# Présentation Thématique CSP
# BIG DATA
# Institut de Mathématique de Toulouse

Francesco Costantino, Fabrice Deluzet, Aurélien Garivier

Institut de Mathématiques de Toulouse
Université Paul Sabatier, Toulouse

18 Avril 2017

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

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

VLDB

XLDB

Big Data

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 |

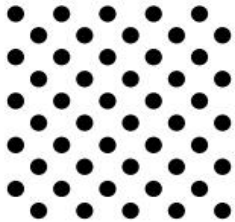
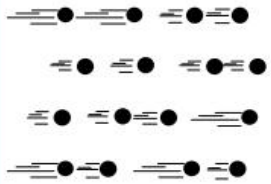
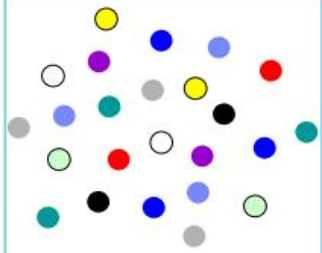
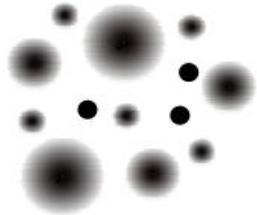
Source: *The Economist* Yotta and Zetta were added in 1991; terms for larger amounts have yet to be established.

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

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

# Complexité multidimensionnele des Big Data

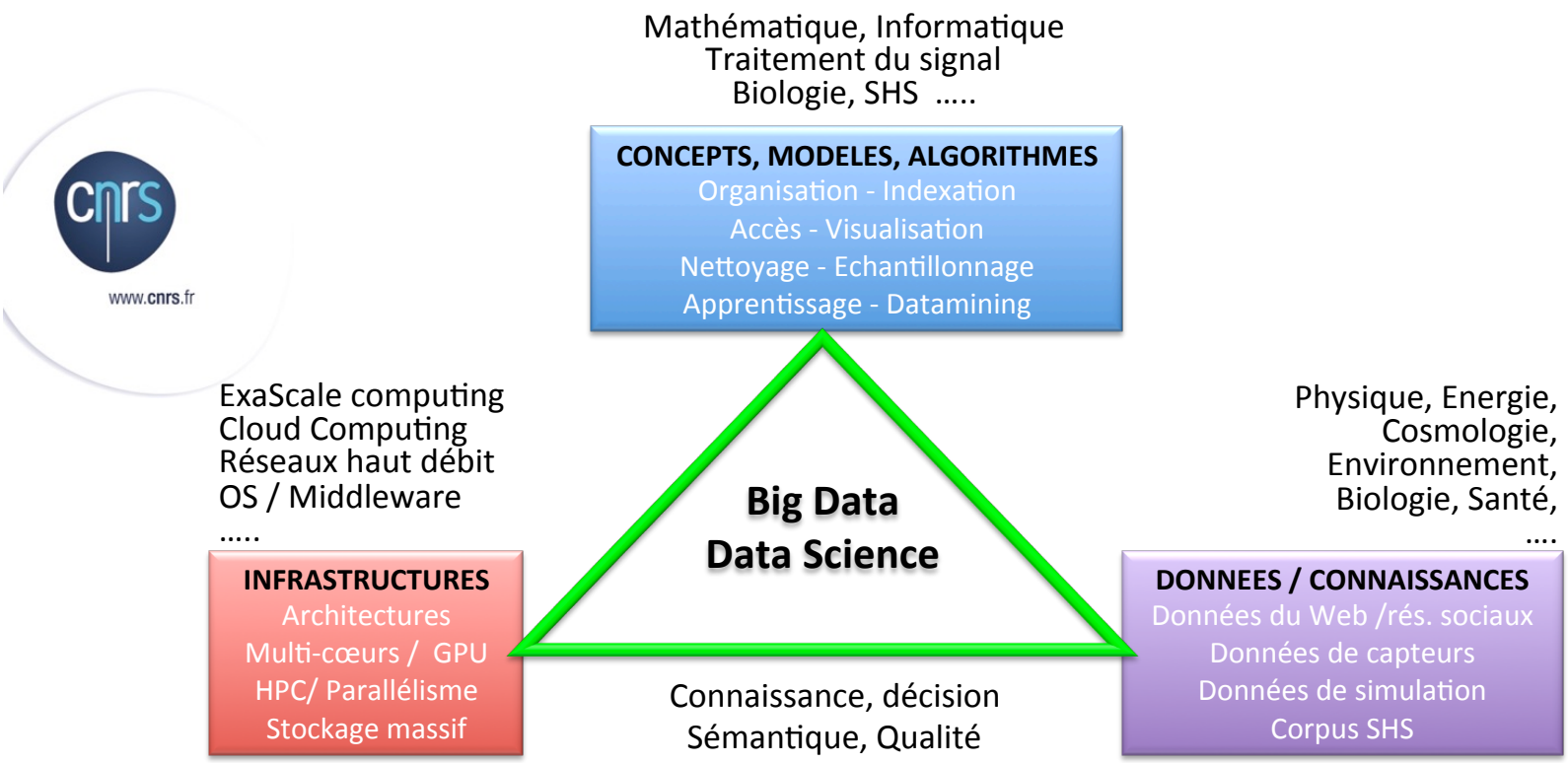
www.cnrs.fr

| Volume | Velocity | Variety | Veracity* |
|---|---|---|---|
|  |  |  |  |
| **Data at Rest** | **Data in Motion** | **Data in Many Forms** | **Data in Doubt** |
| Terabytes to exabytes of existing data to process | Streaming data, milliseconds to seconds to respond | Structured, unstructured, text, multimedia | Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations |
| • Nouvelles archi. de stockage<br>• Nouvelles archi. d'interopérabilité | • Défi pour les réseaux de communication<br>• Nouveaux modèles de calcul sur des flux | • Nettoyage et transformation<br>• Fusion de données | Nouveaux modèles de qualité (données & processus de traitement) |

