Report on the First International Workshop on Data Systems Education (DataEd '22)

Efthimia Aivaloglou
University of Leiden
the Netherlands
e.aivaloglou@liacs.leidenuniv.nl

George Fletcher
Eindhoven University of Technology
the Netherlands
g.h.l.fletcher@tue.nl

Michael Liut
University of Toronto
Mississauga, Canada
michael.liut@utoronto.ca

Daphne Miedema
Eindhoven University of Technology
the Netherlands
d.e.miedema@tue.nl

ABSTRACT

This report summarizes the outcomes of the first international workshop on Data Systems Education: Bridging Education Practice with Education Research (DataEd '22). The workshop was held in conjunction with the SIGMOD '22 conference in Philadelphia, USA on June 17, 2022. The aim of the workshop was to provide a dedicated venue for presenting and and discussing data management systems education experiences and research by bringing together the database and the computing education research communities to share findings, to cross-pollinate perspectives and methods, and to shed light on opportunities for mutual progress in data systems education. The program featured two keynote talks, ten research paper presentations, a discussion session, and an industry panel discussion. In this report, we present the workshop's main results, observations, and emerging research directions.

1. INTRODUCTION

Interest in data systems education (DSE) is increasing, especially with the rise in demand for well trained and re-trained data scientists. Data systems education is foundational in programs such as computer science, data science, and information systems and science. A continual focus since the 1970's in the database research community is the place in curricula and best practices for teaching data systems concepts. There is also a long tradition in both the computing education (CSE) and computer science education (CDEd) community on research into how students learn data systems concepts. For example, a recent research space in the community is understanding the difficulties students face when learning query languages such as SQL and how teachers might improve query language educational practices [12, 17]. Both the DSE and CSE communities, and adjacent communities, e.g., in statistics education, have complementary perspectives and experiences to share with each other, and there is much to be gained by bringing them together: to share findings, to cross-pollinate perspectives and methods, and to shed light on opportunities for mutual progress.

The DataEd workshop[1] was organized as a dedicated venue for the presentation and discussion of data systems education research. DataEd focused on the broad area of data systems education: the teaching and learning of databases, data management, and data systems topics, ranging across the whole field, from classical topics, such as physical design, query optimization, data modeling, data integration, visual analytics, and query languages) to contemporary topics, such as ML & AI for data management systems, data management for ML & AI, very large data science applications/pipelines, and responsible data management.

DataEd '22 had a strong focus on encouraging interaction among the participants. It took place as a full day workshop consisting of:

1. a keynote talk *Data-Centricity: Rethinking Introductory Computing to Support Data Science* by Kathi Fisler (Brown University)
2. a keynote talk *Teaching Responsible Data Science* by Julia Stoyanovich (NYU Tandon School of Engineering)
3. ten research paper presentations & discussions
4. an industry panel on industry perspectives on education and training for emerging roles in data organized by Juan Sequeda (data.world), with panelists Sarah Krasnik (independent) and Emilie Schario (Amplify)
5. a discussion session on topics prioritized by the attendees, including curriculum placement & content of data systems topics and assessment types

[1] https://dataedinitiative.github.io/DataEd22

In the following section, we present the themes that emerged from the various workshop activities.

2. WORKSHOP THEMES

2.1 Course and Curriculum Design

The design and curriculum integration of courses related to data, data systems, and data management was one of the main themes that came up in the workshop. Those topics were addressed in Kathi Fisler's keynote [6], three papers, plus the discussion session on curriculum placement and content of data management topics.

In *Data-Centricity: Rethinking Introductory Computing to Support Data Science* [6], Kathi Fisler presented a novel introductory computing course which combines data science, data structures, and socially-responsible computing. The course is data-centric, focusing on data science and engineering topics, while covering the necessary content for an introductory computing course. The course is designed to support students across majors, introducing computing concepts via data in familiar formats, such as images and two-dimensional tables, before moving to more advanced data types. It starts with Pyret (a functional programming language with Pythonesque syntax which has been developed for education) before introducing Python.

In *Piloting Data Engineering at Berkeley* [8], Joseph M. Hellerstein and Aditya G. Parameswaran present the Data Engineering course that was designed targeting the Data Science major at Berkeley. The course focuses on fluency of data models and transformation tasks, using SQL as the primary language. Moving from course design to specific aspects, paper *Instructional Design for Teaching Relational Query Optimization to Undergraduates* by Karen C. Davis [4] presents a module designed to fit within database systems courses to teach query optimization. The module includes both logical optimization concepts and physical optimization ones, and includes quizzes and a project covering computation of database statistics, and performance of selection and join algorithms. A timely intervention is proposed by Alan G. Labouseur in *Managing Data... and Covid – An Experience Report* [10], which discusses experiences with integrating real-world practice in a data management course by utilizing a Covid-19 screening database for practicing with queries, aggregates, joins, stored procedures, and reports.

