# Data systems education: curriculum recommendations, course syllabi, and industry needs.

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

2024

# Data Systems Education: Curriculum Recommendations, Course Syllabi, and Industry Needs


Daphne Miedema*
d.e.miedema@uva.nl
University of Amsterdam
Amsterdam, the Netherlands
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, USA

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

Data systems have been an important part of computing curricula for decades, and an integral part of data-focused industry roles such as software developers, data engineers, and data scientists. However, the field of data systems encompasses a large number of topics ranging from data manipulation and database distribution to creating data pipelines and data analytics solutions. Due to the slow nature of curriculum development, it remains unclear *(i)* which data systems topics are recommended across diverse higher education curriculum guidelines, *(ii)* which topics are taught in higher education data systems courses, and *(iii)* which data systems topics are actually valued in data-focused industry roles. In this study, we analyzed computing curriculum guidelines, course contents, and industry needs regarding data systems to uncover discrepancies between them. Our results show, for example, that topics such as data visualization, data warehousing, and semi-structured data models are valued in industry, yet seldom taught in courses. This work allows professionals to further align curriculum guidelines, higher education, and data systems industry to better prepare students for their working life by focusing on relevant skills in data systems education.



## CCS Concepts

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


*Working group leader.




*ITiCSE-WGR 2024, Milan, Italy*



ACM ISBN 979-8-4007-1208-1/24/07
https://doi.org/10.1145/3689187.3709609

## Keywords

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

**ACM Reference Format:**
Daphne Miedema, Toni Taipalus, Vangel V. Ajanovski, Abdussalam Alawini, Martin Goodfellow, Michael Liut, Svetlana Peltsverger, and Tiffany Young. 2025. Data Systems Education: Curriculum Recommendations, Course Syllabi, and Industry Needs. In *2024 Working Group Reports on Innovation and Technology in Computer Science Education (ITiCSE-WGR 2024), July 8–10, 2024, Milan, Italy.* ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3689187.3709609


## 1 Introduction

Data Systems Education has long been recognized as an important component of various information technology programs in higher education [45, 88, 121]. In recent years, the industry's need for well-trained and re-trained data engineers, data scientists, and business analysts has reignited growing interest in this field. There is no shortage of tool support, with new tools [78], languages [46], and paradigms [114] for manipulating data emerging constantly. Furthermore, knowledge of traditional environments such as relational databases, remains highly relevant for data professionals [22]. These data-related roles are also often intertwined with other fields that depend on the efficient utilization of data [24].

Computing curricula typically guide the education of future data professionals and are often based on guidelines published by various organizations such as ACM, IEEE, and AIS. Although the field of information technology in general advances relatively rapidly, these guidelines are often released with relatively long cycles, for example, computer science curriculum recommendations for 2013 in December 2013 [88] and, for 2023 in June 2024 [64]. Additionally, it arguably takes time to renew curricula and syllabi based on new guidelines. Nevertheless, it is unclear to what extent the guidelines are utilized in higher education, how fast syllabi are adjusted based

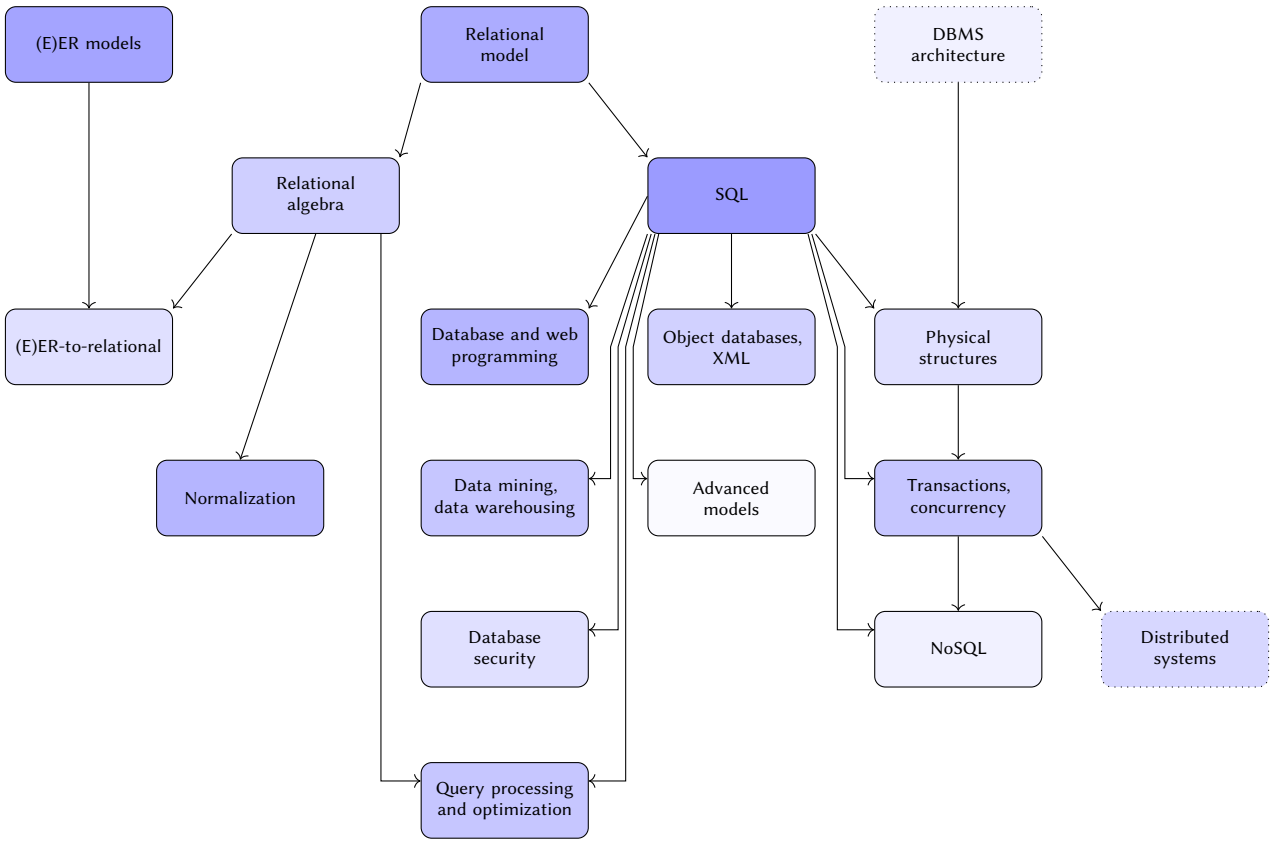


**Figure 1: Textbook topics, modified from *Fundamentals of Database Systems* [40], where arrows represent which topics are prerequisite for other topics; the darker colors represent to which extent topics are covered in twelve other data systems textbooks [27, 32, 37, 47, 55, 56, 61, 63, 96, 99, 130, 131]; dotted rectangles represent recurring topics which are not covered in *Fundamentals of Database Systems* itself.**

on the guidelines, and how much educators account for industry needs in data systems education.

To address the possible discrepancies between curriculum guidelines, curricula, and industry needs, and to understand how data systems educators design their courses, we set out to analyze curriculum recommendations, as well as survey expert opinions of data systems educators and practitioners. The research questions for this study are as follows.

**RQ1**: Which topics do curriculum guidelines recommend to be included in data systems education at the university level?
**RQ2**: Which topics are a part of data systems courses and how they are taught?
**RQ3**: What are the motivations behind educators' choices for syllabus design in data systems courses?
**RQ4**: Which data-related skills are valued for industry roles such as software developers, data engineers, and data scientists?
**RQ5**: To what extent does data systems education in its current form conform to curriculum guidelines and industry preferences?

