2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA 2018)

Turin, Italy
1-3 October 2018

IEEE Catalog Number: CFP18DSB-POD
ISBN: 978-1-5386-5091-2

Copyright © 2018 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number: CFP18DSB-POD
ISBN (Print-On-Demand): 978-1-5386-5091-2
ISBN (Online): 978-1-5386-5090-5

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

2018 IEEE 5th International Conference on Data Science and Advanced Analytics

DSAA 2018

Table of Contents

Message from the DSAA 2018 General Co-Chairs ..... **xiv**
Message from the DSAA 2018 Program Co-Chairs ..... **xvi**
DSAA 2018 Organizing Committee ..... **xvii**
DSAA 2018 Program Committee ..... **xix**
DSAA 2018 Keynotes ..... **xxiv**

Research Track

Estimating Heterogeneous Causal Effects in the Presence of Irregular Assignment Mechanisms 1
Falco Joannes Bargagli Stoffi (IMT School for Advanced Studies) and Giorgio Gnecco (IMT School for Advanced Studies)

Citizen Contributions and Minor Heritage: Feedback on Modeling and Visualising an Information Mash-up 11
Jean-Yves Blaise (UMR CNRS/MC 3495 MAP), Iwona Dudek (UMR CNRS/MC 3495 MAP), and Gamze Saygi (UMR CNRS/MC 3495 MAP)

Unification of Deconvolution Algorithms for Cherenkov Astronomy 21
Mirko Bunse (TU Dortmund), Nico Piatkowski (TU Dortmund), Katharina Morik (TU Dortmund), Tim Ruhe (TU Dortmund), and Wolfgang Rhode (TU Dortmund)

Recommendation of Points-of-Interest Using Graph Embeddings 31
Giannis Christoforidis (Aristotle University), Pavlos Kefalas (Aristotle University), Apostolos Papadopoulos (Aristotle University), and Yannis Manolopoulos (Aristotle University)

Adaptive Threshold for Outlier Detection on Data Streams 41
James Clark (Lockheed Martin), Zhen Liu (Guangdon Pharmaceutical University), and Nathalie Japkowicz (American University)

Neuro-Ensemble for Time Series Data Classification 50
Soukaina Filali Boubrahimi (Georgia State University), Ruizhe Ma (Georgia State University), and Rafal Angryk (Georgia State University)

Estimating Causal Effects on Social Networks 60
Laura Forastiere (Yale University), Fabrizia Mealli (University of Florence), Albert Wu (Harvard University), and Edoardo M. Airoldi (Harvard University)

Hoeffding Trees with Nmin Adaptation . 70
Eva García-Martín (Blekinge Institute of Technology), Niklas Lavesson (Blekinge Institute of Technology and Jönköping University), Håkan Grahn (Blekinge Institute of Technology), Emiliano Casalicchio (Blekinge Institute of Technology and Sapienza University of Rome), and Veselka Boeva (Blekinge Institute of Technology)

Explaining Explanations: An Overview of Interpretability of Machine Learning . 80
Leilani H. Gilpin (Massachusetts Institute of Technology), David Bau (Massachusetts Institute of Technology), Ben Z. Yuan (Massachusetts Institute of Technology), Ayesha Bajwa (Massachusetts Institute of Technology), Michael Specter (Massachusetts Institute of Technology), and Lalana Kagal (Massachusetts Institute of Technology)

Multivariate Time Series Early Classification Using Multi-Domain Deep Neural Network . 90
Huai-Shuo Huang (National Chiao Tung University), Chien-Liang Liu (National Chiao Tung University), and Vincent S. Tseng (National Chiao Tung University)

