FutureTDM
The Future of Text and Data Mining

iDSC 2017

1st International Data Science Conference (iDSC 2017) Conference Program

June 12th - 13th 2017
Salzburg, Austria

www.idsc.at

organized by

fhs Fachhochschule Salzburg University of Applied Sciences

INFO PRO Information Professionals Working Strategies.

sponsored by

cloudera

Imprint

Company wording:

Salzburg University of Applied Sciences
Fachhochschule Salzburg GmbH

Legal form:

Gesellschaft mit beschränkter Haftung (Limited Corporation)

Line of business:

According to §1 FHStG (Stf: BGBl 1993/340 idF BGBl 2011/74):

Accomplishment of University of Applied Science study paths and courses for further education; accomplishment of scientific research and development tasks in the proper sense of the University of Applied Science study law and the organization of scientific lectures, seminars, conferences, courses, scientific symposia and discussions.

Company identification number: ATU 44554503

Commercial register number: FN 166054y

Commercial register court: Landesgericht Salzburg (Regional Court Salzburg)

Address:

Urstein Süd 1
Austria – 5412 Puch/Salzburg

Contact:

T +43-(0)50-2211-1330
F +43-(0)50-2211-1349

E-Mail: office@idsc.at

Web: http://www.idsc.at

Pictures courtesy of:

Fachhochschule Salzburg GmbH
Information Professionals GmbH
Salzburg Tourismus
unsplash.com

Contents

Conference Chairs

Peter Haber - Salzburg University of Applied Sciences
Manfred Mayr - Salzburg University of Applied Sciences
John A. Thompson - Information Professionals GmbH
Thomas J. Lampoltshammer - Danube University Krems

Organization Commitee

Peter Haber - Salzburg University of Applied Sciences
Manfred Mayr - Salzburg University of Applied Sciences
John A. Thompson - Information Professionals GmbH
Thomas J. Lampoltshammer - Danube University Krems
Astrid Karnutsch - Salzburg University of Applied Sciences
Susanne Schnitzer - Information Professionals GmbH
Maximilian Tschuchnigg - Salzburg University of Applied Sciences

Program Committee

David C. Anastasiu - San Jose State University
Günther Eibl - Salzburg University of Applied Sciences
Siegfried Reich - Salzburg Research Forschungsgesellschaft mbH
Elena Lloret Pastor - University of Alicante
Giuseppe Manco - University of Calabria
Robert Merz - Salzburg University of Applied Sciences
Eric Rozier - Iowa State University
Andreas Unterweger - Salzburg University of Applied Sciences
Johannes Scholz - Graz University of Technology
Johann Höchtl - Danube University Krems
Karl Entacher - Salzburg University of Applied Sciences
Gabriela Viale Pereira - Fundação Getúlio Vargas - EAESP
Robert Krimmer - University of Tallinn
Peter Ranacher - University of Zurich
Stefan Wegenkittl - Salzburg University of Applied Sciences
Mohammad Ghoniem - Luxembourg Institute of Science and Technology
Peter Wild - Austrian Institute of Technology
Edison Pignaton de Freitas - Federal University of Rio Grande do Sul
Mark-David McLaughlin - Bentley University
Cody Ryan Peeples - Cisco
Elmar Kiesling - Vienna University of Technology
Radboud Winkels - University of Amsterdam
Stefanie Wiegand - IT Innovation Centre / University of Southampton
Michael Leitner - Louisiana State University
Vera Andrejcenko - University of Antwerp
Eveline Wandl-Vogt - Austrian Academy of Sciences
Charlotte Gerritsen - Netherlands Institute for the Study of Crime and Law Enforcement (NSCR)
Anneke Zuiderwijk - van Eijk - Delft University of Technology
Peer Kröger - Ludwig-Maximilians-Universität München
Christian Bauckhage - University of Bonn
Jürgen Umbrich - Vienna University of Economics and Business
Martin Kaltenböck - Semantic Web Company
Markus Breunig - Rosenheim University of Applied Sciences
Süleyman Eken - University Kocaeli

The way we deal with information is rapidly changing how we work, how we learn, and how we do business. It is vitally important that we come together to discuss issues and to develop solutions to the challenges of a data-driven world.

For this reason, the Salzburg University of Applied Sciences and Information Professionals GmbH have joined together to host the 1st International Data Science Conference. We welcome you to this unique forum, a platform for research and industry to discuss critical questions of our time.

A special thanks goes to our sponsors, all of them organizations which are deeply engaged in the innovative world of data. Their support has made this conference possible, providing a place for us to reach out to one another.

Let's talk data!

