# EDITORIAL: RESEARCH ON DATA SCIENCE EDUCATION

SPECIAL ISSUE OF THE STATISTICS EDUCATION RESEARCH JOURNAL

*Guest Editors*

ROLF BIEHLER
*Paderborn University*
*biehler@math.upb.de*

RICHARD DE VEAUX
*Williams College*
*rdeveaux@williams.edu*

JOACHIM ENGEL
*University of Education Ludwigsburg*
*engel@ph-ludwigsburg.de*

SIBEL KAZAK
*Pamukkale University*
*skazak@pau.edu.tr*

*Special Edition Editor*

DANIEL FRISCHEMEIER
*University of Münster*
*dfrische@uni-muenster.de*

A very warm welcome to this Special Issue of the Statistics Education Research Journal (SERJ) on data science education. Our hope is to give an overview of selected theoretical thoughts and empirical studies on data science education from a statistics education research perspective. Data science education is rapidly developing but research into data science education is still in its infancy. The current issue presents a snapshot of this developing field.

Data science is an emerging discipline at the intersection of statistics, computer science, and application domains. From a statistics education perspective, this discipline requires learners to gain "new skills" to explore and make sense of large and messy datasets, so-called big data (Gould, 2017; Ridgway, 2016). With the proliferation of big data in many areas such as industry or social sciences, competent reasoning about data has become even more critical (Biehler et al., 2018). With the influence of big data infiltrating nearly every sphere of social interaction, changes are necessary at many levels of the education,including both high school and university. Having the citizenry gain a robust understanding of data in areas such as migration, global warming, health, and poverty, so-called civic statistics (Engel, 2017), has become critical for ensuring the sustainability of democratic processes, debates and conversations into the future. Unfortunately, however, data do not necessarily come in a tidy format, and so data handling and management skills have become more important to address this need. Due to these trends, modeling, machine learning methods, and digital tools have been given even more prominent roles. Educating students in data science goes beyond teaching about algorithms, skills of manipulating data sets, selecting and applying appropriate analyses, and creating and interpreting visual representations of data. It also involves raising a critical understanding of how data are produced and how they can be used for particular purposes, including the role of context in interpreting data. It emphasizes developing an awareness for data ethics, and considering the implications for policy and society when powerful algorithms are used.

Data science is one of the research subjects of the 21st century, and in times of fake news and alternative facts, etc., interpreting and critically appraising data has become more essential than ever. Fortunately, data science provides the powerful 21st century skills needed for problem solving and creating new knowledge in many domains and tackling the issues societies face; therefore, data science has become a research subject in its own right. The large participation in recent conferences indicates the strong interest of the statistics education community in topics related to data science and it is exciting to see how research is addressing the aforementioned issues.

The field "data science", however, is still somewhat nebulous from a statistics education perspective. There are a wide variety of curricular approaches, including such programs as the International Data Science in Schools Project IDSPP (http://www.idssp.org/), the IDS project of the University of California (https://www.idsucla.org/), and the ProDaBi project (Biehler et al., 2018; see also www.prodabi.de/en). This Special Issue aims to synthesize curricular approaches and research findings on data science education, and discusses state of the art and future trends of data science

*Statistics Education Research Journal, 21(2), Article 1. https://doi.org/10.52041/serj.v21i2.606*

(education) that we hope will inspire ideas for the teaching and learning of data science at all levels—from secondary and tertiary levels of education as well as into the workplace. Relevant questions which are tackled in the contributions of this special issue are:

- How do we prepare people to cope with the complexity of big data?
- What knowledge, skills, and dispositions are required to develop data acumen in data science?
- What are the roles of statistics, computer science, and domain knowledge in a data science curriculum?
- Which new topics should be included in the curriculum (e.g., machine learning, predictive modeling)?
- How differently should traditional topics be taught from the perspective of data science?
- What are new ways to engage students in studying data science?
- What are the challenges of integrating data science into the school/undergraduate statistics courses or designing a data science curriculum at the school level/undergraduate statistics?
- What are effective ways to support teachers/instructors implementing aspects of data science in schools/at the tertiary level?

