Machine Learning

A Panorama

ML Team

Centre International de Mathématiques et Informatique de Toulouse
Université de Toulouse

ML Team Creation Seminar, 2016

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

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.

AI: composition

Central goals of AI:

- reasoning
- knowledge
- planning
- learning
- natural language processing
- perception
- general intelligence

Central approaches of AI:

- traditional symbolic AI
- statistical methods
- computational intelligence / soft computing

Draws upon:

- computer science
- mathematics
- psychology
- linguistics
- philosophy
- neuroscience
- artificial psychology

Tools:

- mathematical optimization
- logic
- methods based on probability
- economics

Outline

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

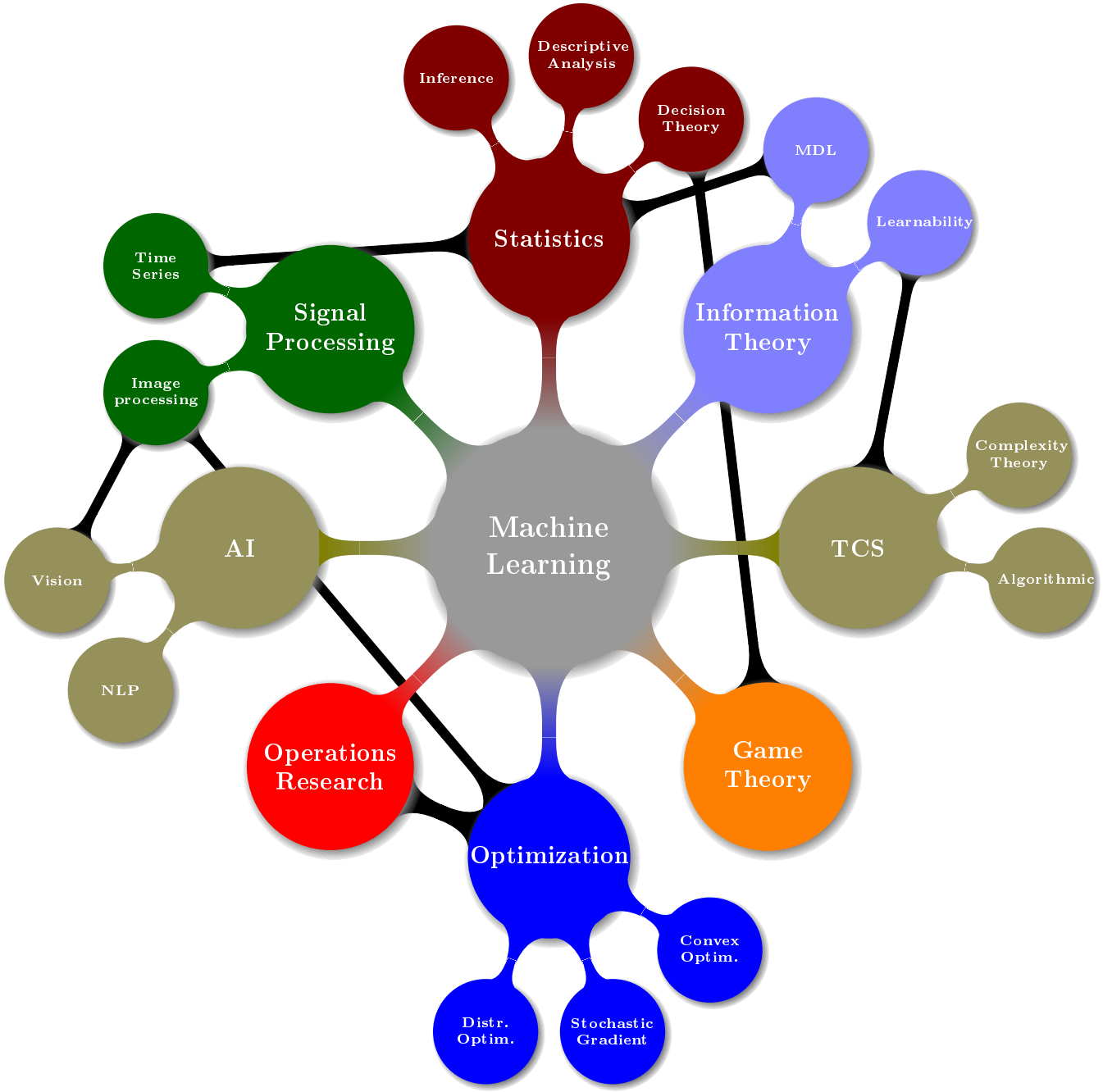
- spam filtering, text classification
- optical character recognition (OCR)
- search engines
- recommendation platforms
- speach recognition software
- computer vision
- bio-informatics, DNA analysis, medicine

For each of this task, 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

ML and its neighbors



Machine Learning: an overview

Supervised Classification: Statistical Framework

Definition	ex: OCR numbers
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 $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$ Regularized empirical risk minimizer $\hat{h}_n = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h) + \lambda C(h)$	

Empirical Risk Minimization

Hoeffding's inequality: with probability at least $1-\eta$,

$$\left|R(h) - \hat{R}_n(h)\right| \leq \sqrt{\frac{1}{2n}\log\left(\frac{2}{\eta}\right)}\,.$$

Problem: true for *every fixed* h but *not for* $\hat{h}$!
Ex: Prediction of 10 digits
Ex: polynomial regression $\rightarrow$ overfitting
Curse of dimensionality

Structural Risk Minimization

$\rightarrow$ *uniform* law of large numbers: Vapnik-Chervonenkis inequality: if $\mathcal{H}$ has *VC dimension* $d_{\mathcal{H}}$, then

$$\sup_{h\in\mathcal{H}} \left|R(h)-\hat{R}_n(h)\right| \leq O\left(\sqrt{\frac{1}{2n}\log\left(\frac{2}{\eta}\right)+\frac{d_{\mathcal{H}}}{n}\log\left(\frac{n}{d_{\mathcal{H}}}\right)}\right) .$$

Structure:

$$\mathcal{H}=\bigcup_m \mathcal{H}_m$$

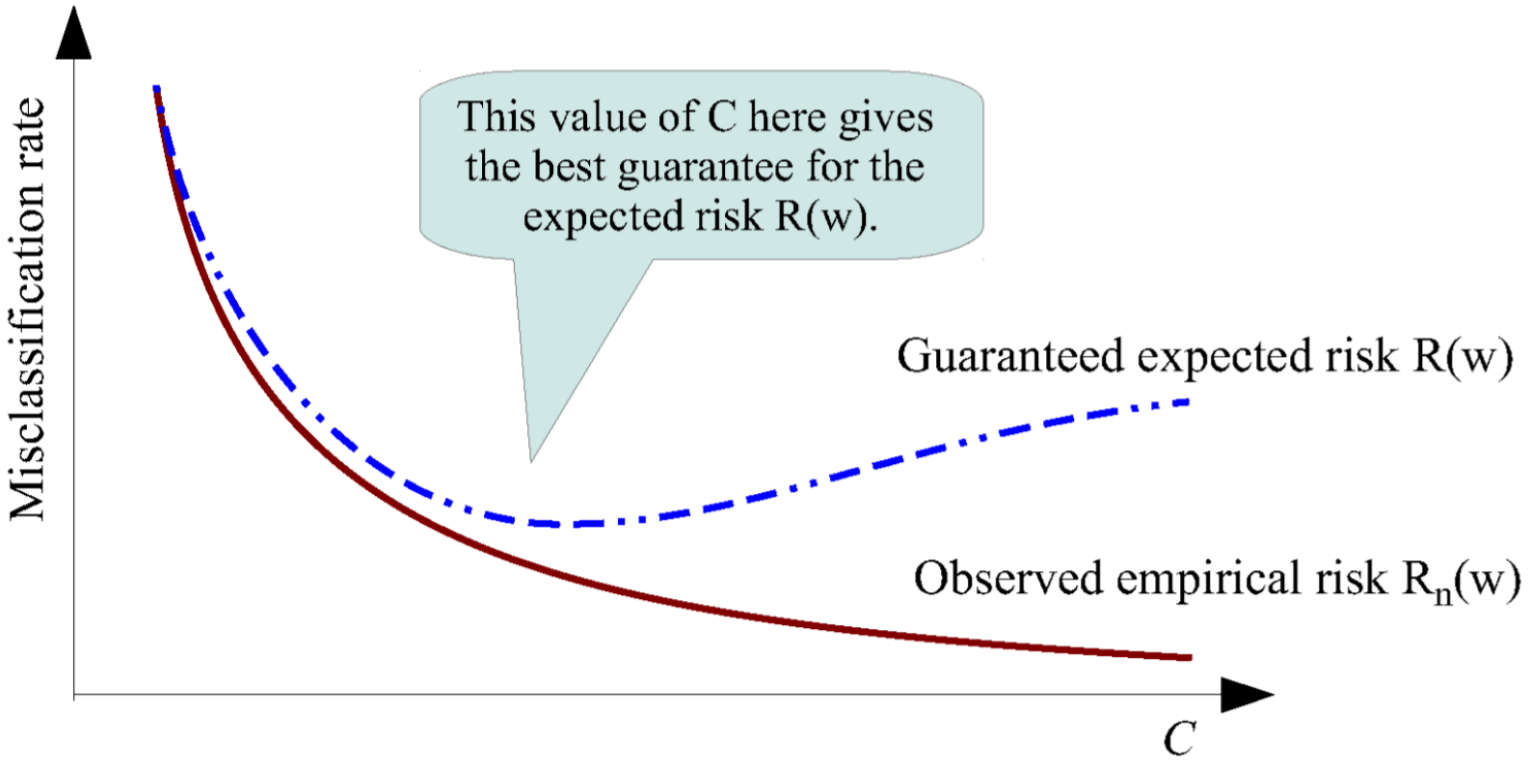
Ex: polynomials/splines of degree m, trees of depth m,...
Bias-variance decomposition of the risk.
Structural risk minimization:

$$\hat{h}_n = \arg\min_{h\in\mathcal{H}} \hat{R}_n(h)+\lambda K(h)$$

or

$$\hat{h}_n = \arg\min_{K(h)\leq C} \hat{R}_n(h)$$

Structural Risk Minimization Tradeoff



Source: Bottou et al. tutorial on optimization

Learning methodology: Choosing C

- Division of the sample set:
  - **Training set**
  - **Validation set**
  - **Testing set**
- **Cross-validation**
- **Early stopping**

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 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 seperation 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).

Theory of Learnability: PAC-Learnability (Valiant '84)

- $\mathcal{X}$ = instance space (ex: set of B/W images)
- *Concept* c=subset of $\mathcal{X}$ (ex: all images of a '3')
- *Concept class* C = set of concepts (ex: all images with connected 1-components)
- tolerance parameter $\epsilon > 0$, risk parameter $\delta > 0$
- given a probability P on $\mathcal{X}$, algorithm A PAC-learns concept c if, given an sample of polynomial size $p(1/\epsilon, 1/\delta)$, A outputs an *hypothesis* $h \in C$ such that

$$P\big(\mathrm{err}\big(h(x), c(x)\big) \leq \epsilon\big) \geq 1 - \delta$$

- if A PAC-learns every $c \in C$ for every distribution P and every $\epsilon, \delta > 0$, then C is **PAC-learnable**.

Vapnik

Under some regularity conditions these three conditions are equivalent:

1. The concept class C is PAC learnable.
2. The VC dimension of C is finite.
3. C is a uniform Glivenko-Cantelli class.

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Simple Analysis

- **Statistical Learning Literature:**
"It is good to optimize an objective function than ensures a fast estimation rate when the number of examples increases."

- **Optimization Literature:**
"To efficiently solve large problems, it is preferable to choose an optimization algorithm with strong asymptotic properties, e.g. superlinear."

- **Therefore:**
"To address large-scale learning problems, use a superlinear algorithm to optimize an objective function with fast estimation rate.
Problem solved."

The purpose of this presentation is. . .

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Too Simple an Analysis

- **Statistical Learning Literature:**
"It is good to optimize an objective function than ensures a fast estimation rate when the number of examples increases."

