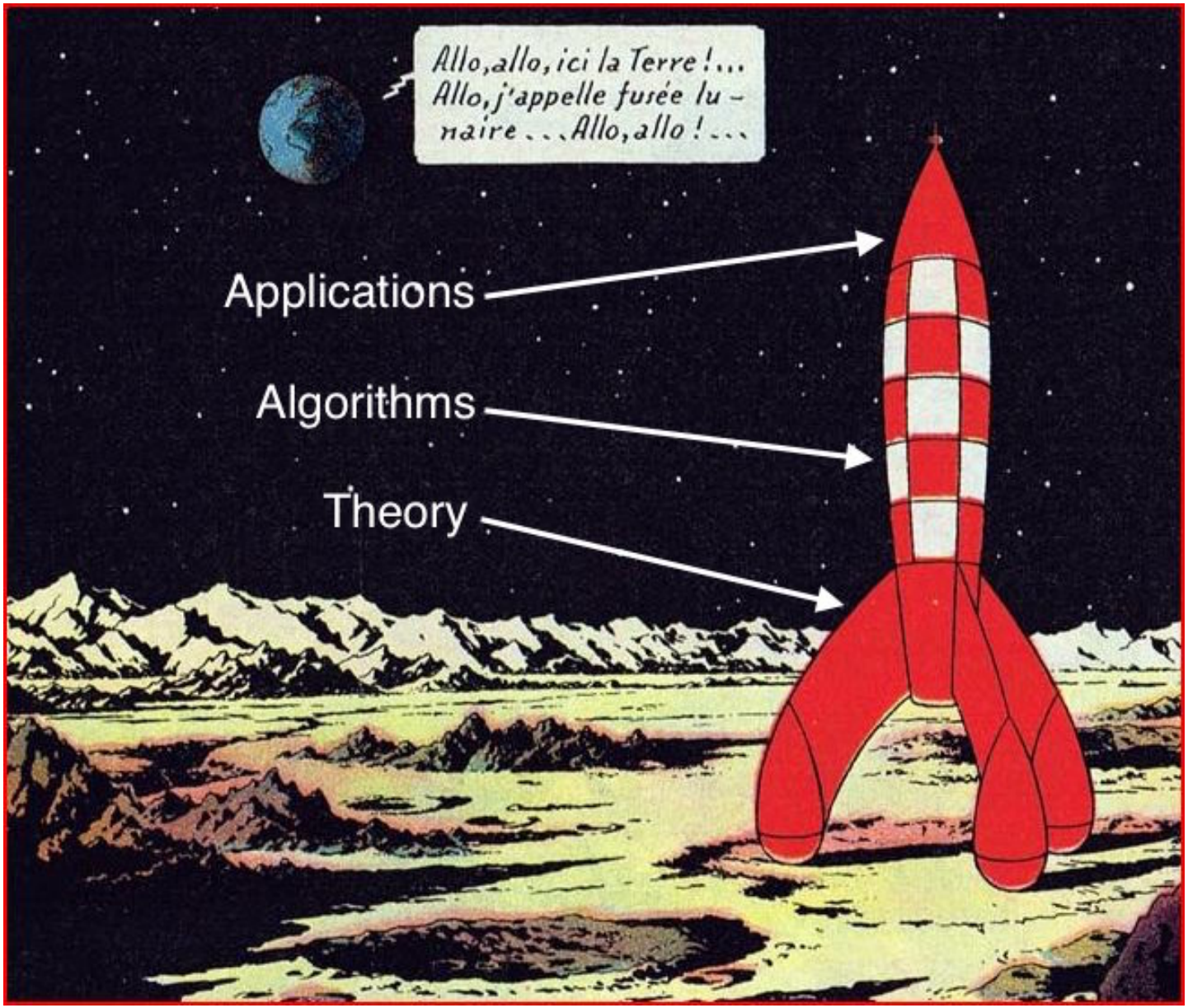
Reconnaissance de caractères manuscrits par factorisation de matrices

Le point de vue quasi-bayésien

Benjamin Guedj, Ph.D.

https://bguedj.github.io
Inria Lille - Nord Europe


Allo, allo, ici la Terre !...
Allo, j'appelle fusée lu-
naire ...Allo, allo !...
Applications
Algorithms
Theory

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

In the (~~rather~~ not so?) long term: mimic the inductive functioning of the humain brain to develop an artificial intelligence.

