



Séminaire Thématique “Traitement et analyse statistique des données massives”

Impacts Sociétaux des Technologies de l'Information

*Du mercredi 19 novembre au vendredi 21 novembre 2014
campus du Futuroscope*

Programme du mercredi 19 novembre

14h30-15h00	<i>Accueil :</i> Samuel Boissière , Professeur à l'Université de Poitiers, directeur-adjoint de l'Ecole Doctorale S2IM Ladjel Bellatreche , Professeur à l'ENSMA, directeur-adjoint de l'Ecole Doctorale S2IM
15h00-16h00	<i>Conférence :</i> Nicolas Vayatis , Professeur à l'ENS Cachan “ <i>La transition numérique : applications du machine learning dans les sciences de l'ingénieur, le e-marketing et la médecine</i> ”
16h00-16h30	<i>Pause café</i>
16h30-17h30	<i>Présentation des posters des doctorants :</i> Jérémy Hyvert (XLIM/SIC) Zahira Chouiref (LIAS) Clément Chesseboeuf (LMA)
17h30-18h30	<i>Exposés des doctorants :</i> Selma Bouarar (LIAS) “ <i>La variabilité dans les bases de données avancées</i> ” Ahcene Boukorca (LIAS/ENSMA) “ <i>A scalable structure for big queries management</i> ”
18h45-20h00	Dîner au restaurant ESEN

Programme du jeudi 20 novembre

9h30-10h30	<i>Conférence :</i> Cristian Preda , Professeur à Polytech'Lille "Modèles de régression avec données fonctionnelles dans des espaces de Hilbert à noyau reproduisant"
10h30-11h00	<i>Pause café</i>
11h00-12h00	<i>Conférence :</i> Nicolas Le Roux , Scientific Program Manager, CRITEO France "15 millions de prédictions par seconde: les défis scientifiques et techniques de Criteo"
12h30-14h00	<i>Déjeuner au RU Gémini</i>
14h00-15h00	<i>Conférence :</i> Stéphane Mallat , Professeur à l'ENS Ulm "Apprendre en grande dimension: de l'image à la chimie quantique"
15h00-15h20	Anne Philippe , Professeure à l'Université de Nantes <i>Présentation de l'AMIES</i>
15h20-16h00	<i>Présentation des posters des doctorants :</i> Zakaria Zettaf (XLIM/SIC) Vincent Aide (LMA)
16h00-16h30	<i>Pause café</i>
16h30-18h00	<i>Présentation des posters des doctorants :</i> Nadir Guetmi (LIAS) Zoé Amblard (XLIM/DMI) Malek Mohamed (XLIM/SIC)
18h30-21h00	<i>Buffet dinatoire</i>

Programme du vendredi 21 novembre

9h30-10h30	<i>Conférence :</i> Sebastiao Correia , Directeur du développement de produits, TALEND France “ <i>Big Data, un nouveau paradigme et de nouveaux challenges</i> ”
10h30-11h00	<i>Pause café</i>
11h00-11h20	<i>Présentation des posters des doctorants :</i> Zouhir Djilani (LIAS)
11h20-12h20	<i>Exposés des doctorants :</i> Geraud Fokou (LIAS/ENSMA) “ <i>Les systèmes de réponses coopératives dans le contexte du big data</i> ” Daouda Sawadogo (L3I) “ <i>Gestion consolidée de ressources numériques dans un environnement personnalisé</i> ”
12h20-12h30	<i>Clôture</i>
12h45-14h00	<i>Déjeuner au restaurant ESEN</i>

Résumés des conférences

Nicolas Vayatis, “*La transition numérique : applications du machine learning dans les sciences de l'ingénieur, le e-marketing, et la médecine*”

