Lecture Notes in Computer Science 12751

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

Marcello La Rosa · Shazia Sadiq ·
Ernest Teniente (Eds.)

Advanced Information Systems Engineering

33rd International Conference, CAiSE 2021
Melbourne, VIC, Australia, June 28 – July 2, 2021
Proceedings

Springer

Editors
Marcello La Rosa
The University of Melbourne
Melbourne, VIC, Australia

Shazia Sadiq
The University of Queensland
St Lucia, QLD, Australia

Ernest Teniente
Universitat Politècnica de Catalunya
Barcelona, Spain

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-030-79381-4 ISBN 978-3-030-79382-1 (eBook)
https://doi.org/10.1007/978-3-030-79382-1

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

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

Preface

The 33rd International Conference on Advanced Information Systems Engineering (CAiSE'21) was organized to be held in Melbourne, Australia, during June 28 – July 2, 2021. Due to the COVID-19 global pandemic, the conference was moved online and held virtually over the same time period.

The CAiSE conference continues its tradition as the premiere venue for innovative and rigorous research across the whole spectrum of information systems (IS) engineering. This year, the conference focussed on the theme of Intelligent Information Systems, acknowledging the high level of uncertainty that organizations have to deal with, and the increasing need to respond through Intelligent Information Systems that provide trusted, adaptive, agile, and autonomous solutions. In the backdrop of recent advancements in IoT, big data analytics, artificial intelligence, machine learning, and blockchain, the Information Systems Engineering research community is ideally positioned to bring together the technical and empirical aspects of Information Systems and contribute to progress in the field.

The CAiSE'21 program included three invited keynotes by Professor Michael Rosemann (Queensland Institute of Technology, Australia), Professor Felix Naumann (HPI, University of Potsdam, Germany), and Professor Sudha Ram (University of Arizona, USA). The call for papers solicited research papers in the categories of Technical, Empirical and Exploratory papers, in all areas of IS engineering, including novel approaches to IS engineering; models, methods, and techniques in IS engineering; architectures and platforms for IS engineering; and domain-specific and multi-aspect IS engineering. 172 full paper submissions were received. We followed the selection process consolidated in the previous years, where each paper was initially reviewed by at least two Program Committee (PC) members; papers with only negative evaluations were rejected; all papers with at least one positive evaluation were reviewed by a member of the Program Board (PB); all reviewers then engaged in an online discussion led by another PB member; finally, during an all-hands meeting of the PB held virtually over-two days in February 2021, the final decision was made about the acceptance or rejection of each paper. The overall evaluation process of the papers resulted in the selection of 33 high-quality papers, which amounted to an acceptance rate of 19%. The final program of CAiSE'21 was complemented by the CAiSE Forum, workshops, co-located working conferences, tutorials and panels, and a PhD consortium. For each of these events, separate proceedings were published.

We would like to thank the general chair, Professor Marcello La Rosa, the organization chair, Laura Juliff, and the whole organization team at The University of Melbourne for their support and incredible work. We thank also the Forum chairs, Axel Korthaus and Selmin Nurcan, the Workshop chairs, Artem Polyvyanyy and Stefanie Rinderle-Ma, the Tutorial/Panel chairs, Pierluigi Plebani and Arthur ter Hofstede, the Doctoral Consortium chairs, Chun Ouyang, John Krogstie and Jolita Ralyté, and the publicity chairs, Abel Armas-Cervantes, Fabrizio Maggi, Kate Revoredo Lin Liu, and

Pnina Soffer, for their extraordinary and professional work. We thank all PC and PB members, who played a fundamental role in the selection process. Finally, we would like to express our deepest gratitude to all those who served as organizers, session chairs, and hosts, and went above and beyond to ensure that CAiSE continues to provide an engaging and high value forum for scientific exchange and networking within the Information Systems engineering community, in spite of challenges posed by the online setting.

