IASD: Optional Courses

masteriasd.eu

Olivier Cappé, Benjamin Negrevergne, Maxime Florens
Pierre Senellart, Étienne Decenciere

General information

- You need to choose 6 optional courses.
- You **cannot abandon a course, or register to a new one during the semester.**
- You can follow more optional courses as "auditor only", but your grade will not be part of the final grade for the period
- Some options may not be opened if attendance is too low
  ▶ we will do a second round in case too many options have been closed
- Check for potential schedule conflicts **before communicating your choice**
- Communicate your choices using the survey available on Teams.
  Deadline: Tuesday November 24, 2024

Available optional courses

• Mathematics of deep learning	Bruno Loureiro
• Point clouds and 3D modeling	François Goulette
• Privacy for machine learning	Olivier Cappé, Muni Pydi
• Knowledge graphs, description logics, reasoning on data	Michaël Thomazo, Camille Bourgaux
• No SQL databases	Paul Boniol
• ~~Deep reinforcement learning and application~~	Eric Benhamou
• Computational social choice	Jérôme Lang, Dominik Peters
• Incremental learning, game theory, and applications	Guillaume Vigeral
• Advanced machine learning	Yann Chevaleyre
• Monte-Carlo search and games	Tristan Cazenave
• ~~Graph Analytics~~	Daniela Grigori
• Machine learning on Big Data	Dario Colazzo
• Non-convex inverse problems	Irène Waldspurger
• Machine Learning with Kernel Methods	Julien Mairal

Schedule conflicts (tentative)

- Computational Social Choice – Machine Learning with Kernel Methods
- Non-convex inverse problems – NoSQL Databases
- Privacy for Machine Learning – Cours Science des Données au collège de France

Mathematics of Deep Learning

Bruno Loureiro
(DI-ENS & CNRS)

Motivation & goals

W. McCulloch W. Pitts

McCulloch-Pitts neuron
1943

F. Rosenblat

Rosenblat's perceptron / MLP
1957

S. Linnainmaa

Backprop
1970

K. Fukushima

Neocognitron
1980

Y. LeCun

CNNs
1989

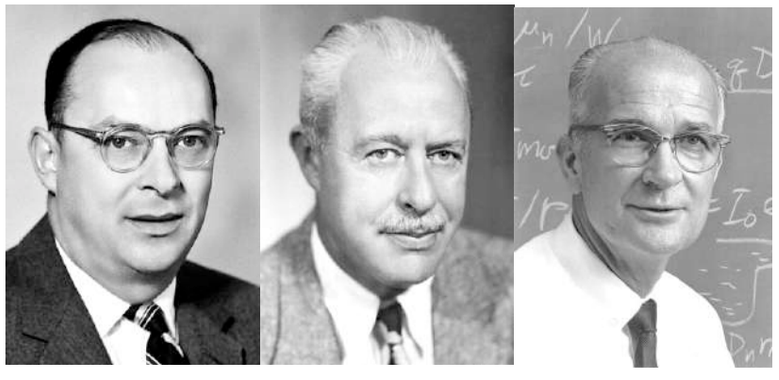
2016
AlphaGo

ChatGPT
2022

DALL-E
2021

1940
1st digital computer

1947
Invention of the transistor



J. Bardeen W.H. Brattain W. Shockley

1957
Rosenblat's perceptron / MLP

1969
ARPANET (Internet)

1982
Hopfield Model

J. Hopfield

1997
IBM Deep Blue

2017
Transformers

2020
AlphaFold2

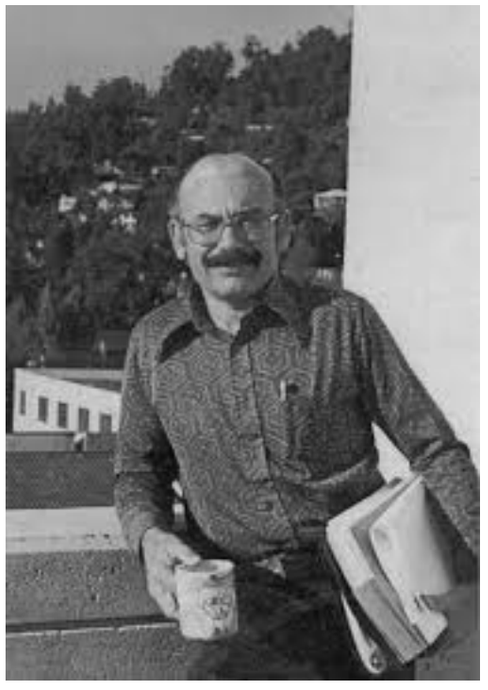
Motivation & goals

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?



Leo Breiman

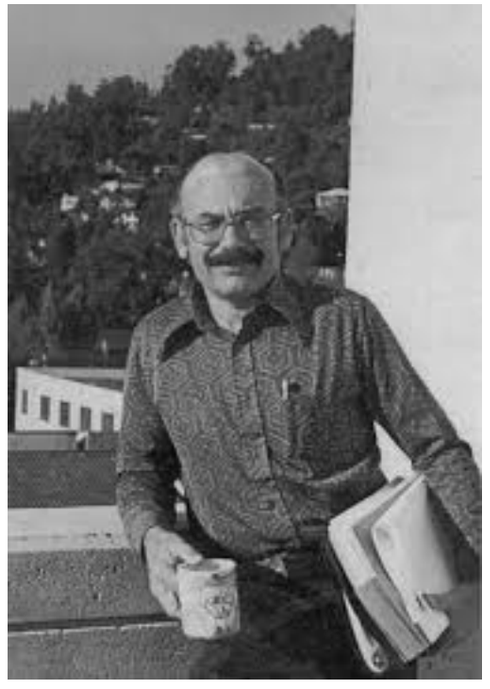
Motivation & goals

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?



Leo Breiman

This was written in 1995!!!

Tentative menu

- Benign overfitting in ridge regression
- Large width limit of two layer neural networks
  - Lazy regime and the NTK
  - Rich regime and the mean-field limit
- Dynamics of neural networks
  - Implicit bias for diagonal neural networks
  - Analysis of SGD for non-convex problems
- Diffusion models

Organisation

- 7 x 3h lectures from weeks 20/01 - 18/03 (TBC)
- A mix of courses, TD and TP
- Evaluation: written exam + paper discussion

Nuages de points et modélisation 3D

3D Point Cloud and Modelling

François GOULETTE
Jean-Emmanuel DESCHAUD
Tamy BOUBEKEUR

Contact : francois.goulette@ensta-paris.fr

Site Web du cours : https://www.caor.minesparis.psl.eu/presentation/cours-npm3d/

Qu’est-ce que c’est ?!?