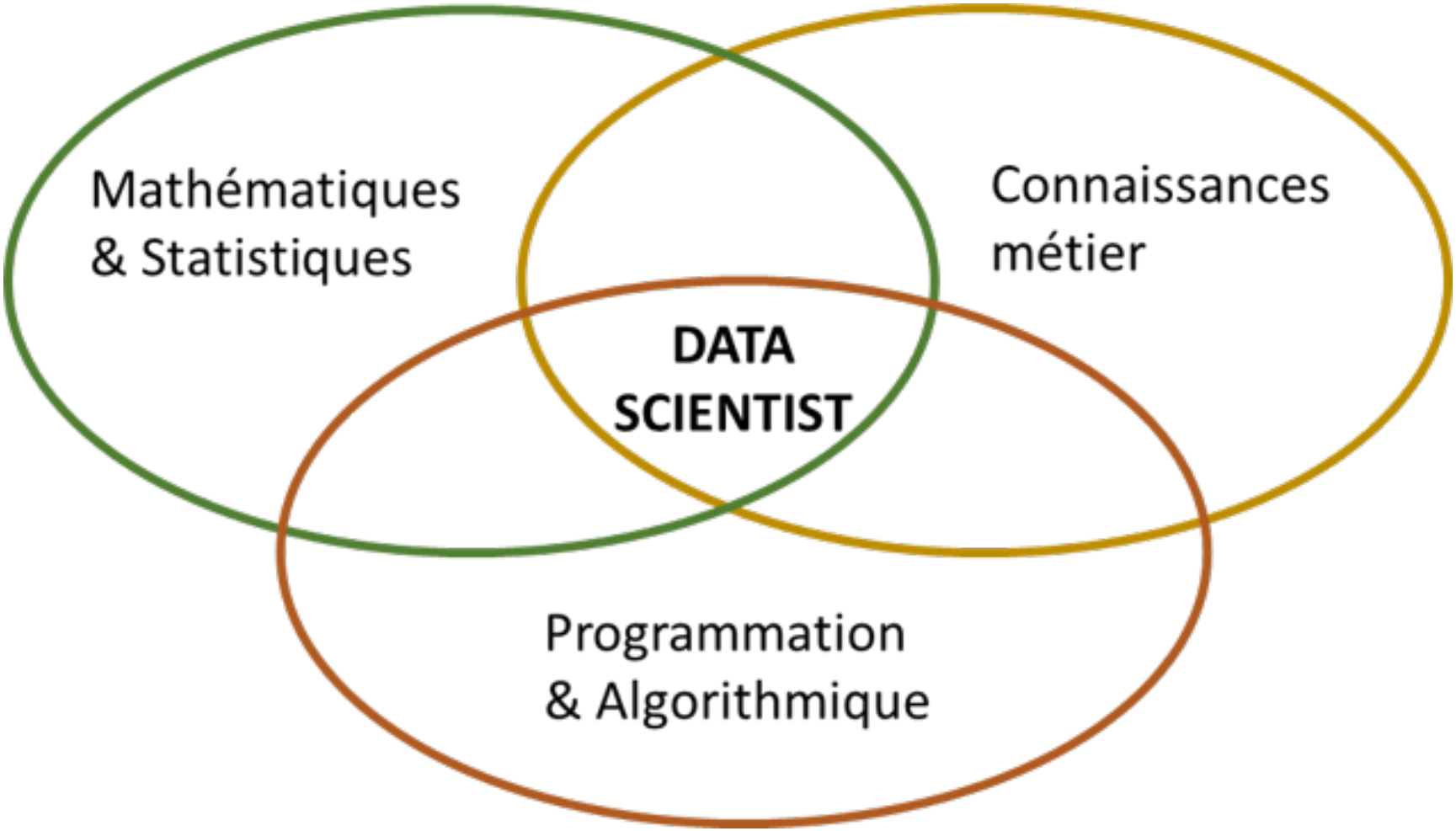
http://www.datasciencecentral.com/profiles/blogs/data-veracity

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

# Conclusion



# Qu'est-ce qu'un "data scientist" ?



# Sous-thèmes considérés et exclus

## Sujets abordés

- *Machine Learning et optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

## Sujets non abordés

- *Sécurité*
- *Grandes bases de données et leur stockage, préservation, architectures de systèmes*
- *Algorithmes distribués*
- *Raisonnement sur les données (logique)*
- *Fouille de données (algorithmique)*
- *Big Science*
- *Visualisation*
- *Aspects juridiques*
- . . .

## Sujets abordés

- *Machine Learning et optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

# Intelligence Artificielle (IA) : définition

## Intelligence des machines

- simuler les capacités cognitives des humains
  (big data : les humains apprennent en utilisant des sources de données très abondantes et diverses).
- une machine mime les fonctions cognitives que les humains associent à l'esprit humain, tels que apprendre ou résoudre un problème.

## Machine intelligente idéale =

agent rationnel flexible qui perçoit son environnement et qui prend des décisions qui maximisent ses chances de succès pour un but donné.

## Fondé sur le postulat que l'intelligence humaine

peut être décrite si précisément qu'on peut construire une machine la simulant.

# Intelligence Artificielle : entre rêve et réalité

## Buts opérationnels

- Robots autonomes pour réaliser des tâches pas trop spécialisées
- En particulier, vision + compréhension et production de langage (naturel)

## Tension entre les objectifs opérationnels et les buts philosophiques

- Au fur et à mesure que les machines accomplissent de plus en plus de tâches, des compétences qu'on pensait relever de l'intelligence sont progressivement retirées de la liste. Par exemple, la reconnaissance de caractères n'est plus considérée comme relevant de l'IA, mais comme une technologie de routine.
- Parmi les compétences encore classées en IA, il y a le jeu de go ou les voitures autonomes...

# ML : Apprendre des données et faire des prédictions

- Les algorithmes construisent un modèle à partir d'exemples donnés en entrée, dans but de faire des prédictions ou de prendre des décisions...
- ...plutôt que de suivre strictement une suite statique d'instructions : c'est utile quand il serait impossible ou inefficace de concevoir et de programmer de tels algorithmes.

## Analyse de données (Data Analytics)

- Le ML est utilisé pour concevoir des modèles complexes et des algorithmes qui conduisent eux-même à des prédictions - le mot commercial est souvent 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.
- évolution à partir de la reconnaissance de motifs (pattern recognition) de la computational learning theory en IA.

# Machine Learning : problèmes-types

- filtrage de spams, classification de textes
- reconnaissance de caractères (OCR)
- moteurs de recherche
- plateformes de recommandation
- outils de reconnaissance de la parole
- vision par ordinateur
- bio-informatique, analyse du génome, médecine (prédictive)

# Disciplines connexes

- **Statistique computationnelle** : centré sur la prédiction obtenue par l'usage de modèles statistiques nécessitant des calculs numériques intensifs (ex : méthodes bayésiennes)
- **Apprentissage statistique** : ML basé sur des méthodes statistiques, avec un point de vue statistique (garanties probabilistes : consistence, inégalités oracles, minimax...) $\rightarrow$ plus axés sur la corrélation, et moins sur la causalité
- **Data Mining** (apprentissage non supervisé) centré plutôt sur l'analyse exploratoire des données et la découverte de propriétés inconnues des données.
- Importance des méthodes basées sur les **probabilités** et les **statistiques** $\rightarrow$ **Data Science** (Michael Jordan)
- Liens très forts avec l'**optimisation mathématique**, qui fournit des méthodes, des concepts et des applications au ML.

