DS

Split, Croatia
October 28 - 30, 2019

THE 22nd INTERNATIONAL CONFERENCE ON **DISCOVERY SCIENCE**

PROGRAM

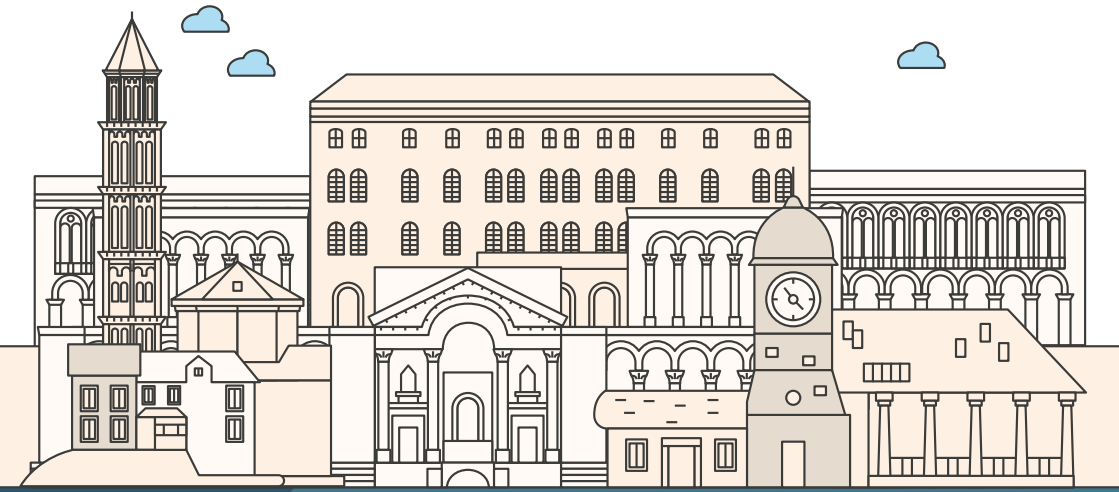
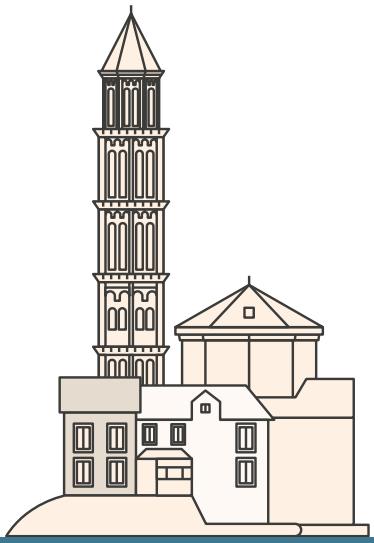


TABLE OF CONTENTS



WELCOME

Sašo Džeroski, General chair and
Petra Kralj Novak and **Tomislav Šmuc**, Program chairs
On behalf of the Discovery Science 2019 Organizing team

Dear colleagues,

Welcome to Croatia, to the Adriatic coast, welcome to Split, and welcome to Discovery Science 2019!

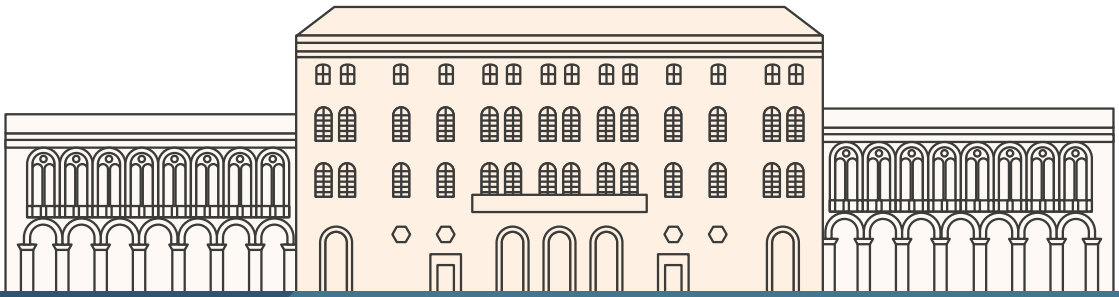
This is the 22nd edition of Discovery Science, this year first time as a stand-alone conference. For its first twenty editions, DS was co-located with the International Conference on Algorithmic Learning Theory (ALT). In 2018 it was co-located with the 24th International Symposium on Methodologies for Intelligent Systems (ISMIS). The 22nd Discovery Science conference received 63 international submissions. Each submission was reviewed by at least three program committee members. The program committee decided to accept 21 regular papers and 19 short papers.

We are glad to be able to present a diverse and scientifically relevant program this year, enriched with highly-profiled keynotes and additional tracks: PhD Symposium and Late Breaking Contributions. The compact scientific program will be accompanied with poster sessions and social program including a welcome reception and a conference dinner. We hope that all this will provide ample opportunities for exciting discussions, exchange of ideas, networking but also pleasure and fun that this venue and city of Split can offer in abundance.

A significant number of people have put considerable hours of work to make this event possible, and to them we express our hearty gratitude: this includes program committee members, awards committee, production and publication chairs, local organizers, proceedings chairs. Furthermore, we would like to thank our partners and sponsors for their generous financial support. At the end, and most importantly, we want to thank Discovery Science community, to all authors that submitted their work for presentation at DS 2019.

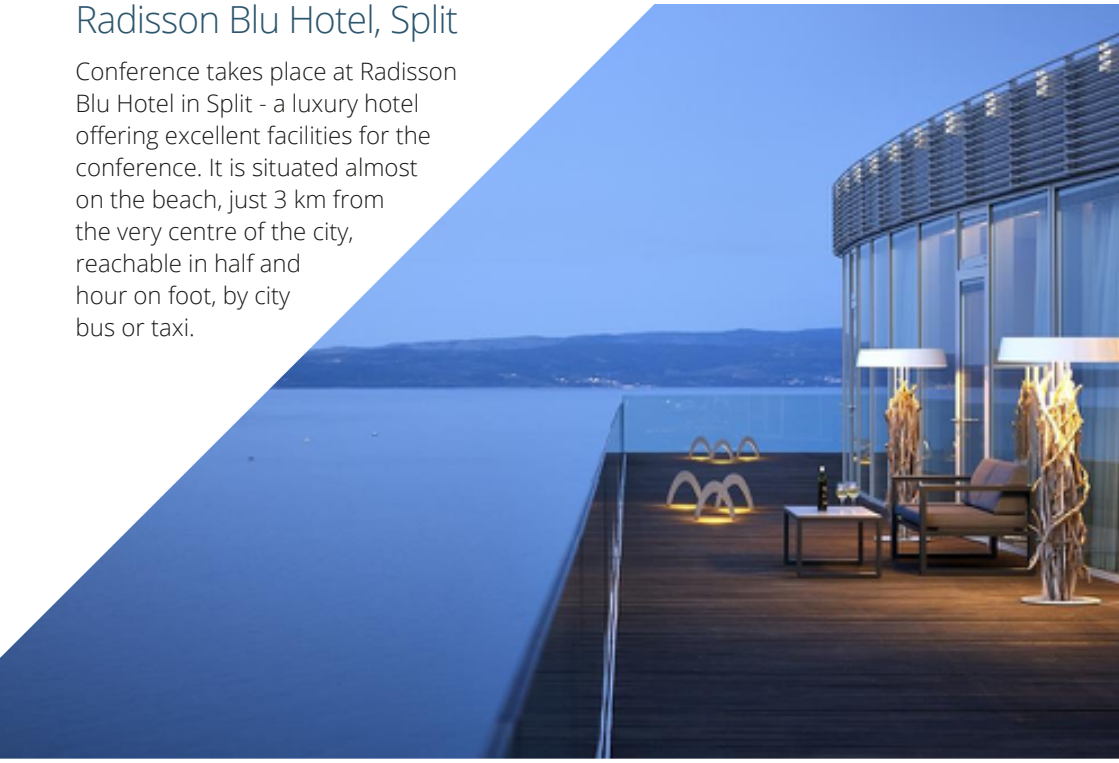
Finally, we would like to thank you for coming to the conference and helping us make it a memorable event, and wish you to enjoy the conference and your stay in Split!

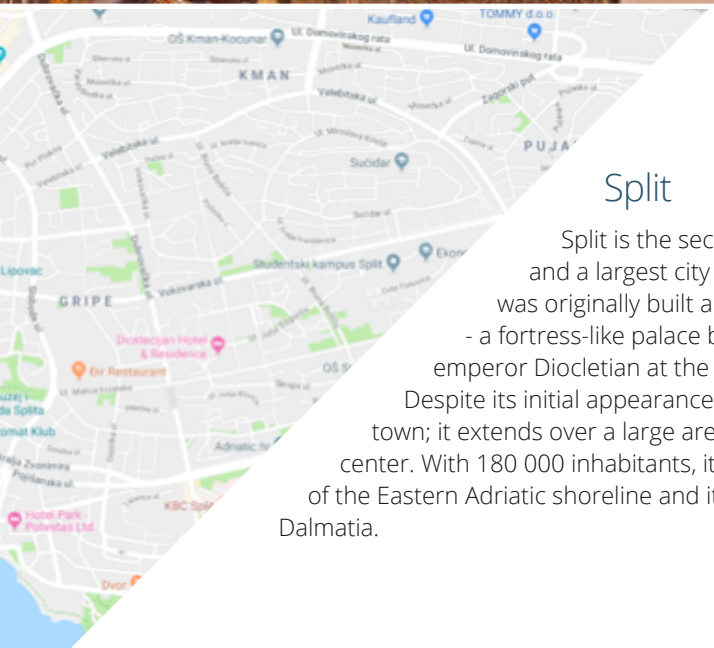
CONFERENCE VENUE



Radisson Blu Hotel, Split

Conference takes place at Radisson Blu Hotel in Split - a luxury hotel offering excellent facilities for the conference. It is situated almost on the beach, just 3 km from the very centre of the city, reachable in half and hour on foot, by city bus or taxi.





