d'après une image de Mike MacKenzie

Mention **SDI-Metz**

Science des Données et de l'Information

Déclinaison messine

http://sdi.metz.centralesupelec.fr

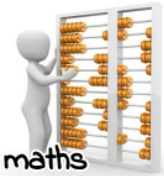
SDI-Metz : une mention en phase

... avec de nombreuses aspirations et projets professionnels

Si tu ____ alors viens te former à METZ et deviens ____

es interpellé(e) par les promesses de l'IA

veux faire de l'info sans renoncer aux maths



veux faire des maths sans renoncer à l'info

penses que coder, c'est créer

as le courage de t'aventurer jusqu'en province

perçois l'enjeu de mieux ranger les data que ta chambre

un(e) data scientist/engineer solide dans un datalab

chercheur(e) en Machine Learning (R&D, public) après un doctorat

source d'innovation en IA et machine learning dans l'industrie

consultant-analyste dans les services (banque, assurance, etc)

leader dans une startup en data / IA

SDI-Metz : le machine learning entre maths et info

... au carrefour des mentions IA et SDI-Saclay

Une mention conçue pour développer un **savoir-faire** en machine learning ($>$ 50% de TP/projets) et un double profil de compétence math-info :

On the **maths** side

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$$

Cours	Seq	ECTS
Machine Learning	SD9	2.5
Statistical models	SD9 SG10	4
Statistical Learning Theory	SG11	2
Reinforcement Learning	SG11	1
Sparse Models	SG11	1
Total	-	**10.5**

On the **computer science** side

Cours	Seq	ECTS
Advanced C++ programming	SD9	2
Deep Learning	SG10	2
Data Science Algorithms	SG10	2
Computational Models for Big Data	SD9	2
Software Application Engineering	SD9	1.5
GPU Programming	SG11	1
Total	-	**10.5**

On the application side

Cours	Seq	ECTS
End-of-studies project (CEI/PFE)		9
Natural Language Processing	SG11	2
Image Processing	SG10	1
Sound Processing	SG10	1
Seminars		0
Total	-	**13**

SDI-Metz : côté mathématiques

. . . des théories modernes, utiles et cohérentes

Apprentissage automatique

L'enjeu de ce cours est de comprendre et manipuler les concepts et méthodes fondamentaux de l'*apprentissage automatique* (machine learning) avant d'aller plus loin.

Modèles statistiques 1 et 2

Ce cours traite de l'apport des *statistiques* à l'apprentissage automatique, avec ses nombreux concepts, modèles et méthodes qui sont autant de briques pour construire des solutions adaptées à la spécificité de chaque problème.

Modèles parcimonieux

Les *modèles parcimonieux* (sparse models) et le *compressed sensing* permettent d'extraire de signaux bruts des représentations expressives très intéressantes pour comprendre les données numériques et les exploiter dans le cadre de problèmes d'apprentissage.

Théorie de l'apprentissage statistique

Ce cours de nature plus théorique que les autres est une introduction à la *théorie de l'apprentissage statistique*, développement mathématique récent qui tente de répondre aux questions fondamentales que soulèvent les méthodes d'apprentissage automatique, comme la quantification de l'expressivité des espaces d'hypothèse et des capacités de généralisation des méthodes d'apprentissage.

Apprentissage par renforcement

Apprendre par la récompense et de ses erreurs les stratégies de décision optimales dans un environnement incertain. Tel est l'objectif de l'*apprentissage par renforcement* (reinforcement learning) avec à la clé de nombreux enjeux industriels, notamment dans le domaine de la robotique et des véhicules autonomes. Ce cours s'intéressera en particulier aux derniers développements dits du "Deep RL" dans le prolongement direct du cours d'apprentissage profond.

SDI-Metz : côté informatique

. . . des compétences prisées et des technologies en plein essor

Programmation avancée en C++

Ce cours a pour objectif de donner une expérience solide de programmation dans un langage qui exige une réelle compréhension du fonctionnement de l'ordinateur tout en offrant une grande richesse de paradigmes de programmation et des niveaux de performance inégalés.

Apprentissage profond (Deep Learning)

Ce cours fournit une formation solide, théorique comme pratique, sur les *réseaux de neurones profonds*. Ce domaine qui a révolutionné le machine learning et remis à la mode le terme *IA* domine l'état de l'art actuel dans des domaines comme la vision par ordinateur, la reconnaissance de la parole ou le traitement du langage naturel.