All the studies presented in this paper to answer the research questions above were approved by the IRBs of the working group leaders' institutions.

The results indicate, for example, that data visualization is often not taught in the graduate or undergraduate curriculum at all. Database normalization, on the other hand, is taught at various levels of education. Furthermore, the industry survey and job advertisement analysis reveal a consensus on the importance of foundational skills like SQL and data modeling for database roles, while highlighting a gap between the evolving market demand for advanced skills such as database scalability and cloud computing and the current priorities of industry professionals. Finally, industry in general seems to be in need of more in-depth data system skills from new job seekers, emphasizing the need for more education on topics such as semi-structured data models, data mining, data privacy and ethics, data warehousing, and data visualization.

The remainder of this paper is structured as follows. In the next section, we discuss educational research on data systems topics, data systems curricula and the role of data systems in various computing curricula, and data-focused industry roles. In Sections 3, 4, and 5, we discuss the methods and results of our curricula analysis, and educator and industry surveys, respectively. Section 6 contains

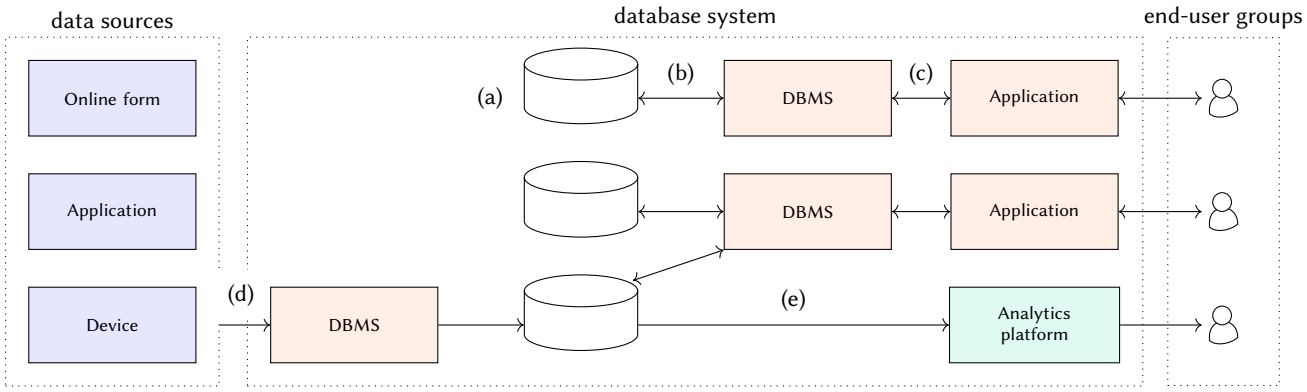


**Figure 2: A general example of data systems components and their interrelationships; arrows represent the flow of data; each of the components and relationships represents one or several data systems topics, for example, the databases depicted in the middle are closely related to conceptual, logical and physical database design**

the comparisons of topics recommended in the aforementioned sections, as well as our recommendations and the limitations of this study. Section 7 concludes the study.

## 2 Background

*Data systems* can be considered a broad topic that encompasses data, databases (DB), database systems (DBS), database management systems (DBMS), and many closely connected concepts such as data engineering, data science, and database programming. Elmasri and Navathe [40] mapped these topics and their dependencies as taught in 13 textbooks, an adapted version can be seen in Figure 1. As the differences in color depth indicate, not all topics all covered to the same extent. As such, the exact definition of data systems varies from researcher to researcher.

### 2.1 Data Systems Education Research

Educational research in data systems has been relatively scarce when compared to other research topics such as programming education. In this section, we describe educational research on data systems topics, divided into subsections common in data systems textbooks. Topics of this subsection are summarized in Fig. 2.

*2.1.1 Conceptual modeling.* Conceptual modeling is used in translating relevant domain needs into conceptual data structures (Fig. 2a), typically in the initial stages of system design. The most popular notations are arguably the Entity-Relationship (ER) model [23] with its extensions such as the similarly named Extended ER (EER) models, as well as the Unified Modeling Language (UML) [100].

Many studies agree that conceptual data modeling is a challenging topic to learn, perhaps due to the open-ended nature of conceptual modeling [110], or the existence and shortcomings of different notations [25]. In order to understand the challenges, some studies have tried to identify the common problems in conceptual modeling by categorizing student mistakes [19, 57, 97]. To facilitate the learning of conceptual modeling, scholars have suggested using a design pattern approach similar to object-oriented programming [132] or using concept maps to reduce cognitive load in modeling large domains [93, 107].

Several studies have proposed tools for learning conceptual modeling of databases, for example, `ERM-VLE` [51], `COLER` [35], `MonstER Park` [103], `#EER` [20], `ERDoc Playground` [69], and `KERMIT` [110], as well as some unnamed tools [39, 133]. While some of these tools aim to foster deeper learning or student collaboration, others aim to ease teacher workload by automatically checking the conceptual models for correctness. Many studies that introduced a tool for conceptual modeling have also evaluated the tool, usually with positive results when compared to learning without such a tool.

*2.1.2 Database design.* Database design consists typically of designing both logical and physical database structures (Fig. 2a) to serve the needs of the end-users, often via a user-facing software system. The few studies on database design education have suggested tying database design tasks to real-world scenarios with project-based learning and a constructivist approach [28, 86], gamification [38], and integrating database design as a part of a more holistic software project [62]. Recently, studies have also investigated how *teachers* should design databases that are engaging for students who are learning querying [81, 117]. In summary, educational insights on effective teaching methods, common student mistakes, or supporting tools are scarce, and almost exclusively focused on *logical* design.

*2.1.3 Query languages.* Query languages are used for retrieving and manipulating data in the database (Fig 2b). Research on query language education has probably been the most prominent theme in data systems education research. Expectedly, most of the work has focused on SQL [118], but there have also been studies on the education of *Query by Example* [58], and different NoSQL query languages [10, 12, 66], as well as the theoretical foundations behind relational query languages [48, 52].

Perhaps the most widely studied theme in educational research on query languages has been the errors committed in query formulation. Studies have explored which errors are the most prominent [7, 119, 134], why errors occur [80, 106, 108, 112], and what affects the occurrence of different errors [98, 111]. Additionally, helping novices in query writing has been a prominent research topic. Several scholars have worked on tools and techniques for visualizing

queries to either facilitate query formulation or to help novices understand more complex queries [31, 65, 70, 79]. Furthermore, there have been efforts in understanding the role of enhanced SQL error messages in education [115, 116], develop more intuitive methods of expressing complex query concepts [75], and the gamification of SQL education [13, 84, 102]. Recently, the integration of language model agents has also been suggested to support students in query writing tasks [92, 94], and also for assisting in creating assignments for students [6].

*2.1.4 Query processing.* Query processing and optimization is a common topic in data systems education, especially in advanced courses (Fig. 2b). The aim is to understand how database queries operate, often through physical database operations and structures such as bitmap indices, hash joins and nested loops, and sequential scans in order to write efficient queries.

Educational research on query processing has seen several recent efforts. Some studies have outlined what topics should be included in a query optimization [33] course, and others have highlighted the opportunities in considering novices when formulating query execution plans [113]. The bulk of educational research on query optimization has proposed tools for facilitating novice understanding of query execution plans in different ways. For example, RelaX[1] and `MOCHA` [120] visualize the physical operations and intermediate relations of SQL queries. `NEURON` [68] and `LANTERN` [129] simplify the interpretation of query execution plans by translating them into simplified natural language. `ARENA` [128] presents students with alternative query execution plans for queries, which aims to help students understand how the selection of different physical operations affects query efficiency.