In the Big Data Era, very dynamic field at the crossroads of Computer Science and Statistics. Strategic focus at Inria!

Probabilistic framework: n-sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(X, Y) \in \mathcal{X} \times \mathcal{Y}.$$

Probabilistic framework: n-sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(X, Y) \in \mathcal{X} \times \mathcal{Y}.$$

We want to infer the link between the explanatory variable X and the response variable Y, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(X)$ is a "good" approximation of Y.

Probabilistic framework: n-sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(X, Y) \in \mathcal{X} \times \mathcal{Y}.$$

We want to infer the link between the explanatory variable X and the response variable Y, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(X)$ is a "good" approximation of Y.

- Classification: $\mathcal{Y}$ is discrete.
- Regression: $\mathcal{Y}$ is a continuum.

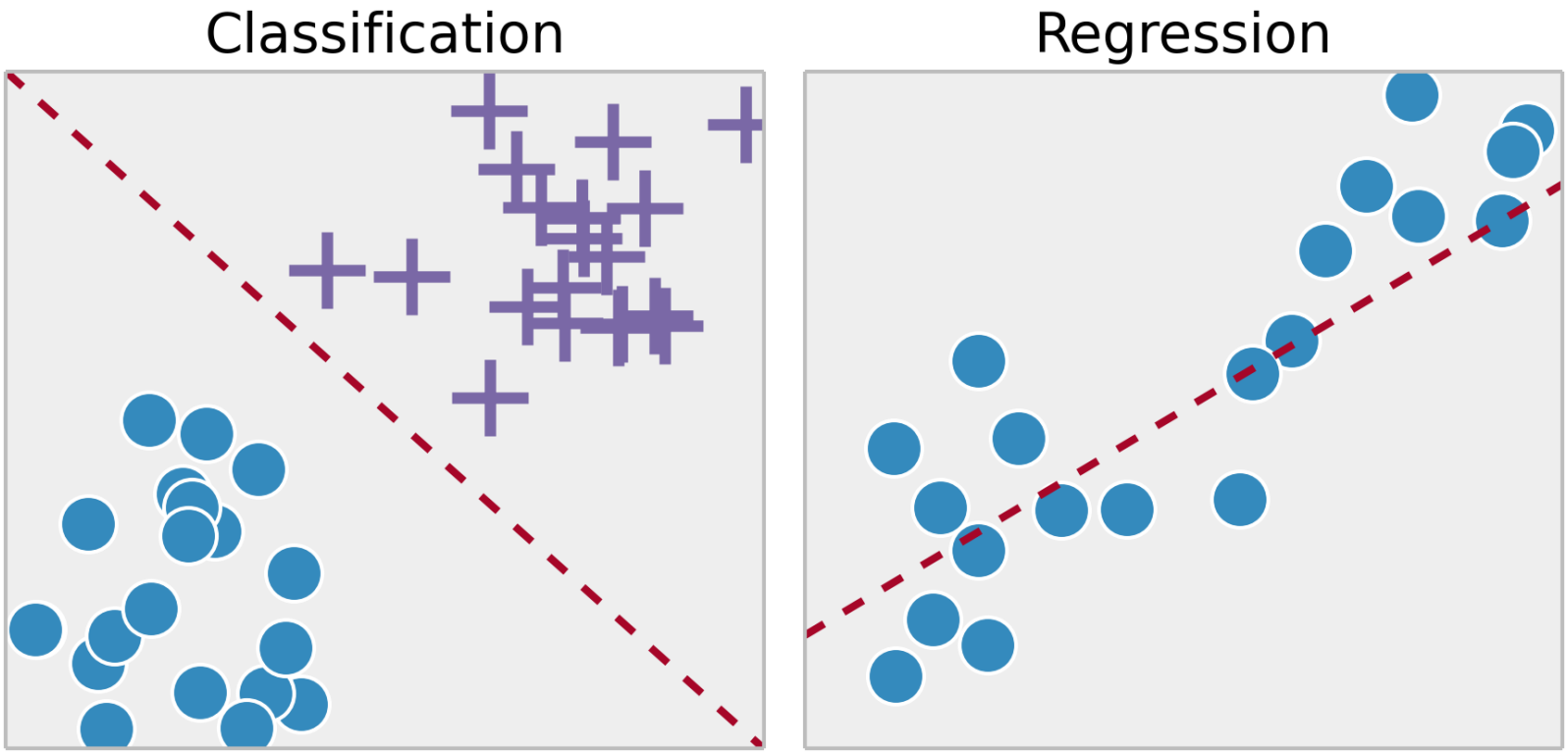
Probabilistic framework: n-sample $\mathcal{D}_n = (X_i, Y_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(X, Y) \in \mathcal{X} \times \mathcal{Y}.$$