- **Optimization Literature:**
"To efficiently solve large problems, it is preferable to choose an optimization algorithm with strong asymptotic properties, e.g. superlinear."

- **Therefore:** *(error)*
"To address large-scale learning problems, use a superlinear algorithm to optimize an objective function with fast estimation rate.
Problem solved."

. . . to show that this is completely wrong !

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Objectives and Essential Remarks

- Baseline large-scale learning algorithm

Randomly discarding data is the simplest way to handle large datasets.

  - What are the statistical benefits of processing more data?
  - What is the computational cost of processing more data?

- We need a theory that joins Statistics and Computation!

- 1967: Vapnik's theory does not discuss computation.
- 1981: Valiant's learnability excludes exponential time algorithms, but (i) polynomial time can be too slow, (ii) few actual results.
- We propose a simple analysis of approximate optimization. . .

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Learning Algorithms: Standard Framework

- Assumption: examples are drawn independently from an unknown probability distribution $P(x, y)$ that represents the rules of Nature.
- Expected Risk: $E(f) = \int \ell(f(x), y)\, dP(x, y)$.
- Empirical Risk: $E_n(f) = \frac{1}{n} \sum \ell(f(x_i), y_i)$.
- We would like f^* that minimizes $E(f)$ among all functions.
- In general $f^* \notin \mathcal{F}$.
- The best we can have is $f^*_{\mathcal{F}} \in \mathcal{F}$ that minimizes $E(f)$ inside $\mathcal{F}$.
- But $P(x, y)$ is unknown by definition.
- Instead we compute $f_n \in \mathcal{F}$ that minimizes $E_n(f)$.
  Vapnik-Chervonenkis theory tells us when this can work.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Learning with Approximate Optimization

Computing $f_n = \arg\min_{f \in \mathcal{F}} E_n(f)$ is often costly.

Since we already make lots of approximations,

why should we compute f_n exactly?

Let's assume our optimizer returns $\tilde{f}_n$

such that $E_n(\tilde{f}_n) < E_n(f_n) + \rho$.

For instance, one could stop an iterative optimization algorithm long before its convergence.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Decomposition of the Error (i)

$$
\begin{aligned}
E(\tilde{f}_n) - E(f^*) &= E(f^*_{\mathcal{F}}) - E(f^*) && \text{Approximation error} \\
&+ E(f_n) - E(f^*_{\mathcal{F}}) && \text{Estimation error} \\
&+ E(\tilde{f}_n) - E(f_n) && \text{Optimization error}
\end{aligned}
$$

Problem:

Choose $\mathcal{F}$, n, and ρ to make this as small as possible,

subject to budget constraints $\left\{ \begin{array}{l} \text{maximal number of examples } n \\ \text{maximal computing time } T \end{array} \right.$

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Decomposition of the Error (ii)

Approximation error bound: (Approximation theory)
– decreases when $\mathcal{F}$ gets larger.

Estimation error bound: (Vapnik-Chervonenkis theory)
– decreases when n gets larger.
– increases when $\mathcal{F}$ gets larger.

Optimization error bound: (Vapnik-Chervonenkis theory plus tricks)
– increases with ρ.

Computing time T: (Algorithm dependent)
– decreases with ρ
– increases with n
– increases with $\mathcal{F}$

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Small-scale vs. Large-scale Learning

We can give *rigorous definitions*.

- **Definition 1:**
  We have a **small-scale learning** problem when the **active budget constraint is the number of examples n.**

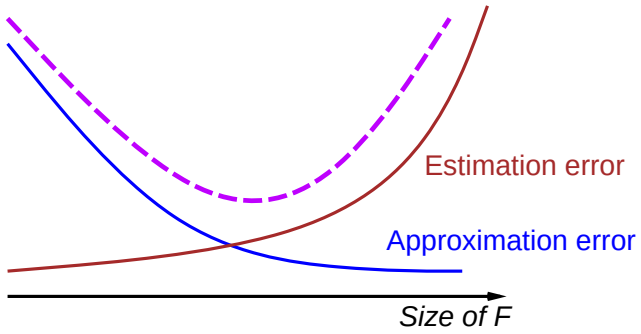
- **Definition 2:**
  We have a **large-scale learning** problem when the **active budget constraint is the computing time T.**

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Small-scale Learning

The active budget constraint is the number of examples.

- To reduce the estimation error, take n as large as the budget allows.
- To reduce the optimization error to zero, take $\rho = 0$.
- We need to adjust the size of $\mathcal{F}$.



See Structural Risk Minimization (Vapnik 74) and later works.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

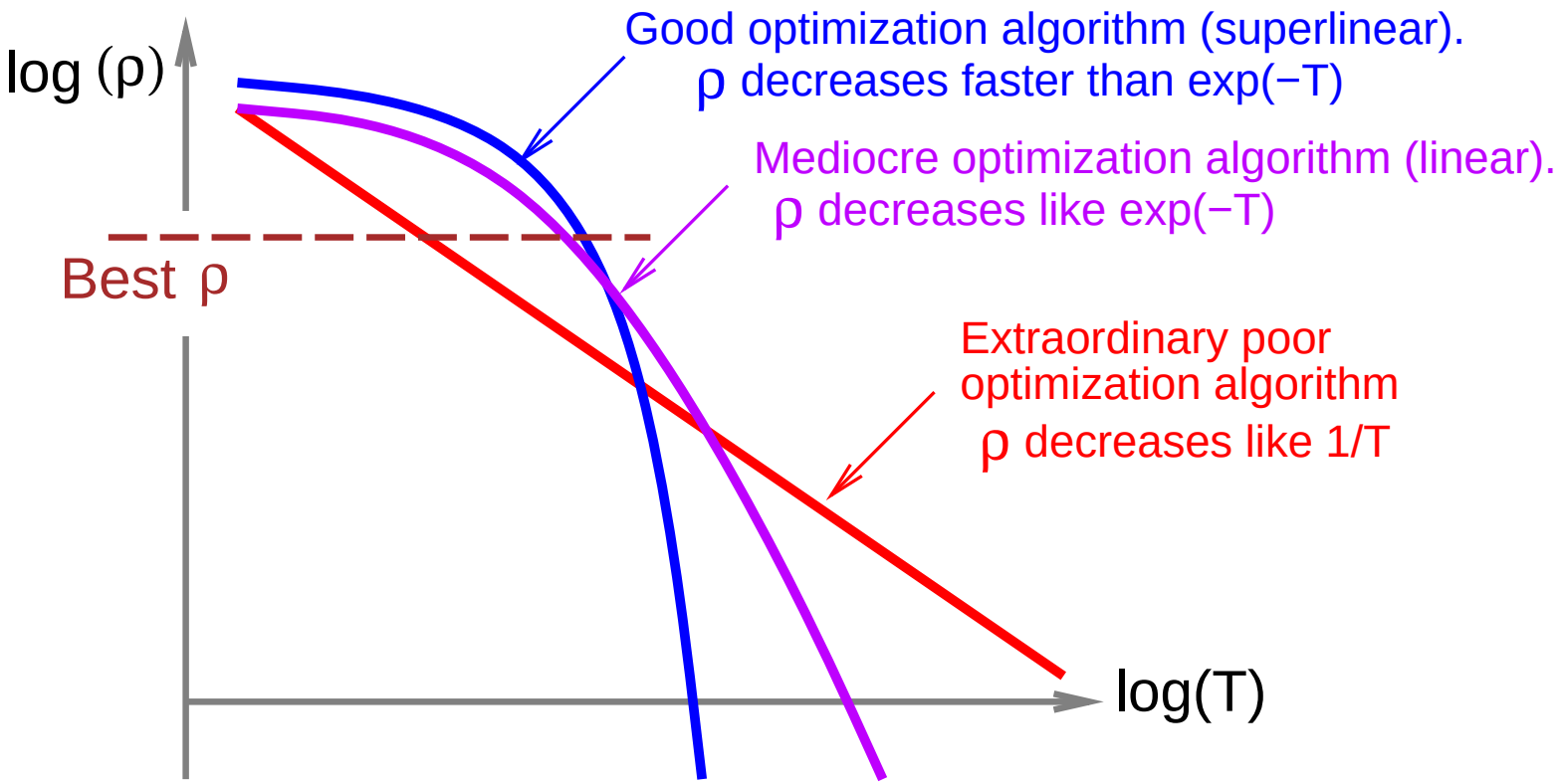
Large-scale Learning

The active budget constraint is the computing time.

- More complicated tradeoffs.
  The computing time depends on the three variables: $\mathcal{F}$, n, and ρ.
- Example.
  If we choose ρ small, we decrease the optimization error. But we must also decrease $\mathcal{F}$ and/or n with adverse effects on the estimation and approximation errors.
- The exact tradeoff depends on the optimization algorithm.
- We can compare optimization algorithms rigorously.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Executive Summary



[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Asymptotics: Estimation

Uniform convergence bounds (with capacity $d+1$)

$$\text{Estimation error} \leq \mathcal{O}\left(\left[\frac{d}{n} \log \frac{n}{d} \right]^{\alpha} \right) \text{ with } \frac{1}{2} \leq \alpha \leq 1 \ .$$

There are in fact three types of bounds to consider:

- Classical V-C bounds (pessimistic): $\mathcal{O}\left(\sqrt{\frac{d}{n}} \right)$
- Relative V-C bounds in the realizable case: $\mathcal{O}\left(\frac{d}{n} \log \frac{n}{d} \right)$
- Localized bounds (variance, Tsybakov): $\mathcal{O}\left(\left[\frac{d}{n} \log \frac{n}{d} \right]^{\alpha} \right)$

Fast estimation rates are a big theoretical topic these days.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

Asymptotics: Estimation+Optimization

Uniform convergence arguments give

$$\text{Estimation error} + \text{Optimization error} \leq \mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha} + \rho\right) .$$

This is true for all three cases of uniform convergence bounds.

Scaling laws for ρ when $\mathcal{F}$ is fixed

The approximation error is constant.

- No need to choose ρ smaller than $\mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right)$.
- Not advisable to choose ρ larger than $\mathcal{O}\left(\left[\frac{d}{n}\log\frac{n}{d}\right]^{\alpha}\right)$.

[Src: Bottou, *The Tradeoffs of Large-scale Learning*, http://leon.bottou.org/slides/]

. . . Approximation+Estimation+Optimization

When $\mathcal{F}$ is chosen via a λ-regularized cost

- Uniform convergence theory provides bounds for simple cases
  (Massart-2000; Zhang 2005; Steinwart et al., 2004-2007; . . .)
- Computing time depends on both λ and ρ.
- Scaling laws for λ and ρ depend on the optimization algorithm.

When $\mathcal{F}$ is realistically complicated

Large datasets matter

- because one can use more features,
- because one can use richer models.

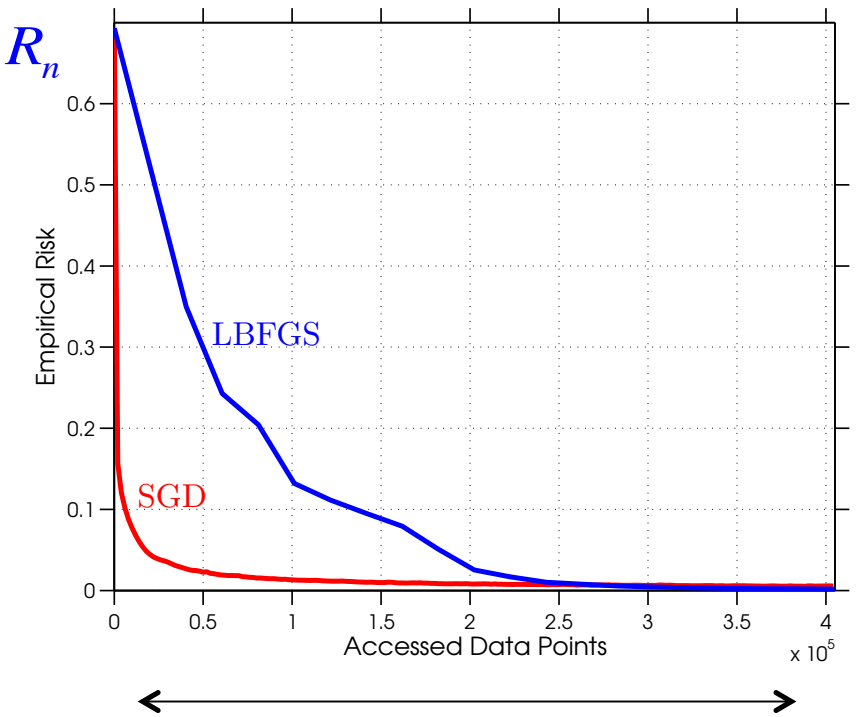
Bounds for such cases are rarely realistic enough.

