European Data Science Conference

November 07-08, 2016
in Luxembourg

We are grateful for the support of our sponsors

Fondation
Alphonse Weicker

Fonds National de la
Recherche Luxembourg

Table of Contents

Welcome Notes 4

Committees 6

Attendees 7

Practical Information 8

Scientific Programme 12

Abstracts 15

Welcome Note of the President of the University of Luxembourg, Rainer Klump

Dear conference participants,

Welcome to Luxembourg and to the European Data Science Conference 2016. The University of Luxembourg is a multilingual European research university. We are a young institution, proud of our personal atmosphere, close to European institutions, innovative companies and the financial sector. With nearly 6,200 students and about 1,600 employees from all over the globe, we offer a unique mix of international excellence and national relevance, delivering knowledge for society and businesses. Our strategic framework defines ambitious goals for the next decade. We strive for international leadership in curiosity and mission-driven research and aim to produce entrepreneurial-driven graduates enabled to face challenges of the 21st century.

The 21st century is characterised by an increasing complexity that is driven by the ever-growing data-landscape. Managing this information overflow is one of the key challenges of our time. Universities, as the main providers of higher education, play a crucial role in assuring that societies, including all actors from the private and institutional sector, are equipped with the appropriate skill-set to transform these challenges into opportunities.

At the University of Luxembourg, we have the potential to actively contribute to this transformation: Constant innovation is visually reflected by the relocation from the University's former three sites Kirchberg, Limpertsberg, and Walferdange to the "City of Science" in Belval. In Belval, Luxembourg is redefining itself. Starting at the beginning of the 20th century as a steel region, reforming to a financial centre in the 1960s, Luxembourg has now become a protagonist of the modern knowledge society. Innovation is also embedded in our strategic framework, which foresees a comprehensive digitalisation of all parts of the university as well as the transversal teaching of 21st century skills, including digital literacy. Computational Sciences as an interdisciplinary research field has been a priority of the University since 2014 and enjoys an international reputation.

The European Data Science Conference is set in this exciting scientific environment with close connections to the Luxembourg Centre of Systems Biomedicine and the Interdisciplinary Centre for Security, Reliability and Trust, as well as the Max Planck Institute Luxembourg for International, European and Regulatory Procedural Law, STATEC and Eurostat; institutions of central importance for the agenda of this conference. It will provide a forum for the exchange of important developments within the field of Data Science, which are needed to analyse empirical data in a great number of disciplines.

I address my special thanks to the conference organisers Sabine Krolak-Schwerdt, Matthias Böhmer and their teams, as well as to colleagues from the Luxembourg School of Finance, the Centre for Research in Economics and Management, and the Computer Science research unit. I wish you an inspiring conference and a pleasant time in Luxembourg.

Sincerely,

Prof. Dr. Rainer Klump

President of the University of Luxembourg

Welcome Note of the President of the European Association for Data Science (EuADS), Sabine Krolak-Schwerdt

Dear Colleagues,

It is both my privilege and my pleasure to welcome you to the European Data Science Conference (EDSC 2016) in Luxembourg. The conference is hosted by the University of Luxembourg, a highly international institution that aims at excellence in research and education. The EDSC 2016 is the inaugural conference of the European Association for Data Science (EuADS) which was founded one year ago in Luxembourg. EuADS is a forum for research, scientific conduct, promoting transparent access to data, policy making as well as support of early career scientists. Scientifically, the association promotes the development of models, methods and instruments of Data Science and provides a unique environment for the presentation and discussion of new methods and models in this field. With regard to networking, it aims to bring together data scientists and decision-makers in this field to address the challenges of Data Science in the 21st century.

Correspondingly, the major purpose of the EDSC 2016 conference is to initiate and foster discussion on relevant priorities in Data Science that are of interest to empirical disciplines, and therefore, the scientific programme committee has invited outstanding scientists in the field of acquisition and analysis of (big) empirical data, representatives from industry and other stakeholders from the domain of Data Science. The EDSC 2016 brings together more than 90 renowned experts and offers a scientific programme of selected plenary talks, associated symposia and panels as well as an opening for the interested public. The major areas involved are Biomedicine and Biostatistics, Computational Sciences, Economics, Educational Sciences, Law, Medical/Health Sciences, Psychology, Social Sciences, Statistics and Mathematics as well as Business. The programme of the conference focuses on the following five thematic fields which evolved from a careful analysis of approaches and unsolved problems across disciplines:

- Science, Statistics and Society
- Legal Dimensions of Data Science
- Structure of Data Science
- Medicine and Healthcare
- What makes Data Science different?

In each of the fields, we are proud to present internationally leading experts. Organising such a conference requires the coordination of myriad people and topics; I would like to thank the members of the EDSC scientific programme committee and the local organisation committee as well as the board of the EuADS that have worked tirelessly to realise the European Data Science Conference. I would also like to express my thanks to the Fonds National de la Recherche, the Fondation Alphonse Weicker, IBM and PricewaterhouseCoopers Luxembourg for their generous support of the conference.

I wish you a productive, inspiring conference and a pleasant stay in Luxembourg!

Sabine Krolak-Schwerdt

Prof. Dr. Sabine Krolak-Schwerdt

Conference General Chair and EuADS Chair Person

Scientific Programme Committee

Serge Allegrezza, STATEC

Eric Dubois, Luxembourg Institute of Science and Technology

Peter A. Flach, University of Bristol

Burkhard Hess, Max Planck Institute Luxembourg

Katja Ickstadt, TU Dortmund University

Sabine Krolak-Schwerdt, University of Luxembourg

Berthold Lausen, University of Essex

Hilmar Schneider, Institute for the Study of Labor

Myra Spiliopoulou, Otto von Guericke University Magdeburg

Claus Weihs, TU Dortmund University

Local Organising Committee

Serge Allegrezza, STATEC

Matthias Böhmer, University of Luxembourg

Eric Dubois, Luxembourg Institute of Science and Technology

Burkhard Hess, Max Planck Institute Luxembourg

Sabine Krolak-Schwerdt, University of Luxembourg

Hilmar Schneider, Institute for the Study of Labor

Attendees

Serge Allegrezza, STATEC / Cordula Artelt, University of Bamberg / Rudi Balling, LCSB - Luxembourg Centre for Systems Biomedicine, University of Luxembourg / Oumayma Banouar, Cadi Ayyad University / Michael Berthold, University of Konstanz / Dirk Betz, GESIS - Leibniz-Institute for the Social Sciences / Bernd Bischl, LMU Munich / Claudia Biniossek, GESIS - Leibniz-Institute for the Social Sciences / Hendrik Blockeel, KU Leuven / Matthias Böhmer, University of Luxembourg / Stéphane P. A. Bordas, University of Luxembourg / Manel Brinchi, Luxembourg Institute of Science and Technology / Pierrick Bruneau, Luxembourg Institute of Science and Technology / Rüdiger Budde, RWI - Leibniz-Institut für Wirtschaftsforschung / Pedro Cardoso-Leite, University of Geneva / Christelle Cocco, University of Lausanne / Mark Cole, University of Luxembourg / Ronald Cornet, Academic Medical Center, Amsterdam; Linköping University / Ricardo João Cruz Correia, University of Porto / Pedro Cruz Villalón, Autonomous University of Madrid / Ronny Bölter, ZPID - Leibniz Institute for Psychology Information / Sergio V. Davalos, University of Washington / Luc De Raedt, KU Leuven / Reinhold Decker, Bielefeld University / Eric Dubois, Luxembourg Institute of Science and Technology / Arnaud Dupuy, University of Luxembourg / Monjed Ezzedine, EDHEC Business School / Daniela Ferreira dos Santos, University of Porto / Peter A. Flach, University of Bristol / Dandolo Flumini, ZHAW Zurich University of Applied Sciences / Rudolf Marcel Füchslin, ZHAW Zurich University of Applied Sciences / Andreas Geyer-Schulz, Karlsruhe Institute of Technology / Fosca Giannotti, Information Science and Technology Institute of the National Research Council – ISTI-CNR / Bart Goethals, University of Antwerp / Audrey Guinchard, University of Essex / Mateusz Guzek, Goodyear Innovation Center* Luxembourg / Georg Herde, Deggendorf Institute of Technology / José Hernández-Orallo, Technical University of Valencia / Burkhard Hess, Max Planck Institute Luxembourg / Paul Heuschling, University of Luxembourg / Lars Hofmann, IT.NRW / Katja Ickstadt, TU Dortmund University / Göran Kauermann, LMU Munich / Martin Kerwer, ZPID - Leibniz Institute for Psychology Information / Hans Kestler, University of Ulm / Joost Kok, Leiden University / Roland Krause, LCSB - Luxembourg Centre for Systems Biomedicine, University of Luxembourg / Sabine Krolak-Schwerdt, University of Luxembourg / Michel Lang, TU Dortmund University / Catherine Larue, Luxembourg Institute of Health / Berthold Lausen, University of Essex / Christophe Ley, Ghent University / Aline Muller, Luxembourg Institute of Socio-Economic Research / Fionn Murtagh, University of Derby, Goldsmiths University of London / Benoît Otjacques, Luxembourg Institute of Science and Technology / James Pang, Business Analytics Center, School of Computing, National University of Singapore / Dino Pedreschi, University of Pisa / Niels Peek, University of Manchester / Pedro Pereira Rodrigues, University of Porto / Walter Radermacher, Eurostat, European Commission / Peter Ryan, SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg / Abdel Salhi, University of Essex / John Samuel, University of Lyon / Jorge Sanz, Business Analytics Center, National University of Singapore / Gilbert Saporta, Conservatoire National des Arts et Métiers, Paris / Heather Savory, Office for National Statistics / Jang Schiltz, University of Luxembourg / Hilmar Schneider, Institute for the Study of Labor / Michèle Sebag, National Center for Scientific Research – CNRS / Arno Siebes, Utrecht University / Myra Spiliopoulou, Otto von Guericke University Magdeburg / Gerald Spindler, University of Göttingen / Detlef Steuer, Helmut Schmidt University / Rudolf Walter Strohmeier, Publications Office of the European Union, European Commission / Paula Souza, University of Luxembourg / Antonio J. Tallón-Ballesteros, University of Seville / Hannu Toivonen, University of Helsinki / Katrijn van Deun, Tilburg University / Philippe van Kerm, Luxembourg Institute of Socio-Economic Research / Maurizio Vichi, Sapienza University of Rome / Marco Viviani, University of Milano-Bicocca / Jan von Hein, University of Freiburg / Claudia Wagner, GESIS - Leibniz-Institute for the Social Sciences / Claus Weihs, TU Dortmund University / Nico Weydert, STATEC