Split

Split is the second largest city in Croatia and a largest city in the Dalmatia region. City was originally built around the Diocletian's Palace - a fortress-like palace built for the retired Roman emperor Diocletian at the turn of the fourth century AD. Despite its initial appearance, the city is not a small tourist town; it extends over a large area, well beyond the ancient city center. With 180 000 inhabitants, it represents economic hub of the Eastern Adriatic shoreline and it is an unofficial "capital" of Dalmatia.

GENERAL INFORMATION

Name tags
Participants will obtain name tags at registration. Please bring it with you at all conference events. Additional tickets for accompanying persons to attend social events can be purchased at the registration desk.

Coffee and lunch breaks
Coffee, food and drinks served in the breaks area as well as lunch in the hotel restaurant are included in the registration fee.

Wi-Fi
Wi-Fi will be available at the Conference venue during the whole conference period.

Transport
The Radison Blu hotel is situated 3km from the city centre. You can take a public bus (station nearby), taxi or even walk by the sea (~35min).
Bus transport will be organized for the transport of participants to the City Tour/ Conference dinner. Detailed information about it will be announced on the second day of the conference.

Mobile phones
Please make sure to switch off your mobile phones during the sessions. Note that roaming charges may apply in case of making phone calls, sending SMS and when using data transfer.

Currency
The official currency of Croatia is Kuna; its exchange rate with the EURO is around: 1 EUR=7.4 kuna. ATM machines are available on many places around the city. International credit cards are accepted for payments in most shops, restaurants, hotels.

Emergency contacts
Emergency calls related to the conference matter should be directed to:
+385 91 516 7547 (Matija Piškorec, Registration Desk, Local Organization)
+385 98 186 9605 (Tomislav Šmuc, Local Organization Chair)
The Emergency Call Center in Croatia uses the number **112**.

INSTRUCTIONS
FOR SPEAKERS AND CHAIRS

Instructions for speakers

- The conference room will have a projector and a laptop. The laptop will have Powerpoint and PDF viewing software preinstalled.
- Session speakers should be in the conference room 10 minutes before the start of their session and report to their Session Chair.
- Session speakers will be asked to copy their presentations to the laptop and to check if everything works fine. All of the presentations will be deleted after the end of the session.
- A volunteer will be available to each speaker to provide assistance.
- The times allocated to presentation/speaker, including time for setup and questions is:
    - Regular papers: 20 minutes
    - Short papers: 15 minutes
    - PhD Symposium and Late Breaking Contributions: according to instructions of the chair

Instructions for chairs

- Please make sure to be in the conference hall at least 10 minutes before the session to check that equipment works.
- Ensure that presenters stick to the schedule. If speaker does not show up, announce a short break.
- Moderate the presentation and questions in a way to ensure that the time allocated to a particular presentation is respected (20 minutes for regular papers and 15 minutes for short papers).
- Start sessions on time.

Instructions for poster presentation

- Joint poster session follows short presentations (pitch talks) at PhD Symposium and Late Breaking Contributions session, and is scheduled on the afternoon of 28th October.
- Poster session of the will take place in the conference hall. Panels with presenter name tags will be available on the morning of 28th October.
- Posters should be setup by presenters before the session, preferably during lunch break or afternoon coffee break.
- The poster size should be A0 (841 x 1189 mm), in portrait orientation.
- The posters can stay on panels during the conference, until the last day, to allow informal discussions with presenters during coffee breaks.

KEYNOTE SPEAKERS

DINO PEDRESCHI

Universita di Pisa, Istituto di Scienza e Tecnologie dell'Informazione CNR, (ISTI-CNR)

Monday, 28th October, 9:00

Data and Algorithmic Bias: Explaining the network effect in opinion dynamics and the training data bias in machine learning

Data science and network science are creating novel means to study the complexity of our societies and to measure, understand and predict social phenomena. My talk gives an overview of recent research at the Knowledge Discovery (KDD) Lab in Pisa within the SoBigData.eu research infrastructure, targeted at explaining the effects of data and algorithmic bias in different domains, using both data-driven and model-driven arguments. First, I introduce a model showing how algorithmic bias instilled in an opinion diffusion process artificially yields increased polarisation, fragmentation and instability in a population. Second, I focus on the urgent open challenge of how to construct meaningful explanations of opaque AI/ML black-box decision systems, introducing the local-to-global framework for the explanation of ML classifiers as a way towards explainable AI. The two cases show how the combination of data-driven and model-driven interdisciplinary research has a huge potential to shed new light on complex phenomena like discrimination and polarisation, as well as to explain how decision making black-boxes, both human and artificial, actually work. I conclude with an account of the open data science paradigm pursued in SoBigData.eu Research Infrastructure and its importance for interdisciplinary data driven science that impacts societal challenges.

About the speaker: *Dino Pedreschi* is a professor of computer science at the University of Pisa, and a pioneering scientist in data science. He co-leads the Pisa KDD Lab – Knowledge Discovery and Data Mining Laboratory http://kdd.isti.cnr.it, a joint research initiative of the University of Pisa and the Information Science and Technology Institute of the Italian National Research Council. His research focus is on big data analytics and mining and their impact on society. He is a founder of the Business Informatics MSc program at Univ. Pisa, a course targeted at the education of interdisciplinary data scientists, and of SoBigData.eu, the European H2020 Research Infrastructure "Big Data Analytics and Social Mining Ecosystem" www.sobigdata.eu. Dino has been a visiting scientist at Barabasi Lab (Center for Complex Network Research) of Northeastern University, Boston, and earlier at the University of Texas at Austin, at CWI Amsterdam and at UCLA. In 2009, Dino received a Google Research Award for his research on privacy-preserving data mining. Dino is a member of the expert group in AI of the Italian Ministry of research and the director of the Data Science PhD program at Scuola Normale Superiore in Pisa. Dino is a co-PI of the 2019 ERC grant XAI – *Science and technology for the explanation of AI decision making (PI: Fosca Giannotti)*

GUIDO CALDARELLI

IMT Lucca,
European Centre for Living Technology Venice
Tuesday, 29th October, 9:00
Chair: Petra Kralj Novak

The Structure of Financial Networks

Financial inter-linkages play an important role in the emergence of financial instabilities and the formulation of systemic risk can greatly benefit from a network approach. In this talk, we focus on the role of linkages along the two dimensions of contagion and liquidity, and we discuss some insights that have recently emerged from network models. With respect to the issue of the determination of the optimal architecture of the financial system, models suggest that regulators have to look at the interplay of network topology, capital requirements, and market liquidity. With respect to the issue of the determination of systemically important financial institutions, the findings indicate that both from the point of view of contagion and from the point of view of liquidity provision, there is more to systemic importance than just size. In particular for contagion, the position of institutions in the network matters and their impact can be computed through stress tests even when there are no defaults in the system. We present an overview of the use of networks in Finance and Economics. We show how this approach enables us to address important questions as, for example, the stability of financial systems and the systemic risk associated with the functioning of the interbank market. For example with DebtRank, a novel measure of systemic impact inspired by feedbackcentrality we are able to measure the nodes that become systemically important at the peak of the crisis. Moreover, a systemic default could have been triggered even by small dispersed shocks. The results suggest that the debate on too-big-to-fail institutions should include the even more serious issue of too-central-to-fail. All these results are new in the eld and allow for a better understanding and modelling of different Financial systems.

About the speaker: Guido Caldarelli is Full Professor in Theoretical Physics at IMT School for Advanced Studies Lucca, and is Research associate at the European Centre for Living Technology, Venice. His main scientific activty is the study of networks, mostly analysis and modelling of Financial networks. Author of more than 200 publication on the subject and three books, he is currently the president of the Complex Systems Society. He has been coordinator of FET IP Project MULTIPLEX: Foundational Research on Multilevel Complex Networks and Systems (2012 2016). Coordinator of FET OPEN Project FoC: Forecasting Financial Crises (2010 2014); coordinator of FET OPEN Project COSIN: Coevolution and Self Organization in Complex Networks (2002 2005). Guido Caldarelli received his Ph.D. from SISSA, after which he was a postdoc in the Department of Physics and School of Biology, University of Manchester. He then worked at the Theory of Condensed Matter Group, University of Cambridge. He returned to Italy as a lecturer at National Institute for Condensed Matter (INFM) and later as Primo Ricercatore in the Institute of Complex Systems of the National Research Council of Italy. In this period, he was also the coordinator of the Networks subproject, part of the Complexity Project, for the Fermi Centre. He also spent some terms at University of Fribourg (Switzerland) and in 2006 he has been visiting professor at cole Normale Suprieure in Paris. More information and a complete CV are available at: http://www.guidocaldarelli.com.

MARINKA ŽITNIK

Stanford University
Wednesday, 29th October, 9:00

Chair: Tomislav Šmuc

Representation Learning as a new Approach to Biomedical Research

Large datasets are being generated that can transform science and medicine. New machine learning methods are necessary to unlock these data and open doors for scientific discoveries. In this talk, I will argue that machine learning models should not be trained in the context of one particular dataset. Instead, we should be developing methods that combine data in their broadest sense into knowledge networks, enhance these networks to reduce biases and uncertainty, and then learn and reason over the networks. My talk will focus on two key aspects of this goal: representation learning and network science for knowledge networks. I will show how realizing this goal can set sights on new frontiers beyond classic applications of neural networks on biomedical image and sequence data. I will start by presenting a framework that learns deep models by embedding knowledge networks into compact embedding spaces whose geometry is optimized to reflect network topology, the essence of networks. I will then describe two applications of the framework to drug discovery and medicine. First, the framework allowed us to, for the first time, predict the safety of drug combinations at scale. We embedded a knowledge network of molecular, drug, and patient data at the scale of billions of interactions for all medications in the U.S. Using the embeddings, the approach can predict unwanted side effects for any combination of drugs that patients take, and we can validate predictions in the clinic using real patient data. Second, I will discuss how the framework enabled us to predict what diseases a new drug could treat. I will show how the new approach can make correct predictions for many recently repurposed drugs and can operate even on the hardest, yet critical, diseases for which no good treatments exist. I will conclude with future directions for learning over interaction data and translation of machine learning methods into solutions for biomedical problems.