The integration of data management topics in the curriculum also came up in the *discussion session*. Commenting on the placement of the databases or data systems courses, most participants indicated that these are towards the middle or second half of the CS Bachelors studies in their institutions. While this may limit reuse and practice opportunities via subsequent projects, other advanced courses such as web systems are sometimes ran in parallel to allow for such opportunities. The core elements of the databases courses were found to be mostly uniform across institutions, with outlying topics being Datalog, database security and SQL injections, especially if they are not covered by other courses in the curriculum. Discussing the concepts and topics that should be considered as core, even if their difficulty is high, participants mentioned declarativeness and conceptual modelling.

2.2 Learning Instruments, Tools, and Practices

A fundamental component to database systems, more specifically DSE, is the necessity to utilise novel learning instruments and tools to aid in the delivery of key material concepts but also better the student experience in data systems and database courses. Over the last decade there have been substantial improvements in the DSE space, much about the curriculum, methods, and tools [9]. It is clear that we need to modernize our courses, as suggested by Kathi Fisler in her keynote [6], but also there is the general need to boost engagement in computer science through novel teaching approaches [2, 7, 13], as well as the shifting change and training needs for students to succeed in industry (see Subsection 2.5).

The research presented at DataEd '22 is a direct extension to this progress, as several DSE learning instruments were presented; a gamified experience [14] for students to learn about SQL injection attacks, engaging learners with a graphical user interface [1] to teach data models, and leveraging the community at large for datathons [11] to allow students to practice their data science skills with real use cases and datasets. There is even research being conducted into student reflections and their impact on students' ability to learn, retain, and apply knowledge after being prompted to reflect on material [15], as well as experience reports and improvements into the curricula [5].

The overarching motive of all the papers in this theme is the need for student engagement. This has been undertaken in many different ways by the authors presenting at the workshop, such as novel tools and creative assignments. However, there are many more avenues of engaging database education research which are open to take on, and the need to continue to collaborate with the community at-large remains important to ensure we stay on track to best educate our students.

2.3 Ethics and Responsibility

Responsible computing, data science, and AI in data education was another major theme of the workshop. The highlight of this discussion was Julia Stoyanovich's keynote *Teaching Responsible Data Science* [16]. In this talk, an overview of the critical and widespread ethical, legal, data quality, fairness, transparency, privacy, and data protection challenges of contemporary data science was presented. In response to these vital challenges, two new courses developed by Julia Stoyanovich at NYU were presented in depth. The first is a technical course for undergraduate and graduate students on "Responsible Data Science", which introduces these challenges through theory and hands on work, striking a balance between techno-optimism (solutionism) and techocriticism. The second is a public education course "We are AI: Taking Control of Technology", based on a peer-learning format for a non-technical general audience. A special feature of the courses is their open format, freely available online to the public, with ample rich materials to be adopted and extended in a wide variety of settings[2]. These critical efforts by Julia Stoyanovich in data education are already filling a vital need in development of education and training for those working with data in public and private enterprises.

In Kathi Fisler's keynote, discussed in Section 2.1, topics of ethics and responsibility are also already confronting students in introductory CS education. Issues such as data cleaning and data quality are used as platforms to highlight the social context of data-centric work, introducing students to these critical and difficult challenges from the early stages of their education.

2.4 Formative and Summative Assessment

The final theme that came up in the workshop was that of assessment approaches for data systems education. The main three avenues for this theme were two papers [3, 18], and the discussion session on teaching.

In *Analyzing Student SQL Solutions via Hierarchical Clustering and Sequence Alignment Scores* [18], the authors aim to explore the problem-solving behavior of their students. They do this by 1) calculating the alignment between the n$^{\text{th}}$ solution and the final solution and 2) clustering solutions to determine different approaches used by students to solve the problem. Their system offers these metrics to the instructors of the course to visualize their students' learning progress. The authors hope that this can help the instructors identify SQL concepts that warrant more in-depth instruction.