Pattern-Based Automatic Parallelization of Representative-Based Clustering Algorithms . 99
Saiyedul Islam (Birla Institute of Technology and Science Pilani), Sundar Balasubramaniam (Birla Institute of Technology and Science Pilani), Shruti Gupta (Birla Institute of Technology and Science Pilani), Shikhar Brajesh (Birla Institute of Technology and Science Pilani), Rohan Badlani (Birla Institute of Technology and Science Pilani), Nitin Labhishetty (Birla Institute of Technology and Science Pilani), Abhinav Baid (Birla Institute of Technology and Science Pilani), Poonam Goyal (Birla Institute of Technology and Science Pilani), and Navneet Goyal (Birla Institute of Technology and Science Pilani)

Generalized Bayesian Factor Analysis for Integrative Clustering with Applications to Multi-Omics Data . 109
Eun Jeong Min (University of Pennsylvania), Changgee Chang (University of Pennsylvania), and Qi Long (University of Pennsylvania)

Kernel Regression on Manifold Valued Data . 120
Alexander Kuleshov (Skolkovo Institute of Science and Technology), Alexander Bernstein (Skolkovo Institute of Science and Technology), and Evgeny Burnaev (Skolkovo Institute of Science and Technology)

Fully Heterogeneous Collective Regression . 130
David Liedtka (United States Naval Academy) and Luke McDowell (United States Naval Academy)

Towards Efficient Closed Infrequent Itemset Mining Using Bi-Directional Traversing . 140
Yifeng Lu (LMU Munich) and Thomas Seidl (LMU Munich)

SMOTEBoost for Regression: Improving the Prediction of Extreme Values . 150
Nuno Moniz (University of Porto), Rita Ribeiro (University of Porto), Vitor Cerqueira (University of Porto), and Nitesh Chawla (University of Notre Dame)

Automatic Semantic Labelling of Images by Their Content Using Non-Parametric Bayesian Machine Learning and Image Search Using Synthetically Generated Image Collages . 160
Michael Niemeyer (University of St. Andrews) and Ognjen Arandjelovi (University of St. Andrews)

Cohort Representation and Exploration .169
Behrooz Omidvar-Tehrani (University of Grenoble Alpes), Sihem Amer-Yahia (University of Grenoble Alpes), and Laks V.S. Lakshmanan (University of British Columbia)

Predicting Worker Disagreement for More Effective Crowd Labeling .179
Stefan Räbiger (Sabancı University), Gizem Gezici (Sabancı University), Yücel Saygın (Sabancı University), and Myra Spiliopoulou (Otto-von-Guericke University)

Gradient Reversal against Discrimination: A Fair Neural Network Learning Approach .189
Edward Raff (Booz Allen Hamilton) and Jared Sylvester (Booz Allen Hamilton)

Forest of Normalized Trees: Fast and Accurate Density Estimation of Streaming Data .199
Patrick Rehn (Johannes Gutenberg-Universität), Zahra Ahmadi (Johannes Gutenberg-Universität), and Stefan Kramer (Johannes Gutenberg-Universität)

Model Adaptation and Personalization for Physiological Stress Detection .209
Aaqib Saeed (Eindhoven University of Technology), Tanir Ozcelebi (Eindhoven University of Technology), Johan Lukkien (Eindhoven University of Technology), Jan van Erp (University of Twente), and Stojan Trajanovski (Philips Research)

Random Forest Framework Customized to Handle Highly Correlated Variables: An Extensive Experimental Study Applied to Feature Selection in Genetic Data .217
Christine Sinoquet (University of Nantes) and Kamel Mekhnacha (ProbaYes)

Parallel Continuous Outlier Mining in Streaming Data .227
Theodoros Toliopoulos (Aristotle University), Anastasios Gounaris (Aristotle University), Kostas Tsichlas (Aristotle University), Apostolos Papadopoulos (Aristotle University), and Sandra Sampaio (University of Manchester)

Learning to Make Predictions on Graphs with Autoencoders .237
Phi Vu Tran (Booz Allen Hamilton)