# Classification supervisée : cadre statistique

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

# Minimisation du risque empirique

**Inégalité de Hoeffding :** avec probabilité au moins $1 - \eta$,

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

Problème : vrai pour chaque $h$ fixé mais pas pour $\hat{h}_n$ !

Ex : Prédiction of 10 lancers de Pile ou Face

Ex : régression polynomiale $\rightarrow$ sur-apprentissage

**Fléau de la dimension**

# Minimisation structurelle du risque

$\rightarrow$ loi des grands nombres uniforme — inégalité de Vapnik-Chervonenkis : si $\mathcal{H}$ a une dimension de VC $d_{\mathcal{H}}$, alors

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

Structure :

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

Ex : polynômes/splines de degré $m$, arbres de décision de profondeur $m$,...

Décomposition du risque en biais–variance

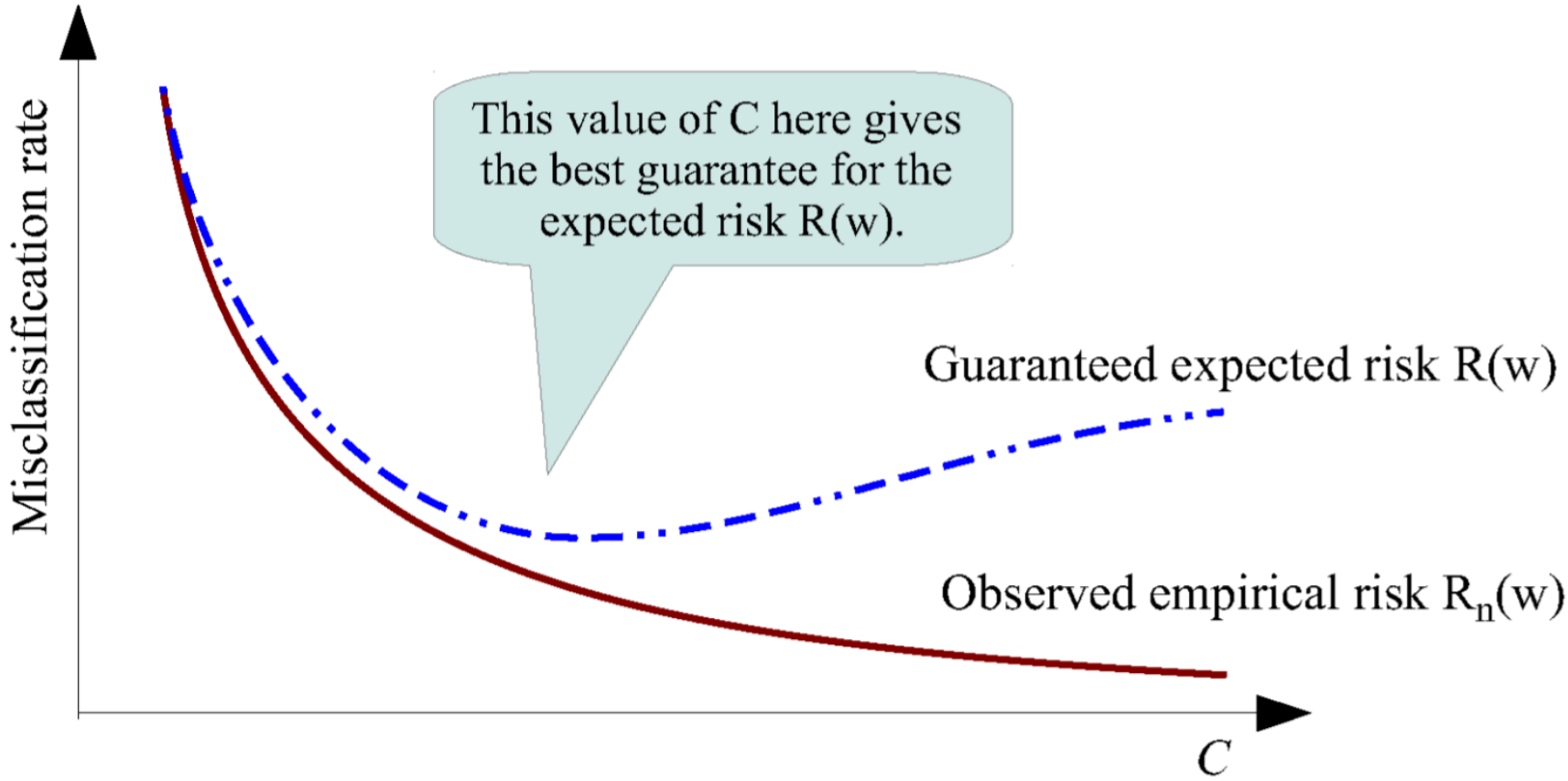
Minimisation structurelle du risque :

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

ou

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

# Structural Risk Minimization Tradeoff




Source : Bottou et al. tutorial on optimization

# Machine Learning et Statistique

- L'analyse de données (inférence, description) est le but des statistiques depuis longtemps.
- Le Machine Learning a des buts plus **opérationels** (ex : la consistence est importante en statistique mais moins en ML).
  Les modèles (quand il y en a) sont instrumentaux.
  Ex : modèle linéaires (jolie théorie mathématique) vs Random Forests (utilisation massive de modèles pauvres et sans signification propre).
- Machine Learning pour les big data : plus de séparation entre modélisation stochastique et optimisation (contrairement aux statistiques classiques).
- En ML, les données sont souvent là a priori (malheureusement).
- Pas de frontière infranchissable (la statistique aussi évolue).

# Quelle statistique pour les big data ?

- inégalités de concentrations et applications
- méthodes robustes (insensibles à des données polluées)
- tests multiples
- régression avec $n$ et $p$ grands
- parcimonie
- matrices aléatoires
- méthodes non paramétriques (notamment bayésiennes)
- traitement parallèle des données
- apprentissage transductif
- etc...