In *Collaborative Learning in an Introductory Database Course: A Study with Think-Pair-Share and Team Peer Review* [3], the authors aim to evaluate whether the application of collaborative learning techniques can be beneficial in a data systems course. They selected Think-Pair-Share to test in lectures and lab sessions, and team peer-review for projects. Participation in these collaborative activities was optional, which meant that the authors could compare the course results of both a test and control group. They found that students who passed the course and had participated in collaborative activities had more homogeneous results than students who had passed the course but not participated. Furthermore, the students rated usefulness of the activities on a scale from 1 (definitely useless) to 4 (definitely useful). All three activities were, on average, ranked higher than 3, showing that students appreciated the activities. The authors hypothesize based on the results, that collaborative activities (and the summative assessment within it) lead to more balanced learning efforts compared to self-regulated learning.

In the *discussion session*, the two main points of discussion were assessment types and assessment creation.

Typical assessments in data systems include the creation of Entity Relationship Diagrams from a textual description and writing queries (in various languages) based on set requirements. However, as one participant noted, students make lots of mistakes in SQL query formulation. As such, they decided to first include it in a small project such that the student can be scaffolded into the more standard exercises mentioned above. Another supporting mechanism is to use group work, such that students can discuss what parts of their answers might be (in)correct.

Finally, assessment creation is seen as a challenge. Coming up with new questions on interesting topics is a drain on creative resources of lecturers and teaching assistants. However, some of our participants came up with ingenious ways of finding subjects for their questions. Some ideas include: asking your kid(s) to come up with a topic, choosing a random Wikipedia category, or using pop culture references such as Disney and Marvel. One way in which this issue could be abated is to create a shared repository of questions, where many teachers add theirs, such that we end up with a resource of thousands of questions. However, students will be able to find it, which might be a problem in case of online exams.

Overall, it seems that most innovations in data systems assessment are based on trial-and-error. From the teacher perspective, we are looking for efficient creation of assignments. Here the findings from Yang et al. [18] may help to identify the (SQL specific) topics that students need more practice in. On the student side, we are looking for in-depth understanding. The findings from Catania et al. [3] suggest that applying collaborative active learning techniques may help students learn in a more balanced way.

2.5 Industry Perspectives on Data Management Knowledge and Skills

The workshop closed with a panel discussion on industry perspectives on education and training for emerg-

[2] `https://dataresponsibly.github.io/we-are-ai/`

ing roles in data, organized by Juan Sequeda (Principal Scientist, data.world), with panelists Sarah Krasnik (Data Engineering and Analytics Advisor, independent) and Emilie Schario (Data Strategist in Residence, Amplify). George Fletcher moderated the discussion.

Several themes arose during the opening position presentations by the panelists and the ensuing discussion with the workshop attendees.

A central theme was gaps between practice and university curricula. Example topics in this gap highlighted were: relatively little coverage of topics in dynamic and streaming data management; principled methods and frameworks for choosing which solutions and technologies to use in a given practical data engineering task; data integration solutions; (re)aligning emerging and established data roles (such as data engineer, analytics engineer, data analyst, machine learning engineer) with university curricula; and, mapping between academic education and industry norms around data workflows and the "modern" data stack. A second related theme was that of balancing generality and ideas which transcend current practice, on the one hand, and mapping these general concepts and perspectives to current practice, on the other hand. What are the perennial ideas with practical impact which are currently poorly covered in curricula? A final theme which arose was that of the role of education and academic curricula with respect to professional and non-technical skills. What should be covered in education? Which topics are better learned outside the classroom? How do we bridge training and practice? This lively discussion was a perfect way to close a very productive and stimulating day.

3. CONCLUSIONS AND EMERGING RESEARCH DIRECTIONS

We identified five main themes which arose during the workshop: course and curriculum design, learning instruments, ethics and responsibility, assessment, and industry requirements. Each of the subsections describing these themes can be seen as illustration of an upcoming research direction within data science education. Clearly, much more work is needed in each of these areas. We hope that DataEd will continue to inspire research efforts on data systems education, in both the aforementioned themes, as well as in new directions.

We aim to continue DataEd as a workshop under SIGMOD, with an ultimate goal to create a space in the data management research community for both computing and computer science education research. Furthermore, under the DataEdInitiative umbrella[3], we aim to bridge the gap between CSE/CSEd and Data Systems researchers with 'sister' activities to DataEd. As such, we will be organizing activities in the CSEd community in the future, with one possible avenue being a working group (for instance, those that occur at ITiCSE[4]).

Acknowledgement

The workshop's advisory board members supported setting the scope of the workshop and provided valuable feedback at all stages of its preparation. We also appreciate the efforts of DataEd's program committee in reviewing the submissions. We would like to thank all of them for their valuable support and encouragement. In addition, we would like to thank all participants for their valuable contributions, discussions, and feedback to DataEd '22.