Practical Information

Conference venue

Abbey of Neumünster Cultural Exchange Center
28, Rue Münster
L-2160 Luxembourg-Grund
Phone: +352 26 20 52 1

How to get to the venue

By foot

At the plateau "Saint-Esprit" in the city centre of Luxembourg a public lift gives access to the Grund quarter. When leaving the lift, go straight, cross the small bridge ahead, then turn left into *rue Münster*. Neumünster Abbey is located at the end of the street.

By car

We strongly recommend using the upper city parking facilities or the car park "Brasserie" (Rives de Clausen, L-2165 Luxembourg). From the car park "Saint-Esprit" (Plateau Saint Esprit, L-1475 Luxembourg) – the first car park in the city centre after the Old Bridge (Luxembourg Viaduct) – a public lift gives access to the Grund quarter. When leaving the lift, go straight, cross the small bridge ahead, then turn left into *rue Münster*. Neumünster Abbey is located at the end of the street. From the car park "Brasserie", walk up *rue de la Tour Jacob* to the stairs of the Wenzel pathway which descends towards Neumünster Abbey.

By bus

From Luxembourg central station

When leaving the train station, turn right and go to the bus terminal. Line 23 (destination Beggen/Henri Dunant) leaves every 20 minutes from platform 7. Leave the bus at the bus stop "Stadgronn-Bréck", cross the small bridge, then turn left into *rue Münster*. Neumünster Abbey is located at the end of the street.

From Luxembourg airport

You'll find the bus stop in front of the airport terminal. Take line 16 or 29 to Luxembourg central station. At the station, look for platform 7 and take line 23 (destination Beggen/Henri Dunant) which runs every 20 minutes. Leave the bus at the bus stop "Stadgronn-Bréck", cross the small bridge, then turn left into rue Münster. Neumünster Abbey is located at the end of the street.

Bus tickets can be purchased from the driver and are available at ticket vending machines. A short-term ticket ('billet courte durée') is 2 EUR and is valid for two hours throughout the country.

By taxi

Webtaxi – Telephone: (+352) 27 515; online booking available at http://www.webtaxi.lu
Taxi Colux – Telephone: (+352) 48 22 33
ALOTAXI – Telephone: (+352) 28 37 18 73

Internet access at the venue

Free WiFi access is available during the whole conference.

Network: neimenster
Login: free
Password: sekpekig

Venue for conference dinner

MAHO Rive droite
2, Place Sainte Cunégonde
L-1367 Luxembourg – Clausen
Phone: +352 27 04 83 71

For your convenience we are happy to inform you that a shuttle service will be provided:

The bus to the restaurant will leave at 19:10 sharp at the Golden Lady ("Gëlle Fra") Memorial (Place de la Constitution, 1478 Luxembourg). Please arrive at the meeting point a few minutes earlier and don't forget to bring your shuttle voucher which you will find in your conference folder.

Overview of the Scientific Programme

Monday, 07 November 2016	
12:00-13:00	Registration and Light Lunch
13:00-13:20	Opening Ceremony
13:20-14:15	Public Talk "Big Data and Business Analytics: Accelerating Digital Transformation in Enterprises and Industries"
14:15-14:30	Coffee Break
14:30-16:00	Symposium "Legal dimensions of Data Science"
16:00-17:00	Plenary Talk: "3S: Science, Statistics and Society"
17:00-18:00	Poster Session
19:30	Conference Dinner at MAHO restaurant

Tuesday, 08 November 2016	
09:00-10:00	Plenary Talk: "Data Science for the Masses: Mission Impossible?"
10:00-11:30	Discussion Session "Structure of Data Science"
11:30-12:30	Walking Lunch and Round Table with Early-Career Scientists
12:30-13:30	Workshop "Data Science in Medicine and Healthcare" Part 1: Plenary debate on "Three controversies about Data Science for Medicine and Healthcare"
13:30-14:30	Workshop "Data Science in Medicine and Healthcare" Part 2: Panel Session on "Data Science for Personalized Medicine"
14:30-15:00	Coffee Break
15:00-16:00	Panel Session "What makes Data Science different?"
16:00-16:30	Plenary Wrap Up and Closing
16:30	Vin d'honneur

Scientific Programme

Monday, 07 November 2016		Page
12:00-13:00	**Registration and Light Lunch**	
13:00-13:20	**Opening Ceremony** Sabine Krolak-Schwerdt, President of the European Association for Data Science Rainer Klump, President of the University of Luxembourg Michèle Weber, Programme Manager & Science Communicator at the FNR Luxembourg	
13:20-14:15	**Public Talk** James Pang (Business Analytics Center, School of Computing, National University of Singapore) and Jorge Sanz (National University of Singapore) "Big Data and Business Analytics: Accelerating Digital Transformation in Enterprises and Industries" Session Chair: Sabine Krolak-Schwerdt (University of Luxembourg)	16
14:15-14:30	**Coffee Break**	
14:30-16:00	**Symposium "Legal dimensions of Data Science"** Topic 1 "Social media and the protection of privacy" by Jan von Hein (University of Freiburg), Discussant: Pedro Cruz Villalón (Autonomous University of Madrid) Topic 2 "Cross border exchanges of data" by Gerald Spindler (University of Göttingen) Discussant: Mark Cole (University of Luxembourg) Session Chair: Burkhard Hess (Max Planck Institute Luxembourg)	17
16:00-17:00	**Plenary Talk: "3S: Science, Statistics and Society"** Walter Radermacher (Eurostat, European Commission) Session Chair: Peter Ryan (SnT - Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg)	19

17:00-18:00	**Poster Session** with Oumayma Banouar (Cadi Ayyad University), Dirk Betz and Claudia Biniossek (GESIS - Leibniz-Institute for the Social Sciences), Christelle Cocco (University of Lausanne), Ronny Bölter and Martin Kerwer (ZPID - Leibniz Institute for Psychology Information), Monjed Ezzedine (EDHEC Business School), Daniela Ferreira dos Santos (University of Porto), Dandolo Flumini (ZHAW Zurich University of Applied Sciences), Andreas Geyer-Schulz (Karlsruhe Institute of Technology), Audrey Guinchard (University of Essex), Georg Herde (Deggendorf Institute of Technology), Michel Lang (TU Dortmund University), Fionn Murtagh (University of Derby / Goldsmiths University of London), Abdel Salhi (University of Essex), John Samuel (University of Lyon), Antonio J. Tallón-Ballesteros (University of Seville) and Marco Viviani (University of Milano-Bicocca)	21
19:30	**Conference Dinner at MAHO restaurant**	