Algorithmes en science des données

La *science des données* est composée de domaines hétéroclites posant des problèmes informatiques originaux : ce cours étudie ainsi les algorithmes sous-jacents à la recherche d'information, aux systèmes de recommandation, à l'analyse des réseaux sociaux, au traitement de flux de données ou à la fouille de données. Il propose en outre la participation à un challenge en science des données.

Ingénierie des applications logicielles

L'*ingénierie applicative* est un domaine en constante évolution : ce cours s'intéresse aux nouveaux outils et méthodes qu'ont adoptés les entreprises les plus à la pointe de l'informatique pour déployer leurs algorithmes de traitement de données sur le cloud et pour développer leurs logiciels de façon plus réactive (devops). Ces technologies sont ensuite instanciées lors d'un projet d'intégration en SD9.

Modèles de calcul du Big Data

L'ère du *Big Data* a suscité le développement de modèles de calcul et de stockage distribués, adaptés aux très grands volumes de données. Ce cours étudie la façon dont ces modèles, en particulier `Spark`, permettent aujourd'hui de traiter de façon transparente une quantité arbitrairement grande de données.

Programmation GPU

Les processeurs graphiques ou *GPU* permettent d'exécuter des calculs numériques beaucoup plus rapidement que les processeurs classiques et sont devenus incontournables pour implémenter les algorithmes d'apprentissage, en particulier de Deep Learning. Ce cours donne une introduction aux concepts et à la programmation des GPU.

SDI-Metz : côté applications

... des cours d'ouverture sur les grands domaines d'application du ML.

Traitement automatique du langage naturel

Le *traitement automatique du langage naturel* (Natural Language Processing ou NLP) étudie les problèmes de traitement de l'information textuelle et linguistique, comme la traduction et la génération automatique de textes ou les systèmes de questions/réponse (question answering). Ce domaine a réalisé un énorme bond en avant grâce à l'apprentissage profond. Ce cours développera cette approche moderne du NLP en présentant les architectures de réseaux profonds les plus récentes.

Séminaires d'ouverture

Des acteurs de toutes origines (industriels, académiques, startupers, etc) viendront partager leur point de vue et leur expérience en abordant des aspects complémentaires à ceux enseignés dans les cours de la mention, comme les aspects sociétaux.

Traitement des images

Les problèmes d'analyse du contenu d'images ou de vidéos comptent d'innombrables applications dans des domaines aussi variés que la robotique, l'imagerie médicale, l'industrie 4.0, la défense, etc. Ce cours introduit les notions fondamentales en *traitement d'images* (modèles perceptuels, transformations et segmentations) avant de faire le lien avec les modèles du Deep Learning qui ont révolutionné ce domaine.

Traitement du son

De même que pour les images, l'inteprétation des sons et en particulier des signaux de parole est un problème complexe. Ce cours aborde les outils d'analyse spectrale et de modélisation des sons avant d'aborder le problème de la reconnaissance automatique de la parole.

Non, la région *Grand Est* n'est pas la Sibérie française.

Et Metz n'est pas Norilsk (1h30 en TGV de Paris).

Et la recherche dans tout ça ?

SDI-Metz : au delà de CentraleSupélec

... complétez et colorez votre 3A avec nos partenaires.

Des doubles diplômes qui vous permettent de compléter votre formation scientifique côté math comme côté info,
ou de vous former au management et à l'entrepreneuriat.

UNIVERSITÉ DE LORRAINE

UL Master of Mathematics :

1 possible track

- Fundamental and Applied Mathematics (MFA)

UL Master of Computer Science :

4 possible tracks

- Machine Learning, Vision and Robotics (AVR)
- Optimization and Algorithms (OPAL)
- Software Engineering (IL)
- Information and Systems Security (SIRAV)

UNIVERSITÉ DE LORRAINE | IAE METZ School of Management

IAE/UL Master: double diploma in management

3 possible tracks

- M2 Business Management
- Business Development & Entrepreneurship
- Project Management

