Curriculum analysis for data systems education.

MIEDEMA, D., TAIPALUS, T., AJANOVSKI, V.V., ALAWINI, A., GOODFELLOW, M., LIUT, M., PELTSVERGER, S. and YOUNG, T.

2024

Curriculum Analysis for Data Systems Education

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

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

Vangel V. Ajanovski
ajanovski@gmail.com
Ss. Cyril and Methodius University
Skopje, North Macedonia

Abdussalam Alawini
alawini@illinois.edu
University of Illinois
Urbana-Champaign
Urbana, IL, Champaign

Martin Goodfellow
martin.h.goodfellow@strath.ac.uk
University of Strathclyde
Glasgow, Scotland

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

Svetlana Peltsverger
speltsve@kennesaw.edu
Kennesaw State University
Marietta, GA, USA

Tiffany Young
t.young3@rgu.ac.uk
Robert Gordon University
Aberdeen, Scotland

ABSTRACT

The field of data systems has seen quick advances due to the popularization of data science, machine learning, and real-time analytics. In industry contexts, system features such as recommendation systems, chatbots and reverse image search require efficient infrastructure and data management solutions. Due to recent advances, it remains unclear *(i)* which topics are recommended to be included in data systems studies in higher education, *(ii)* which topics are a part of data systems courses and how they are taught, and *(iii)* which data-related skills are valued for roles such as software developers, data engineers, and data scientists. This working group aims to answer these points to explain the state of data systems education today and to uncover knowledge gaps and possible discrepancies between recommendations, course implementations, and industry needs. We expect the results to be applicable in tailoring various data systems courses to better cater to the needs of industry, and for teachers to share best practices.

CCS CONCEPTS

• **Applied computing** → **Education**; • **Information systems** → **Data management systems**; • **Social and professional topics** → *Computing industry*; *Model curricula.*

KEYWORDS

data systems, education, database, curriculum, industry, knowledge gap, skill set, student

*Working group leader.

ITiCSE 2024, July 8–10, 2024, Milan, Italy

1 BACKGROUND AND MOTIVATION

Data systems education has always been part of various information technology curricula in higher education. During the last years, however, interest in data systems has been re-invigorated by industry needs of well-trained and re-trained data engineers, data scientists and business analysts, all of whom spend much of their efforts and time working closely with data. Tool support is not sparse: new tools, languages, paradigms, and environments for manipulating data have emerged [12, 15], and knowledge of old and traditional environments, such as relational databases, are increasingly relevant requirements for today's data professionals [2]. Additionally, these professions are often closely related to other fields that rely on utilizing data efficiently [3].

What we know today about data systems education mostly comes from the databases research community. Several tools have been proposed to assist in databases courses by supporting students with tasks such as database query formulation [8] and database design [14]. Many researchers have focused on the difficulties that SQL poses to students by examining the types of errors in the queries they have created [1, 9, 16]. Valuable insights have been gained about what can go wrong without, however, knowing exactly why.

It is increasingly crucial for data systems education research to get input from the computing education community, as well as from industry professionals working with data systems. We need to evaluate teaching practices, understand student's mental models, and understand the difficulties that educators and learners face. More importantly, we need to explore interventions, instructional approaches and specific teaching practices by looking through the lens of education research and utilizing the rich toolbox that we have from programming education research. In order to focus on

the most important aspects first, we also need to understand the relationship between curriculum recommendations for data systems, which recommendations are actually implemented in higher education courses, and what are the topics that data systems professionals value the most in their work.

2 GOALS AND DELIVERABLES

The primary objectives of the working group are to *(i)* understand the relationship between higher education data systems curriculum recommendations [e.g., 5, 6, 10, 11] and implemented courses, i.e., do current syllabi follow curriculum recommendations and to what extent, and *(ii)* understand the data systems skills professional value the most and how they differ from curriculum recommendations and course contents. By understanding the current state of data systems education, we aim to explore the gaps in our current knowledge and to form research directions that hold promise towards educating students from computer science and other majors in data systems. Specifically, the working group will investigate the curricula of data systems courses regarding their topics, instruction methods, and assessment types. The working group aims to deliver three contributions:

Part 1: A synthesis of different data systems curriculum recommendations to be used in the subsequent parts of the study. The synthesis highlights core data systems education topics as recommended by various guidelines in different fields that can be interpreted as data systems, or fields that subsume data systems. The synthesis can be used in assessing data system course suitability for different majors such as those of computer science, software engineering, information systems, and data science.

Part 2: A spreadsheet containing syllabus analysis as inspired by Cunningham et al.'s study [4], and a set of syllabi from data systems educators. The spreadsheet can be used to compare how much course topics align with curriculum recommendations, and which educational approaches and assessment methods are used.

Part 3: A comparison between the possible agreements and disagreements between curriculum recommendations, course topics, and the opinions of industry professionals. We expect the comparison to highlight the knowledge gaps of new data-oriented information technology professionals, similarly to, e.g., Garousi et al. [7]. These results may be utilized in tailoring data systems courses to provide higher education students with skills needed in data-related jobs. We hope that these insights both help future graduated students to have a data-related skill set that is valued by their future employers, and that employers require less time in familiarizing new employees.