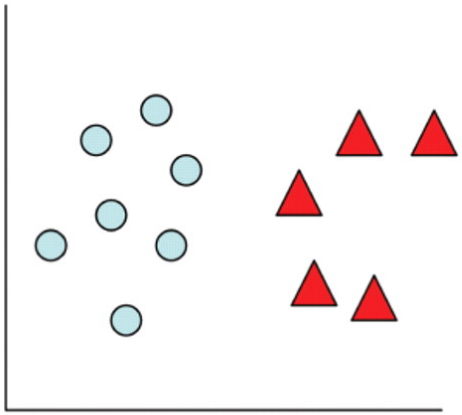
We want to infer the link between the explanatory variable X and the response variable Y, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(X)$ is a "good" approximation of Y.

- Classification: $\mathcal{Y}$ is discrete.
- Regression: $\mathcal{Y}$ is a continuum.
- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

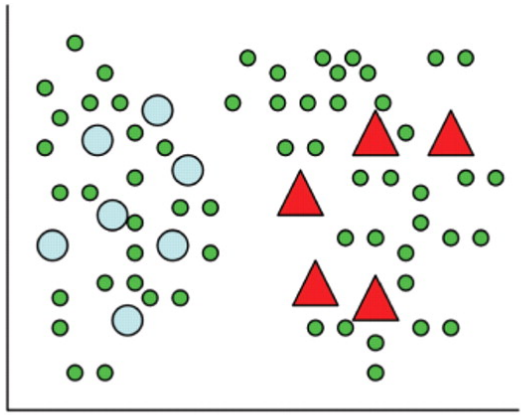
- Supervised learning: all of the Y_is are observed.



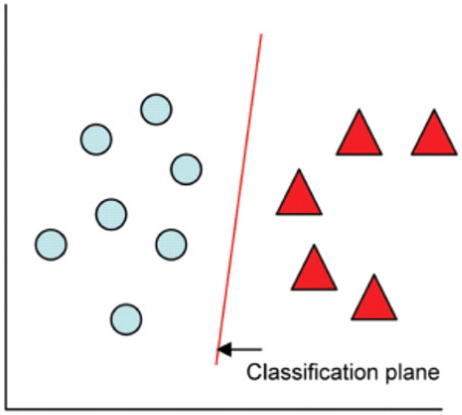
- Semi-supervised learning: some of the Y_is are observed (labeling is expensive or difficult).



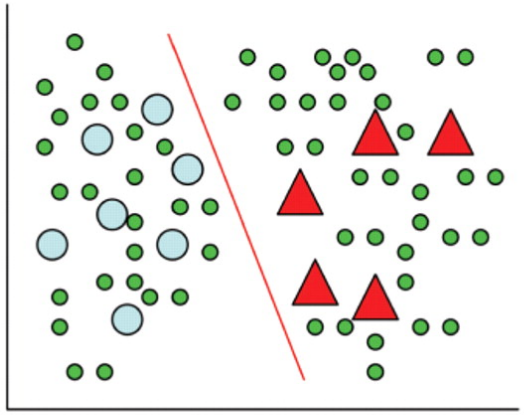
Labeled Data
(a)