## Sujets abordés

- *Machine Learning et optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

# Apprentissage actif et séquentiel

- Traiter les données en flux
- Faire face à la non-stationarité :
  - détection de rupture
  - données ”adversariales”
- Choisir les données pour accélérer l’apprentissage
- Décision automatique en temps réel

→ Apprentissage par renforcement

- Lien avec le contrôle optimal stochastique

# Optimisation et Machine Learning

- Nouvelle décomposition du risque :

  risque = approximation + estimation *+ optimisation*

- Intérêt nouveau pour les méthodes de gradient stochastique (contre des données iid, ou bien ”adversariales”)
- Optimisation convexe (ex : régression logistique) ou non-convexe (ex : deep learning)
- Deep learning : nécessité d’une théorie de l’approximation
- Early Stopping : une autre façon d’éviter le sur-apprentissage

# Statistique et confidentialité

- Motivation : plusieurs exemples spectaculaires de désanonymisation de données (Netflix prize, Massachusetts Group Insurance Commission, etc.)
- Objectif : empêcher la désanonymisation tout en conservant la possibilité de faire des études statistiques
- Principe : introduire de l'aléa dans les données
- Objectif mathématiques : caractériser ce qu'il est possible de faire pour l'estimation fonctionnelle sous contrainte de confidentialité différentielle (et comment le faire !)
- cf théorie de l'information (débit-distorsion)

# Décision automatique et contraintes juridiques / éthiques

- **anonymisation des données**
  cf slide précédent
- **explicabilité des décisions**
  ex : pour les systèmes de référencement, la loi République numérique adoptée en oct 2016 oblige à préciser les modalités de référencement
  Quel contrôle ?
- **non-discrimination** des sous-populations
  création d'une plateforme de dénonciation des mauvaises expériences avec les algorithmes
- **distorsion de concurrence**
  barrière à l'entrée du fait d'avoir les données
- **ouverture/transparence** versus protection du secret d'affaire

# Ex : Projet "Algebra and approximation for ML"

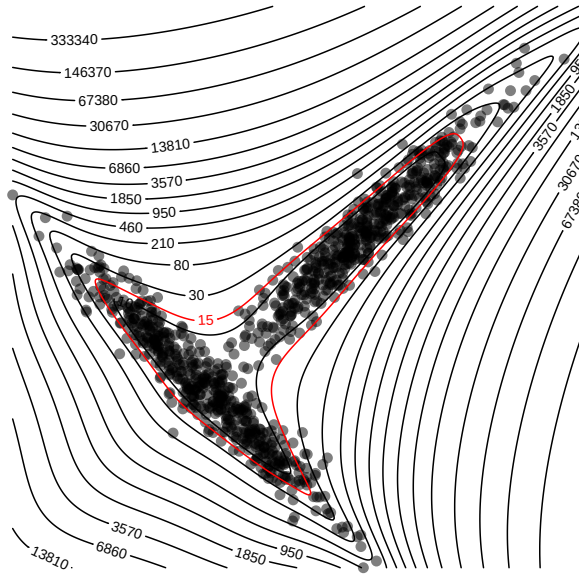
**Contexte :** Utilisation des *fonctions de Christoffel*.

A : Un objet connu depuis longtemps en *théorie de l'approximation* et en *algèbre*.

B : Capture d'importantes propriétés des *lois de probabilité*...

- localisation : *géométrie* du *support*,
- vraisemblance relative : *densité*

C : ... qui sont centrales dans beaucoup d'applications en *statistique* et en Machine Learning.




**Motivation :** A et B ne sont pas exploités pour C :

**Nouveaux outils pour le ML, mais nécessite :**
- Compréhension statistique
- Outils de calcul computationnels
- Meilleure compréhension algébrique

**Applications :**
- détection de nouveauté, d'outlier, de fraude
- apprentissage non supervisé
- reconnaissance de formes.

# Sources de financement

- nombre important d'équipes, de projets, etc.
- ex : appel d'offre Mastodons (1 projet CIMI-AOC)
- fonctionnement sur la base de challenges - ex : Kaggle
- une recherche "machine learning" dans la liste des projets ANR financés donne 40 réponses, et c'est loin d'être exhaustif...
- une recherche "machine learning" dans la liste des projets ERC financés donne 102 réponses (qui ont l'air pertinentes)
- 1 à Toulouse : ERC FACTORY : New paradigms for latent factor estimation (C. Févotte IRIT-ENSEEIHT+AOC)
- 1 GDR spécialement dédié aux "big data" en général : MaDICS

**Masses de données scientifiques**

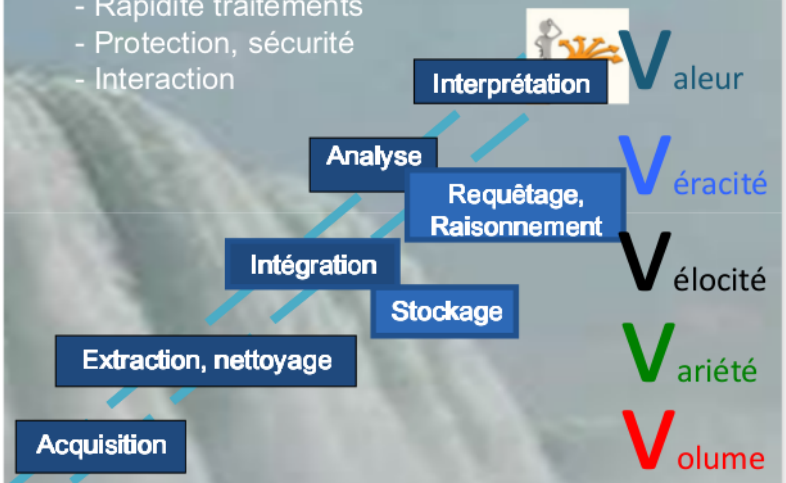
provenant d'instruments, de simulations numériques, de multiple dispositifs de collecte de données …

**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 transverses**

- Passage à l'échelle
- Rapidité traitements
- Protection, sécurité
- Interaction




*repenser les outils algorithmiques et mathématiques*

**Créer un écosystème pour impulser une dynamique de rapprochement entre**

- **chercheurs de différentes disciplines / communautés**
- **scientifiques en lien avec des masses de données**

**sur de nouvelles méthodes et outils pour la gestion, l'exploitation et la valorisation des données des Sciences**

- **Actions d'animation scientifique interdisciplinaires**
  **Journées d'animation thématiques, écoles, ateliers, séminaires, …**
  **Etudes de prospective, réponses collectives au niveau européen, …**
  **Participation souhaitée des industriels**
- **Lieu d'expertise pour les décideurs, favoriser la valorisation**
  **Identification de compétences, analyses stratégiques**
  **Liens avec le monde socio-économique**
- **Formation des « data scientists »**
  **Référentiel MOOC, espace doctorants**

# Principaux laboratoires dans le monde

- 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

# Principaux laboratoires en France 1/2

- **ENS, Paris**, including INRIA-Sierra
  Learning and optimization, LASSO, SGD
  Francis Bach, Jean Ponce, ~~Marc Lelarge~~,...
- **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

# Principaux laboratoires en France 2/2

- **ENSAE** ParisTech Arnak Dalalyan, Olivier Catoni, 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
- **ENS Cachan** Vianney Perchet, Nicolas Vayatis
- **Grenoble** Optimization, Statistical models
  Anatoli Judistski, Florence Forbes,...
- **Huawei** France : Noah's Ark Paris lab
  Moez Draïef,...
- **Facebook** France
  Yann Ollivier, Alessandro Lazaric,...

