PROGRAMME

Tuesday, 10 July

- 9h00 : Welcome
- 9h30 : **Opening Speeches -** Alain Beretz (DGRI - MESRI), Mr. Toshihiko Horiuchi, Minister, Embassy of Japan in France
- 9h45 : **Symposium DATAIA and CREST Program Introduction: Scope & Objectives** - Nozha Boujemaa (Director of DATAIA Institute, Research Director Inria) & Masaru Kitsuregawa (Director General, National Institute of Informatics / Professor, The University of Tokyo)
- 10h30 : **Keynote** "Bayesian Machine Learning" - Eric Moulines, professeur au Centre de mathématiques appliquées de l'École polytechnique, membre de l'Académie des Sciences
- 11h00 : Coffee-Break
- 11h30 : **Keynote** "Shallow and Deep learning at TAU" - Marc Schoenauer, directeur de recherche Inria
- 12h00 : **Oral Plenary Session**
    - Tatsuya Harada (The University of Tokyo) - Harada Team (CREST Bigdata Core technology) : "Recognition, Summarization and Retrieval of Large-Scale Multimedia Data".
    - Gaël Varoquaux & Nicolas Prost (Inria, Institut DATAIA) : "MissingBigData: missing data in the big data era"

- 12h30 : Lunch Cocktail
- 14h00 : **Oral Plenary Session**
    - Satoshi Matsuoka (Tokyo Institute of Technology) & Osamu Tatebe (University of Tsukuba) - Matsuoka Team (CREST Bigdata Core technology) : "EBD: Extreme Big Data – Convergence of Big Data and HPC for Yottabyte Processing"
    - Philippe Caillou, (LRI - Institut DATAIA) : "Valorisation of Data for Job Research (VADORE)"
    - Masato Oguchi (Ochanomizu University), Seigo Arita (INSTITUTE of INFORMATION SECURITY), Yoshiko Yasumura (Waseda University) & Sari Handa (INSTITUTE of INFORMATION SECURITY) - Yamana Team (CREST Bigdata Core technology) : "Secure Data Sharing and Distribution Platform for Integrated Big Data Utilization"
    - Célia Zolynski (UVSQ - Institut DATAIA) & Nicolas Anciaux (Inria - Institut DATAIA) : "RGPD and Personal Cloud: from Empowerment to Responsibility (GDP-ERE)"
    - Florence d'Alché-Buc (Telecom ParisTech) : "Structured Output Learning with Abstention: application to opinion prediction"
- 15h30 : Coffee-Break
- 16h00 : **Oral Plenary Session**
    - Teruaki Hayashi (The University of Tokyo) - Yamanishi Team (CREST Bigdata Core technology) : "Knowledge Structuring for Cross-disciplinary Data Exchange and Collaboration"
    - Bertrand Thirion (Inria) : "Toward rigorous e-science: statistical inference on high-dimensional models"
    - Masaya Saito (The Institute of Statistical Mathematics) - Nishiura Team (CREST Bigdata Applications) : "Detecting premonitory signs and real-time forecasting of pandemic using big biological data"
    - Claire Nédellec (MaIAGE-INRA) : "Ontology-based text mining for microbiology research"
- 17h00 : **Break-Out Session**
- 18h00 : End

Wednesday, 11 July

- 9h30 : **Keynote** " Co-experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration" - Norihiro Hagita, Board Director, Director, Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International

- 10h00 : **Keynote** "AI and HCI - Towards Human-Computer Partnerships" - Michel Beaudouin-Lafon, Professeur d'Informatique à l'Université Paris-Sud et membre senior de l'Institut Universitaire de France
- 10h30 : **Keynote** "Big Data Applications: Opportunities and Challenges" Yuzuru Tanaka 1) 2) ,Research Supervisor, JST CREST Program on "Big Data Applications", Professor Emeritus, Hokkaido University
  1) MaDIS, NIMS 2) Faculty of Engineering, Hokkaigakuen University

- 11h00 : Coffee-Break

- 11h30 : **Oral Plenary Session**
  - Koichi Kise (Osaka Prefecture University) & Olivier Augereau (Osaka Prefecture University) - Kise Team (CREST Intelligent Information Processing) : "Quantified Learning by Deeply Sensing Learner's Behavior"
  - Laurence Devillers (LIMSI-CNRS - Institut DATAIA) : "Bad Nudge - Bad Robot ? : Nudge and Ethics in human-machine verbal interaction"
  - Yuji Matsumoto (Nara Institute of Science and Technology) & Ken Satoh (National Institute of Informatics) - Matsumoto Team (CREST Bigdata Applications) : "Knowledge Discovery through Structural Document Understanding"
  - Pierre Zweigenbaum (LIMSI) : "Natural Language Processing for e-Health"
  - Yutaka Yoshimoto, Deputy Director General (METI)

