equivalent to 10 billion copies of *The Economist*
Zettabyte (ZB)	1,000EB; 2^{70} bytes	The total amount of information in existence this year is forecast to be around 1.2ZB
Yottabyte (YB)	1,000ZB; 2^{80} bytes	Currently too big to imagine

The prefixes are set by an intergovernmental group, the International Bureau of Weights and Measures. Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

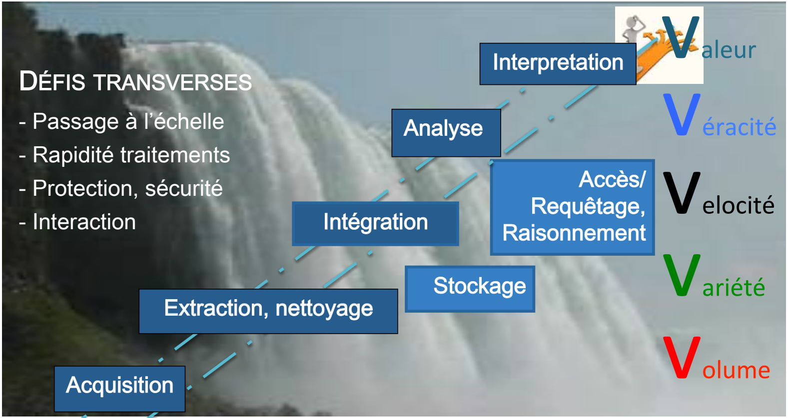
Source: *The Economist*

Grandes Conf du domaine: VLDB, XLDB, ICDE, EDBT, …

[Src: Bouzeghoub, *Mastodons: Une approche interdisciplinaire des Big Data*]



Défis accompagnant les chgts



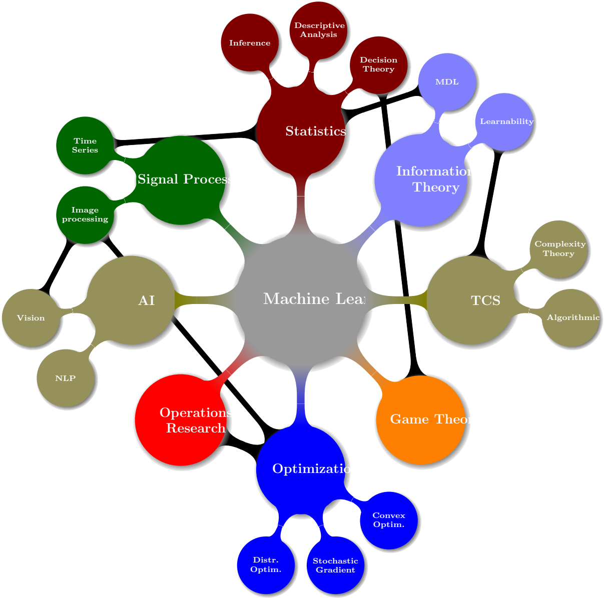
repenser les outils algorithmiques et mathématiques

inspired by "Big Data and Its Technical Challenges, Communications of the ACM, July 2014, vol 57, n°7", © H.V. Jagadish et all.

Machine Learning and Statistics

- Data analysis (inference, description) is the goal of statistics for long.
- Machine Learning has more **operational** goals (ex: consistency is important in the statistics literature, but often makes little sense in ML).
  Models (if any) are *instrumental*.
  Ex: linear model (nice mathematical theory) vs Random Forests.
- Machine Learning/big data: no separation between statistical modelling and optimization (in contrast to the statistics tradition).
- In ML, data is often here before (unfortunately)
- No clear separation (statistics evolves as well).

