Report on the Second International Workshop on Data Systems Education (DataEd '23)

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

Efthimia Aivaloglou
Delft University of Technology
Delft, the Netherlands
e.aivaloglou@tudelft.nl

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
Grenoble, France
ameryahs@univ-grenoble-alpes.fr

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

Michael Mior
Rochester Institute of Technology
Rochester, United States of America
mjmvcs@rit.edu

Toni Taipalus
Tampere University
Tampere, Finland
toni.taipalus@tuni.fi

ABSTRACT

This report summarizes the outcomes of the second international workshop on Data Systems Education: Bridging Education Practice with Education Research (DataEd '23). The workshop was held in conjunction with the SIGMOD '23 conference in Seattle, USA on June 23, 2023. 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, eight research paper presentations, and a discussion session. In this report, we present the workshop's main results, observations, and emerging research directions.

1. INTRODUCTION

Data systems education is foundational in programs such as computer science, data science, and information systems. The DataEd workshop[1] is organized as a dedicated venue for the presentation and discussion of data systems education research. DataEd took place for the first time at SIGMOD 2022 and provided the opportunity to discuss a broad range of topics, including data systems course and curriculum design, learning instruments, tools, and practices, ethics and responsibility, formative and summative assessment, and industry perspectives on data management knowledge and skills [1].

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

DataEd focuses 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 machine learning for data management systems, data management for machine learning, large data science applications and pipelines, and responsible data management.

DataEd'23 took place at SIGMOD 2023 as a second iteration with a full-day workshop consisting of:

1. A keynote talk *Human Learners of Relational Query Processing: Who Cares?* by Sourav Bhowmick (Nanyang Technological University)
2. A keynote talk *SQL: A Trojan Horse Hiding a Decathlon of Complexities* by Toni Taipalus (University of Jyväskylä)
3. Eight research and tool paper presentations with accompanying discussions (50% acceptance rate)

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

2. WORKSHOP THEMES

2.1 Difficulties in Learning Query Languages

A common theme in Data Systems Education research is the analysis of student errors. As such, errors and difficulties were also a prominent theme in the second iteration of DataEd.

In *Human Learners of Relational Query Processing: Who Cares?* [10], Sourav S. Bhowmick discussed how query optimization is a challenging task to learn let alone

master, and novices struggle with learning DBMS internals. Challenges include understanding query execution plans, incorrect ordering of steps, missing intermediate relations, and understanding cost estimation. What's more, DBMS vendors primarily target enterprise users for SQL training due to financial incentives, consequently disregarding novices learning query processing and optimization. Off-the-shelf DBMSs simply do not offer enough feedback on their query execution plans, possibly making learning query processing too challenging for effective learning, given learners of different abilities and backgrounds. To assist in learning query processing through more helpful, natural language feedback and query visualization tools such as ARENA [14], LANTERN [4] and MOCHA [13] were suggested.

In *SQL: A Trojan Horse Hiding a Decathlon of Complexities* [12], Toni Taipalus argued for acknowledging the gap between how simple SQL *looks* and how complicated the language actually *is* by highlighting ten complexities arising from the discrepancies between SQL's underlying principles, the language as it is defined by the SQL Standard, and implementations of SQL in various DBMSs. By acknowledging convolutions such as conflicts between the theory of three-valued logic and how three-valued logic is implemented, confusing error messages, and strange conventions regarding grouping, novices can better brace themselves for learning a challenging language from the get-go. Similarly to Bhowmick, Taipalus suggested using query visualization tools, and additionally focusing on teaching one SQL dialect with a DBMS that closely conforms to the SQL Standard, and refraining from treating SQL like a natural language.

Some of these concerns are further highlighted in *Student's Learning Challenges with Relational, Document, and Graph Query Languages* presented by Abdussalam Alawini [2]. The authors highlighted several challenges in learning query languages faced by students by analyzing over 350,000 student submissions. The authors found that semantic errors (incorrect results) in SQL queries were indeed a problem, with 35% of student submissions experiencing some kind of logic error. This problem was even more prevalent with MongoDB queries, where two-thirds of submissions contained a semantic error. Cypher queries on Neo4j also exhibited a significant error rate, with 40% of submissions containing a semantic error. The authors posit that students may have a more difficult time forming a mental model for document databases compared to relational and graph databases. This suggests a need to further explore methods for teaching query languages beyond SQL, particularly for NoSQL databases.

As one possible approach to furthering query language education, Michael Mior proposed *Relational Playground: Teaching the Duality of Relational Algebra and SQL*[8]. Relational Playground aims to provide students with the ability to thoroughly explore the connection between relational algebra and SQL. Students can enter SQL queries against a sample database and view the corresponding relational algebra expressions. The interface also makes it possible to view intermediate results at each stage of execution as well as the effects of some basic query optimization techniques. Further evaluation of the tool is required, but it shows promise at cementing student understanding of SQL query processing.