CAiSE'21 was organized with the support of the School of Computing and Information Systems at The University of Melbourne, the Melbourne Convention Bureau, and Springer.

May 2021

Shazia Sadiq
Ernest Teniente

Organization

Program Chairs

Shazia Sadiq	The University of Queensland, Australia
Ernest Teniente	Universitat Politècnica de Catalunya, Spain

General Chair

Marcello La Rosa	The University of Melbourne, Australia

Workshop Chairs

Artem Polyvyanyy	The University of Melbourne, Australia
Stefanie Rinderle-Ma	University of Vienna, Austria

Forum Chairs

Axel Korthaus	Swinburne University of Technology, Australia
Selmin Nurcan	Université Paris 1 Panthéon-Sorbonne, France

Tutorial/Panel Chairs

Pierluigi Plebani	Politecnico di Milano, Italy
Arthur ter Hofstede	Queensland University of Technology, Australia

Doctoral Consortium Chairs

Chun Ouyang	Queensland University of Technology, Australia
John Krogstie	Norwegian University of Science and Technology, Norway
Jolita Ralyté	University of Geneva, Switzerland

PhD Award Chair

Eric Dubois	LIST, Luxembourg

Publicity Chairs

Abel Armas-Cervantes (Coordinator)	The University of Melbourne, Australia
Fabrizio Maggi	Free University of Bozen-Bolzano, Italy
Kate Revoredo	Vienna University of Economics and Business, Austria

Lin Liu	Tsinghua University, China
Pnina Soffer	University of Haifa, Israel

Organization Chair

Laura Juliff	The University of Melbourne, Australia

Conference Steering Committee Chairs

Johann Eder	Alpen Adria Universität Klagenfurt, Austria
John Krogstie	Norwegian University of Science and Technology, Norway
Eric Dubois	LIST, Luxembourg

Conference Advisory Board

Janis Bubenko	KTH Stockholm, Sweden
Oscar Pastor	Universidad Politécnica de Valencia, Spain
Barbara Pernici	Politecnico di Milano, Italy
Colette Rolland	Université Paris 1 Pantheon-Sorbonne, France
Arne Solvberg	Norwegian University of Science and Technology, Norway

Program Board

Valeria De Antonellis	University of Brescia, Italy
Eric Dubois	Luxembourg Institute of Science and Technology. Luxembourg
Johann Eder	Alpen Adria Universität Klagenfurt, Austria
Xavier Franch	Universitat Politècnica de Catalunya, Spain
Matthias Jarke	RWTH Aachen University, Germany
John Krogstie	Norwegian University of Science and Technology, Norway
Massimo Mecella	Sapienza University of Rome, Italy
Jan Mendling	Wirtschaftsuniversität Wien, Austria
Selmin Nurcan	Université Paris 1 Panthéon-Sorbonne, France
Oscar Pastor Lopez	Universitat Politècnica de València, Spain
Barbara Pernici	Politecnico di Milano, Italy
Geert Poels	Ghent University, Belgium
Jolita Ralyté	University of Geneva, Switzerland
Manfred Reichert	University of Ulm, Germany
Hajo A. Reijers	Utrecht University, the Netherlands
Stefanie Rinderle-Ma	University of Vienna, Austria
Antonio Ruiz-Cortés	University of Seville, Spain
Camille Salinesi	Université de Paris 1 Panthéon-Sorbonne, France
Pnina Soffer	University of Haifa, Israel

Barbara Weber	University of St. Gallen, Switzerland
Matthias Weidlich	Humboldt-Universität zu Berlin, Germany
Jelena Zdravkovic	Stockholm University, Sweden