*2.1.5 Database programming.* Database programming (Fig. 2c) refers to connecting the DBMS to the software application by either embedding query languages into the host language, or by leveraging host language or third-party libraries for retrieving and manipulating the database. Additionally, some viewpoints consider using SQL's procedural extensions such as T-SQL or PL/SQL to create user-defined functions in the DBMS, which are then called from the host language.

The papers in this area have usually been experience reports on course contents and the teachers' expert opinions on how their approaches worked in the classroom. One study [71] provided a general description of how to incorporate database programming into web programming using PHP and a data abstraction layer for querying and manipulating the database. Another study [91] applied modern software engineering tools and frameworks in supporting the use of database programming with Java. Some studies have discussed database security education in MongoDB [50], and proposed gamification in teaching SQL injection [104, 109]. Finally, two studies [95, 127] have discussed database programming as teaching PL/SQL, yet focused on high-level teaching philosophies rather than understanding the *database programming* aspect specifically. It seems that basic topics such as connecting the DBMS to the rest of the software system, as well as important topics such as object-relational mappers and object-relational impedance mismatch have not received scientific attention from the perspective of educational research.

*2.1.6 Data analytics.* Data analytics is a general term encompassing concepts such as descriptive, prescriptive, and inferential analytics, all of which typically aim at producing business value from collected data (Fig. 2e). Tangential terms are data mining and data science. The analysis of data is typically preceded by extracting, cleaning, and transforming data from various sources [26] (Fig. 2d) into a database where it is feasible to perform the analyses.

Educational research on data analytics has been scarce and diverse. Tangential studies have highlighted the differences between math and tool-focused data science and visualization-oriented data analytics [4]. Furthermore, various research teams have recognized that data science and analytics courses are often taught across disciplines by educators who do not have formal knowledge on the topic, to students who do not have CS backgrounds [72, 82]. Some studies have also proposed high-level frameworks for teaching data analytics [87].

## 2.2 Data Systems Curriculum Analyses

In the area of computing, various curriculum guidelines exist. These are specified by organizations that discuss and evaluate computing as a field of science. In this study, we were unable to identify any curriculum guidelines specific for Data Systems, yet data systems topics are a part of several computing curriculum guidelines such as computer science, software engineering, and information systems. For a full list of curricula considered in this study, see Section 3.

Computer science curricula have been analyzed in various existing works. One study developed a framework for data education [29], identifying core competencies such as data management and data analysis, as well as higher-order competencies such as ethics and governance, and inter-disciplinary competencies such as communication and critical thinking. Another studied data analytics programs and their contents, discovering that courses are often rebranded with more modern names such as data warehousing to big data [59]. Finally, one study built a data science program [101]. The DataEd workshop at SIGMOD also attracts both educators and industry professionals, leading to conversations surrounding data systems curricula [8]. Other papers on computer science curricula did not focus on data-intensive topics, but instead surveyed introductory programming course curricula [17, 73, 74, 85].

Methods of data-led curriculum analysis have been investigated previously [9, 76, 77, 105]. These methods were based on topic modeling. Finally, a syllabus analysis of computer science education courses revealed 550 educational topics extracted from syllabi and categorized them under seven different themes, including theories of thinking and learning, and technological applications [30]. Their research setup inspired this work.

## 2.3 Data-focused Industry Roles

In the rapidly evolving field of data systems, it is crucial to align educational curricula with the skills and knowledge demanded by the industry. As computing disciplines continuously change, educational programs in information technology and related fields must adapt to ensure that graduates are prepared to contribute

[1] https://dbis-uibk.github.io/relax/landing

value to enterprises. Understanding the industry's perspective helps ensure that graduates are well-prepared to meet the challenges and requirements of real-world data systems roles. This alignment not only enhances the employability of graduates but also ensures that educational programs remain relevant and up-to-date with the latest technological advancements and industry practices.

Many roles in industry are focused on data, and these roles have seen several changes in job descriptions and titles over the years [8]. From titles such as *data engineer*, *database administrator*, *database designer*, or *data analyst*, the connection to data system topics is relatively obvious. However, many *software developers* are involved in database design, query optimization, and especially database programming. In contrast, some seemingly data-focused titles such as *data scientists* may focus on tasks closer to statistics than data systems as they are understood in this study.

Incorporating industry insights into educational strategies can bridge the gap between academic learning and practical application, ultimately fostering a more competent and industry-ready workforce. The CS2023 Task Force, a reputable body involved in shaping computing curricula, surveyed 110 academics and 865 industry practitioners to identify the characteristics of computer science graduates [64]. This comprehensive study highlights the critical skills and knowledge areas that industry professionals value, providing a robust foundation for justifying the integration of industry perspectives into educational curricula.

Recognizing the importance of professional practice is essential, as most students in information technology programs will enter the workforce upon graduation [90]. This integration of industry perspectives into the curriculum supports the development of graduates who can effectively contribute to their respective fields.

Graduate employment rates underscore the importance of aligning educational curricula with industry demands. For instance, in Scotland, the graduate employment rate stands at 82% [53], while in England, it is slightly higher at 87.7% [43]. The European Union reports an employment rate for recent graduates of 83.5% in 2023 [41], and Australia boasts an even higher rate, with 88.3% of undergraduate graduates in employment in 2022[42]. These statistics highlight the need for educational programs to adapt and evolve continuously to ensure that graduates are well-equipped to enter the labor market successfully. By integrating industry perspectives into curricula, educational institutions can better prepare their students for the workforce, contributing to higher employment rates and more successful career outcomes for graduates.

# 3 Curriculum Guidelines

In this research effort, we try to understand how data systems courses are taught throughout the world and how they meet industry needs and expectations. International standards, or in the absence of any, widely used guidelines from international professional organizations can give us a point of reference (or validity, or even a coordinate system) upon which we can measure and compare. In this section, we investigate **RQ1**: Which topics do curriculum guidelines recommend to be included in data systems education at the university level?

We aim to uncover both the commonalities among these guidelines as well as their differences.

## 3.1 Methodology

Our data collection efforts followed a two-part approach. Initially, our working group conducted a comprehensive web search to identify data curricula related to Computer Science, Data Science, Information Science, and other pertinent subjects. Secondly, we leveraged the authors' international network of data systems educators to uncover guidelines that may not be readily accessible online. Our efforts resulted in identifying 19 guidelines from 12 different countries and global organizations, including ACM, IEEE, and AIS.

To facilitate the analysis of these guideline documents, we first translated all non-English documents into English. Subsequently, we conducted a detailed examination of the curriculum recommendations to extract vital information pertinent to data systems curricula. This included identifying the responsible agency, document title, educational level (whether undergraduate or graduate), target program (such as data science), and providing a link to the original document.

In the second phase, we aimed to identify prevalent educational topics within data systems as referenced in these guidelines. We extracted topics associated with data systems from each of the global and national guidelines we collected and categorized them into relevant groups. We used this categorization to perform a structured analysis of the common themes and areas of focus.

Below, we provide an overview of these guidelines and in Section 3.4, we present the result of the analysis we conducted on this dataset.

## 3.2 Overview: Global Computing Organizations

The landscape of data systems education is shaped by curriculum guidelines and recommendations developed by various global organizations, such as ACM, AIS, and IEEE, as well as national education bodies from several countries. The main goal of these recommendations is to ensure that educational programs remain relevant and aligned with the fast-evolving demands of industry and national strategic plans. Here, we provide an overview of these guidelines, highlighting their key components and recommendations. We list the examined curriculum guidelines in reverse chronological order (i.e. from newest to oldest).