Luckily there are interesting things to say for $\mathcal{F}$ fixed.

Optimization: Stochastic Gradient

- Second-order methods are too costly (even 1 iteration)
- Even first-order methods are too costly with massive data
- SG employs information more efficiently than batch method

Practical Experience



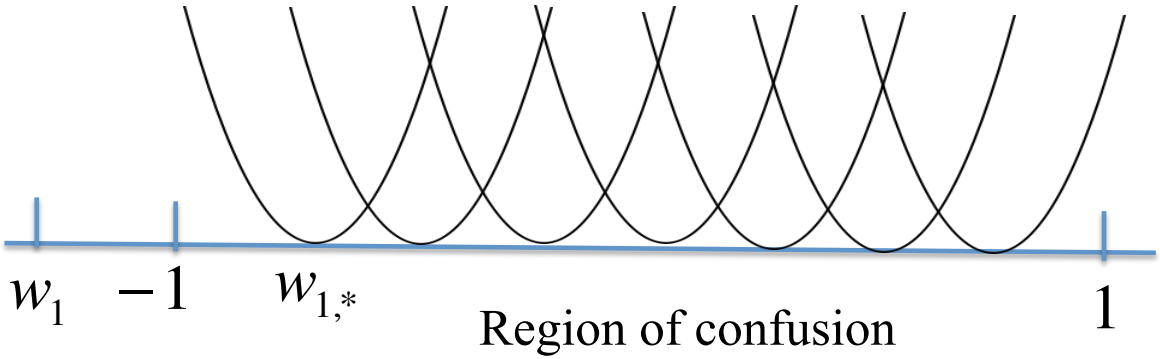
10 epochs

Fast initial progress of SG followed by drastic slowdown

Can we explain this?

Example by Bertsekas

$$R_n(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$



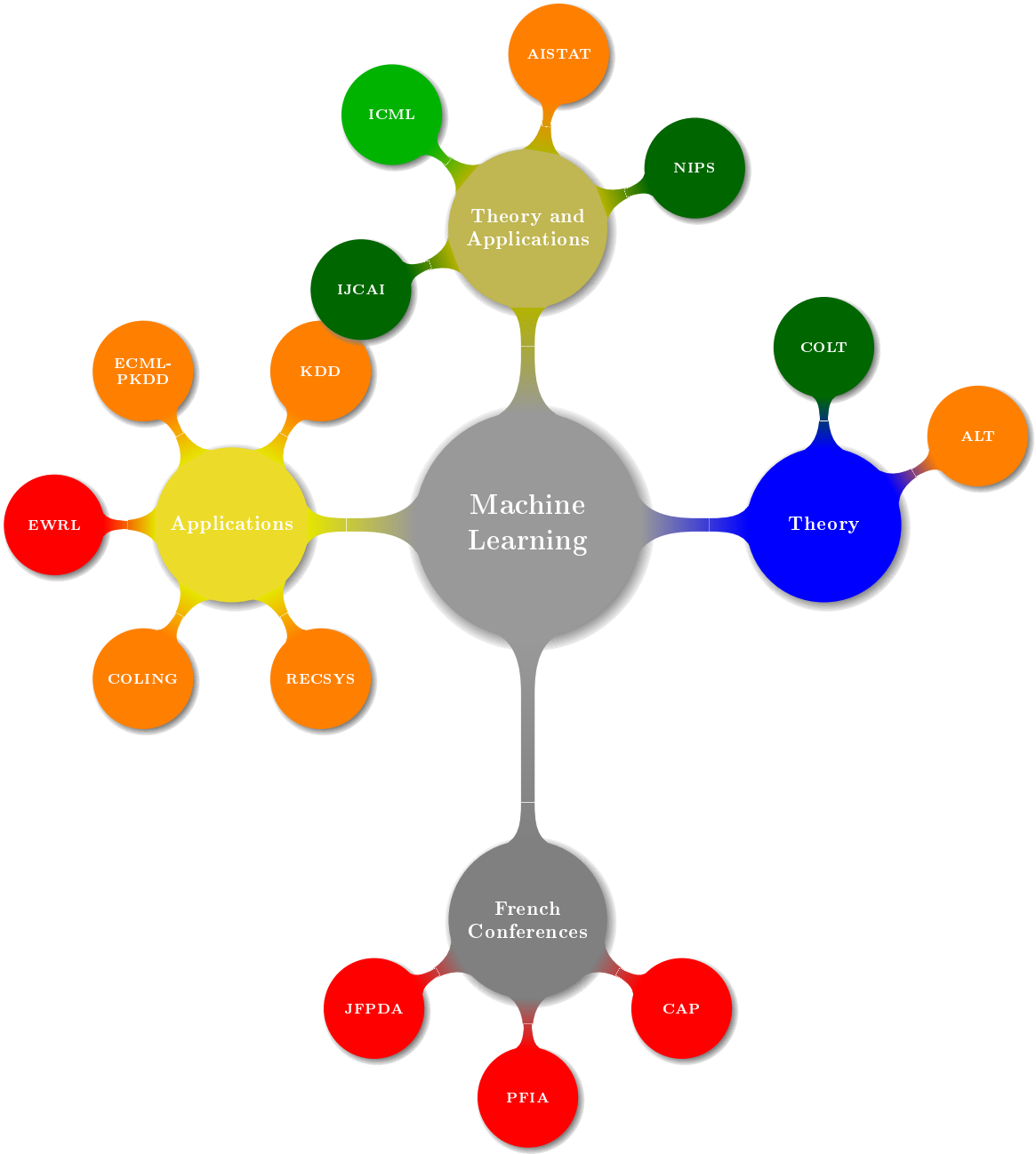
Note that this is a geographical argument

Analysis: given w_k what is the expected decrease in the objective function R_n as we choose one of the quadratics randomly?

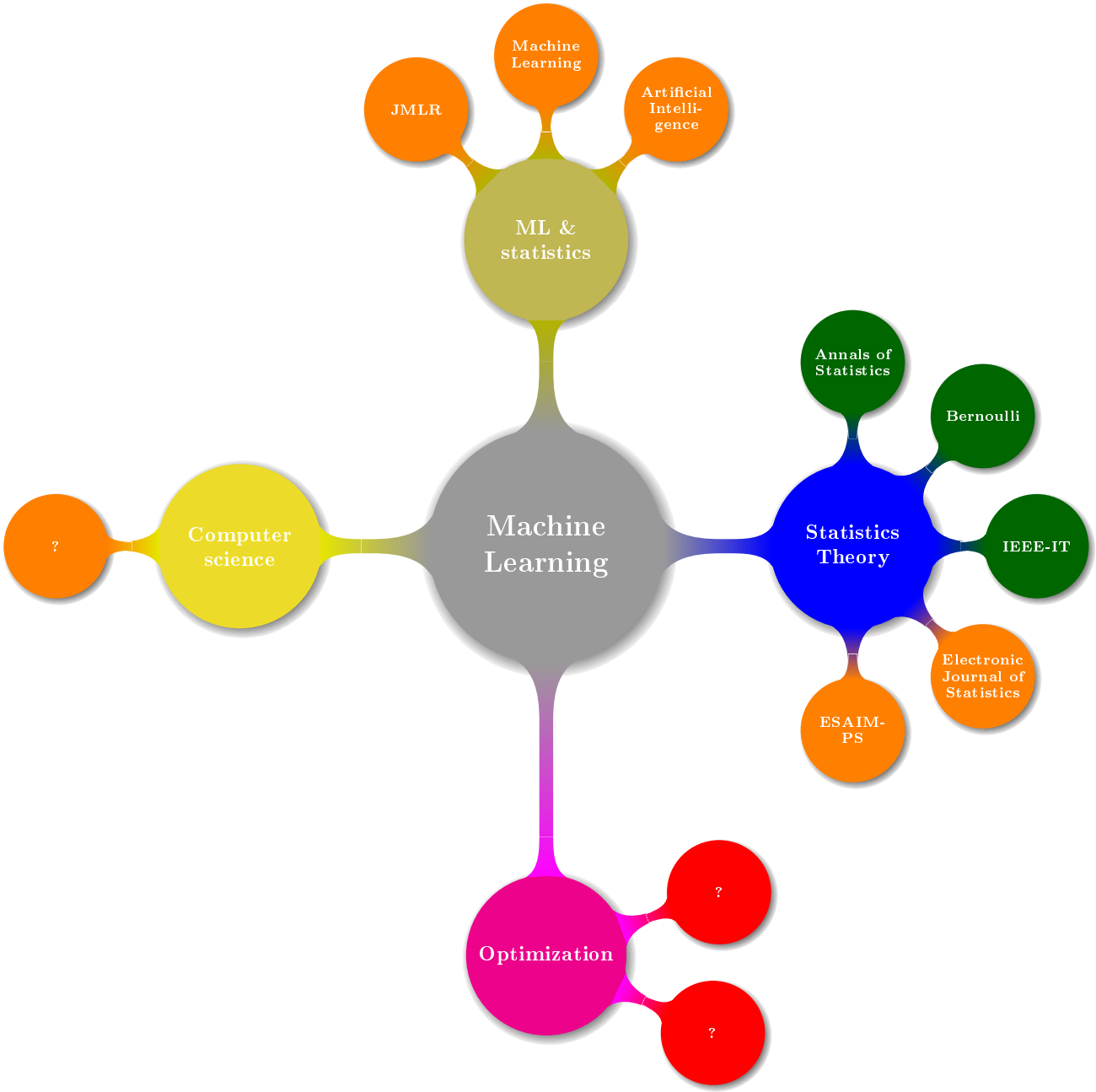
Machine Learning and Conferences

- Very active / frenetic domain of research
- Mass meetings
- Fashion trends, passing fads