Program Committee

Raian Ali	Hamad Bin Khalifa University, Qatar
Joao Araujo	Universidade NOVA de Lisboa, Portugal
Marko Bajec	University of Ljubljana, Slovenia
Alistair Barros	Queensland University of Technology, Australia
Boualem Benatallah	The University of New South Wales, Australia
Alex Borgida	Rutgers University, USA
Sjaak Brinkkemper	Utrecht University, the Netherlands
Andrea Burattin	Technical University of Denmark, Denmark
Cristina Cabanillas	University of Seville, Spain
Cinzia Cappiello	Politecnico di Milano, Italy
Josep Carmona	Universitat Politècnica de Catalunya, Spain
Fabiano Dalpiaz	Utrecht University, the Netherlands
Ernesto Damiani	University of Milan, Italy
Maya Daneva	University of Twente, the Netherlands
Adela Del Río Ortega	University of Seville, Spain
Claudio Di Ciccio	Sapienza University of Rome, Italy
Oscar Diaz	University of the Basque Country, Spain
João Falcão E. Cunha	University of Porto, Portugal
Pablo Fernández	University of Seville, Spain
Agnès Front	Grenoble Alpes University, France
Giancarlo Guizzardi	Federal University of Espirito Santo, Brazil
Jennifer Horkoff	Chalmers University of Technology and the University of Gothenburg, Sweden
Jan Jürjens	Fraunhofer Institute for Software and Systems Engineering ISST and University of Koblenz-Landau, Germany
Marite Kirikova	Riga Technical University, Latvia
Agnes Koschmider	Kiel University, Germany
Sander J. J. Leemans	Queensland University of Technology, Australia
Henrik Leopold	Kühne Logistics University, Germany
Fabrizio Maria Maggi	Free University of Bozen-Bolzano, Italy
Andrea Marrella	Sapienza University of Rome, Italy
Florian Matthes	Technical University of Munich, Germany
Raimundas Matulevicius	University of Tartu, Estonia
Patrick Mikalef	Norwegian University of Science and Technology, Norway
Marco Montali	Free University of Bozen-Bolzano, Italy
Haralambos Mouratidis	University of Brighton, UK
John Mylopoulos	University of Toronto, Canada
Andreas L. Opdahl	University of Bergen, Norway

Xavier Oriol	Universitat Politècnica de Catalunya, Spain
Jeffrey Parsons	Memorial University of Newfoundland, Canada
Anna Perini	Fondazione Bruno Kessler, Italy
Pierluigi Plebani	Politecnico di Milano, Italy, Italy
Klaus Pohl	University of Duisburg-Essen, Germany
Artem Polyvyanyy	The University of Melbourne, Australia
Henderik A. Proper	Luxembourg Institute of Science and Technology, Luxembourg
Gil Regev	Ecole Polytechnique Fédérale de Lausanne, Switzerland
Iris Reinhartz-Berger	University of Haifa, Israel
Manuel Resinas	University of Seville, Spain
Marcela Ruiz	Zurich University of Applied Sciences, Switzerland
Flavia Santoro	Universidade do Estado do Rio de Janeiro, Brazil
Samira Si-Said Cherfi	Conservatoire National des Arts et Métiers, France
Kari Smolander	Lappeenranta University of Technology, Finland
Monique Snoeck	Katholieke Universiteit Leuven, Belgium
Janis Stirna	Stockholm University, Sweden
Arnon Sturm	Ben-Gurion University, Israel
Boudewijn Van Dongen	Eindhoven University of Technology, the Netherlands
Panos Vassiliadis	University of Ioannina, Greece
Ingo Weber	TU Berlin, Germany
Hans Weigand	Tilburg University, the Netherlands
Lijie Wen	Tsinghua University, China
Mathias Weske	University of Potsdam, Germany
Jian Yang	Macquarie University, Australia

Additional Reviewers

Affia, Abasi-Amefon
Ahmadian, Amir Shayan
Bondel, Gloria
Burke, Adam
Ehl, Marco
Ehrendorfer, Matthias
Elnaggar, Ahmed
Estrada Torres, Irene Bedilia
Farshidi, Siamak
Flake, Julian
Fumagalli, Mattia
Gianola, Alessandro
Haarmann, Stephan
Heindel, Tobias
Hobeck, Richard
Hyrynsalmi, Sonja
Iqbal, Mubashar
Jansen, Slinger
Kalenkova, Anna
Kostova, Blagovesta
Ladleif, Jan
Lux, Marian
Lāce, Ksenija
Mamudu, Azumah