- 12h30 : Lunch Cocktail

- 14h00 : **Oral Plenary Session**
  - Albert Bifet (LTCI) : "Machine Learning for Data Streams"
  - Cédric Gouy-Pailler (CEA - Institut DATAIA) : "StreamOps : Open Source Platform for Research and Integration of Algorithms for Massive Time Series Flow Analysis"

- Naoki Katoh (Kwansei Gakuin University) & Chako Takahashi (Tohoku University) - Katoh Team (CREST Bigdata Core technology) : "Foundations of Innovative Algorithms for Big Data"
- Hossam Afifi (Télécom SudParis - Institut DATAIA), Jordi Badosa (Ecole polytechnique - Institut DATAIA) & Florence Ossart (CentraleSupélec - Institut DATAIA) : "PEPER: prediction of prosummers with reinforcement deep learning"
- Takemasa Miyoshi (RIKEN) - Miyoshi Team (CREST Bigdata Applications) : "Innovating "Big Data Assimilation" technology for revolutionizing very-short-range severe weather prediction"

- 15h30 : Coffee-Break
- 16h00 : **Oral Plenary Session**
  - Shuichi Onami (RIKEN) & Koji Koyamada (Kyoto University) - Onami Team (CREST Bigdata Applications) : "Data-driven analysis of the mechanism of animal development"
  - Vincent Fromion (MaIAGE, INRA, Université Paris-Saclay) : "Representation of biological data, information and knowledge: opportunities offered by systemic biology"
  - Masayuki Hirafuji (The Universiy of Tokyo) & Guo Wei (The Universiy of Tokyo) - Hirafuji Team (CREST Bigdata Applications) : "Knowledge Discovery by Constructing AgriBigData"
  - Hervé Monod (MIA department, INRA) : "#DigitAg, the French Digital Agriculture Convergence Lab"
- 17h00 : **Break-Out Session**
- 18h00 : End

Thursday, 12 July

- 9h30 : **Oral Plenary Session**
    - Christophe Prieur (Telecom ParisTech) : "HistorIA : Large historical databases. Data mining, exploration and explicability"
    - Takeaki Uno (National Institute of Informatics) & Kunihiro Wasa (National Institute of Informatics) - Uno Team (CREST Bigdata Core technology) : "Data Particlization for Next Generation Data Mining"
    - David Restrepo Amariles (HEC Paris) : "Smart Lawyer: Rating Legal Services in the Courtroom"
- 10h30 : Coffee-Break
- 11h00 : **Oral Plenary Session**
    - Shunichi Koshimura (Tohoku University) - Koshimura Team (CREST Bigdata Applications) : "Establishing the most advanced disaster reduction management system by fusion of real-time disaster simulation and big data assimilation"
    - Gregory Blanc & Mustafizur Shahid (Télécom SudParis) : "Machine Learning Based Intrusion Detection System for IoT Network"
    - Masaharu Munetomo (Hokkaido university) - Aida Team (CREST Bigdata Core technology) : "Optimal Resource Selection in Application-Centric Overlay Cloud Utilizing Inter-Cloud"
- 11h45 : **Wrap-up Discussion-** Nozha Boujemaa & Masaru Kitsuregawa
- 12h45 : **Closing of the event**

PROGRAMME WITH ABSTRACTS

Tuesday, 10 July

- 9h00 : Welcome
- 9h30 : **Opening Speeches -** Alain Beretz (DGRI - MESRI), Mr. Toshihiko Horiuchi, Minister, Embassy of Japan in France
- 9h45 : **Symposium DATAIA and CREST Program Introduction: Scope & Objectives** - Nozha Boujemaa (Director of DATAIA Institute, Research Director Inria) & Masaru Kitsuregawa (Director General, National Institute of Informatics / Professor, The University of Tokyo)
- 10h30 : **Keynote** "Bayesian Machine Learning" - Eric Moulines, professeur au Centre de mathématiques appliquées de l'École polytechnique, membre de l'Académie des Sciences

Abstract: Stochastic Gradient Langevin Dynamics (SGLD) has emerged as a key MCMC algorithm for Bayesian learning from large scale datasets. While SGLD with decreasing step sizes converges weakly to the posterior distribution, the algorithm is often used with a constant step size in practice and has demonstrated spectacular successes in machine learning tasks.

The current practice is to set the step size inversely proportional to N where N is the number of training samples. As N becomes large, we show that the SGLD algorithm has an invariant probability measure which significantly departs from the target posterior and behaves like as Stochastic Gradient Descent (SGD). This difference is inherently due to the high variance of the stochastic gradients.

Several strategies have been suggested to reduce this effect; among them, SGLD Fixed Point (SGLDFP) uses carefully designed control variates to reduce the variance of the stochastic gradients. We show that SGLDFP gives approximate samples from the posterior distribution, with an accuracy comparable to the Langevin Monte Carlo (LMC) algorithm for a computational cost sublinear in the number of data points.

We provide a detailed analysis of the Wasserstein distances between LMC, SGLD, SGLDFP and SGD and explicit expressions of the means and covariance matrices of their invariant distributions. Our findings are supported by limited numerical experiments.

