

Peter Haber
Thomas Lampoltshammer
Manfred Mayr *Eds.*

Data Science — Analytics and Applications

Proceedings of the 1st International Data
Science Conference — iDSC2017

EBOOK INSIDE

 Springer Vieweg

Data Science – Analytics and Applications

Peter Haber · Thomas Lampoltshammer · Manfred Mayr
(Eds.)

Data Science – Analytics and Applications

Proceedings of the 1st International Data Science Conference – iDSC2017

Editors

Peter Haber

Informationstechnik & System-Management
Fachhochschule Salzburg Puch/Salzburg, Österreich

Manfred Mayr

Informationstechnik & System-Management
Fachhochschule Salzburg Puch/Salzburg, Österreich

Thomas Lampoltshammer

Department für E-Governance in Wirtschaft und Verwaltung
Donau-Universität Krems, Krems an der Donau / Österreich

ISBN 978-3-658-19286-0

ISBN 978-3-658-19287-7 (eBook)

<https://doi.org/10.1007/978-3-658-19287-7>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Springer Vieweg

© Springer Fachmedien Wiesbaden GmbH 2017

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist Teil von Springer Nature

Die eingetragene Gesellschaft ist Springer Fachmedien Wiesbaden GmbH

Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Preface

It is with deep satisfaction that we write this foreword for the Proceedings of the 1st International Data Science Conference (iDSC) held in Salzburg, Austria, June 12th - 13th 2017. The conference program and the resulting proceedings represent the efforts of many people. We want to express our gratitude towards the members of our program committee as well as towards our external reviewers for their hard work during the reviewing process.

iDSC proofed itself as an innovative conference, which gave its participants the opportunity to delve into state-of-the-art research and best practice in the fields of Data Science and data-driven business concepts. Our research track offered a series of presentations by Data Science researchers regarding their current work in the fields of Data Mining, Machine Learning, Data Management, and the entire spectrum of Data Science.

In our industry track, practitioners demonstrated showcases of data-driven business concepts and how they use Data Science to achieve organisational goals, with a focus on manufacturing, retail, and financial services. Within each of these areas, experts described their experience, demonstrated their practical solutions, and provided an outlook into the future of Data Science in the business domain.

Besides these two parallel tracks, a European symposium on Text and Data Mining has been integrated into the conference. This symposium highlighted the EU project FutureTDM, granting insights into the future of Text and Data Mining, and introducing overarching policy recommendations and sector-specific guidelines to help stakeholders overcome the legal and technical barriers, as well the lack of skills that have been identified.

Our sponsors had their own, special platform via workshops to provide hands-on interaction with tools or to learn approaches towards concrete solutions. In addition, an exhibition of products and services offered by our sponsors took place throughout the conference, with the opportunity for our participants to seek contact and advice.

Completing the picture of our program, we proudly presented keynote presentations from leaders in Data Science and data-driven business, both researchers and practitioners. These keynotes provided all participants the opportunity to come together and shared views on challenges and trends in Data Science.

In addition to the contributed papers, five invited keynote presentations were given by: Euro Beinat (CS Research, Salzburg University), Mario Meir-Huber (Microsoft Austria), Mike Olson (Cloudera), Ralf Klinkenberg (RapidMiner) and Janek Strycharz (Digital Center Poland). We thank the invited speakers for sharing their insights with our community.

The conference chair John Thompson has also helped us in many ways setting up the industry track, for which we are grateful. We would especially like to thank our two colleagues, Astrid Karnutsch and Maximilian Tschuchnig, for their enormous and constructive commitment to organizing and conducting the conference. The paper submission and reviewing process was managed using the EasyChair system.

These proceedings will provide scientists and practitioners with an excellent reference to current activities in the Data Science domain. We trust also that this will be an impetus to stimulate further studies, research activities and applications in all discussed areas ensured by the support of our publisher Springer / Vieweg Wiesbaden Germany.

Finally, again, the conference would not be possible without the excellent papers contributed by our authors. We thank them for their contributions and their participation at iDSC'17.

Peter Haber, Thomas Lampoltshammer and Manfred Mayr

Conference Chairs

Future TDM Symposium Recap

FutureTDM is a European project focusing on reducing barriers and increasing uptake of Text and Data Mining (TDM) for research environments in Europe. The outcomes of the project were presented in the Symposium which has also served to connect key actors and interest groups and promote open dialogue via discussion panels and informal workshops. The FTDM Symposium was scheduled alongside iDSC 2017, given that both events address similar target groups and share a common perspective: they both aimed at creating a communication network among the members of the TDM community, where experts can exchange ideas and share the most up-to-date research results, as well as legal and industrial advances relevant to TDM. The audience targeted by the iDSC conference was the broad community of researchers and industry practitioners as well as other practitioners and stakeholders, making it ideal for disseminating the project's results.

