

*Commenced Publication in 1973*

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

## Editorial Board

David Hutchison

*Lancaster University, Lancaster, UK*

Takeo Kanade

*Carnegie Mellon University, Pittsburgh, PA, USA*

Josef Kittler

*University of Surrey, Guildford, UK*

Jon M. Kleinberg

*Cornell University, Ithaca, NY, USA*

Friedemann Mattern

*ETH Zurich, Zurich, Switzerland*

John C. Mitchell

*Stanford University, Stanford, CA, USA*

Moni Naor

*Weizmann Institute of Science, Rehovot, Israel*

C. Pandu Rangan

*Indian Institute of Technology Madras, Chennai, India*

Bernhard Steffen

*TU Dortmund University, Dortmund, Germany*

Demetri Terzopoulos

*University of California, Los Angeles, CA, USA*

Doug Tygar

*University of California, Berkeley, CA, USA*

Gerhard Weikum

*Max Planck Institute for Informatics, Saarbrücken, Germany*

More information about this series at <http://www.springer.com/series/7409>

Wouter Duivesteijn · Arno Siebes  
Antti Ukkonen (Eds.)

# Advances in Intelligent Data Analysis XVII

17th International Symposium, IDA 2018  
's-Hertogenbosch, The Netherlands, October 24–26, 2018  
Proceedings

*Editors*

Wouter Duivestijn  
Eindhoven University of Technology  
Eindhoven  
The Netherlands

Antti Ukkonen   
University of Helsinki  
Helsinki  
Finland

Arno Siebes  
Department of Information  
and Computing Sciences  
University Utrecht  
Utrecht  
The Netherlands

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Computer Science  
ISBN 978-3-030-01767-5              ISBN 978-3-030-01768-2 (eBook)  
<https://doi.org/10.1007/978-3-030-01768-2>

Library of Congress Control Number: 2018956595

LNCS Sublibrary: SL3 – Information Systems and Applications, incl. Internet/Web, and HCI

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

We are proud to present the proceedings of the 17th International Symposium on Intelligent Data Analysis (IDA 2018), which was held during October 24–28, 2018, in 's-Hertogenbosch, The Netherlands. The series started in 1995 and was held biannually until 2009. In 2010 the symposium refocused to support papers that go beyond established technology and offer genuinely novel and potentially game-changing ideas in the field of data analysis. The call for papers for this 2018 conference was formulated as follows:

“Complementary to other mainstream conferences in data science, IDA's mission is to promote ideas over performance: A solid motivation can be as convincing as exhaustive empirical evaluation. To this end, IDA creates an open atmosphere that encourages discussion and promotes innovative ideas in data analysis novel and game-changing ideas.”

But, clearly, not all novel ideas are good ideas. To ensure the quality of all accepted papers, standard rigorous, single-blind peer evaluation of all papers was performed by the Program Committee (PC) consisting of established researchers in the field who evaluated the papers against the requirements set out in the call for papers. As in previous editions, this process was complemented by the PC advisors, a select set of senior researchers with a multi-year involvement in the IDA conference series. Whenever a PC advisor flagged a paper as both good and presenting an interesting, novel, idea with an informed, thoughtful, positive review, the paper was accepted irrespective of the other reviews.

As in previous installments, this somewhat special focus of IDA has resulted in the submission and acceptance of a number of highly innovative papers that would have had a hard time in the mainstream conferences. In fact, we are pleased and proud to have put together a very strong program. We received 65 paper submissions out of which 29 could be accepted. Every submission was reviewed by at least two PC members, and the majority of submissions had at least three reviews. Each accepted paper was offered a slot for oral presentation and, new this year, also offered a poster at a specially organized poster session to foster deeper discussions than the brief Q&A minutes often offered right after the presentation.

We were honored that the regular program was complemented by three distinguished invited speakers who fulfilled IDA's quest for novel, game-changing ideas:

- Tuuli Toivonen (University of Helsinki) talked about how modern data science and machine learning methods can be used for analyzing and understanding human accessibility and mobility in urban and natural environments.
- Luc de Raedt (KU Leuven) talked about his ERC project to automate data science. More specifically, he discussed how automated data wrangling approaches can be used for pre-processing and how both predictive and descriptive models can in principle be combined to automatically complete spreadsheets and relational databases.