About the speaker: Marinka Zitnik is a postdoctoral scholar in Computer Science at Stanford University. She will join Harvard University as a tenure-track assistant professor in December 2019. Her research investigates machine learning for sciences. Her methods have had a tangible impact in biology, genomics, and drug discovery, and are used by major biomedical institutions, including Baylor College of Medicine, Karolinska Institute, Stanford Medical School, and Massachusetts General Hospital. She received her Ph.D. in Computer Science from University of Ljubljana while also researching at Imperial College London, University of Toronto, Baylor College of Medicine, and Stanford University. Her work received several best paper, poster, and research awards from the International Society for Computational Biology. She was named a Rising Star in EECS by MIT and also a Next Generation in Biomedicine by The Broad Institute of Harvard and MIT, being the only young scientist who received such recognition in both EECS and Biomedicine. She is also a member of the Chan Zuckerberg Biohub at Stanford.

BEST PAPERS AWARDS

Efficient Discovery of Expressive Multi-label Rules using Relaxed Pruning

Yannik Klein, Michael Rapp and Eneldo Loza Mencía

Time: 28th October, 13:45-14:15

Abstract: Being able to model correlations between labels is considered crucial in multi-label classification. Rule-based models enable to expose such dependencies, e.g., implications, subsumptions, or exclusions, in an interpretable and human-comprehensible manner. Albeit the number of possible label combinations increases exponentially with the number of available labels, it has been shown that rules with multiple labels in their heads, which are a natural form to model local label dependencies, can be induced efficiently by exploiting certain properties of rule evaluation measures and pruning the label search space accordingly. However, experiments have revealed that multi-label heads are unlikely to be learned by existing methods due to their restrictiveness. To overcome this limitation, we propose a plug-in approach that relaxes the search space pruning used by existing methods in order to introduce a bias towards larger multi-label heads resulting in more expressive rules. We further demonstrate the effectiveness of our approach empirically and show that it does not come with drawbacks in terms of training time or predictive performance.

Sparse Robust Regression for Explaining Classifiers

Anton Björklund, Andreas Henelius, Emilia Oikarinen, Kimmo Kallonen and Kai Puolamäki

Time: 29th October, 13:45-14:15

Abstract: Real-world datasets are often characterised by outliers, points far from the majority of the points, which might negatively influence modelling of the data. In data analysis it is hence important to use methods that are robust to outliers. In this paper we develop a robust regression method for finding the largest subset in the data that can be approximated using a sparse linear model to a given precision. We show that the problem is NP-hard and hard to approximate. We present an efficient algorithm, termed SLISE, to find solutions to the problem. Our method extends current state-of-the-art robust regression methods, especially in terms of scalability on large datasets. Furthermore, we show that our method can be used to yield interpretable explanations for individual decisions by opaque, black box, classifiers. Our approach solves shortcomings in other recent explanation methods by not requiring sampling of new data points and by being usable without modifications across various data domains. We demonstrate our method using both synthetic and real-world regression and classification problems.

CONFERENCE PROGRAM

DAY 1
Monday, 28th October

8:00-18:00	**Registration**
8:45-9:00	**Conference opening:** Sašo Džeroski, Petra Kralj Novak, Tomislav Šmuc
9:00-10:00	**Keynote talk 1: Dino Pedreschi** Title: ***Data and Algorithmic Bias: Explaining the Network Effect in Opinion Dynamics and the Training Data Bias in Machine Learning*** Chair: Sašo Džeroski
10:00-10:30	Coffee Break
10:30-12:00	**SESSION 1: Advanced machine learning I**
	Chair: Dino Ienco
10:30-10:50	Colin Bellinger, Paula Branco and Luis Torgo ***The CURE for Class Imbalance***
10:50-11:10	Vu-Linh Nguyen, Sébastien Destercke and Eyke Huellermeier ***Epistemic Uncertainty Sampling***
11:10-11:30	Elena Battaglia and Ruggero G. Pensa ***Parameter-less Tensor Co-clustering***
11:30-11:45	André Correia, Carlos Soares and Alípio Jorge ***Dataset Morphing to Analyze the Performance of Collaborative Filtering***
11:45-12:00	Takayasu Fushimi, Kiyoto Iwasaki, Seiya Okubo and Kazumi Saito ***Construction of Histogram with Variable Bin-width based on Change Point Detection***
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)
13:45-15:15	**SESSION 2: Interpretable Machine Learning**
	Chair: Blaž Zupan
13:45-14:15	**BEST PAPER I** Yannik Klein, Michael Rapp and Eneldo Loza Mencía ***Efficient Discovery of Expressive Multi-label Rules using Relaxed Pruning***
14:15-14:35	Michael Rapp, Eneldo Loza Mencía and Johannes Fürnkranz ***On the Trade-off Between Consistency and Coverage in Multi-label Rule Learning Heuristics***
14:35-14:55	Fabrizio Angiulli, Fabio Fassetti, Luigi Palopoli and Cristina Serrao ***A density estimation approach for detecting and explaining exceptional values in categorical data***
14:55-15:15	Martin Atzmueller, Stefan Bloemheuvel and Benjamin Klöpper ***A Framework for Human-Centered Exploration of Complex Event Log Graphs***
15:15-15:45	Coffee break and poster exhibition

15:45-17:10	**SESSION 3: Data and Knowledge Representation**
	Chair: Michelangelo Ceci
15:45-16:00	Ana Kostovska, Ilin Tolovski, Fatima Maikore, Larisa Soldatova and Pance Panov ***Neurodegenerative Disease Data Ontology***
16:00-16:20	Blaž Škrlj, Nada Lavrač and Jan Kralj ***Symbolic Graph Embedding using Frequent Pattern Mining***
16:20-16:40	Pavlin Gregor Policar, Martin Strazar and Blaz Zupan ***Embedding to Reference t-SNE Space Addresses Batch-Effects in Single-Cell Classification***
16:40-17:00	Dino Ienco and Ruggero G. Pensa ***Deep Triplet-Driven Semi-Supervised Embedding Clustering***
17:10-19:00	PhD Symposium & Late Breaking Contributions (pitch talks + poster session)
	Chair: Tomislav Lipić
19:10-20:00	Steering Committee Meeting
20:00-22:00	Welcome Reception at Radison Blue Terrace

DAY 2
Tuesday, 29th October

8:00-18:00	Registration
8:55-9:00	**Daily announcements**
9:00-10:00	**Keynote talk 2: Guido Caldarelli** Title: ***The Structure of Financial Networks*** Chair: Petra Kralj Novak
10:00-10:30	Coffee Break and Poster Exhibition

10:30-11:55	**SESSION 4: Pattern Discovery, Time Series**
	Chair: Martin Atzmueller
10:30-10:50	Vitor Cerqueira, Luis Torgo and Carlos Soares ***Layered Learning for Early Anomaly Detection: Predicting Critical Health Episodes***
10:50-11:10	Bozhidar Stevanoski, Dragi Kocev, Aljaž Osojnik, Ivica Dimitrovski and Saso Dzeroski ***Predicting thermal power consumption of the Mars Express satellite with data stream mining***
11:10-11:25	Kemilly Dearo Garcia, Elaine Ribeiro de Faria, Cláudio Rebelo de Sá, João Mendes-Moreira, Charu C. Aggarwal, André C.P.L.F de Carvalho and Joost N. Kok ***Ensemble Clustering For Novelty Detection In Data Streams***
11:25-11:40	Samaneh Khoshrou and Mykola Pechenizkiy ***Adaptive Long-term Ensemble Learning from Multiple High-dimensional Time-series***
11:40-11:55	Aljaž Osojnik, Pance Panov and Saso Dzeroski ***Utilizing Hierarchies in Tree-based Online Structured Output Prediction***
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)

13:45-15:15	**SESSION 5: Advanced Machine Learning II; Interpretable Machine Learing II**
	Chair: Luis Torgo
13:45-14:15	**BEST PAPER II** Anton Björklund, Andreas Henelius, Emilia Oikarinen, Kimmo Kallonen and Kai Puolamäki ***Sparse Robust Regression for Explaining Classifiers***
14:15-14:35	Matej Petković, Saso Dzeroski and Dragi Kocev ***Ensemble-Based Feature Ranking for Semi-supervised Classification***
14:35-14:50	Mohsen Ahmadi Fahandar and Eyke Hüllermeier ***Feature Selection for Analogy-Based Learning to Rank***
14:50-15:05	Cláudio Rebelo de Sá ***Variance-based Feature Importance in Neural Networks***
15:05-15:20	Nyoman Juniarta, Miguel Couceiro and Amedeo Napoli ***A Unified Approach to Biclustering Based on Formal Concept Analysis and Interval Pattern Structure***
15:20-15:35	Sandy Moens, Boris Cule and Bart Goethals ***A Sampling-based Approach for Discovering Subspace Clusters***
15:35-16:00	Coffee Break and Poster Exhibition
16:00-17:30	**SESSION 6: Networks**
	Chair: Boris Cule
16:00-16:20	Domenico Mandaglio and Andrea Tagarelli ***A Combinatorial Multi-Armed Bandit based method for Dynamic Consensus Community Detection in Temporal Networks***
16:20-16:40	Kazumi Saito, Kouzou Ohara, Masahiro Kimura and Hiroshi Motoda ***Resampling-based Framework for Unbiased Estimator of Node Centrality over Large Complex Network***
16:40-17:00	Angelo Impedovo, Michelangelo Ceci and Toon Calders ***Efficient and Accurate Non-exhaustive Pattern-based Change Detection in Dynamic Networks***
17:00-17:15	Sofia Fernandes, Hadi Fanaee-T and Joao Gama ***Evolving social networks analysis via tensor decompositions: from global event detection towards local pattern discovery and specification***
17:15-17:30	Vincent Branders, Guillaume Derval, Pierre Schaus and Pierre Dupont ***Mining a maximum weighted set of disjoint submatrices***
17:30-18:15	Community Meeting
18:30-19:45	City Tour (Diocletian Palace Tour): Chairs: Matija Piškorec, Ana Vidoš
20:00-22:00	Gala Dinner (TBA)