*3.2.1 Computer Science Curricula 2023 [CS2023].* The ACM, IEEE, and the Association for Advancement of Artificial Intelligence collaborated on developing updates to the guidelines for the computer science discipline from 2013 [88]. The latest guidelines [64] have changed the names of knowledge areas on information management to data management and intelligent systems to artificial intelligence, mainly to avoid confusion or to reflect more recent nomenclature. The guidelines propose adopting learning approaches that stress practical applications and interdisciplinary projects. They also recommend a continuous process for updating curricula to keep up with advancements in the AI field. Our main interest is in the data management area, but we also investigated the other areas having intersections with data management and AI fundamentals, data management, and machine learning in the *core*, and advanced elective topics, such as deep learning, big data analytics, and AI ethics.

*3.2.2 Computing Competencies for Undergraduate Data Science Curricula [CCDS2021].* An ACM task force developed curriculum

guidelines for the emerging data science field [44]. As data science is inherently an interdisciplinary field, such are the guidelines, which include knowledge areas from analysis and presentation, artificial intelligence, data management, machine learning, and software development – aiming to cover everything a data-science project would encompass and everything a data-science specialist should be aware of. Most of these topics are relevant to our pursuit and we investigated several in more detail.

*3.2.3 IS2020 - A Competency Model for Undergraduate Programs in Information Systems [IS2020].* The ACM and the Association of Information Systems jointly created guidelines and recommendations targeting the information systems discipline [45]. Here there is also a focus on information and data management, and business analytics. Key components of the past guidelines [123] focused on core courses in the IS discipline, but in the latest edition, the perspective has changed in favor of competence areas. Our focus is mainly on the data competence realm and the data and information management competence areas, which also include data mining, business analytics, and data visualization. The guidelines incorporate an *interdisciplinary stance* to bridge technical skills with business needs. The guidelines also recommend collaboration between industry and academia to ensure that educational programs meet workforce demands.

*3.2.4 Information Technology Curricula 2017 [IT2017].* These guidelines were prepared by a task force from the ACM and IEEE-CS [90]. The guidelines focus on the information technology discipline to create an update of the prior guidelines published in 2008. In the last revision, the guidelines focused on the skills and dispositions that should complement the body of knowledge, defined from the Enterprise Information Technology Body of Knowledge report developed by the IEEE-CS. Our focus in these guidelines was mainly the information management domain, and we also investigated the other domains that covered tangentially relevant topics to our interests, like the integrated systems technology domain.

*3.2.5 Cybersecurity Curricula 2017 [CSEC2017].* These guidelines were prepared by a joint task force from the ACM, IEEE-CS, AIS SIGSEC, and IFIP WG 11.8 [89], focusing on topics related to various aspects of security – ranging from component and system security to human and societal security. We have investigated all areas, looking for topics that could relate to our interest focus of data systems education.

*3.2.6 Software Engineering 2014 [SE2014].* These guidelines were prepared by a task force from the IEEE-CS and ACM and focused on the software engineering discipline [14]. These guidelines included some interconnections to our focus – education about data systems from the point of view of software development.

## 3.3 Overview: National Curriculum Guidelines and Advisories

Although we were able to identify various national curriculum guidelines concerning computer science, we must note that not all of these were mandatory for teachers to work with. In addition, some were meant for other educational levels such as primary school or high school. In the text below we provide a general analysis of these guidelines, but only the three applicable bodies of work (for the United Kingdom, Brazil and Australia) are included in our systematic results further below.

Educational bodies from several countries developed guidelines that align with their countries' strategic needs. In the Netherlands, the Stichting Leerplanontwikkeling (SLO) has developed comprehensive guidelines for data systems education, focusing on digital literacy and data competencies for primary and secondary education [16, 49, 122]. Key components of these guidelines focus on the early introduction (primary and secondary level) of fundamental knowledge in Data Management, Programming, and Data Ethics. At the higher levels (undergraduate and graduate), the guidelines recommend the introduction of advanced topics, such as Data Analytics, Machine Learning, and Data Engineering. Another advisory to secondary education comes from France, whose guidelines contain a few pages on managing data [83]. Their preferred skills include knowledge of DBMSs and the relational model, as well as building queries using various SQL clauses. The Danish highschool objectives [21] also include specific sections on representation and manipulation of data, mentioning skills such as building web systems to display data, dealing with various data types and mentions of ER diagrams and SQL.

Australia identifies a "core body of knowledge" upon which they base educational accreditation for ICT higher education degrees [15]. They focus on both core ICT knowledge, professionalism in its general form, as well as professionalism as it applies in ICT (such as the ethics involved in data storage and processing). The core ICT knowledge extensively mentions data management concepts. The guidelines published in the United Kingdom by several educational bodies and industry partners also emphasize a focus on teaching data management and analytics skills in higher education [1, 2].

Furthermore, various countries outside the global west have also established guidelines to enhance data systems education. Brazil's guidelines focus on integrating data systems topics with the Information and Communication Technologies (ICT) curriculum [34]. They emphasize the offering of Data Management, Programming, and Data Analytics as early as the secondary school level. Saudi Arabia's curricula (for example [126]) are aligned with the Vision 2030 initiative [3], emphasizing data literacy and advanced analytics. Tunisia [36] includes programming, database management, and data science applications in its curriculum, preparing students for careers in data-driven fields. The guidelines emphasize the need for continuous curriculum updates, practical experience through internships, and partnerships with industry to ensure that graduates are well-equipped to meet contemporary demands.

## 3.4 Results

To answer **RQ1**, we extracted topics associated with data systems from each of the global and national guidelines we collected and categorized them into relevant groups, such as data pipeline, data modeling, and big data. This categorization facilitated a structured analysis of common themes and areas of focus within the studied guidelines. The results of the analysis are presented in Fig. 3.

Our analysis indicates that data security and privacy is the most frequently mentioned topic, appearing in six different guidelines. It is followed by big data, data management, data visualization, data


**Figure 3: Mapping global and national computing curriculum guidelines to data systems concepts**

modeling, data analytics, and data ethics, each referenced four times. Several topics were mentioned in both global and national guidelines, including cloud computing, data mining, relational databases as well as other types of models, query formulation and computer science foundations. Additionally, a few topics were specifically highlighted in more than one global guideline: data analytics, distributed databases, cloud computing and web development.

There are also clear differences between the analyzed guidelines. Some of them have very specific core focus, such as the CCSD2021 and the data pipeline, and CSEC20217's and Security. CS2023 seems to present the most complete view of data systems education as

we identify it here, as they consider topics in all themed boxes in Figure 3. Concerning the national guidelines, the United Kingdom guidelines seem to focus on data modeling and foundational knowledge, whereas the Brazilian guidelines have a broader topic spread, seemingly targeting the more applied topics. The Australian guidelines are somewhere in between, including both theory and applied topics.

### 3.5 Discussion

The predominance of data security and privacy as the most frequently mentioned topic underscores its critical importance in the realm of data systems education across diverse educational guidelines. This finding suggests a universal recognition of the need to address security and privacy concerns as foundational elements in the curriculum for future data professionals. The repeated reference to topics like Big Data and Data Management in both global and national guidelines reflects a broad consensus on the essential skills required in data-driven industries. The inclusion of data modeling, data visualization, and data ethics further indicates a growing emphasis on not only handling data effectively but also responsibly and ethically. The distinction between the topics emphasized in global versus national guidelines could signal varying regional priorities or the influence of international standards on local educational frameworks.

## 4 Database Course Content and Teaching Practices

Now that we have learned more about what existing curriculum guidelines describe, we set out to learn whether data systems course syllabi were in line with this. To this end, we reached out to instructors who teach database, data science, and/or data systems courses internationally, inviting them to participate in a comprehensive survey. The survey aimed to gather detailed insights into database content being taught, the teaching methodologies employed, and the resources utilized in these courses. The objective was to understand current teaching practices, identify trends, and share what the greater educational community is doing.