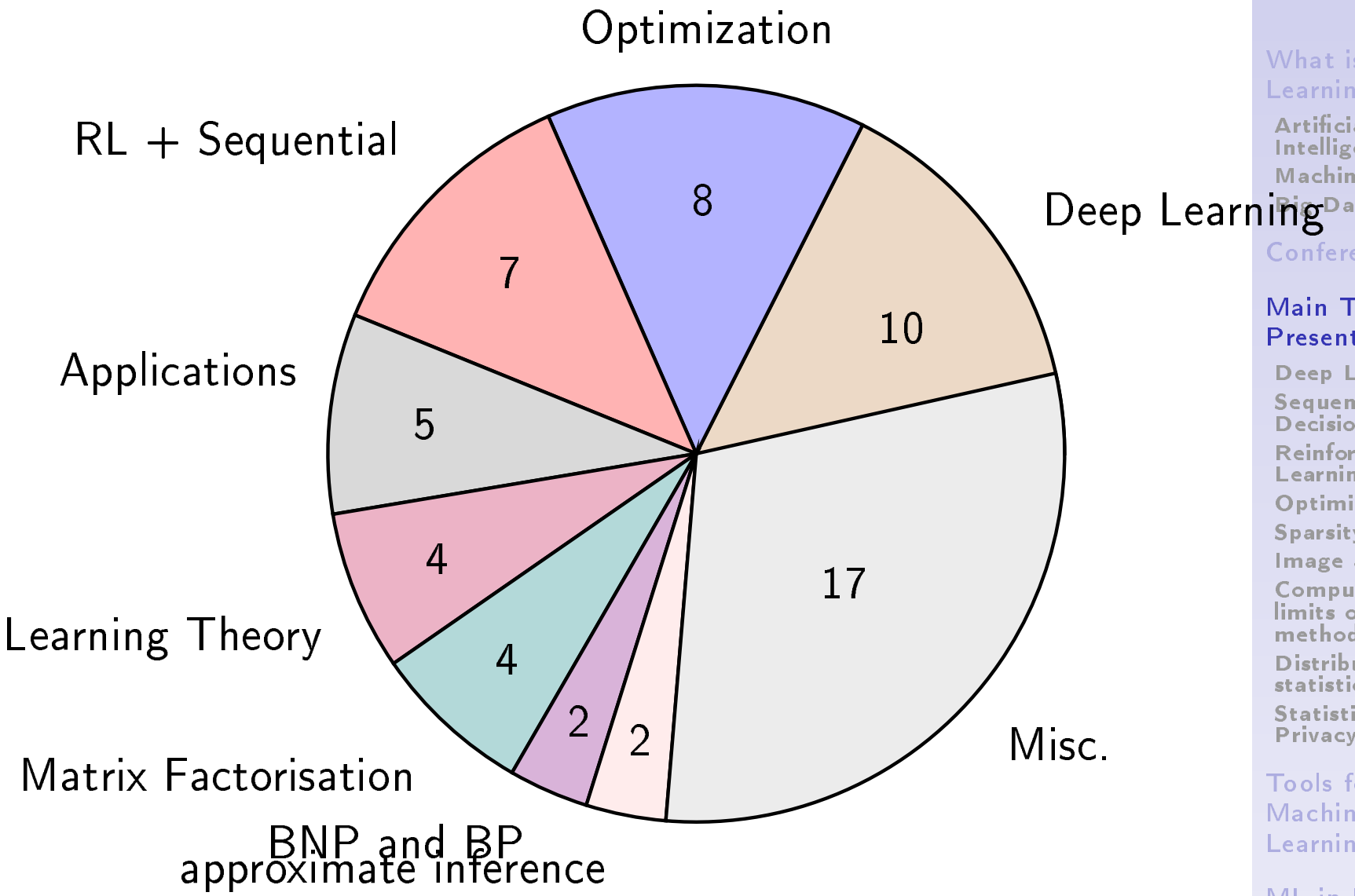
Main Machine Learning Conferences



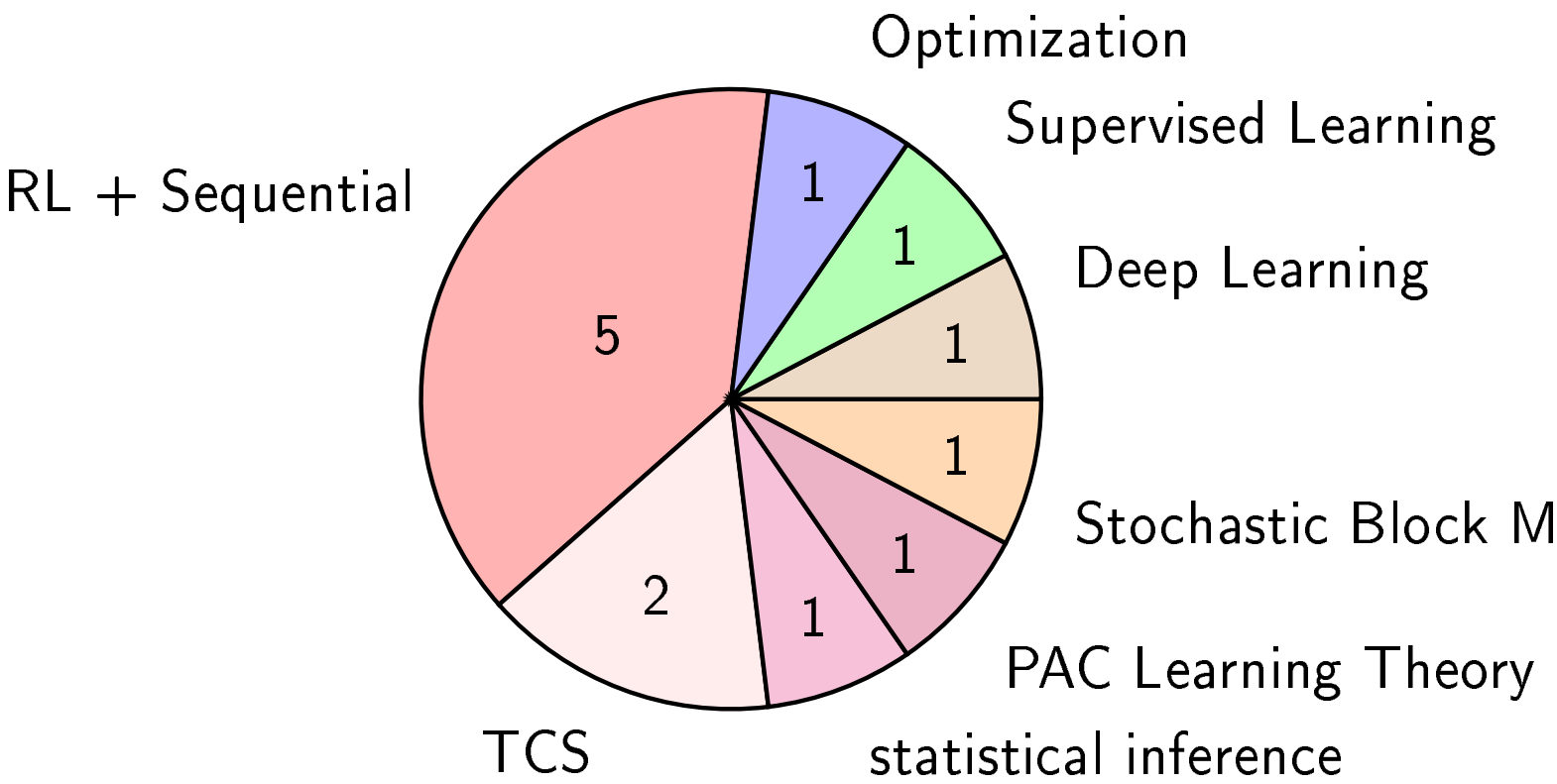
Machine Learning Journals



Sessions at ICML'16



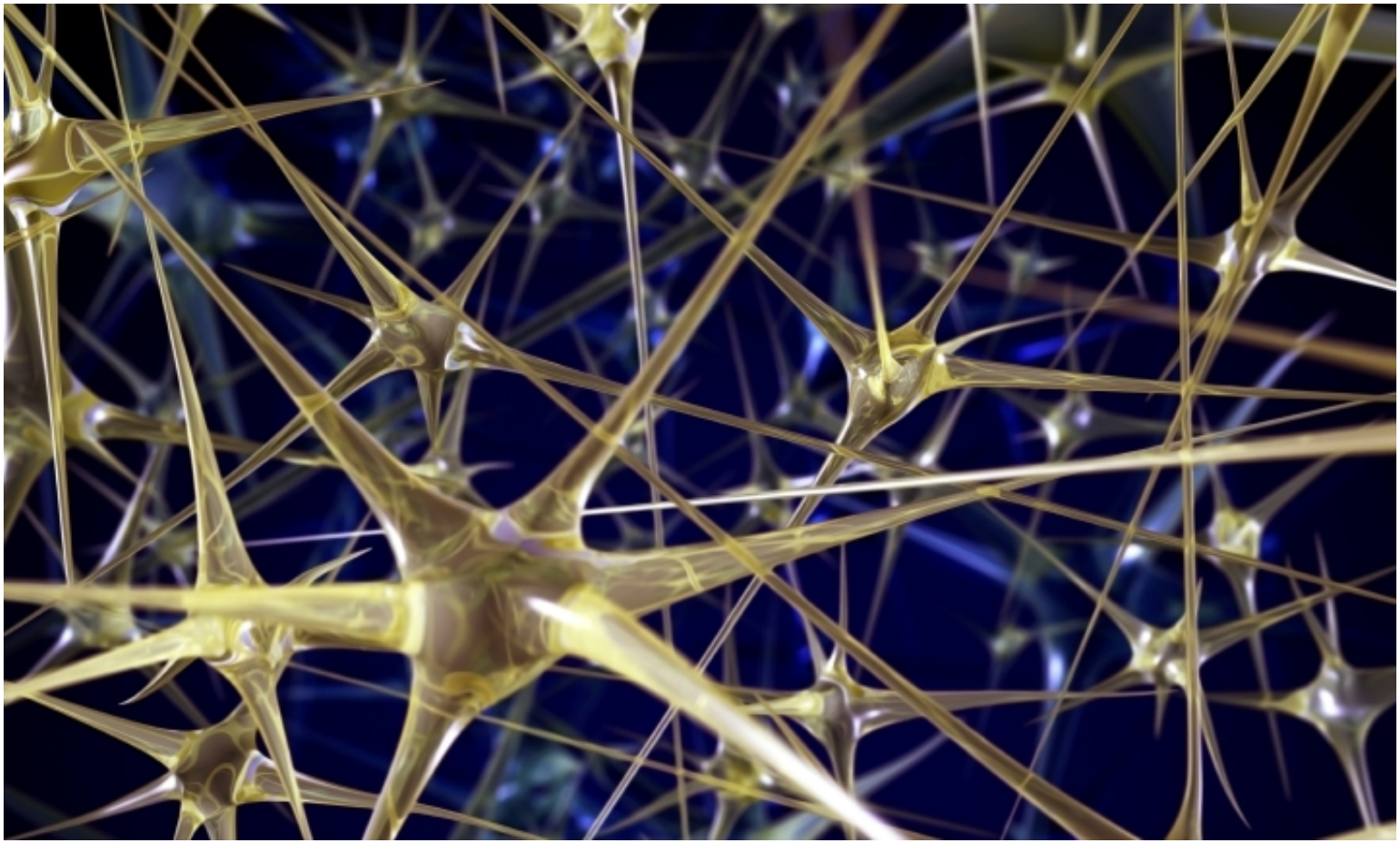
Sessions at COLT'16



Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Réseau mono-couche



Source: http://insanedev.co.uk/open-cranium/

Réseau mono-couche

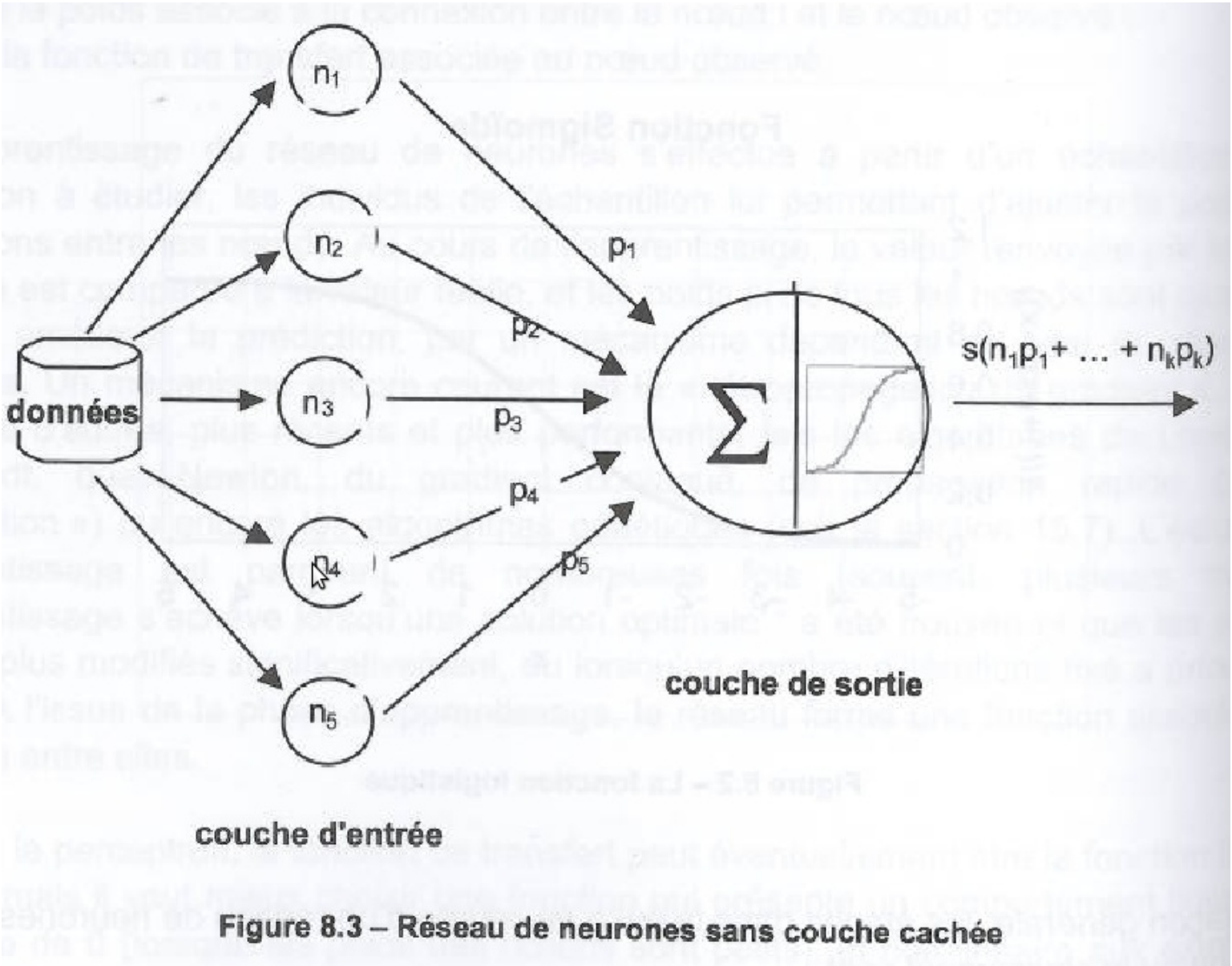
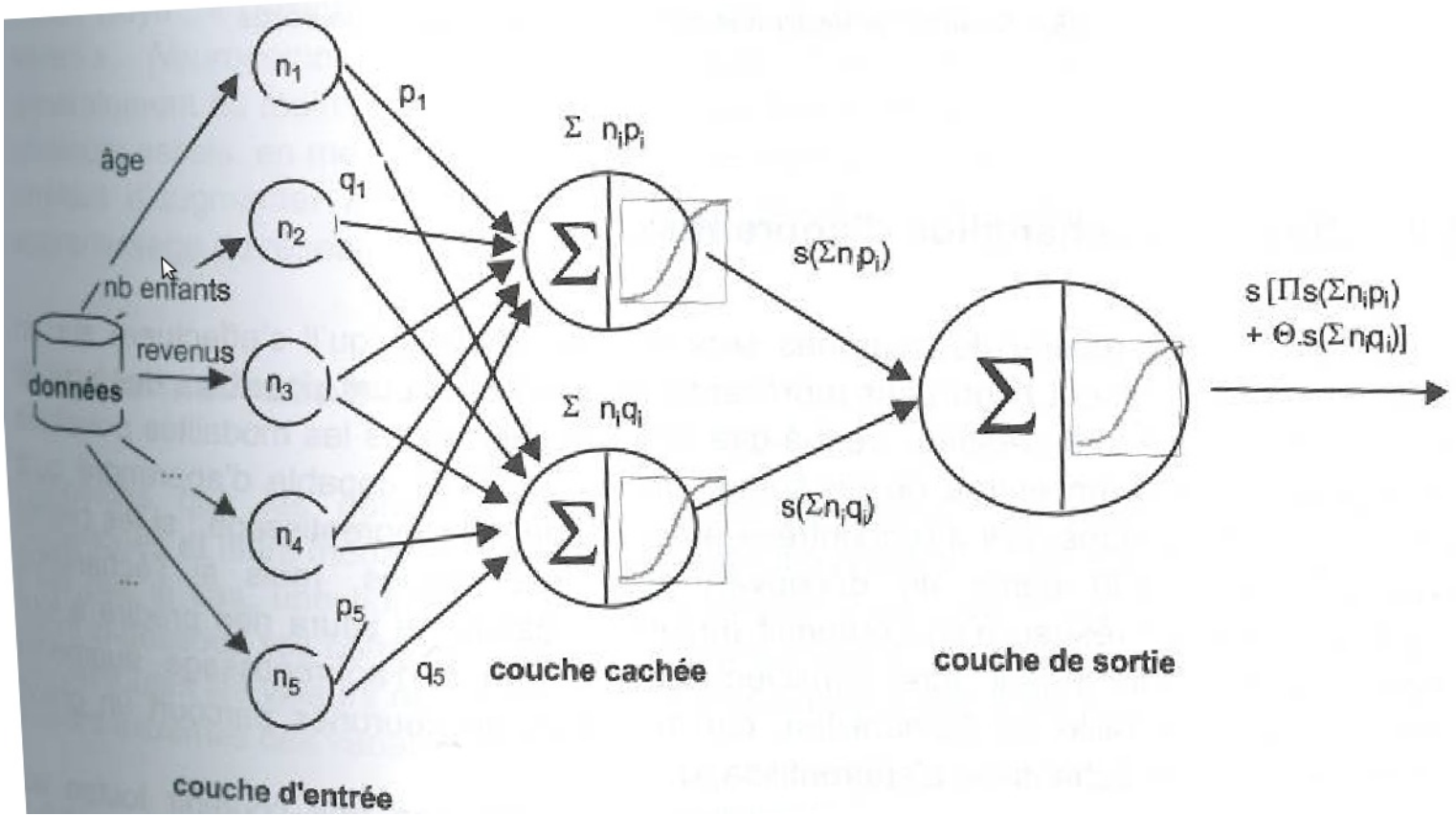


Figure 8.3 – Réseau de neurones sans couche cachée

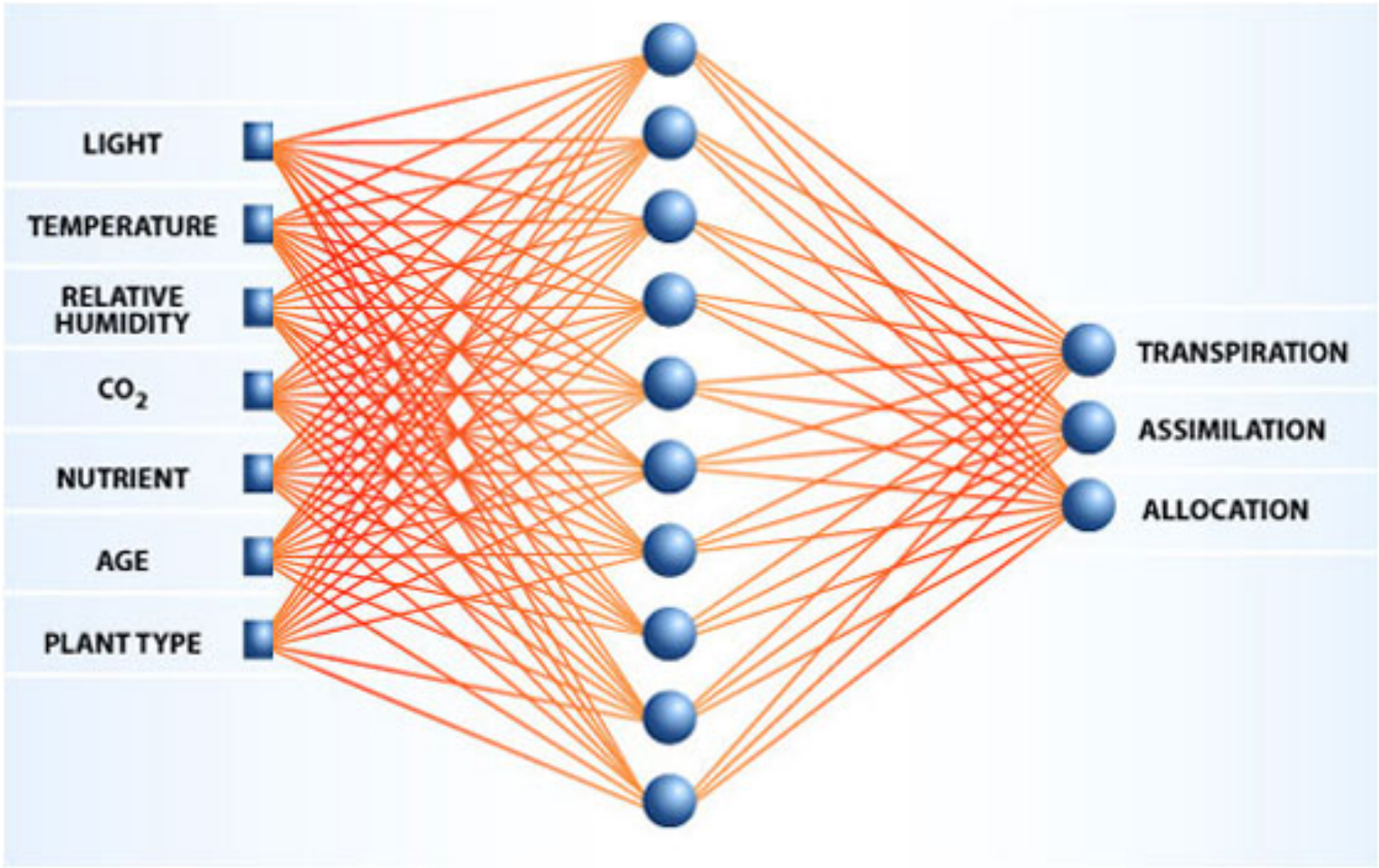
Source: [Tufféry, Data Mining et Informatique Décisionnelle]

Réseau avec couche intermédiaire



Source: [Tufféry, Data Mining et Informatique Décisionnelle]

Réseau mono-couche



Src: http://www.makhfi.com

From neural networks to deep learning

- Deep learning=neural networks + 3 improvements
- Extented scope
  - new activation functions
  - convolution
  - recursivity
- Regularization
  - Dropout
  - Pooling
  - Make possible the learning of complex networks
- Computation
  - GPU
  - massive data
  - Make efficient the learning of complex networks

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Sequential decision making

Sequential protocol: at each time $t = 1, 2, \ldots$, a learner outputs an action $A_t \in \mathcal{A}$ and receives a reward $r_t(A_t) \in \mathbb{R}$.

Regret minimization: the goal of the learner is to output sequential actions $a_1, a_2, \ldots$ that are almost as good as the best fixed action in hindsight, i.e., to minimize the regret

$$R_T := \sup_{a \in \mathcal{A}} \sum_{t=1}^{T} r_t(a) - \sum_{t=1}^{T} r_t(A_t) \,.$$

Various feedback models:

- full information: the whole function $r_t(\cdot)$ is observed;
- bandit feedback: we only observe the reward $r_t(A_t)$ of the played action (exploration is thus needed);
- partial monitoring: we observe a function of r_t and A_t.

Rich literature since the 50's (seminal works by Robbins [12], Blackwell [9], and Hannan [11]). Excellent introduction: book by Cesa-Bianchi and Lugosi [10].