- Johannes Fürnkranz (TU Darmstadt) talked about the need for interpretability biases. Ever since the start of the field, interpretability has been one of the holy grails. Usually this notion is operationalized as simplicity. In this talk, he questioned this assumption, in particular with respect to commonly used rule learning heuristics that aim at learning rules that are as simple as possible.

We also invited all keynote speakers to submit a paper on the topic of their presentation. Professors de Raedt and Fürnkranz decided to take this opportunity. These invited papers appear in a separate Invited Papers section in the beginning of the proceedings. Also, the first selected contribution is a slightly shorter position paper by Leo Lahti about the importance of tools to facilitate open data science.

Finally, the program was completed by the traditional IDA PhD poster session in which PhD students get the opportunity to promote their work.

The conference was held in the former chapel of the Jheronimus Academy of Data Science, and we are grateful for their willingness to host the conference. We wish to express our gratitude to all authors of all submitted papers, for their intellectual contributions; to the PC members and additional reviewers for their efforts in reviewing, discussing, and commenting on all submitted papers; to the program chair advisors for their active involvement; and to the IDA council for their ongoing guidance and support, in particular Elizabeth Bradley, Jaakko Hollmén, and Matthijs van Leeuwen. Also, the program chairs wish to thank the general chair of IDA 2017, David Weston, for his help with practical matters related to preparing the conference proceedings. Finally, we are grateful to our sponsors and supporters: KNIME, which funded the IDA Frontier Prize for the most visionary contribution, as well as The Netherlands Research School for Information and Knowledge Systems (SIKS), the *Artificial Intelligence* journal, and Springer.

August 2018

Wouter Duivesteijn  
Arno Siebes  
Antti Ukkonen

# Organization

## General Chair

Wouter Duivestijn                      Eindhoven University of Technology, The Netherlands

## Program Chairs

Arno Siebes                              Utrecht University, The Netherlands  
Antti Ukkonen                            University of Helsinki, Finland

## Local Chair

Arjan Haring                              Jheronimus Academy of Data Science, The Netherlands

## Frontier Prize Chair

Michael Berthold                        University of Konstanz, Germany

## Advisory Chairs

Allan Tucker                              Brunel University London, UK  
Jaakko Hollmén                            Aalto University, Finland  
Matthijs van Leeuwen                    Leiden University, The Netherlands

## Organizing Committee

Arjan van den Born                        Jheronimus Academy of Data Science, The Netherlands  
Arjan Haring                                Jheronimus Academy of Data Science, The Netherlands  
Laura Niemeijer                            Jheronimus Academy of Data Science, The Netherlands

## Web and Social Media Chair

Simon van der Zon                        Eindhoven University of Technology, The Netherlands

## Program Committee Advisors

Michael Berthold                        University of Konstanz, Germany  
Hendrik Blockeel                         Katholieke Universiteit Leuven, Belgium  
Elizabeth Bradley                         University of Colorado, USA  
Tijl De Bie                                 Ghent University, Data Science Lab, Belgium  
Elisa Fromont                              Université de Rennes 1, France  
Jaakko Hollmén                            Aalto University, Finland

Frank Klawonn	Ostfalia University of Applied Sciences, Germany
Nada Lavrač	Jozef Stefan Institute, Slovenia
Matthijs van Leeuwen	Leiden University, The Netherlands
Panagiotis Papapetrou	Stockholm University, Sweden
Stephen Swift	Brunel University London, UK
Hannu Toivonen	University of Helsinki, Finland
Allan Tucker	Brunel University London, UK