With this data, we investigate **RQ2**: Which topics are a part of data systems courses and how they are taught? and **RQ3** What are the motivations behind educators' choices for syllabus design in data systems courses? Answering these questions and connecting them back to curriculum guidelines, allows us to enhance the quality of education in database-related fields. The data also gives insights that help courses remain relevant and aligned to the evolving needs of the industry (see Section 5).

### 4.1 Methods

We designed a survey consisting of 16 overarching questions, with sub-parts, comprised of 53 questions total (see Table 1 and Table 2). This survey was then distributed to educators in the larger Data Systems Education community[2] via email. In some cases, data systems community members were contacted via email directly, through a list of data systems educators created by the joint networks of this paper's authors, and in other cases, they were contacted through larger email lists (i.e., DBWorld, SIGCSE, and SIGITE) to ensure the larger communities were contacted.

The quantitative data and qualitative data were analyzed by two researchers. Quantitative data were extracted in the form of CSV files. The data were then cleaned, analyzed, and graphed using Python. Note that we separated the results by undergraduate and graduate program levels for reporting. Additionally, the qualitative data were read and summarized by the same two researchers, extracting all survey responses, discussing them, and then reporting the key points and significant information in the results section below.

### 4.2 Results

The survey had 105 responses from post-secondary/tertiary data systems educators, spanning 24 different countries (see Table 3). This includes a mix of both undergraduate and graduate courses (see Table 5), their program level (also in Table 5), and a count of courses based on their size range (see Table 4). Additionally, the average survey response time was approximately 19 minutes and 50 seconds.

Most of the courses we surveyed, both undergraduate and graduate, were intended for computer science majors and minors. The courses are also taken (as electives) by students from other degree programs such as engineering, information technology, data science, and finance. Furthermore, most courses appeared to be called "database systems" or some variation of this and the majority (approx. 65%) appeared to be required for the degree program (with approx. 35% being elective). The average course duration was 13.1 weeks (max: 24 weeks, min: 5 weeks), and educators claimed they had 3.27 hours per week on average of direct and meaningful two-way interaction with individual students.

When asking data systems educators about Q10 ("Is this topic covered in your course?") we explicitly looked for coverage of 38 sub-questions (Table 2), where we first identified whether the topic was "not covered", covered in a "prerequisite" course, or taught based on "Bloom's Taxonomy", and split by course types: "undergraduate" (see Figure 4) and "graduate" (see Figure 5). We then refined the figures to include what level of Bloom's taxonomy each concept is taught, based on educator responses, for the "undergraduate" (see Figure 6) and "graduate" (see Figure 7) levels of these courses.

From the qualitative data, we found that data systems educators taught the following additional topics:

- Vector databases, temporal databases and graph databases
- Database connectivity (e.g., JDBC, ODBC, or alternative connections to applications)
- Triggers, views, and temporary tables
- Data wrangling and analysis
- Algorithms (e.g., PageRank)
- Ontologies (e.g., RDF, OWL)
- Application development (incl. web)
- Reviewing and analysing research papers

[2]The "Data Systems Education" community is defined as a collective of educators, researchers, and practitioners dedicated to advancing the teaching and learning of data management, data science, and information systems, fostering interdisciplinary collaboration and sharing best practices to enhance data literacy and skills across various fields.

| ID | Question | Style | Options/(Restrictions) |
|---|---|---|---|
| Q1 | What is the level of the course? | Radio-Button | Options:<br>— Undergraduate<br>— Graduate<br>— Other (with open-textbox) |
| Q2 | In which country the course is given in? | Open-Textbox | N/A |
| Q3 | What is the title of the data systems course you teach? | Open-Textbox | N/A |
| Q4 | What year is the course? | Open-Textbox | (numbers only) |
| Q5 | What is the general course size in number of students? | Open-Textbox | (numbers only) |
| Q6 | What types of students primarily take this course (e.g., CS Major, CS Minors, other Computing Majors, Non-Computing) | Open-Textbox | N/A |
| Q7 | Is this course required or elective in your program? | Open-Textbox | N/A |
| Q8 | How many weeks is the course? | Open-Textbox | (numbers only) |
| Q9 | How many hours per week has a single student direct meaningful two-way interaction with the teacher? | Open-Textbox | (numbers only) |
| Q10 | Is this topic covered in your course? — see Table 2 for the detailed sub-topics (Q10.1–Q10.38). | | |
| Q11 | Are there any other topics covered in the course? | Open-Textbox | N/A |
| Q12 | Why did you choose these course topics? You can elaborate your answer. | Multi-Select with Open-Textbox | Options:<br>— They were dictated by other courses<br>— I applied curriculum guidelines<br>— Someone else (e.g., in the faculty) dictated them<br>— They are based on industry needs<br>— I chose them myself (i.e., based on intuition, past experiences...)<br>— I inherited the course from a predecessor<br>— They are based on a textbook<br>— Other |
| Q13 | What active learning techniques do you use in your course? | Open-Textbox | N/A |
| Q14 | What structure do you primarily use to deliver material in this course (e.g., lectures, seminars, tutorials, labs, project-based learning, etc.)? | Open-Textbox | N/A |
| Q15 | How do you assess student performance (e.g., exams, projects, assignments, peer reviews)? | Open-Textbox | N/A |
| Q16 | Do you utilize any online platforms or tools to support your teaching? If so, which ones? | Open-Textbox | N/A |

**Table 1: These are the 53 questions of the DataEd teacher survey questions, including the presentation style provided, and any relevant options. All questions were required to complete.**

Furthermore, when analyzing the responses of Q12 ("Why did you choose these course topics?", see Table 6), we observed that most of the choices revolved around ensuring alignment of the curriculum (whether as a pre-requisite course or to meet accreditation purposes), meeting industry needs (to help ensure their alumni can secure a job after graduating), or just because it was easier to teach this way (since the course had been designed by another faculty member or taken directly from a course textbook).

When analyzing Q13 ("What active learning techniques do you use in your course?"), we found the following themes:

- Group work and discussion, for example, collaborative projects, in-class group activities, pair programming, peer review, peer instruction, think-pair-share exercises.
- Interactive methods, for example, flipped classrooms, think-aloud problem solving, mini-lectures with active problem solving, concept mapping, student-led presentations and demonstrations.
- Hands-on activities, for example, practical laboratories, live coding/querying, and completion of active worksheets in lectures.
- Real-world applications, for example, project-based learning, simulations, and case studies.
- Other techniques, for example, gamification, role-playing, real-time quizzes, start-stop-continue, and snowball technique.