Salzburg University of Applied Sciences (SUAS), School of Information Technology and Systems Management (ITS)

Besides teaching, research, and development, which are important pillars of SUAS, the university is known for its outstanding application-oriented performance and close contact with the economy and industry. Within the degree programme ITS, with its 400 students in total, research has been pursued for nearly 20 years in numerous regional, national, and international projects.

By integrating the competences of the research fields software technology, network technology, smart grid and IT security, e-health, industrial signal processing, automation technology and robotics, ITS together with several partners from industry and business has resumed an important role as innovator in the region and beyond.

Information Professionals GmbH

Information Professionals GmbH is a consulting and services company, enabling IT-driven innovation and digital transformation. Our primary focus is data-driven business, especially in the areas of Big Data and Advanced Analytics. Our products include management and technology consulting, as well as data products based on data mining methods.

Though we are active world-wide, we are firmly rooted in the region Salzburg / Southeast Bavaria. We strongly support current efforts to establish a regional network of competence in Data Science and digital transformation, of which iDSC is an important building block.

The **1st International Data Science Conference** (iDSC 2017) is the meeting place for researchers, business managers, and data scientists to discover and share innovative approaches and solutions to the challenges of a data-driven world. Over the course of two days, iDSC 2017 gives its participants the opportunity to delve into state-of-the-art research and up-to-date practice in Data Science and data-driven business.

Our **Research Track** offers a series of short presentations from 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 demonstrate showcases of data-driven business and how they use Data Science to achieve organizational goals, with a focus on manufacturing, retail, and financial services.

Besides these parallel tracks, on the second day, a european symposium on Text and Data Mining has been integrated into the conference. This symposium highlights 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 will have their own, special platform via **Workshops** to provide hands-on interaction with tools or to learn approaches towards concrete solutions. In addition, there will be an **Exhibition** of the sponsors' products and services throughout the conference, with the opportunity for our participants to seek contact and advice.

Rounding out the program, we are proud to present **Keynote Presentations** from leaders in Data Science and data-driven business, both researchers and practitioners. These keynotes provide all participants the opportunity to come together and share views on challenges and trends in Data Science.

The **FutureTDM (FTDM)** project seeks to improve uptake of text and data mining (TDM) in the EU by actively engaging with stakeholders such as researchers, developers, publishers and SMEs. This engagement has been a two way process, involving both information dissemination and stakeholder feedback. FTDM run a series of Knowledge Cafés and Workshops across Europe addressing people working in content mining, big data and/or data analytics.

These consultations have led in identifying barriers to TDM uptake in the following categories: i) Legal & Content, ii) Economy & Incentives, iii) Skills & Education, and iv) Technical & Infrastructure.
Evidence was further provided by the study of case studies and TDM practices carried out by scientific researchers and small scale companies working in different sectors of economy. These case studies involved cross-discipline collaborations, private-public partnerships and not only EU but also international collaborations. In parallel FTDM documented the European status quo of TDM research, development and applications in a number of activity areas across different economic sectors.

The principles underlying the barriers which were identified are Uncertainty, Fragmentation and Restrictiveness. FTDM has produced guidelines for practitioners and recommendations for stakeholders based on counterbalancing principles: Awareness and Clarity, TDM Without Boundaries (for the fragmented TDM landscape) and Equitable Access (for restrictions - either legal, practical, economic or technical - to TDM).

All the FutureTDM outcomes are publicly available under www.futuretdm.eu as well as the FutureTDM Knowledge Base. The knowledge base showcases structured collections of resources on TDM that has been gathered throughout the FutureTDM project phase. The collections encompass experts as projects or organisations focusing on TDM, as well as technologies and resources that are useful for TDM practitioners (i.e. TDM methods and TDM tools).