Labeled and Unlabeled Data
(b)

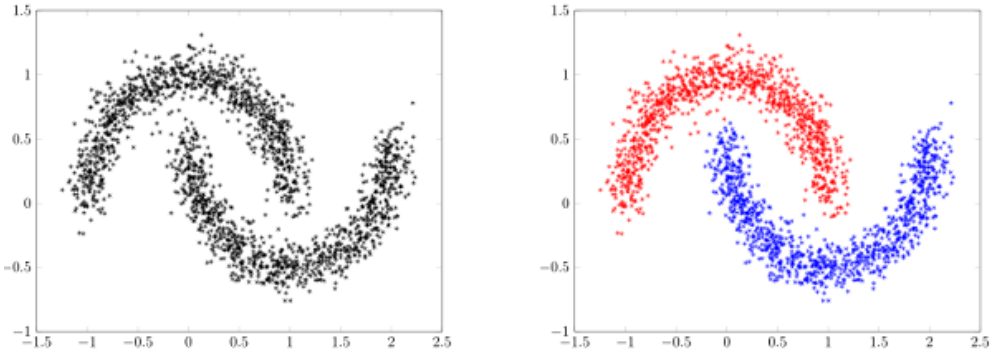


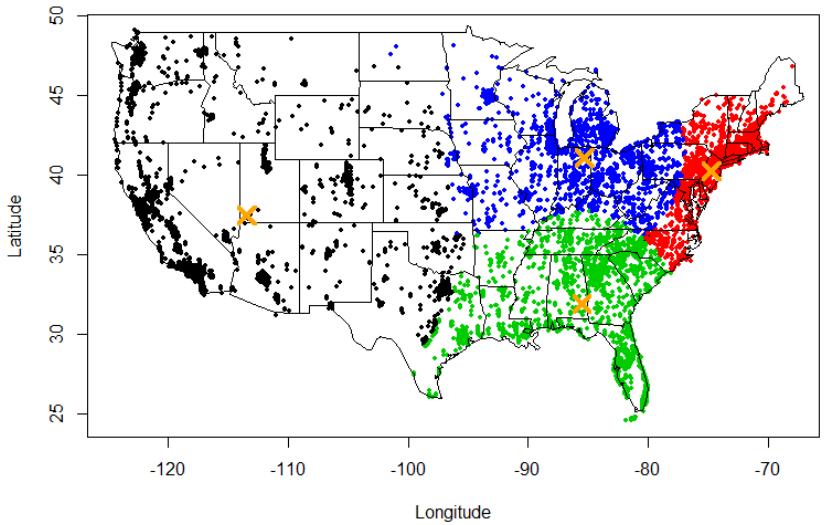
Supervised Learning
(c)



Semi-Supervised Learning
(d)

- Unsupervised learning: none of the Y_is are observed (detect patterns).





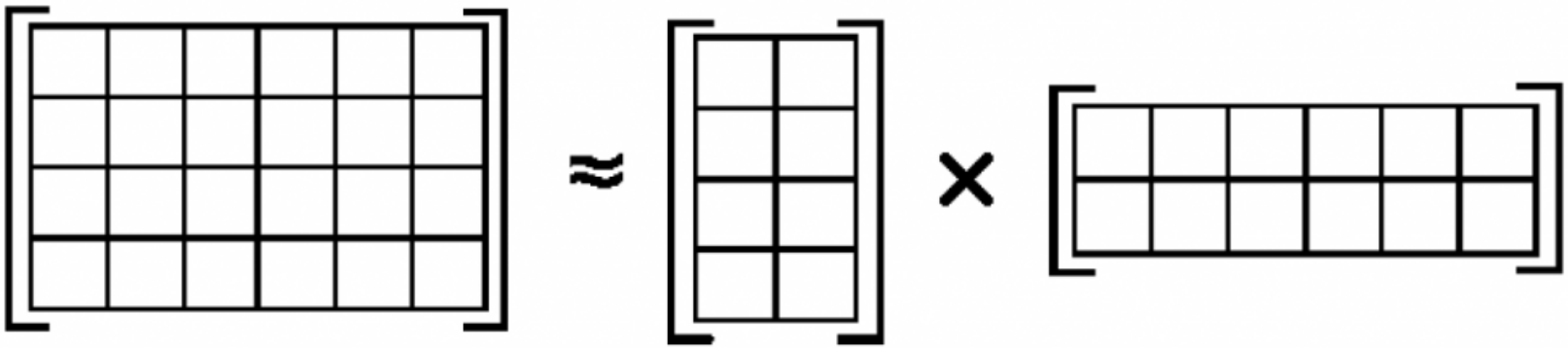
- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