Entity-Level Stream Classification: Exploiting Entity Similarity to Label the Future Observations Referring to an Entity .246
Christian Beyer (Otto-von-Guericke University Magdeburg), Vishnu Unnikrishnan (Otto-von-Guericke University Magdeburg), Pawel Matuszyk (Otto-von-Guericke University Magdeburg), Uli Niemann (Otto-von-Guericke University Magdeburg), Rüdiger Pryss (University Ulm), Winfried Schlee (University Hospital Regensburg), Eirini Ntoutsi (Leibniz University Hannover), and Myra Spiliopoulou (Otto-von-Guericke University Magdeburg)

Forecasting Retweet Count during Elections Using Graph Convolution Neural Networks .256
Raghavendran Vijayan (Indiana University-Purdue University Indianapolis) and George Mohler (Indiana University-Purdue University Indianapolis)

Practical Deep Learning Architecture Optimization .263
Martin Wistuba (IBM Research)

DNA: General Deterministic Network Adaptive Framework for Multi-Round Multi-Party Influence Maximization .273
Tzu-Hsin Yang (National Cheng Kung University), Hao-Shang Ma (National Cheng Kung University), and Jen-Wei Huang (National Cheng Kung University)

DeepClean: Data Cleaning via Question Asking .283
Xinyang Zhang (Lehigh University), Yujie Ji (Lehigh University), Chanh Nguyen (Lehigh University), and Ting Wang (Lehigh University)

Applications Track

Towards Bankruptcy Prediction: Deep Sentiment Mining to Detect Financial Distress from Business Management Reports .293
Zahra Ahmadi (Johannes Gutenberg-Universität), Peter Martens (Johannes Gutenberg-Universität), Christopher Koch (Johannes Gutenberg-Universität), Thomas Gottron (SCHUFA Holding AG), and Stefan Kramer (Johannes Gutenberg-Universität)

Wetting and Drying of Soil: From Data to Understandable Models for Prediction .303
Aniruddha Basak (Carnegie Mellon University), Ole J. Mengshoel (Carnegie Mellon University), Kevin Schmidt (U.S. Geological Survey), and Chinmay Kulkarni (Carnegie Mellon University)

Starting Movement Detection of Cyclists Using Smart Devices .313
Maarten Bieshaar (University of Kassel), Malte Depping (University of Kassel), Jan Schneegans (University of Kassel), and Bernhard Sick (University of Kassel)

The Economic Value of Neighborhoods: Predicting Real Estate Prices from the Urban Environment .323
Marco De Nadai (University of Trento and Fondazione Bruno Kessler) and Bruno Lepri (Fondazione Bruno Kessler)

A Multi-Task Deep Learning Architecture for Maritime Surveillance Using AIS Data Streams .331
Duong Nguyen (IMT Atlantique, Lab-STICC, UBL), Rodolphe Vadaine (CLS-Space and Ground Segments), Guillaume Hajduch (CLS-Space and Ground Segments), René Garello (IMT Atlantique, Lab-STICC, UBL), and Ronan Fablet (IMT Atlantique, Lab-STICC, UBL)

Scalable and Interpretable Predictive Models for Electronic Health Records .341
Amela Fejza (University of Grenoble Alpes, CNRS, Inria), Pierre Genevès (University of Grenoble Alpes, CNRS, Inria), Nabil Layaïda (University of Grenoble Alpes, CNRS, Inria), and Jean-Luc Bosson (University of Grenoble Alpes, CNRS)

Mining Sensor Data for Predictive Maintenance in the Automotive Industry .351
Flavio Giobergia (Politecnico di Torino), Elena Baralis (Politecnico di Torino), Maria Camuglia (General Motors Global Propulsion Systems), Tania Cerquitelli (Politecnico di Torino), Marco Mellia (Politecnico di Torino), Alessandra Neri (General Motors Global Propulsion Systems), Davide Tricarico (General Motors Global Propulsion Systems), and Alessia Tuninetti (General Motors Global Propulsion Systems)

Learning Data Mining .361
Riccardo Guidotti (University of Pisa), Anna Monreale (University of Pisa), and Salvatore Rinzivillo (Institute of Information Science and Technologies CNR)

