EU Projects Forum

Collection of abstracts

Editor: Petra Kralj Novak

Monday, September 18th, 2017

ECML PKDD 2017

SKOPJE, MACEDONIA
18-22 SEPTEMBER

THE EUROPEAN CONFERENCE ON MACHINE LEARNING &
PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN
DATABASES

SKOPJE 2017

Preface

The EU Projects Forum at ECML PKDD 2017 is a novel initiative that encourages the dissemination of EU projects and their results to the targeted scientific audience of conference participants. It provides an opportunity for the EU funded projects to present their vision/work to the conference audience, and an opportunity for the ECML PKDD 2017 audience to learn about the European scientific success stories in their research field. As a satellite event of ECML PKDD 2017, it is envisioned as a machine learning/data mining/big data/data science venue, gathering ERC grantees, EU project consortia, EU project officers, with the ECML PKDD audience including the interested researchers and industrial participants.

- *ERC grantees in the machine learning/data mining field were invited to present their projects and results at the EU Projects Forum at ECML PKDD 2017. The European Research Council (ERC) was established ten years ago by the EU to fund excellent scientists and their most creative ideas. Nearly 7,000 top researchers have been supported, including six who later received Nobel Prizes. ERC grants also created career opportunities for some 50,000 researchers, resulted in numerous scientific breakthroughs and led to over 800 patent applications that lay the foundations for growth and jobs.*
- *Also Horizon 2020 EU projects related to machine learning/data mining were invited to present their vision, disseminate research results, and potentially recruit researchers. Horizon 2020 is the biggest EU Research and Innovation program ever with nearly €80 billion of funding available over 7 years (2014 to 2020) – in addition to the private investment that this money will attract. It promises more breakthroughs, discoveries and world-firsts by taking great ideas from the lab to the market.*

In addition to the ERC grants and H2020 project presentations, this year's EU Projects Forum edition features two invited speakers: Salvatore Spinello – Research Programme Officer at the Research Executive Agency (REA), and Richard Wheeler – private scientific consultant from Edinburgh Scientific.

Welcome to the EU Projects Forum at ECML PKDD 2017 and enjoy a day of interesting presentations, fruitful discussions, insightful discoveries and informal consultations in the Forum participants company.

EC Forum organizers: Petra Kralj Novak and Nada Lavrač

In Ljubljana, August 2017

Schedule for Monday, September 18th, 2017

9:00	**Nada Lavrač, Petra Kralj Novak**	
		Opening
9:10	**Salvatore Spinello**	
	Invited talk	FET Open - main features and evaluation process
10:00	**Tijl de Bie, Jefrey Lijffijt**	
	ERC grant	FORSIED: Formalizing Subjective Interestingness in Exploratory Data Mining
10:30	***Coffee & Tea Break***	
11:00	**Fosca Giannotti**	
	H2020 project	SoBigData Research Infrastructure
11:20	**Hussain Kazmi**	
	H2020 project	REnnovates: Towards Net-Zero Energy Communities and Beyond
11:40	**Martin Žnidaršič**	
	H2020 project	CF-Web: ClowdFlows Data and Text Analytics Marketplace on the Web
12:00	**Brian Mac Namee**	
	H2020 project	AURORA: Advanced User-Centric Efficiency Metrics for Air Traffic Performance Analytics
12:20	**Sašo Džeroski**	
	H2020 project	MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data
12:40	***Lunch Break***	
14:00	**Vincenzo Lagani, Ioannis Tsamardinos**	
	ERC grant	CAUSALPATH: Next Generation Causal Analysis Inspired by the Induction of Biological Pathways from Cytometry Data
14:30	**Richard Wheeler**	
	Invited talk	Hints on how to write a successful project proposals (Part I)
15:40	***Coffee & Tea Break***	
16:00	**Richard Wheeler**	
	Invited talk	Hints on how to write a successful project proposals (Part II)
17:00		Late breaking project presentations
17:30		Closing remarks

Table of Contents

Salvatore Spinello – invited speaker

Research Programme Officer at Research Executive Agency (REA)

FET Open: Main Features and Evaluation Process

FET Open is one of the most attractive research programme under Horizon 2020. FET-Open supports the early-stage, high-risk research around new ideas towards radically new future technologies. It explores an open range of new and disruptive technological possibilities in all areas of Science & Technology, inspired by cutting edge science, unconventional collaborations and pioneering new ways to create the optimum conditions for serendipity to occur.