http://project.futuretdm.eu/

FutureTDM
The Future of Text and Data Mining

Research Track	Industry Track
09:00 Opening and Welcome	
9:15 Ralf Klinkenberg (RapidMiner) **Current Developments and Trends in Machine Learning**	
10:15 Break	
Reasoning and Predictive Analytics Chair: David Anastasiu	**Projects**
10:30 Cesar Ojeda (Fraunhofer IAIS) Circadian Cycles and Work Under Pressure: A Stochastic Process Model for E-learning Population Dynamics **11:00 Rafet Sifa (Fraunhofer IAIS)** Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst **11:30 Norman Ihle** **(OFFIS e.V. – Institute for Information Technology)** Knowledge-based Short-Term Load Forecasting for Maritime Container Terminals	**10:30 Albin Gruber (Wüstenrot Datenservice GmbH)** Data Warehousing@Wüstenrot - Structured Data as Basis for Reporting and Analytics **11:15 Eric Rozier (Iowa State University)** Securing Against Data Integrity Attacks at The World Bank
12:00 Lunch	
Data Analytics in Community Networks Chair: Karl Entacher	**Technology**
13:15 Kostadin Cvejoski (Fraunhofer IAIS) Beyond Spectral Clustering: A Comparative Study of Community Detection for Document Clustering **13:45 Shubham Agarwal** **(Hewlett Packard Enterprise GmbH)** Third Party Effect: Community Based Spreading in Complex Networks **14:15 David Anastasiu (San José State University)** Cosine Approximate Nearest Neighbors	**13:15 Tomas Knap (Semantic Web Company GmbH)** Preparing DBpedia Knowledge Graph using UnifiedViews, an ETL tool for RDF data **14:00 Dmytro Martynenko (The MathWorks GmbH)** Building a Big Engineering Data Analytics System using MATLAB
14:45 Coffee Break	
Data Analytics through Sentiment Analysis Chair: Eric Rozier	**Technology**
15:15 Cornelia Ferner **(Salzburg University of Applied Sciences)** Information Extraction Engine for Sentiment-Topic Matching in Product Intelligence Applications **15:45 Eduardo Brito (Fraunhofer IAIS)**: Towards German Word Embeddings: A Use Case with Predictive Sentiment Analysis	**15:15 Stephan Schiffner, Markus Breunig (F&F GmbH)** Data Analysis with Spark - from Raw Data to Insights
16:15 Break	
16:45 Euro Beinat (University of Salzburg) **From Data Science to A.I.: A Management and Leadership Perspective**	
17:15 Panel Discussion **Data Science Sustainability - From Research to Industry through Education** Euro Beinat (University of Salzburg), Ralf Klinkenberg (RapidMiner), Mihai Lupu (Data Market Austria), Siegfried Reich (Salzburg Research), Stefan Wegenkittl (FH-Salzburg), Ben White (FutureTDM)	
18:15 Closing	

Research Track	Industry Track
9:00 | Mike Olson (Cloudera) **Enabling Data Science in the Enterprise**	
9:45 | Bernhard Jäger (Synyo GmbH) **Introduction to the FutureTDM Project**	
10:00 | Janek Strycharz (Projekt Polska Foundation) **The Economic Potential of Data Analytics**	
10:30 Break	
User/Customer-centric Data Analytics Chair: Johannes Scholz **10:45 Wassim El-Hajj (American University of Beirut)** Feature Extraction and Large Activity-Set Recognition Using Mobile Phone Sensors **11:15 Nikola Obrenovic (Schneider Electric DMS NS Llc.)** The Choice of Metric for Clustering of Electrical Power Distribution Consumers **11:45 Erwin Filtz (Vienna University of Economics and Business)** Evolution of the Bitcoin Address Graph	**Application Scenarios** **10:45 Thomas Thalhammer (SPAR Business Services GmbH)** Digital Business at SPAR Business Services GmbH **11:30 Tamas Molnar (Vodafone Group BI)** Reshaping Ticket Based Services using Process Mining
12:15 Lunch	
13:30 | Mario Meir-Huber (Microsoft) **Creating Value of Data**	
14:15 Break	
Data Analytics in Industrial Application Scenarios Chair: Elena Lloret Pastor **14:30 Edwin Yaqub (RapidMiner GmbH)** A Reference Architecture for Quality Improvement in Steel Production **15:00 David Arnu (RapidMiner GmbH)** Anomaly Detection and Structural Analysis in Industrial Production Environments **15:30 Johannes Scholz (Institute of Geodesy)** Semantically Annotated Manufacturing Data to support Decision Making in Industry 4.0: A Use-Case Driven Approach	**Application Scenarios** **14:30 Fabian Räbiger (Hagleitner Hygiene International GmbH)** Der intelligente Waschraum 4.0 **15:15 John A. Thompson (Information Professionals GmbH)** Creating Business Value with an Inter-Company Data Lake
16:00 Coffee Break	
Student Talks Chair: Andreas Unterweger **16:30 Dorian Prill** Improving Maintenance Processes with Data Science **17:00 Elisabeth Birnbacher, Marco Gruber** ouRframe - Ein graphisches Workflow-Tool für R **17:30 Dominik Hofer** Sentiment Analysis - A Student's Perspective Analysis	**Projects** **16:30 Klaas Wilhelm Bollhöfer (*um - The unbelievable Machine Company GmbH)** The Art of Data Maturity Modeling at Metro Group **17:15 Thomas Soboll (Porsche Informatik GmbH), Norbert Walchhofer (Porsche Austria GmbH & Co OG)** Big Data Applications in Automotive Retail - Lessons Learned
18:15 Farewell	