Large-Scale Railway Networks Train Movements: A Dynamic, Interpretable, and Robust Hybrid Data Analytics System .371
Alessandro Lulli (University of Genoa), Luca Oneto (University of Genoa), Renzo Canepa (Rete Ferroviaria Italiana S.p.A.), Simone Petralli (Rete Ferroviaria Italiana S.p.A.), and Davide Anguita (University of Genoa)

Selection of a Nominal Device Using Functional Data Analysis .381
Nevin Martin (Sandia National Laboratories), Thomas Buchheit (Sandia National Laboratories), and Shahed Reza (Sandia National Laboratories)

A Data Science Approach to Modelling a Manufacturing Facility's Electrical Energy Profile from Plant Production Data .387
Konrad Mulrennan (Institute of Technology Sligo), John Donovan (Institute of Technology Sligo), David Tormey (Institute of Technology Sligo), and Russell Macpherson (GlaxoSmithKline)

Big Data-Driven Platform for Cross-Media Monitoring .392
Liana Napalkova (Eurecat Technology Centre), Pablo Aragón (Eurecat Technology Centre), and Juan Carlos Castro Robles (Eurecat Technology Centre)

Detecting and Counting Panicles in Sorghum Images .400
Peder A. Olsen (IBM Research), Karthikeyan Natesan Ramamurthy (IBM Research), Javier Ribera (Purdue University), Yuhao Chen (Purdue University), Addie M. Thompson (Michigan State University), Ronny Luss (IBM Research), Mitch Tuinstra (Purdue University), and Naoki Abe (IBM Research)

A Case Study on Reducing Auto Insurance Attrition with Econometrics, Machine Learning, and A/B Testing .410
Miguel Paredes (Massachusetts Institute of Technology)

Captioning with Language-Based Attention .415
Anshu Rajendra (Carnegie Mellon University), Ritwik Rajendra (Carnegie Mellon University), Ole J. Mengshoel (Carnegie Mellon University), Ming Zeng (Carnegie Mellon University), and Momina Haider (Carnegie Mellon University)

Developing and Deploying a Taxi Price Comparison Mobile App in the Wild: Insights and Challenges .424
Anastasios Noulas (New York University), Vsevolod Salnikov (University of Namur), Desislava Hristova (University of Cambridge), Cecilia Mascolo (University of Cambridge), and Renaud Lambiotte (University of Oxford)

Classifying Sensitive Content in Online Advertisements with Deep Learning .434
Ashutosh Sanzgiri (AppNexus Inc.), Daniel Austin (Freestar), Kannan Sankaran (AppNexus Inc.), Ryan Woodard (AppNexus Inc.), Amit Lissack (AppNexus Inc.), and Sam Seljan (AppNexus Inc.)

Optimizing New User Experience in Online Services .442
Ken Soong (LinkedIn Corporation), Xin Fu (LinkedIn Corporation), and Yang Zhou (LinkedIn Corporation)

Towards Simulation-Data Science – A Case Study on Material Failures .450
Holger Trittenbach (Karlsruhe Institute of Technology), Martin Gauch (Karlsruhe Institute of Technology), Klemens Böhm (Karlsruhe Institute of Technology), and Katrin Schulz (Karlsruhe Institute of Technology)

Using Data Analytics to Optimize Public Transportation on a College Campus .460
Kurt Zimmer (Indiana University Bloomington), Hasan Kurban (Indiana University Bloomington), Mark Jenne (Indiana University Bloomington), Logan Keating (Indiana University Bloomington), Perry Maull (Indiana University Bloomington), and Mehmet Dalkilic (Indiana University Bloomington)

Special Sessions

Data Science for Social Good

DSAA 2018 Special Session: Data Science for Social Good .470
Daniela Paolotti (ISI Foundation) and Michele Tizzoni (ISI Foundation)