Mangat, Amolkirat Singh
Nägele, Sascha
Padró, Lluís
Peldszus, Sven
Penicina, Ludmila
Ramadan, Qusai
Rivkin, Andrey
Scheibel, Beate
Spijkman, Tjerk
Sànchez-Ferreres, Josep
Turki, Slim
Vuolasto, Jakko
Wang, Qi

Extended Abstracts of Invited Keynote Talks

Designing Intelligent Systems: The Role of Affordances and Trust

Michael Rosemann

Queensland University of Technology, Centre for Future Enterprise, 2 George Street,
Brisbane, 4000, Qld, Australia
m.rosemann@qut.edu.au

Abstract. In a world in which the capabilities of systems grow faster than our ability to comprehend these, we need revised approaches for the design of such increasingly intelligent systems. No longer is a requirements-driven approach the only paradigm. Instead, the affordances of systems provide a rich design space that needs to be explored. Such an affordances-driven approach, however, is still in its infancy. The incomprehensibility of systems also leads to new challenges for system use. Though trust is now a key factor determining the user acceptance of systems, we are still at the beginning of a trusted-by-design discipline. Thus, we need to invest our research efforts into deriving a better understanding of the role and the integration of affordances and trust in contemporary system design.

Keywords: Intelligent systems · System design · Affordances · Trust

1 The Growing Gap

The capabilities of technology in general, and intelligent information systems in particular, are developing rapidly and often exponentially. However, our capability to comprehend this rapid change, i.e. our digital intelligence, is not developing in the same speed [1]. As a result, the gap between what intelligent systems can do and what humans comprehend is growing (Fig. 1).

This growing gap between the capabilities of intelligent systems and our digital intelligence leads to the *problem of incomprehensibly* with two significant design implications.

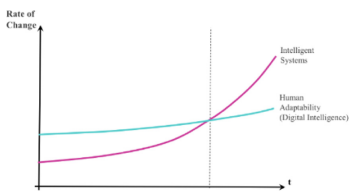

Fig. 1. The growing gap between intelligent systems and digital intelligence (inspired [1]).

First, there is the danger of under-capitalization when it comes to the design of systems. The dominating paradigm of design-follows-requirements ignores that unconscious incompetence prevents us from articulating entirely new design options. As a consequence, a shift needs to occur from a focus on specifying requirements (to the left of the dotted line in Fig. 1) to an exploration of affordances (what is possible?). Affordances-driven approaches to system design, however, are far less understood than the domain of requirements engineering.

Second, if systems have a level of intelligence that is beyond the comprehension of the system's users, trust concerns might emerge as a barrier to the acceptance of such systems. As a result, we need to go beyond a focus on ease-of-use and usefulness, and add trust-building design principles and mechanisms into our design methodologies.

2 Affordances-Driven Design

Affordances are action possibilities arising from the relation between the features of a technology and goal-oriented actors determining how the technology can be used in a value-creating way [2]. Such a definition assumes an actor capable of assessing technological' capabilities. As technology develops rapidly, however, actors will be challenged to identify and capitalise from relevant affordances meaning they remain hidden affordances [3].

One approach to overcome the incomprehensibly of technology is to systematically identify tiered layers of affordances in order to derive higher-order affordances [4, 5]. In particular, we differentiate here the three layers of technical, design and business affordances.