Non-negative Matrix Factorization (NMF)

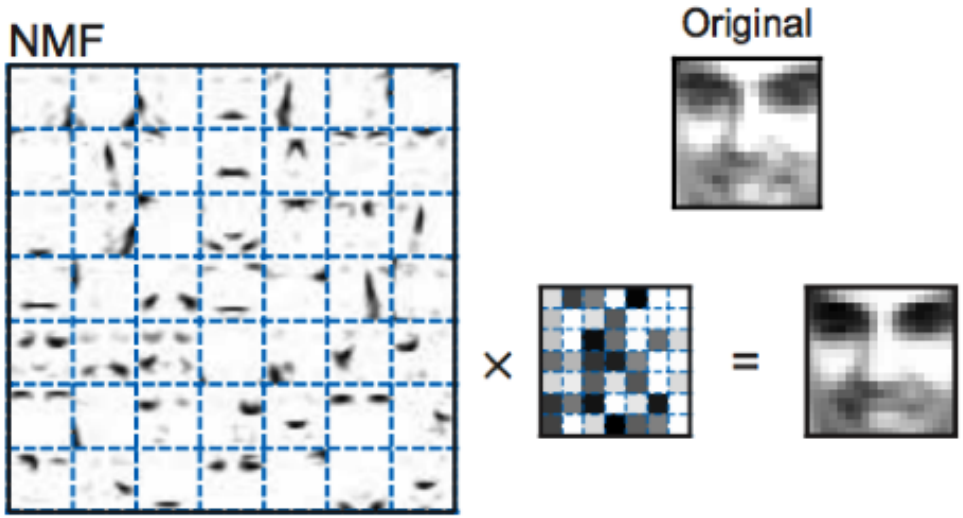
NMF amounts to decompose an $m_1 \times m_2$ matrix M as a product of two low rank matrices with non-negative entries.

$$M \approx UV^\top,$$

where U is $m_1 \times K$ and V is $m_2 \times K$, and $K \ll m_1 \wedge m_2$.



- Audio/video source separation $\longrightarrow$ [Video] [Demo]
- Denoising/restoration
- Speaker recognition
- Music classification
- Image processing: object recognition
- Topics extraction in texts
- Recommender systems
- ...



We observe an $m_1 \times m_2$ matrix Y and we assume

$$Y = M + \mathcal{E}$$

with $\mathbb{E}(\mathcal{E}) = 0$ and $\mathbb{V}(\mathcal{E}) = \sigma^2 \mathrm{Id}$.

The goal is to find a "good" factorization of M, in the sense of the Frobenius norm

$$\|A\|_F = \sqrt{\langle A, A \rangle_F},$$

$$\langle A, B \rangle_F = \mathrm{Tr}(AB^\top) = \sum_{i=1}^{p} \sum_{j=1}^{q} A_{i,j} B_{i,j}.$$

The quasi-Bayesian approach in one slide!

The quasi-Bayesian approach in one slide!

Loss:

$$R_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \phi(X_i)).$$

(examples: $\ell(a,b) = (a-b)^2$, $\ell(a,b) = \|a-b\|_F$, etc.)

The quasi-Bayesian approach in one slide!

Loss:

$$R_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \phi(X_i)).$$

(examples: $\ell(a,b) = (a-b)^2$, $\ell(a,b) = \|a-b\|_F$, etc.)

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

The quasi-Bayesian approach in one slide!

