http://www.univ-tln.fr/13es-Journees-Scientifiques-de-l-Universite-de-Toulon.html / Porte d'Italie, Toulon

PÔLE THÉMATIQUE INTERDISCIPLINAIRE INPS Information, Numérique, Prévention, Santé

UNIVERSITÉ DE TOULON

Data Sciences Symposium

7th & 8th March, University of Toulon
City Center - Free Entrance

Prog. com. : H. Glotin & A. Paiement, LIS - DYNI - PSD - CNRS - U. Toulon
V. Kornilov, M. Komarov & S. Maltseva, HSE Moscow

7th March, amphi FA001

11h00	Présentation du Pôle INPS : *Hervé Glotin*, Université de Toulon
11h10	Conférence invitée INPS : Information Spaces for Big Data Problems: Transforming Sequential Processing into Parallel *Peter Golubtsov*, Dr., Pr., Dept of Mathematics, Faculty of Physics, Lomonosov Moscow State University
11h40	Table ronde : des Pôles interdisciplinaires de recherche à l'université de Toulon, pourquoi et avec qui ?
12h00	Pause déjeuner - salle CO.315 - 3e étage du plot Coudon

7 Jeudi mars 2019

co-org. *H. Glotin*, *A. Paiement*, *V. Kornilov*, *M. Komarov*, *S. Maltseva*, RU-FR BigData

Session INPS - Amphithéâtre FA.110

Method for Big data

13:15 DataMall: a new platform for trading with big data
Vitalii Veliamidov, Engineer, Nat. Center of Cognitive Tech., ITMO University, St Petersbourg, RU

13:35 All-Russian Big Data competition among students
Irina Deeva, Engineer, Nat. Center of Cognitive Tech., ITMO University, St Petersbourg, RU

13:55 High Performance Computing for mono channel source localisation
Julie Patris, Dr, Pr., Dept of Physics, CNRS, LIS, Aix Marseille University

14:15 Long term survey of the Fukushima Nuclear Exclusion Zone : Systems & challenges
Daisuke Shimotoku, PhD Student, Faculty of Computer Science, Univ. Tokyo, JP

14:30 The pain and gain of maintaining the ML models zoo
Natalia Khapaeva, Senior Architect, Mobile TeleSystems PJSC, RU

Learning from Big data

14:40 Big Data in climate: models fitting, analysis & storage of models' results
Nikolay Nikitin, Engineer, Nat. Center of Cognitive Tech., ITMO University, St Petersbourg, RU

15:00 Logic-based boundaries of some Modular Neural Networks for Big Data
Denis Fedyanin, Research Fellow, Int. Lab. for Logic, Linguistics & Formal Philosophy, Nat. Research Univ. Higher Sch. of Economics, Moscow, RU

15:20 Large scale marine mammal monitoring: Learning Orca calls
Hervé Glotin, Dr., Pr., Dept of Computer Science, CNRS, UTLN, Toulon, FR

15:35 On compression, learning & searching regularity in big data
Tatiana Makhalova, PhD Student, Faculty of Computer Science, Nat. Research Univ. Higher Sch. of Economics, Moscow, RU

Deep Learning

16:00 Manifold learning from Stereo Auto-Encoder
Maxence Ferrari, PhD Student, Faculty of Computer Science, CNRS LAMFA LIS, FR

16:20 Convolutional neural networks for automated search of anomalies in metocean fields
Pavel Vychuzhanin, Engineer, Nat.Center of Cognitive Tech., ITMO University, St Petersbourg, RU

16:45 Time series in Astrophysics & Deep Learning
Adeline Paiement, Dr., Pr., Dept of Computer Science, CNRS, UTLN, Toulon, FR

8 Vendredi mars 2019

co-org. *H. Glotin*, *A. Paiement*, *V. Kornilov*, *M. Komarov*, *S. Maltseva*, RU-FR BigData

Session INPS - Amphithéâtre FA.110

Deep Learning

9:00 Bioacoustics & Deep Learning
Jan Schlüter, Dr., Researcher, LIS Lab & OFAI, CNRS, UTLN, Toulon, FR & Vienna, Austria

Forecasting

9:45 Decisions for Large Retail Network - Machine learning for tornado prediction
Fuad Aleskerov, Dr., Pr, Head of Dept of Mathematics, Faculty of Economic Sciences, Nat. Research Univ. Higher Sch. of Economics, Moscow, RU

10:30 Deep learning methods for time series forecasting
Petr Gladilin, Researcher, eScience Research Institute, ITMO University, St Petersbourg, RU

10:50 Forecasting purchase categories by transactional data: a comparative study of classification methods
Egor Shikov, Engineer, Nat. Center of Cognitive Tech., ITMO University, St Petersbourg, RU