FutureTDM	Workshops
9:00 | Mike Olson (Cloudera) **Enabling Data Science in the Enterprise**	
9:45 | Bernhard Jäger (Synyo GmbH) **Introduction to the FutureTDM Project**	**9:45 | Steffen Märkl (Cloudera)** **Cloudera - Powering Possibilities in Machine Learning**
10:00 | Janek Strycharz (Projekt Polska Foundation) **The Economic Potential of Data Analytics**	
10:30 Break	
Data Analytics and the Legal Landscape: Intellectual Property and Data Protection **10:45 Marco Caspers (University Amsterdam)** Dealing with the Legal Bumps on the Road to Further TDM Uptake **12:05 Flash Presentations**	**10:45 | Federica Fusco (F&F GmbH)** **TensorFlow step-by-step**
12:15 Lunch/Demo Session	
13:30 | Mario Meir-Huber (Microsoft) **Creating Value of Data**	
14:15 Break	
Startups to Multinationals: An Overview of Future TDM case studies **14:30 Freyja van den Boom (Open Knowledge / Content Mine)** Stakeholder Consultations - the Highlights **Universities, TDM and the need for strategic thinking on educating researchers** **15:15 Kiera McNeice (British Library)** Supporting TDM in the Education Sector	
16:00 Coffee Break	
Technologies and infrastructures supporting Text and Data Analytics: challenges and solutions **16:30 Maria Eskevich (Radboud University):** The TDM Landscape: Infrastructure and Technical Implementation **17:15 Kiera McNeice (British Library)** Next Steps: A Roadmap to Promoting Greater Uptake of Data Analytics in Europe **17:40 Bernhard Jäger (Synyo GmbH)** Beyond FutureTDM	
18:15 Farewell	

Euro Beinat (CS Research, Salzburg University)

Euro Beinat holds a Masters in Computer Science and a PhD in Economics. He is the Managing Director/ Chairman of CS Research and teaches Geoinformatics and Data Science at the University of Salzburg, where he is a Board Member of the GI Doctoral College. Euro has an extensive experience in industry and business, ranging from startups to global corporations: until recently he held the position of Vice President Internet of Things for Zebra Technologies Corporation in Chicago. He has worked for more than 15 years in data science, predictive analysis, and emerging technologies such as the Internet of Things. As a recognized expert in these fields, he advises executives in public and private organizations on technology innovation and strategy making. Euro is a regular speaker at scientific, business and media conferences and has published over fifty papers and books on topic ranging from geospatial databases, data analytics, natural resource management and strategy making.

Mario Meir-Huber (Microsoft Austria)

Mario Meir-Huber, Solution Professional at Microsoft, is a professional at Big Data technology. He has a lot of experience in the technological aspects of Big Data as well as the process of consulting. In his career Mario has lead several Consulting teams which offered Big Data and Hadoop Solutions to customers. Mario is the author of 2 books about Cloud Computing and several e-books with the topic Big Data. He also writes the Blog – cloudvane.com – which is on the list of top 100 blogs about Big Data.

Mike Olson (Cloudera)

Mike Olson co-founded Cloudera in 2008 and served as its CEO until 2013 when he took on his current role of chief strategy officer (CSO). As CSO, Mike is responsible for Cloudera's product strategy, open source leadership, engineering alignment and direct engagement with customers. Prior to Cloudera Mike was CEO of Sleepycat Software, makers of Berkeley DB, the open source embedded database engine. Mike spent two years at Oracle Corporation as vice president for Embedded Technologies after Oracle's acquisition of Sleepycat in 2006. Prior to joining Sleepycat, Mike held technical and business positions at database vendors Britton Lee, Illustra Information Technologies and Informix Software. Mike has a Bachelor's and a Master's Degree in Computer Science from the University of California, Berkeley.

Ralf Klinkenberg (RapidMiner

Ralf Klinkenberg is the Co-Founder and Head of Data Science Research at RapidMiner. Ralf is a data scientist and consultant with more than 20 years of experience in data mining and predictive analytics. He holds a Master of Science degree from Technical University of Dortmund in Germany and Missouri University of Science and Technology in the USA, and worked as a machine learning researcher at both universities. He co-founded both the open source data mining project RapidMiner, as well as the company RapidMiner, for which he currently serves as Head of Data Science Research. Ralf Klinkenberg has more than 20 years of experience in machine learning, data mining, text mining, web mining, predictive analytics, and their applications in diverse sectors and use cases.