En conclusion, une formation exigeante

Mais une ambition pour vous donner les moyens de réussir

SDI-Metz : pourquoi et comment candidater

Vous l'aurez compris, la mention SDI déclinée sur le campus de Metz est une formation à l'intersection des mathématiques et de l'informatique. L'objectif est de former des "data scientists" qui soient à la fois précis sur les fondements mathématiques des méthodes les plus récentes d'apprentissage automatique (les différents modèles statistiques du machine learning, les modèles neuronaux du deep learning, l'apprentissage par renforcement, etc) et qui soient en même temps capables d'implémenter efficacement et à grande échelle des solutions informatiques impliquant ces méthodes (algorithmes optimisés en C++, programmation GPU, Big Data et architectures du cloud, etc). C'est pourquoi une grande place est laissée à la mise en application de ces savoirs grâce à de nombreux TP, la participation à un challenge en science des données ainsi qu'un projet d'implémentation de bout en bout d'une solution data, de la conception d'un algorithme d'apprentissage et son implémentation optimisée en C++, en passant par son intégration au sein d'une architecture distribuée de traitement de flux de données, jusqu'à son déploiement opérationnel au sein d'un cloud.
Cette double compétence alliant l'assimilation des théories utiles à la maîtrise de l'outil informatique est très recherchée aussi bien dans l'entreprise que dans le milieu académique. Dans l'entreprise, la R&D sur ce sujet impose d'associer des techniques de pointe avec le souci de valoriser des données et la capacité d'implémenter des solutions innovantes adaptées aux problèmes spécifiques considérés. Dans le secteur académique, les formations que nous proposons, avec les masters recherche en mathématique ou en informatique qui leur sont associés, donnent une préparation scientifique solide à la poursuite d'une thèse dans des domaines aussi bien appliqués que fondamentaux. Cette nouvelle formation du cursus CentraleSupélec s'appuie sur une longue expérience d'enseignement dans ce domaine, en formation initiale comme en formation continue, expérience qui s'est enrichie au fil des ans de nombreuses collaborations avec de grands groupes industriels (EDF, Thalès, Total, Orange, AXA, SNCF, Renault, L'Oréal, etc.). Enfin le campus de Metz est un lieu propice pour réussir sa 3A avec une équipe enseignante disponible ainsi qu'un cadre de vie agréable à seulement 1H30 en TGV du centre de Paris.

Alors n'hésitez plus, nous vous attendons !

Hervé Frezza-Buet et Frédéric Pennerath, responsables de la mention SDI-Metz

Plus d'information pour candidater à

http://sdi.metz.centralesupelec.fr

Course overview

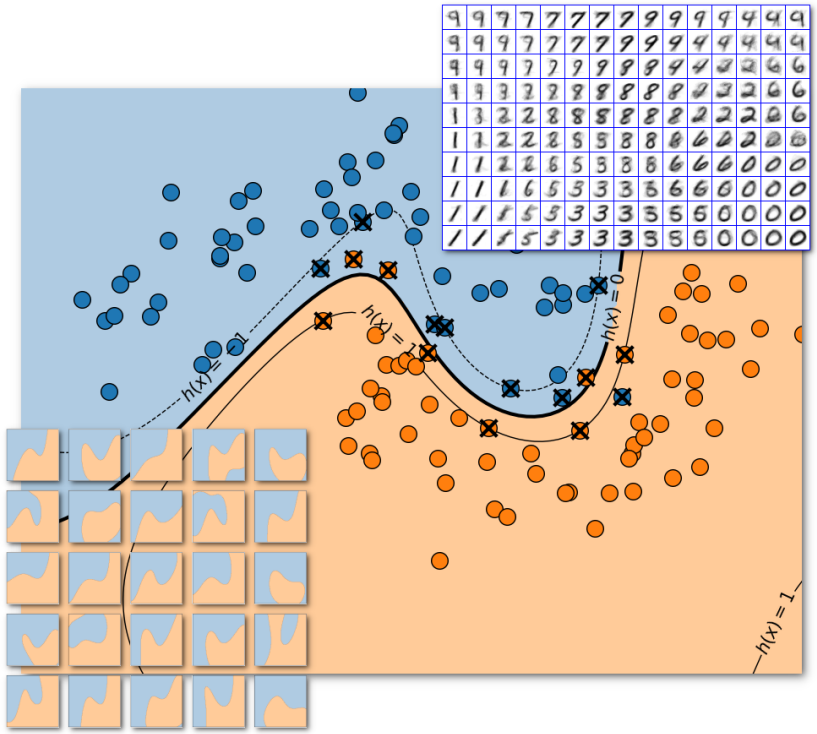
Open the black boxes of machine learning algorithms!

An overview of the whole field

- Different ways of learning
- Introduction to the statistical approach (risks, overfitting, ...)
- Main principles of data manipulation (dimensionality reduction, ...)