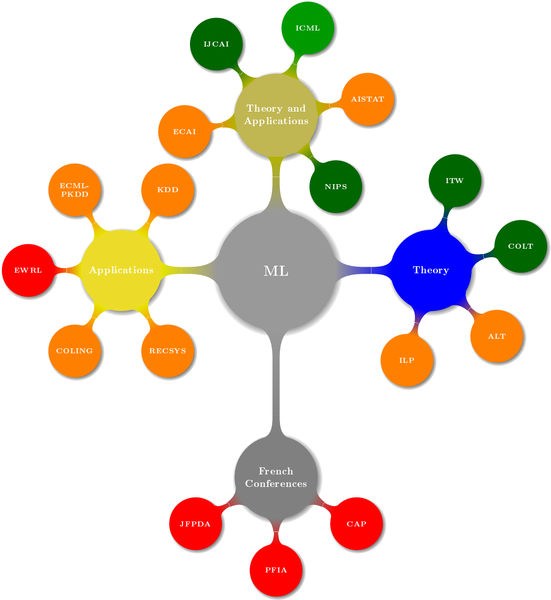
ML and its neighbors



ML journals

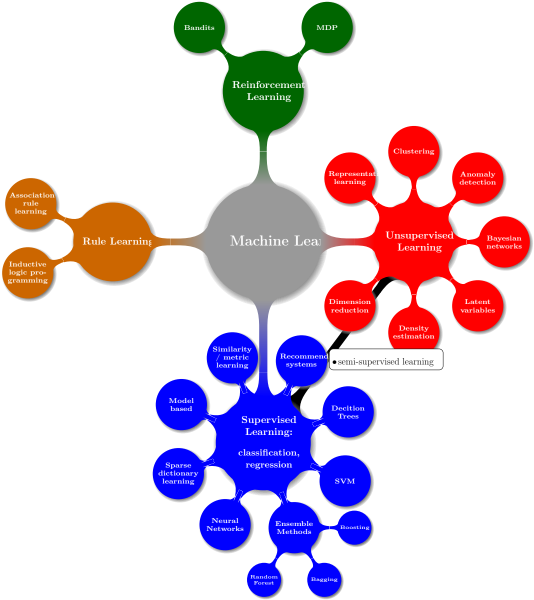

Machine Learning
ML & statistics
JMLR
Machine Learning
Artificial Intelligence
Mathematical Statistics and Learning
Statistics Theory
Annals of Statistics
Bernoulli
IEEE-IT
Electronic Journal of Statistics
ESAIM-PS
Computer science
TCS
Optimization
Optimization series of Imperial College Press
Journal of Optimization Theory and Applications
SIAM journal on optimization
Mathematics of Operations Research

ML conferences



Outline

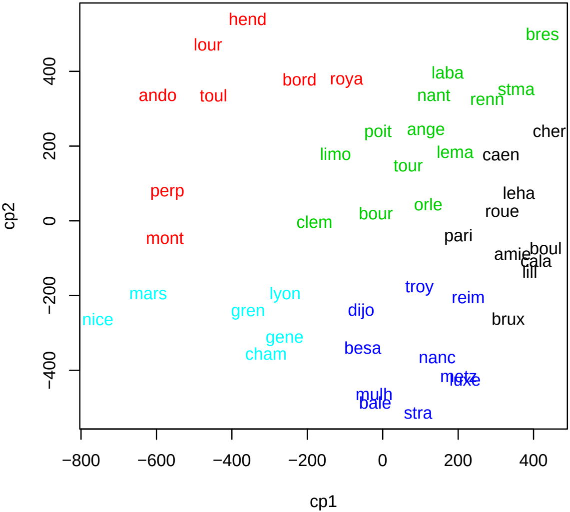
What ML is composed of



Unsupervised Learning

- (many) observations on (many) individuals
- need to have a simplified, structured overview of the data
- *taxonomy*: untargeted search for *homogeneous clusters* emerging from the data
- Examples:
  - customer segmentation
  - image analysis (recognizing different zones)
  - exploration of data

Example


hend
lour
bres
laba
bord
roya
ando
toul
nant
renn
stma
poit
ange
cher
limo
lema
caen
tour
perp
leha
orle
bour
roue
clem
pari
mont
amie
boul
cala
lill
troy
mars
lyon
reim
gren
dijo
nice
brux
gene
besa
cham
nanc
metz
luxe
mulh
bale
stra
400
200
0
−200
−400
cp2
−800
−600
−400
−200
0
200
400
cp1

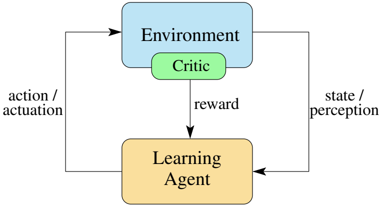
Supervised Learning

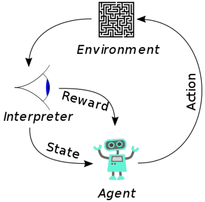
- observations = pairs (X_i, Y_i)
- goal = learn to *predict* Y_i *given* X_i
- regression (when Y is continuous)
- classification (when Y is discrete)
- statistical technique: linear models, and much more!