Tuesday, 08 November 2016		**Page**
09:00-10:00	**Plenary Talk** Michael Berthold (University of Konstanz) "Data Science for the Masses: Mission Impossible?" Session Chair: Peter A. Flach (University of Bristol)	28
10:00-11:30	**Discussion Session "Structure of Data Science"** with position statements from Manel Brichni (Luxembourg Institute of Science and Technology), Ronald Cornet (Academic Medical Center, Amsterdam / Linköping University), José Hernández-Orallo (Technical University of Valencia), Katja Ickstadt (TU Dortmund University) and Claus Weihs (TU Dortmund University), Benoît Otjacques (Luxembourg Institute of Science and Technology), Luc de Raedt (KU Leuven), Gilbert Saporta (Conservatoire National des Arts et Métiers, Paris) Session Chairs: Peter A. Flach (University of Bristol) and Claus Weihs (TU Dortmund University)	29
11:30-12:30	**Walking Lunch and Round Table with Early-Career Scientists**	

12:30-13:30	**Workshop "Data Science in Medicine and Healthcare"** Part 1: Plenary debate by Niels Peek (University of Manchester) and Pedro Pereira Rodrigues (University of Porto) on "Three controversies about Data Science for Medicine and Healthcare" Session Chairs: Berthold Lausen (University of Essex) and Myra Spiliopoulou (Otto von Guericke University Magdeburg)	40
13:30-14:30	**Workshop "Data Science in Medicine and Healthcare"** Part 2: Panel Session on "Data Science for Personalized Medicine" with Rudi Balling (LCSB - Luxembourg Centre for Systems Biomedicine, University of Luxembourg), Ricardo João Cruz Correia (University of Porto), Katja Ickstadt (TU Dortmund University), Hans Kestler (University of Ulm), Niels Peek (University of Manchester), Pedro Pereira Rodrigues (University of Porto) Session Chairs: Berthold Lausen (University of Essex) and Myra Spiliopoulou (Otto von Guericke University Magdeburg)	40
14:30-15:00	**Coffee Break**	
15:00-16:00	**Panel Session "What makes Data Science different?"** with Stéphane P. A. Bordas (University of Luxembourg), Andreas Geyer-Schulz (Karlsruhe Institute of Technology), Christophe Ley (Ghent University), Fionn Murtagh (University of Derby / Goldsmiths University of London) and Arno Siebes (Utrecht University) Session Chairs: Peter A. Flach (University of Bristol) and Berthold Lausen (University of Essex)	41
16:00-16:30	**Plenary Wrap Up and Closing** by Sabine Krolak-Schwerdt (University of Luxembourg)	
16:30	Vin d'honneur	

Monday - 07 November 2016

Big Data and Business Analytics: Accelerating Digital Transformation in Enterprises and Industries

James Pang and Jorge Sanz

National University of Singapore

Abstract. Business Analytics (BA) is an interdisciplinary domain that requires expertise from data analytics, IT and line-of-business knowledge in specific industries (including key processes that govern operations in enterprises). In this talk, we will review the recent development and progress in BA applications to different industries. It will cover typical BA use cases in different industries or lines-of-business, key challenges in BA technologies and industry competitive landscape. The future outlook and trends of BA will be also discussed.

Social Media and the Protection of Privacy

Jan von Hein

University of Freiburg

Abstract. The protection of privacy in the context of social media raises intricate and so far not fully resolved legal questions. Social media platforms such as Facebook or Twitter operate in an international environment; frequently, at least from a European point of view, the seat of a service provider, which is often based in the US, and the habitual residence of a user of social media diverge. Moreover, social media platforms are in most cases open to users from different jurisdictions. In cases of a violation of privacy (including claims for defamation), lawyers thus have to solve the problem which law applies to the dispute. The answer may be found in international contract law (regarding claims against the service provider), tort law (regarding other media users), or even international successions law (concerning the right to terminate a facebook account of a deceased person, for example). In addition, data protection rules belong to public law and may be characterized as overriding mandatory provisions that have to be taken into account regardless of the otherwise applicable law. In light of the inherently transnational character of the internet, the utility of traditional territorial connecting factors is significantly diminished, leading to the question as to whether new objective criteria should be developed or whether the scope of party autonomy should be further extended in cyberspace cases. In his presentation, the author analyzes these challenges from the perspective of European private international law.

Cross Border Exchanges of Data

Gerald Spindler

University of Göttingen

Abstract. The Internet is characterized by a global flow of data. In contrast data protection (as well as other regulations) usually centers on territories. Given the high level of data protection in Europe, one of the major issues always has been to safeguard the adequacy level of data protection in third-party countries outside the EU. Whereas the Date Protection Directive of 1995 focused still on the notion of establishment, thus creating troubles how to cope with IT players located outside the EU (such as Facebook), the new General Data Protection Regulation makes use of a de minimis approach. It will be then sufficient for applying the GDPR that goods or services are offered to EU citizens, regardless of the place where data processing is happening. Moreover, even a mere monitoring of behaviour of EU citizens will trigger the application of the GDPR, thus extending European Data Protection to IT players situated outside the EU. Hence, tools for ensuring the adequacy of data protection in third countries are crucial for any data flow between the EU and these countries (regarding personal data!). These tools encompass acknowledgements by the EU Commission of adequacy, standard contract clauses, binding corporate rules, or specific treaties such as the former Safe Harbour Treaty between the U.S. and the EU. However, as the ECJ lays stress on the fact in her decision on the Safe harbour Treaty that every tool has to ensure the same level of protecting fundamental rights of data protection. Hence, not only the safe harbour treaty has been attacked but also standard contract clauses etc. Crucial for any acknowledgement has become the effective enforcement of data protection – which is not given just by mere self-certification.

3S: Science, Statistics and Society

Walter Radermacher

Eurostat, European Commission

Abstract. "Science and technology permeate the culture and politics of modernity".[1] Social science literature in the field of science and technology (STS) has made considerable progress, however without much attention from the community of official statisticians, which in itself is a surprise.

Essentially, statistics is the science of learning from data. Certainly, it is a modern technology that is part of the standards of today's information age and society and is used in a wide array of practical fields. 'Official statistics' is one of these domains of applications, belonging to those with the longest history. Since the beginning of the 19th century, official statistics - as one child of the enlightenment - has grown and developed side by side with the different forms of the (modern) state. Desrosières uses the term "mutual co-construction" for three interlinked phenomena, A) a theory of the state (economy), B) interventions of the state and C) statistical measurements of 'variables' specifically targeted by policy measures[2].

The relationship between official statistics as a technology and the society has been analysed by a small community of historically interested scholars working in the intersection of statistics, sociology, political and historical sciences. Standard literature, such as "The Politics of Large Numbers - A History of Statistical Reasoning"[3], "Trust in Numbers: The Pursuit of Objectivity in Science and Public Life"[4] or „The Mutual Construction of Statistics and Society"[5] provides for a profound understanding of this two-way dynamic interaction between this specific technology on the one hand and societies, politics and economies on the other.

In order for official statistics to function as a language, a 'boundary object'[6] for all kinds of societal interactions and decision-making, it is essential that the quality of products and services is outstanding, an authority in itself. This is the brand-mark and the competitive advantage of official statistics. Once this authority is undermined, be it through real quality problems or only through perception, trust in official statistics will be replaced by suspicion and statistics will become part of political fights and games.

Against this background, it is important to define quality of official statistics with a much wider scope, including not only the production but also the use side of statistical information and how these two sides are interacting in a dynamic relationship.

The era of the data revolution has started, significantly changing the picture on both sides. On the one hand, the availability of enormous amounts of data gives the statistical business a completely new push into a direction that is not yet sufficiently understood. On the other hand, new demand in terms of 'evidence based decision making', new (public) management etc. create a forceful driving force on the pull side.

While uncertainties and risks are constantly growing in the eyes of citizens and while the impacts of globalisation become more and more visible, it appears as if people had enough of experts[7] and as if 'post-truth-politics[8]' would gain credibility and support, opening opportunities for populist and nationalist activists of all kind. In this context, a profound epistemological shift is needed since complexity and irreversibility undermine the idea that science can provide single, objective and exhaustive answers. In the late modernity of risk societies, there is the epistemic and methodological necessity to empower citizens with the appropriate insight in order to enable them to make appropriate decisions for achieving sustainability and pursuing resilience in a complex world: *"The tools for thought of the Enlightenment no longer suffice for mastering the challenges of the present. The course European societies are taking can be compared to the exploration journeys of bygone days. Maps, which ought to provide orientation and security, seem to have lost their value. We are journeying into the uncertain and have yet to discover new paths and routes in many areas.[9]"*

Under these conditions it seems to be absolutely necessary and urgent to launch a (scientific) debate and reflection in the community of official statisticians that is focussed on this area.