For example, for blockchain we identified via an empirical study of the practices of the 30 largest financial institutions globally, the immediate *technical affordances* tokenization, tracing and triggering [6]. Technical affordances are explicit as they can be perceived directly by a technology-aware user. *Design affordances* are action possibilities an organization can embed when designing with blockchain in mind. These include in the context of blockchain integrity, validity and compatibility. Design affordances are implicit and on a higher order of abstraction. Finally, *business affordances* are possibilities to create new value for customers using technical and design affordances. For example, blockchain's business affordances are micro-fulfilment, synergistic delivery and sovereignty.

Affordances-driven design requires embedding such affordances into the specification of requirements. Such requirements would be proactive requirements as they do not (reactively) emerge from an organization's demands, but are inspired by new design opportunities made available by external enablers.

3 Trusted-by-Design

The interactions of users with contemporary systems are becoming more trust-intensive. There are three reasons why trust increasingly matters for systems' acceptance.

First, the move from offline to online transactions is reducing *tangibility* (e.g., online grocery shopping) and as a result leads to new trust concerns. Second, there is limited *visibility* of the implications in those cases where a user contributes, directly or indirectly, personal data to the interactions with a system. Third, the sophistication of the intelligence embedded in contemporary systems (e.g., Amazon Go) does not only lead to new levels of convenience and experience, but also raises the issue of *explainability*. However, despite this increasing relevance of trust as a design goal, the overall trust literacy is still low.

Like affordances, trust is a relational concept. It describes the willingness of a trustor (e.g., a customer) to rely on a trustee (e.g., an organization, a business process, a system) in light of uncertainty. However, the notion of trust in organizations (e.g., ability, integrity, benevolence [7]) and trust in systems (trustworthiness) varies.

Therefore, it is suggested to decompose trust design more broadly into reducing uncertainty and increasing confidence [8]. A system that is trusted-by-design is low in uncertainty. Uncertainty an be further broken down into the elements of systemic uncsertainty, behavioral uncertainty, perceived uncertainty and vulnerability. While a design targeting uncertainty directly changes the system, confidence is about the perception of the system. Various confidence mechanisms need to be differentiated (e.g., confidence derived from peers, experts or previous experiences) and context-specifically be activated.

Therefore, trust designers, a new species, will need to develop skills to manage the uncertainty of and the confidence in a system. Advanced trust design will not only be needed to ensure that systems perform according to expectations (*core trust*), but, and especially in the context of intelligent systems, to facilitate entirely new forms of extreme trust. *Extreme trust* is the situation in which a system makes decisions on behalf of the user, and the user expects and accepts that this is the case. Examples for such systems can already be found in the domain of personalized healthcare and entertainment (e.g., music streaming), but are also emerging in areas such as banking, insurance, transportation and retail. Extremely trusted systems are a design option for intelligent systems. They are grounded in the business affordance proactivity; organizations increasingly have more data and higher algorithmic capabilities than their customers which provides them with the action possibility to ultimately make better and faster decisions than their customers.

References

1. Friedman, Th.L.: Thank You for Being Late. An Optimist's Guide to Thriving in the Age of Accelerations. Farrar, Strauss and Giroux. New York (2016)
2. Volkoff, O., Strong, D. M. Critical realism and affordances: theorizing IT-associated organizational change processes. MIS Q. **37**(3), pp. 819–834 (2013)
3. Gaver, W.W.: Technology affordances. In: Robertson, S.P., et al. (eds.) Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, New Orleans, pp. 79–84, ACM: New York (1991)

4. Bygstad, B., Munkvold, B.E.: In search of mechanisms: conducting a critical realist data analysis. In: Beath, C., et al. (eds.) Proceedings of the 32nd International Conference on Information Systems (ICIS 2011), Shanghai, 4–7 December 2011
5. Ostern, N., Rosemann, M.: A framework for digital affordances. In: Matook, S., et al. (eds.) Proceedings of the 29th European Conference on Information Systems (ECIS 2021), Marrakech, 14–16 June 2021
6. Ostern, N., Rosemann, M., Moormann, J.: Determining the idiosyncrasy of blockchain: an affordances perspective. In: Proceedings of the 41st international Conference on Information Systems (ICIS 2020). Hyderabad, 13–16 December 2020
7. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An integrative model of organizational trust. Acad. Manage. Rev. **20**(3), 707–734 (1995)
8. Rosemann M.: Trust-aware process design. In: Hildebrandt T., van Dongen B., Röglinger M., Mendling J. (eds.) Business Process Management. BPM 2019. LNCS, vol. 11675, pp. 305–321. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-26619-6_20