Example: Character Recognition

Input space $\mathcal{X}$ Output space $\mathcal{Y}$ Joint distribution $P(x, y)$	64×64 images $\{0, 1, \ldots, 9\}$?
Prediction function $h \in \mathcal{H}$ Risk $R(h) = P(h(X) \neq Y)$	
Sample $\{(x_i, y_i)\}_{i=1}^n$ Empirical risk $\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{h(x_i) \neq y_i\}$	MNIST dataset
Learning algorithm $\phi_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ Expected risk $R_n(\phi) = \mathbb{E}_n[R(\phi_n)]$	NN,boosting...
Empirical risk minimizer $\hat{h}_n = \mathrm{argmin}_{h \in \mathcal{H}} \hat{R}_n(h)$ Regularized empirical risk minimizer $\hat{h}_n = \mathrm{argmin}_{h \in \mathcal{H}} \hat{R}_n(h) + \lambda C(h)$	

Reinforcement Learning





[Src: https://en.wikipedia.org/wiki/Reinforcement_learning]

- area of machine learning inspired by behaviourist psychology
- how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.
- Model: random system (typically : Markov Decision Process)
  - agent
  - state
  - actions
  - rewards
- sometimes called approximate dynamic programming, or neuro-dynamic programming

Markov decision process

A Markov Decision Process is defined as a tuple $M = (X, A, p, r)$:

- X is the state space,
- A is the action space,
- $p(y|x, a)$ is the transition probability with

$$p(y|x, a) = \mathbb{P}(x_{t+1} = y|x_t = x, a_t = a),$$

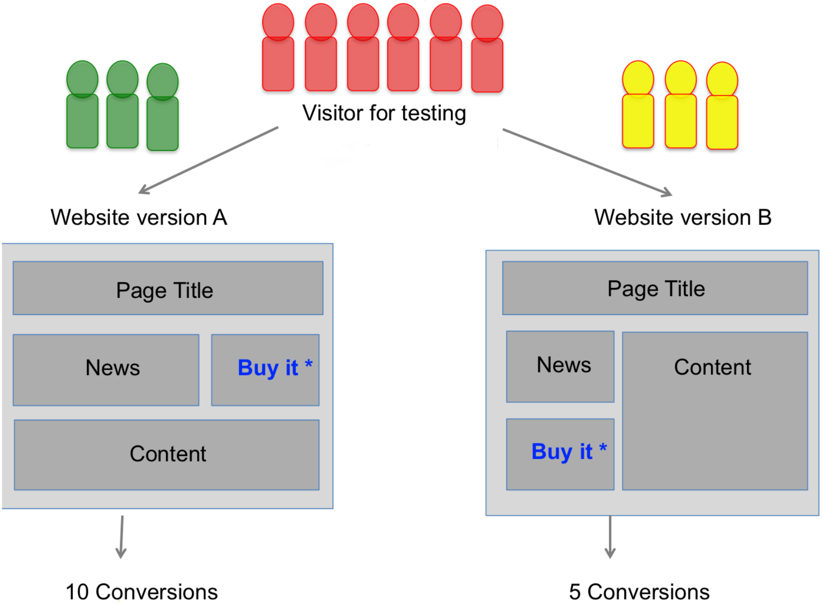
- $r(x, a, y)$ is the reward of transition (x, a, y).

Example: the Retail Store Management Problem

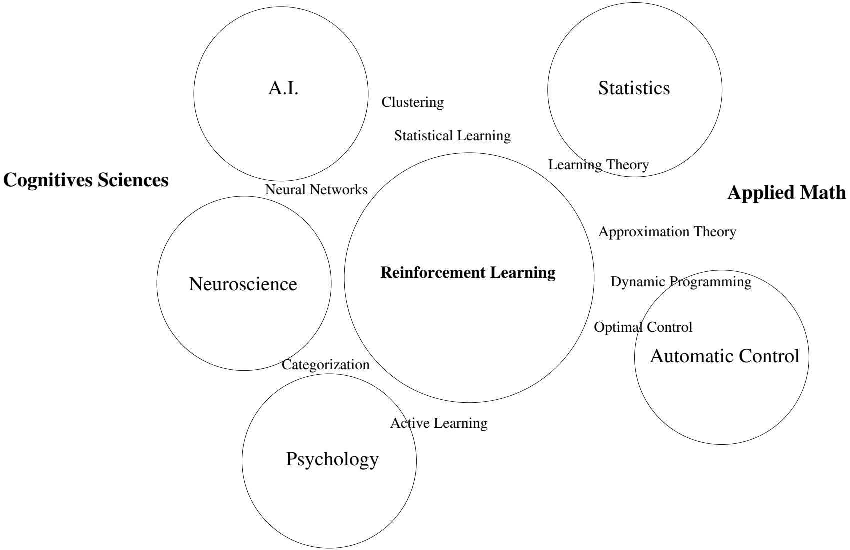
At each month t, a store contains x_t items of a specific goods and the demand for that goods is D_t. At the end of each month the manager of the store can order a_t more items from his supplier. Furthermore we know that:

- The cost of maintaining an inventory of x is $h(x)$.
- The cost to order a items is $C(a)$.
- The income for selling q items is $f(q)$.
- If the demand D is bigger than the available inventory x, customers that cannot be served leave.
- The value of the remaining inventory at the end of the year is $g(x)$.
- Constraint: the store has a maximum capacity M.