In these first 3 years of Horizon 2020, a total of 2.648 proposals were submitted to the FET-Open programme and covered a wide range of disciplines: from Physics to Life Sciences, from Information Sciences and Engineering to Chemistry. Most proposals show indeed high degree of interdisciplinarity.

During my presentation I will focus on Research and Innovation-Actions (RIA) and the so-called "gatekeepers" that every excellent proposal should address. I will then present the evaluation process that allows the selection of the best proposals, resulting in a continuously growing portfolio of high quality interdisciplinary projects. I will conclude providing some statistics in terms of country and organization participations, scientific fields covered and interdisciplinarity.

Salvatore Spinello's profile:

Salvatore Spinello started his Ph.D. in Computer Science in 1997 at the University of Catania (Italy). He moved to Germany in 1999 where I finished his Ph.D. in collaboration with the University of Erlangen-Nuremberg. He then moved to London for his first Post-Doc (UCL) and after one year to Bordeaux (France) for his second Post-Doc (Inria).

In 2004 he joined a small company, the French leader in the distribution of Virtual Reality's products. He held the position of Director of the R&D Department managing two European Projects.

In 2006 he joined Inria, the French Institute for Research in Computer Science and Control. He was in charge of identifying knowledge and technologies from research teams which were transferable to the external world (both industry and academic partners), enabling the transfer typically through licensing or joint R&D projects, protecting whenever appropriate the underlying intellectual property. He was also in charge of activities which aimed to facilitate the participation of researchers to National and European collaborative projects.

In 2013 he took the position of Project Officer at the Aquitaine Regional Council (France). He negotiated grant agreements; he monitored his portfolio from administrative, financial and technical aspects; he assessed technological progress and the fulfilment of contractual obligations; he managed the correct use of resources allocated to the projects, ensuring that the work was been carried out as planned; he monitored the overall performance (technical, dissemination, exploitation) and the strategic impact of projects.

In 2015 he joined the Research Executive Agency (REA) as Research Programme Officer where he is participating to the evaluation and selection of proposals submitted to the FET Open Programme and monitoring several funded projects.

Tijl de Bie and Jefrey Lijffijt
Universiteit Gent, Belgium

FORSIED: Formalizing Subjective Interestingness in Exploratory Data Mining

Link:
http://cordis.europa.eu/project/rcn/185593_en.html

Funding scheme:
ERC-CG - ERC Consolidator Grants

Abstract:
The rate at which research labs, enterprises and governments accumulate data is high and fast increasing. Often, these data are collected for no specific purpose, or they turn out to be useful for unanticipated purposes: Companies constantly look for new ways to monetize their customer databases; Governments mine various databases to detect tax fraud; Security agencies mine and cross-associate numerous heterogeneous information streams from publicly accessible and classified databases to understand and detect security threats. The objective in such Exploratory Data Mining (EDM) tasks is typically ill-defined, i.e. it is unclear how to formalize how interesting a pattern extracted from the data is. As a result, EDM is often a slow process of trial and error.

In this project we are developing the mathematical principles of what makes a pattern interesting in a very subjective sense. Crucial in this endeavour is research into automatic mechanisms to model and duly consider the prior beliefs and expectations of the user for whom the EDM patterns are intended, thus relieving the users of the complex task to attempt to formalize themselves what makes a pattern interesting to them.

The results of this project may radically change the way in which EDM research is done. Currently, researchers typically imagine a specific purpose for the patterns, try to formalize interestingness of such patterns given that purpose, and design an algorithm to mine them. However, given the variety of users, this strategy has led to a multitude of algorithms. As a result, users need to be data mining experts to understand which algorithm applies to their situation. To resolve this, we are developing a theoretically solid framework for the design of EDM systems that model the user's beliefs and expectations as much as the data itself, so as to maximize the amount of useful information transmitted to the user. This will ultimately bring the power of EDM within reach of the non-expert.