# IMT

- Statistics Modelling : Fabrice Gamboa, Jean-Michel Loubes
- Sequential Methods : Aurélien Garivier, Sébastien Gerchinovitz
- Model Selection : Xavier Gendre, Béatrice Laurent-Bonneau, 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, Pierre Weiss
- 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

# mais aussi...

- LAAS : Optimisation (Jean-Bernard Lasserre, Didier Henrion, etc...)
- Robotique (planification)
- ENAC
- ISAE
- ONERA
- SIGFOX ( ?)
- etc... ?

# Recruter en Machine Learning

- Grande difficulté d'AOC à répondre aux sollicitations de projets (ex : Airbus, Sigfox)
- Thématique très concurrentielle : de plus en plus de très bons profils, mais beaucoup de centres de recherche compétitifs (ex : Ecole Polytechnique, ENSAE, Telecom ParisTech, Google, Facebook, etc.)
- Grande volatilité et variabilité au niveau MCF, cf. équipes mentionnées plus haut. Ex : Ilaria Giulini, Alain Durmus
- Candidats PR potentiels :
  - Joseph Salmon (Telecom ParisTech)
  - Michal Valko (INRIA Lille)
  - Marianne Clausel (LJK Grenoble)
  - Poste "image" à l'INSA
  - ...

## Sujets abordés

- *Machine Learning et optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

# Un exemple : ITER (International Thermonuclear Experimental Reactor)

Produire de l'énergie (électrique) de la fusion nucléaire.
Point critique : maintenir le confinement du plasma.

ITER : « The physics is complicated, the mathematics is hard, the computations are extremely demanding [a] »

a. D. Batchelor, ORNL

- « The curse of dimensionality » : Eq. cinétique 7D.
- Problème multi-échelle.
- Anisotropie très sévère : $\mathcal{O}(10^8)$.

# Big Data, but how big ?

## ≪ Modélisation intégrée ≫ multi-physique, multi-modèle, multi-échelle

Une simulation directe (avec résolution uniforme des plus petites échelles) d'ITER sur une machine Peta ($10^{15}$) FLOPS nécessiterait $10^{6-12}$ l'âge de l'univers et toujours $10^{3-9}$ sur une machine Exa ($10^{18}$) FLOPS [a].

---

*a.* J.R. Cary et al. ≪ Concurrent, Parallel, Multiphysics Coupling in the FACETS Project ≫, in Journal of Physics : Conference Series 28, (2009).

# Big Data, but how big ?

## Objectif des codes de simulations

- Paramètres de discrétisation et ressources [a]
  - $4096 \times 16384 \times 32 \sim 10^9$ mailles en espace.
  - $100 \times 100 \sim 10^4$ mailles en vitesse (code GC) ou $10^{12}$ particules.
  - $10^6\ \Delta t$.
  - $10^6$ cœurs de calcul.
- Sauvegardes de reprise : 2Go par cœur toutes les 10 min. (=MTBF pour $10^6$ cœurs) soit 3.3 To/sec.
  Système de fichiers actuels : 0.2 To/Sec (Perf. crête).
- ITER Gyro-Cinétique Digital : $10^{18}$ FLOPS (ExaFLOPS)
  ITER Cinétique Digital : $>$ 10 EF.

---

*a.* VPIC, OISIRIS, M3D-K

# Besoin en mathématiques

## Analyse / Modélisation

- Analyse (Pb bien posé à la limite « Asymptotic well-posedness », convergence, stabilité, précision) de couplages de codes avec des échelles d'espace et de temps très différentes (turbulence/transport).
- Réduction de modèles, Analyse asymptotique, Modèles Hybrides (Couplage Fluide/Cinétique, Compressible/Non compressible, etc...).
- Quantification d'incertitude.
- Préconditoneurs basés sur la physique / propriétés des opérateurs.
- Construction de coordonnées adaptées (ou presque) à l'anisotropie : Champ magnétique dépendant du temps, formation d'îlots magnétiques, singularités.

# Besoin en mathématiques

## Méthodes numériques

- Méthodes hautes fidélités (précision en temps long, conservativité, positivité), Algorithmes symplectiques, réduction du bruit (méthode PIC).
- Compression de données (matrice hiérarchique, sparse grids).
- Méthode numériques ≪ efficientes ≫ : massivement parallèles, minimisant les mouvements de données.
- Raffinement de maillage adaptatif, analyse d'erreur.
- Méthodes implicites, Résolutions efficaces de systèmes non linéaires (JFNK), Préconditioneurs.

# Les « Digital twins »

## Simulation / Contrôle temps-réel grâce aux réseaux neuronaux.

BIG-DATA-DRIVEN Support-Vector-ML pour la prédiction des perturbations dans JET : détection de 80% à 90% des évènements avec 98% de fausses alarmes (<3% pour ITER).

- « Real-Time capable first principle based modeling of tokamak turbulent transport », Nucl. Fusion, 2015.
- « Modeling of transport phenomena in tokama plasmas with neural networks », Phys. Plasmas, 2014

## Questions actuelles

Comment appliquer les techniques du « machine learning » aux systèmes (données) décrit(e)s par des équations.

## Big Data and extreme scale computing (BDEC)

- www.exascale.org
- www.ipam.ucla.edu/programs/workshops/big-data-meets-computation/
- ...

# Simulation grande échelle (cinétique ) Financements / Projets

## EUROfusion (H2020)

- Enabling research 2014 : 42 projets ($\sim$ 500 kEuros/an/projet)
- Enabling research 2015-2017 : 18 projets ($\sim$ 500 kEuros/an/projet)

## Fédération de Recherche pour la Fusion par Confinement Magnétique

- Appel à Projets 2017 : 120 kEuros,
- Work Package EDUCATION 2017 : 465 kEuros.

## Projets INRIA

- TONUS : TOkamak NUmerical Simulations (Strasbourg, P. Helluy)
- CASTOR : Control, Analysis and Simulations for TOkamak Research (Nice, J. Blum).
- INRIA PROJECT LAB FRATRES (2015-2018) : Fusion ReAcTors REsearch and Simulations (Nice, H. Guillard).