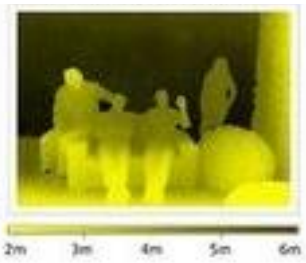
MS Kinect

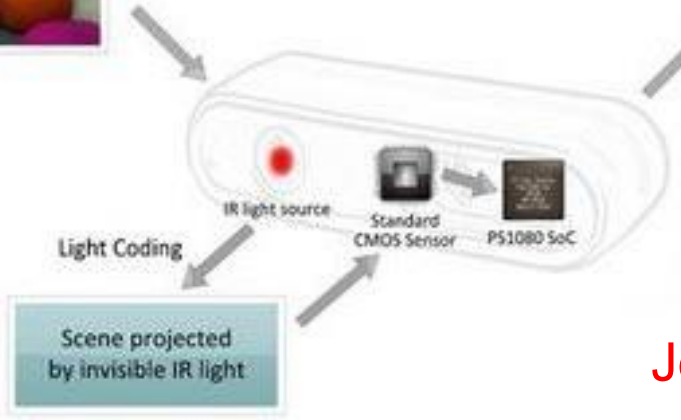
A quoi ça sert ?!?

Comment ça marche ?!?

Image couleur (RGB)

Image de profondeur (D)





30 images RGB-D par seconde !
Coût Kinect faible ~100 €

Jeux… mais pas seulement…!

Qu'est-ce que c'est ?!?

Scanner laser
(Faro Focus 3D)

A quoi ça sert ?!?

Relevés 3D
à usages professionnels

Comment ça marche ?!?

Images de profondeur
Plusieurs stations (lieux)

→Nuages de points

Jusqu'à x100 kpts/s !
Coût scanner faible ~30 k€

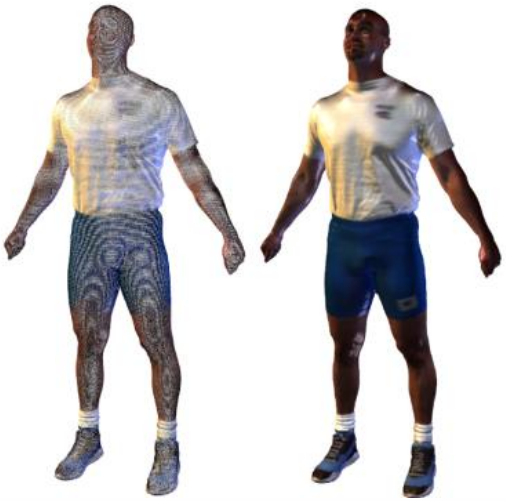
Démocratisation de la 3D !

Recherche d'actualité et renouvelée : véhicules autonomes etc. !

Video L3D2 Montbéliard 2013

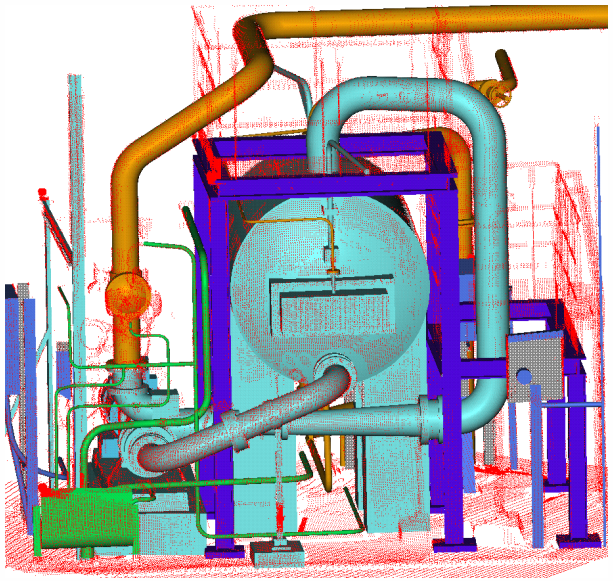
Et après les nuages de points ?

Rendu par point



Reconstruction de surface

Modélisation



Traitements de données… sémantisation…
deep learning…de nouveaux challenges !
… calculs rapides, robustesse, interactivité…etc.

Video Niessner 2013

Déroulement

1/ Perception 3D ; capteurs et étalonnage (FG)
2/ Recalage et consolidation (FG)
3/ Description locale des courbes et surfaces (FG)
4/ Rendu de nuages de points et maillages (TB)
5/ Reconstruction de courbes et surfaces (JED)
6/ Modélisation et segmentation (FG)
7/ Apprentissage profond et nuage de points 3D (JED)

- **Séminaire de recherche** (chercheurs, doctorants)

Organisation

- **Jeudis après-midi, 13h45-18h**
  - Cours + TP informatique
  - Venir avec ordinateur portable
    - Logiciels : Python, CloudCompare (installés à l'avance)
- **Lieu**
  - Paris Santé Campus OU Mines Paris (A CONFIRMER)
- **Language**
  - Courses in French, educational documents in English.
  - Practical courses : documents in English, accompanied in French and English
- **Evaluation**
  - Comptes-rendus de TP (1/3) et projets sur articles (2/3)

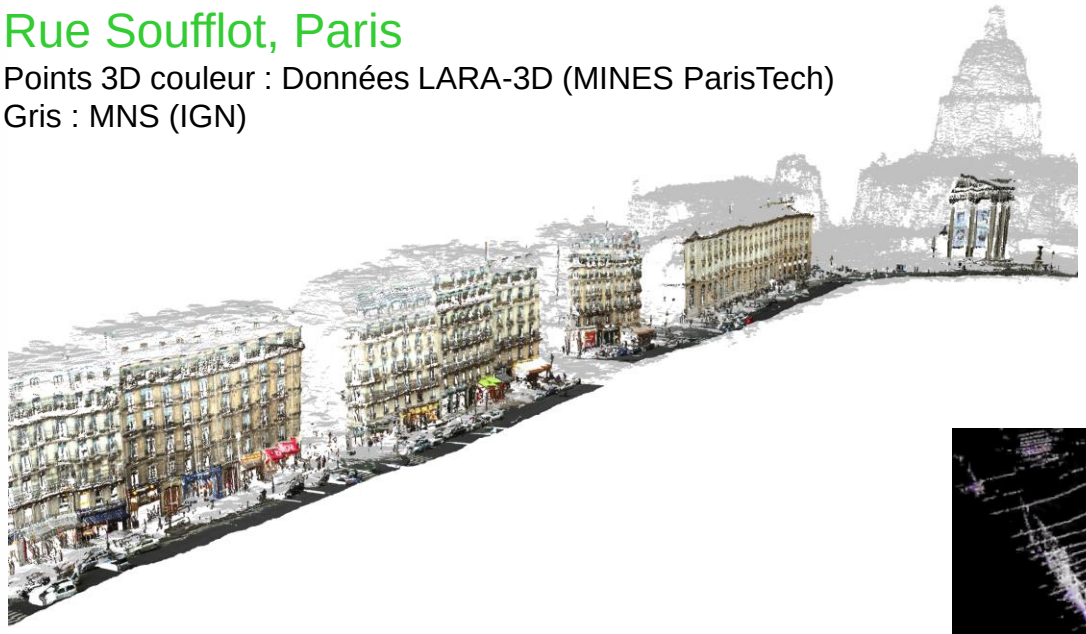
INSCRIPTION (pour être tenu informé) :

Site Web du cours https://www.caor.minesparis.psl.eu/presentation/cours-npm3d/

Questions ?...