When analyzing Q14 ("What structure do you primarily use to deliver material in this course?"), we saw that lectures, labs, and tutorials were popular. Additionally, assignments and regular homework exercises, projects, and asynchronous learning components (including flipped classrooms) appeared to be popular among database educators. This closely relates to Q15 ("How do you assess student performance?"), which appears to include: exams (in 60% of courses), projects (in 55% of courses), assignments (in 53% of courses), quizzes (in 16% of courses), and laboratories (in 53% of courses). Other minor occurrences include tests (in 8% of courses)

| ID | Question | Style | Options/(Restrictions) |
|---|---|---|---|
| **Q10** | **Is this topic covered in your course?**— the continuation from Table 1 | | |
| Q10.1 | —relational theory: relations, tuples and attributes | Radio-Button | Options:<br>— It is not covered.<br>— It is covered on a prerequisite course.<br>It is covered with the learning outcome of:<br>—Remember,<br>—Understand,<br>—Apply,<br>—Analyze,<br>—Evaluate, or<br>—Create. |
| Q10.2 | — tuple relational calculus | | |
| Q10.3 | — relational algebra | | |
| Q10.4 | — data visualization | | |
| Q10.5 | — database optimization: indexing | | |
| Q10.6 | — database optimization: query execution plans | | |
| Q10.7 | — database optimization: query optimization | | |
| Q10.8 | — database scalability: replication | | |
| Q10.9 | — database scalability: sharding | | |
| Q10.10 | — NoSQL database management systems | | |
| Q10.11 | — logical and physical data independence | | |
| Q10.12 | — database management system components | | |
| Q10.13 | — functions and stored procedures | | |
| Q10.14 | — data modeling: conceptual modeling | | |
| Q10.15 | — data modeling: mapping conceptual models to logical models | | |
| Q10.16 | — data modeling: creating tables and columns | | |
| Q10.17 | — database normalization: functional dependency, candidate and super keys | | |
| Q10.18 | — database normalization: normal forms up to BCNF | | |
| Q10.19 | — database normalization: multivalued dependency | | |
| Q10.20 | — database normalization: join dependency | | |
| Q10.21 | — object-oriented data models | | |
| Q10.22 | — semi-structured traditional data models (e.g., XML) | | |
| Q10.23 | — SQL: select, project, join | | |
| Q10.24 | — SQL: insert, update, delete | | |
| Q10.25 | — SQL: aggregation and group by | | |
| Q10.26 | — SQL subqueries | | |
| Q10.27 | — SQL: common table expressions | | |
| Q10.28 | — transaction processing | | |
| Q10.29 | — concurrency control and isolation levels | | |
| Q10.30 | — database back-ups and recovery | | |
| Q10.31 | — distributed database management systems | | |
| Q10.32 | — data mining: algorithms | | |
| Q10.33 | — data mining: associative and sequential patterns | | |
| Q10.34 | — data mining: data cleaning | | |
| Q10.35 | — data mining: market basket analysis | | |
| Q10.36 | — data privacy and ethics | | |
| Q10.37 | — data security and database access management | | |
| Q10.38 | — data warehousing | | |

**Table 2: Continuation of DataEd teacher survey questions (Table 1), specifically all 38 of Q10's sub-topics (Q10.1–Q10.38). Some questions are grouped in green or blue highlight, for the purposes of linking the industry survey. With respect to "learning outcome", we refer to the classification seen in Bloom's Taxonomy [18].**

and participation (in 2% of courses). Interestingly, some instructors tried forms of graded peer reviews, noting that the quality was often poor and needed to be discontinued. One instructor noted that they use oral exams.

When analyzing Q15 ("Do you utilize any online platforms or tools to support your teaching?"), we observed that learning management systems (LMS, a.k.a., VLE: virtual learning environment) were popular (e.g., Moodle, Canvas, Blackboard, Brightspace, Aula, Opal, PrarieLearn, Gradescope, Exam.net, and other custom in-house eLearning platforms). Additionally, video conferencing tools (e.g., Zoom, BigBlueButton, Microsoft Teams), coding and data analysis tools (e.g., Jupyter Notebooks, Google Collab, and DBMSs – such as MySQL, DB/2, PostgreSQL, MongoDB, and Oracle Database), specialized tools and platforms (e.g., entity-relationship diagram drawing tools Draw.io, ERDPlus, ERDoc), relational algebra simulators, runestone academy, B-Tree simulators, RapidMiner, Mentimeter, quiz platforms (e.g., Quizziz, Wooclap, and Kahoot!), discussion boards (e.g., Piazza), project and collaboration tools (e.g., GitHub, Google Drive, Facebook Messenger, Discord, Slack, Microsoft Teams), and cloud platforms.

## 4.3 Discussion

This subsection discusses two of the research questions stated in Section 1. Specifically, we will address:

**RQ2**: Which topics are a part of data systems courses and how they are taught?
**RQ3**: What are the motivations behind educators' choices for syllabus design in data systems courses?