The project's objective has been to detect the barriers to TDM, reveal best practices and put together sets of recommendations for TDM practitioners through a collaborative knowledge and open information approach. The barriers recorded were grouped around four pillars: a) **legal**, b) **economic**, c) **skills**, d) **technical**. These categories emerged after discussions with respective stakeholders such as researchers, developers, publishers and SMEs during Knowledge Cafés run across Europe (the Netherlands, the United Kingdom, Italy, Slovenia, Germany, Poland etc) and two workshops held in Brussels¹ (on September, 27th 2016 and March, 29th 2017).

The Symposium² was a chance to invite experts from all over Europe to share their experience and expertise in different domains. It was also a great opportunity to announce the guidelines and recommendations formulated in order to increase TDM uptake. It started with a brief introduction by Bernhard Jäger (SYNYO)³ underlying the need to bring together different groups of stakeholders, such as policy makers and legislators, developers and users who would benefit from the project's findings and the respective recommendations formed by the FTDM working groups. It continued with a keynote speech by Janek Strycharz (Projekt Polska Foundation) dedicated to the Economic Potential of Data Analytics. Janek Strycharz elaborated on different types of Big Data and the variety of possibilities they offer and explained how that at a global and European scale there could be a benefit from Big Data and TDM (the European GDP alone would be increased by USD 200 billion).

¹ FutureTDM Workshop I and II outcomes can be found at <http://www.futuretdm.eu/knowledge-cafes/futuretdm-workshop/>
<http://www.futuretdm.eu/knowledge-cafes/futuretdm-workshop-2/>

² All presentation slides are available online at www.slideshare.net/FutureTDM/presentations

³ Presentation on Introduction to the FutureTDM project is available at <http://www.slideshare.net/FutureTDM/introduction-to-the-future-tdm-project>

The first session entitled “**Data Analytics and the Legal Landscape: Intellectual Property and Data Protection**” included Freyja van den Boom, researcher from Open Knowledge International/Content Mine who presented the legal barriers identified and the respective recommendations created under the subject "Dealing with the legal bumps on the road to further TDM uptake". The focus of the presentation was on the principles identified to counterbalance barriers: Awareness and Clarity, TDM Without Boundaries, and Equitable Access. The session was chaired by Ben White (Head of Intellectual Property at the British Library) and included the following panelists: i) Duncan Campbell (John Wiley & Sons, Inc.), representing the publisher’s perspective, ii) Prodromos Tsiavos (Onassis Cultural Centre/IP Advisor), providing an organization’s point of view, iii) Marie Timmermann (Science Europe), offering her point of view as the EU Legislation and Regulatory Affairs Officer and iv) Romy Sigl (AustrianStartups) sharing her experience from startUps. The discussion revolved around regulations which must address the implementation of the law and its exceptions, copyright issues, the distinction between commercial and noncommercial activities, the need for better communication between different groups of stakeholders and the importance and value of TDM for publishers.

During the following session the projects ContentMine (Stefan Kasberger), PLAZI (Donat Agosti), CORE (Petr Knoth), RapidMiner (Ralf Klinkenberg), clarin:el (Maria Gavrilidou) and ALCIDE (Alessio Palmero Aprosio) were introduced and the presenters were accessible for a more detailed presentation of their work to the attendees who would be interested in learning more. The researchers shared their experience on technical and legal problems they have encountered demonstrating the TDM applications and infrastructures they had created.

The next session offered an **overview of FTDM case studies from Startups to Multinationals**. The presentation entitled "Stakeholder consultations - The Highlights" was given by Freyja van den Boom (Open Knowledge International/Content Mine) who talked about the findings from continuous stakeholder consultations throughout the project. The session was chaired by Maria Eskevich (Radboud University) and included as panelists Donat Agosti (PLAZI), Petr Knoth (CORE), Kim Nilsson (PIVIGO), and Peter Murray-Rust (ContentMine). The issues raised during discussion pinpointed the need for realistic solutions to infrastructures, community engagement, and open source and data.

Kiera McNeice (British Library) was the presenter in the fourth session and her presentation was entitled "Supporting TDM in the Education Sector". The session focusing on “**Universities, TDM and the need for strategic thinking on educating researchers**” was chaired by Ben White (Head of Intellectual Property at the British Library) and panelists Claire Sewell (Cambridge University Library), Jonas Holm (Stockholm University Library), and Kim Nilsson (PIVIGO). The discussion which followed touched upon issues such as the future of Data Science and the nature of Data Scientists. Some of the key concepts which were discussed were that of inclusion and diversity, gender imbalance and nationality characteristics, which all affect access to Data Science and the ability to become a Data Scientist. Concerns were expressed as to whether anyone could become a Data Scientist, and whether the focus should be on becoming a Data Scientist or a more efficient TDM user.

