Applied Data Science

Martin Braschler • Thilo Stadelmann •
Kurt Stockinger
Editors

Applied Data Science

Lessons Learned for the Data-Driven Business

Springer

Editors
Martin Braschler
Inst. of Applied Information Technology
ZHAW Zurich University
of Applied Sciences
Winterthur, Switzerland

Thilo Stadelmann
Inst. of Applied Information Technology
ZHAW Zurich University
of Applied Sciences
Winterthur, Switzerland

Kurt Stockinger
Inst. of Applied Information Technology
ZHAW Zurich University
of Applied Sciences
Winterthur, Switzerland

ISBN 978-3-030-11820-4 ISBN 978-3-030-11821-1 (eBook)
https://doi.org/10.1007/978-3-030-11821-1

Preface

In early 2013, the three editors of this volume were instrumental in founding the ZHAW Datalab, a Data Science Research Institute (DSRI)[1] at Zurich University of Applied Sciences.[2] "Big data" was big in the public press,[3] but DSRIs were still rare, especially in Europe. Both for our colleagues and us, it was the natural thing to do: joining forces to create synergies internally and demonstrate critical mass outwardly to ultimately facilitate better applied research projects (for which selecting project team members and acquiring funding becomes much easier in a larger group). The initial idea was to form a network of experts[4] that would engage in regular project-based collaborations without much administrative overhead. The goal of that partnership had been to perform applied research projects between academia and industry at the interface of structured and unstructured data analysis. The already existing strong partnership not only gave us confidence in the validity of the approach, it also made explicit the very need that led to the foundation of the ZHAW Datalab: the need for a concise description of what we were doing. Let us explain.

[1] A DSRI is a university-wide initiative integrating researchers from different disciplines predominantly occupied with a wide range of aspects surrounding the analysis of data.

[2] A snapshot of these beginnings is contained in Fig. 1. The founders of the ZHAW Datalab also published an early position paper (Stadelmann, Stockinger, Braschler, Cieliebak, Baudinot, Dürr and Ruckstuhl, "*Applied Data Science in Europe—Challenges for Academia in Keeping Up with a Highly Demanded Topic*", ECCS 2013). See also www.zhaw.ch/datalab

[3] The community that formed around the term was also very active, as can be seen, for example, in the history of meetings of the Swiss Big Data User Group (see https://www.meetup.com/swiss-big-data/). One of the editors (K.S.) even gave a talk at the first meetup of this group when it was still called "Swiss Hadoop User Group."

[4] A similar concept has been demonstrated by the Network Institute of the Vrije Universiteit Amsterdam (see http://www.networkinstitute.org/).

How We Became Data Scientists

Basically, we were able to capitalize on the emergence of the new data science field at exactly the right time. The ZHAW School of Engineering had been very successful in executing many applied research projects for more than 10 years prior to the inception of the ZHAW Datalab in 2013. Most of these projects were firmly "located" at the interfaces of the disciplines that today make up data science. In particular, computer scientists and statisticians joined forces and worked on problems integrating and analyzing structured and unstructured data. This was applied data science at work "par excellence." However, there was no "elevator pitch" for the kinds of problems we were working on together with our colleagues, no easy way to describe the ideas to funding agencies and prospective industry partners. If nothing less, the term "data science" delivered a concise description of what we perceived to be the field we were working in (Fig. 1).

One of the first joint activities within Datalab was to organize a workshop to perform a reality check on the potential of the topic of data science.[5] SDSI2014, the first Swiss Workshop on Data Science already exceeded our proudest expectation of attendees by a factor of 2 (see also Fig. 2); since then, the workshop has grown into a

Fig. 1 Five of the seven founders of the ZHAW Datalab in one of their first board meetings, with two of the editors (K.S. and T.S.) in the back row and the third (M.B.) taking the picture. The bottom row shows Gerold Baudinot (left), Andreas Ruckstuhl and Oliver Dürr (right), while Mark Cieliebak is missing (picture courtesy of T.S.)

[5]While a search conducted on LinkedIn for the terms "data scientist switzerland" returns more than 1500 hits as of early 2018, in 2013 it found only two persons (this credit goes to Violeta Vogel of PostFinance, and Daniel Fasel then of Swisscom: https://www.linkedin.com/in/violeta-vogel-3a556527/, https://www.linkedin.com/in/danielfasel/).