Sequential decision making (2)

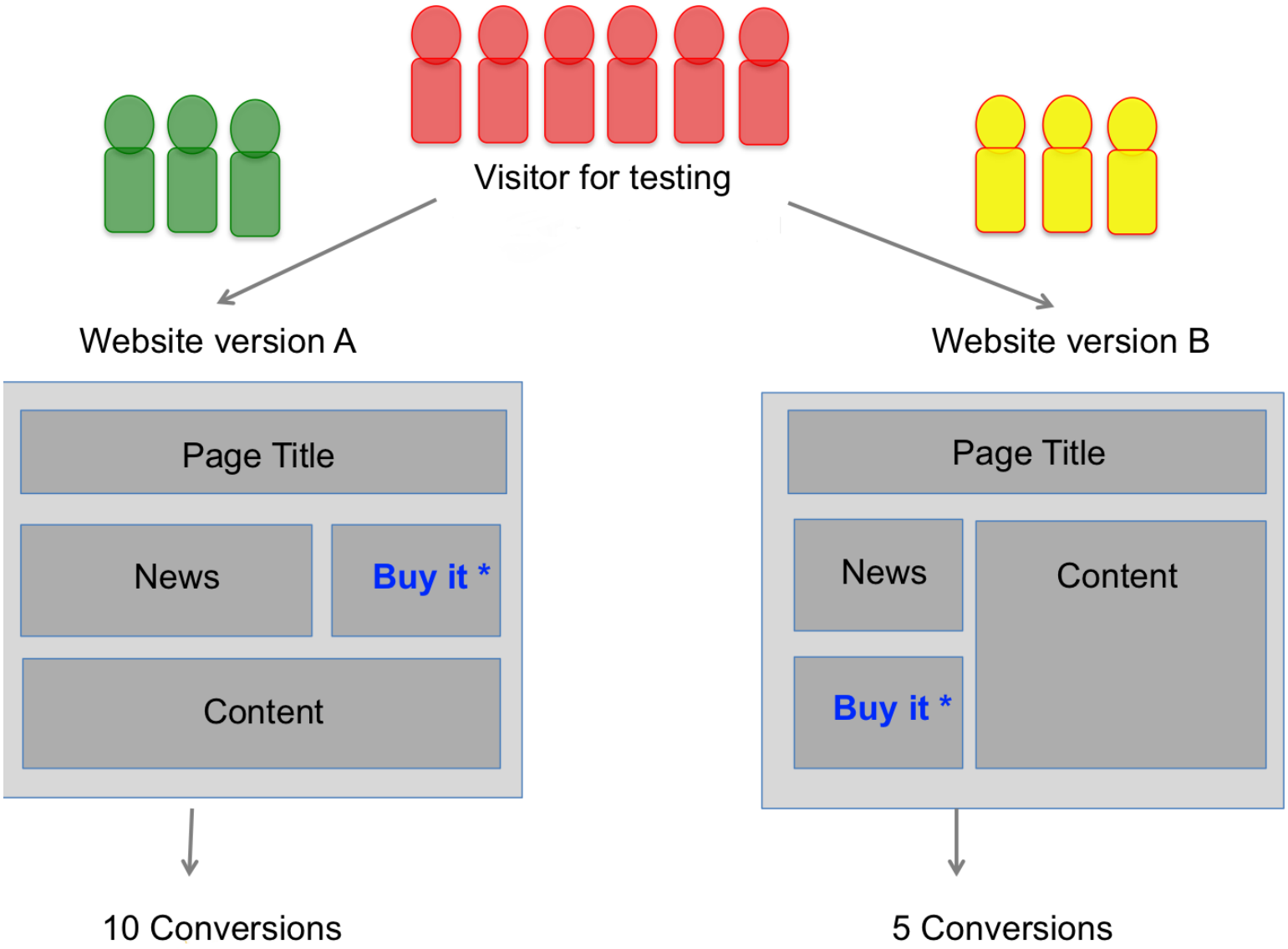
Various models for the environment:

- Stochastic environment: the reward functions r_t are drawn i.i.d. according to an unknown probability distribution.
- Adversarial environment: the reward functions r_t are chosen by an adversary that may react to the learner's past actions.
- There are of course intermediate settings between these two extreme (but easy-to-analyze) sets of assumptions.

Flavor of theoretical contributions: sequential algorithms with regret guarantees $R_T = o(T)$ under mild conditions on the reward functions r_t. The growth of the regret R_T depends on:

- the feedback model
- whether the environment is stochastic or adversarial (or in between);
- the curvature of the reward functions (e.g., strong convexity implies small regret);
- how large is the set of possible reward functions.

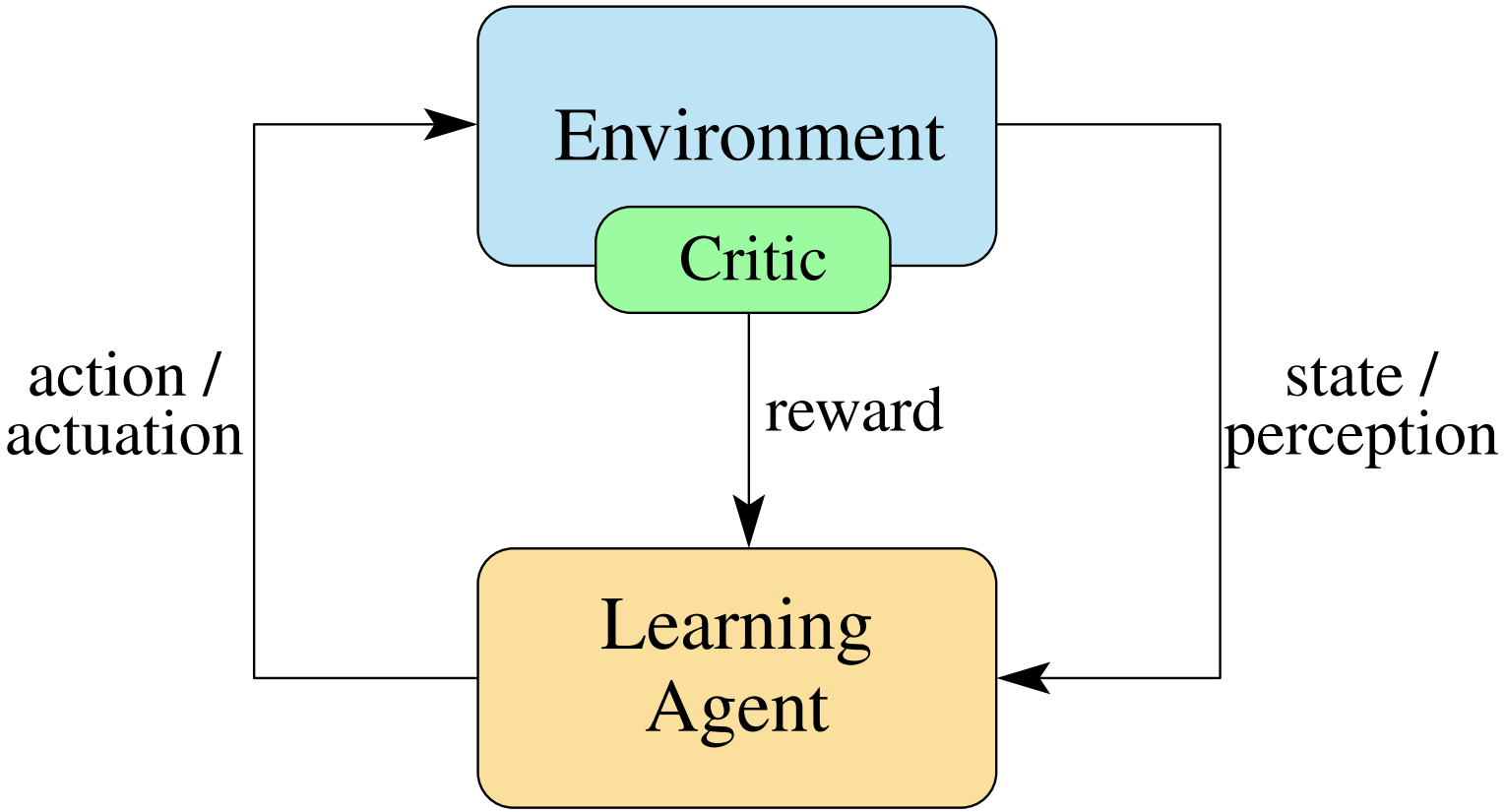
A/B Testing



Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Reinforcement Learning



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

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

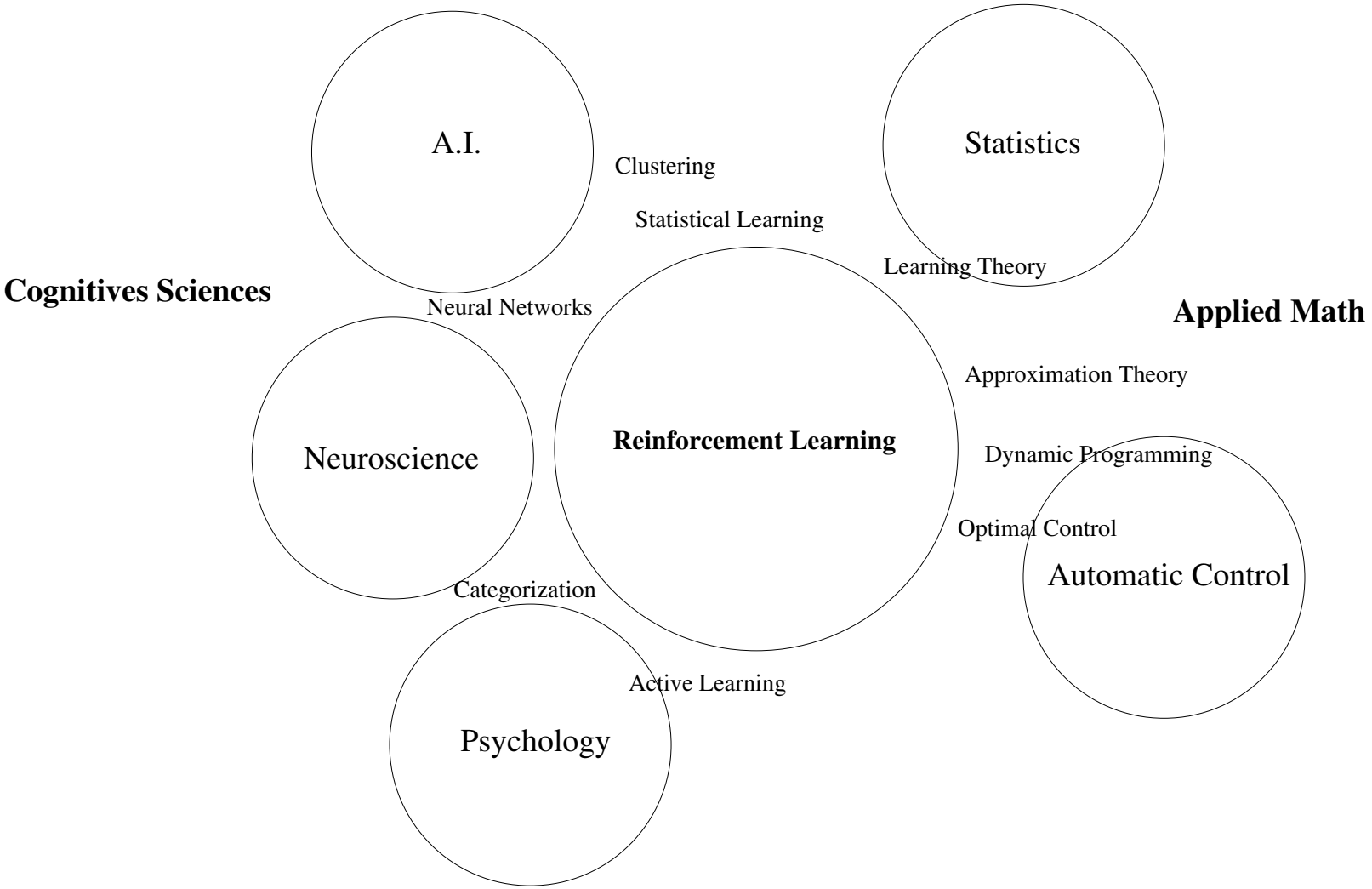
- $r(x, a, y)$ is the reward of transition (x, a, y).

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.

Reinforcement Learning



Example: the Retail Store Management Problem

- State space: $x \in X = \{0, 1, \ldots, M\}$.
- Action space: it is not possible to order more items that the capacity of the store, then the action space should depend on the current state. Formally, at state x, $a \in A(x) = \{0, 1, \ldots, M - x\}$.
- Dynamics: $x_{t+1} = [x_t + a_t - D_t]^+$.
  Problem: the dynamics should be Markov and stationary!
- The demand D_t is stochastic and time-independent. Formally, $D_t \overset{i.i.d.}{\sim} \mathcal{D}$.
- Reward: $r_t = -C(a_t) - h(x_t + a_t) + f([x_t + a_t - x_{t+1}]^+)$.

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Computational limits of learning methods

Recently mathematical papers started studying the fundamental limits of statistical learning methods when constrained to be computationally efficient.