## Program Committee

Ana Aguiar	University of Porto, Portugal
Fabrizio Angiulli	DEIS, University of Calabria, Italy
Mahir Arzoky	Brunel University London, UK
Martin Atzmueller	Tilburg University, The Netherlands
José Luis Balcázar	Universitat Politècnica de Catalunya, Spain
Gustavo Batista	University of São Paulo, Brazil
Maria Bielikova	Slovak University of Technology in Bratislava, Slovakia
Christian Borgelt	Otto von Guericke University Magdeburg, Germany
Ulf Brefeld	Leuphana Universität Lüneburg, Germany
Paula Brito	University of Porto, Portugal
Ricardo Cachucho	Leiden University, The Netherlands
Loïc Cerf	Universidade Federal de Minas Gerais, Brazil
Edward Cohen	Imperial College London, UK
Paulo Cortez	University of Minho, Portugal
Bruno Cremilleux	Université de Caen, France
Andre de Carvalho	University of São Paulo, Brazil
José Del Campo-Ávila	Universidad de Málaga, Spain
Anton Dries	Katholieke Universiteit Leuven, Belgium
Brett Drury	SciCrop, Brazil
Nuno Escudeiro	Instituto Superior de Engenharia do Porto, Portugal
Ad Feelders	Utrecht University, The Netherlands
Peter Flach	University of Bristol, UK
Johannes Fürnkranz	TU Darmstadt, Germany
Tias Guns	Vrije Universiteit Brussel, Belgium
Andreas Henelius	Aalto University, Finland
Pedro Henriques Abreu	FCTUC-DEI/CISUC, Portugal
Frank Höppner	Ostfalia University of Applied Sciences, Germany
Ulf Johansson	Jönköping University, Sweden
Alipio M. Jorge	University of Porto, Portugal
Arno Knobbe	Leiden University, The Netherlands
Irena Koprinska	University of Sydney, Australia
Petra Kralj Novak	Jozef Stefan Institute, Slovenia
Rudolf Kruse	University of Magdeburg, Germany
Niklas Lavesson	Jönköping University, Sweden
Jose A. Lozano	University of the Basque Country, Spain
Ling Luo	CSIRO, Australia



George Magoulas	Birkbeck College, Knowledge Lab, University of London, UK
Vera Migueis	University of Porto, Portugal
Mohamed Nadif	Paris Descartes University, France
Andreas Nuernberger	Otto von Guericke University Magdeburg, Germany
Kaustubh Raosaheb Patil	Massachusetts Institute of Technology, USA
Mykola Pechenizkiy	Eindhoven University of Technology, The Netherlands
Ruggero G. Pensa	University of Torino, Italy
Marc Plantevit	LIRIS - Université Claude Bernard Lyon 1, France
Lubos Popelinsky	Masaryk University, Czech Republic
Alexandra Poulouvassilis	Birkbeck College, University of London, UK
Miguel A. Prada	Universidad de Leon, France
Ronaldo Prati	Universidade Federal do ABC - UFABC, Brazil
Antonio Salmeron	University of Almería, Spain
Vítor Santos Costa	University of Porto, Portugal
Roberta Siciliano	University of Naples Federico II, Italy
Myra Spiliopoulou	Otto von Guericke University Magdeburg, Germany
Frank Takes	University of Amsterdam and Leiden University, The Netherlands
Melissa Turcotte	Los Alamos National Laboratory, USA
Peter van der Putten	Leiden University and Pegasystems, The Netherlands
Jan N. van Rijn	Leiden University, The Netherlands
Veronica Vinciotti	Brunel University London, UK
Jilles Vreeken	Max Planck Institute for Informatics and Saarland University, Germany
Leishi Zhang	Middlesex University, UK
Albrecht Zimmermann	Université Caen Normandie, France
Indre Zliobaite	University of Helsinki, Finland

# Contents

## Invited Papers

Elements of an Automatic Data Scientist . . . . .	3
<i>Luc De Raedt, Hendrik Blockeel, Samuel Kolb, Stefano Teso, and Gust Verbruggen</i>	
The Need for Interpretability Biases . . . . .	15
<i>Johannes Fürnkranz and Tomáš Kliegr</i>	

## Selected Contributions

Open Data Science . . . . .	31
<i>Leo Lahti</i>	
Automatic POI Matching Using an Outlier Detection Based Approach. . . . .	40
<i>Alexandre Almeida, Ana Alves, and Rui Gomes</i>	
Fact Checking from Natural Text with Probabilistic Soft Logic. . . . .	52
<i>Nouf Bindris, Saatviga Sudhahar, and Nello Cristianini</i>	
ConvoMap: Using Convolution to Order Boolean Data . . . . .	62
<i>Thomas Bollen, Guillaume Leurquin, and Siegfried Nijssen</i>	
Training Neural Networks to Distinguish Craving Smokers, Non-craving Smokers, and Non-smokers . . . . .	75
<i>Christoph Doell, Sarah Donohue, Cedrik Pätz, and Christan Borgelt</i>	
Missing Data Imputation via Denoising Autoencoders: The Untold Story. . . . .	87
<i>Adriana Fonseca Costa, Miriam Seoane Santos, Jastin Pompeu Soares, and Pedro Henriques Abreu</i>	
Online Non-linear Gradient Boosting in Multi-latent Spaces. . . . .	99
<i>Jordan Frery, Amaury Habrard, Marc Sebban, Olivier Caelen, and Liyun He-Guelton</i>	
MDP-based Itinerary Recommendation using Geo-Tagged Social Media . . . . .	111
<i>Radhika Gaonkar, Maryam Tavakol, and Ulf Brefeld</i>	
Multiview Learning of Weighted Majority Vote by Bregman Divergence Minimization . . . . .	124
<i>Anil Goyal, Emilie Morvant, and Massih-Reza Amini</i>	