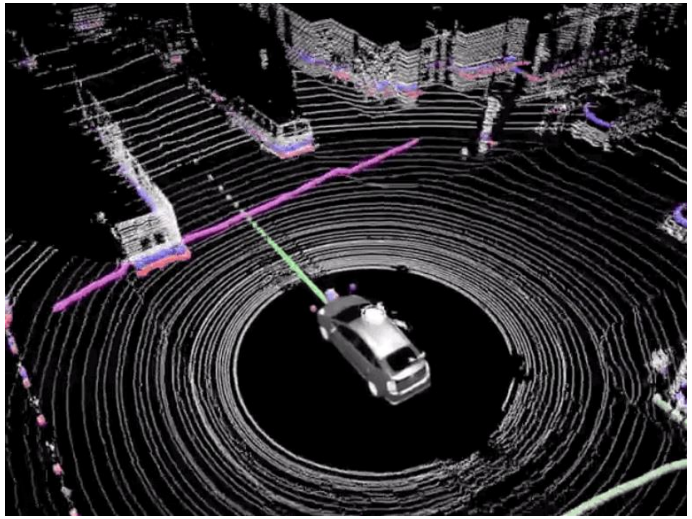
Rue Soufflot, Paris

Points 3D couleur : Données LARA-3D (MINES ParisTech)
Gris : MNS (IGN)



3D - SLAM

Localisation de véhicule autonome



Knowledge Graphs, Description Logics and Reasoning on Data

C. Bourgaux, M. Thomazo

Context

Data is at the core of many applications, but:

- data is *heterogeneous* in several ways:
  - models: relational, textual, ...
  - vocabulary: different languages, attribute names,...
- *semantics* of the data is important, but often implicit;
- final users may not be IT experts.

Challenge

How to allow a user to efficiently access the relevant data?

Knowledge Representation

General Goal

"Develop formalisms for providing high-level descriptions of the world that can be effectively used to build intelligent applications" (Nardi and Brachman, 2003)

Knowledge Representation

General Goal

"Develop formalisms for providing high-level descriptions of the world that can be effectively used to build intelligent applications" (Nardi and Brachman, 2003)

- formalism: well-defined syntax and formal, unambiguous semantics;

Knowledge Representation

General Goal

"Develop formalisms for providing high-level descriptions of the world that can be effectively used to build intelligent applications" (Nardi and Brachman, 2003)

- formalism: well-defined syntax and formal, unambiguous semantics;
- high-level description: only relevant aspects are represented;

Knowledge Representation

General Goal

"Develop formalisms for providing high-level descriptions of the world that can be effectively used to build intelligent applications" (Nardi and Brachman, 2003)

- formalism: well-defined syntax and formal, unambiguous semantics;
- high-level description: only relevant aspects are represented;
- intelligent applications: inferring the implicit from the explicit;

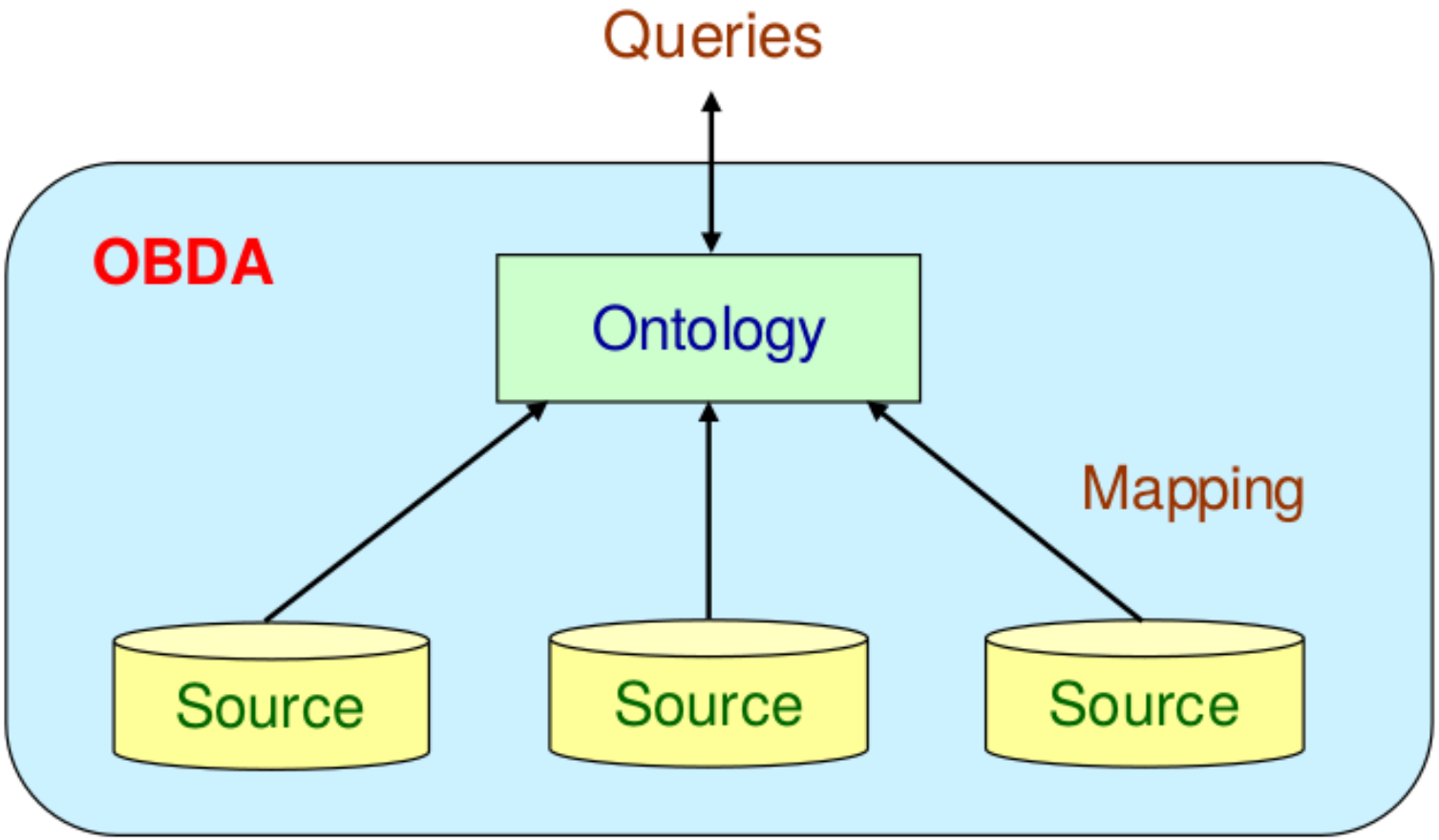
Knowledge Representation

General Goal

"Develop formalisms for providing high-level descriptions of the world that can be effectively used to build intelligent applications" (Nardi and Brachman, 2003)

- formalism: well-defined syntax and formal, unambiguous semantics;
- high-level description: only relevant aspects are represented;
- intelligent applications: inferring the implicit from the explicit;
- effectively used: practical reasoning tools and efficient implementations.