In this presentation, we will first present our results so far, with a focus on overarching theoretical insights, as well as specific new approaches for mining subjectively interesting patterns in relational data, in networks, and in high-dimensional data. We will conclude with a look to the future, during the remainder of the FORSIED project, and beyond.

Fosca Giannotti

Istituto di Scienza e Tecnologie dell'Informazione, National Research Council of Italy (ISTI-CNR)

SoBigData Research Infrastructure

Link:

www.sobigdata.eu

Grant:

Project ID: 654024

Funding scheme:

RIA - Research and Innovation action

Abstract:

The SoBigData Research Infrastructure provides an integrated ecosystem for ethic-sensitive scientific discoveries and advanced applications of social data mining on the various dimensions of social life. In addition, as an open research infrastructure, SoBigData promotes repeatable and open science. Its mission is to support data science research projects by providing:

▪ An ever-growing, distributed data ecosystem for procurement, access and curation and management of big social data, to underpin social data mining research within an ethic-sensitive context. This is based on innovative strategies for acquiring social big data for research purposes, using both opportunistic means offered by social sensing technologies and participatory means based on user involvement as prosumers of social data and knowledge.

▪ An ever-growing, distributed platform of interoperable, social data mining methods and associated skills: tools, methodologies and services for mining, analyzing, and visualizing complex and massive datasets, harnessing the techno-legal barriers to the ethically safe deployment of big data for social mining.

▪ An ecosystem where protection of personal information and the respect for fundamental human rights can coexist with a safe use of the same information for scientific purposes of broad and central societal interest. SoBigData strengthens research concerning the protection of personal data as a fundamental right, while at the same time boosting the free flow of personal data as a common good. To this aim a Legal and Ethical board is established to the aim to serve all users of the RI.

The talk will present the SoBigData Vision, Goal, Organization and opportunities for researchers to be users and partners of this initiative.

Partners: CNR (coordinator, Italy); University of Sheffield (UK); Fraunhofer: IAIS and IGD Institutes e Gottfried Wilhelm Leibniz Universitaet Hannover (DE); the centre E-Gov.data, at University of Tartu (Estonia), Aalto University: Sociophysics and Data Mining laboratories (Finland); ETH Zurich; Technische Universiteit Delft (NL), University of Pisa, Scuola Normale Superiore di Pisa, Scuola IMT (Italy).

Hussain Kazmi
Enervalis, Belgium

REnnovates: Flexibility Activated Zero Energy Districts

Link:
http://rennovates.eu/

Grant:
Project ID: 680603

Funding scheme:
IA - Innovation action

Abstract:
REnnovates is a deep renovation concept that develops smart energy-based communities resulting in energy-neutral housing – up to and beyond Zero Net Energy (i.e. the household produces as much energy annually (or more) as it consumes). This is done by reducing energy consumption and maximizing the use of renewable energy. The project started in September 2015 and is co-funded by the European Commission in the H2020 Programme. Over the course of the project, multiple demo projects will be developed to investigate the concept, with a real world pilot already well underway in the Netherlands.

This is important because residential buildings account for roughly 20% of the total European energy consumption. Reductions in this figure can be achieved by:

1. Improving the building's thermal characteristics (e.g. through better façade insulation)
2. Improving the operational behavior of the building (e.g. through smart control)
3. Changing occupant behavior (e.g. through recommender systems)

REnnovates aims to address all three of these concerns to create energy efficient buildings and well-informed occupants. The improved building envelope relies on a thorough design process and a subsequent deep renovation which can reduce energy consumption by 60%. Machine learning algorithms can complement this and provide substantial additional value through smart control and providing advice to building occupants. The key ML challenge that we address in this project is that no prior knowledge about the building or its thermal systems is assumed and models for these are learnt in completely online settings. Additional operational synergies to reduce the load on the electric grid as a result of these Net Zero Energy Buildings are also investigated.

Methodology
In the REnnovates project, the machine learning component is divided into two distinct parts: an operational aspect and a recommendation component.