A computational lower bound is a minimax lower bound where whe only look to estimators $\hat{h} \in \mathcal{H}$ with a reasonable computational complexity (e.g., polynomial-time algorithms):

$$\inf_{\hat{h} \in \mathcal{H}} \sup_{\theta} \mathbb{E}\left[R_\theta\left(\hat{h}\right)\right] \geq \dots$$

Example in high-dimensional linear regression: Yuchen, Wainwright and Jordan (COLT'14) proved in [13] that the restricted eigenvalue that appears in Lasso oracle bounds is necessary for all polynomial-time algorithms (assuming NP not in P/poly).

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Distributed statistics

Suppose we want to perform a statistical task on a very large dataset X_i, $1 \leq i \leq N$. Due to space or time limitations this dataset is split accross m different machines.

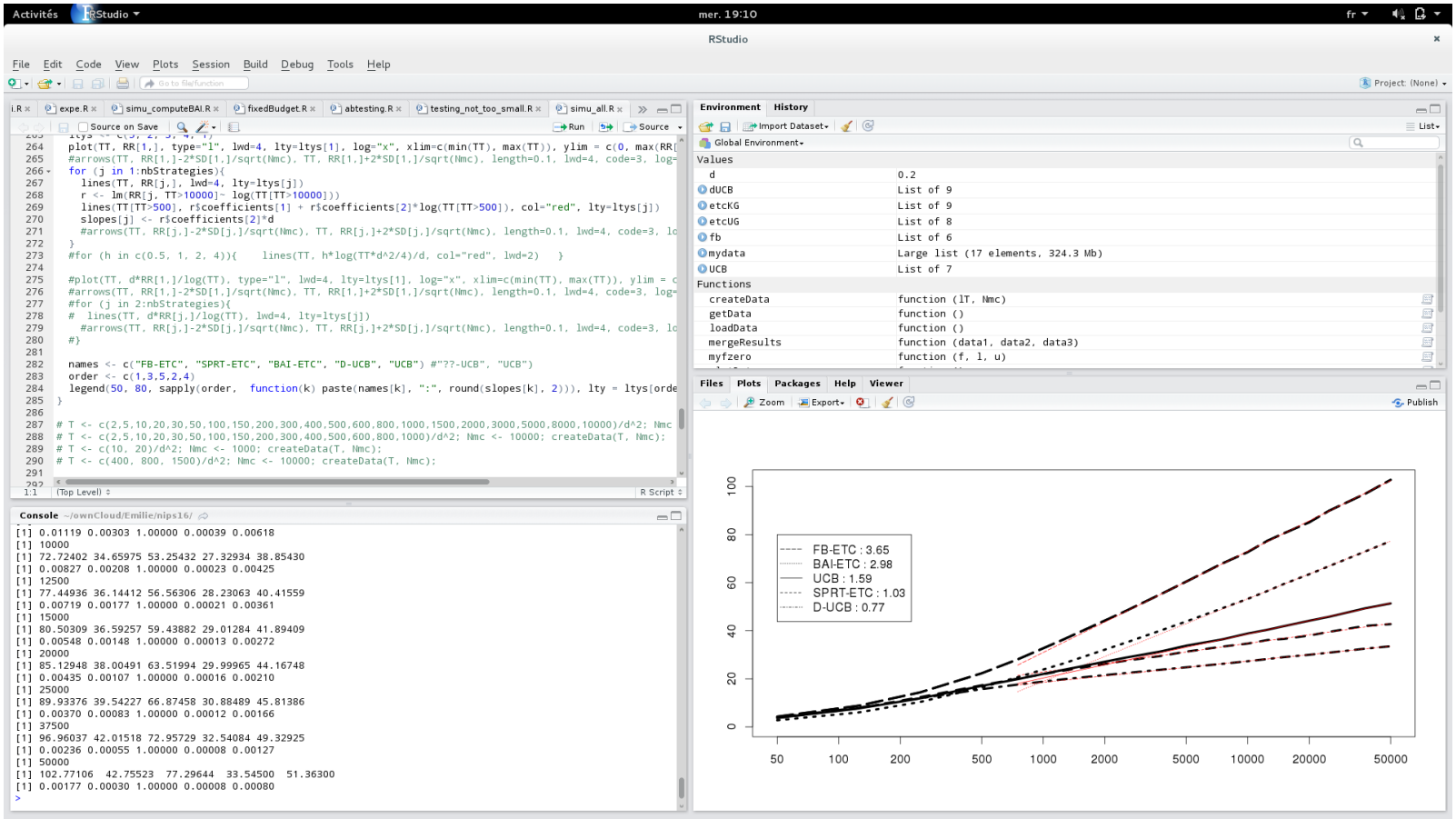
- Each machine j processes its own dataset (of size N/m) and communicates its answer $\hat{h}_j$ to a central node.
- Suppose we have a constraint B on the overall communication budget (compare to previous slide: computational constraint).
- Question: what is the best statistical performance we can guarantee with a finite communication budget B?

Recent theoretical contributions: upper and lower bounds for various statistical tasks by Zhang et al. [14, 15].

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

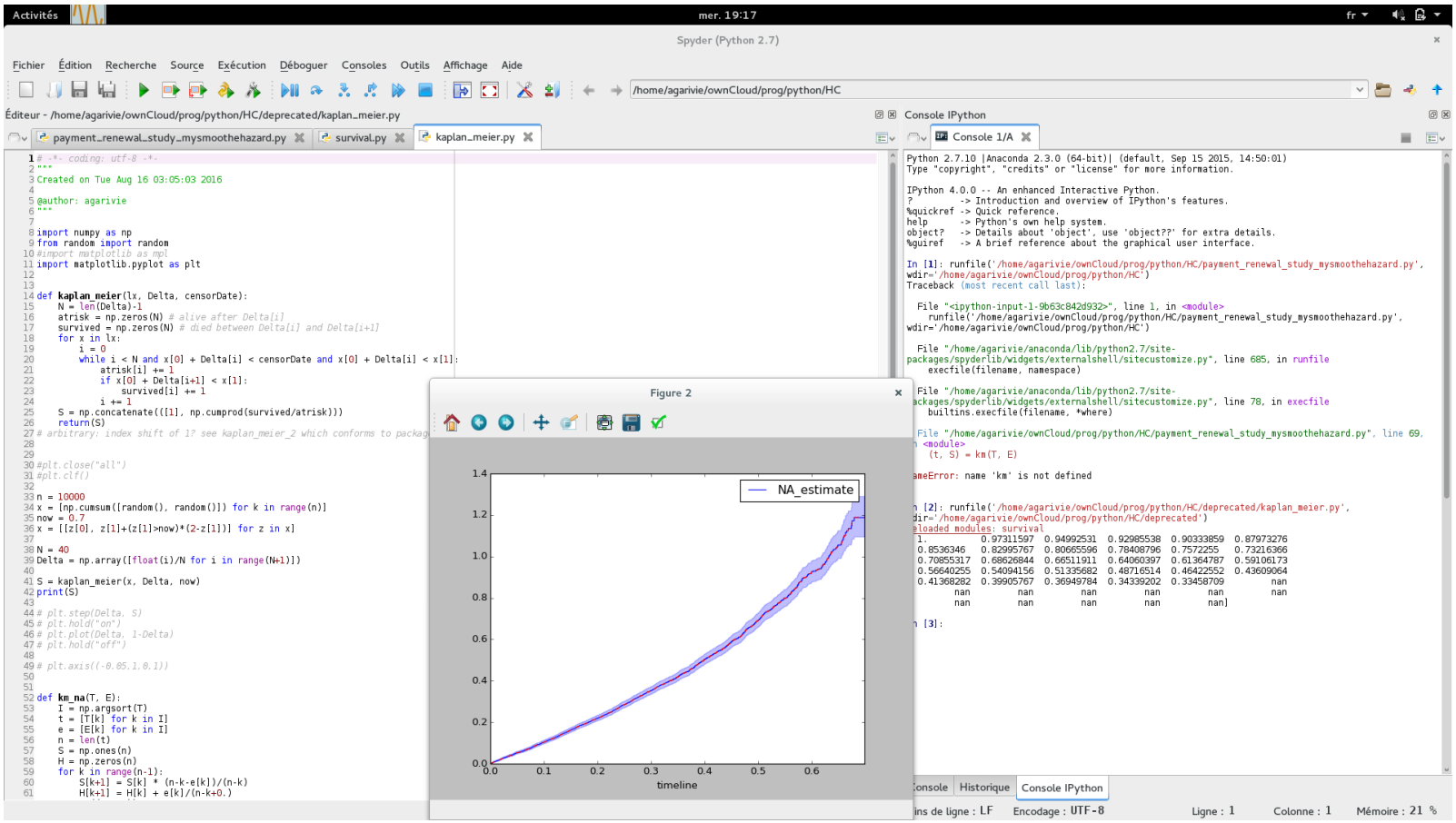
R



-: old language, slow

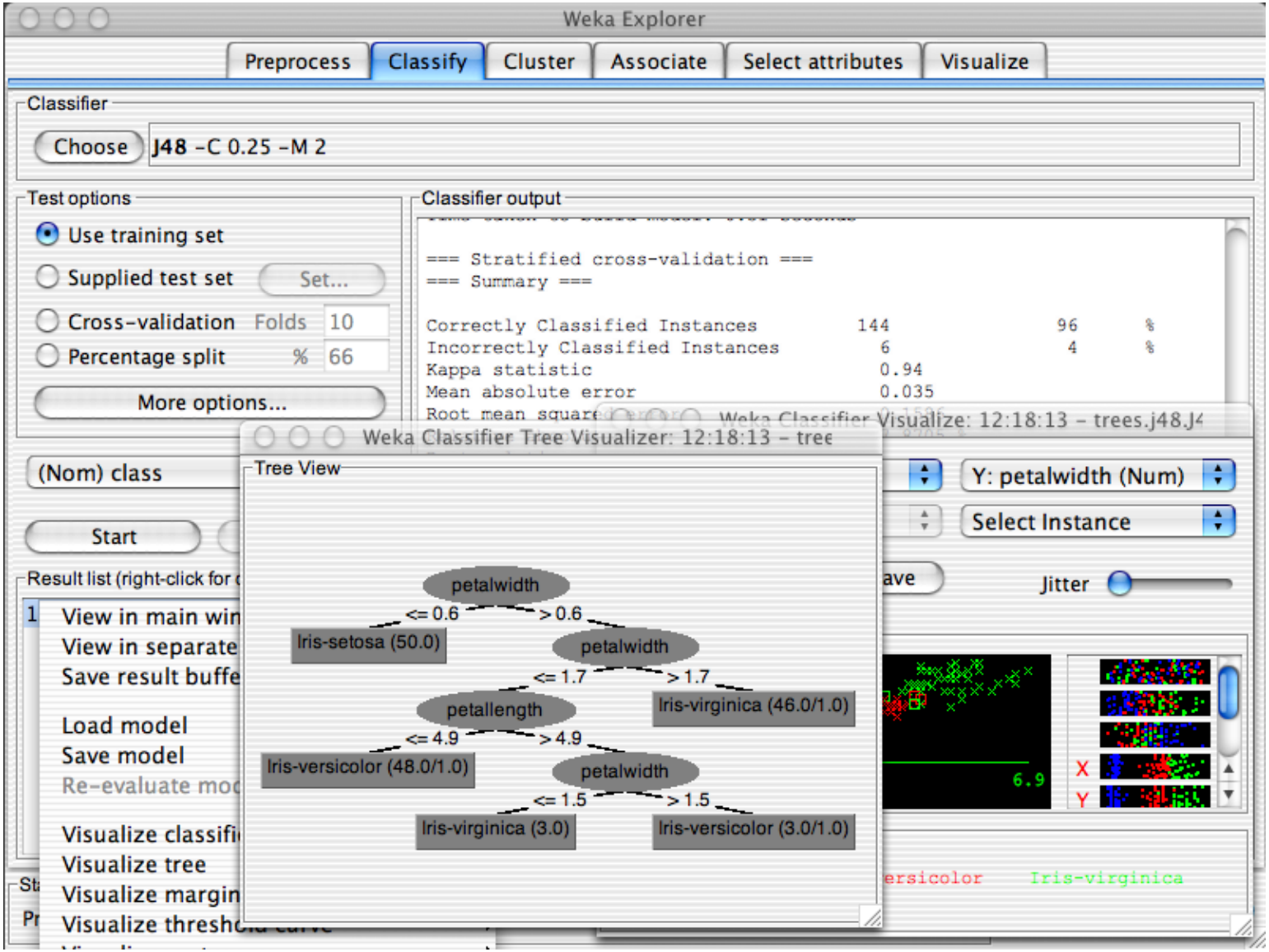
+: huge library, well-known

Python



+ general-purpose language, growing fast
- smaller library, not (yet?) so good for statistical analysis

Knime, Weka and co



\+ conceptual overview of processes, click-and-see

\- less control and analysis

Data Repositories

ex: MNIST, iris, adult, abalone,...
Kaggle Challenges, ...