Leveraging Artificial Intelligence and Big Data to Address Grand Challenges

Sudha Ram

Anheuser-Busch Professor of MIS, Entrepreneurship & Innovation, Director, INSITE Center for Business Intelligence and Analytics, Eller College of Management University of Arizona, Tucson, AZ 85721
ram@eller.arizona.edu

Abstract. The phenomenal growth of social media, mobile applications, sensor-based technologies and the Internet of Things is generating a flood of "Big Data" and disrupting our world in many ways. Simultaneously, we are seeing many interesting developments in machine learning and Artificial Intelligence (AI) technologies. In this keynote I will examine the paradigm shift caused by recent our society. Using examples from health care, smart cities, education, and businesses in general, this talk will highlight challenges and research opportunities to address problems that have social implications.

Keyword: Big data · Machine learning · Artificial intelligence · Prediction models · Social good

Introduction

We live in an exciting world of the Fourth Industrial Revolution where we are witnessing a convergence of technological advances and a data deluge. These advances are merging the physical and worlds in ways developments in AI and Big Data and ways to harness their power to address grand challenges facing that create a paradigm shift and hold great promise for the future. The phenomenal growth of social media, mobile applications, sensor based and wearable devices, and the Internet of Things, is generating a flood of "Big Data" and disrupting our world in many ways. Simultaneously, we are seeing many interesting developments in machine learning and Artificial Intelligence (AI) technologies and methods. Organizations and individuals as well as societies are in a position to harness these advancements in AI and Big data analytics to identify grand challenges facing us and to solve them in new ways.

Paradigm Shift in Big Data and AI

The term "Big Data" is often understood to reflect characteristics such as volume, velocity, and variety. While these may appear to be terms that describe big data, we need to dig deeper to truly appreciate its potential. These characteristics do not do justice to explain how big data can be harnessed. I will go beyond these terms to reflect

on why big data is changing our world to create a paradigm shift. Specifically, I will focus on three specific properties of big data related to the "datafication" of the world, dissolution of the line between the physical and digital world, and the temporal and spatial characteristics of granular big data. These three characteristics are fundamental to big data and can be harnessed in multiple ways to creatively solve problem.

Simultaneously there has been a paradigm shift in machine learning and artificial intelligence (AI). The founding fathers of AI coined the term Artificial Intelligence in 1956 and predicted great optimism for the field. Two distinct approaches were proposed for AI – one mathematical using deductive reasoning or statistical using inductive reasoning and the other biological or psychological to create reasoning akin to the human brain.

One paradigm started dominating in AI soon after, with its focus on using symbolic logic where computers were taught given symbols and operators. This approach has been now supplanted by another paradigm i.e., sub symbolic approach, inspired by psychologists such as Rosenblatt. This approach proposed the idea of "perceptrons" which were inspired by the functioning of the human brain and which needed to fire neurons based on weights and thresholds. The symbolic approach was transparent and interpretable, while the sub symbolic approach is not. Development of large-scale computational power and availability of large amounts of data with the advent of the WWW have spurred the sub symbolic approach and consequent AI advancements. Today we have multilayer neural networks (deep learning methods) such as Convolutional neural nets or Recurrent neural nets. While these types of neural nets started as black boxes, we have "attention mechanisms" that can now open up these neural nets to some extent. However, these are all still supervised techniques in that they need data and examples to learn. Emerging areas now include unsupervised methods such as reinforcement learning which start with a goal and learning to progress toward that goal.