Janek Strycharz (Digital Center Poland)

Janek is an economist with a long record of informing and evaluating public policies and programs. He is especially interested in the digitalization of societies and was involved in the assessment of Poland's programs aimed at digitalization of culture and development of digital competencies. He collaborates with the Digital Center (Poland) whose aim is to inform public decision makers on important social and economic issues connected to digitalization. In this capacity he worked on various analyses aiming at assessing digital school reforms in Poland as well as co-authored a report calculating the economic worth of Polish Open Data. He also co-founded Workshop for Social Innovation – an NGO that aims at connecting innovations with societal goals and needs.

The iDSC Industry Track showcases practitioners exploring the way Data Science is used in business today. In the mornings and the afternoons of the conference, the focus is laid on one of three broad areas, providing an opportunity for in-depth reflection and comparison of approaches.

Within each of these focus areas, experts describe their experience, show their practical solutions, and dare a look into the future of Data Science in business.

Projects

Which new skills and new development processes are required to effectively use Data Science?

Klaas Wilhelm Bollhöfer, Chief Evangelist (*um - The unbelievable Machine Company, Berlin & Vienna):
The Art of Data Maturity Modeling at Metro Group

Albin Gruber, Product Owner Management Information Systems (Wüstenrot Datenservice GmbH, Salzburg, Austria):
Data Warehousing@Wüstenrot - Structured Data as Basis for Reporting and Analytics

Eric Rozier, Assistant Professor (Iowa State University, Department of Computer Science, Ames, Iowa, USA):
Securing Against Data Integrity Attacks at The World Bank

Thomas Soboll, Department Head, Product Management Automotive Smart Devices; Digital Innovation Manager (Porsche Informatik GmbH, Salzburg, Austria)
Norbert Walchhofer, Digital Solutions (Porsche Austria GmbH & Co OG, Salzburg, Austria):
Big Data Applications in Automotive Retail - Lessons Learned

Technology

How can innovative tools and technology be applied to solve real-world problems?

Tomas Knap, Software Architect & Researcher (Semantic Web Company GmbH, Vienna, Austria):
Preparing DBpedia Knowledge Graph using UnifiedViews, an ETL tool for RDF data

Dmytro Martynenko, Application Engineer (The MathWorks GmbH, Aachen, Germany):
Building a Big Engineering Data Analtyics System Using MATLAB

Stephan Schiffner, Head of Big Data & Advanced Analytics;
Markus Breunig, Lead Consultant Big Data & Advanced Analytics;
(F&F GmbH, Munich, Germany):
Data Analysis with Spark – From Raw Data to Insights

Application Scenarios

What results have been achieved with Data Science in meeting stakeholder needs and creating business value?

Tamas Molnar, Advanced Analytics Team Leader (Vodafone Group BI, Budapest, Hungary):
Reshaping Ticket Based Services Using Process Mining

Fabian Räbiger, Produktmanager waschraumHYGIENE (HAGLEITNER HYGIENE INTERNATIONAL GmbH, Zell am See, Austria):
Der intelligente Waschraum 4.0

Thomas Thalhammer, Head of IT Hervis, Enterprise Architect (SPAR Business Services GmbH, Salzburg, Austria):
Digital Business at SPAR Business Services GmbH

John A. Thompson, Co-Founder and Managing Director (Information Professionals GmbH, Freilassing, Germany):
Creating Business Value with an Inter-Company Data Lake

Workshop 1 - Tuesday, June 13th, 9:45

Cloudera: Powering Possibilities in Machine Learning

Steffen Märkl, Systems Engineer, Cloudera

Machine learning is all about the data, but it's often out of reach for analytics teams working at scale. Cloudera customers such as Wargaming.net can store, process and analyse 550 million events each day to help them improve gamers' experiences and increase customer lifetime value.

Whether you are new to machine learning and advanced analytics, or you already take advantage of the possibilities, this interactive workshop will explore practical examples and give you some new ideas to take away. Discover how enterprise organisations can accelerate machine learning from exploration to production by empowering their data scientists with R, Python, Apache Spark and more in one secure, unified and collaborative platform.