Loss:

$$R_n(\phi) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, \phi(X_i)).$$

(examples: $\ell(a,b) = (a-b)^2$, $\ell(a,b) = \|a-b\|_F$, etc.)

Set of candidates $\mathcal{F}$ equipped with a probability measure π (prior).

Quasi-posterior

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right)\pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

In general, $\exp\left(-\lambda R_n(\cdot)\right)$ is not a likelihood (hence the term quasi-Bayesian).

Quasi-Bayesian learning

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right) \pi(\cdot)$$

Quasi-Bayesian learning

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right) \pi(\cdot)$$

MAQP

$$\widehat{\phi}_\lambda \in \underset{\phi \in \mathcal{F}}{\arg\max}\ \widehat{\rho}_\lambda(\phi).$$

Quasi-Bayesian learning

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right) \pi(\cdot)$$

MAQP

$$\widehat{\phi}_\lambda \in \underset{\phi \in \mathcal{F}}{\arg\max}\ \widehat{\rho}_\lambda(\phi).$$

Mean

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \widehat{\rho}_\lambda(\mathrm{d}\phi).$$

Quasi-Bayesian learning

$$\widehat{\rho}_\lambda(\cdot) \propto \exp\left(-\lambda R_n(\cdot)\right) \pi(\cdot)$$

MAQP

$$\widehat{\phi}_\lambda \in \underset{\phi \in \mathcal{F}}{\arg\max}\ \widehat{\rho}_\lambda(\phi).$$

Mean

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \widehat{\rho}_\lambda(\mathrm{d}\phi).$$

Realization

$$\widehat{\phi}_\lambda \sim \widehat{\rho}_\lambda.$$

And so on.

Probably Approximately Correct (PAC) oracle inequalities

Probably Approximately Correct (PAC) oracle inequalities

For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$.

Probably Approximately Correct (PAC) oracle inequalities

For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha}\right\}\right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$.

Probability (Error of our estimator $\leq$

Error of the oracle + Decaying reminder term) $\geq 1 - \epsilon$.

Probably Approximately Correct (PAC) oracle inequalities

For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha}\right\}\right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$.

Probability (Error of our estimator $\leq$
Error of the oracle + Decaying reminder term) $\geq 1 - \epsilon$.

Typical rates in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

Probably Approximately Correct (PAC) oracle inequalities

For any $\epsilon > 0$,

$$\mathbb{P}\left(R\left(\widehat{\phi}_\lambda\right) - R^\star \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^\star + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$.

Probability (Error of our estimator $\leq$
Error of the oracle + Decaying reminder term) $\geq 1 - \epsilon$.

Typical rates in the literature

- $\alpha = \frac{1}{2}$ (slow rate)
- $\alpha = 1$ (fast rate)

Let $d = \dim(\mathcal{X})$

- $\Delta(\phi, \epsilon) \propto d + \log \frac{1}{\epsilon}$
- $\Delta(\phi, \epsilon) \propto \log d + \log \frac{1}{\epsilon}$

The PAC-Bayesian theory

The PAC-Bayesian theory

PAC-Bayesian Theory

$$= \text{Quasi-Bayesian learning} + \text{PAC oracle inequalities}$$

- Shawe-Taylor and Williamson (1997). A PAC analysis of a Bayes estimator, *COLT*
- McAllester (1998). Some PAC-Bayesian theorems, *COLT*
- McAllester (1999). PAC-Bayesian model averaging, *COLT*
- Catoni (2004). Statistical Learning Theory and Stochastic Optimization, Springer
- Catoni (2007). PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, IMS

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg\min_x f(x)$).

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg\min_x f(x)$).

Input: tolerance ϵ, initialization x_0, step size α.

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg\min_x f(x)$).

Input: tolerance ϵ, initialization x_0, step size α.

While $f'(x_k) \geq \epsilon$

$$x_{k+1} = x_k - \alpha f'(x_k)$$

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg\min_x f(x)$).