Dans toutes les branches de l'activité humaine, l'ubiquité des capteurs et l'accumulation de l'information digitale soulèvent de nouveaux défis intellectuels, des rêves technologiques mais aussi des craintes quant au respect de la vie privée, au libre-arbitre, à la diversité culturelle, ... Les avancées récentes dans les domaines de la statistique en grande dimension, la description et la modélisation de processus sur des réseaux, et la deuxième (?) vie de la théorie de l'optimisation ont conduit à des concepts et des algorithmes qui rendent possible l'inférence à partir de données complexes et ouvrent des perspectives quant à l'interaction entre experts et scientifiques dans divers champs d'application. Un point de tension dans cette démarche est l'équilibre à trouver entre la personnalisation des solutions et la reproductibilité des travaux. Pourtant ces deux aspects devront, selon moi, guider les innovations futures dans le domaine du machine learning. Dans l'exposé, j'illustrerai ces idées en parcourant quelques réalisations récentes résultant de projets interdisciplinaires dans les domaines comme la mécanique des fluides, du marketing digital, et de l'éthomique.

Cristian Preda, “*Modèles de régression avec données fonctionnelles dans des espaces de Hilbert à noyau reproduisant*”

Nous considérons le problème de l'estimation de la fonction de régression avec prédicteur une variable aléatoire à valeurs dans un espace fonctions. La recherche des estimateurs se fait dans la classe des fonctionnelles appartenant à un espace de Hilbert à noyau reproduisant. Des résultats numériques sont présentés.

Nicolas Le Roux, *“15 millions de prédictions par seconde: les défis scientifiques et techniques de Criteo”*

Criteo affiche chaque seconde des dizaines de milliers de bannières publicitaires et enregistre chaque jour des milliards d'événements pour construire ses modèles de prédiction. Cela introduit des contraintes fortes sur les algorithmes de prédiction et d'optimisation et nécessite à la fois des solutions techniques et scientifiques. La plupart de ces affichages se faisant au travers d'enchères, il existe également de nombreuses problématiques proches de la théorie des jeux que Criteo se doit de traiter afin de pouvoir être efficace.

Stéphane Mallat, *“Apprendre en grande dimension: de l'image à la chimie quantique”*

Le déluge de données, l'explosion des capacités mémoires et calculs, ouvrent un domaine scientifique qui semblait inaccessible il y a quelques années. Certains algorithmes semblent être capables de contourner la malédiction liée à la très grande dimensionnalité des données. En particulier, les réseaux de neurones profonds atteignent des performances spectaculaires pour la classification d'images et de données numériques très diverses.

Pourtant l'apprentissage ne semble pas être le noyau de l'histoire. Nous montrerons que ces réseaux non-linéaires calculent des invariants multi-échelles, permettant de caractériser des propriétés géométriques complexes aussi bien que des processus aléatoires. Au-delà des applications à la classification d'images, nous montrerons que ces architectures peuvent apprendre des fonctionnelles physiques, telles que l'énergie de molécules en chimie quantique.

Sebastiao Correia, *“Big Data, un nouveau paradigme et de nouveaux challenges”*

La révolution big data a bouleversé notre façon d'appréhender la donnée. La donnée arrive maintenant dans un flot continu et incessant. Pour la traiter, il a fallu un changement de paradigme qui rend obsolètes les anciens systèmes et nécessite de nouveaux outils. Hadoop fait parti de ces nouveaux systèmes qui permettent de gérer le déluge de données.

Une nouvelle architecture technique voit le jour, l'architecture Lambda, dans laquelle les données sont immuables. Les mises à jour des données sont supprimées, les schémas des données sont souples (schema-on-read) et les données elles-mêmes ne sont plus ou peu déplacées une fois qu'elles sont entrées dans le système. Au contraire, ce sont les algorithmes qui sont déportés au plus près de la donnée. Cette approche prend le contre-pied des anciens systèmes qui recevaient la donnée afin de la traiter (SGBD).

De nouvelles utilisations de la donnée sont apparues grâce à cette nouvelle possibilité de traiter une quantité d'information jusque-là inenvisageable. Mais, si cette architecture résout plusieurs problèmes liés à la volumétrie et au flux des données, d'autres problèmes ayant trait à la qualité de l'information surgissent. Sans parler des questions touchant à la vie privée.