4. REFERENCES

[1] Abdussalam Alawini, Peilin Rao, Leyao Zhou, Lujia Kang, and Ping-Che Ho. Teaching Data Models with TriQL. In *1st International Workshop on Data Systems Education*, DataEd '22, page 16–21. ACM, 2022.

[2] Ricardo Caceffo, Guilherme Gama, and Rodolfo Azevedo. Exploring active learning approaches to computer science classes. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education*, pages 922–927, 2018.

[3] Barbara Catania, Giovanna Guerrini, and Daniele Traversaro. Collaborative Learning in an Introductory Database Course: A Study with Think-Pair-Share and Team Peer Review. In *1st International Workshop on Data Systems Education*, DataEd '22, page 60–66. ACM, 2022.

[4] Karen Collins Davis. Instructional Design for Teaching Relational Query Optimization to Undergraduates. In *1st International Workshop on Data Systems Education*, DataEd '22, page 44–50. ACM, 2022.

[5] Alan Fekete. Teaching Data Management Concepts for Data in Files. In *1st International Workshop on Data Systems Education*, DataEd '22, page 51–55. ACM, 2022.

[6] Kathi Fisler. Data-Centricity: Rethinking Introductory Computing to Support Data Science. In *1st International Workshop on Data Systems Education*, DataEd '22, page 1–3. ACM, 2022.

[7] Michail N Giannakos, John Krogstie, and Nikos Chrisochoides. Reviewing the flipped classroom research: reflections for computer science education. In *Proceedings of the computer science education research conference*, pages 23–29, 2014.

[3] https://dataedinitiative.github.io

[4] https://sigcse.org/events/workinggroups.html

[8] Joseph M. Hellerstein and Aditya G. Parameswaran. Piloting Data Engineering at Berkeley. In *1st International Workshop on Data Systems Education*, DataEd '22, page 38–43. ACM, 2022.

[9] Muhammad Ishaq, Adnan Abid, Muhammad Shoaib Farooq, Muhammad Faraz Manzoor, Uzma Farooq, Kamran Abid, and Mamoun Abu Helou. Advances in database systems education: Methods, tools, curricula, and way forward. *Education and Information Technologies*, pages 1–45, 2022.

[10] Alan Labouseur. Managing Data...and Covid An Experience Report. In *1st International Workshop on Data Systems Education*, DataEd '22, page 56–59. ACM, 2022.

[11] Antonella Longo, Marco Zappatore, Angelo Martella, and Chiara Rucco. Enhancing Data Education with Datathons: An Experience with Open Data on Renewable Energy Systems. In *1st International Workshop on Data Systems Education*, DataEd '22, page 26–31. ACM, 2022.

[12] D. Miedema, E. Aivaloglou, and G. Fletcher. Identifying SQL Misconceptions of Novices: Findings from a Think-Aloud Study. In *Proc. ICER 2021*, page 355–367. ACM, 2021.

[13] Thomas L. Naps, Guido Rößling, Vicki Almstrum, Wanda Dann, Rudolf Fleischer, Chris Hundhausen, Ari Korhonen, Lauri Malmi, Myles McNally, Susan Rodger, and J. Ángel Velázquez-Iturbide. Exploring the role of visualization and engagement in computer science education. *SIGCSE Bull.*, 35(2):131–152, jun 2002.

[14] Johannes Schildgen and Jessica Rosin. Game-Based Learning of SQL Injections. In *1st International Workshop on Data Systems Education*, DataEd '22, page 22–25. ACM, 2022.

[15] Naaz Sibia and Michael Liut. The Positive Effects of Using Reflective Prompts in a Database Course. In *1st International Workshop on Data Systems Education*, DataEd '22, page 32–37. ACM, 2022.

[16] Julia Stoyanovich. Teaching Responsible Data Science. In *1st International Workshop on Data Systems Education*, DataEd '22, page 4–9. ACM, 2022.

[17] T. Taipalus and V. Seppänen. SQL Education: A Systematic Mapping Study and Future Research Agenda. *ACM Trans. Comput. Educ.*, 20(3), 2020.

[18] Sophia Yang, Geoffrey L. Herman, and Abdussalam Alawini. Analyzing Student SQL Solutions via Hierarchical Clustering and Sequence Alignment Scores. In *1st International Workshop on Data Systems Education*, DataEd '22, page 10–15. ACM, 2022.