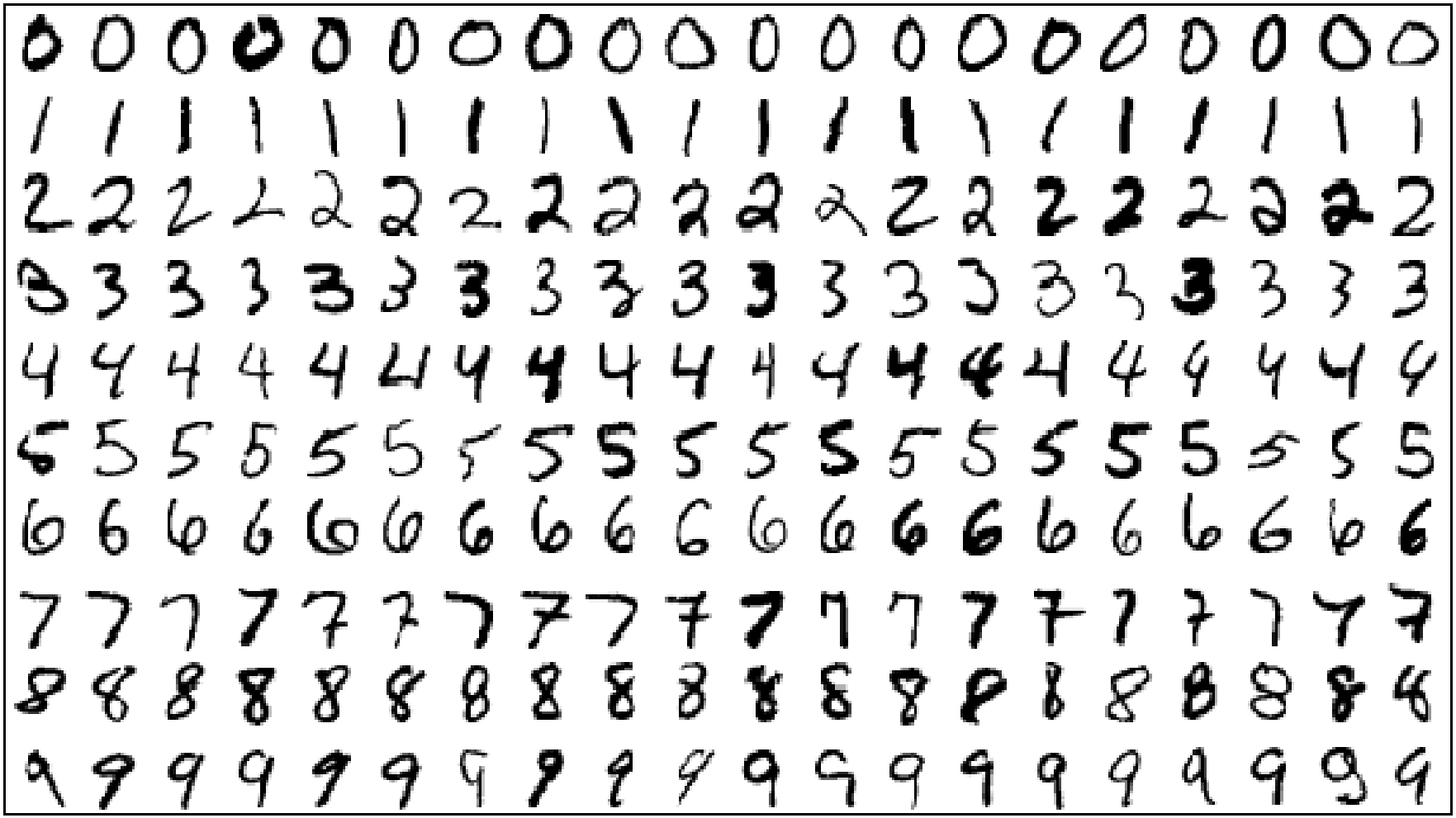
Input: tolerance ϵ, initialization x_0, step size α.