References

1 JASANOFF S.: States of knowledge – The coproduction of science and social order, Abington, 2004, pg. 1 and http://sts.hks.harvard.edu/about/whatissts.html

2 DESROSIÈRES A.: Birth of a new statistical language between 1940 and 1960. Courrier des statistiques, (English series 11), 39, 2005b:

3 DESROSIÈRES A.: The Politics of Large Numbers - A History of Statistical Reasoning, Cambridge, Mass 1998

4 PORTER T.M.: Trust in Numbers: The Pursuit of Objectivity in Science and Public Life, Princeton 1995

5 RUDINOV SAETNAN A. / MORK LOMELL H. / HAMMER S.: The Mutual Construction of Statistics and Society, New York 2011

6 https://en.wikipedia.org/wiki/Boundary_object

7 http://www.ft.com/cms/s/0/3be49734-29cb-11e6-83e4-abc22d5d108c.html#axzz4FgnrQMVK

8 https://en.wikipedia.org/wiki/Post-truth_politics

9 http://www.alpbach.org/en/forum2016/programme-2016/new-enlightenment-an-introduction-by-the-presidents-of-the-european-forum-alpbach/

Smart Contracts and Data Science

Dandolo Flumini

ZHAW Zurich University of Applied Sciences

Abstract. Equipping smart contracts with data driven functionalities (e.g. data analytics, visualization) opens up a variety of new possibilities. The execution of smart contracts can be based on their own history or from other data sources. It is further possible to evaluate and compare running contracts based on visual representations of their state and or historic data. We shall discuss examples of how smart contracts can benefit from said possibilities and what the difficulties are in implementing these features, also in the context of specialized (e.g. verified, not Turing complete) smart contract languages.

Automatization and Standardization of e-Research Processes by Knowledge-Based Research Process Configurations

Andreas Geyer-Schulz

Karlsruhe Institute of Technology

Abstract. Today's most profitable companies (e.g. Apple, Google) are scientific companies. A scientific company's core business processes ARE research processes. But, as a matter of fact, most companies do not fall into this category. And, what is even worse, most scientists do not consider their own research processes as subject to business process optimization.

However, scientific research processes offer a huge improvement potential when viewed from an automation and standardization perspective: Less qualified workers in companies can act as scientists provided:

- the research tasks are well-defined and coordinated,
- sufficiently small,
- organized as routine processes,
- supported by an IT-service (e-research) infrastructure,
- and that they have been carefully introduced to the necessary theoretical background.

When one views a standardized research process as a complex variant configuration problem, the development of knowledge-based e-research configurators holds promise for a revolution in research processes.

An example of survey processes using causal models operationalized by various types of scales shows the potential of the reuse of standard models (e.g. TAM and its variants), the automatic comparison with competing cognitive models, questionnaire generation and deployment, code plan documentation, data set generation, and generation of the standard statistical model code as well as meta-analysis and visualization.

Trading Big Data: Regaining the Trust of Individuals?

The Potentials of Establishing a Futures Market in Data from a Consumer Point of View?

Audrey Guinchard

University of Essex

Abstract. As networked digital technologies become more pervasive and diverse, it is possible to collect personal data generated across the Internet and to gather information on the different facets or personas of a person's patterns of consumption and lifestyle behaviours. Currently, this digital footprint of personhood data is fragmented, residing in various repositories across the Internet, often held by big companies that 'freely' mine data to generate advertisement revenues. A shortcoming of this advertisement-centric business model is that the individual whose personal data is being monetized by third-parties has little control over the collection process and the subsequent exploitation of his/her digital-personhood-data and that has direct consequence on his privacy. Furthermore, technology intermediaries capture only a small subset of the individual's digital-personas, which are not necessarily accurate.

An alternative to the advertising-centric model is for consumers to sell their analysed data (personas) directly to companies that want to buy it through an exchange. The aim of the 'micro-Persona eXchange' ("MPX") is to be a real alternative to the advertising-centric model by giving prosumers both actual control (as opposed to nominal control) over their data and real economic benefit of the data they generate. Even though this alternative will not resolve all problems related to the existing commodification of data, it has the potential to empower data subjects by exploiting the economic value of their digital self in a new tradable market in personhoods, while at the same time ensuring an adequate flow of data and safeguarding individual privacy.

The MPX is a project funded by Research Council EPSRC Digital Prosumer -- Establishing a 'Futures Market' for Digital Personhood Data (April 2013 - March 2016) – website: http://digitalprosumer.co.uk

European Data Science Conference EDSC

Georg Herde

Deggendorf Institute of Technology

Abstract. The "Deggendorf Forum of Digital Data Analysis" (DFDDA) founded in 2005 at the Deggendorfer Institute of Technology (DIT) – Technische Hochschule Deggendorf, was and is still concerned about different aspects of digital data analysis.

Concerning data science, we placed our questions and interests to the requirements of auditors/controllers (internal and external) which have to audit and analyze company processes. The topic "Structure of Data Science" rises many questions, for which we try to find answers and solutions at our conferences, proceedings and educational programs.

- Aims and Objectives: In our world we are faced with an increasing amount of data, even economic data. These data are the basis of automated business transactions and also the starting point of business data analysis. The point here is to bring together the increasing requirements on transparency of the business with the growing complexity of the data.
- Technology: How to get the data for an audit? We discuss questions like how to access and extract mass data of company systems without influencing the day to day business of the organization: How to make sure to get the right data, in the right quantity to the fully extend for the relevant period the auditor need for their analysis? Which tools of data analysis are appropriate for auditors/controllers? Comparing among others ACL, IDEA, SQL-Server, Excel we check which of those tools fulfill the needs of those professionals best in respect of transparency (can they proof their findings), performance (analyzing mass data under time restriction of an audit) and convenience for the user (do they have to be IT-Professionals to use these tools).
- People: How do we have to qualify the next generation of auditors/controllers in respect of IT, mathematical and statistical knowledge? This bears the consequence to find ways to make (mass-) data for trainings purposes available to students, so that they can gather their own experiences of data analysis by using real business data. Those data should be realistic data with a business context and not artificial generate pseudo data.

The topic, "What makes Data Science different" concerning data analysis is, to a certain extent, a consequence of the first topic, already. Following the above questions and statements we understand that data science is or have to become an interdisciplinary approach. There have to be involved computer scientist, mathematicians, statistician but also economist and lawyers who know the business processes and have legal issues, like personal privacy, in mind.

As long as we talk about structured data in data analysis project, the complexity of data will grow exponentially the more tables and attribute (in an SAP-System we find single tables with more than 1.000 different attributes) are involved. That implies we have to find new methods to classify, analyze, visualize and interpret mass data for audit purposes. E.g. from our respect it is not enough to state: "The probability that fraud happens in a specific company stands by 7 %", we want to know "Where and how often fraudulent activities can be observed, and are we able to detect them while transactions are entered in the systems".

Structure of Data Science

Fionn Murtagh

University of Derby / Goldsmiths University of London

Abstract. Given the successes of data science in many domains and in many sectors, there are growing numbers of MSc in Data Science degree courses, and increasing planning for, and initiating, BSc in Data Science degree courses.

Firstly, it would be very interesting to establish a list of all of these courses, proceeding then to summarize publicly and openly available attributes of these courses. Such course attributes include, foremostly, the modules, any public information on accreditation (certification, chartered status, etc. in professional bodies), internships and placements, and unique attributes of the courses.

Secondly, the sharing of course content materials would be valuable for those involved in the provisioning, teaching, and training of these courses. The sharing of such course content materials can also help with the standards and quality of course materials. By course content materials is here intended: source files of all presentation materials, source data, full specifications of all coursework and other evaluation, and clear listing of links, in priority order, to externally related materials, such as books, articles, other published literature, video and other online materials, environments and frameworks supporting interactive engagement by learners, distributed learning environment details, including MOOC access points.

I propose that sharing of course materials be a club of contributors to this initiative. That is, all contents are accessible to those who are registered. Registration is obtained through making one or more contributions of course content materials.