In total, this special issue includes eleven contributions from different perspectives analyzing and discussing current or innovative practices about the teaching and learning of data science, developing theories and frameworks for teaching and research in data science (education), or offering empirical insights in learners´ reasoning when dealing with data science problems. These papers address issues related to teaching data science at all levels: secondary school students, undergraduate students, teachers, and practicing data scientists. We identify and distinguish three categories of contributions:

1) Theoretical contributions and frameworks for data science education in different fields (school, teacher education, undergraduate studies).
2) Design-based and empirical contributions to teaching and learning data science with secondary school students or in teacher education.
3) Issues relating to teaching and learning data science in different fields at the undergraduate level and the workplace.

*Theoretical contributions.* We can observe theoretical considerations and approaches to data science education from different perspectives in the first three articles. In the first article of this special issue, Richard De Veaux, Roger Hoerl, Ronald Snee and Paul Velleman argue for an holistic approach to teaching data science. Their concept of an holistic data science curriculum is placed in the context of applications and solving specific problems. It emphasizes the meaning inherent in the data, their quality and background of the data, and brings ethical considerations and implications of data science solutions for society into focus while spending less time on algorithms and technology.

Taking the perspective of teaching and learning data science at the school level, Hollylynne Lee, Gemma Mojica, Emily Thrasher, and Peter Baumgartner identify key practices and processes for K–12 data science education. More specifically, the authors present a data investigation process framework that includes six phases: *frame problem*, *consider & gather data*, *process data*, *explore & visualize data*, *consider models*, and *communicate & propose action*. According to the authors, this framework can be implemented in K–12+ classroom settings where students learn and apply data science to investigate issues across various domains and curricula strands (e.g., mathematics, statistics, sciences, social sciences, humanities, engineering).

Concerning data science education at the undergraduate level, the idea of data science projects is a point of discussion in the paper of Mine Çetinkaya-Rundel, Mine Dogucu, and Wendy Rummerfield. The authors argue for the value and importance of data science projects for introductory data science courses and specify their notion of data science projects along the so-called 5Ws (What? Why? Who? When? Where?) and 1H (How?). Taking up these W-questions words, the authors clarify—amongst other things—what they mean by a data science project, what the learning goals are, why an introductory data science course includes a project, and who works on the projects. In addition, Çetinkaya-Rundel et al. discuss when data science projects should take place and where student projects can be shared to give valuable practical information on how to include data science projects in the teaching and learning of data science.

*Design-based and empirical studies*. Investigating the implementation of data science education at school and for teacher education can be found in the contributions of Heinzman, Podworny et al., Fleischer et al., and Fergusson and Pfannkuch. Heinzman reports on a case study of student experiences in the introduction of a data science course (IDS) at the University of California. Based on the framework of self-determination theory, the findings of the case study suggest IDS students found meaning and empowerment in data science and expressed a new sense of confidence, agency, and belonging in contrast to previous mathematics and statistics courses they had attended.

The two contributions from the ProDaBi project (Project Data Science and Big Data in school, see www.prodabi.de/en) report on experiences of implementing data science in secondary school classrooms in Germany. In the contribution, "A place for a data science introduction in school: Between statistics and programming," Susanne Podworny, Sven Hüsing, and Carsten Schulte describe a data science teaching unit focusing on the analysis of environmental data experience with digital programming tools. A new insight-driven programming approach was implemented for Grade 9 students (aged 14–16) using *Jupyter Notebook* and the programming language *Python* for the data analysis. Podworny et al. investigated how the lower secondary school students coped with the programming within Jupyter Notebook for doing statistical investigations and stated that worked examples proved to be an appropriate pedagogical tool that empowered the students to go beyond methods usually used in Grade 9.

In another contribution from the ProDaBi-project, Yannik Fleischer, Rolf Biehler, and Carsten Schulte investigate how far upper secondary students can model with machine learning algorithms, in this case, with automatically created decision trees. More specifically, this study explored how secondary students applied machine learning methods using Jupyter Notebook and how they documented the modeling process in the form of a so-called computational essay, which took into account the different steps of the CRISP-DM cycle. Their empirical study shows that their participants were able to adopt and adapt the code from worked examples to create a decision tree model and create visualizations for evaluation based on test data. In addition to that, all of their students established a narrative for their machine learning process.