- 11h00 : Coffee-Break
- 11h30 : **Keynote** "Shallow and Deep learning at TAU" - Marc Schoenauer, directeur de recherche Inria

- 12h00 : **Oral Plenary Session**
    - Tatsuya Harada (The University of Tokyo) - Harada Team (CREST Bigdata Core technology) : "Recognition, Summarization and Retrieval of Large-Scale Multimedia Data"

Abstract: Our goal is development of fundamental technologies to recognize and summarize the large-scale multimedia data and enable users to understand the content of the data quickly without checking all of them. It is impossible for human to manually examine the huge amount of multimedia data such as image, video, audio, text etc. in the internet since they are uploaded every day from the physical world and are increasing rapidly. If the system automatically recognizes and summarizes the multimedia data, the system not only helps users to grasp their content, but also extends the existing text based search engine to retrieve them using natural language queries.

    - Gaël Varoquaux & Nicolas Prost (Inria, Institut DATAIA) : "MissingBigData: missing data in the big data era"

Abstract: "big data", often observational and compound, rather than experimental and homogeneous, poses missing-data challenges: missing values are structured, non independent of the outcome variables of interest. We propose to use more powerful models that can benefit from the large sample sizes, specifically autoencoders, to impute missing values. To avoid biasing conclusions, we will study multiple imputation and conditions on the dependencies in the data.

Our project will enable proper causal interpretations of the risk factors despite data missing not at random. We seek an operational solution, from the methodology to the implementation, that integrate the diversity of data and of questions while dealing with larger data. Indeed, combining predictive models with causal inferences is classic and existing methodologies focus on one or other of the problems, though the missing-data methodology impacts the whole process. We will also depart from classic studies by considering multiple types of missing data and MNAR on multiple variable. This will be a first, but seems feasible in view of the theoretical results of Mohan and Pearl (2018).

- 12h30 : Lunch Cocktail

- 14h00 : **Oral Plenary Session**
    - Satoshi Matsuoka (Tokyo Institute of Technology) & Osamu Tatebe (University of Tsukuba) - Matsuoka Team (CREST Bigdata Core technology) : "EBD: Extreme Big Data – Convergence of Big Data and HPC for Yottabyte Processing"

Abstract: Althought the data being handled in the current "Big Data" infrastructure is often actually not so "big" by HPC standards, in the future they are expect to explode by several orders of magnitude, both in terms of their capacity and complexity. This poses immence problems for the exisiting IDC/Cloud "Big Data" infrastructures due to their lack of system bandwidth and processing capacity, as well as for HPC/Supercomputers because of their lack of real-time processing capabilities etc. Our work will focus on the next generation "Extreme Big Data" infrastructure by developing a set of technologies and the resulting system attaining their convergence through co-designing with representatiive future big data applications, aiming for up to 100,000 times improvement in the data processing capabilities over the next 10 years.

- Philippe Caillou, (LRI - Institut DATAIA) : "Valorisation of Data for Job Research (VADORE)"

Abstract: Our project focuses on unemployment in France. Unemployment has many causes, and they involve mainly factors limiting labor supply and demand. Unemployment can also hinge on the efficiency of the matching process that can pair demand with supply. In many cases there might be imperfections in the process, leading to the notion of frictional unemployment. Frictions in the labor market correspond to the case in which in a « micro labor market » there are both unfilled vacancies and jobseekers who would be willing to take them. These frictions in the labor market are related to imperfect information, and mostly due to the cost of collecting, processing and disseminating information; information asymmetry between employers and jobseekers; cognitive limitations of individuals that prevent them from scanning large numbers of job ads.

The central idea of the project is to mobilize all available information to improve the matching of jobseekers and vacancies. The project relies on the mobilization of the considerable body of information available at the Public employment Service in France on jobseekers and companies, some of which (textual data in particular) are still unexploited. This information will be used to develop two functionalities aimed at improving the matching process in the labor market for both jobseekers and firms: first a recommendation engine, and second an interactive personalized map for a jobseeker that will help her see the job market in her region/domain of competence and better appreciate her opportunities. One important aspect of the project is that the two functionalities will be evaluated using randomized control trial. The evaluation will be performed so as to identify the impact on inequalities in the labor market, e.g., whether using the tools leads to displacement effects.

- Masato Oguchi (Ochanomizu University), Seigo Arita (INSTITUTE of INFORMATION SECURITY), Yoshiko Yasumura (Waseda University) & Sari Handa (INSTITUTE of INFORMATION SECURITY) - Yamana Team (CREST Bigdata Core technology) : "Secure Data Sharing and Distribution Platform for Integrated Big Data Utilization"

Abstract: To promote and boost the development of big data applications, the infrastructure, such that contents providers are able to provide any data securely and that contents users are able to use their analyzed result without any doubt, becomes indispensable. To respond the requirement, we will construct a new infrastructure which handles all the data with encryption by shaking ourselves free from anonymization or secure communication. Our final goal is to speed-up encrypted calculation over 1,000 times over current methods by theoretical and computer architecture optimization-approaches. Here, the encrypted calculation consists of fully homomorphic encryption, proof of storage, and attribute-based encryption.