Data Science and Operational Research

Abdel Salhi

University of Essex

Abstract. Although OR has a number of branches with a high theoretical content, it is primarily a practical discipline which strives to solve real world problems using mathematical and computational tools. As such, it relies heavily on data. For instance, unless you are given the instance of some problem or model, it is often not possible to solve it. Data, therefore, is central to the practice of OR. Here, however, I am interested in highlighting the importance of the OR approach, and in particular the optimisation aspect of it, when dealing with the so called Data Science issues and questions such as optimum sampling, data summarisation, self-affine time-series partitioning, dimensionality reduction etc… These problems can no doubt be formulated as optimisation problems and solved using optimisation and other OR techniques. To give a practical example, I refer to the Logistics industry, which is a heavy user of the OR methodology. For instance, solving operations problems in container ports, drawing distribution routes on land, scheduling deliveries on sea, all rely on the OR approach. A lot of data is generated by this industry too and Data Science has an important role to play if efficiencies are to gained. But, it is almost impossible to imagine Data Science solving the supply chain problems without reliance on the OR approach which is the main stream management technology in this global industry.

Tuesday - 08 November 2016

Data Science for the Masses: Mission Impossible?

Michael Berthold

University of Konstanz

Abstract. The vision of "Data Science for the Masses" is simple: allow non-data scientists to make use of the power of data science without really understanding the machinery under the hood. While this is possible for narrow disciplines in which a nice GUI can hide complexity, it is not so simple for serious data science. First of all, it is already hard enough to make use of the wisdom of fellow data scientists who use their own favorite tools. Secondly, and still very much an open problem, is the issue of injecting feedback from casual users. Using a well known analytics platform, I will demonstrate methods to encapsulate, reuse, and deploy analytical procedures at various levels of abstraction - giving data scientists control to expose select parts of the analytics processes to user unfamiliar with the underlying tool or without the in-depth knowledge of the analytical algorithms used.

From Business Intelligence to Business Analytics

Manel Brichni

Luxembourg Institute of Science and Technology

I. Introduction

The analytics revolution is continuously evolving. In this context, one promising breakthrough is the application of Data Science (DS) analytics in order to support decision making (Wang, Kung, & Byr, 2016). DS is about leveraging value from an immense volume, variety and velocity of data (Larsona & Chang, 2016). Aware of the increasing breadth of opportunities of such a complex domain, in this position paper, we are interested in discussing how analytics have evolved with DS as well as its capabilities and requirements.

II. Analytics Capabilities Evolution

Data science is an interdisciplinary field that allows discovering hidden insights from massive amounts of structured and unstructured data, using methods such as statistics, machine learning, data mining, and predictive analytics in order to support decision making and gain competitive advantage (IBM Analytics, 2016).

In the literature DS is linked to other related concepts (Ayankoya, Calitz, & Greyling, 2014), for example, Business Intelligence (BI) and Business Analytics (BA). This is the reason why in this section, we discuss how analytics capabilities have evolved with DS and how it is related to such concepts.

The idea to provide in depth insights and to make decisions based on analytics has emerged since the 1950s. Since then, several technologies and concepts are continuously evolving. Business Intelligence (BI) is one of the first analytics solutions, generally used for reporting processes. It was a real progress in reaching an objective, deep understanding and giving its users the fact-based comprehension when making decisions (Davenport, 2013). A data warehouse is the core technical solution to design BI platforms (Brichni, Dupuy-Chessa, Gzara, Mandran , & Jeannet, 2015). It is the repository of large volume of data from various transactional information systems for analytical purpose. However, several limits have been identified with data warehousing solutions, mainly due to the evolution of business requirements and to the way data is generated nowadays. Actually, while 80% of the world's data is unstructured (Schneider, 2016), BI capabilities are limited to transactional data, limiting therefore the ability to consider non- traditional data sources. In addition, the high volume digital flood of data that is being generated at ever- higher velocities and varieties in current business domains adds complexity to the equation (Wang, Kung, & Byr, 2016).

As a result, volume, variety and velocity were considered as the main boundaries of BI solutions and are the reasons why new trends such as Data Science (DS) have emerged. The emergence of DS has evolved the analytics capabilities, while allowing organizations focus on predictive and prescriptive analysis versus descriptive ones with BI. In fact, we believe that DS came to expand the scope of internal transactional systems and analytics purpose considered in BI. This evolution is itself due to the evolution of business requirements and needs, which leads to a new analytics concept called Business Analytics (BA) as a part of the DS field.

Belonging to the DS field, BA provides the set of technologies, techniques and methods able to achieve and address the DS challenges and objectives. In order to understand its scope, in the following, we provide a description of DS analytics capabilities and requirements.

III. Data Science Analytics Capabilities and Requirements

Three categories of DS analytics capabilities can be defined (Acito & Khatri, 2014) (Wang, Kung, & Byr, 2016). First, analytical capability refers to the set of techniques and methods that allow process an immense volume, variety and velocity of data, in order to support descriptive, prescriptive and predictive analysis. We notice, for example, regression analysis, machine learning and simulation. Second, decision making support capability refers to the set of deliverables able to help users making decisions, for example, historical reports, queries, dashboards, etc. Third, data and information management capability refers to the ability to integrate, organize, store, track and share data assets in real time. To this end, NoSQL and the Hadoop distributed file system are ones of the most used technologies.

Moreover, the challenge in DS is the ability to align analytical tasks with management strategies to take advantage of its capabilities (Davenport, 2013). On the one hand, DS requires considering various types and volumes of data sets and data sources, with embedded and automated data exploration and analytics platforms into operational and business processes. On the other hand, new strategies and management approaches are required to promote DS capabilities. For example, new way of managing and collaboration with a variety of disciplinary data teams ensures clearer requirements, an understanding of data, joint accountability, and produce higher quality results (Larsona & Chang, 2016). To summarize, DS analytics capabilities and requirements make it a challenging field that can be more explored and applied to various fields, particularly, by means of BA solutions.

IV. Conclusion

Data Science is being the business revolution that has been applied to several fields, such as supply chain, healthcare or banking. While opportunities for employing DS continue to grow, a further investigation is required to take advantage of its analytics capabilities provided particularly by BA solutions.

References

ACITO, F., & KHATRI, V. (2014). Business analytics: Why now and what next? Business Horizons, 57(5), 565 - 570.

AYANKOYA, K., CALITZ, A., & GREYLING, J. (2014). Intrinsic relations between data science, big data, business analytics and datafication. Proceedings of the Southern African Institute for Computer Scientist and Information Technologists Annual Conference, 192-198.

BRICHNI, M., DUPUY-CHESSA, S., GZARA, L., MANDRAN, N., & JEANNET, C. (2015). Business Intelligence for Business Intelligence: A. IEEE Eleventh International Conference on Research Challenges in Information Science, 239-249.

DAVENPORT, T. (2013). Analytics 3.0. Harvard Business Review, 1-12.

IBM ANALYTICS. (2016). IBM Insight at World of Watson 2016. Mandalay Bay, Las Vegas: IBM Analytics. Retrieved from http://www.ibm.com/analytics/us/en/technology/data-science/

LARSONA, D., & CHANG, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. International Journal of Information Management, 36, 700-710.

PROVOST , F. (2013). Data Science for Business. O'Reilly Media.

SCHNEIDER, C. (2016). IBM Watson. https://www.ibm.com/blogs/watson/2016/05/biggest-data- challenges-might-not-even-know/.

WANG, Y., KUNG, L., & BYR, T. (2016). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 1-11.

Patient Participation, Profit, and Privacy Protection

Ronald Cornet

Academic Medical Centre, Amsterdam / Linköping University

Abstract. With great powers come great responsibilities. Therefore, there is a need for Responsible Data Science [http://www.responsibledatascience.org/]. Responsible Data Science (RDS) is a joint collaboration of expert researchers from 11 knowledge institutions across the Netherlands.

It has its focus on aspects fairness, accuracy, confidentiality, and transparency. The can be described by the following questions:

- Fairness: How to avoid unfair conclusions even if they are true?
- Accuracy: How to answer questions with a guaranteed level of accuracy?
- Confidentiality: How to answer questions without revealing secrets?
- Transparency: How to clarify answers such that they become indisputable?

In healthcare, responsible data science is crucial, but not the enough. We need to be able to feed information back to individual patients and their careers, which may be challenging because of the privacy protection. For example, if contra-indications or side effects are detected, applicable patients should be informed about any potential risks. However, with anonymous data it is likely impossible that these patients can be traced.