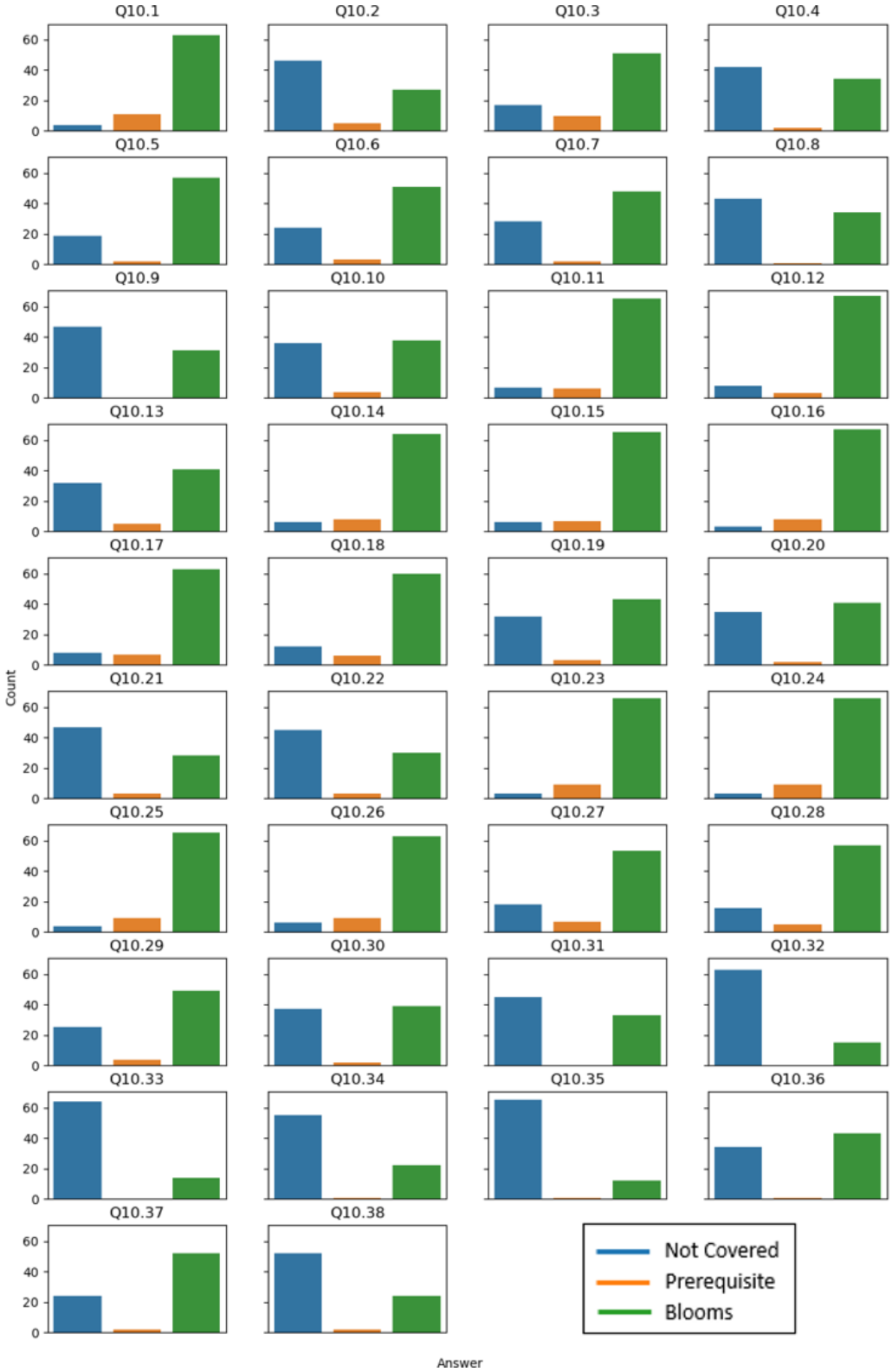


**Figure 4: Topics covered in undergraduate courses, as per Table 2.**

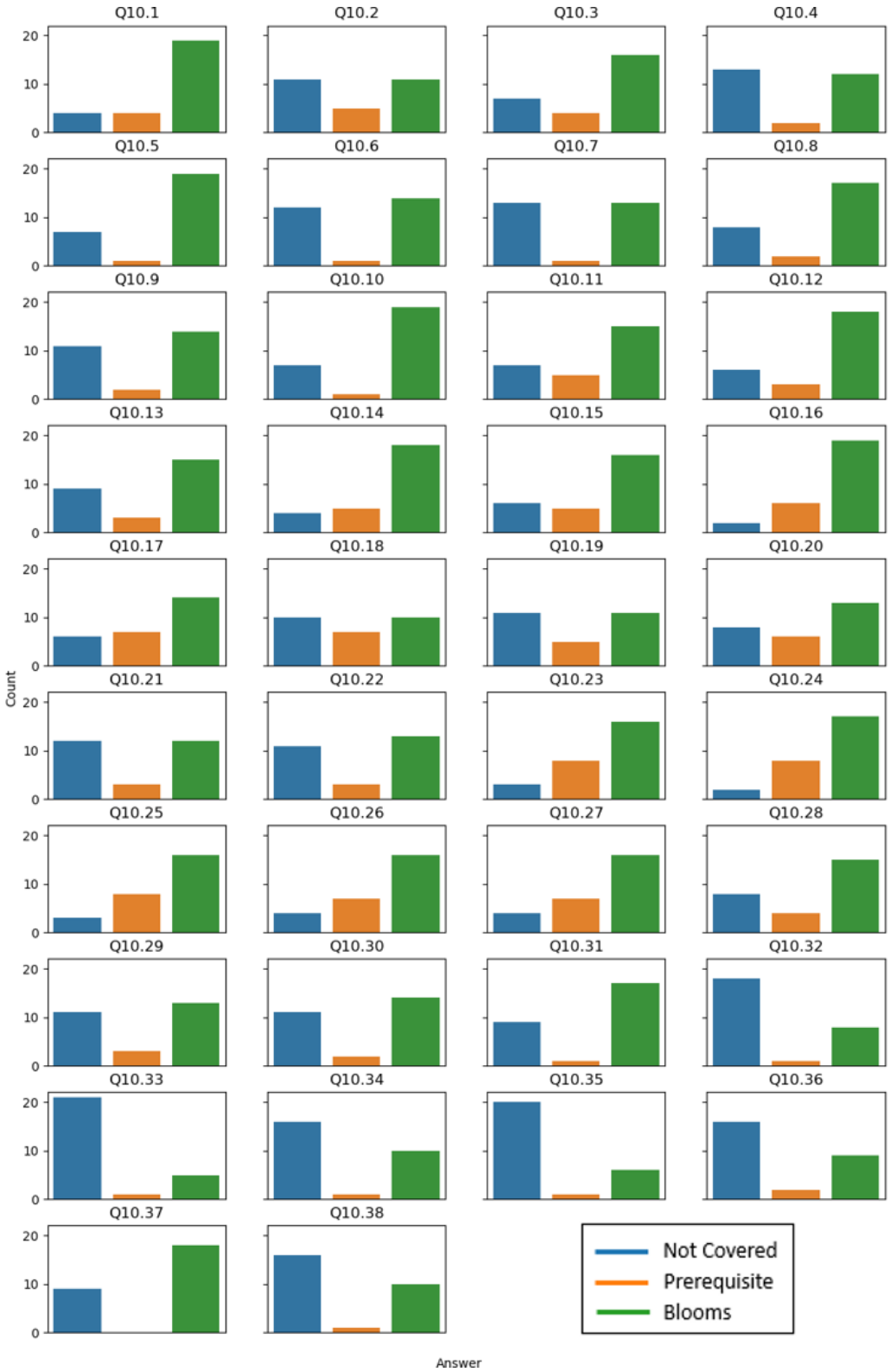


**Figure 5: Topics covered in graduate courses, as per Table 2.**

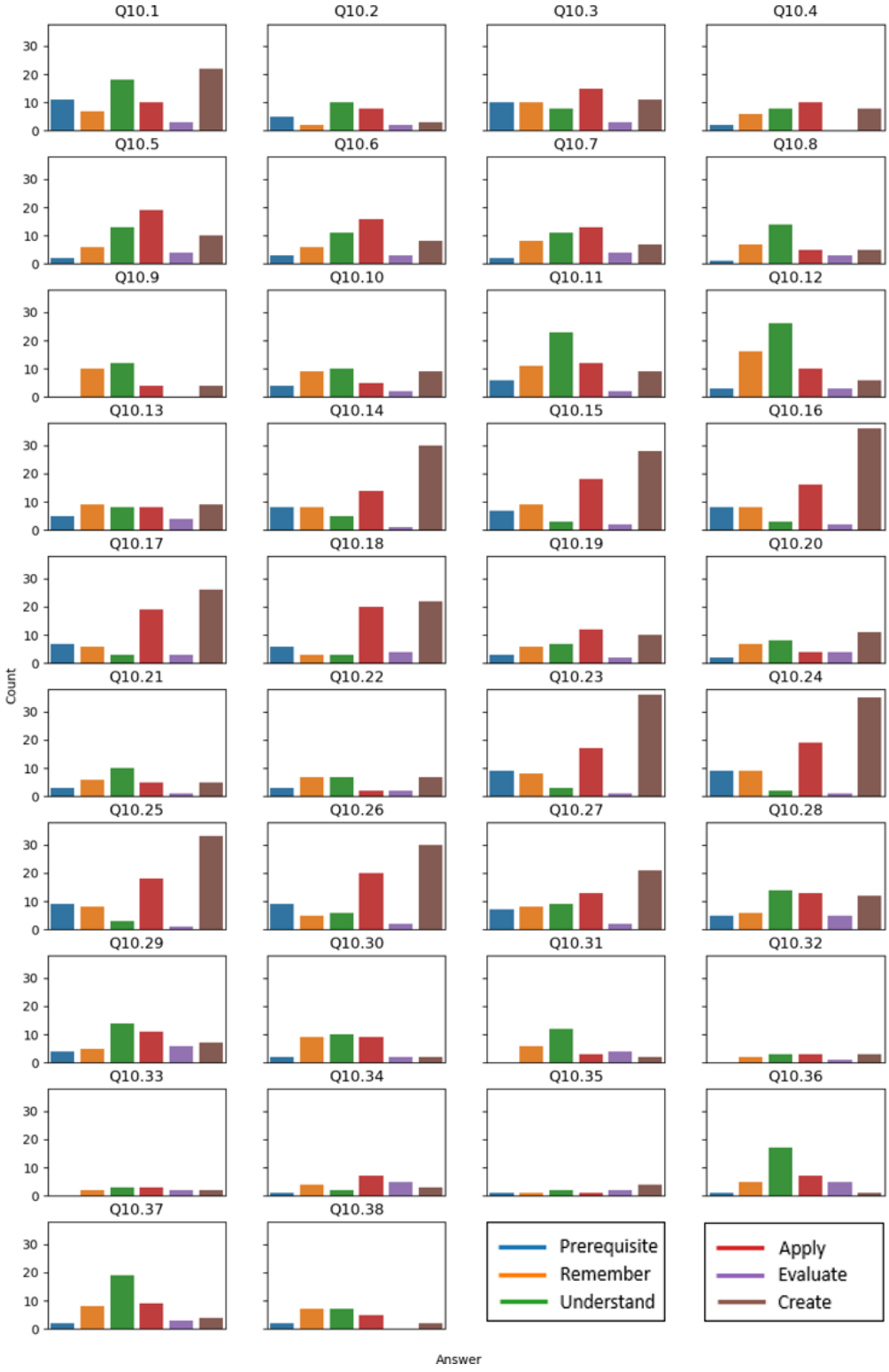


**Figure 6: Topics covered in undergraduate courses, with a refined breakdown of Bloom's taxonomy sub-components, as per Table 2.**