Weak Nodes Detection in Urban Transport Systems: Planning for Resilience in Singapore .472
Michele Ferretti (King's College London), Gianni Barlacchi (University of Trento and Fondazione Bruno Kessler), Luca Pappalardo (Institute of Information Science and Technologies CNR), Lorenzo Lucchini (University of Trento and Fondazione Bruno Kessler), and Bruno Lepri (Fondazione Bruno Kessler)

A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care .481
Qiwei Han (Nova School of Business and Economics), Mengxin Ji (University of California, Davis), Inigo Martinez de Rituerto de Troya (Nova School of Business and Economics), Manas Gaur (Wright State University), and Leid Zejnilovic (Nova School of Business and Economics)

Coolabilities API .491
David Nordfors (i4j), Sudipto Shankar Dasgupta (i4j), Ganapathy Subramanian (i4j), V.R. Ferose (i4j), Chally Grundwag (i4j), and Behrang Zandi (i4j)

Global Income Inequality and Savings: A Data Science Perspective .496
Kiran Sharma (Jawaharlal Nehru University), Subhradeep Das (Jawaharlal Nehru University), and Anirban Chakraborti (Jawaharlal Nehru University)

Fine-Grained Analysis of Cyberbullying Using Weakly-Supervised Topic Models .504
Yue Zhang (State University of New York Binghamton) and Arti Ramesh (State University of New York Binghamton)

Data Fusion to Describe and Quantify Search and Rescue Operations in the Mediterranean Sea .514
Katherine Hoffmann Pham (New York University), Jeremy Boy (United Nations Global Pulse), and Miguel Luengo-Oroz (United Nations Global Pulse)

Data Science in Computational Psychiatry and Pyschiatric Research

Data Science Challenges in Computational Psychiatry and Psychiatric Research .524
Daniel Stahl (King's College London) and Daniel Stamate (Goldsmiths, University of London)

MixDir: Scalable Bayesian Clustering for High-Dimensional Categorical Data .526
Constantin Ahlmann-Eltze (Heidelberg University) and Christopher Yau (CCB University of Birmingham & Alan Turing Institute London)

Latent Dirichlet Allocation in Discovering Goals in Patients Undergoing Bladder Cancer Surgery .540
Thomas M. Atkinson (Memorial Sloan Kettering Cancer Center)

Analysing Activities, Contextualized for General Health, Depression and Demographics .547
Fionn Murtagh (University of Huddersfield)

MRI-Based Diagnostics of Depression Concomitant with Epilepsy: In Search of the Potential Biomarkers .555
Maksim Sharaev (Skolkovo Institute of Science and Technology), Alexey Artemov (Skolkovo Institute of Science and Technology), Ekaterina Kondrateva (Skolkovo Institute of Science and Technology), Svetlana Sushchinskaya (Skolkovo Institute of Science and Technology), Evgeny Burnaev (Skolkovo Institute of Science and Technology), Alexander Bernstein (Skolkovo Institute of Science and Technology), Renat Akzhigitov (Moscow Research and Clinical Center for Neuropsychiatry), and Alexander Andreev (Deloitte Analytics Institute)

Outcome-Weighted Learning for Personalized Medicine with Multiple Treatment Options .565
Xuan Zhou (University of North Carolina at Chapel Hill), Yuanjia Wang (Columbia University), and Donglin Zeng (University of North Carolina at Chapel Hill)

EnGeoData'2018: Environmental and Geo-Spatial Data Analytics

Environmental and Geo-Spatial Data Analytics (EnGeoData'2018) .575
Maguelonne Teisseire (TETIS, Irstea), Mathieu Roche (TETIS, Cirad), and Diana Inkpen (University of Ottawa)

Airpred: A Flexible R Package Implementing Methods for Predicting Air Pollution .577
M. Benjamin Sabath (Harvard University), Qian Di (Harvard University), Danielle Braun (Harvard University), Joel Schwartz (Harvard University), Francesca Dominici (Harvard University), and Christine Choirat (Harvard University)