As increasingly the (chronic) patient becomes the expert of his/her own disease, patient participation is required. Patients can contribute by continuously providing health data via sensors, apps, or questionnaires, and patients can determine in which studies or types of research they want to be involved, which data they want to share, etc. Involving them in the results of research based on their data will incentivize them to contribute, and gives them the opportunity to maximally profit from these results and any new knowledge emerging from data science.

The 'Bachelorisation' of Data Science: What's the Role of EuADS for Standard Curricula?

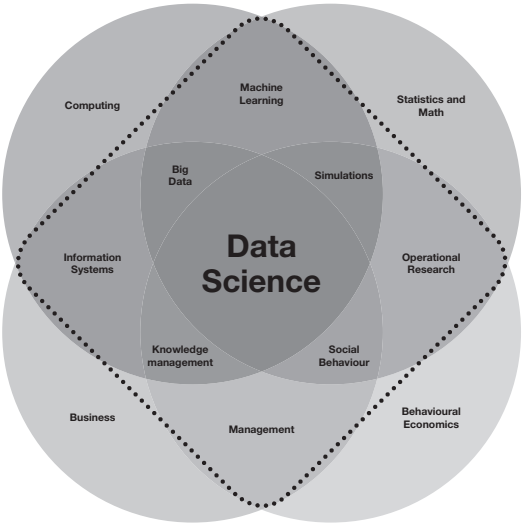
José Hernández-Orallo

Technical University of Valencia

Abstract. It's already a cliché to say that data science is bound to be one of the new professions of the early XXIst century. However, in order to reach the status of other well-established professions, data science must be associated with well-delineated curricula at the bachelor level, in such a way that companies, public institutions and society in general recognise a "data scientist" in a distinctive way. In the past two decades, and more strongly in the last five years –boosted by the popularity of the term "big data"–, the number of master degrees in data science and related areas has grown exponentially. It is only in the past two years when we are witnessing a consolidation of this offer at the level of bachelors, after realising that the reconversion of a different profession into a data scientist through a master degree is inefficient, and that young data scientists can be trained from the very first year of university.

Are 3-year or 4-year bachelor degrees the right setting to configure the profession? In what way should the profile of a bachelor degree be different from a profile of a master degree? What is the role of associations such as EuADS in shaping standard curricula as other associations have done in other disciplines (e.g., ACM for the computer science curricula)?

For instance, the following figure reflects the main areas data science is usually based on.



Data science goes beyond a mixture of statistics/maths with computer science, moving from the top of the figure to a more central, but complex location. But precisely because of this, it requires a better delineation to prevent it from being a fuzzy term that could vanish with time and never catch on as a real profession (such as data mining, KDD or other related areas, which consolidated as research disciplines but neither as professions nor identified as bachelor degrees in general). Also, the key characteristics of a data scientist go beyond the disciplines in the previous figure, covering some special soft skills, such as being proactive, curious and inquisitive, being able to tell a story about the data and visualise the insights appropriately, use the scientific method and focus on traceability and trust. Some of these skills are perhaps better achieved from early years with a project-oriented methodology. Similarly, we have to decide whether a bachelor in data science should prepare the students to lead data science projects, or this should be relegated for master degrees.

I will set out the discussion over these questions and the perspective of how the offer in master and bachelor degrees has evolved in the past two years. For instance, in 2014 there were only two bachelor degrees in data science in Europe; now there are dozens and many more are being set on the table. I will link this to the recent experience of my university, which, after setting up several master degrees in big data, data analytics and associated areas in the past, is now preparing a new 4-year bachelor in data science. I can present its basic structure at the conference if so requested, and also summarise the discussions we had about the main structure of the degree, the share of ECTS credits for different areas, and why and how they are justified could have been benefited by the existence of standard curricula and an accreditation institution.

Data Science – Is the Impact of Statistics Significant?

Katja Ickstadt and Claus Weihs

TU Dortmund University

Abstract. *Premise:* Statistics is the discipline to provide tools and methods to find structure in and to give deeper insight into data!

Data Science as a scientific discipline is influenced by computer science, mathematics, operations research, and statistics as well as the applied sciences. Statistical methods are crucial in all basic steps of data science:

- *Data achievement and enrichment:* Design of experiments should be utilized for systematically reducing data bases (Schiffner & Weihs, 2009) and for tuning algorithms, as well as for imputation of missing data to fill gaps in data bases.
- *Data Exploration:* Exploratory statistics is basic for preprocessing to learn about the content of a data base.
- *Data analysis:* Statistical methods are fundamental to find structure in data and to make predictions: These methods comprise
  - Multiple testing to correct for multiple usage of data,
  - Classification methods to find and predict groups of data (Weihs, 2016),
  - Regression methods to find global and local relationships between features,
  - Network analysis to learn network structure (Wieczorek et al., 2015),
  - Stochastic differential and difference equations to handle models from the applied sciences (Weihs et al., 2010),
  - Time series analysis to understand and predict temporal structure.
- *Integrative Data Analysis:* Local models (Bischl et al., 2013, Klein et al. 2014) can be combined to global models (Weihs&Ligges, 2005) by hierarchical and chain-like modelling.
- *Differential Data Analysis:* Tests for comparing models (Vatolkin & Weihs , 2017), perturbation experiments, and analyzing structural breaks.
- *Model validation and model selection:* Resampling methods (Bischl et al., 2012) and Meta-analysis as well as model averaging and the analysis of concept drifts (Mejri et al., 2016) are important tools.
- *Deployment:* Visualization to interpret found structures, storing of models in an easy-to-update form.

Hypothesis: Yet, the role of statistics in data science is underrepresented as, e.g., compared to computer science, in particular in the areas of data achievement and enrichment as well as in advanced modelling.

Questions: What is on the data scientist`s wish list for the statistician? What would data scientists suggest to strengthen the role of statistics in the field?

References

BISCHL, B., MERSMANN, O., TRAUTMANN, M., AND WEIHS, C. (2012): Resampling Methods in Model Validation. Evolutionary Computation Journal 20 (2), 249-275

BISCHL, B., SCHIFFNER, J., AND WEIHS, C. (2013). Benchmarking local classification methods. Computational Statistics, 28(6):2599–2619

KLEIN, H.-U., SCHÄFER, M., PORSE, B.T., HASEMANN, M.S., ICKSTADT, K, AND DUGAS, M. (2014): Integrative analysis of histone ChIP-seq and transcription data using Bayesian mixture models. Bioinformatics 30 (8), 1154-1162

MEJRI, D., LIMAM, M., AND WEIHS, C. (2016): A new Dynamic Weighted Majority Control Chart for data streams; Journal of Soft Computing; Online first: DOI: 10.1007/s00500-016-2351-3

SCHIFFNER, J., AND WEIHS, C. (2009): D-optimal Plans for Variable Selection in Data Bases; SFB 475, Technical Report 14/09, https://www.statistik.tu-dortmund.de/fileadmin/user_upload/Lehrstuehle/MSind/SFB_475/2009/tr14-09.pdf

VATOLKIN, I., AND WEIHS, C. (2017): Evaluation; Chapter 13 in Weihs, C., Jannach, D., Vatolkin, I., Rudolph, G. (Eds.): „Music Data Analysis – Foundations and Applications", Chapman&Hall/CRC, 329–363

WEIHS, C., AND LIGGES, U. (2005), From local to global analysis of music time series. In: K. Morik, A. Siebes, J.-F. Boulicault (Eds): Detecting Local Patterns, Springer Lecture Notes in Artificial Intelligence 3539, Springer, Heidelberg, 233-245

WEIHS, C., MESSAOUD, A., AND RAABE, N. (2010): Control Charts Based on Models Derived from Differential Equations; Quality and Reliability Engineering International 26, 807-816

WEIHS, C. (2016): Big Data Classification - Aspects on Many Features; In: Michaelis, S. , Piatkowski, N. , Stolpe, M. (Eds.): Solving Large Scale Learning Tasks. Challenges and Algorithms; Springer Lecture Notes in Artificial Intelligence, 9580, 139-147

WIECZOREK, J., SHERIFF, R.S.M., FERMIN, Y., GRECCO, H.E., ZAMIR, E., AND ICKSTADT, K. (2015): Uncovering distinct protein-network topologies in heterogeneous cell populations. BMC Systems Biology, 9, 24Texte

Visual Analytics to Support Research and Applications for a Sustainable Future

Benoît Otjacques

Luxembourg Institute of Science and Technology

Abstract. The data deluge heavily advertised some years ago has become a fact notably in the environmental sciences and in the green business. Vast amounts of environmental data are collected by satellites, in-situ sensors, field observations or citizen scientists. Large datasets are also produced by complex simulation runs on clusters of servers or High-Performance Computers. These datasets may take various forms: huge tables of experimental results mixing numerical or categorical variables, dynamic biological networks, large text corpora of scientific papers, or geo-located images and videos.