Workshop 2 - Tuesday, June 13th, 10:45

F&F GmbH: TensorFlow step-by-step

Dr. Federica Fusco, Data Scientist, F&F GmbH

Nowadays deep learning is everywhere and along with it TensorFlow, Google's deep learning library, which has widely spread since its first release at the end of 2015. Learning TensorFlow can be quite challenging mainly for two reasons. On the one hand the low-level of this tool introduces some obvious difficulties, on the other hand the available code examples are sometimes hard to understand for beginners, especially if they are related to a previous version of the software. Hence, in this presentation we will introduce TensorFlow from scratch. We will start from a couple of simple lines of code, and will arrive at the implementation of a full neural network. Every line of code will be described and live executed. Our purpose is to give a deep insight into TensorFlow, in order to make it accessible to everybody who is interested in learning this tool. A standard TensorFlow execution starts with the definition of the computational graph. This graph defines the placeholders of the input parameters, which have to be fed with data on execution, and the operations that one wants to execute. Once the definition is complete, a session, i.e. a runtime context for executing the graph, is launched. We will start the presentation from a very simple graph definition, in order to solve a multinomial logistic regression. This example will give the opportunity to understand the basics of graphs, sessions, and more generally TensorFlow itself. The way we implement the logistic regression, it can be seen as a very simple neural network (NN), consisting of one layer of neurons, a softmax activation function and a categorical crossentropy cost function. Therefore, to conclude our presentation we will present a NN, which will be composed of a few convolutional layers. In particular, we will focus on the architecture of the network, the input and the output of every layer, and we will use it to perform a simple classification task.

Cesar Ojeda, Rafet Sifa and Christian Bauckhage (Fraunhofer IAIS): Circadian Cycles and Work Under Pressure: A Stochastic Process Model for E-learning Population Dynamics

Abstract—Web analytics techniques designed to quantify Web usage patterns allow for a deeper understanding of human behavior. Recent models of human behavior dynamics have shown that, in contrast to randomly distributed events, people engage in activities which show bursty behavior. In particular, participation in online courses often shows periods of inactivity and procrastination followed by frequent visits shortly before examination deadlines. Here, we propose a stochastic process model which characterizes such patterns and incorporates circadian cycles of human activities. We validate our model against real data spanning two years of activity on a university course platform. We then propose a dynamical model which accounts for both periods of procrastination and work under pressure. Since circadian and procrastination-pressure cycles are fundamental to human activities, our method can be extended to other tasks such as analyzing browsing behaviors or customer purchasing patterns.

Cesar Ojeda, Rafet Sifa and Christian Bauckhage (Fraunhofer IAIS): Investigating and Forecasting User Activities in Newsblogs: A Study of Seasonality, Volatility and Attention Burst

The study of collective attention is a major topic in the area of Web science as we are interested to know how a particular news topic or meme is gaining or losing popularity over time. Recent research focused on developing methods which quantify the success and popularity of topics and studyied their dynamics over time. Yet, the aggregate behavior of users across content creation platforms has been largely ignored even though the popularity of news items is also linked to the way users interact with the Web platforms. In this paper, we present a novel framework of research which studies the shift of attentions of population over newsblogs. We concentrate on the commenting behavior of users for news articles which serves as a proxy for attention to Web content. We make use of methods from signal processing and econometrics to uncover patterns in the behaviour of users which then allow us to simulate and hence to forecast the behavior of a population once an attention shift occurs. Studying a data set of over 200 blogs with 14 million news posts, we found periodic regularities in the commenting behavior. Namely, cycles of 7 days as well as 24 days of activity which may be related to known scales of meme lifetimes.

Norman Ihle and Axel Hahn (OFFIS e.V. – Institute for Information Technology): Knowledge-based Short-Term Load Forecasting for Maritime Container Terminals

With the rise of Demand Response and Demand Side Management in modern energy systems Short-Term Load Forecasting for single industrial customers is getting increased attention. For industrial sites it seems promising to integrate the knowledge of planned operations into the next day's energy-consumption forecasting process. In the case of a maritime container terminal these operation plans are based on the list of ship arrivals and departures. In this paper two approaches are introduced that integrate this knowledge in different ways: while Case-Based Reasoning as a lazy-learner uses it during the forecasting process, an Artificial Neural Network has to be trained before the actual forecasting process can occur. It can be shown that the integration of more knowledge into the forecasting process enables better results in terms of forecast accuracy

Cesar Ojeda, Rafet Sifa, Christian Bauckhage and Kostadin Cvejoski (Fraunhofer IAIS): **Beyond Spectral Clustering: A Comparative Study of Community Detection for Document Clustering**

Document clustering is an ubiquitous problem in data mining as text data is one of the most common forms of communication. The richness of the data require methods tailored to different tasks, depending on the characteristics of the information to be mined. In recent years graph-based methods have appeared that allow for hierarchical, fuzzy, and non Gaussian density features to identify structures in complicated data sets. In this paper we present a novel methodology for the clustering of documents based on a graph defined over a vector space model. We make use of an overlap hierarchical algorithm and show the equivalence of our quality function to that of Ncut. We compare our method to spectral clustering and other graphbased models and find that our method provides a good and flexible alternative for news clustering, whenever fine grained details between topics are required.