11:20 On the Ethics of Big Data
Olga Tsukanova, Dr Pr HSE Moscow, with Tatyana Telysheva

11:45 Discussion
Svetlana Maltseva, Dr., Pr., Head of Sch. of Business Informatics, Fac. of Business & Managt., Nat. Research Univ. Higher Sch. of Economics, Moscow, RU

12:30 Pause déjeuner - salle CO.315 - 3e étage du plot Coudon

14:05 End to End SpliNet Learning
Randall Balestriero, PhD Student, Faculty of Computer Science, Rice Univ., Houston, USA

14:45 City-scale event prediction using convolutional neural networks & adaptive geogrids
Ksenia Mukhina, Engineer, Nat.Center of Cognitive Tech., ITMO University, St Petersbourg, RU

15:15 Discussion

16:00 End

ABSTRACTS

Fuad Aleskerov	Dr., prof., Head of Department of Mathematics, Faculty of Economic Sciences	National Research University Higher School of Economics	**Decisions for Large Retail Network**	
We study large retail network with 1.5 mln customers, more than 400 thousand goods, and data collected for more than one year for each customer. The problem was to find a segmentation of customers on the basis of their basket consumption to evaluate up-sell potential as well as find churning customers. New methods to solve these problems as well as general view on the problems are discussed. The results were applied for one of the European countries and large European retail network.			https://www.hse.ru/en/org/persons/140159 Professor, Member of Academia Europaea, Honorary Worker of Science and Education of Russia Member of Editorial Board of 14 journals 10 books, more than 200 articles, more than 100 in peer-reviewed journals and volumes Copyright certificates, patents - 6 Invited speaker for more than 100 conferences and workshops	
Denis Fedyanin	Junior Research Fellow	V.A.Trapeznikov Institute of Control Sciences	**Logic-based boundaries of some Modular Neural Networks for BigData**	
Neural Networks (NN) could be divided into several levels. Sometimes it is possible to hold the weights of lowest levels and teach only the highest levels when we want to apply the NN to another problems. A good example is recognition of faces, where one can take predefined NN and adjust it to specific group of people. Such division a large NN into smaller Networks could be very useful for BigData. We can host a unique NN for each big part of BigData. And then we can just integrate results of these local NN into one result by special NN without teaching NN for entire dataset for the very beginning each time. These distributed NN has an ability to replace local NN when we need it. It gives us high flexibility. Such architecture has some limits and properties. We will focus on its boundaries based on using some recursive methods and show connections to dynamic epistemic logic and Kripke semantics.			Denis Fedyanin currently works at the Institute of Control Sciences. Denis does research in Logic and Foundations of Mathematics, Applied Mathematics and Game Theory.	
Peter Golubtsov	Dr., prof. Department of Mathematics, Faculty of Physics	Lomonosov Moscow State University	**Information Spaces for Big Data Problems: Transforming Sequential Processing into Parallel**	
The procedures of sequential updating of information is important for “big data streams” processing because it avoids accumulating and storing large data sets. It is shown that processing can be simplified by introducing a special intermediate form of information representation. Thanks to the rich algebraic properties of the corresponding information space, such approach allows unifying and increasing the efficiency of the information update. It also leads to various parallelization options for the inherently sequential Bayesian procedure, which are ideally suited for distributed data processing platforms			Doctor of Science, Professor of National Research University Higher School of Economics and Lomonosov Moscow State University. Main areas of research: mathematical foundations of information theory, information processes in Big Data systems, information effects in dynamic stochastic games with applications to economics and optimal management of natural resources.	