DAY 3
Wednesday, 30th October

8:00-18:00	**Registration**
8:55-9:00	**Daily announcements**
9:00-10:00	**Keynote talk: Marinka Žitnik** Title: ***Representation Learning as a New Approach to Biomedical Research*** Chair: Tomislav Šmuc
10:00-10:30	Coffee Break
10:30-12:00	**SESSION 7: Applications**
	Chair: Larisa Soldatova
10:30-10:50	Adriano Rivolli, Catarina Amaral, Luis Guardão, Cláudio Rebelo de Sá and Carlos Soares ***KnowBots: Discovering Relevant Patterns in Chatbot Dialogues***
10:50-11:05	Vladimir Kuzmanovski, Mika Sulkava, Taru Palosuo and Jaakko Hollmen ***Temporal analysis of adverse weather conditions affecting wheat production in Finland***
11:05-11:20	Andreia Conceição and João Gama ***Main Factors Driving the Open Rate of Email Marketing Campaigns***
11:20-11:35	Erik Dovgan, Bojan Leskošek, Gregor Jurak, Gregor Starc, Maroje Sorić and Mitja Luštrek ***Enhancing BMI-Based Student Clustering by Considering Fitness as Key Attribute***
11:35-11:50	Hoang Son Pham, Siegfried Nijssen, Kim Mens, Dario Di Nucci, Tim Molderez, Coen De Roover, Johan Fabry and Vadim Zaytsev ***Mining Patterns in Source Code using Tree Mining Algorithms***
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)
15:15-15:30	**SESSION 8: Time Series, Applications**
	Chair: Joao Gama
13:45-14:05	Abhina Sharma, Jan N. van Rijn, Frank Hutter and Andreas Mueller ***Hyperparameter Importance for Image Classification by Residual Networks***
14:05-14:25	Amin Azari, Panagiotis Papapetrou, Stojan Denic and Gunnar Peters ***Cellular Traffic Prediction and Classification: a comparative evaluation of LSTM and ARIMA***
14:25-14:45	Natasa Sarafijanovic-Djukic and Jesse Davis ***Fast Distance-based Anomaly Detection in Images Using an Inception-like Autoencoder***
14:45-15:00	Julian Vexler and Stefan Kramer ***Integrating LSTMs with Online Density Estimation for the Probabilistic Forecast of Energy Consumption***
15:00-15:15	Sascha Krstanovic and Heiko Paulheim ***Fourier-based Parametrization of CNNs for Robust Time Series Forecasting***
15:15-15:30	Qianqian Gu and Ross King ***On Recognizing Cats and Dogs in Chinese Paintings***
15:30-15:45	Closing of the Conference

SESSIONS WITH ABSTRACTS

Advanced Machine Learning I

Chair: Dino Ienco

Time: 28th October, 10:30-10:50

The CURE for Class Imbalance

Colin Bellinger, Paula Branco and Luis Torgo

Addressing the class imbalance problem is critical for several real world applications. The application of pre-processing methods is a popular way of dealing with this problem. These solutions increase the rare class examples and/or decrease the normal class cases. However, these procedures typically only take into account the characteristics of each individual class. This segmented view of the data can have a negative impact. We propose a new method that uses an integrated view of the data classes to generate new examples and remove cases. ClUstered REsampling (CURE) is a method based on a holistic view of the data that uses hierarchical clustering and a new distance measure to guide the sampling procedure. Clusters generated in this way take into account the structure of the data. This enables CURE to avoid common mistakes made by other resampling methods. In particular, CURE prevents the generation of synthetic examples in dangerous regions and undersamples safe, non-borderline, regions of the majority class. We show the effectiveness of CURE in an extensive set of experiments with benchmark domains. We also show that CURE is a user-friendly method that does not require extensive ne-tuning of hyper-parameters.

Time: 28th October, 10:50-11:10

Epistemic Uncertainty Sampling

Vu-Linh Nguyen, Sébastien Destercke and Eyke Huellermeier

Various strategies for active learning have been proposed in the machine learning literature. In uncertainty sampling, which is among the most popular approaches, the active learner sequentially queries the label of those instances for which its current prediction is maximally uncertain. The predictions as well as the measures used to quantify the degree of uncertainty, such as entropy, are almost exclusively of a probabilistic nature. In this paper, we advocate a distinction between two different types

of uncertainty, referred to as epistemic and aleatoric, in the context of active learning. Roughly speaking, these notions capture the reducible and the irreducible part of the total uncertainty in a prediction, respectively. We conjecture that, in uncertainty sampling, the usefulness of an instance is better reflected by its epistemic than by its aleatoric uncertainty. This leads us to suggest the principle of "epistemic uncertainty sampling", which we instantiate by means of a concrete approach for measuring epistemic and aleatoric uncertainty. In experimental studies, epistemic uncertainty sampling does indeed show promising performance.

Time: 28th October, 11:10-11:30

Parameter-less Tensor Co-clustering

Elena Battaglia and Ruggero G. Pensa

Tensors co-clustering has been proven useful in many applications, due to its ability of coping with high-dimensional data and sparsity. However, setting up a co-clustering algorithm properly requires the specification of the desired number of clusters for each mode as input parameters. This choice is already difficult in relatively easy settings, like flat clustering on data matrices, but on tensors it could be even more frustrating. To face this issue, we propose a tensor co-clustering algorithm that does not require the number of desired co-clusters as input, as it optimizes an objective function based on a measure of association across discrete random variables (called Goodman and Kruskal's τ) that is not affected by their cardinality. The effectiveness of our algorithm is shown on both synthetic and real-world datasets, also in comparison with state-of-the-art co-clustering methods based on tensor factorization.

Time: 28th October, 11:30-11:45

Dataset Morphing to Analyze the Performance of Collaborative Filtering

André Correia, Carlos Soares and Alípio Jorge

Machine Learning algorithms are often too complex to be studied from a purely analytical point of view. Alternatively, with a reasonably large number of datasets one can empirically observe the behavior of a given algorithm in different conditions and hypothesize some general characteristics. This knowledge about algorithms can be used to choose the most appropriate one given a new dataset. This very hard problem can be approached using meta-learning. Unfortunately, the number of datasets available may not be sufficient to obtain reliable meta-knowledge. Additionally, datasets may change with time, by growing, shrinking and editing, due to natural actions like people buying in a e-commerce site. In this paper we propose dataset morphing as the basis of a novel methodology that can help overcome these drawbacks and can be used to better

understand ML algorithms. It consists of manipulating real datasets through the iterative application of gradual transformations (morphing) and by observing the changes in the behavior of learning algorithms while relating these changes with changes in the meta-features of the morphed datasets. Although dataset morphing can be envisaged in a much wider framework, we focus on one very specific instance: the study of collaborative filtering algorithms on binary data. Results show that the proposed approach is feasible and that it can be used to identify useful meta-features to predict the best collaborative filtering algorithm for a given dataset.

Time: 28th October, 11:45-12:00

Construction of Histogram with Variable Bin-width based on Change Point Detection

Takayasu Fushimi, Kiyoto Iwasaki, Seiya Okubo and Kazumi Saito

For a given set of samples with a numeric variable and a set of nominal variables, we address a problem of constructing a histogram drawn by K bins with variable widths, so as to have relatively large numbers of narrow bins for some ranges where numeric values distribute densely and change substantially, while small numbers of wide bins for the other ranges, together with the characteristic nominal values for describing these bins as annotation terms. For this purpose, we propose a new method, which incorporates a change point detection method to numeric values based on an L_1 or L_2 error criterion, and an annotation terms identification method for these bins based on the z-score with respect to the distribution of nominal values. In our experiments using four datasets of humidity deficit (HD) collected from vinyl greenhouses, we show that our proposed method can construct more natural histograms with appropriate variable bin widths than those with an equal bin width constructed by the standard method based on square-root choice or Sturges' formula, the histograms constructed with the L_1 error criterion has more desirable property than those with the L_2 error criterion, and our method can produce a series of naturally interpretable annotation terms for the constructed bins.

Interpretable Machine Learning

Chair: Blaž Zupan

BEST PAPER I

Time: 28th October, 13:45-14:15

Efficient Discovery of Expressive Multi-label Rules using Relaxed Pruning

Yannik Klein, Michael Rapp and Eneldo Loza Mencía

Being able to model correlations between labels is considered crucial in multi-label classification. Rule-based models enable to expose such dependencies, e.g., implications, subsumptions, or exclusions, in an interpretable and human-comprehensible manner. Albeit the number of possible label combinations increases exponentially with the number of available labels, it has been shown that rules with multiple labels in their heads, which are a natural form to model local label dependencies, can be induced efficiently by exploiting certain properties of rule evaluation measures and pruning the label search space accordingly. However, experiments have revealed that multi-label heads are unlikely to be learned by existing methods due to their restrictiveness. To overcome this limitation, we propose a plug-in approach that relaxes the search space pruning used by existing methods in order to introduce a bias towards larger multi-label heads resulting in more expressive rules. We further demonstrate the effectiveness of our approach empirically and show that it does not come with drawbacks in terms of training time or predictive performance.

Time: 28th October, 14:15-14:35

On the Trade-off Between Consistency and Coverage in Multi-label Rule Learning Heuristics

Michael Rapp, Eneldo Loza Mencía and Johannes Fürnkranz

Recently, several authors have advocated the use of rule learning algorithms to model multi-label data, as rules are interpretable and can be comprehended, analyzed, or qualitatively evaluated by domain experts. Many rule learning algorithms employ a heuristic-guided search for rules that model regularities contained in the training data and it is commonly accepted that the choice of the heuristic has a significant impact on the predictive performance of the learner. Whereas the properties of rule learning heuristics have been studied in the realm of single-label classification, there is no such work taking into account the particularities of multi-label classification. This is surprising, as the quality

of multi-label predictions is usually assessed in terms of a variety of different, potentially competing, performance measures that cannot all be optimized by a single learner at the same time. In this work, we show empirically that it is crucial to trade off the consistency and coverage of rules differently, depending on which multi-label measure should be optimized by a model. Based on these findings, we emphasize the need for configurable learners that can flexibly use different heuristics. As our experiments reveal, the choice of the heuristic is not straight-forward, because a search for rules that optimize a measure locally does usually not result in a model that maximizes that measure globally.