- Célia Zolynski (UVSQ - Institut DATAIA) & Nicolas Anciaux (Inria - Institut DATAIA) : "RGPD and Personal Cloud: from Empowerment to Responsibility (GDP-ERE)"

Abstract : In a world disrupted by Artificial Intelligence and the exploitation of personal data, the role of individuals and the control of their data is a central issue in the new European regulation (GDPR), enforced on 25th May 2018, after the French Numeric Republic regulation adopted in 2016. Data portability is a new right provided under those regulations, introduced as a key element for powerful legal and technical tools. Building upon many projects such as Blue Button (health data) and Green Button (energy consumption data) in the US, MiData (energy, financial, telecommunications and retail data) in the UK or MesInfos in France, data portability allows citizens to retrieve their personal data from the companies and governmental agencies that

collected them, in an interoperable digital format. Data portability thus gives the individual the ability to get out of a captive ecosystem, to regain control of his or her personal data towards empowerment and informational self-determination. It also opens to important societal benefits when individuals collectively decide to make their data available to public service missions, citizen actions (e.g., « in vivo » tests of an algorithm presenting risks of bias) or a scientific studies (e.g., epidemiological surveys). Finally, data portability represents a new vector of development for innovative and virtuous personal data economy beyond the existing de facto monopolistic positions.

The consequence of this new right is the design and deployment of technical platforms, commonly known as Personal Cloud. Individuals are thus able to retrieve all their personal data from different data silos and group them in a single system, with the ability to control all access and usage in favor of innovative services. Yet, managing all the « digital assets » of an individual in one platform obviously raises security and privacy issues. But personal cloud architectures are very diverse, ranging from cloud based solutions where millions of personal cloud are managed centrally, to self-hosting solutions where the individual installs a personal server on his or her own own equipment. These architectural choices are not neutral both in terms of security (risk of massive attacks for centralised solutions versus lack of IT expertise of individuals for decentralised solutions) and from the point of view of the chain of liabilities. This last point particularly deserves to be studied. For example, considering a context of self-hosting under the angle of liabilities imposes to reexamine the role of every actor involved (controller, processor, third party) and redefine their respective prerogatives and obligations. The GDP-ERE project tends to solve those issues in an interdisciplinary approach by the involvement of jurists and computers scientists. Two main objectives are sought under this project: (i) analysis of the effects of the personal cloud architectures on legal liabilities, enlightened by the analysis of the rules provided under the GDPR; (ii) proposals of legal and technological evolutions to highlight the share of liability between each relevant party, and create adapted tools to endorse those liabilities.

GDP-ERE project builds on an existing collaboration between jurists researchers and computer scientists. Researcher activities are supported by some national authorities (including the French data protection authority, CNIL) and personal cloud platform providers.

- Florence d'Alché-Buc (Telecom ParisTech) : "Structured Output Learning with Abstention: application to opinion prediction"

Abstract: Motivated by Supervised Opinion Analysis, we propose a novel framework devoted to Structured Output Learning with Abstention (SOLA). The structure prediction model is able to abstain from predicting some labels in the structured output at a cost chosen by the user in a flexible way. For that purpose, we decompose the problem into the learning of a pair of predictors, one devoted to structured abstention and the other, to structured output prediction. To compare fully labeled training data with predictions potentially containing abstentions, we define a wide class of asymmetric abstention-aware losses. Learning is achieved by surrogate regression in an appropriate feature space while prediction with abstention is performed by solving a new pre-image problem. Thus, SOLA extends recent ideas about Structured Output Prediction via surrogate problems and calibration theory and enjoys statistical guarantees on the resulting excess risk. Instantiated on a hierarchical abstention-aware loss, SOLA is shown to be relevant for fine-grained opinion mining and gives state-of-the-art results on this task.

- 15h30 : Coffee-Break

- 16h00 : **Oral Plenary Session**
    - Teruaki Hayashi (The University of Tokyo) - Yamanishi Team (CREST Bigdata Core technology) : "Knowledge Structuring for Cross-disciplinary Data Exchange and Collaboration"

Abstract: The Bigdata that we deal with today is not only huge but also extremely complex. In order to utilize such complex data, it is important to discover knowledge that exists in the latent space that is not observable but lies deeply behind the data. We call such knowledge the "deep knowledge." In this research project, we aim at developing novel mathematical methodologies for discovering deep knowledge from complex data and creating values from it. We are concerned explicitly with deep knowledge discovery from a complex network in which a massive number of small data sets are connected, and each data is heterogeneous and dynamic. The Data Jacket (DJ) is one of the solutions. The idea underlying the DJ is to share "a summary of data" as metadata without sharing the data itself. Metadata is conventionally a data description format for enhancing the readability of machines. On the contrary, the DJ is a method for describing summarized information about data for humans to read and understand the utility of data. In this presentation, we introduce our latest technologies for activating cross-disciplinary data exchange and collaboration by structuring the knowledge of data utilization using DJs.