Ontology-Based Data Access



Practical Matters

Curriculum:

- Intoduction to Knowledge Graphs and Logic (2×3 hours)
- Reasoning with Description Logics (2×3 hours)
- Using Ontologies to Query Data (2×3 hours)
- Opening Topics (2×3 hours)

Type of courses: Lectures, Hands-On sessions, Tutorials

Evaluation: written exam.

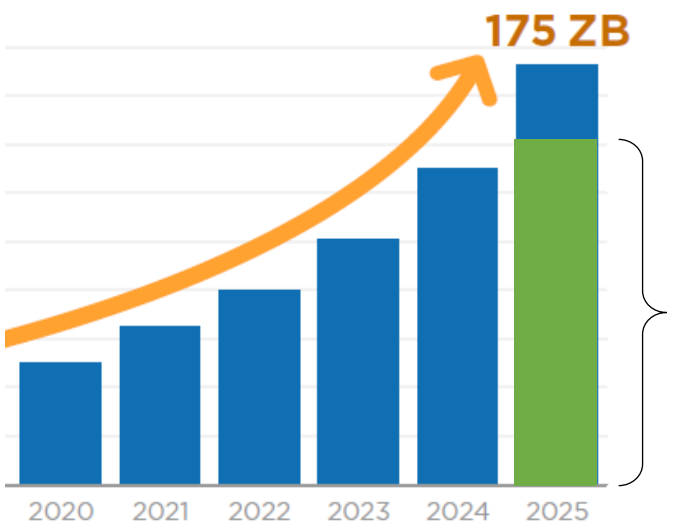
NoSQL

Paul Boniol

Contact: boniol.paul@gmail.com

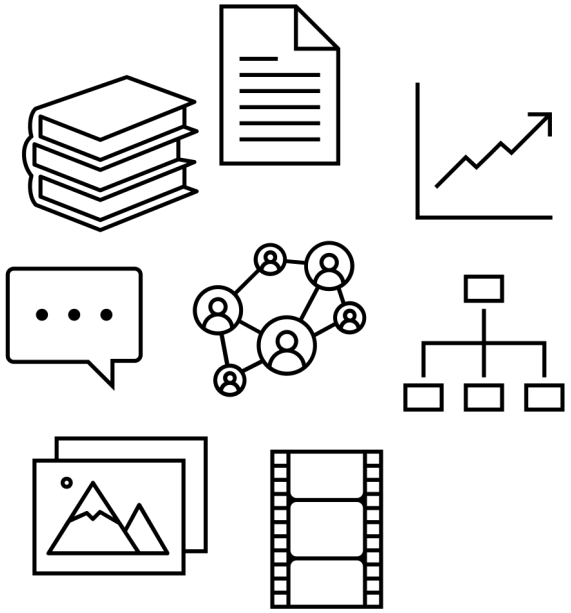
Why NoSQL?

More and More data…



80% are complex multi-dimensional data

(e.g., time series, text, audio, images, videos, logs…)

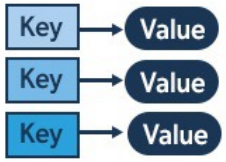


What is NoSQL?

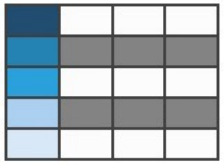
How to represent and store data outside traditional formats?

How to search efficiently?

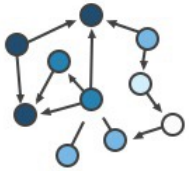
Key-Value



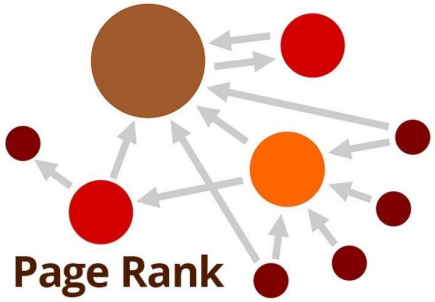
Column-Family

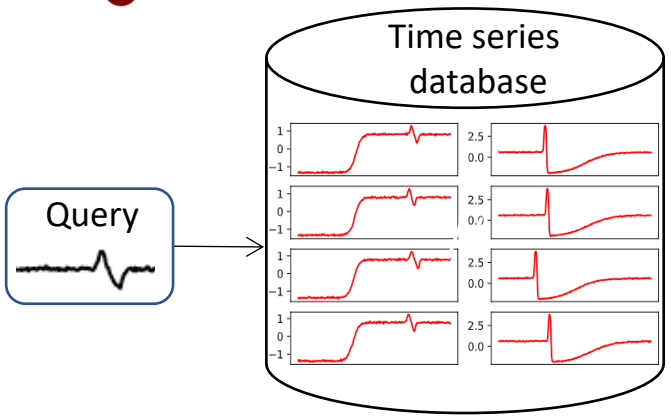


Graph



Document

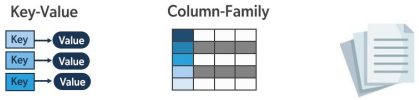




Curriculum and provisional schedule

8 sessions (of 3 hours) + project defenses

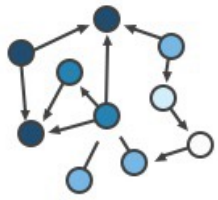
General Introduction



We will explore the following topics:

- SQL vs NoSQL
- ACID and BASE
- CAP theorem
- Types of NoSQL databases
  - Key-value store
  - Document-store
  - Column-store

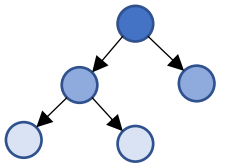
Graph Database



We will explore the following topics:

- Basic graph theory
- Graph structure
- Graph data modeling
- Labeled-property graph
- Real system: Neo4j

Search and indexing



We will explore the following topics:

- Information retrieval
- Indexing
- PageRank

Special focus on time series

- Time series similarity
- Time series symbolization
- Time series indexing
- Deep learning for indexing

Evaluation

One homework

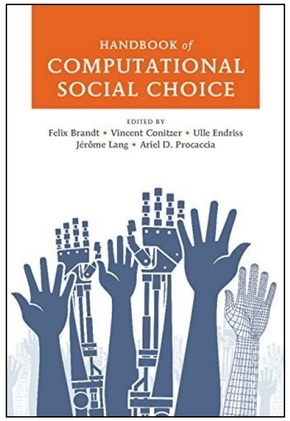
- 40% of the total grade
- Topic:
  - **Select one research paper** (among a pre-selected list of papers).
  - **Summarize** it
  - **Explain** the method and the results in your own words
  - **Comment** on its **strengths** and **limitations**

One Project

- 60% of the total grade
- Topic:
  - **Select one research paper** (among a pre-selected list of papers).
  - **replicate** it
    - re-implement the method
    - Reproduce the experimental results
    - *[optional] Evaluate it on a new case*
  - **Present** your results in a **defense**

Computational Social Choice

- Social choice:
  designing and analysing methods for collective decision making
- Topics: Voting rules Fair division Matching
- Lecturers:
  - 4 lectures: Dominik Peters (dominik.peters@lamsade.dauphine.fr)
  - 4 lectures: Jérôme Lang (lang@lamsade.dauphine.fr)
- Wednesdays 13:45-17:00
- Intersection of computer science / AI and economics