A comprehensive approach of algorithms

- Support vector machines and kernel methods
- Ensemble methods (boosting, bagging, décision trees)
- Vector quantization (growing neural gas, self-organizing maps)



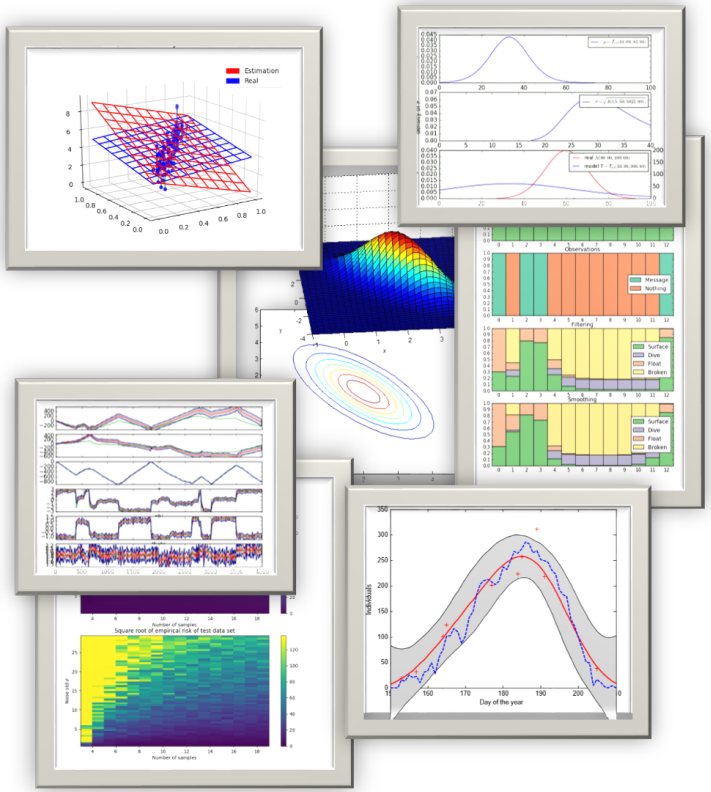
Become a data scientist solid and credible in statistics

Get a good grounding in statistics.

- Understand fundamental notions (frequentist and Bayesian estimation, belief networks, exp. models, causality, stochastic processes, etc).
- Gather these notions as theoretical ingredients to build up statistical models and efficient algorithms to learn their parameters from data.
- Discover useful models (NB, LDA/QDA, GLM, GP, HMM, KF, etc) and their application scopes.

Become a practitioner in statistical modeling.

- Play and tune models on datasets.
- Implement and test by yourself some models in Python.
- Design your own statistical models.



Efficient programming can be smart!

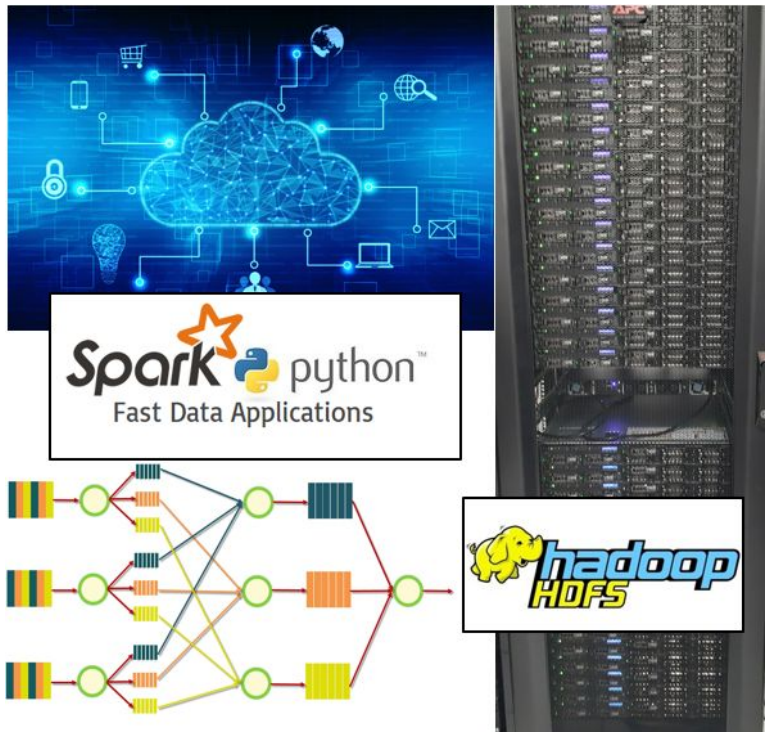
Strong typing in C++

- C++11/14/17/20... offers smart concepts. Typing is crucial.
- Standard Template Library design is efficient and close to theoretical algorithmics.
- Type checking annoys you at compiling time... but makes your code robust!
- Get the power of templates, it sounds like maths.
- Think functional!

System

- Learn to invoke system features for efficient executions.
- Understand how memory works.
- Use threads and shared data.