Time: 28th October, 14:35-14:55

A density estimation approach for detecting and explaining exceptional values in categorical data

Fabrizio Angiulli, Fabio Fassetti, Luigi Palopoli and Cristina Serrao

In this work we deal with the problem of detecting and explaining exceptional behaving values in categorical datasets. As a first main contribution we provide the notion of frequency occurrence which can be thought as a form of Kernel Density Estimation applied to the domain of frequency values. As a second contribution, we define an outlierness measure for categorical values that, leveraging the cdf of the density described above, decides if the frequency of a certain value is rare if compared to the frequencies associated with the other values. This measure is able to simultaneously identify two kinds of anomalies called outliers and outliers, namely exceptionally low or high frequent values. The experiments highlight that the method is scalable and able to identify anomalies of different nature from traditional techniques.

Time: 28th October, 14:55-15:15

A Framework for Human-Centered Exploration of Complex Event Log Graphs

Martin Atzmueller, Stefan Bloemheuvel and Benjamin Klöpper

Graphs can conveniently model complex multi-relational characteristics. For making sense of such data, effective interpretable methods for their exploration are crucial, in order to provide insights that cover the relevant analytical questions and are understandable to humans. This paper presents a framework for human-centered exploration of attributed graphs on complex, \ie large and heterogeneous event logs. The proposed approach is based on specific graph modeling, graph summarization and local pattern mining methods. We demonstrate promising results in the context of a real-world industrial dataset.

Data and Knowledge Representation

Chair: Michelangelo Ceci

Time: 28th October, 15:45-16:00

Neurodegenerative Disease Data Ontology

Ana Kostovska, Ilin Tolovski, Fatima Maikore, Larisa Soldatova and Pance Panov

In this paper, we report on the ontology for the representation of brain diseases data - NDDO. The proposed ontology facilitates semantic annotation of datasets containing neurodegenerative diagnostic data (i.e. clinical, imaging, biomarker, etc.) and disease progression data collected on patients by the hospitals. Rich semantic annotation of datasets is essential for efficient support of data mining, for example for the identification of suitable algorithms for data analytics, text mining, and reasoning over distributed data and knowledge sources. To address the data analytics perspective, we reused and extended our previous work on ontology of data types (OntoDT) and ontology of core data mining entities (OntoDM-core) to represent specific domain datatypes that occur in the domain datasets. We demonstrate the utility of NDDO on two use cases: semantic annotation of datasets, and incorporating information about clinical procedures used to produce neurodegenerative data.

Time: 28th October, 16:00-16:20

Symbolic Graph Embedding using Frequent Pattern Mining

Blaž Škrlj, Nada Lavrač and Jan Kralj

Relational data mining is becoming ubiquitous in many fields of study. It offers insights into behavior of complex, real-world systems which cannot be modeled directly using propositional learning. We propose Symbolic Graph Embedding (SGE), an algorithm aimed to learn symbolic node representations. Built on the ideas from the field of inductive logic programming, SGE first samples a given node's neighborhood and interprets it as a transaction database, which is used for frequent pattern mining to identify logical conjuncts of items that co-occur frequently in a given context. Such patterns are in this work used as features to represent individual nodes, yielding interpretable, symbolic node embeddings. The proposed SGE approach on a venue classification task outperforms shallow node embedding methods such as DeepWalk, and performs similarly to metapath2vec, a black-box representation learner that can exploit node and edge types in a given graph. The proposed SGE approach performs especially well when small amounts of data are used for learning, scales to graphs with millions of nodes and edges, and can be run on an of-the-shelf laptop.

Time: 28th October, 16:20-16:40

Embedding to Reference t-SNE Space Addresses Batch-Effects in Single-Cell Classification

Pavlin Gregor Policar, Martin Strazar and Blaz Zupan

Dimensionality reduction techniques, such as t-SNE, can construct informative visualizations of high-dimensional data. When working with multiple data sets, a straightforward application of these methods often fails; instead of revealing underlying classes, the resulting visualizations expose data set-specific clusters. To circumvent these batch effects, we propose an embedding procedure that uses a t-SNE visualization constructed on a reference data set as a scaffold for embedding new data points. Each data instance in the secondary data is embedded independently, and does not change the reference embedding. This prevents any interactions between instances in the secondary data and implicitly mitigates batch effects. We demonstrate the utility of this approach by analyzing six recently published single-cell gene expression data sets with up to tens of thousands of cells and thousands of genes. The batch effects in our studies are particularly strong as the data comes from different institutions and was obtained using different experimental protocols. The visualizations constructed by our proposed approach are cleared of batch effects, and the cells from secondary data sets correctly co-cluster with cells of the same type from the primary data.

Time: 28th October, 16:40-17:00

Deep Triplet-Driven Semi-Supervised Embedding Clustering

Dino Ienco and Ruggero G. Pensa

In most real world scenarios, experts dispose of limited background knowledge that they can exploit for guiding the analysis process. In this context, semi-supervised clustering can be employed to leverage such knowledge and enable the discovery of clusters that meet the analysts' expectations. To this end, we propose a semi-supervised deep embedding clustering algorithm that exploits triplet constraints as background knowledge within the whole learning process. The latter consists in a two-stage approach where, initially, a low-dimensional data embedding is computed and, successively, cluster assignment is refined via the introduction of an auxiliary target distribution. Our algorithm is evaluated on real-world benchmarks in comparison with state-of-the-art unsupervised and semi-supervised clustering methods. Experimental results highlight the quality of the proposed framework as well as the added value of the new learnt data representation.

Pattern Discovery, Time Series

Chair: Martin Atzmueller

Time: 29th October, 10:30-10:50

Layered Learning for Early Anomaly Detection: Predicting Critical Health Episodes

Vitor Cerqueira, Luis Torgo and Carlos Soares

Critical health events represent a relevant cause of mortality in intensive care units of hospitals, and their timely prediction has been gaining increasing attention. This problem is an instance of the more general predictive task of early anomaly detection in time series data. One of the most common approaches to solve this problem is to use standard classification methods. In this paper we propose a novel method that uses a layered learning architecture to solve early anomaly detection problems. One key contribution of our work is the idea of pre-conditional events, which denote arbitrary but computable relaxed versions of the event of interest. We leverage this idea to break the original problem into two layers, which we hypothesize are easier to solve. Focusing on critical health episodes, the results suggest that the proposed approach is advantageous relative to state of the art approaches for early anomaly detection. Although we focus on a particular case study, the proposed method is generalizable to other domains.

Time: 29th October, 10:50-11:10

Predicting thermal power consumption of the Mars Express satellite with data stream mining

Bozhidar Stevanoski, Dragi Kocev, Aljaž Osojnik, Ivica Dimitrovski and Saso Dzeroski

Orbiting Mars, the European Space Agency (ESA) operated spacecraft - Mars Express (MEX), provides extraordinary science data for the past 15 years. To continue the great contribution, MEX requires accurate power modeling, mainly to compensate for aging and battery degradation. The only unknown variable in the power budget is the power provided to the autonomous thermal subsystem, which in a challenging environment, keeps all equipment under its operating temperature. In this paper, we address the task of predicting the thermal power consumption (TPC) of MEX on all 33 thermal power lines, having available the stream of its telemetry data. Considering the problem definition, we face the task of multi-target regression, learning from data streams. To analyze such data streams, we use the incremental Structured Output Prediction tree (iSOUP-Tree) and the Adaptive Model Rules from High Speed Data Streams (AMRules) to model the

power consumption. The evaluation aims to investigate the potential of the methods for learning from data streams for the task of predicting satellite power consumption and the influence of the time resolution of the measurements of thermal power consumption on the performance of the methods.

Time: 29th October, 11:10-11:25

Ensemble Clustering For Novelty Detection In Data Streams

Kemilly Dearo Garcia, Elaine Ribeiro de Faria, Cláudio Rebelo de Sá, João Mendes-Moreira, Charu C. Aggarwal, André C.P.L.F de Carvalho and Joost N. Kok

In data streams new classes can appear over time due to changes in the data statistical distribution. Consequently, models can become outdated, which requires the use of incremental learning algorithms capable of detecting and learning the changes over time. However, when a single classification model is used for novelty detection, there is a risk that its bias may not be suitable for new data distributions. A solution could be the combination of several models into an ensemble. Besides, because models can only be updated when labeled data arrives, we propose two unsupervised ensemble approaches: one combining clustering partitions using the same clustering technique; and other using different clustering techniques. We compare the performance of the proposed methods with well-known novelty detection algorithms. The methods were tested on datasets commonly used in the novelty detection literature. The experimental results show that proposed ensembles have competitive performance for novelty detection in data streams.

Time: 29th October, 11:25-11:40

Adaptive Long-term Ensemble Learning from Multiple High-dimensional Time-series

Samaneh Khoshrou and Mykola Pechenizkiy

Learning from multiple time-series over an unbounded time-frame has received less attention despite the key applications (such as video analysis, home-assisted) generating this data. Inspired by never-ending approaches, this paper presents an algorithm to continuously learn from multiple high-dimensional un-regulated time-series, in a framework based on ensembles which with respect to drift level develops over time in order to reflect the latest concepts. Here, we explicitly look into video surveillance problem as one of the main sources of high-dimensional data in daily life and extensive experiments are conducted on multiple datasets that demonstrate the advantages of the proposed framework in terms of accuracy and complexity over several baseline approaches.

Time: 29th October, 11:40-11:55

Utilizing Hierarchies in Tree-based Online Structured Output Prediction

Aljaž Osojnik, Pance Panov and Saso Dzeroski

Methods for online prediction of structured values are becoming more and more popular. However, hierarchical prediction, which has recently been shown to produce good results in terms of predictive performance in the batch learning setting, has not yet been applied in the online learning setting. We address the recently introduced task of hierarchical multi-target regression. To this end, we propose a hierarchical extension of iSOUP-Tree, which can address online multi-target regression. The extension weighs the split evaluation heuristic according to the location of the targets in the hierarchy. We design the experimental setup to ascertain whether the additional information contained in the hierarchy can be utilized to improve the predictive performance in the leaf targets. The proposed method shows promising results, producing potential improvements that should be investigated further.