3 PRACTICAL IMPLICATIONS

The results potentially yielded by this working group may shape curriculum guidelines and syllabi to better meet industry needs of various data systems related positions. Such needs may be related to technical topics such as deep theoretical understanding concerning algorithms, statistics, or data structures, as well as technical tools such as those used in data engineering tasks. While many of such tools and even wider topics change at a rapid pace in industry, we are not aware of any empirical scientific efforts of mapping the discrepancies between guidelines, syllabi, and industry needs. Perhaps the most closely related report is that of Raj et al. [13], which focused on the technical tools needed and syllabi provided for data science and -engineer students, concluding that textbooks and curricula have not caught up with the needs of industry.

Adding more contents to a course or more mandatory courses to a syllabus is not possible if something is not removed in turn. To this end, we hope to deliver a ranking of topics based on all deliverables of the working group including industry relevance. Through these we can provide recommendations which are based on realistic modifications of courses and course contents, rather than recommending mere additions. In addition, the ranking can be used to (re)design courses with different contact- and lecture hour requirements.

REFERENCES

[1] Alireza Ahadi, Julia Prior, Vahid Behbood, and Raymond Lister. 2015. A Quantitative Study of the Relative Difficulty for Novices of Writing Seven Different Types of SQL Queries. In *Proceedings of the 2015 ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE)*. 201–206.

[2] Stephen Cass. 2022. SQL Should Be Your Second Language. *IEEE Spectrum* 59, 10 (2022), 20–21. https://doi.org/10.1109/MSPEC.2022.9915547

[3] Filippo Chiarello, Paola Belingheri, and Gualtiero Fantoni. 2021. Data science for engineering design: State of the art and future directions. *Computers in Industry* 129 (2021), 103447. https://doi.org/10.1016/j.compind.2021.103447

[4] Kathryn Cunningham, Miranda C. Parker, and Jonathan Zhang. 2023. The Landscape of Computer Science Education Courses: A Syllabi Analysis. In *Koli '23: Koli Calling*.

[5] ACM Data Science Task Force. 2021. Computing Competencies for Undergraduate Data Science Curricula.

[6] The Joint ACM/AIS IS2020 Task Force. 2020. A Competency Model for Undergraduate Programs in Information Systems.

[7] Vahid Garousi, Gorkem Giray, and Eray Tuzun. 2019. Understanding the Knowledge Gaps of Software Engineers: An Empirical Analysis Based on SWEBOK. *ACM Transactions on Computing Education* 20, 1, Article 3 (nov 2019), 33 pages. https://doi.org/10.1145/3360497

[8] Daphne Miedema and George Fletcher. 2021. SQLVis: Visual Query Representations for Supporting SQL Learners. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*.

[9] Daphne Miedema, George Fletcher, and Efthimia Aivaloglou. 2022. So many brackets! An analysis of how SQL learners (mis) manage complexity during query formulation. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*. 122–132.

[10] Joint Task Force on Computer Engineering Curricula. 2016. Curriculum Guidelines for Undergraduate Degree Programs in Computer Engineering.

[11] Joint Task Force on Computer Science Curricula. 2013. Curriculum Guidelines for Undergraduate Degree Programs in Computer Science.

[12] Bo Pang, Erik Nijkamp, and Ying Nian Wu. 2020. Deep Learning With TensorFlow: A Review. *Journal of Educational and Behavioral Statistics* 45, 2 (2020), 227–248. https://doi.org/10.3102/1076998619872761

[13] Rajendra K. Raj, Allen Parrish, John Impagliazzo, Carol J. Romanowski, Sherif G. Aly, Casey C. Bennett, Karen C. Davis, Andrew McGettrick, Teresa Susana Mendes Pereira, and Lovisa Sundin. 2019. An Empirical Approach to Understanding Data Science and Engineering Education. In *Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education* (Aberdeen, Scotland Uk) *(ITiCSE-WGR '19)*. Association for Computing Machinery, New York, NY, USA, 73–87. https://doi.org/10.1145/3344429.3372503

[14] Christos Stefanidis and Georgia Koloniari. 2016. An interactive tool for teaching and learning database normalization. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics* (Patras, Greece) *(PCI '16)*. Association for Computing Machinery, New York, NY, USA, Article 18, 4 pages. https://doi.org/10.1145/3003733.3003790

[15] Toni Taipalus. 2024. Vector database management systems: Fundamental concepts, use-cases, and current challenges. *Cognitive Systems Research* 85 (2024), 101216. https://doi.org/10.1016/j.cogsys.2024.101216

[16] Toni Taipalus, Mikko Siponen, and Tero Vartiainen. 2018. Errors and Complications in SQL Query Formulation. *ACM Transactions on Computing Education* 18, 3 (2018).