Cesar Ojeda, Shubham Agarwal, Rafet Sifa and Christian Bauckhage (Fraunhofer IAIS): **Third Party Effect: Community Based Spreading in Complex Networks**

A substantial amount of network research has been devoted to the study of spreading processes and community detection without considering the role of communities in the characteristics of spreading processes. Here, we generalize the SIR model of epidemics by introducing a matrix of community infecting rates to capture the heterogeneous nature of the spreading as defined by the natural characteristics of communities. We find that the spreading capabilities of one community towards another is influenced by the internal behavior of third party
communities. Our results provide insights into systems with rich information structure and into populations with diverse immunology responses.

David Anastasiu (San José State University): **Cosine Approximate Nearest Neighbors**

Cosine similarity graph construction, or all-pairs similarity search, is an important kernel in many data mining and machine learning methods. Building the graph is a difficult task. Up to n2 pairs of objects should be naively compared to solve the problem for a set of n objects. For large object sets, approximate solutions for this problem have been proposed that address the complexity of the task by retrieving most, but not necessarily all, of the nearest neighbors. We propose a novel approximate graph construction method that leverages properties of the object vectors to effectively select few comparison candidates, those that are likely to be neighbors. Furthermore, our method leverages filtering strategies recently developed for exact methods to quickly eliminate unpromising comparison candidates, leading to few overall similarity computations and increased efficiency. We compare our method against several state-of-the-art approximate and exact baselines on six real-world datasets. Our results show that our approach provides a good tradeoff between efficiency and effectiveness, showing up to 35.81x efficiency improvement over the best alternative at 0.9 recall.

Cornelia Ferner, Werner Pomwenger, Martin Schnöll, Veronika Haaf, Arnold Keller and Stefan Wegenkittl (Salzburg University of Applied Sciences): Information Extraction Engine for Sentiment-Topic Matching in Product Intelligence Applications

Product reviews are a valuable source of information for companies and customers alike. While companies use this information to improve their products, customers need it for decision support. Online shops often provide reviews, opinions and additional information to encourage customers to buy on their site. However, current online review implementations often lack a quick overview of how well certain product components meet customer preferences making product comparison difficult. Therefore, we have developed a product intelligence tool that combines state-of-the-art technologies into a natural language processing engine. The engine is capable of collecting and storing product-related online data, extracting metadata and analyzing sentiments. The engine is applied to technical online product reviews for component-level sentiment analysis. The fully automated process crawls the web for expert reviews, extracts sequences related to product components and aggregates the sentiment values from the reviews.

Eduardo Brito, Rafet Sifa, Kostadin Cvejoski, Cesar Ojeda and Christian Bauckhage (Fraunhofer IAIS): Towards German Word Embeddings: A Use Case with Predictive Sentiment Analysis

Despite the research boom on words embeddings and their text mining applications from the last years, the vast majority of publications focus only on the English language. Furthermore, hyperparameter tuning is a rarely well documented process (specially for non English text) that is necessary to obtain high quality word representations. In this work, we present how different hyperparameter combinations impact the resulting
German word vectors and how these word representations can be part of more complex models. In particular, we perform first an intrinsic evaluation of our German word embeddings, which are later used within a predictive sentiment analysis model. The latter does not only serve as an extrinsic evaluation of the German word embeddings but also shows the feasibility of predicting preferences only from document embeddings.

Wassim El-Hajj, Ghassen Ben Brahim, Cynthia El-Hayek and Hazem Hajj (American University of Beirut): Feature Extraction and Large Activity-Set Recognition Using Mobile Phone Sensors

In this work, the problem of activity recognition using data collected from the user's mobile phone is being addressed. We start with reviewing and discussing the limitations of most state of the art activity recognition approaches for mobile phone devices. Then, we present our approach recognizing a large set of activities that are comprehensive enough to cover most activities users engage in. Moreover, multiple environments are supported, for instance, home, work, and outdoors. Our approach suggests a single-level classification model that is accurate in terms of activity classification, comprehensive in terms of the large number of activities being covered, and applicable in the sense that it can be used in real settings. In the literature, these three properties are not existent altogether in a single approach. Existing approaches normally optimize their models for either one or, at a maximum two of the following properties: accuracy, comprehensiveness and applicability.