    - Bertrand Thirion (Inria) : "Toward rigorous e-science: statistical inference on high-dimensional models"

Abstract: Medical imaging involves high-dimensional data, yet their acquisition is obtained for limited samples. Multivariate predictive mod- els have become popular in the last decades to fit some external variables from imaging data, and standard algorithms yield point estimates of the model parameters. It is however challenging to attribute confidence to these parameter estimates, which makes solutions hardly trustworthy. In this presentation we discus a new algorithm that assesses parameters statisti- cal significance and that can scale even when the number of predictors p ≥ 10 5 is much higher than the number of samples n ≤ 10 3, by lever- aging structure among features. Our algorithm combines three main ingredients: a powerful inference procedure for linear models –the so-called Desparsified Lasso– feature clustering and an ensembling step. We provide theoretical and empirical evidence for this approach.

    - Masaya Saito (The Institute of Statistical Mathematics) - Nishiura Team (CREST Bigdata Applications) : "Detecting premonitory signs and real-time forecasting of pandemic using big biological data"

Abstract: We aim to achieve detection of premonitory signs and real-time forecasting of pandemic, thereby elucidating the most effective countermeasures. Analyzing big data, we will (i) develop a prediction model of an epidemic, (ii) establish a risk model of premonitory signs of a pandemic, and (iii) take advantages of early detection and forecasts in infectious disease control. Early detection and forecasting will be practicalized in our daily life.

    - Claire Nédellec (MaIAGE-INRA) : "Ontology-based text mining for microbiology research"

- 17h00 : **Break-Out Session**
- 18h00 : End

Wednesday, 11 July

- 9h30 : **Keynote** " Co-experience Knowledge and Wisdom with Human-Machine Harmonious Collaboration" - Norihiro Hagita, Board Director, Director, Intelligent Robotics and Communication Laboratories, Advanced Telecommunications Research Institute International

Abstract: This talk will introduce ongoing research projects on co-experience knowledge and wisdom with human-machine harmonious collaboration. These include 'situated services' with human-robot interaction and collaboration, wearable sensors, crowdsourcing and co-experience sharing, social concensus. It is anticipated that the resultant advances in the projects will lead to great strides in the intellectual activities of individuals and groups with ethical, legal, social and economic perspectives on harmonization with human society.

- 10h00 : **Keynote** "AI and HCI - Towards Human-Computer Partnerships" - Michel Beaudouin-Lafon, Professeur d'Informatique à l'Université Paris-Sud et membre senior de l'Institut Universitaire de France

Abstract: The classic approach to Artificial Intelligence treats the human being as a cog in the computer's process -- the so-called "human-in-the-loop". By contrast, the classic approach to Human-Computer Interaction seeks to create a 'user experience' with the computer. We seek a third approach, a true human-computer partnership that takes advantage of machine learning, but leaves the user in control. I describe how we can create interactive systems that are discoverable, appropriable and expressive, drawing from the principles of instrumental interaction and reciprocal co-adaptation. Our goal is to create robust interactive systems that grow with the user, with a focus on augmenting human capabilities.

- 10h30 : **Keynote** "Big Data Applications: Opportunities and Challenges" Yuzuru Tanaka 1) 2) ,Research Supervisor, JST CREST Program on "Big Data Applications", Professor Emeritus, Hokkaido University
  1) MaDIS, NIMS 2) Faculty of Engineering, Hokkaigakuen University

Abstract: This talk first focuses on the two potential application directions dealing with big data. They are urban-scale social CPSs (Cyber-Physical Systems) for secure, sustainable, and better social life, and data-driven sciences, i.e., a paradigm shift from "X" science to "X" informatics for varieties of "X". Then it shows an outline of JST CREST Program on "Big Data Applications" in which nine projects were carefully selected to form a good portfolio to cover challenging big data applications, and for each one to work as a flagship project of each different application domain. Through his involvement in varieties of big data application projects as a researcher or a supervisor, the speaker observed the heterogeneity of training data sets as a general characteristic of practical applications, which, if neglected, may lead us to conclude a single regression model for the mixture of subsets following different mathematical models. We need exploratory visual analytics to appropriately segment the heterogeneous training data set into homogeneous ones before the regression analysis of each of them. Even for a large training set, its segmentation may often make each homogeneous data set quite small. As a typical case, we focus on the discovery of new inorganic materials through machine learning, and propose a challenging future research direction.

- 11h00 : Coffee-Break
- 11h30 : **Oral Plenary Session**
    - Koichi Kise (Osaka Prefecture University) & Olivier Augereau (Osaka Prefecture University) - Kise Team (CREST Intelligent Information Processing) : "Quantified Learning by Deeply Sensing Learner's Behavior"

Abstract: Most of the problems we face in our life have already been experienced and solved by others. In this research, focusing on this point, we realize human-machine harmonious collaboration by recording experiences as digital data, store them in "experience bank" for their distribution. In this process, we transform the digital record of experiences into a user-adapted form called "experiential supplement" taking into account users' cognitive biases. This makes the experiences more acceptable by users to change their behavior. We demonstrate the effectiveness of the above process in the fields of learning, health care, sports and entertainment.