The challenges and solutions regarding technologies and infrastructures supporting Text and Data Analytics was the topic of the fifth session, the main presenter of which was Maria Eskevich (Radboud University). She focused on "The TDM Landscape: Infrastructure and Technical Implementation" and touched upon the business and scientific perspectives on TDM by showing the investment made by the EU in the five economic sectors. She also talked about the barriers/challenges encountered in terms of accessibility and interoperability of infrastructures, sustainability of data and digital readiness of language resources. The following discussion, chaired by Stelios Piperidis (ARC) with Mihai Lupu (Data Market Austria), Maria Gavrilidou (clarin: el) and Nelson Silva (know-centre) revolved around real TDM problems and the solutions the researchers came up with and close with the requirements of an effective TDM infrastructure.

The final session of the Symposium was dedicated to the **Next Steps: A Roadmap to promoting greater uptake of Data Analytics in Europe**. A presentation was made by Kiera McNeice (British Library) who briefly summarised what the project has achieved so far and focussed on the key principles from the FutureTDM Policy Framework⁴ which must underlie all the efforts to be made in the future in Legal Policies, Skills and Education, Economy and Incentives and Technical and Infrastructure.

The Symposium close with a presentation of Bernhard Jäger and Burcu Akinci (SYNYO) of the FutureTDM platform (<http://www.futuretdm.eu/>), which is populated with the project outcomes and findings. The platform will continue to exist after the end of the project and will be continuously revised and updated in order to maintain a coherent and up-to-date view on the TDM landscape open to the public.

Kornella Pouli

Athena RIC/ILSP, Athens

Burcu Akinci

SYNYO GmbH, Vienna

⁴ <http://www.futuretdm.eu/policy-framework/>

Organisation

Organising Institutions

Salzburg University of Applied Sciences
Information Professionals GmbH

Conference Chairs

Peter Haber
Thomas J. Lampoltshammer
Manfred Mayr
John A. Thompson

Salzburg University of Applied Sciences
Danube University Krems
Salzburg University of Applied Sciences
Information Professionals GmbH

Organising Committee

Peter Haber
Astrid Karnutsch
Thomas J. Lampoltshammer
Manfred Mayr
John A. Thompson
Susanne Schnitzer
Maximilian E. Tschuchnig

Salzburg University of Applied Sciences
Salzburg University of Applied Sciences
Danube University Krems
Salzburg University of Applied Sciences
Information Professionals GmbH
Information Professionals GmbH
Salzburg University of Applied Sciences

Program Committee

David C. Anastasiu
Vera Andrejcenko
Christian Bauckhage
Markus Breunig
Stefanie Cox
Werner Dubitzky
Günther Eibl
Süleyman Eken
Karl Entacher
Edison Pignaton de Freitas
Bernhard Geissler
Charlotte Gerritsen

Mohammad Ghoniem
Peter Haber
Johann Höchtl
Martin Kaltenböck
Astrid Karnutsch
Elmar Kiesling
Robert Krimmer
Peer Kröger
Thomas J. Lampoltshammer

San Jose State University
University of Antwerp
University of Bonn
Rosenheim University of Applied Sciences
IT Innovation Centre
University of Ulster, Coleraine
Salzburg University of Applied Sciences
University Kocaeli
Salzburg University of Applied Sciences
Federal University of Rio Grande do Sul
Danube University Krems
Netherlands Institute for the Study of Crime and Law
Enforcement (NSCR)
Luxembourg Institute of Science and Technology
Salzburg University of Applied Sciences
Danube University Krems
Semantic Web Company
Salzburg University of Applied Sciences
Vienna University of Technology
University of Tallinn
Ludwig-Maximilians-Universität München
Danube University Krems

Michael Leitner
Giuseppe Manco
Manfred Mayr
Mark-David McLaughlin
Robert Merz
Elena Lloret Pastor
Cody Ryan Peebles
Gabriela Viale Pereira
Peter Ranacher
Siegfried Reich
Eric Rozier
Johannes Scholz
Maximilian E. Tschuchnig
Jürgen Umbrich
Andreas Unterweger
Eveline Wandl-Vogt
Stefan Wegenkittl
Stefanie Wiegand
Peter Wild
Radboud Winkels
Anneke Zuiderwijk - van Eijk