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

Main ML labs in the world

- Berkeley / Stanford / UCLA
- MIT, Carnegie Mellon, UPenn, Caltech, Harvard, Georgia Tech, Duke,...
- China (with HK)
- France: INRIA, Paris, Lille, Toulouse, Marseille,...
- Israel: Weizmann / Technion / Tel Aviv
- Oxford, Cambridge, UCL
- Zurich, EPFL
- Amsterdam
- google / deepmind
- Microsoft research (Theory+ML Seattle, Bangalore)
- University of Alberta

Main ML labs in France

- **INRIA-Sierra** ENS, Paris
  Learning and optimization, LASSO, SGD
  Francis Bach
- **INRIA-Sequel** + **Modal** Cristal, Lille
  sequential learning, decision making under uncertainty, bandit problems, reinforcement learning
  Alessandro Lazaric, Rémi Munos, Michal Valko, Philippe Preux, Emilie Kaufmann
- **INRIA-Sequel** + **Modal** Cristal, Lille
  generative models
  Christophe Biernacki, Alain Célisse, Benjamin Guedj,...
- **INRIA-TAO** Saclay
  Optimization, Evolutionary Algorithms, Geometry
  Isabelle Guyon, Michelle Sebag, Olivier Teytaud, Odalric Maillard, Yann Ollivier, Balazs Kegl (ass)

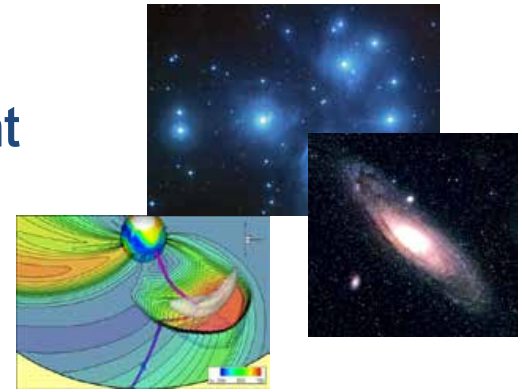
Main ML labs in France

- **ENSAE** ParisTech Arnak Dalalyan, Vianney Perchet, Pierre Alquier
- **Saclay** Univ Orsay, X
  Sélection de modèles
  Chritophe Giraud, Syvlain Arlot, Pascal Massart, Stéphane Gaïffas
- **Telecom** ParisTech Ranking
  Stéphane Clémençon
- **groupe "ML" Mines de Saint-Etienne** Marc Sebban
  Metric learning, transfer learning, anomaly detection, data mining
- **Grenoble** Optimization, Statistical models
  Anatoli Judistski, Florence Forbes,
- **Huawei** France
  ML and communication
  in construction

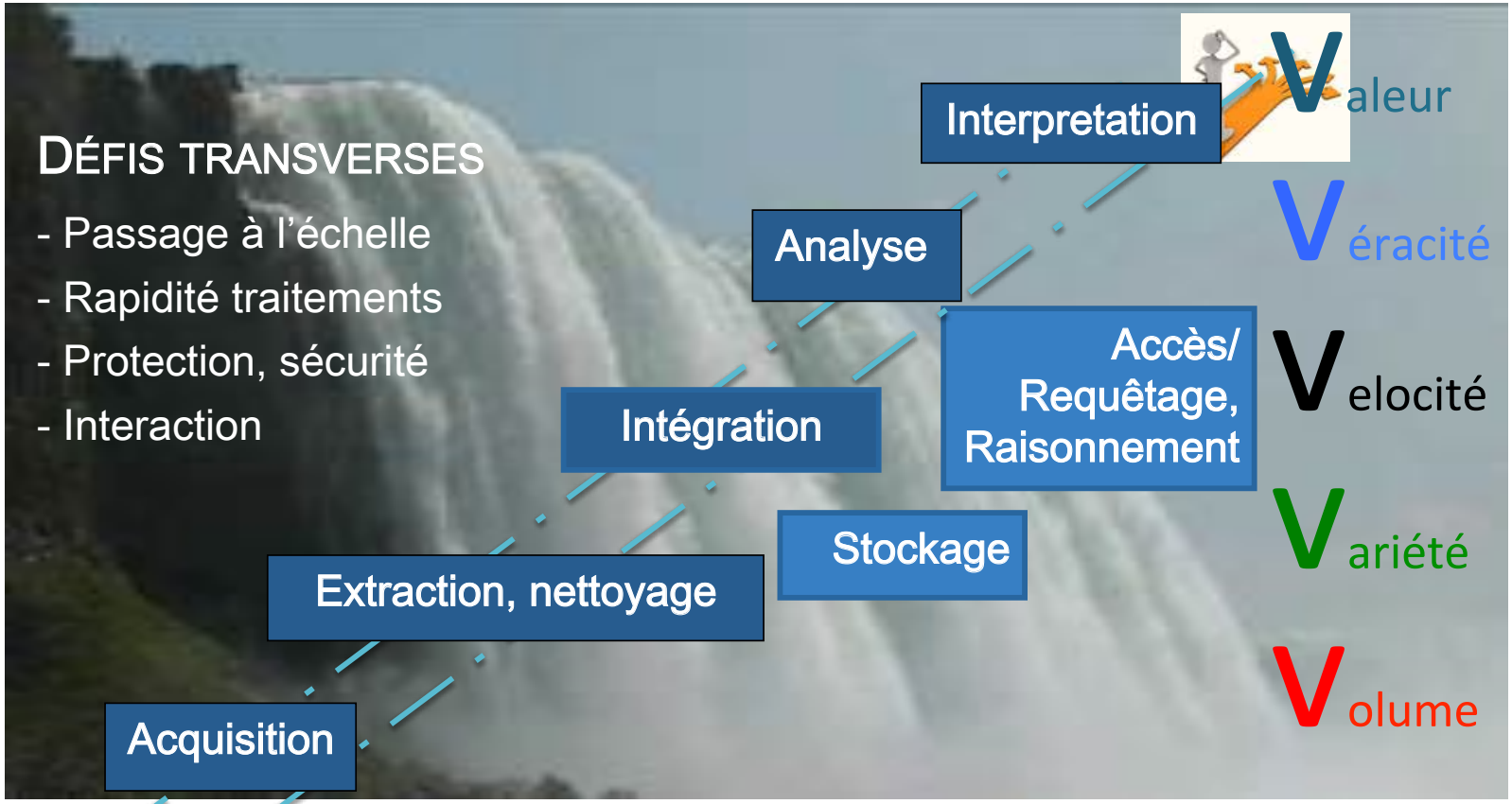
Contexte

Masses de données scientifiques provenant d'instruments, de simulations numériques, de multiple dispositifs de collecte de données

- volume, vélocité, variété, véracité
- changement de paradigme de traitement
  - approche traditionnelle : les besoins métiers guident la conception de la solution
  - approche par les données : les sources de données guident la découverte



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.

Objectifs du GdR

- Créer un écosystème pour impulser une dynamique de rapprochement / travail entre
  - chercheurs de différentes disciplines / communautés

Action

coordonne diverses activités interdisciplinaires.

Manifestation : journée animation, groupe de travail, études, écoles, ateliers … (problématiques, applications, données)
Participation d'industriels possible et souhaitée.

Réseau :

- formation : contribuer à la formation de chercheurs/spécialistes en Sciences des Données (espace doctorants)
- innovation : liens avec le monde socio-économique

- Participer à la formation des «data scientist » (informatique, statistique, mathématiques, intelligence métier)

Outline

What is Machine Learning?
Artificial Intelligence
Machine Learning
Big Data
Conferences
Main Topics of Present Interest
Deep Learning
Sequential Decision Making
Reinforcement Learning
Optimization
Sparsity
Image and Vision
Computational limits of learning methods
Distributed statistics
Statistics and Privacy
Tools for Machine Learning
ML in France, Toulouse, CIMI
ML Community
ML in Toulouse

IMT

- Statistics Modelling: Fabrice Gamboa, Jean-Michel Loubes
- Sequential Methods: Aurélien Garivier, Sébastien Gerchinovitz
- Model Selection: Xavier Gendre, Béatrice Laurent-Bonnneau, Cathy Maugis
- ML&bio: Philippe Besse, Pierre Neuvial
- Gaussian Processes: Jean-Marc Azaïs, François Bachoc
- Optim: Nicolas Couellan, Sophie Jan, Aude Rondepierre
- Image: Jérôme Fehrenbach, François Malgouyres, Laurent Risser,
- ML&probability: Gersende Fort, Jonas Kahn
- + specialists of statistics, stochastic processes, concentration,...

IRIT

- ML et IA: Mathieu Serrurier, Gilles Richard, Jérome Mengin, Henri Prade...
- NPL: Stergos Afantenos, Philippe Muller, Tim van de Cruys...
- Optimization: Édouard Pauwels
- ML&SP: Jean-Yves Tourneret, Cédric Févotte...
- planning: Maarike Verloop
- Spectral Clustering: Sandrine Mouysset, Daniel Ruiz
- Recommender Systems & Data Mining: Yoann Pitarch (IRIS), Josiane Mothe SIG...
- Speech analysis: SAMOVA
- Image processing: TCI

LAAS & al.

- Optim: Jean-Bernard Lasserre, Didier Henrion...
- Robotics (planning)
- ENAC
- ...?

For Further Reading I

Wikipedia.
The free encyclopedia.

Léon Bottou.
The Tradeoffs of Large-scale Learning.
NIPS'07 Tutorial.

Léon Bottou, Franck E. Curtis, Jorge Nocedal.
Optimization Methods for Large-Scale Machine Learning
Arxiv:1606.04838, NIPS'16 tutorial, 2016.

Massih-Reza Amini
Apprentissage machine, De la théorie à la pratique - Concepts fondamentaux en Machine Learning
Editions Eyrolles, 2015.

R. Duda, P. Hart, D. Stork
Pattern Classification
Wiley Interscience, 2001

For Further Reading II

T. Hastie, R. Tibshirani, J. Friedman
The Elements of Statistical Learning
Springer, 2001
online: `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

S. Tufféry
Data Mining
Technip

P. Besse et al.
`http://wikistat.fr/`

D. Blackwell.
An analog of the minimax theorem for vector payoffs.
Pacific J. Math., 6(1):1–8, 1956.

N. Cesa-Bianchi and G. Lugosi.
Prediction, Learning, and Games.
Cambridge University Press, 2006.

For Further Reading III

J. Hannan.

Approximation to Bayes risk in repeated play.

Contributions to the theory of games, 3:97–139, 1957.

H. Robbins.

Some aspects of the sequential design of experiments.

Bulletin of the American Mathematical Society, 55: 527–535, 1952.

Yuchen Zhang, Martin J. Wainwright, and Michael I. Jordan

Lower bounds on the performance of polynomial-time algorithms for sparse linear regression.

Proceedings of COLT'14.

For Further Reading IV

Y. Zhang, J. C. Duchi, M. I. Jordan, and M. J. Wainwright.
Information-theoretic lower bounds for distributed statistical estimation with communication constraints.
Proceedings of NIPS'13.

Y. Zhang, J. C. Duchi, and M. J. Wainwright.
Communication-efficient algorithms for statistical optimization.
Journal of Machine Learning Research 14, pp. 3321-3363, 2013.