On Non-Routine Places in Urban Human Mobility .584
Christian Quadri (Università degli Studi di Milano), Matteo Zignani (Università degli Studi di Milano), Sabrina Gaito (Università degli Studi di Milano), and Gian Paolo Rossi (Università degli Studi di Milano)

Convolutional Neural Networks for Disaggregated Population Mapping Using Open Data .594
Luciano Gervasoni (University of Grenoble Alpes), Serge Fenet (University of Lyon 1), Regis Perrier (University of Grenoble Alpes), and Peter Sturm (INRIA)

Crowdsourcing Landforms for Open GIS Enrichment .604
Rocio Nahime Torres (Politecnico di Milano), Darian Frajberg (Politecnico di Milano), Piero Fraternali (Politecnico di Milano), and Sergio Luis Herrera Gonzales (Politecnico di Milano)

Opportunities and Risks for Data Science in Organizations: Banking, Finance, and Policy

Opportunities and Risks for Data Science in Organizations: Banking, Finance, and Policy - Special Session Overview .612
Antonia Azzini (Consorzio per il Trasferimento Tecnologico C2T), Stefania Marrara (Consorzio per il Trasferimento Tecnologico C2T), Amir Topalovi (Consorzio per il Trasferimento Tecnologico C2T), Mirjiana Peji Bach (University of Zagabria), and Matthew Rattigan (University of Massachusetts Amherst)

Determinants of Growth in Micro and Small Enterprises: Empirical Evidence from Jordan .614
Faisal Awartani (Alpha International for Research, Polling, and Informatics) and Bryanna Millis (FHI360)

Network-Based Models to Improve Credit Scoring Accuracy .623
Branka Hadji Misheva (University of Pavia), Paolo Giudici (University of Pavia), and Valentino Pediroda (University of Trieste)

Predictive Modeling for Identifying Return Defaulters in Goods and Services Tax .631
Priya Mehta (Indian Institute of Technology Hyderabad), Jithin Mathews (Indian Institute of Technology Hyderabad), K. Suryamukhi (Indian Institute of Technology Hyderabad), K. Sandeep Kumar (Indian Institute of Technology Hyderabad), and Ch. Sobhan Babu (Indian Institute of Technology Hyderabad)

SeCredISData: Sentiment, Emotion, and Credibility of Information in Social Data

SeCredISData 2018: Special Session on Sentiment, Emotion, and Credibility of Information in Social Data .638
Farah Benamara (Paul Sabatier University), Cristina Bosco (University of Turin), Elisabetta Fersini (University of Milano-Bicocca), Gabriella Pasi (University of Milano-Bicocca), Viviana Patti (University of Turin), and Marco Viviani (University of Milano-Bicocca)

A Methodological Template to Construct Ground Truth of Authentic and Fake Online Reviews .641
Snehasish Banerjee (University of York)

Behind Our Words: Psychological Paths Underlying the Un/Supportive Stance Toward Immigrants in Social Media .649
Francesca D'Errico (Roma Tre University), Marinella Paciello (Uninettuno University), and Matteo Amadei (Roma Tre University)

Portability of Aspect Based Sentiment Analysis: Thirty Minutes for a Proof of Concept 657
Luca Dini (Innoradiant), Paolo Curtoni (Innoradiant), and Elena Melnikova (Innoradiant)

Willingness to Share Emotion Information on Social Media: Influence of Personality and Social Context 665
Damien Dupré (Insight Centre for Data Analytics), Gary McKeown (Queen's University Belfast), Nicole Andelic (Sensum Ltd.), and Gawain Morrison (Sensum Ltd.)

An Empirical Analysis of the Role of Amplifiers, Downtoners, and Negations in Emotion Classification in Microblogs 673
Florian Strohm (University of Stuttgart) and Roman Klinger (University of Stuttgart)

Utilizing Tweet Content for the Detection of Sentiment-Based Interaction Communities on Twitter 682
Alron Jan Lam (De La Salle University Manila) and Charibeth Cheng (De La Salle University Manila)

Author Index 693