Advanced Machine Learning II, Interpretable Machine Learning II

Chair: Luis Torgo

BEST PAPER II

Time: 29th October, 13:45-14:15

Sparse Robust Regression for Explaining Classifiers

Anton Björklund, Andreas Henelius, Emilia Oikarinen, Kimmo Kallonen and Kai Puolamäki

Real-world datasets are often characterized by outliers, points far from the majority of the points, which might negatively influence modelling of the data. In data analysis it is hence important to use methods that are robust to outliers. In this paper we develop a robust regression method for finding the largest subset in the data that can be approximated using a sparse linear model to a given precision. We show that the problem is NP-hard and hard to approximate. We present an efficient algorithm, termed SLISE, to find solutions to the problem. Our method extends current state-of-the-art robust regression

methods, especially in terms of scalability on large datasets. Furthermore, we show that our method can be used to yield interpretable explanations for individual decisions by opaque, black box, classifiers. Our approach solves shortcomings in other recent explanation methods by not requiring sampling of new data points and by being usable without modifications across various data domains. We demonstrate our method using both synthetic and real-world regression and classification problems.

Time: 29th October, 14:15-14:35

Ensemble-Based Feature Ranking for Semi-supervised Classification

Matej Petković, Saso Dzeroski and Dragi Kocev

In this paper, we propose three feature ranking scores (Symbolic, Genie3, and Random Forest) for the task of semi-supervised classification. In this task, there are only a few labeled examples in a dataset and many unlabeled. This is a highly relevant task, since it is increasingly easy to obtain unlabeled examples, while obtaining labeled examples is often an expensive and tedious task. Each of the proposed feature ranking scores can be computed by using any of three approaches to learning predictive clustering tree ensembles (bagging, random forests, and extra trees). We extensively evaluate the proposed scores on 8 benchmark datasets. The evaluation finds the most suitable ensemble method for each of the scores, shows that taking into account unlabeled examples improves the quality of a feature ranking, and demonstrates that the proposed feature ranking scores outperform a state-of-the-art semi-supervised feature ranking method SEFR. Finally, we identify the best performing pair of a feature ranking score and an ensemble method.

Time: 29th October, 14:35-14:50

Feature Selection for Analogy-Based Learning to Rank

Mohsen Ahmadi Fahandar and Eyke Hüllermeier

Learning to rank based on principles of analogical reasoning has recently been proposed as a novel method in the realm of preference learning. Roughly speaking, the method proceeds from a regularity assumption as follows: Given objects A, B, C, D, if A relates to B as C relates to D, and A is preferred to B, then C is presumably preferred to D. This assumption is formalized in terms of so-called analogical proportions, which operate on a feature representation of the objects. Consequently, a suitable feature representation is an important prerequisite for the success of analogy-based learning to rank. In this paper, we therefore address the problem of feature selection and adapt common feature selection techniques, including forward selection, correlation-based filter techniques, as

well as Relief-based methods, to the case of analogical learning. The usefulness of these approaches is shown in experiments with synthetic and benchmark data.

Time: 29th October, 14:50-15:05

Variance-based Feature Importance in Neural Networks

Cláudio Rebelo de Sá

This paper proposes a new method to measure the relative importance of features in Artificial Neural Networks (ANN) models. Its underlying principle assumes that the more important a feature is, the more the weights, connected to the respective input neuron, will change during the training of the model. To capture this behavior, a running variance of every weight connected to the input layer is measured during training. For that, an adaptation of Welford's online algorithm for computing the online variance is proposed. When the training is finished, for each input, the variances of the weights are combined with the final weights to obtain the measure of relative importance for each feature. This method was tested with shallow and deep neural network architectures on several well-known classification and regression problems. The results obtained confirm that this approach is making meaningful measurements. Moreover, results showed that the importance scores are highly correlated with the variable importance method from Random Forests (RF).

Time: 29th October, 15:05-15:20

A Unified Approach to Biclustering Based on Formal Concept Analysis and Interval Pattern Structure

Nyoman Juniarta, Miguel Couceiro and Amedeo Napoli

In a matrix representing a numerical dataset, a bicluster is a submatrix whose cells exhibit similar behavior. Biclustering is naturally related to Formal Concept Analysis (FCA) where concepts correspond to maximal and closed biclusters in a binary dataset. In this paper, a unified characterization of biclustering algorithms is proposed using FCA and pattern structures, an extension of FCA for dealing with numbers and other complex data. Several types of biclusters - constant-column, constant-row, additive, and multiplicative - and their relation to interval pattern structures is presented.

Time: 29th October, 15:20-15:35

A Sampling-based Approach for Discovering Subspace Clusters

Sandy Moens, Boris Cule and Bart Goethals

Subspace clustering aims to discover clusters in projections of highly dimensional numerical data. In this paper, we focus on discovering small collections of interesting subspace clusters that do not try to cluster all data points, leaving noisy data points unclustered. To this end, we propose a randomised method that first converts the highly dimensional database to a binarised one using projected samples of the original database. This database is then mined for frequent itemsets, which we show can be translated back to subspace clusters. In our extensive experimental analysis, we show on synthetic as well as real world data that our method is capable of discovering highly interesting subspace clusters.

Networks

Chair: Boris Cule

Time: 29th October, 16:00-16:20

A Combinatorial Multi-Armed Bandit based method for Dynamic Consensus Community Detection in Temporal Networks

Domenico Mandaglio and Andrea Tagarelli

Community detection in temporal networks is an active field of research, which can be leveraged for several strategic decisions, including enhanced group-recommendation, user behavior prediction, and evolution of user interaction patterns in relation to real-world events. Recent research has shown that combinatorial multi-armed bandit (CMAB) is a suitable methodology to address the problem of dynamic consensus community detection (DCCD), i.e., to compute a single community structure that is conceived to be representative of the knowledge available from community structures observed at the different time steps. In this paper, we propose a CMAB-based method, called MYLAGO, to solve the DCCD problem. Unlike existing approaches, our algorithm is designed to provide a solution, i.e., dynamic consensus community structure that embeds both long-term changes in the community formation and newly observed community structures. Experimental evaluation based on publicly available real-world and ground-truth-oriented synthetic networks, with different structure and evolution rate, has confirmed the meaningfulness and key benefits of the proposed method, also against competitors based on evolutionary or consensus approaches.

Time: 29th October, 16:20-16:40

Resampling-based Framework for Unbiased Estimator of Node Centrality over Large Complex Network

Kazumi Saito, Kouzou Ohara, Masahiro Kimura and Hiroshi Motoda

We address a problem of efficiently estimating value of a centrality measure for a node in a large network, and propose a sampling-based framework in which only a small number of nodes that are randomly selected are used to estimate the measure. The error estimator we derived is an unbiased estimator of the approximation error defined as the expectation of the difference between the true and the estimated values of the centrality. We experimentally evaluate the fundamental performance of the proposed framework using the closeness and betweenness centralities on six real world networks from different domains, and show that it allows us to estimate the approximation error more tightly and more precisely than the standard error estimator traditionally used based on i.i.d. sampling, i.e., with the confidence level of 95% for a small number of sampling, say 20% of the total number of nodes.

Time: 29th October, 16:40-17:00

Efficient and Accurate Non-exhaustive Pattern-based Change Detection in Dynamic Networks

Angelo Impedovo, Michelangelo Ceci and Toon Calders

Pattern-based change detectors (PBCDs) are non-parametric unsupervised change detection methods that are based on observed changes in sets of frequent patterns over time. In this paper we study PBCDs for dynamic networks; that is, graphs that change over time, represented as a stream of snapshots. Accurate PBCDs rely on exhaustively mining sets of patterns on which a change detection step is performed. Exhaustive mining, however, has worst case exponential time complexity, rendering this class of algorithms inefficient in practice. Therefore, in this paper we propose non-exhaustive PBCDs for dynamic networks. The algorithm we propose prunes the search space following a beam-search approach. The results obtained on real-world and synthetic dynamic networks, show that this approach is surprisingly effective in both increasing the efficiency of the mining step as in achieving higher detection accuracy, compared with state-of-the-art approaches.

Time: 29th October, 17:00-17:15

Evolving social networks analysis via tensor decompositions: from global event detection towards local pattern discovery and specification

Sofia Fernandes, Hadi Fanaee-T and Joao Gama

Existing approaches for detecting anomalous events in time-evolving networks usually focus on detecting events involving the majority of the nodes, which affect the overall structure of the network. Since events involving just a small subset of nodes usually do not affect the overall structure of the network, they are more difficult to spot. In this context, tensor decomposition based methods usually beat other techniques in detecting global events, but fail at spotting localized event patterns. We tackle this problem by replacing the batch decomposition with a sliding window decomposition, which is further mined in an unsupervised way using statistical tools. Via experimental results in one synthetic and four real-world networks, we show the potential of the proposed method in the detection and specification of local events.

Time: 29th October, 17:15-17:30

Mining a maximum weighted set of disjoint submatrices

Vincent Branders, Guillaume Derval, Pierre Schaus and Pierre Dupont

The objective of the maximum weighted set of disjoint submatrices problem is to discover K disjoint submatrices that together cover the largest sum of entries of an input matrix. It has many practical data-mining applications, as the related biclustering problem, such as gene module discovery in bioinformatics. It differs from the maximum-weighted submatrix coverage problem by the explicit formulation of disjunction constraints: submatrices must not overlap. In other words, all matrix entries must be covered by at most one submatrix. The particular case of K=1, called the maximal-sum submatrix problem, was successfully tackled with constraint programming. Unfortunately, the case of K > 1 is more challenging to solve as the selection of rows cannot be decided in polynomial time solely from the selection of K sets of columns. It can be proved to be NP-hard. We introduce a hybrid column generation approach using constraint programming to generate columns. It is compared to a standard mixed integer linear programming (MILP) through experiments on synthetic datasets. Overall, fast and valuable solutions are found by column generation while the MILP approach cannot handle a large number of variables and constraints.