Large-scale data analysis requires the use of distributed platforms for data storage and processing, such as "Spark clusters" or "Clouds". Specific distributed programming models must then be used, and optimization of algorithms and codes is necessary to successfully "scale up".



Content

- Hadoop distributed file system (HDFS)
- Spark programming model on PC cluster
- Data analysis framework on Cloud
- "Scaling" metric

Develop your data science application like a pro!

Topics

- software application architecture, service-oriented architecture, middleware
- continuous delivery, continuous deployment, cloud, virtualization, containers
- continuous integration

Approach

- 4 lectures to present concepts
- 5 tutorials to get you started on widely used tools (Kafka, Docker, Kubernetes, GitLab)
- 6 lab sessions on a live wire example (track most active Wikipedia pages over time), deployed on a cluster

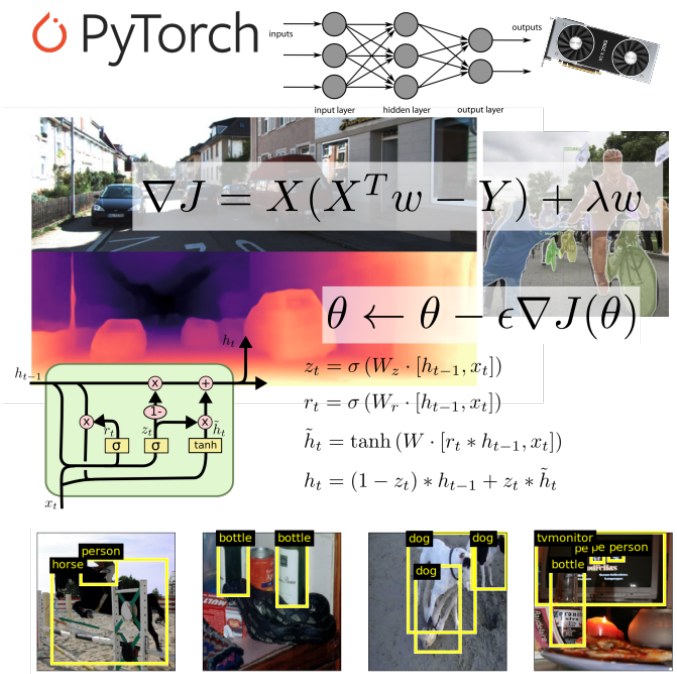
Learn and practice with deep neural networks

Understanding every single piece

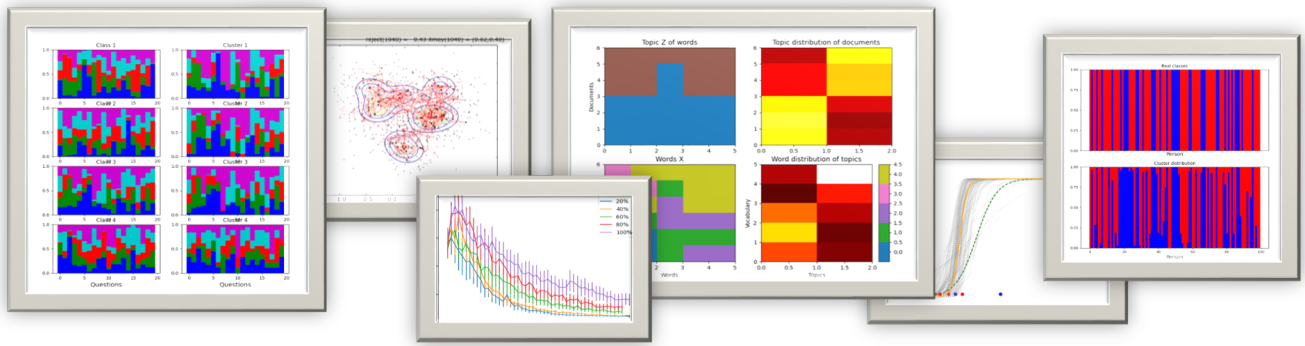
- Neural network architectures (FFNN, CNN, RNN)
- Computational graphs
- Initialization and learning with gradient descents

Practice on the GPUs with pytorch

- Introduction to deep learning with pytorch
- Object detection with convolutional neural networks
- Sequence processing with recurrent neural networks
- Generative neural networks



Discover how to apply statistical models to complex and large problems.