Anna Fergusson and Maxine Pfannkuch describe the application of a design-based research approach to develop web-based tasks to introduce high school statistics teachers to predictive modeling and APIs using code-driven tools. The tasks were implemented within a professional development workshop involving six high school teachers. First analyses reveal that the web-based task supported the development of new statistical and computational ideas related to predictive modeling and APIs. Furthermore, all six teachers in their study were able to use a code-driven tool to interact with APIs and develop a model that generated prediction intervals.

*Teaching and learning data science in different fields*. Finally, the papers by Vance et al., Mike and Hazzan, Bolch and Crippen, and Bilgin et al. discuss the teaching and learning of data science at the undergraduate levels and in the workplace. In the contribution of Eric Vance, David Glimp, Nathan Pieplow, Jane Garrity, and Brett Melbourne, the reader learns how to integrate the humanities into data science education. Vance et al. introduce an interdisciplinary data science course (IIIDS) aiming to merge STEM and humanities perspectives at the beginning of the data science curriculum. The data science course aimed to attract a broader range of students and instill data acumen into the humanities.

In their article, Koby Mike and Orit Hazzan report on how to teach machine learning to non-major data science students. Specifically, they point out how to realize a white-box approach. They suggest a pedagogical method and a learning module based on hands-on tasks to support white-box understanding of machine learning algorithms for learners who do not necessarily bring to the task the required mathematical knowledge. In an accompanying study based on a survey and applying the process-object theory to the analysis of the survey data, the authors provide evidence of the effectiveness of the approach.

Charlotte Bolch and Kent Crippen used a Delphi method to understand the experiences of data scientists regarding common skills and strategies for interpreting and creating data visualizations. Their study with researchers in the field of data science and researchers whose projects involves components of data science, provides empirical evidence for a consensus of skills and strategies that data scientists show when interpreting and creating visualizations. The work of Ayse Bilgin, Angela Powell, and

Deborah Richards deals with the novel field of work-integrated learning in data science. In their contribution, the authors present an assessment framework for a data science unit and—amongst other things—possibilities to weight students' final marks in work-integrated learning, which can be valuable for courses in a similar context.

All eleven contributions come from different perspectives, have different backgrounds, involve different learners, and take different approaches to data science (education), which makes this SERJ issue special indeed. However, this special issue should only be seen as the beginning of further and more profound empirical research in data science education. More qualitative, design-based, and quantitative research is needed to investigate the integration of data science at all educational levels and beyond.

We sincerely hope that you will find all these papers useful, interesting and thought provoking.

## ACKNOWLEDGMENTS

Within the process of editing this special issue, many people have supported us immensely. We are incredibly grateful to Noleine Fitzallen, who has helped us tremendously as the SERJ assistant editor in improving the quality, the formatting, and the language of the papers of the non-native authors with her very helpful, constructive, and intensive feedback. We thank all reviewers for their support in writing constructive and helpful reviews to improve the quality and writing of the papers in this special issue. Last but not least, we are very grateful to all authors who have contributed to this special issue.

## REFERENCES

Biehler, R., Budde, L., Frischemeier, D., Heinemann, B., Podworny, S., Schulte, C., & Wassong, T. (Eds.) (2018). *Paderborn symposium on data science education at school level 2017: The collected extended abstracts.* Universitätsbibliothek Paderborn. https://doi.org/10.17619/UNIPB/1-374

Biehler, R., Frischemeier, D., Reading, C. & Shaughnessy, M. (2018). Reasoning about data. In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 139–192). Springer. https://doi.org/10.1007/978-3-319-66195-7_5

Engel, J. (2017). Statistical literacy for active citizenship: A call for data science education. *Statistics Education Research Journal, 16*(1), 44–49. https://doi.org/10.52041/serj.v16i1.213

Gould, R. (2017). Data literacy is statistical literacy. *Statistics Education Research Journal, 16*(1), 22–25. https://doi.org/10.52041/serj.v16i1.209

Ridgway, J. (2016). Implications of the data revolution for statistics education. *International Statistical Review, 84*(3), 528–549. https://doi.org/10.1111/insr.12110