Applications

Chair: Larisa Soldatova

Time: 30th October, 10:30-10:50

KnowBots: Discovering Relevant Patterns in Chatbot Dialogues

Adriano Rivolli, Catarina Amaral, Luis Guardão, Cláudio Rebelo de Sá and Carlos Soares

Chatbots have been used in business contexts as a new way of communicating with customers. They use natural language to interact with the customers, whether while offering products and services, or in the support of a specific task. In this context, an important and challenging task is to assess the effectiveness of the machine-to-human interaction, according to business' goals. Although several analytic tools have been proposed to analyze the user interactions with chatbot systems, to the best of our knowledge they do not consider user-defined criteria, focusing on metrics of engagement and retention of the system as a whole. For this reason, we propose the KnowBots tool, which can be used to discover relevant patterns in the dialogues of chatbots, by considering specific business goals. Given the non-trivial structure of dialogues and the possibly large number of conversational records, we combined sequential pattern mining and subgroup discovery techniques to identify patterns of usage. Moreover, a friendly user-interface was developed to present the results and to allow their detailed analysis. Thus, it may serve as an alternative decision support tool for business or any entity that makes use of this type of interactions with their clients.

Time: 30th October, 10:50-11:05

Temporal analysis of adverse weather conditions affecting wheat production in Finland

Vladimir Kuzmanovski, Mika Sulkava, Taru Palosuo and Jaakko Hollmen

Growing conditions of agricultural crops are increasingly affected by global climate change. Not only the overall agro-climatic conditions are changing, but also climatic variability and the occurrence of extreme weather events are becoming more frequent. This will affect crop yields and impact food supply both locally and globally. Located in the north, with short growing seasons and long days, Finland is not an exception. Drought- and temperature-related adverse events have been identified as most harmful abiotic factors on the production. Farmers try to mitigate with a range of management options. However, they need to adapt them over time as the climate is changing. This study aims

to identify the most adverse weather events that affect the spring wheat production in Finland and to ascertain if there have been changes on the most harmful abiotic weather-related factors during the last decades. Adverse weather conditions studied include frequency and length of periods with exceptional snow, drought, intensive rainfall and extreme heat. This was studied by modeling the wheat production using the adverse weather events as predictors with different lengths of training period (consecutive number of years) using LASSO regression. The results reveal clear shift from early season drought and periodical intensive rainfall to the adverse effects of frequent and long periods of extremely high temperatures during later development stages.

Time: 30th October, 11:05-11:20

Main Factors Driving the Open Rate of Email Marketing Campaigns

Andreia Conceição and João Gama

Email Marketing is one of the most important traffic sources in Digital Marketing. It yields a high return on investment for the company and offers a cheap and fast way to reach existent or potential clients. Getting the recipients to open the email is the first step for a successful campaign. Thus, it is important to understand how marketers can improve the open rate of a marketing campaign. In this work, we analyze what are the main factors driving the open rate of financial email marketing campaigns. For that purpose, we develop a classification algorithm that can accurately predict if a campaign will be labeled as Successful or Failure. A campaign is classified as Successful if it has an open rate higher than the average, otherwise it is labeled as Failure. To achieve this, we have employed and evaluated three different classifiers. Our results showed that it is possible to predict the performance of a campaign with approximately 82% accuracy, by using the Random Forest algorithm and the redundant filter selection technique. With this model, marketers will have the chance to sooner correct potential problems in a campaign that could highly impact its revenue. Additionally, a text analysis of the subject line and preheader was performed to discover which keywords and keyword combinations trigger a higher open rate. The results obtained were then validated in a real setting through A/B testing.

Time: 30th October, 11:20-11:35

Enhancing BMI-Based Student Clustering by Considering Fitness as Key Attribute

Erik Dovgan, Bojan Leskošek, Gregor Jurak, Gregor Starc, Maroje Sorić and Mitja Luštrek

The purpose of this study was to redefine health and fitness categories of students, which were defined based on body mass index (BMI). BMI enables identifying overweight and obese persons, however, it inappropriately classifies overweight-and-fit and normal-

weight-and-non-fit persons. Such a classification is required when personalized advice on healthy life style and exercises is provided to students. To overcome this issue, we introduced a clustering-based approach that takes into account a fitness score of students. This approach identifies fit and not-fit students, and in combination with BMI, students that are overweight-and-fit and those that are normal-weight-and-non-fit. These results enable us to better target students with personalized advice based on their actual physical characteristics.

Time: 30th October, 11:35-11:50

Mining Patterns in Source Code using Tree Mining Algorithms

Hoang Son Pham, Siegfried Nijssen, Kim Mens, Dario Di Nucci, Tim Molderez, Coen De Roover, Johan Fabry and Vadim Zaytsev

Discovering regularities in source code is of great interest to software engineers, both in academia and in industry, as regularities can provide useful information to help in a variety of tasks such as code comprehension, code refactoring, and fault localization. However, traditional pattern mining algorithms often find too many patterns of little use and hence are not suitable for discovering useful regularities. In this paper we propose EFT, a new algorithm for mining patterns in source code based on the FT tree mining algorithm. First, we introduce several constraints that effectively enable us to find more useful patterns; then, we show how to efficiently include them in FT. To illustrate the usefulness of the constraints we carried out a case study in collaboration with software engineers, where we identified a number of interesting patterns in a repository of Java code.

Time Series, Applications

Chair: Joao Gama

Time: 30th October, 13:45-14:05

Hyperparameter Importance for Image Classification by Residual Networks

Abhina Sharma, Jan N. van Rijn, Frank Hutter and Andreas Mueller

Residual neural networks (ResNets) are among the state-of-the-art for image classification tasks. With the advent of automated machine learning (AutoML), automated hyperparameter optimization methods are by now routinely used for tuning various network types. However, in the thriving field of deep neural networks, this progress is not yet matched by equal progress on rigorous techniques that yield information beyond performance-optimizing hyperparameter settings. In this work, we aim to answer the following question: Given a residual neural network architecture, what are generally (across datasets) its most important hyperparameters? In order to answer this question, we assembled a benchmark suite containing 10 image classification datasets. For each of these datasets, we analyze which of the hyperparameters were most influential using the functional ANOVA framework. This experiment both confirmed expected patterns, and revealed new insights. With these experimental results, we aim to form a more rigorous basis for experimentation that leads to better insight towards what hyperparameters are important to make neural networks perform well.

Time: 30th October, 14:05-14:25

Cellular Traffic Prediction and Classification: a comparative evaluation of LSTM and ARIMA

Amin Azari, Panagiotis Papapetrou, Stojan Denic and Gunnar Peters

Prediction of user traffic in cellular networks has attracted profound attention for improving the reliability and efficiency of network resource utilization. In this paper, we study the problem of cellular network traffic prediction and classification by employing standard machine learning and statistical learning time series prediction methods, including long short-term memory (LSTM) and autoregressive integrated moving average (ARIMA), respectively. We present an extensive experimental evaluation of the designed tools over a real network traffic dataset. Within this analysis, we explore the impact of different parameters on the effectiveness of the predictions. We further extend our analysis to the problem of network traffic classification and prediction of traffic bursts. The results, on the one hand, demonstrate the superior performance of LSTM over

ARIMA in general, especially when the length of the training dataset is large enough and its granularity is fine enough. On the other hand, the results shed light onto the circumstances in which, ARIMA performs close to the optimal with lower complexity.

Time: 30th October, 14:25-14:45

Fast Distance-based Anomaly Detection in Images Using an Inception-like Autoencoder

Natasa Sarafijanovic-Djukic and Jesse Davis

The goal of anomaly detection is to identify examples that deviate from normal or expected behavior. We tackle this problem for images. We consider a two-phase approach. First, using normal examples, a convolutional autoencoder (CAE) is trained to extract a low-dimensional representation of the images. Here, we propose a novel architectural choice when designing the CAE, an Inception-like CAE. It combines convolutional filters of different kernel sizes and it uses a Global Average Pooling (GAP) operation to extract the representations from the CAE's bottleneck layer. Second, we employ a distanced-based anomaly detector in the low-dimensional space of the learned representation for the images. However, instead of computing the exact distance, we compute an approximate distance using product quantization. This alleviates the high memory and prediction time costs of distance-based anomaly detectors. We compare our proposed approach to a number of baselines and state-of-the-art methods on four image datasets, and we find that our approach resulted in improved predictive performance.

Time: 30th October, 14:45-15:00

Integrating LSTMs with Online Density Estimation for the Probabilistic Forecast of Energy Consumption

Julian Vexler and Stefan Kramer

In machine learning applications in the energy sector, it is often necessary to have both highly accurate predictions and information about the probabilities of certain scenarios to occur. We address this challenge by integrating and combining long short-term memory networks (LSTMs) and online density estimation into a real-time data streaming architecture of an energy trader. The online density estimation is done in the MiDEO framework, which estimates joint densities of data streams based on ensembles of chains of Hoeffding trees. One attractive feature of the solution is that queries can be sent to the here-called forecast-based point density estimators (FPDE) to derive information from a compact representation of two data streams, leading to a new perspective to the problem. The experiments indicate promising application possibilities of FPDE, including but not limited to the estimation of uncertainties, early model evaluation and the

simulation of alternative scenarios.

Time: 30th October, 15:00-15:15

Fourier-based Parametrization of CNNs for Robust Time Series Forecasting

Sascha Krstanovic and Heiko Paulheim

Classical statistical models for time series forecasting most often make a number of assumptions about the data at hand, therewith, requiring intensive manual preprocessing steps prior to modeling. As a consequence, it is very challenging to come up with a more generic forecasting framework. Extensive hyperparameter optimization and ensemble architectures are common strategies to tackle this problem, however, this comes at the cost of high computational complexity. Instead of optimizing hyperparameters by training multiple models, we propose a method to estimate optimal hyperparameters directly from the characteristics of the time series at hand. To that end, we use Convolutional Neural Networks (CNNs) for time series forecasting and determine a part of the network layout based on the time series' Fourier coefficients. Our approach significantly reduces the amount of required model configuration time and shows competitive performance on time series data across various domains. A comparison to popular, state of the art forecasting algorithms reveals further improvements in runtime and practicability.