Fig. 2 Impressions from SDSI2014, the first Swiss Workshop on Data Science: Michael Natusch (left) delivers his insight into the core values of big data in front of parts of the audience (right; pictures courtesy of T.S.)

series of conferences that attracts a majority of the Swiss data science community. The growing interest in data science resulted in a significant increase of applied research projects that were initiated by the members of Datalab. Reflecting on the ever-growing number of people identifying themselves as data scientists and projects being described as data science projects led us to identify an additional need.

Why This Book Is Relevant

While data science builds on foundations from other disciplines, it is inherently an interdisciplinary and applied endeavor. The goal of data science is not only to work in one of its constituting sub-disciplines per se (e.g., machine learning or information systems), but to apply such methods and principles to build data products for specific uses cases that generate value from data. While very valuable textbooks exist on the individual subdisciplines,[6] the data science literature is missing a volume that acknowledges the applied science context of data science by systematically showing the connection between certain principles and methods, on the one end, and their application in specific use cases, on the other. One of the major goals of this book is to provide the reader with relevant lessons learned from applied data science projects at the intersection of academia and industry.

How to Read the Book

This book is organized into three parts: Part I pays tribute to the interdisciplinary nature of data science and provides a common understanding of data science terminology for readers with different backgrounds. The book is not a replacement for classical textbooks (i.e., it does not elaborate on fundamentals of certain methods

[6]See for example http://www.learndatasci.com/free-data-science-books/

and principles described elsewhere), but defines applied data science, the work of a data scientist, and the results of data science, namely, data products. Additionally, Part I sheds light on overarching topics such as legal aspects and societal risks through widely applied data science. These chapters are geared toward drawing a consistent picture of data science and are predominantly written by the editors themselves. We recommend the reader to work through the first four chapters in order.

Part II broadens the spectrum by presenting views and insights from diverse authors—some from academia, some from industry, some from Switzerland, some from abroad. These chapters describe a fundamental principle, method, or tool in data science by means of analyzing specific use cases and drawing concrete lessons learned from them. The presented case studies as well as the applied methods and tools represent the nuts and bolts of data science. The chapters in Part II can be read in any order, and the reader is invited to focus on individual chapters of interest.

Part III is again written from the perspective of the editors and summarizes the lessons learned of Part II. The chapter can be viewed as a meta study in data science across a broad range of domains, viewpoints and fields. Moreover, the chapter provides answers to the following question: What are the mission critical factors for success in different data science undertakings? Part III is written in a concise way to be easily accessible even without having read all the details of the case studies described in Part II.

Who Should Read the Book

While writing and editing the book, we had the following readers in mind: first, practicing data scientists in industry and academia who want to broaden their scope and enlarge their knowledge by assimilating the combined experience of the authors. Second, decision-makers in businesses that face the challenge of creating or implementing a data-driven strategy and who want to learn from success stories. Third, students of data science who want to understand both the theoretical and practical aspects of data science vetted by real case studies at the intersection of academia and industry.

Thank You

We thank you, the reader, for taking the time to learn from the collected insights described in this book. We as editors are university lecturers and researchers in our primary job; it is an immense pleasure and honor for us to be able to convey our insights. We are also very grateful for the trust and patience we received from our publisher, Springer, specifically impersonated by Ralf Gerstner. We want to thank our coauthors that contributed excellent work that is fundamental for making this

book a success. Finally, we thank our students, colleagues, and partners from the Datalab, the Master of Advanced Studies in Data Science Program, and the Swiss Alliance for Data-Intensive Services for providing the environment in which this book project (and some of the reported use cases) could flourish.

Specifically, I (Martin Braschler) thank my co-editors for consistently engaging and stimulating discussions, Vivien Petras and the team of the Berlin School of Library and Information Science at the Humboldt-Universität zu Berlin for hosting me during part of the period I worked on this book, my colleagues that I have collaborated with in past projects and who have thus informed my understanding of data science topics, and last but not least my family, who provides for me the much needed balance to life as a university teacher and researcher.

I (Thilo Stadelmann) thank my co-editors and Michael Brodie for helpful discussions and valuable lessons in collaboration. Thank you for your patience and collegiality. I learned a lot. Thanks go to Geri Baudinot for enabling the ZHAW Datalab and further developments by his vision, patronage, and mentorship. My final "thank-you" is best expressed with a quote adapted from Reuben Morgan: "*Freely you gave it all to me… great is the love, poured out for all, this is my god.*"

I (Kurt Stockinger) thank my wife Cinthia and my two little kids Luana and Lino for the ability to work on the book during calm evening hours—after having changed diapers and read several good night stories that did not contain data science topics.

Winterthur, Switzerland
Spring 2019

Martin Braschler
Thilo Stadelmann
Kurt Stockinger

Contents

Part III Lessons Learned and Outlook