2.2 Tools/Automating Assessment

A second major theme is that of tool development. In this iteration of DataEd, we had four papers on tooling for education, to support students, teachers, or both.

The first talk was by Daniel Kocher, who discussed *Feedforward-Aided Course Designs for Similarity*[5]. The course designs he presented are called project-based learning and task-based learning and are of use for other teachers to reflect on and potentially adapt. In both course designs, they employ an auto-grader to provide students with automated and instant feedforward. This helps both students and lecturers, as it removes the workload of grading. It also supported students working with different programming languages, creating even more flexibility. However, they found that the heterogeneity of students also led to struggles, such as in different levels of programming and conceptualization knowledge. The discussion afterward led to suggestions for future development, such as creating individual feedforward and reducing the opportunities for gaming the system.

The second talk in this category was by Sihem Amer-Yahia, titled *Adaptive Test Recommendation for Mastery Learning* [3], where the authors presented several solutions for adaptive educational systems. First, the authors formalized the Adaptive Upskilling Problem (ADUP) as a multi-objective optimization problem to select a batch of tests that maximize expected performance and aptitude while minimizing the skill gap for learners. Second, a heuristic algorithm based on Hill Climbing was developed to find a subset of Pareto solutions for the ADUP Problem. This algorithm helps in selecting the most appropriate tests for learners to improve their skills efficiently. Finally, the authors used simulations to compare different variants of the problem and confirmed the effectiveness of the proposed approach in achieving skill mastery [9].

Next, Ruben Mayer discussed *pTA: An Automated Teaching Assistant for Lab Courses* [11]. Lab courses allow for active learning, which is more effective than learning passively from lectures. However, the workload associated with this is much higher too. At the Technical University of Munich, they developed pTA,

which can evaluate students' solutions to a Cloud Database course, allowing for instant feedback to the students. Previously, the course was at maximum capacity, but pTA created space for more students due to the lowered teaching load. It also helped by generating a grade report for the submissions and allowed students to work more independently by studying the logs. In the later discussion, Ruben elaborated on some of the avenues for future work, such as connecting to other universities to test the platform and improving error messages within the system.

Finally, Sophia Yang discussed work on *Mining SQL Problem Solving Patterns using Advanced Sequence Processing Algorithms*[15]. This is a follow-up work on their DataEd'22 paper, where they found that global analysis was too time-consuming. Therefore, they decided to re-encode students' problem-solving patterns (gathered from solution scores) employing symbols. They then performed sequence compaction to shorten repeated equal patterns. They found that some patterns represent code testing, and others represent incorrect thought patterns. The encoding also allows one to easily extract questions with a single answer, which form suspicious behavior. Overall, the encoding of student behavior with symbols can easily be adopted by other teachers to make students' learning progress more actionable.

Overall, this iteration of DataEd provided a new set of tools to support teaching and assessment.

2.3 DataEd for Non-Computing Students

The last major theme concerned non-computing students; students majoring in something other than computer science or computer engineering. This student population is increasingly interested in data science, as it can provide them with tools to work more effectively with data. For teachers, it is important to distinguish between majoring and non-majoring students, as their context is different. As such, teachers likely need to approach these students differently.

The first talk highlighting this was presented by Erik Golen, titled *Offering Data Science Education to Non-Computing Majors* [7]. The central theme in this work is how traditional (data science) courses are not suitable for non-computing majors. Erik and his colleagues suggest taking a hands-on, low-code approach. However, they also find that, for low-code courses, some problems in course design include the definition of achieving desirable learning outcomes, solving problems in varying domains, and creating accessible lab environments. The course that they designed to work around these problems involved using datasets from the students' own domain, such as film, communication, political science, and history. The full course design is described in the paper and was well received by the students.

A set of web-based visualization tools was presented by Sean Kross for the paper *Teaching Data Science by Visualizing Data Table Transformations: Pandas Tutor for Python, Tidy Data Tutor for R, and SQL Tutor* [6]. The work is motivated by the observation that enrollments in introductory data classes in many universities have increased due to interest from across disciplines and that it is hard for students to understand individual code statements and the semantic differences between data languages. The presented set of tools[2] support this for introductory data courses by rendering step-by-step diagrams, from input to output, of data table transformations such as filtering, sorting, reshaping, pivoting, grouping, and joining, expressed in Python, R, and SQL. Behind the tools is a table visualization library that illustrates the relationships between rows, columns, and cells of operations' input and output tables. At the time of the workshop, deployment was limited, and so we cannot report on students' experiences, but instructor interest has been high.

Discussion within this topic moved to the influence of programming experience on performance. Anecdotally for the DataEd community, students without experience seem to regularly perform better. Some attendees speculate that this might be due to students with experience being complacent, and students without the experience being aware they start off 'behind'.