Get the fundamental ingredients

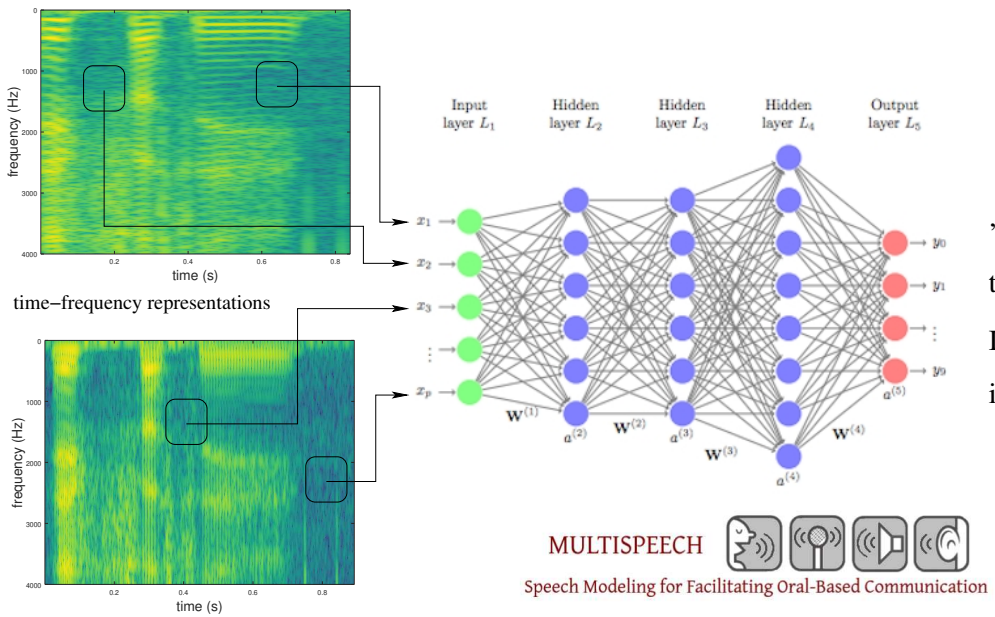
- Learn algorithmic techniques for approximate estimation based on sampling (MC, MCMC, PF, etc) and variational inference.
- Discover applications of these techniques to different problems (Latent Dirichlet Alloc., Bayesian deep learn., variat. autoencoders, etc)
- Implement these methods to solve problems.

Learn fundam. algorithms in data science

- Information retrieval
- Recommender systems
- Social network analysis
- Stream processing and data sketching
- Data mining

Learn and practice numerical sound processing.

- Sound perception and production
- Speech processing, including speech recognition, acquisition, manipulation, storage, transfer and synthesis
- Human machine communication

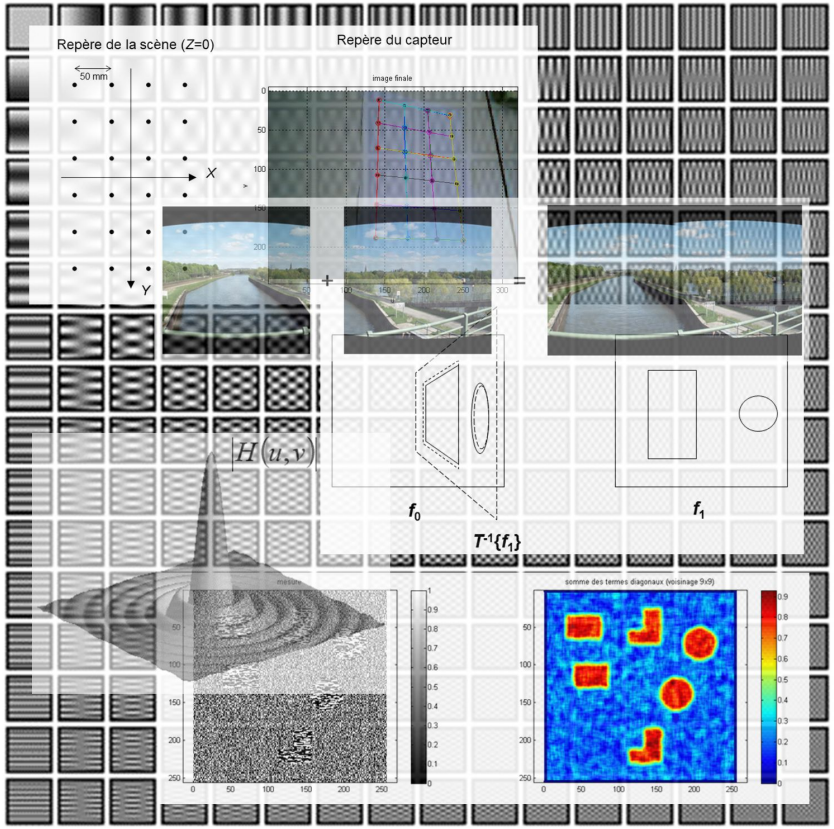


''This is the thing that I think is thrilling''