Interdisciplinary Approaches to Address Grand Challenges

Given these developments in AI and Big data, we are perfectly positioned to make contributions to solving grand challenges particularly to address problems that have social implications [1].

The Information Systems and Computer Science fields have a unique opportunity to lead by embracing a new research approach for identifying and solving interesting problems [2]. Many opportunities abound for data science based research that exploits the temporal and spatial characteristics of big data, the datafication phenomenon and the dissolving line between the physical and digital world. These research methods can also exploit the developments in deep learning methods particularly to identify and remove bias in results of predictions from machine learning. Research that is able to open up the "black boxes" of deep learning to explain the results is also very important. Finally this is an opportunity to develop interdisciplinary collaborations to solve grand challenges in areas that include health care, environmental, and social justice challenges.

References

1. Ram, S., Goes, P.: Focusing on programmatic high impact information systems research, not theory, to address grand challenges, MIS Q. **45**(1) 479-483 (2021)
2. Zhang, W., Ram, S.: A comprehensive analysis of risk factors for asthma: based on machine learning and large heterogeneous data sources. MIS Q. **44**(1), 305–349 (2020). Special Issue on the Role of Information Systems in Chronic Disease Prevention and Management

Bad Files, Bad Data, Bad Results: Data Quality and Data Preparation

Felix Naumann

Hasso Plattner Institute, University of Potsdam, Germany
felix.naumann@hpi.de

Abstract. A significant obstacle when developing and deploying data science solutions is the poor state of data: *files* will not load, schemata are outdated, *data* are ill-formatted, incorrect, or simply missing. Data stewards, data scientists, and developers spend too much time finding, wrangling, and cleaning their training and test data to ensure reliable *results*. Only recently has our community begun to recognize such shortcomings as a research (and tooling) opportunity. We examine data quality problems through all stages of the data science pipeline – from the mundane, such as unexpected field delimiters, to the complex, such as violations of data dependencies. We explore methods to discover and repair such problems and point to the still many open research challenges in the field of data quality and data preparation.

1 Bad Files and Bad Data

Raw data come in many shapes and forms, most of which are not what a data engineer, data scientist, or an analytics tool expects. And more often than not, even after massaging the data into an amenable format, the data themselves might contain errors, have missing values, or are outdated. Incorrectly read files and poorly cleaned data lead to incorrect or poor decisions – by humans analyzing the data or by machines building models based on that data, following the well-known garbage-in-garbage-out principle.

Information systems research has developed a rich foundation on the topic of *information quality*, encompassing a wide range of quality dimensions [15] to be assessed [12] and potentially improved through organizational and technical measures. Database research has traditionally focused on the data quality dimension of *accuracy*, essentially devising methods to identify erroneous data, such as duplicates [1] or violations of data dependencies [8].

Raw data are rarely in a shape that can be directly consumed by down-stream applications. Rather, they need to be prepared. In fact, Trifacta's data preparation study shows that 72% of respondents indicated that data preparation by data users is critical [14]. Data scientists spend approximately 80% of the time on collecting and preparing data and about 20% on actual model implementation and deployment [4, 9, 13].

Yet beyond "bad data", a new dimension of information quality has only recently been identified and is only beginning to be systematically addressed: "bad files". Typical problems in csv-files include multiple tables in a single file, titles, footnotes

and other metadata mingling among the data [6], aggregate rows, uncaught reserved characters, heterogeneous delimiters, empty rows, and many other issues that deviate from the (rather loose) standard for csv-files [11]. For instance, among 23k open data files a study identified 14 different encodings, five different delimiters, and up to 226 tables in a single file [2]. Such files typically cannot even be loaded into the target system. Further, even when data can be loaded from raw files, many data preparation tasks along the data-engineering pipeline remain, such as standardizing formats [7], splitting columns, or detecting disguised missing values [10].