    - Laurence Devillers (LIMSI-CNRS - Institut DATAIA) : "Bad Nudge - Bad Robot ? : Nudge and Ethics in human-machine verbal interaction"
    - Yuji Matsumoto (Nara Institute of Science and Technology) & Ken Satoh (National Institute of Informatics) - Matsumoto Team (CREST Bigdata Applications) : "Knowledge Discovery through Structural Document Understanding"

Abstract: Through deepening the text and document analysis technologies necessary for document understanding, this project aims to develop foundations of content understanding of large scale technical documents, integration of acquired knowledge, and semantic similarity at structural levels of contents and documents. In collaboration with the experts in Biological Science, Material Science, Neuroscience, Law, and Artificial Intelligence, we will develop an integrated environment for content-based document retrieval and summarization, knowledge discovery and survey generation from large scale technical documents.

    - Pierre Zweigenbaum (LIMSI) : "Natural Language Processing for e-Health"

Abstract: From the patient record to scientific publications, natural language has an important position in electronic health (e-Health). Exploiting information and knowledge conveyed by natural language texts raises issues of Natural Language Processing. I will exemplify applications of Natural Language Processing to the analysis of text found in patient records, in medical Web sources, in the scientific biomedical literature, and in doctor-patient dialogues. I will then zoom in on methods for normalizing (also known as linking) medical concepts obtained from text, which are essential for interoperability.

    - Yutaka Yoshimoto, Deputy Director General (METI)
- 12h30 : Lunch Cocktail

- 14h00 : **Oral Plenary Session**
    - Albert Bifet (LTCI) : "Machine Learning for Data Streams"

Abstract: Advanced analysis of big data streams from sensors and devices is bound to become a key area of machine learning research as the number of applications requiring such processing increases. Dealing with the evolution over time of such data streams, i.e., with concepts that drift or change completely, is one of the core issues in stream mining. In this talk, I will present an overview of data stream mining, and I will introduce some popular open source tools for data stream mining.

    - Cédric Gouy-Pailler (CEA - Institut DATAIA) : "StreamOps : Open Source Platform for Research and Integration of Algorithms for Massive Time Series Flow Analysis"

Abstract : In the last few years, streaming platforms have become increasingly popular. This trend has been driven by requirements to quickly process never-ending flows of human-generated data or physical measurements, and supported by huge efforts in the open-source community (sometimes coupled with initiatives from web-oriented companies). Beyond the need to offer a cross-fertilized robust and scalable streaming platform to researchers, the current context poses new challenges for such a tool: in particular we will present the idea of a privacy-aware and accountable streaming platform, and develop related implications in terms of algorithms design.

    - Naoki Katoh (Kwansei Gakuin University) & Chako Takahashi (Tohoku University) - Katoh Team (CREST Bigdata Core technology) : "Foundations of Innovative Algorithms for Big Data"

Abstract: A lot of attention has been paid to "Big Data" from the beginning of this century. Because of huge data volume, theory of algorithm is now faced with a fundamental innovation. Under such circumstances, our project proposes a new computation paradigm called "Sublinear Time Paradigm". In order to realize such paradigm, we dvelop three fundamental technologies; sublinear time algorithm, sublinear data structure, and sublinear modeling. Integrating these three technologies, we establish the foundations of algorithms for big data.

    - Hossam Afifi (Télécom SudParis - Institut DATAIA), Jordi Badosa (Ecole polytechnique - Institut DATAIA) & Florence Ossart (CentraleSupélec - Institut DATAIA) : "PEPER: prediction of prosummers with reinforcement deep learning"

Abstract: Today, the world of electric power is facing major structural changes: the use of electricity is constantly increasing and the climate challenges impose an increase in the share of renewable energy (solar and wind). Since these energies are by nature intermittent, uncertain and distributed over the entire territory, the centralized electricity grid is evolving towards a decentralized structure, made up of subsets that combine production, storage and consumption at the local scale (microgrids, building scale). They should cooperate to cover as much as possible, the needs on a larger scale (city, region, country). Prosumers (consumption behavior adaptation according to the energy produced and available) is then a key point to ensure the balance of the network. The efficient and cooperative energy management of such a system is based on the prediction of the behavior of the various actors of the network (producers and consumers), the exchange of data between them (cooperation), and this at different time and cost scales.

PEPER project will contribute to this interdisciplinary challenge.

The project will gather relevant data from different actors of the network, and exploit the deep learning techniques to develop algorithms for forecasting the production and consumption of each actor, then provide solutions for the cooperation between them . These algorithms integrate data of different natures in the past and present: geographical positions, meteorological measurements, production and energy consumption profiles, dynamic mobility of the populations at each position. They then produce consumption predictions, recommendations on consumption-related adjustments or recommendations on the complementarity between different geographical zones according to their production and consumption profiles.