SG10

Image Processing (Jean-Luc Collette)

- Visual perception, photometry: Description of a scene. Definition of physical quantities in the field of optics.
- Image sensor: Physical and geometric modeling. Calibration problem. Generalization.
- Continuous systems, digitization: Linear modeling of optical systems. Sampling.
- Geometric transformations: Optimization methods for image registration. Choice of degrees of freedom.
- Linear transformations: Orthogonal transformations, wavelets, filter banks. Application to coding and compression.
- Features on images Problem of detection and classification. Statistical, geometrical features.

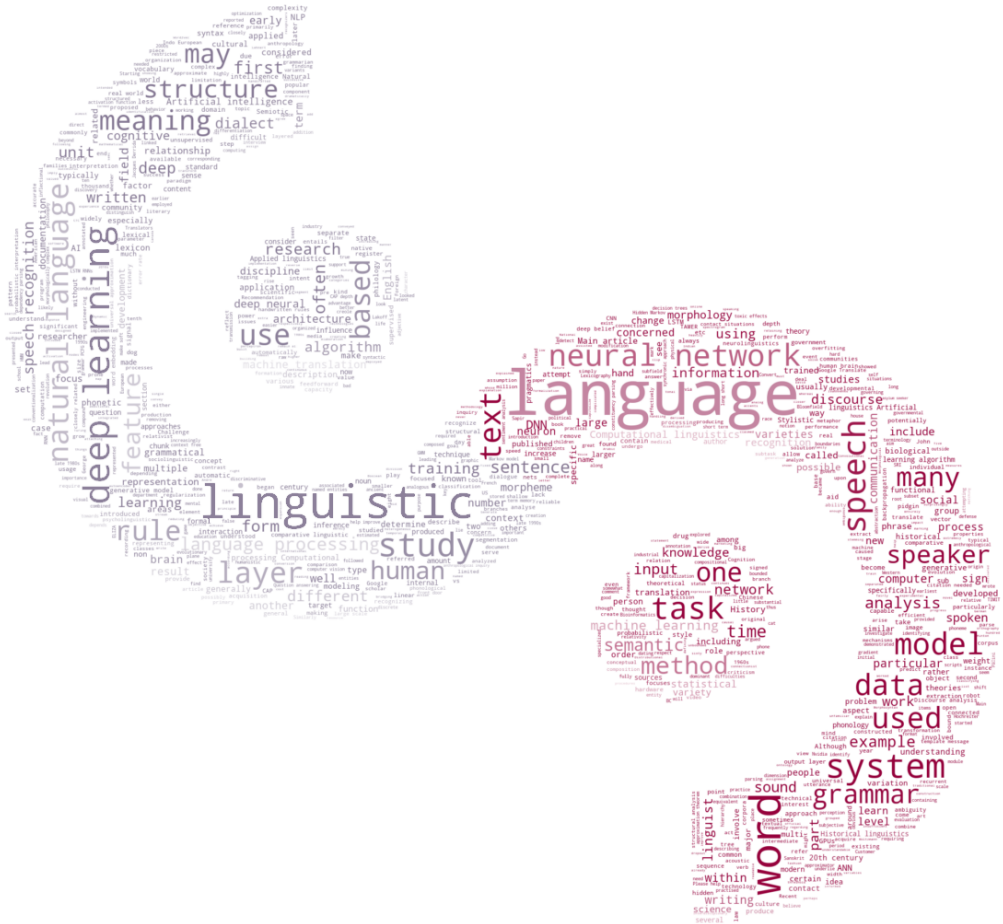


Topics

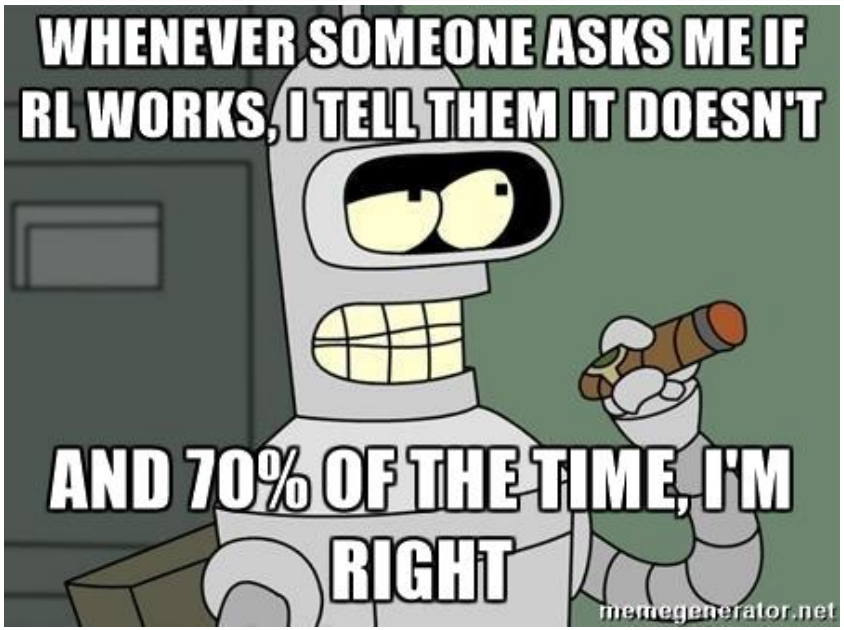
- Understand the theoretical foundations for conceptualizing and modelling linguistic phenomena.
- Acquire autonomy for the automatic processing of large-scale textual content.
- Using state of the art deep learning approaches (Recurrent and recursive neural networks, seq-to-seq, attention based models, ...)

Approach

- 9 lectures to present theoretical concepts
- Each of them directly followed by a practical session
- A final project based on a real industrial problem using real data

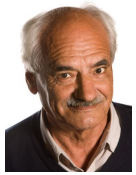


Using Deep Reinforcement Learning, computers can beat human experts at video games (Atari, Starcraft, Dota) and even at the game of Go. Such achievements were not within reach only 5 years ago. What has changed ? What is the theory behind all this ? And still, why is the following picture quite right ?



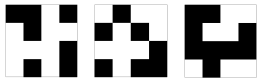