Example: A/B testing



Reinforcement Learning and its neighbors



Outline of the training program

Day 1 Panorama of machine learning.
Unsupervised learning:

- Principal Component Analysis
- Agglomerative Hierarchical Clustering
- k-means, k-medoids, and variants
- Overview of other methods: spectral clustering, Affinity Propagation, dbscan

Day 2 Supervised learning 1/2:

- Gaussian linear model, logistic regression, model selection
- LASSO and variants
- Support Vector Machines

Outline of the training program

Day 3 Supervised learning 2/2:

- Decision trees
- Bagging, Random Forests, Boosting
- Neural networks, introduction to deep learning

Day 4 Other learning paradigms:

- Sequential learning, multi-armed bandit problems
- Super-learning and expert aggregation
- Reinforcement learning (introduction)

Outline

ML Data

n-by-p matrix X

- n examples = observation points
- p features = characteristics measured for each example

Questions to consider:

- Are the features centered?
- Are the features normalized? bounded?

In `scikitlearn`, all methods expect a 2D array of shape (n, p) often called

```
X (n_samples, n_features)
```

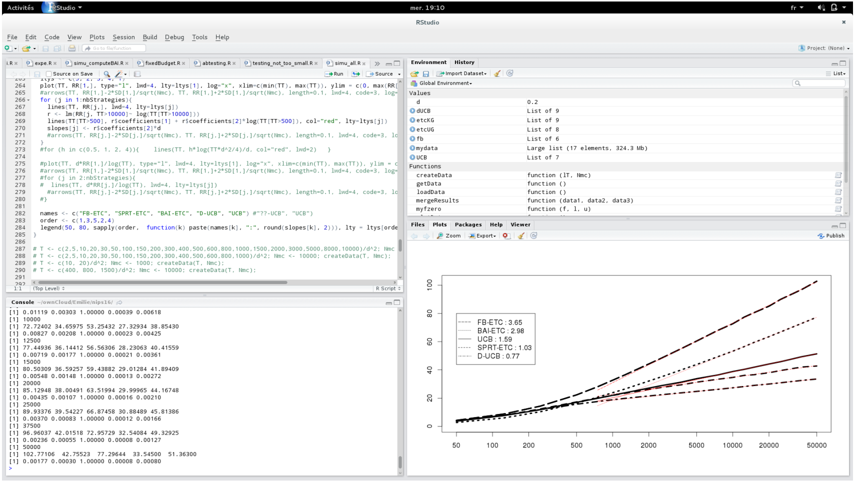
Data repositories

- Inside R: package `datasets`
- Inside scikitlearn: package `sklearn.datasets`
- UCI Machine Learning Repository
- Challenges: Kaggle, etc.

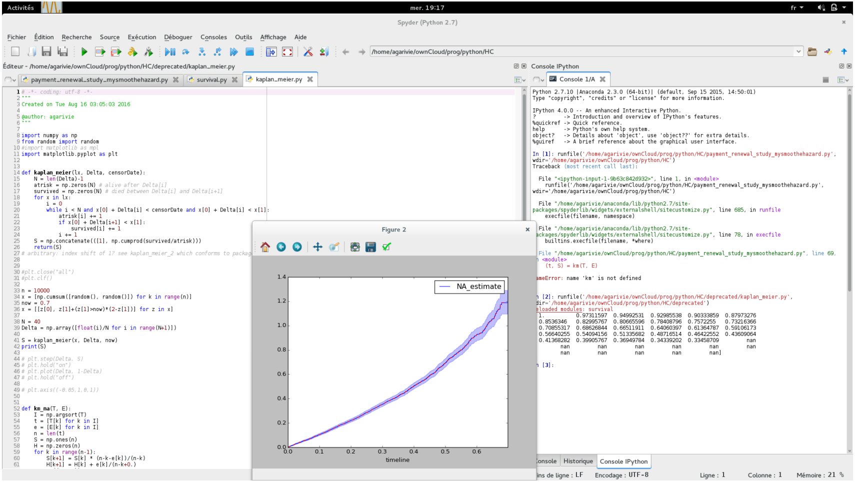
The big steps of data analysis

1. Extracting the data to expected format
2. Exploring the data
   - detection of outliers, of inconsistencies
   - descriptive exploration of the distributions, of correlations
   - data transformations
3. Random partitioning of the data: (see also: cross-validation)
   - learning sample
   - validation sample
   - test sample
4. For each algorithm: parameter estimation using training and validation samples
5. Choice of final algorithm using testing sample, risk estimation

Machine Learning tools: R



Machine Learning tools: python



scikitlearn:

http://scikit-learn.org/stable/index.html

Knime, Weka and co: integrated environments