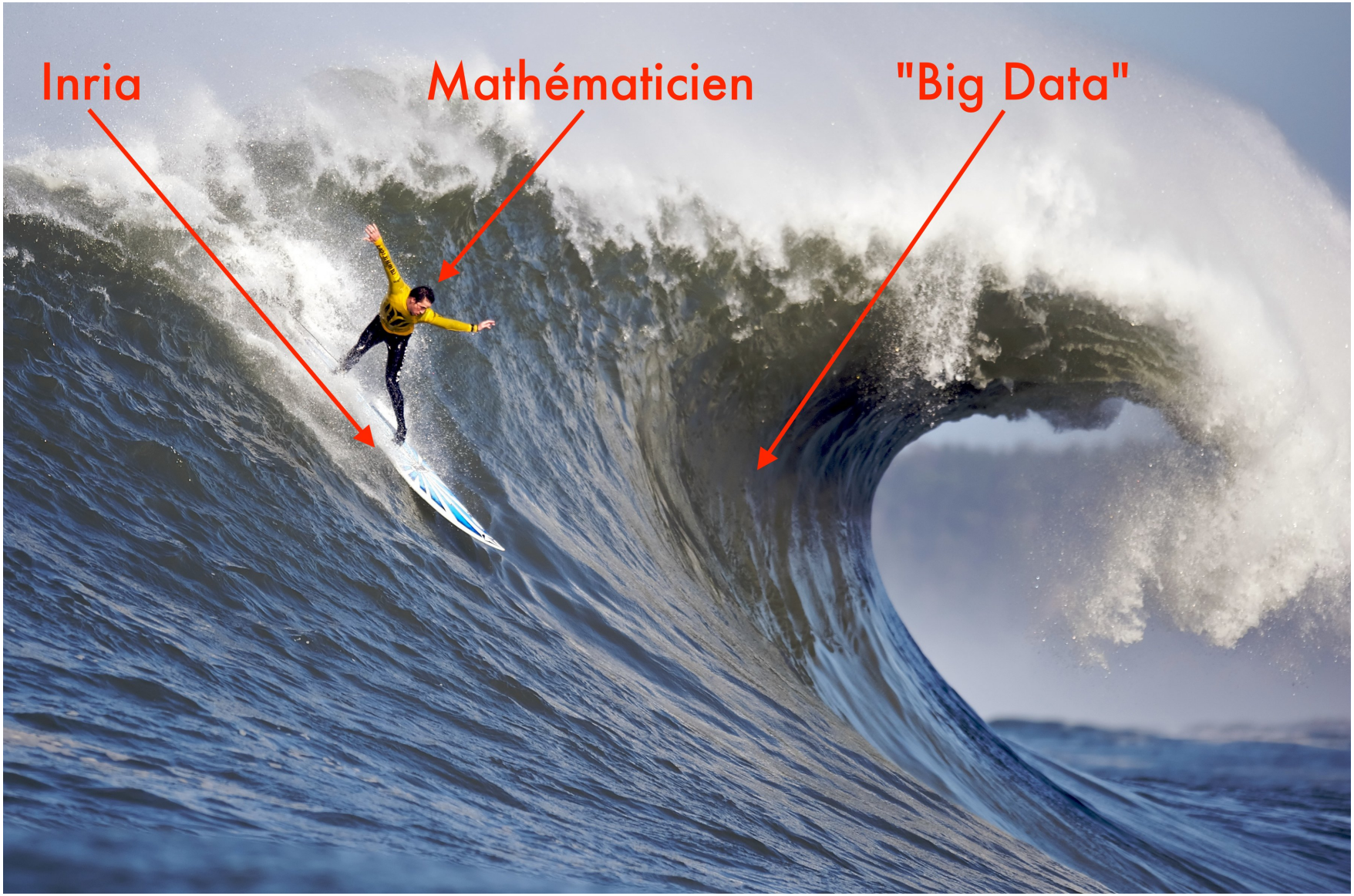
While $f'(x_k) \geq \epsilon$

$$x_{k+1} = x_k - \alpha f'(x_k)$$

Output: $(x_k)_{k=0}^K$ and $x_K \simeq \arg\min_x f(x)$.

Digits recognition




Inria
Mathématicien
"Big Data"