Technically, this poses several scientific obstacles: choice of types of traces, or sources of most relevant data, choice of learning algorithms, cooperation technique between these algorithms to provide the expected results. The results of these learning techniques will then be compared to those of the mathematical tools currently used by the project's research teams. The partners will publish the algorithms developed in the PEPER project and their results will be deployed on a real physical testbed consisting of several new equipped buildings in Polytechnique zone and its neighbourhood.

- Takemasa Miyoshi (RIKEN) - Miyoshi Team (CREST Bigdata Applications) : "Innovating "Big Data Assimilation" technology for revolutionizing very-short-range severe weather prediction"

Abstract: Data assimilation plays a key role in numerical weather prediction, combining computer models with real-world data through dynamical systems theory and statistical mathematics. Computing, sensing, and information/communication technologies are advancing rapidly, and data assimilation is becoming more popular as a means of cyber-physical fusion in broader science and technology fields. At RIKEN, the Japan's flagship research institute for all sciences, we have been pioneering the future possibilities of numerical weather prediction by taking advantage of the powerful K computer and Big Data from advanced sensing technologies such as the phased array weather radar and Himawari-8 geostationary satellite, at the scales ranging from global to convective. We developed innovative "Big Data Assimilation" (BDA) technology, enabling 30-second-update severe weather prediction at 100-m resolution, two orders of magnitude more rapidly than the currently operational numerical weather prediction systems. I will talk about some excitement of our BDA effort in numerical weather prediction, and a perspective toward data assimilation as a science hub – from severe weather forecasting and beyond.

- 15h30 : Coffee-Break

- 16h00 : **Oral Plenary Session**
    - Shuichi Onami (RIKEN) & Koji Koyamada (Kyoto University) - Onami Team (CREST Bigdata Applications) : "Data-driven analysis of the mechanism of animal development"

Abstract: This project develops data driven-methods to understand a full picture of the mechanisms of animal development by integrating the world largest spatiotemporal biological dynamics data of gene knockdown embryos, genomic information, and information of spatiotemporal expression and interaction of biological molecules such as mRNA, proteins and metabolites, and by combining advanced statistics and scientific visualization technologies. The project will establish the foundations of data-driven biology and transform biology into information science.

    - Vincent Fromion (MaIAGE, INRA, Université Paris-Saclay) : "Representation of biological data, information and knowledge: opportunities offered by systemic biology"

Abstract: High-throughput technologies produce huge amounts of heterogeneous biological data at all cellular levels. Structuring these data together with biological knowledge is a critical issue in biology and requires integrative tools and methods such as bio-ontologies to extract and share valuable information. In parallel, the development of recent whole-cell models using a systemic cell description opened alternatives for data integration. We recently proposed to integrate a systemic cell description within BiPON , a bio-ontology, to progress in whole-cell data integration and modeling synergistically. Altogether, BiPON opens up promising perspectives for knowledge integration and sharing by biologists, systems and computational biologists, and the emerging community of whole-cell modeling.

    - Masayuki Hirafuji (The Universiy of Tokyo) & Guo Wei (The Universiy of Tokyo) - Hirafuji Team (CREST Bigdata Applications) : "Knowledge Discovery by Constructing AgriBigData"

Abstract: Plants grow dynamically under changing environmental conditions. So time series data such as plant growth and soil moisture should be collected simultaneously for long term. Such time-series data can be analyzed as phenomena of complex systems, and the results enable objective evaluation and control in farm management and breeding. We will develop a method to construct agricultural big data (AgriBigData) automatically using our field monitoring technologies, that is, sensor networks and drones. Also we will develop methods on the AgriBigData to discover new knowledge, which can be applied for optimal farming and rapid breeding.

    - Hervé Monod (MIA department, INRA) : "#DigitAg, the French Digital Agriculture Convergence Lab"

Abstract: Many initiatives arise in France and Europe about digital agriculture. My talk will focus on #DigitAg, the French Digital Agriculture Convergence Lab which was born in 2017 in Montpellier (South of France) with two satellite sites in Toulouse and Rennes. It brings together 360 researchers and higher education teachers from leading French organizations. #DigitAg works on new digital tools and services to be transferred to the agricultural sector. It aims to develop Information and Communication Technologies in a farmer-friendly way to help agriculture to be competitive and sustainable.

- 17h00 : **Break-Out Session**
- 18h00 : End

Thursday, 12 July

- 9h30 : **Oral Plenary Session**
    - Christophe Prieur (Telecom ParisTech) : "HistorIA : Large historical databases. Data mining, exploration and explicability"

Abstract: The development of computational approaches to social sciences has stimulated new ambitious projects in historical research. However, the drastically new ways of dealing with historical sources come with high criticism and distrust from many historians who feel they might lose an essential contact with their work material.
Our project, HistorIA, gathers researchers from history, computational social sciences and information visualization. Its aim is to develop large data bases from historical sources with data mining analyses, along with iterative exploration tools, putting the main focus on the explicability of algorithms with visualization interfaces based on progressive analysis. The workflow will involve iterative steps of algorithm design by computer scientists and visual exploration by historians.