Natalia Khapaeva	Senior Architect	Mobile TeleSystems PJSC	**The pain and gain of maintaining the ML models zoo**	75 1438311
When we talk about the Data Science models, the most recent topics are algorithm selection, feature engineering, etc. Thus, the «fog of war» covers almost half of the model lifecycle. However, deploying, monitoring and quality assessment stages of the model lifecycle can be the critical bottleneck of the whole project. This talk covers sustainable deployment and monitoring in a web production environment.			https://www.linkedin.com/in/nmkh/ Senior Big Data Architect at MTS Group, Data Governance, Data Quality and DS Products Operation projects.	
Svetlana Maltseva	Dr., prof., Head of School of Business Informatics, Faculty of Business and Management	National Research University Higher School of Economics		
			https://www.hse.ru/en/org/persons/67823	
Tatiana Makhalova	PhD Student, Faculty of Computer Science	National Research University Higher School of Economics	**On compression, learning and searching regularity in data**	
Data Science (DS) is an umbrella term that encompasses a big variety of methods, processes, and algorithms aimed at getting insight from data. The “insight” is understood differently and results in different approaches having, at first glance, completely different objectives. More than that, being developed independently the fields of DS are constantly moving apart. It seems unlikely that well-adapted to solve distinct tasks, the methods from one field can contribute to the development of the other fields. We consider the most common fields in DS, namely Knowledge Discovery, Machine Learning and Data Compression, and show the interconnection between them using the formalism of Formal Concept Analysis. Our goal is to draw a clear interconnection between distinct tasks and to show how the best practices (the objectives) of one field can contribute to others.			PhD Student, Faculty of Computer Science, National Research University Higher School of Economics	
Petr Gladilin	Researcher, eScience Research Institute	ITMO University	**Deep learning methods for time series forecasting**	
The application of machine learning to time series forecasting problems has achieved accurate prediction results in lots of case studies. This study presents a review of modeling approaches for time series forecasting and comparison of predictive accuracy of different models. Also we propose a method for fine tuning of recurrent neural networks parameters such as history lag and corresponding predictive horizon for a better performance on a given time series depending on its properties.			Peter works at the eScience Research Institute of ITMO University, Saint Petersburg. Areas of scientific interests: machine learning, computer vision, time-series forecasting, generative neural networks and deep-art.	
Egor Shikov	Engineer, National Center of Cognitive Technologies	ITMO University	**Forecasting purchase categories by transactional data: a comparative study of classification methods**	
Forecasting purchase behavior of bank clients allows for development of new recommendation and personalization strategies and results in better Quality-of-Service and customer experience. In this study, we consider the problem of predicting purchase categories of a client for the next time period by the historical transactional data. We study the predictability of expenses for different Merchant Category			Egor Shikov graduated from St.Petersburg Polytechnic University with a master's degree in applied nuclear physics. Engineer of National Center for Cognitive Technologies, works in "Algorithmics of Complex Systems" research group where he focuses on predictive modeling in finance with neural networks.	

Codes (MCCs) and compare the efficiency of different classes of machine learning models including boosting algorithms, long-short term memory networks and convolutional networks.				
Ksenia Mukhina	Engineer, National Center of Cognitive Technologies	ITMO University	**City-scale event prediction using convolutional neural networks and adaptive geogrids**	
Active development of modern cities requires not only efficient monitoring systems but furthermore forecasting systems that can predict future state of the urban area with high accuracy. In this work, we present a method for urban area prediction based on geospatial activity of users on Instagram. We propose three different deep learning architectures that are able to solve a target problem and show that convolutional neural network based on three-dimensional convolution layers provides the best results with an accuracy of 99%.			Ksenia Mukhina is a Lecturer at High-Performance Computing Department and a Research Assistant at eScience Research Institute in ITMO University. Ksenia works in "GSS infrastructure" research team where she focuses on social networks analysis. She is an author of more than 20 scientific papers.	
Vitalii Veliamidov	Engineer, National Center of Cognitive Technologies	ITMO University	**DataMall: a new platform for trading with big data**	
Complex and interconnected processes in modern business world leads to emergence of those who has data but don't know how to or can not extract value for himself and those who wants to improve his own business or knows how to help others but lacks appropriate data. DataMall is a new platform dedicated to bring them all together and solve the problem of data access and data selling. Its design has been developed with 4 main aspects in mind: ● safety of data with adjustable degree of control for their owner; ● flexible access to elastic computing power provided by clouds; ● different monetization policies of data access - wholesales for large business and "pay-as-you-go" policy for small and medium business; ● easy-to-use web hub that serves as a showcase for customers to help find desired data in desired quantities and buy it. The talk contains overview of the platform and its general architecture along with examples of use-case scenarios			Vitalii Veliamidov has graduated from St. Petersburg State University with M.D. (dep. of Information and Analytical Systems). His professional interests include distributed data storage and processing systems and now he is working on the DataMoll platform at National Center of Cognitive Technologies, ITMO University	
Nikolay Nikitin	Engineer, National Center of Cognitive Technologies	ITMO University	**Big Data in climate modelling: models fitting, analysis and storraging of models' results**	
The Big Data concept becomes extremely important for metocean data analysis and storage. The experience of semantic data storage's development in a frame of long-term Arctic ocean simulation project is presented. Several data engineering solutions that allow providing real-time data processing are concerned. Also, the TerraXT software for the dynamic environmental data visualisation is described.			Nikolay Nikitin is an Engineer of National Center for Cognitive Technologies, Teaching assistant and PhD Student at High-Performance Computing Department of ITMO University. Nikolay works in "Natural Systems Simulation" research group where he focuses on models identification and calibration.	