2 Data Preparation and Data Cleaning

To achieve high quality results and insights from data, they must usually undergo many syntactic and semantic transformations: data preparation and data cleaning. Data preparation is the set of operations performed in early stages of a data processing pipeline, i.e., transformations at the structural and syntactical levels, which are independent of the data content. In contrast, data cleaning concerns subsequent data transformations and corrections at the semantic level, i.e., correcting erroneous data. Figure 1 shows this spectrum.

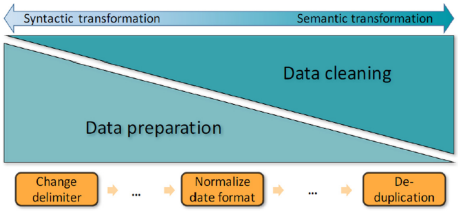

Fig. 1. Data preparation vs. data cleaning from [3].

While there is a rich literature in the field of data cleaning [5] and commercial products abound, the field of data preparation is only budding [3], despite its great potential both in automation opportunities and in time-savings for data engineers and data scientists [4]. Open or yet unsatisfyingly solved challenges include the automatic extraction of data from human-readable files, the standardization of data values and row formats, the automatic suggestion of preparation steps, and finally the ability to properly load any relevant data file into a system without human intervention.

References

1. Christen, P.: Data Matching. Springer Verlag, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31164-2
2. Christodoulakis, C., Munson, E., Gabel, M., Brown, A.D., Miller, R.J.: Pytheas: pattern-based table discovery in CSV files. PVLDB **13**(11), 2075–2089 (2020)

3. Hameed, M., Naumann, F.: Data preparation: a survey of commercial tools. SIGMOD Record **49**(3), 18–29 (2020)
4. Hellerstein, J.M., Heer, J., Kandel, S.: Self-service data preparation: research to practice. IEEE Data Eng. Bull. **41**(2), 23–34 (2018)
5. Ilyas, I.F., Chu, X.: Data cleaning. ACM (2019)
6. Jiang, L., Vitagliano, G., Naumann, F.: Structure detection in verbose CSV files. In: Proceedings of the International Conference on Extending Database Technology (EDBT), pp. 193–204 (2021)
7. Jin, Z., Anderson, M.R., Cafarella, M.J., Jagadish, H.V.: Foofah: transforming data by example. In: Proceedings of the International Conference on Management of Data (SIGMOD), pp. 683–698 (2017)
8. Pena, E.H.M., Filho, E.R.L., de Almeida, E.C., Naumann, F.: Efficient detection of data dependency violations. In: Proceedings of the International Conference on Information and Knowledge Management (CIKM), pp. 1235–1244 (2020)
9. Press, G.: Cleaning data: most time-consuming, least enjoyable data science task. Forbes, March 2016
10. Qahtan, A.A., Elmagarmid, A., Castro Fernandez, R., Ouzzani, M., Tang, N.: FAHES: a robust disguised missing values detector. In: Proceedings of the International Conference on Knowledge discovery and data mining (SIGKDD), pp. 2100–2109 (2018)
11. RFC 4180. https://tools.ietf.org/html/rfc4180. Accessed 12 Mar 2021
12. Sadiq, S., et al.: Data quality – the role of empiricism. SIGMOD Record **46**(4), 35–43 (2018)
13. Terrizzano, I.G., Schwarz, P.M., Roth, M., Colino, J.E.: Data wrangling: The challenging journey from the wild to the lake. In: Proceedings of the Conference on Innovative Data Systems Research (CIDR) (2015)
14. Trifacta end user data preparation. https://www.trifacta.com/wp-content/uploads/2018/02/End-User-Data-Preparation-Market-Study-2018.pdf. Accessed 19 Sept 2019
15. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. Manage. Inf. Syst. **12**(4), 5–34 (1996)

Contents

Process Understanding

Theory Development and Use

Platforms and Architectures

Models, Methods and Tools