Scientists and engineers use these types of datasets to solve scientific or technological problems. However, several paths exist to go from raw data to the answer of a question. The visual analytics paradigm is an emerging promising approach to deal with these issues. Basically, it can be defined as the combined use of data analytics methods, data visualisation techniques and advanced interaction in order to gain insight into complex data.

The visual analytics approach is being used in the e-Science Unit of LIST (Luxembourg Institute of Science and Technology) in various domains, in particular to support progress towards a more sustainable future. For instance, advanced visual analytics software tools have been designed to better understand the process of biogas production, to better calibrate climate models, to deal with massive amounts of data in plants biology, to analyse indoor air quality or to analyse some processes in the industry.

Dr Ir Benoît Otjacques is leading the e-Science Unit of LIST. He has 20 years of experience in (applied) research in computer science and especially regarding data-related issues. The e-Science Unit of LIST regroups around 25 experts in data sciences covering data analytics as well as data visualization topics. In the recent years the applications in the environmental domain have significantly grown and represent today a strategic part of the e-Science portfolio of activities.

Can We Automate Data Science?

Luc de Raedt

KU Leuven

Abstract. AI has been successful in automating scientific reasoning processes in e.g. the life science (with the Robot Scientists). The question that I want to ask is whether it is possible to automate the processes involved in data science? I also want to answer that question in the course of our ERC AdG project SYNTH on "Synthesising Inductive Data Models".

To start the discussion on this topic, it is useful to look at the famous knowledge discovery cycle, where one typically starts from raw data, select and pre-process the data, identify the data mining task, use the right data mining algorithms, and then interpret the results and possibly iterate. It turns out that most of the existing approaches to automating this process, such as the automated statistician and meta-learning, algorithm portfolio and configuration approaches assume the learning task is known and we only need to identify the right algorithm and parameters to find the optimal task. It is well-known in the data mining community that this step takes typically only about 20% of the time, while the preprocessing and task identification take 80% of the time.

The question that I am interested in is what we can do to automate the pre-processing and task identification aspects, particularly for non-experts in data science.

For more details, see http://synth.cs.kuleuven.be/ Synthesising Inductive Data Models (ERC AdG Luc De Raedt).

Training Data Scientists: A Few Challenges

Gilbert Saporta

Conservatoire National des Arts et Métiers, Paris

Abstract. Linked with the Big Data phenomenon, the need of thousands of data scientists in the next years has qualitative and quantitative impacts on the educational system.

Academic curricula should include much more information technologies (parallel and distributed computing). Learning in Big commercial software (SAS, SPSS, etc.) becomes less important compared to free environments (R, Python, ScikitLearn, Spark etc.). A data scientist should have three main skills: statistics, computer science, communication.

Other curricula are concerned: e.g. economics (see Hal Varian, Big Data: New tricks for econometrics, Journal of Economic Perspectives, 2014), and masters in official statistics: the syllabus of the European Master in Official Statistics (EMOS) needs to be updated.

However initial training by universities will not be enough to provide quickly enough specialists. A large part of the solution has to be found in continuous education (or long life training) of statisticians and computer scientists already employed. At CNAM we have developed a professional certificate in Big Data analytics, for statisticians and engineers. Online and distance education (e.g. MOOC) should be developed and there is a clear need for cooperation and mutualisation of efforts, at least at the european level. Learned societies and federations (FenSTATS, EuADS) could be stakeholders in promoting and labelling european courses.

Three Controversies about Data Science for Medicine and Healthcare

Niels Peek[1] and Pedro Pereira Rodrigues[2],

[1]University of Manchester and [2]University of Porto

Abstract. During this plenary talk the presenters will discuss three controversies around data science in medicine and healthcare. The first controversy is Johan van der Lei's 1st Law of Medical Informatics, "Data shall be used only for the purpose for which they were collected". Although this law was formulated in the early 1990s, it is still subject to fierce debate. The second controversy concerns the question to which extend innovations in analytical methods have alleviated the need to conduct expensive randomised clinical trials. While data science enthusiasts claim that Big Data and Machine Learning can be used to answer all research questions, traditionalists claim that there is no replacement for randomised experiments when we are interested in causation. The third controversy, finally, addresses access to data – time and again a highly divise issue. Many researchers have the opinion that all medical and healthcare data should made freely available to them without any restrictions, in order to accelerate research and improve medical knowledge. But data custodians fear privacy breaches and loss of public trust: in recent years they have become more restrictive rather than more lenient when it comes to sharing health data. Before and after discussing each controversy the audience will be asked to express their opinion through a vote.

Computational Sciences: Data-Driven Modelling and Simulation

Stéphane P. A. Bordas[1] and Christophe Ley[2]

[1]University of Luxembourg and [2]Ghent University

Abstract. Science is defined as the activity concerned with the systematic acquisition of knowledge and is an enterprise that builds and organises knowledge in the form of testable explanations and predictions about the universe. Engineering is the application of scientific and practical knowledge for the benefits of mankind. For example, Theodore von Kármán, a leading mathematician, aerospace engineer and physicist developed theories for aerodynamics, in particular supersonic and hypersonic airflow characterisation, which have been essential to the design and fabrication of modern jet engines and rockets.

To produce new knowledge and apply this knowledge to practical fields, scientists and engineers use the "scientific method" which tests statements that are logical consequences of scientific hypotheses (theories or computer models and simulations) through repeatable experiments and observations. This production of knowledge has been fuelled by a significant revolution which has taken place over the last 50 years, through which a new, inherently multi-disciplinary pillar of science has emerged to complement these theories and observations: Computational Sciences. Computational Sciences is the tri-disciplinary endeavour concerned with the use of computational methods and devices to enable scientific discovery and engineering applications in science. But we speak today about a new revolution that could be even more impactful and already permeates our lives. This is the Data revolution; we are living in the "Big Data era".

In this new era, the wealth of Data has transformed the world of scientific discovery and engineering innovation. The combination of computational sciences with data science will lie at the core of future scientific and engineering research. A new ability will play a central role, namely that of extracting knowledge from this wealth of information by storing, compressing, classifying, ordering and analyzing Data.

In particular, we will see the emergence of smart systems, able to adapt to their environment through advanced data gathering and treatment approaches. These developments will be multi-disciplinary with mathematics, in particular statistics and numerical analysis as well as computer science at its core.

Illustrations of this paradigm shift will range from aerospace engineering to life sciences. For example, predicting the three-dimensional structure of a protein from its amino acid sequence is a holy grail problem in bioinformatics with, in case of success, far-reaching impacts on drug design, biotechnology and optimisation. Massive amounts of protein data are available nowadays, yet the problem remains extremely challenging. Major recent advances in this field have been achieved through a crafty combination of directional statistics, Bayesian statistics, advanced modeling, graphical computational methods, and machine learning.

Our specific contribution to the conference will be a discussion towards suitable approaches for machine learning and Bayesian statistics based multi-scale model selection. From engineering to medicine, such approaches could fuel a "digital twin" concept enabling models to learn from real-time data acquired during the life of the system, accounting for the actual environmental conditions within predictions. As such, the "digital twin", continuously fed sensor data from the actual system, would live its own digital life which would enable scientists to investigate various scenarii about its future life and make predictions on its behaviour.

The Problem of Finding Interesting Patterns

Andreas Geyer-Schulz

Karlsruhe Institute of Technology

Abstract. A standard argument of many evangelists of Big Data applications (based on massive data collection) is that interesting patterns wait to be discovered in these data sets and their discovery will lead to new breakthrough applications. This is also a major promise of Data Science. However, with regard to the real operationalization of finding interesting patterns these visionary papers remain remarkably silent.

Percy Diaconis (1985) has described this kind of thinking as magical thinking as exhibited e.g. by the South Sea Cargo Cult or as the attempt of guessing the outcome of a sequence of random numbers.

A first attempt at finding interesting patterns is data crunching which helps to find interesting patterns in big noisy data by systematic search. In this approach interesting patterns must be defined by some search criteria. And this essentially leaves the problem unsolved.

It is the position of this author (and of Percy Diaconis) that the finding of interesting patterns must rely on scientific thinking. Data Science can only live up to its promise, if a scientific theory adequate to the problem is used for operationalizing the concept of an interesting pattern. For example, for library recommender systems, a normative model of a decision-maker without preferences models all (random) behavioral patterns which are not of interest. In the field of end-consumer Internet product configurators, interesting patterns are irrational choices of consumers which can be modeled as deviations from Von Neumann – Morgenstern axiomatic utility theory.

