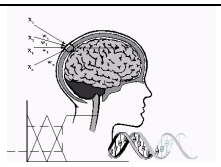
International
Innovation in Knowledge Based and Intelligent Engineering Systems



INVITED SESSION SUMMARY

Title of Session:
Novel information levels in scientific discovery: sharing, refining, and fusion of complex results in federated architectures of deep learning and causal analysis

Name, Title and Affiliation of Chair:

Chair: Péter Antal, PhD, associate professor at the Department of Measurement and Information Systems, Budapest University of Technology and Economics

Co-chair: Gábor Hullám, PhD, associate professor at the Department of Measurement and Information Systems, Budapest University of Technology and Economics

Co-chair: Bence Bolgár, PhD, post-doc researcher at the Department of Measurement and Information Systems, Budapest University of Technology and Economics

Details of Session (including aim and scope):

Aim of the session:
Sharing experimental data and expert knowledge are central pillars of modern science, especially in life sciences. However, the multi-stage and multi-party nature of modern scientific research, data heterogeneity, personal privacy, and institutional confidentiality require the sharing of intermediate information between data and knowledge, especially computational results of statistical analysis. Particular examples are the wide range of summary statistics in genetic association analysis, "edge posteriors" from systems biology, which are frequently reported as approximate results in causal discovery, and shared neural network (sub)models in multi-party and federated learning. The characterization of such distinct information levels, tentatively called "data-analytic levels" is challenging from an epistemological point of view. However, recently emerged information levels have the following properties: (1) they are intended for computational integration and not for a direct human interpretation, (2) single information items can be highly uncertain and weakly or even non-significant, nonetheless (3) the set of information items corresponding to a given abstraction level is well defined, and at least in principle, can be fully reported. These ephemeral results are frequently wasted, despite their considerable costs in terms of data, knowledge, and computation. The admittance and explicit management of such intermediate information levels are also in line with the blurring boundaries between databases and knowledge bases. Additionally, such novel information levels provide further ingredients for the support and automation of science by recycling an increasingly wider range of complex results. The Bayesian framework offers a principled formalism for this online meta-learning or batch-learning: sharing "posteriors" and incorporating "posterior priors" in the sequential process of the scientific discovery.

The aim of the special session is to investigate such novel levels of shared information in multi-party and federated scientific discovery, especially in deep neural network learning and causal analysis.

Topics of interest include (but are not limited to):
- Practical application of novel intermediate information levels in multi-party approaches, such as reporting posteriors of multivariate causal relations beyond edge posteriors, effect sizes for interactions beyond main effects, and Bayesian neural networks beyond single structures with real-valued parameters.
- Theoretical models of large-scale multi-party scientific infrastructures using statistical meta-analysis, sequential/online learning, and federated learning. How can we adopt the concept of sufficient statistics to guide the design of shared information levels, especially regarding sample size and data quality?
- Novel machine learning and discovery methods using data-analytic information levels. Novel technologies, e.g. probabilistic extensions of semantic technologies, to support the management of data-analytic

information levels.

- How can we compress the volume of shared information, while also preserving negative results? What are the optimal trade-offs between complexity and performance improvement in long-term learning? Can we select the optimal level of complexity for a given threshold of uncertainty? Consider for example the complexity of preserved structural network features, the quantized real-valued parameters of deep neural networks, and the application of probabilistic graphical models with varying complexity to represent the uncertainty of information items.
- What are the quantitative advantages and disadvantages of using intermediary data-analytic information levels compared to learning using the unification of data sets, and to mixture of expert integration of models from each data set?

Contributions describing work in progress, as well as position papers, are invited. Of particular interest are papers that quantitively evaluate the effect of using an intermediary data-analytic information level compared to full data integration before learning and integration of models after learning.

The session will be of interest to researchers working in artificial intelligence, machine learning, multi-party and federated learning, discovery systems, and bioinformatics.

Deadlines: To be published later.
Tentative schedule:
Submission of papers: 3 May, 2019
Notification of acceptance: 20 May, 2019
Receipt of publication files: 1 June, 2019

Main Contributing Researchers / Research Centres (tentative, if known at this stage):s

To submit papers, we will invite research centers (among others) from Hungary, Belgium, Switzerland, France, Great-Britain, Germany, US.

Website URL of Call for Papers (if any):

To be published later if the proposal is accepted.

Email & Contact Details:

Péter Antal, antal@mit.bme.hu, Budapest University of Technology and Economics, Budapest, Hungary, http://www.mit.bme.hu/eng/general/staff/antal

Short CVs of key persons:

Péter Antal received his M.Sc. degree in Computer Science (Informatics Engineer) in 1995 from the Faculty of Electrical Engineering and Informatics, Technical University of Budapest (BME). Between 1998-2002 he was an international scholar and Ph.D student at the Department of Electrical Engineering, Katholieke Universiteit Leuven, where he received his Ph.D. Between 2002-2007 he was a research assistant, later an assistant professor at the Department of Measurement and Information Systems (MIS) at the BME. Between 2008-2011 he was a postdoctoral researcher at OTKA (Hungarian Scientific Research Fund). From 2011 he is an associate professor at MIS and the head of the Department's Computational Biomedicine Laboratory (ComBine Lab) and the Artificial Intelligence Group. His research interests include artificial intelligence, machine learning, Bayesian approaches, causality research, biostatistics, chemo- and bioinformatics.

Gábor Hullám received the M.Sc. degree in Computer Science from the Budapest University of Technology and Economics (BME) in 2005, and earned the Ph.D. degree from BME in 2016. In 2008, he joined the Department of Measurement and Information Systems at BME, where he currently holds a position as an associate professor. He is a member of the Laboratory of Computational Biomedicine and Bioinformatics (ComBine Lab). He has done research on probabilistic modeling, Bayesian statistical methods for data analysis (Bayesian relevance and effect size measures, particularly for the analysis of genetic association studies), causal modelling and discovery. Since 2013 he is also a member of MTA-SE Neuropsychopharmacology and Neurochemistry Research group,

where he participated in depression research investigating the genetic background and corresponding environmental factors of various depression phenotypes. Recently, he has been involved in a depression-related study based on the analysis of large-scale health data (UKBiobank).

Bence Bolgár received his MD degree in 2012 from the Semmelweis University, Budapest, Hungary. He received his Ph.D at the Department of Measurement and Information Systems, Budapest University of Technology and Economics, where currently he is a post-doc researcher. His research focuses on Bayesian methods and kernel machines in massively parallel computing environments, with applications in bio- and chemoinformatics.