Operational control: here, the controller is responsible for provision of hot water and space heating to the building in a way that doesn't adversely impact occupant comfort while optimizing towards a set objective. The objectives investigated in REnnovates include building energy efficiency, maximization of solar self-consumption and electricity grid interaction. We demonstrate reinforcement agents which learn appropriate control actions to achieve these objectives without adversely affecting occupant comfort. This leads to potential savings of hundreds of kilo-watt hours per year per house and substantial greenhouse gas reductions. Smart controlling to reduce electric grid interaction on the other hand has the potential to substantially reduce grid reinforcement costs which would otherwise be necessitated.

Recommender system: for building a recommender system, the model learnt by the reinforcement learner can be used to advise users about taking appropriate control actions. This serves two purposes, firstly it informs users about the consequences of their actions. This can, despite the low elasticity of electricity demand, help achieve more rational use of energy. Secondly, lateral comparisons can be provided to people to introduce an aspect of gamification to the system. This is an on-going research direction.

Since the project is a forerunner to EU-wide rollout, both the operational control and recommender system are generalizable by design. In this way, the developed framework will be applicable to any building which has a roughly similar set of sensor data available. Since only standard sensor data is used primarily in this research, this doesn't lead to loss of generality. Machine learning in general and reinforcement learning in particular provides a tractable, cost-effective framework for doing this which help complement the design optimizations.

Martin Žnidaršič

Jožef Stefan Institute, Ljubljana, Slovenia

CF-Web: ClowdFlows Data and Text Analytics Marketplace on the Web

Link:
http://cordis.europa.eu/project/rcn/208432_en.html

Grant:
Project ID: 754549

Funding scheme:
CSA - Coordination and support action

Abstract:
CF-Web is a FET Innovation Launchpad project which aims at commercializing ClowdFlows platform as a software component marketplace.

ClowdFlows is an open source cloud based platform that supports the composition and execution of data processing workflows on the Web. It has a user-friendly graphical interface and does not require any client side installation as it runs in any modern browser. ClowdFlows supports data analytics and text mining that enable discovery of patterns and regularities which potentially lead to new insights. Data and text processing in ClowdFlows is easily managed by connecting processing components into an executable workflow executed in the cloud. Once constructed, workflows can be reused to process static data or high volumes of data streams. The web based platform is provided as a hosted service, with the ability for users to install it on a private cloud.

In scope of CF-Web, we will develop a business model in which ClowdFlows serves as a workflow publishing platform: a marketplace of ready-made workflows and data/text processing components, offering components for reuse to companies lacking the expertise and tools to perform machine learning and data analytics tasks. This way, from the end user's perspective, the lack of internal developmental resources is compensated by crowdsourcing of needed new analytical components within the marketplace. From the developer's perspective, the ClowdFlows marketplace allows the developers of processing components to publish and monetize their work, and to connect to companies in need of easy to use cloud based data mining tools.

Brian Mac Namee
University College Dublin

AURORA: Advanced User-centric efficiency metRics for air traffic perfORmance Analytics

Link:
http://aurora-er.eu/

Grant:
Project ID: 699340

Funding scheme:
SESAR-RIA - Research and Innovation action

Abstract:
The perspectives of air navigation service providers (ANSP) and airlines on the efficiency of the routes that flights follow can often differ. While airlines are concerned mainly with schedule adherence and fuel consumption, ANSPs focus on more in-depth components like managing sector capacity, minimizing air traffic controllers' interventions, and reducing emissions and noise. Designing indicators to capture these different perspectives for effective air traffic management (ATM) is challenging. Designing a data processing pipeline that can calculate and update these indicators in almost real-time makes this an even bigger challenge. The goal of the AURORA project is to address these challenges.

Figure 1: A visualization of the flight trajectories of every flight in European airspace in a single day (trajectories are reconstructed from ADS-B data).

Current indicators of route efficiency typically used in ATM rely only on radar track and flight plan data. These data sources, however, are insufficient when trying to take into account issues such as flight fuel consumption and cost; and radar track data itself is limited in coverage and typically only available to the local ANSP that generates it. Automatic dependent surveillance broadcast (ADSB) data is an alternative,

widely available, reliable source of global surveillance data with increasing coverage. ADS-B surveillance data can provide accurate location of flights updated approximately every 4 seconds (Figure 1 shows an illustration of reconstructed trajectories for all flights over Europe in a single day). In AURORA we are using ADS-B data to calculate efficiency indicators that capture the different perspectives of both airlines and ANSPs. The methodology we use is based on comparing the differences between flight trajectories reconstructed from ADS-B surveillance data and ideal trajectories generated based on user preferences (for example trajectories constructed to minimize fuel used or cost).