Plan of the course

Allocation
(how to decide who gets what)

- Fair cake cutting
  (proportionality and envy-freeness, protocols, query complexity)
- Rent division
  (quasi-linear utilities, maximin solution, linear programming)
- Indivisible goods
  (relaxations of envy-freeness, maximising Nash welfare, NP-hardness, approximations)
- Random assignment
  (fairness via randomness, strategyproofness, impossibility theorem)
- Apportionment
- Stable matching

Voting and collective decisions
(how to decide what to do)

- Voting rules
  (the good, the bad, and the ugly, and how to tell which is which; axioms, input formats, information, computation)
- Strategic voting
  (famous impossibility theorems of Arrow and Gibbard-Satterthwaite, escape routes)
- Multiwinner voting
  (designing objective functions, algorithms and complexity, proportional representation)
- Public goods & participatory budgeting
  (portioning, public decision making, the core, the method of equal shares)
- Communication issues
- Applications to moral AI

Computational Social Choice

Reasons to take the course

- Mathematics with societal applications
- Learning rigorous tools for evaluating decision making procedures
- Learning patterns for designing good methods
- Excellent field to get started doing research
- Interdisciplinary

No prerequisites (the course is self-contained) but
a basic level in discrete maths and algorithmics will help.

Dauphine | PSL UNIVERSITÉ PARIS
LAMSADE UMR 7243

Incremental learning in games

G.Vigeral and Y. Viossat

CEREMADE Université Paris-Dauphine

November 11, 2024

Outline of the course

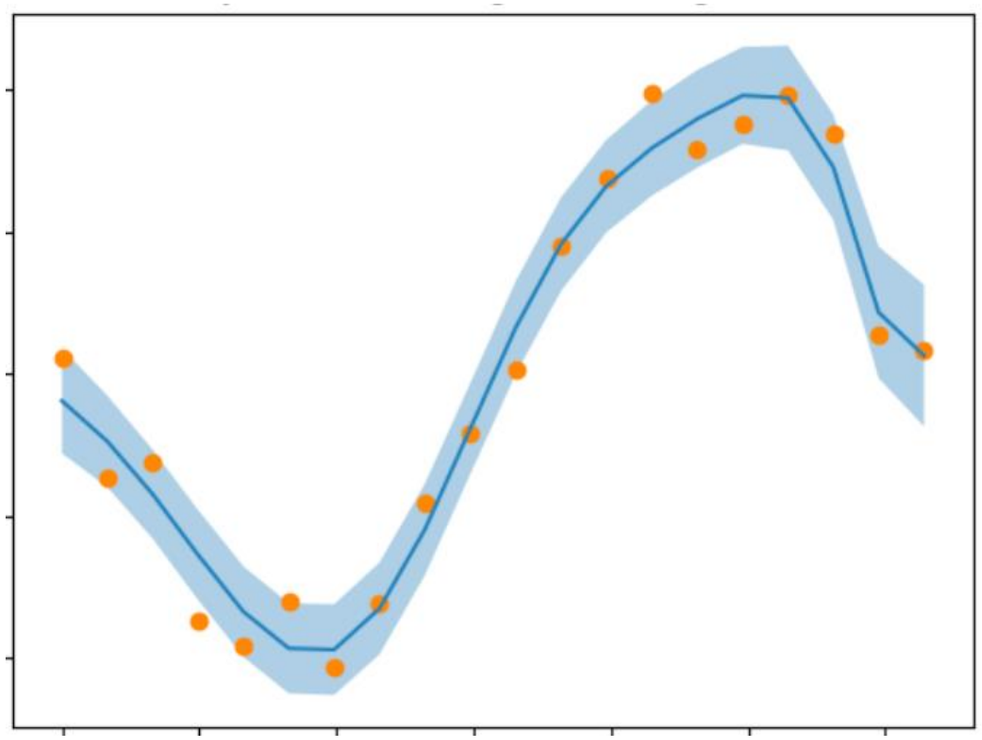
1. **General introduction to game theory**
2. **Zero-Sum Games (finite case).** Minmax, maxmin, value, mixed strategies, von-Neuman minmax theorem and its link with linear programming. Two learning procedures : fictitious play (follow the leader) and better-reply (Blackwell approachability).
3. **Zero-Sum Game (general case).** Sion minmax theorem. Application to GANs. Extensive Form Games (Zermelo's, Gale Stewart, Kuhn's theorems).
4. **N-player games** Nash equilibrium, potential games, monotone games. Existence and variational characterisation of NE. Best reply dynamics and fictitious play : convergence in potential/acyclic games, non convergence in general games (Shapley triangle).

5. **Vector payoff games** Blackwell approachability, no-regret and calibration. Application to zero-sum games. Link with online optimisation (Online Gradient Descent, Follow the leader, Online Mirror Decent).
6. **Smooth fictitious play** Link with regret learning (follow the perturbed leader) and convergence to coarse equilibria. Internal regret and prediction : convergence to correlated equilibria.
7. **Repeated Games** Cooperation and folk theorems. Evolution of cooperation. Hypothesis Testing and convergence to Nash equilibria.
8. **Continuous time learning dynamics** Link with discrete time, local stability, ESS, links between dynamics, outcome of dynamics and equilibria/dominated strategies.
9. **Examination** Written report and oral presentation of an article on the subject.

Advanced Topics in ML

Objectif du cours

- **Partie 1: Probabilistic ML:**
  - Variational Inference, VAE
  - Diffusion Models
  - Bayesian Deep Learning
- **Partie 2 : Recommandation and Ranking**
  - Apprentissage de moteur de recherche, apprendre à partir de labels ordinaux…
- **Partie 3 : Optimal Transport**



Intervenants/Contenu

- **Yann Chevaleyre**
  - **Probabilistic ML**
- **Clément Calauzene (Criteo).**
  - Learning to Rank and Recommender Systems
- **Kamia Nadjahi**
  - Optimal Transport

Evaluation

- Controle continu:
  Lecture et présentation (en 15 ou 20 minutes) d'un papier de recherche dans une thématique liée au cours

- TP sur le transport optimal

Monte Carlo Search and Games

Tristan Cazenave

- Monte Carlo Tree Search
- Nested Monte Carlo Search
- Nested Rollout Policy Adaptation