# Simulation grande échelle (cinétique ) Financements / Projets

## Projets ANR

- Examag (2016-2018) Exascale simulations of the magnetic universe Problems with Multiple Objective Function (Allemagne, France, Japon).
- MOONRISE (2014-2018) MOdels, Oscillations and NumeRIcal SchEmes (Rennes,Toulouse,Cadarache).
- PEPPSI (2013-2017) Plasma Edge Physics and Plasma-Surface Interactions (Strasbourg,Nancy,Toulouse).
- CHROME (2012-2016) Waves in plasmas for heating anf reflectometry (UPMC,Nancy).
- ANEMOS (2011-2015) Advanced Numeric for ELMs : Modeling and Optimized Schemes (Nice,Cadarache,Bordeaux).

# Quelques noms à l'étranger et en France

- USA : Shi Jin (Université de Madison, USA), Quantification d'incertitude pour les modèles cinétiques, Luis Chacon (LANL), Methodes PIC implicites, JFNK, C.S Chang (Courant Institute of Mathematical Scences) Gyro-Kinetic PIC, F. Genko (UCLA), Bonolu (MIT), Hong Qin (Princeton Plasma Plasma Physics Laboratory), ...
- Chine : Institute of Natural Sciences (Shanghai), Harbin Institute of Technology, Beijin University.
- Europe : E. Sonnendrucker (Max-Planck-Institute, Garching, Allemagne), P. Ricci (EPFL, Suisse), ...
- France : CEA Cadarache, IRFM (P. Tamain, V. Grandgirard), Nice (N. Besse, H. Guillard, S. Minjeau), Marseille (M. Bostan, A. Nourri), Paris, Polytechnique (F. Golse), ENS (L. Saint Raymond), UPMC (B. Després), Rennes (N. Crouseilles, M. Lemou), Strasbourg/Nancy (P. Helluy).

# A Toulouse

## Autour des méthodes cinétiques

- Modèles réduits, méthodes numériques muti-échelle / multi-physique, HPC : F. Deluzet , F. Filbet , R. Loubère, J. Narski, C. Negulescu., M.-H. Vignal.
- Interaction avec l'IRIT pour le développement de méthodes scalable et ou de décomposition de domaine (Équipe projet).

## Dans d'autres domaines

Assimilation de données (glaciologie,climatologie) J. Monnier, F. Couderc, L. Amodei.

## Sujets abordés

- *Machine Learning (et deep learning) (Aurélien)*
- *Optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

# Image/Vision

## Représentation et apprentissage d'opérateurs

Images : $10^6 - 10^9$ pixels $\Longrightarrow$ matrices $10^{12} - 10^{18}$ coefficients.

- Représenter des opérateurs de façon compacte ;
- Décomposition pour faire des calculs efficaces ;
- Identification de structures pour faire de l'apprentissage.

## Traitement d'image

- Débruitage
- Recalage ;
- Problèmes inverses ;
- Segmentation ;
- ...

## Vision

- Extraction d'éléments caractéristiques (traitement d'image)
- Classifieurs (ML/DL)

☞ Reconstruction de scènes 3D

☞ Détection d'objets, d'anomalie (chaînes robotiques)

# Représentation/Apprentissage d'opérateurs

## Approximation, Factorisation de Matrices

$$M_1 \dots M_k \approx X \,,$$

- $k = 1$, approximation avec structure (parcimonie, non négativité, ...) ;
- $k = 2$, approximation de rang faible, factorisation de matrice positive, reconstruction de la phase ;
- $k > 2$, factorisation profonde : transformée de Fourier (FFT), en ondelettes (FWT), rotation de Jacobi,...

## Applications de la "deep matrix factorisation"

- Vision, traitement du langage, statistique, théorie de l'information, analyse numérique matricielle ;
- "Compréhension" du Deep learning, construction de dictionnaire, factorisation efficiente numériquement.

# Image / Vision

## Questions scientifiques

- Identifiabilité (Stabilité / des paramètres) ;
- Structure de la fonction objectif (minima locaux, globaux) ;
- Optimisation convexe et non convexe, non lisse, optimisation en grande dimension (stochastique) ;
- Capacité d'approximation (pour une structure de facteurs données) ;
- Apprentissage de dictionnaires (base optimale pour une classe de signaux), de variétés (trouver une variété capable de décrire un ensemble d'opérateurs : flou variable en microscopie).

## Domaines mathématiques

- Optimisation ;
- Analyse harmonique appliquée, théorie de l'approximation ;
- Géométrie des tenseurs de rang 1 (géométrie algébrique, complexité) ;
- Mesure, probabilité et statistique ;
- Traitement de données (images, sons).

# Image / Vision : quelques noms

## En France

- Jean-François Aujol, IMB, Bordeaux, Directeur du GDR MIA (Mathématiques de l'Imagerie et de ses Applications) ;
- Francis Bach, INRIA, ENS, Paris (ERC SIERRA, Sparse Structure Methods for Machine Learning) ;
- Antonin Chambolle, CNRS, Ecole Polytechnique ;
- Albert Cohen, UPMC, Paris ;
- Jalal Fadili, ENSI Caen ;
- Rémi Gribonval, INRIA Rennes (PANAMA, PArsimony & New Algorithms for Audio & Signal Modeling) ;
- Stéphane Mallat, École Polytechnique, Paris ;
- Gabriel Peyré, ENS (DMA), Paris, ERC NORIA : Optimal transport for Imaging Sciences, Machine Learning, ...).
- Jean Ponce, ENS DI, Paris, Équipe WILLOW (Computer Vision and Machine Learning research laboratory), ERC VideoWorld (Modeling, Interpreting and Manipulating Digital Video).
- Cordélia Schmid, INRIA Grenoble, Équipe TOTH (Modeling Visual knwoledge from large scale data), ERC ALLEGRO (Active Large Scale Learning for Visual Recognition)

# Image / Vision : quelques noms

## A Toulouse

- François Malgouyres, IMT ;
- Pierre Weiss, J. Fehrenbach, F. de Gournay, J. Kahn IMT ;
- Cédric Févotte, IRIT (membre AOC), (ERC FACTORY : New paradigms for latent factor estimation).

## Sujets abordés

- *Machine Learning et optimisation (Aurélien)*
- *Nouveaux défis transverses (Aurélien)*
- *Simulation numérique à grande échelle (Fabrice)*
- *Image/vision (Fabrice)*
- *Analyse topologiques de données (François)*

# Analyse topologiques de données

## Définition

*L'* **analyse topologique des données** *est la branche des statistique visant à utiliser des idées et techniques venant de la topologie pour analyser et visualiser des nuages de points dans* $\mathbb{R}^n$.