For the efficiency indicators calculated to be most useful, they should be calculated in near real-time and updated every time new surveillance data for a flight becomes available. This involves performing relatively complex processing of large scale, fast changing ADS-B surveillance data and modern data streaming technologies are being used for this. Figure 2 shows a schematic of the AURORA real-time analytics platform that has been developed. Early experimental results indicate that this platform is capable of calculating sophisticated efficiency indicators in near real-time for large volumes of flights. The current platform is based on Apache Kafka, Spark Streaming and Cassandra technologies, and novel approaches to distributed real-time indicator calculation.

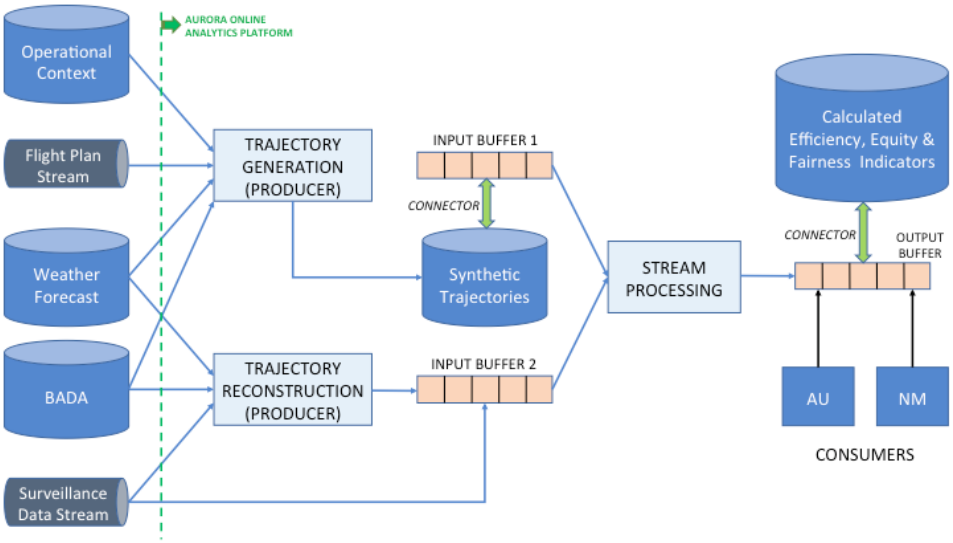

Figure 2: A schematic of the AURORA real-time analytics platform.

The AURORA project is supported by the SESAR Joint Undertaking (SJU) under grant 699340 under the European Union's Horizon 2020 research and innovation programme. The collaborating partners are Centro de Referencia I+D+i ATM (CRIDA), Boeing Research and Technology Europe (BR&TE), Centre for Applied Data Analytics (CeADAR), and Flight Radar 24 (FR24).

Sašo Džeroski

Jožef Stefan Institute, Ljubljana, Slovenia

MAESTRA: Learning from Massive, Incompletely annotated, and Structured Data

Link:
http://maestra-project.eu

Grant:
ICT-2013-612944

Funding scheme:
CP - Collaborative project (generic)

Abstract:
The need for machine learning (ML) and data mining (DM) is ever growing due to the increased pervasiveness of data analysis tasks in almost every area of life, including business, science and technology. Not only is the pervasiveness of data analysis tasks increasing, but so is their complexity. We are increasingly often facing predictive modeling tasks involving one or several of the following complexity aspects: (a) structured data as input or output of the prediction process, (b) very large/massive datasets, with many examples and/or many input/output dimensions, where data may be streaming at high rates, (c) incompletely/partially labelled data, and (d) data placed in a spatio-temporal or network context. Each of these is a major challenge to current ML/DM approaches and is the central topic of active research in areas such as structured-output prediction, mining data streams, semi-supervised learning, and mining network data. The simultaneous presence of several of them is a much harder, currently insurmountable, challenge and severely limits the applicability of ML/DM approaches.