Louisiana State University
University of Calabria
Salzburg University of Applied Sciences
Bentley University
Salzburg University of Applied Sciences
University of Alicante
Cisco
Fundação Getúlio Vargas – EAESP
University of Zurich
Salzburg Research Forschungsgesellschaft mbH
Iowa State University
Graz University of Technology
Salzburg University of Applied Sciences
Vienna University of Economics and Business
Salzburg University of Applied Sciences
Austrian Academy of Sciences
Salzburg University of Applied Sciences
IT Innovation Centre / University of Southampton
Austrian Institute of Technology
University of Amsterdam
Delft University of Technology

Reviewer

David C. Anastasiu
Christian Bauckhage
Markus Breunig
Cornelia Ferner
Werner Dubitzky
Günther Eibl
Karl Entacher
Bernhard Geissler
Martin Kaltenböck
Peer Kröger
Thomas J. Lampoltshammer
Michael Leitner
Elena Lloret Pastor
Manfred Mayr
Robert Merz
Edison Pignaton de Freitas
Siegfried Reich
Eric Rozier
Johannes Scholz
Maximilian E. Tschuchnig
Jürgen Umbrich
Andreas Unterweger
Stefan Wegenkittl

San Jose State University
University of Bonn
Rosenheim University of Applied Sciences
Salzburg University of Applied Sciences
University of Ulster, Coleraine
Salzburg University of Applied Sciences
Salzburg University of Applied Sciences
Danube University Krems Höchtel
Semantic Web Company
Ludwig-Maximilians-Universität München
Danube University Krems
Louisiana State University
University of Alicante
Salzburg University of Applied Sciences
Salzburg University of Applied Sciences
Federal University of Rio Grande do Sul
Salzburg Research Forschungsgesellschaft mbH
Iowa State University
Graz University of Technology
Salzburg University of Applied Sciences
Vienna University of Economics and Business
Salzburg University of Applied Sciences
Salzburg University of Applied Sciences

Sponsors of the conference

Platinum Sponsors



Cloudera GmbH

Apache Hadoop-based software, support and services, and training

www.cloudera.com

Silver Sponsors



The unbelievable Machine Company GmbH

Full-service provider for Big Data, cloud services & hosting

www.unbelievable-machine.com



F&F GmbH

IT consulting, solutions and Big Data Analytics

www.ff-muenchen.de



The MathWorks GmbH

Mathematical computing software

www.mathworks.com



RapidMiner GmbH

Data science software platform for data preparation, machine learning, deep learning, text mining, and predictive analytics

www.rapidminer.com



ITG: innovative consulting and location development

ITG is Salzburg's innovation centre

www.itg-salzburg.at

Table of Content

German Abstracts	1
Full Papers – Double Blind Reviewed	9
Reasoning and Predictive Analytics.....	11
Circadian Cycles and Work Under Pressure: A Stochastic Process Model for E-learning Population Dynamics	13
<i>César Ojeda, Rafet Sifa and Christian Bauckhage</i>	
Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst	19
<i>César Ojeda, Rafet Sifa and Christian Bauckhage</i>	
Knowledge-based Short-Term Load Forecasting for Maritime Container Terminals.....	25
<i>Norman Ihle and Axel Hahn</i>	
Data Analytics in Community Networks.....	31
Beyond Spectral Clustering: A Comparative Study of Community Detection for Document Clustering.....	33
<i>César Ojeda, Rafet Sifa, Kostadin Cvejovski and Christian Bauckhage</i>	
Third Party Effect: Community Based Spreading in Complex Networks.....	39
<i>César Ojeda, Shubham Agarwal, Rafet Sifa and Christian Bauckhage</i>	
Cosine Approximate Nearest Neighbors	45
<i>David C. Anastasiu</i>	
Data Analytics through Sentiment Analysis	51
Information Extraction Engine for Sentiment-Topic Matching in Product Intelligence Applications	53
<i>Cornelia Ferner, Werner Pomwenger, Stefan Wegenkittl, Martin Schnöll, Veronika Haaf and Arnold Keller</i>	
Towards German Word Embeddings: A Use Case with Predictive Sentiment Analysis	59
<i>Eduardo Brito, Rafet Sifa, Kostadin Cvejovski, César Ojeda and Christian Bauckhage</i>	
User/Customer-centric Data Analytics.....	63
Feature Extraction and Large Activity-Set Recognition Using Mobile Phone Sensors	65
<i>Wassim El Hajj, Ghassen Ben Brahim, Cynthia El-Hayek and Hazem Hajj</i>	
The Choice of Metric for Clustering of Electrical Power Distribution Consumers	71
<i>Nikola Obrenović, Goran Vidaković and Ivan Luković</i>	
Evolution of the Bitcoin Address Graph	77
<i>Erwin Filtz, Axel Polleres, Roman Karl and Bernhard Haslhofer</i>	