Time: 30th October, 15:15-15:30

On Recognizing Cats and Dogs in Chinese Paintings

Qianqian Gu and Ross King

Although Deep Learning (DL) image analysis has made recent rapid advances, it still has limitations that indicate that its approach differs significantly from human vision, e.g. the requirement for large training sets, and adversarial attacks. Here we show that DL also differs in failing to generalize well to Traditional Chinese Paintings (TCPs). We developed a new DL object detection method A-RPN (Assembled Region Proposal Network), which concatenates low-level visual information, and high-level semantic knowledge to reduce coarseness in region-based object detection. A-RPN significantly outperforms YOLO2 and Faster R-CNN on natural images ($P < 0.02$). We applied YOLO2, Faster R-CNN and A-RPN to TCPs with a 12.9%, 13.2% and 13.4% drop in mAP compared to natural images. There was little or no difference in recognizing humans, but a large drop in mAP for cats and dogs (27% & 31%), and very large drop for horses (35.9%). The abstract nature of TCPs may be responsible for DL poor performance.

PhD SYMPOSIUM

Time: 28th October, 17:10 – 19:00

Big Data Storage in High Energy Physics
Petra Lončar

Acquisition of Mathematical Knowledge with the Use of a Tangible User Interface
Lea Dujić Rodić and Andrina Granić

Dimensionality Reduction Techniques within Behavioral Analysis
Demijan Grgić and Vedran Podobnik

Data-driven Spatio-temporal Traffic Anomaly Detection on the Urban Road Networks
Leo Tišljarić and Tonči Carić

Inferring Default Cascades in Financial Systems.
Irena Barjašić and Vinko Zlatić

New Approaches for Elucidation of Drug-Target Binding Affinities
Davor Oršolić, Bono Lučić, Višnja Stepanić and Tomislav Šmuc

Late Breaking Contributions

Time: 28th October, 17:10 – 19:00

Predicting Associations Between Proteins and Multiple Diseases
Martin Breskvar and Sašo Džeroski

Gene Function Prediction using Gene Ontology Decomposition
Vedrana Vidulin and Sašo Džeroski

Quantifying the Effects of Gyroless Flying of the Mars Express Spacecraft with Machine Learning
Matej Petković, Luke Lucas, Dragi Kocev, Saso Dzeroski, Redouane Boumghar and Nikola Simidjievski

Rules and Redescriptions as Features in Binary Classification Tasks
Matej Mihelčić and Tomislav Šmuc

Significance of Patterns in Data Visualisations
Rafael Savvides

Translation Between Waves, wave2wave
Tsuyoshi Okita, Hirotaka Hachiya, Sozo Inoue and Naonori Ueda

Machines in a Classroom: Towards Human-like Active Learning
Ilona Kulikovskikh and Tomislav Šmuc

~~***Pajé: End-to-End Data Science***~~
~~*Edesio Alcobaça, Davi Pereira-Santos, Moises R. dos Santos, Gean T. Pereira, Rafael G. Mantovani, Luis P. F. Garcia, Saulo M. Mastelini and Andre de Carvalho*~~

Social Program

Welcome reception will be held at Radison Blu Hotel, 28th October at 20:00.

City Tour: A guided tour of the old city center (Diocletian's palace) is planned on 29th October. Participants will be informed on the details before the conference sessions on day of the tour.

Conference dinner will be immediately after the City Tour, in the restaurant Adriatic Grašo, a 15 minutes walk from the city center, on 29th October at 20:00.

Organization

General Chair
Sašo Džeroski - Jožef Stefan Institute

Program Committee Chairs
Petra Kralj Novak - Jožef Stefan Institute
Tomislav Šmuc - Ruđer Bošković Institute

PhD Symposium Chair
Tomislav Lipić - Ruđer Bošković Institute

Proceedings Chair
Matija Piškorec - Ruđer Bošković Institute

Web and Social Media Chairs
Ratko Mileta - Ruđer Bošković Institute
Matija Piškorec - Ruđer Bošković Institute

Local Arrangements Chairs
Ana Vidoš - Ruđer Bošković Institute
Matija Piškorec - Ruđer Bošković Institut

Program Committee

Annalisa Appice - University of Bari Aldo Moro, Italy
Martin Atzmueller - Tilburg University, The Netherlands
Viktor Bengs - Paderborn University, Germany
Concha Bielza Lozoya - Universidad Politecnica de Madrid, Spain
Albert Bifet - LTCI, Telecom ParisTech, France
Alberto Cano - Virginia Commonwealth University, USA
Michelangelo Ceci - University of Bari Aldo Moro, Italy
Bruno Cremilleux - University of Caen Normandy, France
Claudia d'Amato - University of Bari Aldo Moro, Italy
Nicola Di Mauro - University of Bari Aldo Moro, Italy
Ivica Dimitrovski - Ss. Cyril and Methodius University in Skopje, North Macedonia
Wouter Duivesteijn - Eindhoven University of Technology, The Netherlands
Lina Fahed - IMT Atlantique, France
Hadi Fanaee - University of Oslo, Norway
Nicola Fanizzi - University of Bari Aldo Moro, Italy
Stefano Ferilli - University of Bari Aldo Moro, Italy
Johannes Fürnkranz - Technische Universitat Darmstadt, Germany
Mohamed Gaber - Birmingham City University, United Kingdom
João Gama - University of Porto, Portugal
Dragan Gamberger - Rudjer Bošković Institute, Croatia
Makoto Haraguchi - Hokkaido University, Japan
Kouichi Hirata - Kyushu Institute of Technology, Japan
Jaakko Hollmen - Aalto University, Finland
Eyke Huellermeier - Paderborn University, Germany
Alipio Jorge - University of Porto, Portugal
Masahiro Kimura - Ryukoku University, Japan
Dragi Kocev - Jozef Stefan Institute, Slovenia
Stefan Kramer - Johannes Gutenberg University Mainz, Germany
Vincenzo Lagani - Ilia State University, Georgia
Pedro Larranaga - University of Madrid, Spain
Nada Lavrač - Jozef Stefan Institute, Slovenia
Jurica Levatić - Institute for Research in Biomedicine, Spain
Tomislav Lipić - Rudjer Bošković Institute, Croatia
Francesca Alessandra Lisi - University of Bari Aldo Moro, Italy
Gjorgji Madjarov - Ss. Cyril and Methodius University in Skopje, North Macedonia
Giuseppe Manco - Institute for high performance computing and networking, Italy

Sanda Martinčić-Ipšić - University of Rijeka, Croatia
Elio Masciari - Institute for high performance computing and networking, Italy
Anna Monreale - University of Pisa, Italy
Siegfried Nijssen - Universite catholique de Louvain, Belgium
Rita P. Ribeiro - University of Porto, Portugal
Panče Panov - Jozef Stefan Institute, Slovenia
Ruggero G. Pensa - University of Torino, Italy
Bernhard Pfahringer - University of Waikato, New Zealand
Gianvito Pio - University of Bari Aldo Moro, Italy
Pascal Poncelet - LIRMM Montpellier, France
Jan Ramon - French research institute for digital sciences, France
Chedy Raissi - French research institute for digital sciences, France
Marko Robnik-Šikonja - University of Ljubljana, Slovenia
Kazumi Saito - University of Shizuoka, Japan
Marina Sokolova - University of Ottawa and Institute for Big Data Analytics, Canada
Jerzy Stefanowski - Poznan University of Technology, Poland
Ljupčo Todorovski - University of Ljubljana, Slovenia
Luis Torgo - Dalhousie University, Canada
Herna Viktor - University of Ottawa, Canada
Albrecht Zimmermann - Universite Caen Normandie, France
Blaž Zupan - University of Ljubljana, Slovenia

Partners and sponsors

Ruđer Bošković Institute, Zagreb, Croatia

Croatian Center of Excellence in Data Science and Advanced Cooperative Systems, Croatia

Jožef Stefan Institute, Ljubljana, Slovenia

Springer Nature

Machine Learning journal

Croatian Center of Excellence in Data Science and Advanced Cooperative Systems

Newton Technologies Adria, Zagreb, Croatia

Croatian Society for Information and Communication Technology, Electronics and Microelectronics - MIPRO.

Notes

Notes

Notes

DAY 1
Monday, 28th October

8:00-18:00	**Registration**
8:45-9:00	**Conference opening**
9:00-10:00	Keynote talk 1: Dino Pedreschi
10:00-10:30	Coffee Break
10:30-12:00	**SESSION 1: Advanced machine learning I**
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)
13:45-15:15	**SESSION 2: Interpretable Machine Learning**
	13:45-14:15 BEST PAPER I
15:15-15:45	Coffee Break and Poster Exhibition
15:45-17:10	**SESSION 3: Data and Knowledge Representation**
17:10-19:00	PhD Symposium & Late Breaking Contributions
19:10-20:00	Steering Committee Meeting
20:00-22:00	Welcome Reception at Radison Blue Terrace

DAY 2
Tuesday, 29th October

8:00-18:00	**Registration**
8:55-9:00	**Daily announcements**
9:00-10:00	Keynote talk 2: Guido Caldarelli
10:00-10:30	Coffee Break and Poster Exhibition
10:30-11:55	**SESSION 4: Pattern Discovery, Time Series**
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)
13:45-15:15	**SESSION 5: Advanced Machine Learning II; Interpretable Machine Learing II**
	13:45-14:15 BEST PAPER II
15:35-16:00	Coffee Break and Poster Exhibition
16:00-17:30	**SESSION 6: Networks**
17:30-18:15	Community meeting
18:30-19:45	City Tour (Diocletian Palace)
20:00-22:00	Gala Dinner

DAY 3
Wednesday, 30th October

8:00-18:00	**Registration**
8:55-9:00	**Daily announcements**
9:00-10:00	Keynote talk 3: Marinka Žitnik
10:00-10:30	Coffee Break
10:30-12:00	**SESSION 7: Applications**
12:00-13:45	Lunch (Hotel Radison Blue Restaurant)
15:15-15:30	**SESSION 8: Time Series, Applications**
15:30-15:45	Closing of the Conference