What makes data science different, is the systematic use of scientific theories for operationalizing the search for interesting patterns.

References

Diaconis, P. (1985). Theories of data analysis. From magical thinking through classical statistics.
In Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors, Exploring Data Tables, Trends, and Shapes,
Wiley Series in Probability and Mathematical Statistics, chapter 1, pages 1 – 36. John Wiley & Sons, New York.

What Makes Data Science Different?

Fionn Murtagh

University of Derby / Goldsmiths University of London

Abstract. From my contribution in this paper, which has been taken further in many areas, including, currently, very great social media, Twitter, analytics of (music, cinema, parade) festivals. (F. Murtagh, "Semantic mapping: Towards contextual and trend analysis of behaviours and practices", in K. Balog, L. Cappellato, N. Ferro, C. MacDonald, Eds., Working Notes of CLEF 2016 - Conference and Labs of the Evaluation Forum, Évora, Portugal, 5-8 September, 2016, pp. 1207-1225, 2016.)

N. Keiding and T.A. Louis, "Perils and potentials of self-selected entry to epidemiological studies and surveys", Journal of the Royal Statistical Society A (2016) 179, Part 2, pp. 319–376

Published response by these authors:
Murtagh muses on the geometry and topology of the relationship between the data analysed and big data. Implementing his construct would be challenging, but we thank him for his broad perspective.

Fionn Murtagh (University of Derby and Goldsmiths University of London).
Interesting perspectives that support Keiding and Louis include Friedman et al. (2015), and the following quote, from Laurison and Friedman (2015): '... the GBCS [Great British Class Survey] data have three important limitations. First, the GBCS was a self-selecting web-based survey, ... This means it is not possible to make formal inferences. ... the nationally representative nature of the Labour Force Survey (LFS) along with its detailed and accurate measures ... facilitates a much more in-depth investigation. ...' In a blog posting, Laurison (2015) pointed very clearly to how, just 'Because the GBCS is not a random sample or representative survey', other ways can and are being found to draw great benefit (http://www.thesociologicalreview.com/information/blog/three-myths-and-facts-about-the-great-british-class-survey.html). Another different study on open, free-text questionnaires (Züll and Scholz (2011); see also Züll and Scholz (2015)) notes selection bias, but also 'However, the reasonable use of data always depends on the focus of analyses. So, if the bias is taken into account, then group-specific analyses of open-ended questions data seem appropriate.'

The bridge between the data that are analysed and the calibrating 'big data' is well addressed by the geometry and topology of data. Those form the link between sampled data and the greater cosmos. Eminent quantitative and qualitative sociologist Pierre Bourdieu's concept of field is a prime exemplar. Consider, as noted by Lebaron (2009), how Bourdieu's work involves 'putting his thinking in mathematical terms', and that it 'led him to a conscious and systematic move toward a geometric frame-model'. This is a multi-dimensional, 'structural vision'. Bourdieu's analytics 'amounted to the global [hence big data] effects of a complex structure of interrelationships, which is not reducible to the combination of the multiple [... effects] of independent variables'. The concept of field, here, uses geometric data analysis that is core to the integrated data and methodology approach used in the correspondence analysis platform (Murtagh, 2010). An approach to drawing benefit from big data is precisely as described by Keiding and Louis. The noting of the need for the 'formulation of abstract laws' that bridge sampled data and calibrating big data can be addressed, for the data analyst and for the application specialist, as geometric and topological.

Data Science, a Position Paper for EDSC

Arno Siebes

Utrecht University

(Note that in the interest of brevity I restrict myself to academia, in a setting like industry the arguments would be largely similar. Furthermore, I (mis)use the word science to stand for all academic areas of endeavour).

1. Data Science

   The most important observation one can make regarding Data Science is that its advent signals the end of the era in which Computer Science was under the purview of computer scientists. Due to the digitization – or perhaps more apt, *datafication* – of all academic endeavours, computer science is quickly becoming an integral part of all sciences. Storing, manipulating and analysing vast amounts of data of a bewildering variety of types is becoming the core of many new approaches to science, any kind of science.

   Bioinformatics is probably the first, and arguably the biggest, example of the datafication of a science, but other sciences are following quickly. From the Social Sciences (e.g., social network analysis), to the Humanities (digital humanities) to Education (Learning Analytics) to Astronomy (sky surveys and massive simulations). All of these areas have their own problems, invent their own solutions and have their own view on what the important problems are, what marks a good solution, and have their own publishing standards and culture. Some of these problems, and their solutions, can be classified as "computer science" while others – such as ethical issues – can not.

   In this short position paper, I discus what this observation means both for Data Science and for Computer Science

2. Implications for Data Science

   The most important implication is that Data Science is *not* Computer Science. The close second is that Data Science is not an interdisciplinary – or multidisciplinary – affair. To start with the former, a historian is out to answer historical questions not to answer computer science questions. She may need to solve CS problems to answer her historical questions, but the interest is on the latter. That is, both the problems and their proposed solutions are evaluated from the point of view of a historian.

   This doesn't preclude that an area like history-informatics – that is solely concerned with developing solutions for historians – may arise in the non to distant future. It does preclude that that area will be just another computer science area; a journal like Bioinformatics is really different from a regular computer science journal. Learning Analytics tools that do not align with current pedagogical theories will not be seen as valid.

   This may be seen as an argument to see Data Science as an interdisciplinary field. But it isn't. True, each separate field – like Bioinformatics or History Informatics – may be seen as an interdisciplinary feld, albeit with its own culture and standards; a computer scientist willing to learn enough of the host science to understand the problems and what makes a valid solution can probably make important contributions. But

solutions that are valid in one corner of Data Science – say, Bioinformatics – is not necessary a valid solution in another corner, say History Informatics. This is also the reason why Data Science is not multidisciplinary, biologists with data problems do not have necessarily much in common with historians with data problems.

It may be nice to know what it isn't, but it is more useful to know what it is. So, *what is Data Science?*

That is easy:

> Data Science is a language

A computer science kind of language, its vocabulary contains words like *algorithms, data structures, queries, …* However, the sentences uttered in that language are *not* necessarily computer science.

Rather,

> Data Science is a language in which scientists in disciplines other than Computer Science, can talk about problems and their solutions in their discipline that involve the storage, manipulation, or analysis of (large amounts of) data.

There will be many dialects of Data Science, dialects that allow a scientist to formulate problems in her own science precisely. But all these dialects will contain a common core, viz., algorithmic thinking. A core that allows that scientists to talk about the computational problems and solutions involved in solving her science problem.

3. Implications for Computer Science

Losing the purview of their own field my seem a dystopic future to many of my fellow computer scientists. However, it isn't. Rather it is an illustration of the pervasiveness of our way of thinking. Due to the datafication of all sciences – in fact, of (almost) all aspects of life – means that storing, manipulating, and analysing (vast amounts of) data is quickly becoming an integral part of each and everyone of the sciences.

However hard we would try, it is impossible to train enough computer scientists to satisfy the computational needs of our colleagues in all other departments, let alone in all nooks and crannies of modern society. Rather than solving their (computational) problems we should help them to solve these themselves. That is, we should teach them to speak Data Science.

This puts Computer Science in a position Mathematics has already been in for ages. Mathematics is the language of choice for topics such as Physics, Econometrics, and Theoretical Biology. Not only do these areas use mathematics as a way to state and solve problems – albeit not always in ways that mathematicians approve of – but they also actively contribute to Mathematics proper. The physicist Witten won a Fields medal

for his work on Knot Theory. Many of the well-known analysis methods in Statistics have been developed in areas as diverse as Agriculture, Beer Brewing, and the Social Sciences.

Just as Mathematics, Computer Science will not disappear because of the advent of Data Science. Rather it will both grow and shrink. Grow because many more scientists will contribute, shrink because computer scientists will focus (even) more on computer science and (even) less on its applications. Restricting myself to the area I know most of – Machine Learning/Data Mining – I expect, e.g., more research on fundamental questions – e.g., what can be learned in sublinear time – on algorithmic questions – what is the complexity of learning a classifier – and data structure questions – what is the optimal way to store text data. And less research on social network analysis, recommender systems, and pure applications.

Moreover, Data Science will offer many challenges for computer scientists. Problems, A, B, and C in different Data Science dialects seem to have a lot in common as do their solutions. Can we abstract the differences away and formulate and solve an underlying – more general – computer science problem?