Non-negative Local Sparse Coding for Subspace Clustering . . . . .	137
<i>Babak Hosseini and Barbara Hammer</i>	
Pushing the Envelope in Overlapping Communities Detection . . . . .	151
<i>Said Jabbour, Nizar Mhadhbi, Badran Raddaoui, and Lakhdar Sais</i>	
Right for the Right Reason: Training Agnostic Networks . . . . .	164
<i>Sen Jia, Thomas Lansdall-Welfare, and Nello Cristianini</i>	
Link Prediction in Multi-layer Networks and Its Application to Drug Design. . . . .	175
<i>Maksim Koptelov, Albrecht Zimmermann, and Bruno Crémilleux</i>	
A Hierarchical Ornstein-Uhlenbeck Model for Stochastic Time Series Analysis. . . . .	188
<i>Ville Laitinen and Leo Lahti</i>	
Analysing the Footprint of Classifiers in Overlapped and Imbalanced Contexts . . . . .	200
<i>Marta Mercier, Miriam S. Santos, Pedro H. Abreu, Carlos Soares, Justin P. Soares, and João Santos</i>	
Tree-Based Cost Sensitive Methods for Fraud Detection in Imbalanced Data . . . . .	213
<i>Guillaume Metzler, Xavier Badiche, Brahim Belkasm, Elisa Fromont, Amaury Habrard, and Marc Sebban</i>	
Reduction Stumps for Multi-class Classification . . . . .	225
<i>Felix Mohr, Marcel Wever, and Eyke Hüllermeier</i>	
Decomposition of Quantitative Gaifman Graphs as a Data Analysis Tool . . . .	238
<i>José Luis Balcázar, Marie Ely Piceno, and Laura Rodríguez-Navas</i>	
Exploring the Effects of Data Distribution in Missing Data Imputation . . . . .	251
<i>Justin Pompeu Soares, Miriam Seoane Santos, Pedro Henriques Abreu, Hélder Araújo, and João Santos</i>	
Communication-Free Widened Learning of Bayesian Network Classifiers Using Hashed Fiedler Vectors. . . . .	264
<i>Oliver R. Sampson, Christian Borgelt, and Michael R. Berthold</i>	
Expert Finding in Citizen Science Platform for Biodiversity Monitoring via Weighted PageRank Algorithm. . . . .	278
<i>Zakaria Saoud and Colin Fontaine</i>	

Random Forests with Latent Variables to Foster Feature Selection in the Context of Highly Correlated Variables. Illustration with a Bioinformatics Application. . . . . 290  
*Christine Sinoquet and Kamel Mekhnacha*

Don't Rule Out Simple Models Prematurely: A Large Scale Benchmark Comparing Linear and Non-linear Classifiers in OpenML . . . . . 303  
*Benjamin Strang, Peter van der Putten, Jan N. van Rijn, and Frank Hutter*

Detecting Shifts in Public Opinion: A Big Data Study of Global News Content . . . . . 316  
*Saatviga Sudhahar and Nello Cristianini*

Biased Embeddings from Wild Data: Measuring, Understanding and Removing . . . . . 328  
*Adam Sutton, Thomas Lansdall-Welfare, and Nello Cristianini*

Real-Time Excavation Detection at Construction Sites using Deep Learning . . . . . 340  
*Bas van Boven, Peter van der Putten, Anders Åström, Hakim Khalafi, and Aske Plaat*

COBRAS: Interactive Clustering with Pairwise Queries . . . . . 353  
*Toon Van Craenendonck, Sebastijan Dumančić, Elia Van Wolputte, and Hendrik Blockeel*

Automatically Wrangling Spreadsheets into Machine Learning Data Formats . . . . . 367  
*Gust Verbruggen and Luc De Raedt*

Learned Feature Generation for Molecules . . . . . 380  
*Patrick Winter, Christian Borgelt, and Michael R. Berthold*

**Author Index** . . . . . 393