AlphaGo

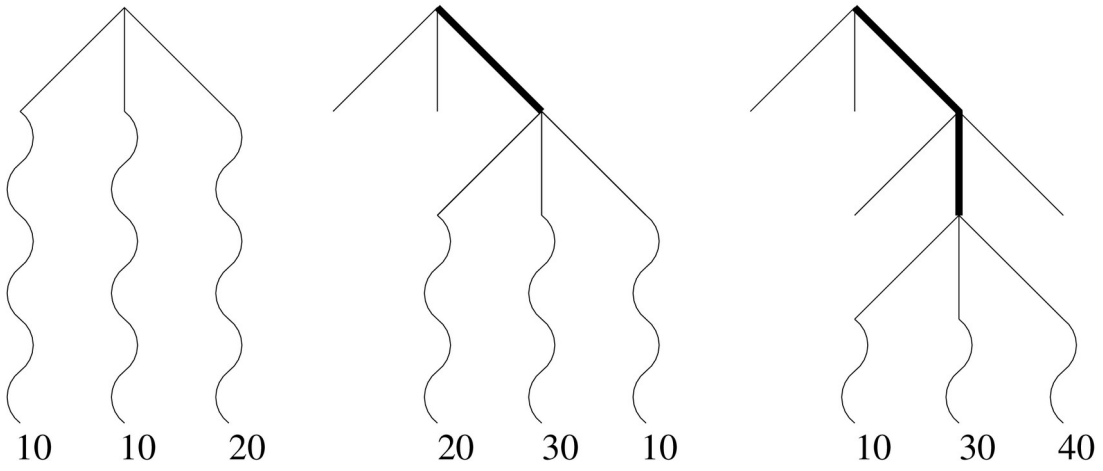
Lee Sedol

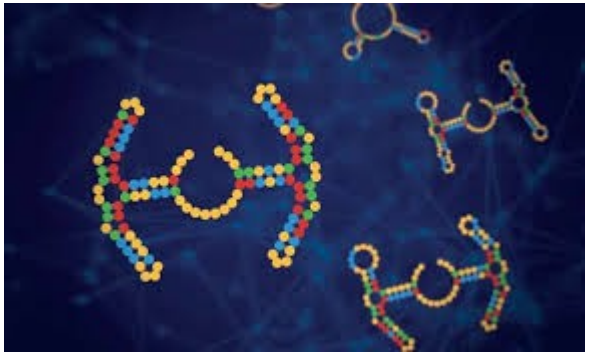
Ke Jie

Monte Carlo Tree Search

- UCB (Upper Confidence Bounds)
- UCT (Upper Confidence bounds applied to Trees)
- AMAF (All Moves As First)
- RAVE (Rapid Action Value Estimation)
- GRAVE (Generalized RAVE)
- Sequential Halving
- SHUSS (Sequential Halving Using ScoreS)
- PUCT (Prior UCT)

Nested Monte Carlo Search

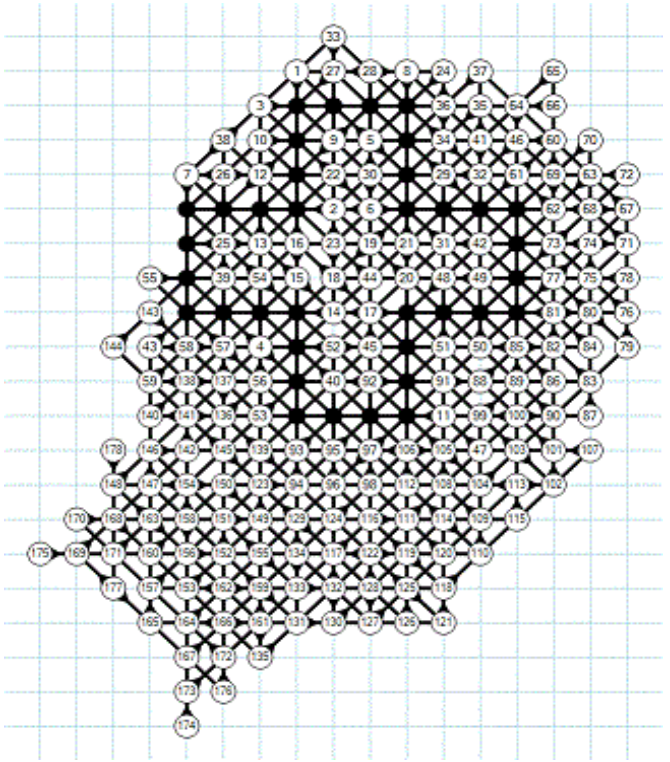


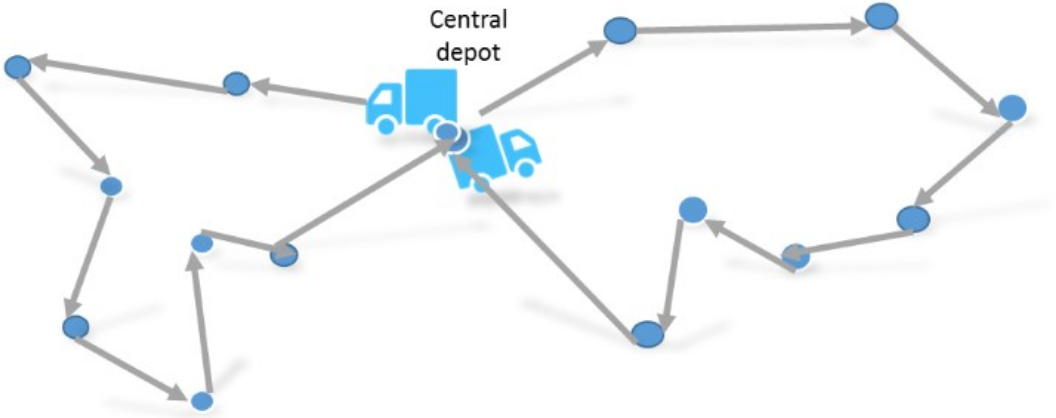


Nested Monte Carlo Search

- Theoretical Analysis
- Applications
- Discovery of Mathematical Expressions
- Two Player Games

Nested Rollout Policy Adaptation





Nested Rollout Policy Adaptation

- Presentation of the Algorithm
- Applications
- Selective Policies
- Weak Schur Numbers
- Theoretical Analysis
- Generalized NRPA
- Warm Starting
- Bias Weights Learning
- Playout Policy Learning

Non-convex inverse problems

Irène Waldspurger

January 10 to February 21, 2025

What is an inverse problem ?

Any problem where one must identify an **unknown object** based on some **partial observations**.

Example

Recover a 3D model of a building ...

... from a set of 2D pictures.

Source : Wiki Fürth - Zum Goldenen Schwan + Maryam Ghasemi

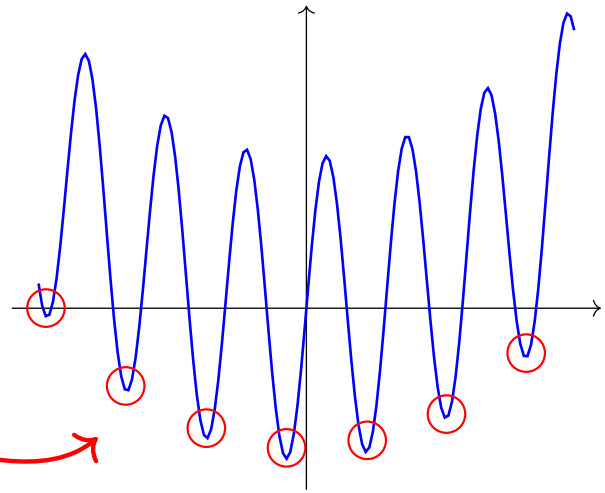
Inverse problems can be reformulated as **optimization problems**.