Nikola Obrenovic, Goran Vidakovic and Ivan Lukovic (Schneider Electric DMS NS Llc.): **The Choice of Metric for Clustering of Electrical Power Distribution Consumers**

An important part of any power distribution management system data model is a model of load type. A load type represents typical load behaviour of a group of similar consumers, e.g. a group of residential, industrial or commercial consumers. A common method for creation of load types is the clustering of individual energy consumers based on their yearly consumption behaviour. To reach the satisfactory level of load type quality, the crucial decision is a choice of proper clustering similarity measure. In this paper, we present a comparison of different metrics, used as similarity measures in our process of load type creation. Additionally, we present a novel metric, also included in the comparison. The metrics and the quality of load types created therewith are assessed by using a real data set obtained from the distribution network smart meters.

Erwin Filtz, Axel Polleres, Roman Karl and Bernhard Haslhofer (Vienna University of Economics and Business): **Evolution of the Bitcoin Address Graph**

Bitcoin is a decentralized virtual currency, which can be used to execute pseudo-anonymous payments globally within a short period of time and comparably low transaction costs. In this paper, we present initial results of a longitudinal study conducted over the Bitcoin address graph, which contains all addresses and transactions from the beginning of Bitcoin in January 2009 until 31st of August 2016. Our analysis reveals a highly-skewed degree distribution with a small number of outliers and illustrates that the entire graph is expanding rapidly. Furthermore, it demonstrates the power of address clustering heuristics for identifying real-world actors, who prefer to use Bitcoin for transferring rather than storing value. We believe that this paper provides novel insight into virtual currency ecosystems, which can inform the design of future analytics methods and infrastructures.

David Arnu, Edwin Yaqub, Claudio Mocci, Valentina Colla, Marcus Neuer, Gabriel Fricout, Xavier Renard, Patrick Gallinari and Christophe Mozzati (RapidMiner GmbH): A Reference Architecture for Quality Improvement in Steel Production

There is a global increase in demand for steel, but steel manufacturing is a highly sophisticated and costly process where good quality is hard to achieve. Improving the quality remains a major challenge faced by the steel industry. The EU project PRESED (Predictive Sensor Data mining for Product Quality Improvement) addresses this challenge by focusing on widespread recurring problems. The variety and veracity of data, as well as the change in properties of the observed material complicates the interpretation of data. In this paper, we present the reference architecture of PRESED, which is being purpose-built to address the vital concerns of managing and operationalizing the data. The architecture leverages big and smart data concepts with data mining algorithms. Data preprocessing and predictive analytics tasks are supported by means of a malleable data model.

The approach allows the users to design processes and evaluate multiple algorithms pertinent to the problem at hand. The concept is to store and harness the complete production data instead of relying on aggregated values. Early results on data modeling show that fine grained preprocessing of time series data through feature extraction and predictions provide superior insights than traditionally used aggregation statistics.

Martin Atzmüller, David Arnu and Andreas Schmidt (RapidMiner GmbH): Anomaly Detection and Structural Analysis in Industrial Production Environments

Detecting anomalous behavior can be of critical importance in an industrial application context. While modern production sites feature sophisticated alarm management systems, they mostly react to single events. Due to the large number and various types of data sources a unified approach for anomaly detection is not always feasible. One prominent type of data are log entries of alarm messages. They allow a higher level of abstraction compared to raw sensor readings. In an industrial production scenario, we utilize sequential alarm data for anomaly detection and analysis, based on first-order Markov chain models. We outline hypothesis-driven and description-oriented modeling options. Furthermore, we provide an interactive dashboard for exploring and visualization of the results.

Stefan Schabus and Johannes Scholz (Graz University of Technology): Semantically Annotated Manufacturing Data to support Decision Making in Industry 4.0: A Use-Case Driven Approach

Smart Manufacturing or Industry 4.0 is a key approach to increase productivity and quality in industrial manufacturing companies by automation and data driven methods. Smart manufacturing utilizes theories from cyber-physical systems, Internet of Things as well as cloud computing. In this paper, the authors focus on ontology and (spatial) semantics that serve as technology to ensure semantic interoperability of manufacturing data. Additionally, the paper proposes to structure production relevant data by the introduction of geography and semantics as ordering dimensions. The approach followed in this paper stores manufacturing data from different IT-systems in a graph database. During the data integration process, the system semantically annotates the data – based on an ontology, developed for that purpose - and attaches spatial information. The approach presented in this paper facilitates an analysis of manufacturing data in terms of semantics and the spatial dimension. The methodology is applied to two use-cases of a semiconductor manufacturing company. The first use-case deals with the data analysis for incident analysis utilizing semantic similarities. The second use-case supports decision making in the manufacturing environment, by the identification of potential bottlenecks in the semiconductor production line.

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