The project concerned with developing predictive modelling methods capable of simultaneously addressing several (ultimately all) of the above complexity aspects. More specifically, we developed various methods for supervised and semi-supervised structured output prediction, learning from data streams and learning from networked data. The developed methods were then applied in various practically relevant problems from life sciences (e.g., association of phenotypic traits with genes, investigation of complementarity of different feature sets for gene function prediction, drug repositioning for M. tuberculosis and S. Typhimurium, habitat modelling of extremophilic funghi), sensor networks (wind and solar power production prediction, public transportation networks) and social/multimedia (e.g., analysis of various types of images, analysis of phone calling patterns, news aggregation based on twitter etc). The presentation at the EU projects forum will outline the major developments and applications from the project.

Vincenzo Lagani, Ioannis Tsamardinos,
Computer Science Department, University of Crete

CAUSALPATH: Next Generation Causal Analysis Inspired by the Induction of Biological Pathways from Cytometry Data

Link:
http://mensxmachina.org/en/projects/causalpath/

Grant:
Project ID: 617393

Funding scheme:
European Research Council (ERC) Consolidation Grant

Abstract:
Controlling a complex system requires a detailed knowledge of the causal mechanism underlying its operation. The field of computational Causal Discovery (CD) provides algorithms that attempt to identify causal relationships from observational data, optionally in conjunction with data resulting from experimental manipulations / interventions. Currently, the application of CD methods on real-world is a non-trivial task, mainly due to a number of limitations these algorithms still suffer.

Biological pathways can be considered as informal causal models representing processes happening at the cellular level. Identifying and organizing all interactions composing a single pathway can require decades of experiments and the work of thousands of scientist. Mass cytometry (Cytometry Time Of Fly, CyTOF) allows the measurement of tens of intra-cellular and surface markers at single-cell level, realistically capturing the multivariate distribution of these molecular quantities. This makes CyTOF data the ideal source for learning pathways by using CD algorithms.

The objective of CAUSALPATH is to further advance the CD field to the point of enabling the induction of biological pathways from cytometry as well as other biological data. “Bridging this gap” requires the development of novel formalisms, algorithms whose assumptions are compatible with the characteristics of biological data, and easy-to-use software tools for disseminating the novel methods across non-expert users. Understanding human T-cells differentiation is the testbed that drives the methodological development in CAUSALPATH, a problem with potential application in the treatment of autoimmune and inflammatory disease, as well as cancer.

With the project approaching its third year of activity, novel methods and tools are now being published and released. SCENERY (scenery.csd.uoc.gr) is the first web application specifically devised for allowing non-expert users applying CD methods on cytometry data, by shelding the researcher from all the technicalities that this task demands. SCENERY allows to easily organize, gate and analyse data with complex experimental design, and it offers a set of analytical tools that spans simple univariate comparisons, trends detection and all CD methods developed in CAUSALPATH.

Richard Wheeler – invited speaker

Edinburgh Scientific

Hints on How to Write a Successful Project Proposal

EU Funding has never been more important, nor harder to get. In this very practical talk, Richard Wheeler will provide an insider's view to securing European Union funding, including tips and tricks on good proposal writing, what really happens in EU review meetings, why most proposals fail, good project management methods, and more. Attendees will have the opportunity to ask questions and discuss their ideas in an informal workshop environment.

Topics will include:

- What makes a good consortium
- Why proposals fail
- The EU review process
- What makes good proposal writing
- Getting started
- Writing: Science and Technology
- Writing: Knowledge Transfer
- Writing: Exchange Programmes
- Writing: Management and Implementation
- Writing: Impact
- Writing: Exploitation
- Writing: Dissemination

Richard Wheeler's profile:

Richard Wheeler is a specialist in artificial intelligence and computer science who has worked for the World Health Organisation in Geneva and The University of Edinburgh, and been a research manager at laboratories in Brussel s in Vienna. He currently runs a private scientific consultancy (Edinburgh Scientific) serving academic and industrial clients across Europe. He is active in the field of renewable energy and scientific management, acts as a chair in a number of EU funding schemes, and recently completed a book "Success with EU Proposals".