3. CLOSING DISCUSSION AND EMERGING RESEARCH DIRECTIONS

DataEd concluded with a discussion of recurring topics gathered throughout the day. The topics and open questions included:

- There is a wealth of theory in psychology and education that we can borrow from and build better educational materials and tools with.
- How do we teach some of the following topics: conceptual modeling, (de)normalization, NoSQL approaches, embedded SQL, and advanced SQL keywords such as JOIN and GROUP BY.
- What topics belong in an introductory data systems course?
- How do tools like Spark and Flink fit into Data Systems Education?
- What is the role of LLMs in Data Systems Education?

These topics provide new challenges and research directions for the field of Data Systems Education.

[2] `https://tidydatatutor.com`, `https://pandastutor.com/`, and `https://cudbg.github.io/sqltutor/`

4. CONCLUSIONS AND IMPLICATIONS

DataEd'23 was another highly successful event, with a good number of submissions, interesting talks, and high attendance. This document describes some of the most pressing topics in Data Systems Education, which researchers in the area can use to further the field.

The research and findings from this second edition of DataEd have various implications, both for industry and educators. For *industry*, the usability of off-the-shelf products could be improved by increasing the level of feedback available to users on inner workings. Some examples include the explanation of query plans and the development of more accessible error messages. For *educators*, the increase of interest in data science from students of different backgrounds leads to challenges in teaching. These students have different backgrounds, and as such may benefit from a different approach to teaching, with increased scaffolding and the use of more engaging data. The tools discussed in subsection 2.2 may help educators by reducing the teaching load.

5. REFERENCES

[1] Efthimia Aivaloglou, George Fletcher, Michael Liut, and Daphne Miedema. Report on the First International Workshop on Data Systems Education (DataEd'22). *ACM SIGMOD Record*, 51(4):49–53, 2023.

[2] Ridha Alkhabaz, Zepei Li, Sophia Yang, and Abdussalam Alawini. Student's learning challenges with relational, document, and graph query languages. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 30–36, New York, NY, USA, 2023. Association for Computing Machinery.

[3] Nassim Bouarour, Idir Benouaret, Cédric D'Ham, and Sihem Amer-Yahia. Adaptive test recommendation for mastery learning. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 18–23, New York, NY, USA, 2023. Association for Computing Machinery.

[4] Peng Chen, Hui Li, Sourav S. Bhowmick, Shafiq R. Joty, and Weiguo Wang. LANTERN: Boredom-conscious natural language description generation of query execution plans for database education. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 2413–2416, New York, NY, USA, 2022. Association for Computing Machinery.

[5] Thomas Hütter and Daniel Kocher. Feedforward-aided course designs for similarity search. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 14–17, New York, NY, USA, 2023. Association for Computing Machinery.

[6] Sam Lau, Sean Kross, Eugene Wu, and Philip J. Guo. Teaching data science by visualizing data table transformations: Pandas tutor for Python, tidy data tutor for R, and SQL Tutor. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 50–55, New York, NY, USA, 2023. Association for Computing Machinery.

[7] Xumin Liu, Erik Golen, Rajendra K. Raj, and Kimberly Fluet. Offering data science coursework to non-computing majors. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 44–49, New York, NY, USA, 2023. Association for Computing Machinery.

[8] Michael J. Mior. Relational playground: Teaching the duality of relational algebra and SQL. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 56–58, New York, NY, USA, 2023. Association for Computing Machinery.

[9] Radek Pelánek and Jiří Řihák. Experimental analysis of mastery learning criteria. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 156–163, 2017.

[10] Sourav S Bhowmick. Human learners of relational query processing: Who cares? In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 1–8, New York, NY, USA, 2023. Association for Computing Machinery.

[11] Jawad Tahir, Raj Mandal, Olha Stefanova, Hans-Arno Jacobsen, Christoph Doblander, and Ruben Mayer. pTA: A programmable teaching assistant for lab courses. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 24–29, New York, NY, USA, 2023. Association for Computing Machinery.

[12] Toni Taipalus. SQL: A Trojan horse hiding a decathlon of complexities. In *Proceedings of the*

2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research, DataEd '23, page 9–13, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Jess Tan, Desmond Yeo, Rachael Neoh, Huey-Eng Chua, and Sourav S Bhowmick. MOCHA: A tool for visualizing impact of operator choices in query execution plans for database education. *Proc. VLDB Endow.*, 15(12):3602–3605, aug 2022.

[14] Hu Wang, Hui Li, Sourav S Bhowmick, and Baochao Xu. ARENA: Alternative relational query plan exploration for database education. SIGMOD '23, page 107–110, New York, NY, USA, 2023. Association for Computing Machinery.

[15] Sophia Yang, Geoffrey L. Herman, and Abdussalam Alawini. Mining SQL problem solving patterns using advanced sequence processing algorithms. In *Proceedings of the 2nd International Workshop on Data Systems Education: Bridging Education Practice with Education Research*, DataEd '23, page 37–43, New York, NY, USA, 2023. Association for Computing Machinery.