Pavel Vychuzhanin	Engineer, National Center of Cognitive Technologies	ITMO University	**CNN for automated search of anomalies in metocean fields**	
The presentation proposes to consider the problem of detecting anomalies in the results of ocean modeling as a pattern recognition tasks. As practical examples, it will be shown how the proposed methods can be implemented for fields with different nature of anomalies, in particular, for fields of sea currents and ice concentration in the Arctic region. The advantage of the presented approach is the ability to validate fields in a fully automatic mode, without expert supervision, even at the stage of labeling.			Pavel Vychuzhanin is an Engineer of National Center for Cognitive Technologies, Master's Student at High-Performance Computing Department of ITMO University. Pavel works in "Natural Systems Simulation" research group where he focuses on data-driven methods in geophysical modeling.	
Irina Deeva	Engineer, National Center of Cognitive Technologies	ITMO University	**All-Russian Big Data competition among students: results and experience**	
Competitive data analysis as a separate unit in the world of machine learning and big data. The experience of the All-Russian competitions in Big Data Analytics. How to cover machine learning and computing infrastructure. Results, statistics and trends among students in the field of big data.			Irina Deeva is an Engineer of National Center for Cognitive Technologies and PhD Student at High-Performance Computing Department of ITMO University. Irina works in "Natural Systems Simulation" research group where she focuses on multiscale and surrogate modeling	
Jan Schlüter	Postdoctoral research fellow	LIS lab, University of Toulon	**Deep Learning for Acoustic Bird Detection and Species Identification**	
Biodiversity, the variety of life on Earth, is constantly reduced by human influences. Monitoring this loss is vital to understand and control it. We explored how well current methods of artificial intelligence may help to monitor biodiversity from audio recordings, supporting or complementing human observations. Using artificial neural networks, we were able to automatically detect bird calls in field recordings, as well as identify the species to some extent, reaching top results in two scientific competitions. I will present the technical solutions in detail, discuss shortcomings, and finish with an outlook on applying the same methodology to underwater recordings of whales.				
Julie Patris	Dr., Pr.	Aix-Marseille University, LIS lab	**High performance computing for mono channel source localization**	
A key-parameter for whale density estimation through passive acoustics is recovery of animal position (see Marques et al. 2013). Most commonly, the localization process requires an array of 4 synchronized hydrophones (Kuperman et al. 2004), which is a more technically complex and expensive option than installing one single sensor. Several attempts have been made towards recovering the position of the singer with only one hydrophone (McDonald et al. 1999 , Kuperman et al. 2001, Bonnel et al. 2014), however, these methods usually permit only range (and sometimes depth) estimation, and are usually possible only when multiple arrivals of a short sound are separable. Here, we use acoustic modeling methods (SPECFEM, Tromp et al. 2008 https://geodynamics.org/cig/software/specfem3d/) to understand the propagation of whale			Dept. Physics, Aix Marseille University	

vocalizations in a specific context. Thanks to this precise time-domain modelling, we explore all the information inside the signal, including modifications produced by environmental complexity (bathymetry, speed variation, etc.). Subsequently we applied a low-cost inversion method based on Green's functions reciprocity principle to reconstruct the whale's position. First tests on simulated data confirm that the technique is well suited to biological signals of very low frequency, such as the Southeast Pacific 2 blue whale song type (described in Buchan et al. 2014). We also show the possibilities of this method for determining the range of an animal in shallow coastal waters. The results of two simulations in a 3D 10 km-wide box lead to the recovery of the range (with a precision greater than 100 m) in 98 % of cases, of the depth in 100 % of cases and of full position (that is range, depth and azimuth, thanks to our 3D modelling, with a resolution greater than 500 m) in 43% of cases. We also present the first results of this method applied to Blue Whales in Northern Chile, where a coupled visual/acoustic study was performed in austral summer 2017.				
Randall Balestriero	PhD candidate	Rice University	**End to end SpliNet learning**	
We propose to tackle the problem of end-to-end learning for raw waveform signals by introducing learnable continuous time-frequency atoms. The derivation of these filters is achieved by defining a functional space with a given smoothness order and boundary conditions. From this space, we derive the parametric analytical filters. Their differentiability property allows gradient-based optimization. As such, one can utilize any Deep Neural Network (DNN) with these filters. This enables us to tackle in a front-end fashion a large scale bird detection task based on the freefield1010 dataset known to contain key challenges, such as the dimensionality of the inputs data (>100,000) and the presence of additional noises: multiple sources and soundscapes.			Computer Science Dept., Rice University, Houston, USA	
Adeline Paiement	Dr., Dept. Computer Science	University of Toulon, LIS lab	**Deep learning for astronomy image analysis**	
This talk will present our recent and ongoing work on astronomy image analysis. We will examine the requirements they impose on the design of new computer vision based analysis methods. We will particularly consider deep learning methods, which have been so far predominantly developed for natural images, and we will illustrate the potential for this paradigm in enhancing the interpretation and exploitation of astronomy images.			Computer Science Dept., Université de Toulon, LIS lab	