This course will try to make you able to answer theses questions and, thus, shine in salon conversation.

- Formalization of supervised learning problems
- PAC learning capacity and uniform convergence
- The bias-complexity trade-off
- The VC (Vapnik-Chervonenkis) dimension of a hypothesis space
- Two fundamental theorems of PAC learning


Alexey Chervonenkis

Statistical learning theory
Supervised learning
Generalization
PAC learning
$\text{dimVC} < \infty$


Vladimir Vapnik


$\mathbf{x}_1$ $\mathbf{x}_2$ $\mathbf{x}_3$

$f(\mathbf{x}_1) = f(\mathbf{x}_2) = f(\mathbf{x}_3) = -1$

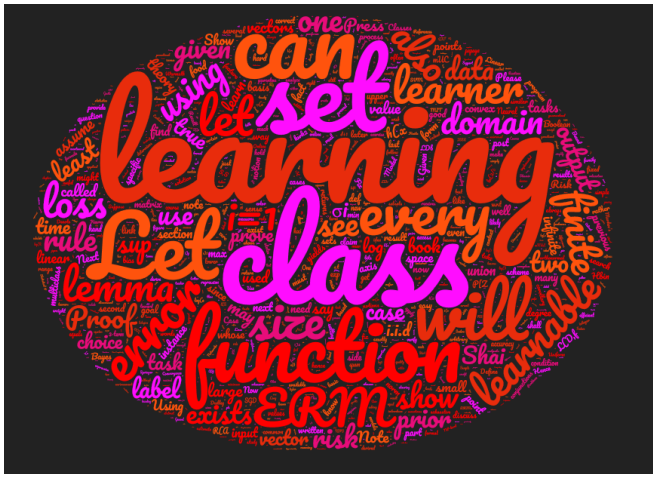

$\mathbf{x}_4$ $\mathbf{x}_5$ $\mathbf{x}_6$

$f(\mathbf{x}_4) = f(\mathbf{x}_5) = f(\mathbf{x}_6) = 1$


$\mathbf{x}_7$

$f(\mathbf{x}_7) = ?$

Learning f from a dataset seems difficult, what is required ?



GPU computing increases the application performance and their power efficiency. Unfortunately, it requires original algorithms and programming, and some computations still remain more efficient on the CPU. Hence the need to develop "hybrid" algorithms and codes running on couple CPU + GPU.

Content

- Massively parallel computing on GPU
- Hybrid computing on CPU+GPU
- CUDA programming
- Clustering algorithms on CPU+GPU architecture

Sparse models for data science

- Data processing tasks are complicated in high-dimension, noise, etc.
- Sparsity (parsimony aka Occam's razor) is useful as a problem-solving principle.
- Meaningful data feature extraction & dimension reduction.
- Efficient statistical models for data analysis & visualization.