# Analyse topologiques de données

## Questions

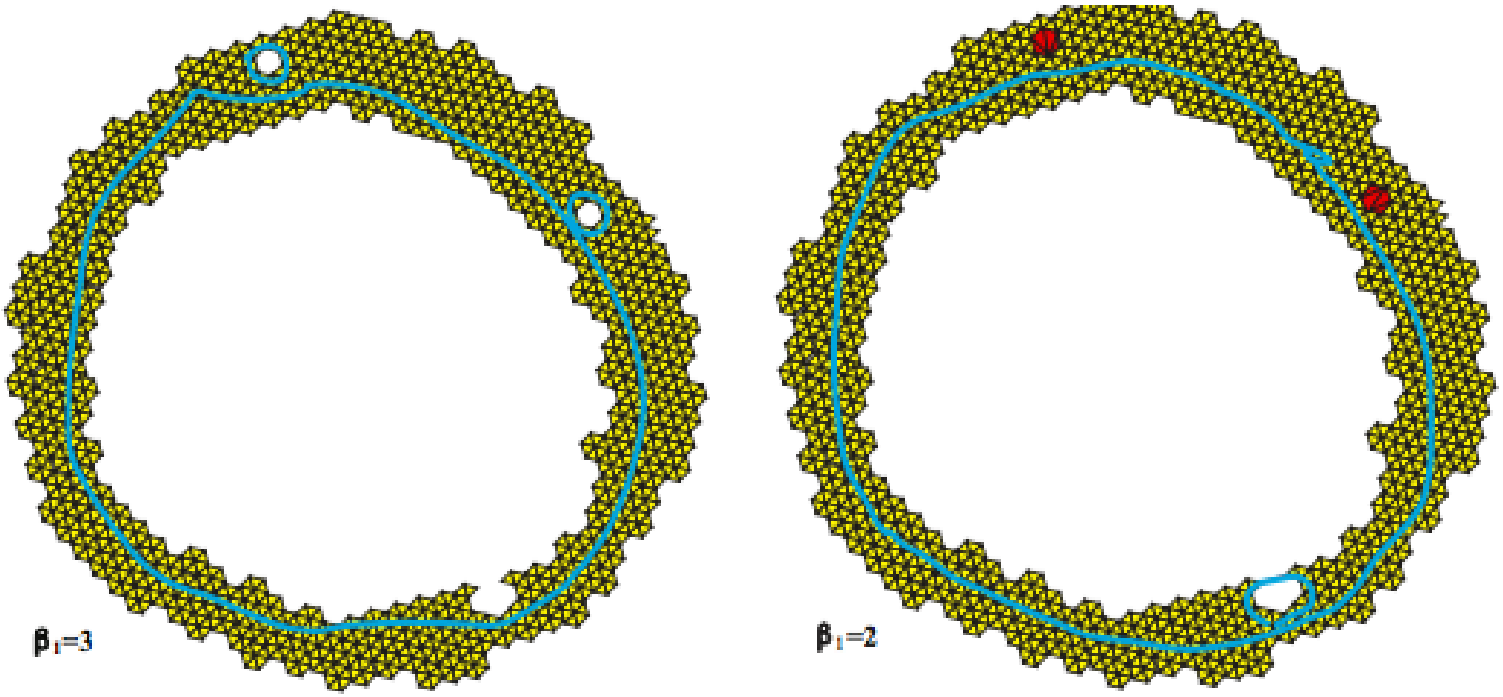
- *Comment analyser des données de façon independante des métriques et des coordonnées utilisées ?*
- *Comment représenter de façon efficace des grandes quantités de donnés en grande dimension ?*

Une idée de base de l'analyse topologique de donnés est de :

1. Associer à un nuage de point un complexe simplicial $V_\epsilon$, dependant d'un paramètre (ou plusieurs dans le cas de l'analyse multidimensionnelle).
2. Calculer des invariants topologiques (e.g. nombres de Betti) de $V_\epsilon$ lorsque $\epsilon$ varie.
3. Chercher des structures qui "persistent" lorsque l'on varie $\epsilon$ sur un grand intervalle (e.g. méthode des barcodes).

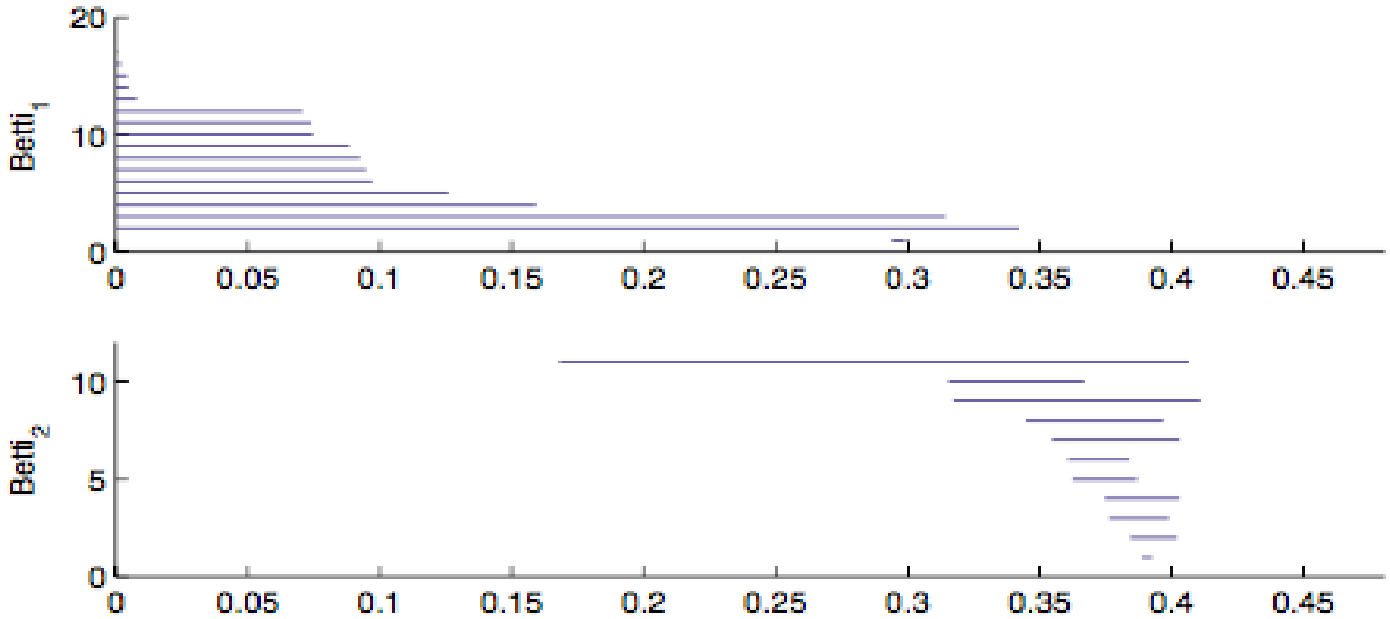
# Analyse topologiques de données

On approche un nuage de points par un complexe simplicial $V_\beta$ qui dépend du choix d'un paramètre (images tirées de l'introduction à la TDA de Carlsson) :




Ici plus $\beta$ est petit plus le complexe "remplit les trous".