    - Takeaki Uno (National Institute of Informatics) & Kunihiro Wasa (National Institute of Informatics) - Uno Team (CREST Bigdata Core technology) : "Data Particlization for Next Generation Data Mining"

Abstract: This project studies ways to clarify small and middle size structures in big data so that complicated big data will become small understandable data. Data analysis such as machine learning becomes easy to design, visualization becomes clear, and the solutions will have connections to human understandings. Particlization discloses latent structures in customer data, economy data, and social networks, and support inovations in new businesses on information technologies through useful data analysis.

    - David Restrepo Amariles (HEC Paris) : "Smart Lawyer: Rating Legal Services in the Courtroom"

Abstract: On 11 May 2017, the French Cour de cassation handed down its decision n° 561 (16-13.669) on which it acknowledged the importance of online services and platforms offering comparative assessments of lawyers and law firms, including through rankings and ratings, for the protection of consumers of legal services. However, the Court also affirmed that such services must ensure a certain level of quality: "[il] leur appartient [...], dans leurs activités propres, de délivrer au consommateur une information loyale, claire et transparente". Indeed, the Court held that providers of ratings, which are designed to inform the behaviour and decisions of consumers, should ensure that such information is loyal, clear and transparent. Unfortunately, to date, providers of these services rely mostly on anecdotal evidence as well as on perception and self-reported data. The lack of reliable information about the quality of legal services delivered by lawyers in the courtroom is a worrying and widespread phenomenon in all jurisdictions across the European Union, but also in other jurisdictions such as the United States and Canada. This project aims to fill this gap by combining legal expertise with data science research. It seeks to develop a meaningful and reliable measurement tools of legal performance in the courtroom that can help

improving access to justice and the quality of legal services, while also helping law firms assess the performance of lawyers and the quality of jurisdictions.

- 10h30 : Coffee-Break

- 11h00 : **Oral Plenary Session**
    - Shunichi Koshimura (Tohoku University) - Koshimura Team (CREST Bigdata Applications) : "Establishing the most advanced disaster reduction management system by fusion of real-time disaster simulation and big data assimilation"

Abstract: Bringing together state-of-the-art disaster science, high-performance computing, information and mathematical sciences, our team will tackle the significant research challenges in disaster response and management issues which remain, particularly in the areas of "Integration of many data sources", "Interpretation/Veracity of data content", "Real-time big data assimilation and Mapping". This research project aims to establish the world's most advanced fusion of high-performance real-time disaster simulation and big data driven disaster management system to enhance ability of societies and social systems to respond promptly, sensibly and effectively to natural disasters, withstanding adversities and exploiting lessons in the future catastrophic disaster and crisis management.

    - Gregory Blanc & Mustafizur Shahid (Télécom SudParis) : "Machine Learning Based Intrusion Detection System for IoT Network"

Abstract: Internet has grown into a deeply complex and diverse network, accommodating a large number of devices and protocols, which is accelerated by the progressive introduction of so-called smart objects, with applications in domains such as homes, hospitals or transportation. Such domains then suffer from common weaknesses found in legacy computer networks, as evidenced by recent attacks that compromise a large number of objects over the Internet and turn them against designated targets, harming the availability of some popular Internet services. The wealth of data generated by these objects is both a blessing and a curse. In fact, monitoring complex networks for detecting malicious behaviors can be heavily clouded by the huge amount of traffic, making it costly to perform. On the other hand, one can leverage these data to detect common behaviors exhibited by legitimate traffic, as well as malware-generated one. Machine learning is hence widely used in the cybersecurity community for many different purposes including network intrusion detection, malware analysis and adversarial learning. In this presentation, we will introduce past and current research works carried out in the field of network intrusion detection, with application to the detection of botnets, large networks of compromised machines, in both the population of computers, and the population of smart objects. In the latter work, we first explain how appropriate ML is to extract characteristic patterns in the so-called Internet of Things, and the preliminary results of our work to detect anomalies in smart home networks.

    - Masaharu Munetomo (Hokkaido university) - Aida Team (CREST Bigdata Core technology) : "Optimal Resource Selection in Application-Centric Overlay Cloud Utilizing Inter-Cloud"

Abstract: In this project, we significantly improve the performance of big data analysis by developing a new infrastructure technology on inter-cloud. The developed technology enables to automatically and quickly build large-scale data analysis platforms that are optimized for user applications. Our goal is to build data analysis platforms and run applications, genome sequencing and fluid acoustic simulation, within days/weeks by using our technology, while it takes months in the current technology.

- 11h45 : **Wrap-up Discussion-** Nozha Boujemaa & Masaru Kitsuregawa
- 12h45 : **Closing of the event**