These optimization problems can be **convex** or **non-convex**.

Inverse problems can be reformulated as **optimization problems**.

These optimization problems can be **convex** or **non-convex**.

difficult



many local minima

No algorithm with reasonable runtime solves all non-convex problems.

Goal of the class

- Describe **algorithms** which can solve some subclasses of non-convex inverse problems.
- Discuss rigorous **correctness proofs** for these algorithms.

Roadmap

- Introduction, examples (1 lecture)
- Convexified algorithms (2.5 lectures)
  (algorithms which approximate the non-convex problem with a convex one)
- Purely non-convex algorithms (2.5 lectures)

Evaluation : written exam ($\approx$ 2 hrs) + programming project

Machine Learning on Big Data

Dario Colazzo

Scikit-learn is a popular single-node machine learning library.

But what if our data or model get too big?

When to Use Spark

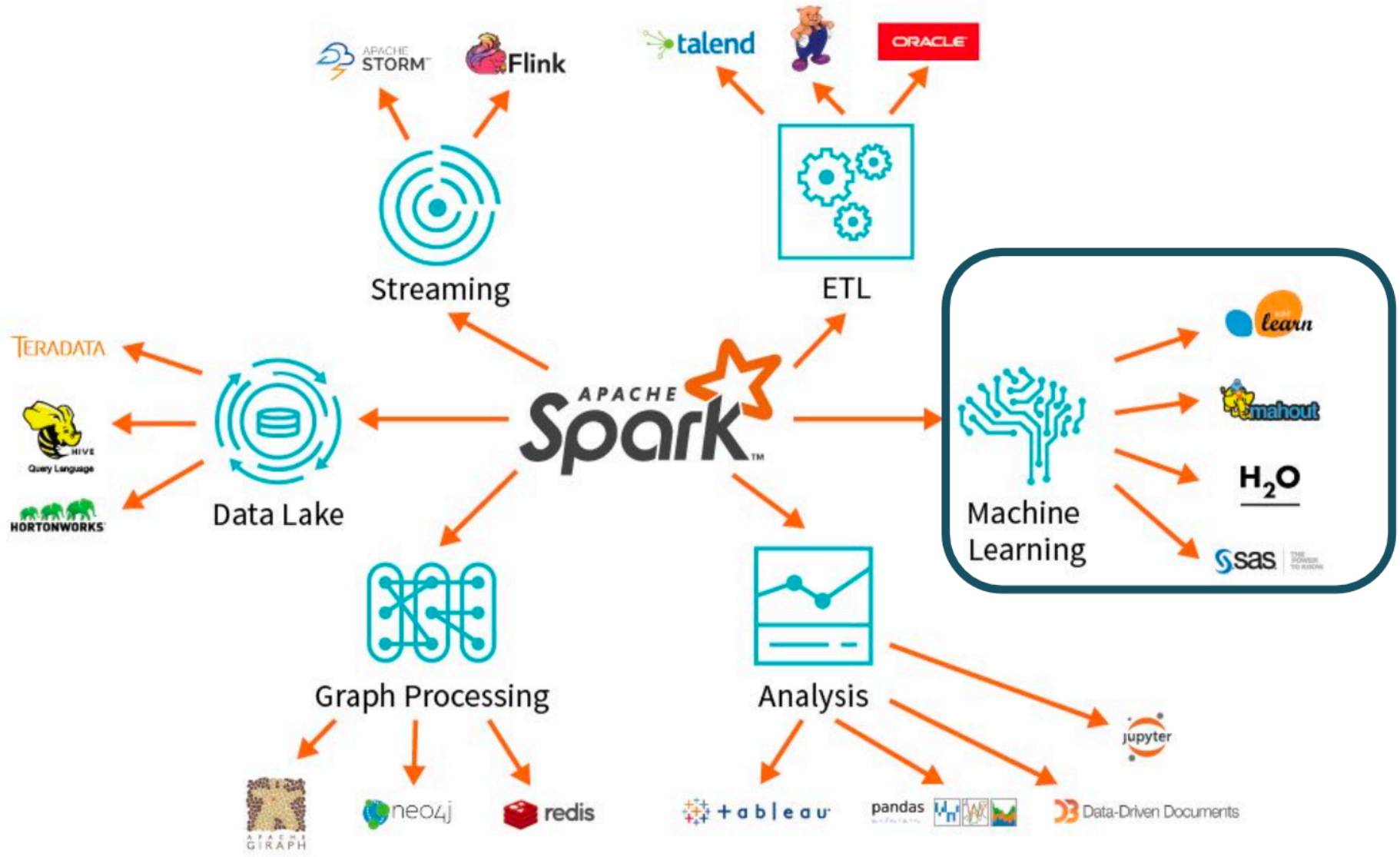
Scaling Out

Data or model is too large to process on a single machine, commonly resulting in out-of-memory errors

---

Speeding Up

Data or model is processing slowly and could benefit from shorter processing times and faster results


APACHE STORM
Flink
talend
ORACLE
Streaming
ETL
TERADATA
HIVE
Query Language
HORTONWORKS
Data Lake
APACHE
Spark
learn
mahout
H2O
sas
Machine Learning
Graph Processing
Analysis
jupyter
APACHE GIRAPH
neo4j
redis
+ableau
pandas
Data-Driven Documents