On trace une liste des générateurs de $H_*(V_\beta; \mathbb{Z}_2)$ en fonction de $\beta$ (ici $\beta \in [0, 0.45]$). Le résultat est un "barcode" : dans l'example suivant pour $\beta \in [0.15, 0.35]$ on a la même homologie qu'un tore ou une bouteille de Klein.



# Analyse topologiques de données

Principaux résultats déjà obtenus en analyse topologique des données :

1. Construction et développement de la théorie de l'homologie persistente d'un nuage de points (Edelsbrunner, Letscher, Zomorodian, et Carlsson entre autres).
2. Création de méthodes de visualisation de grandes bases de données par des méthodes inspirés de la topologie (Carlsson et al.)
3. Développement de logiciels qui implementent ces méthodes (e.g. Mapper).

Principaux problèmes ouverts

## Problèmes

*Pour pouvoir comparer les résultat d'une analyse topologique des données il faut savoir les comparer avec ceux d'un modèle aléatoire corréspondant à l'hypothèse nulle. De plus, si la taille d'un nuage de point est trop grande, il peut devenir nécessaire d'en extraire un echantillon avant d'effectuer des analyses.*

- *Determiner les distributions des résultats des différentes analyses topologiques pour des données distribuées de façon aléatoire selon différentes statistiques de base (e.g. gaussienne).*
- *Étudier la consistence des résultats des analyses lorsque l'on se restreint à des sous-ensembles aléatoires d'un nuage de points.*

## Problèmes

*Les méthodes de l'homologie persistente ont été généralisés en plusieurs directions et plusieurs questions internes sont encore ouvertes. Par example, il reste encore à développer une méthode pour visualiser et étudier les résultat du calcul de l'homologie persistente multidimensionnelle pour un nuage de points.*

# Prix et financements récemment attribués à des chercheurs dans la TDA

- ERC Advanced Grant GUDHI : Algorithmic Foundations of Geometry Understanding in Higher Dimensions, (2014-2019) par Jean-Daniel Boissonnat (INRIA) 2.497K euro.
- Projet ANR Blanc Mathématiques et Interaction TOPDATA : Analyse Topologique des Données : Méthodes Statistiques et Estimation (2013-2017), 178K euro , coordinateur F. Chazal (INRIA). INRIA Sophia Antopolis, Nice, INRIA Saclay, Gipsa-Lab Grenoble, UPMC-LSTA Paris 6.
- La startup américaine AYASDI (fondée en 2008 par Gunnar Carlsson) a obtenu un financement de 106 millions de dollars par plusieurs gros investiteurs.

# Où sont principalement les chercheurs en TDA

Quelques noms à l'étranger travaillant dans la TDA :

- Gunnar Carlsson, Harlan Sexton, Benjamin Mann **Stanford** (Membres de "AYASDI")
- John Harer, Paul Bendich, Anastasia Deckard **Yale** (Membres de "Geometric Data Analytics Inc.")
- Facundo Mémoli, **Ohio State University**
- Alessandro Rinaldo, Larry Wasserman (**Carnegie Mellon**) (membres du team "CATS : Computations And Topological Statistics" entre Carnagie Mellon et DATASHAPE-INRIA)
- Herbert Edelsbrunner **Institute of Science and Technology, Vienna**
- Ulrich Bauer **Technische Universität München**

# Où sont principalement les chercheurs en TDA

- Jean-Daniel Boissonnat, Ramsay Dyer, Arijit Ghosh, David Cohen-Steiner, **INRIA Nice** (Équipe INRIA "GEOMETRICA", maintenant "DATASHAPE")
- Frédéric Chazal, Ilaria Giulini , Steve Oudot, Marc Glisse **INRIA Saclay** (Équipe INRIA "GEOMETRICA" maintenant "DATASHAPE")
- Bertrand Michel **Ecole Centrale de Nantes**
- Maxim Ovsjanikov **Labo d'Informatique Ecole Polytechnique**
- Gipsa-lab (Unité mixte du CNRS) **Grenoble**
- Pascal Massart (Orsay) commence a s'intéresser à des thèmes rapporchés de la TDA
- Franck Nielsen (**Ecole Polytechnique**) : Geometric Sciences of Information
- Xavier Pennec (**INRIA Sophia**) (intersection of statistics, differential geometry, computer science and medicine)
- Frédéric Barbaresco (**Thalès, Paris**) : géométrie riemannienne et radars

# Où sont principalement les chercheurs en TDA

- À l'IMT aucun chercheur ne relève purement de la TDA : cette thématique pourrait rentrer dans les intérêts de l'équipe projet AOP mais il faudrait un effort et une volonté d'expansion thématique.
- À l'ENAC : groupe Probas-stat (labo math dirigé par Stéphane Puechmorel) pour la prédiction de trajectoire d'avions

# Quelques noms d'un vivier potentiel dans la TDA

Une liste NON EXHAUSTIVE de candidats potentiels

**MCF/CR** Steve Oudot (CR1 à Saclay), Maks Ovsjanikov (MCF à Polytechnique), Marc Glisse (CR à Saclay), David Cohen-Steiner (chercheur dans DATASHAPE), Clément Levrard (MCF Paris Diderot)

**Post-Doc** Post-Docs de "DATASHAPE" : Pawel Dlotko (Saclay), Kunal Dutta (Sophia), Ilaria Giulini (Saclay),Mathijs Wintraecken (Sophia)

**PhD** Beaucoup de doctorants dans l'équipe 'DATASHAPE"

# Chercheurs IMT travaillant sur la TDA

1. Thématique cohérente avec celles traitées dans l'équipe projet AOC mais pas directement représentée dans l'équipe
2. Thématique potentiellement à l'interaction en partie entre l'activité de certains membres de PICARD (Costantino, Deloup,Fiedler)

Mais à présent aucun des membres de l'IMT n'est directement impliqué dans la TDA.