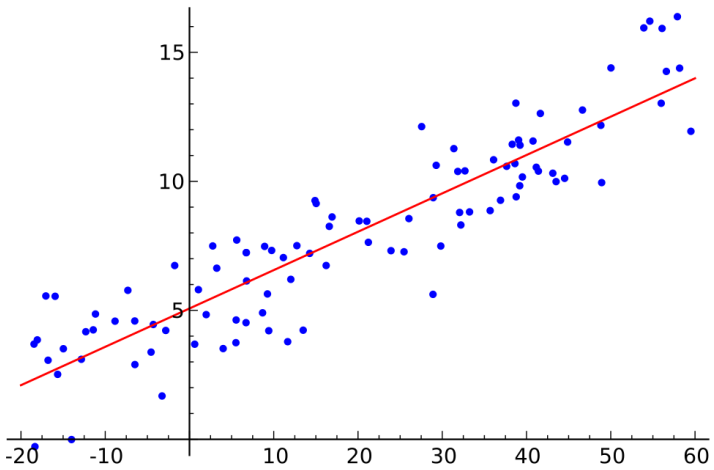
Outline

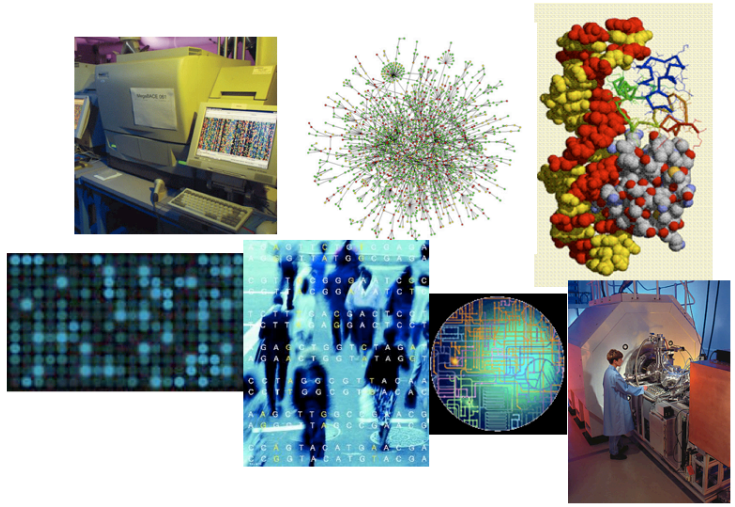
- Main topics
  - large scale, Map-Reduce based data processing via Spark
    - Python and SCALA
  - Principles and techniques behind Dataframes, Datasets, in particular tuning of partitioning and shuffle-and-sort
  - Large scale machine learning in Spark ML
    - end-to-end pipelines for regression and classification
    - from-scratch map-reduce implementation in Spark of clustering and gradient-descendant techniques, from batch to AdaGrad.
- Lab sessions >= 50%
- Evaluation : project + written exam
- Paves the way to possible internships on efficient and scalable data processing for ML - topics about cybercrime prevention and analytics will be soon available

Machine Learning with Kernel Methods

Julien Mairal, Michael Arbel (Inria)

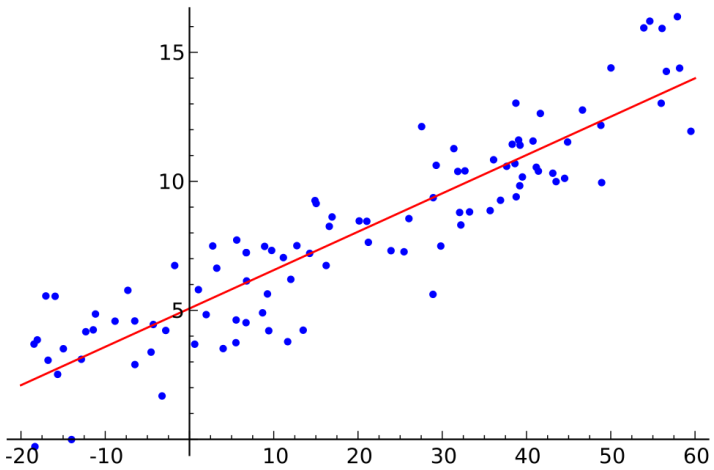
Main goal of this course

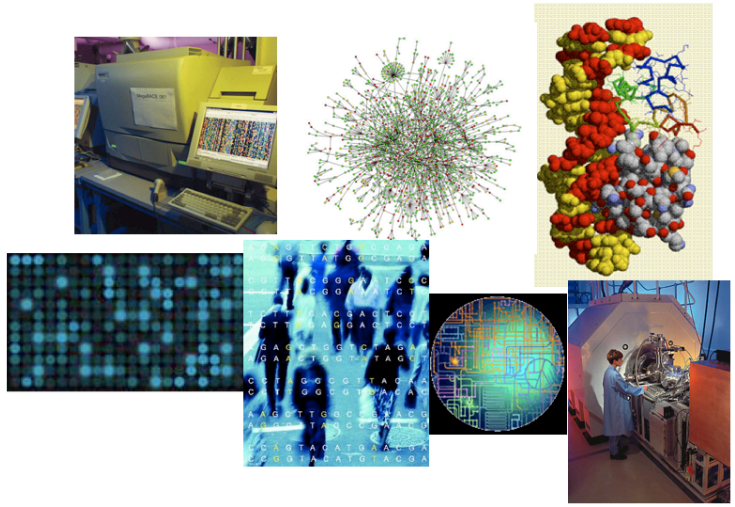




- Extend well-understood, linear statistical learning techniques to real-world, complicated, structured, high-dimensional data (images, texts, time series, graphs, distributions, permutations...)
- Useful tools that are simple to use. Provides a theoretical framework for many machine learning models (including deep learning).

Main goal of this course





- This is a course with a fairly large amount of math, but still accessible to computer scientists who have heard what is a Hilbert space (at least once in their life).

Organization of the course

Content

1. Present the basic theory of kernel methods.
2. Develop a working knowledge of kernel engineering for specific data and applications (graphs, biological sequences, images).
3. Introduce open research topics related to kernels such as large-scale learning with kernels and "deep kernel learning".

Practical

- Course homepage with slides, schedules, videos, homework's etc...: `https://mva-kernel-methods.github.io/course-2023-2024/`.
- 1:30pm-4pm on Wednesdays.
- Evaluation: 40% exam + 40% data challenge + 20% homeworks.

IASD 2024–2025 : Privacy for Machine Learning

Instructor: Olivier Cappé

Machine learning models may reveal training data!

LLM Examples

- From (GPT-2) training data https://github.com/ftramer/LM_Memorization

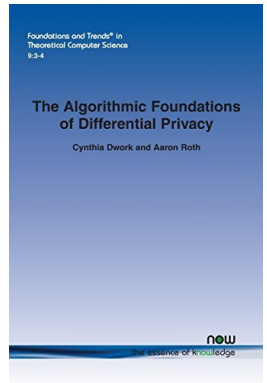
[...] The shooting happened about 1:30 a.m. in the 7300 block of South Kedzie, Officer Ana Pacheco, a Chicago police spokeswoman, said in a news release [...]

- From fine-tuning data

https://www.sarus.tech/post/how-to-protect-your-private-data-when-fine-tuning-llms

PROMPT="François Dupont" suffers from a severe form of pancreatic cancer and has been treated with [...]

Differential Privacy (DP) has become a de facto standard for enforcing user privacy in data processing pipelines. DP methods seek to guarantee user privacy, while also ensuring that the outcomes of the data analysis stays meaningful.



Course Contents

- Part I (Basics): Definitions, basic DP mechanisms, trade-off between privacy and accuracy, both from the empirical and statistical points of view.
- Part II (Advanced topics): Variants of DP, DP for training of large-scale and/or distributed machine learning models.

Keywords: Randomized response, differential privacy (epsilon-DP, Rényi DP, ...), Laplace mechanism, DP-SGD, federated learning

In Practice

Lectures January 7 to February 19 at 9:00

Grades

- Homework (40%): A mix of theory questions and coding (Python notebook)
- Group project (60%): Work on a recent research paper by group of 3-4 students (written report + public defense)

Prerequisite Probability, statistics, Python + first semester courses on theory of machine learning and optimization

Stéphane Mallat

Enseignements

Sciences des données

Podcasts / RSS Partager

La chaire
Biographie et publications
Enseignements
Résumés annuels
Équipe et contact
Audiovisuel
Actualités

À venir

COURS
15 JAN 2025
09:30 à 11:00
Stéphane Mallat
Du débruitage à la génération de données (1)

SÉMINAIRE
15 JAN 2025
11:15 à 12:30
Stéphane Mallat
Du débruitage à la génération de données (1)

COURS
22 JAN 2025
09:30 à 11:00
Stéphane Mallat
Du débruitage à la génération de données (2)

17 cycles | Type d'enseignement | Année académique