**Figure 7: Topics covered in graduate courses, with a refined breakdown of Bloom's taxonomy sub-components, as per Table 2.**

| Country | Count |
|---|---|
| Australia | 1 |
| Bulgaria | 2 |
| Canada | 4 |
| Chile | 1 |
| Czech Republic | 1 |
| Denmark | 1 |
| Finland | 2 |
| France | 3 |
| Germany | 7 |
| Ghana | 2 |
| India | 2 |
| Indonesia | 1 |
| Italy | 1 |
| Lebanon | 1 |
| Libya | 3 |
| Netherlands | 3 |
| New Zealand | 1 |
| North Macedonia | 2 |
| Serbia | 1 |
| Singapore | 1 |
| Slovenia | 1 |
| Sweden | 2 |
| United Kingdom | 14 |
| United States of America | 47 |
| **Total**: | **104** |

**Table 3: Count of participating DataEd educators by country. One omitted due to entry error.**

| Number of Students | Count | Average |
|---|---|---|
| 0 - 100 | 76 | 42 |
| 101 - 250 | 19 | 182 |
| 251 - 500 | 8 | 376 |
| 501+ | 2 | 825 |
| **Average**: | | **108** |

**Table 4: Count of courses by course size.**

*4.3.1 Included topics and teaching methods.* For undergraduate courses, specifically observing Fig. 4, we see that the following questions (identified in Table 2) are taught by more than 50% of data systems educators at the **undergraduate** level:

**Q10.16**: "data modeling: creating tables and columns" (96%)
**Q10.23**: "SQL: select, project, join" (96%)
**Q10.24**: "SQL: insert, update, delete" (96%)
**Q10.1**: "relational theory: relations, tuples and attributes" (95%)
**Q10.25**: "SQL: aggregation and group by" (95%)
**Q10.14**: "data modeling: conceptual modeling" (92%)
**Q10.15**: "data modeling: mapping conceptual models to logical models" (92%)
**Q10.26**: "SQL subqueries" (92%)
**Q10.11**: "logical and physical data independence" (91%)
**Q10.12**: "database management system components" (90%)

| Program Level | Course Level | Count | Total |
|---|---|---|---|
| Undergraduate | 1 | 6 | 78 |
| | 2 | 23 | |
| | 3 | 26 | |
| | 4 | 13 | |
| | Unknown | 10 | |
| Graduate | 1 | 15 | 27 |
| | 2 | 2 | |
| | 4 | 3 | |
| | Unknown | 7 | |
| **Total**: | | **105** | |

**Table 5: Count of courses by program and course level. 17 submissions had entry errors and are reported as "Unknown" (e.g., entering the year the course finished/last happened, instead of the providing the year/level of the course.**

| Question | Count |
|---|---|
| "They were dictated by other courses" | 25 |
| "I applied curriculum guidelines" | 38 |
| "Someone else (e.g., in the faculty) dictated them" | 13 |
| "They are based on industry needs" | 47 |
| "I chose them myself (i.e., based on intuition, past experiences...)" | 50 |
| "I inherited the course from a predecessor" | 42 |
| "They are based on a textbook" | 43 |
| "Other" | 6 |

**Table 6: Counts the number of selections chosen based on Q12: "Why did you choose these course topics?"**

**Q10.17**: "database normalization: functional dependency, candidate and super keys" (90%)
**Q10.18**: "database normalization: normal forms up to BCNF" (85%)
**Q10.28**: "transaction processing" (79%)
**Q10.3**: "relational algebra" (78%)
**Q10.37**: "SQL: common table expressions" (77%)
**Q10.5**: "database optimization: indexing" (76%)
**Q10.6**: "database optimization: query execution plans" (69%)
**Q10.37**: "data security and database access management" (69%)
**Q10.29**: "concurrency control and isolation levels" (68%)
**Q10.7**: "database optimization: query optimization" (64%)
**Q10.13**: "functions and stored procedures" (59%)
**Q10.19**: "database normalization: multivalued dependency" (59%)
**Q10.36**: "data privacy and ethics" (56%)
**Q10.20**: "database normalization: join dependency" (55%)
**Q10.10**: "NoSQL database management systems" (54%)
**Q10.30**: "database back-ups and recovery" (53%)

Interestingly, topics that are mostly not covered, are data mining and warehousing, relational calculus, object-oriented data models, semi-structured traditional data models, database scalability, distributed database management systems, and data visualization.

For graduate courses, specifically observing Fig.5, we see that the following questions (identified in Table 2) are **not** taught by more than 50% of data systems educators at the **graduate** level:

**Q10.34**: "data mining: data cleaning" (41%)
**Q10.38**: "data warehousing" (41%)
**Q10.36**: "data privacy and ethics" (41%)
**Q10.32**: "data mining: algorithms" (33%)
**Q10.35**: "data mining: market basket analysis" (26%)
**Q10.33**: "data mining: associative and sequential patterns" (22%)

Interestingly, when comparing the graduate findings to the undergraduate findings, data mining is the least taught topic. Additionally, Q10.4 ("data visualization") is one that is often thought of going hand-in-hand with data systems education [11, 60], but often not taught in the graduate or undergraduate curriculum at all. While Q10.18 ("database normalization: normal forms up to BCNF"), is unsurprisingly considered to be a pre-requisite for some graduate courses (26% of graduate courses), more surprisingly is that it is still taught at the graduate level (63%).

*For all courses*, we see that the following topics (identified in Table 2) are taught by more than 80% of data systems educators:

**Q10.1**: "relational theory: relations, tuples and attributes"
**Q10.11**: "logical and physical data independence"
**Q10.12**: "database management system components"
**Q10.14**: "data modeling: conceptual modeling"
**Q10.15**: "data modeling: mapping conceptual models to logical models"
**Q10.16**: "data modeling: creating tables and columns"
**Q10.17**: "database normalization: functional dependency, candidate and super keys"
**Q10.23**: "SQL: select, project, join"
**Q10.24**: "SQL: insert, update, delete"
**Q10.25**: "SQL: aggregation and group by"
**Q10.26**: "SQL subqueries"

This isn't surprising as the majority of responses related to undergraduate courses and these topics can be considered fundamentals of data systems. Conversely, the following topics are taught by less than 25% of classes:

**Q10.32**: "data mining: algorithms"
**Q10.33**: "data mining: associative and sequential patterns"
**Q10.35**: "data mining: market basket analysis"

This could be similarly explained by these being considered more advanced topics. The only other data mining topic *data mining: data cleaning* was covered by more classes but still only 41% of graduate courses and less than 30% of undergraduate courses.

In the DataEd workshop in 2022, they found that unique topics among the instructors present were "Datalog, database security and SQL injections" [8]. This is reflected in our findings, as our topic list, formed from the curriculum analysis in section 3, does not have separate items for Datalog and SQL injections. On the other hand, we can map database security to topic 10.37 (data security and database access management), which was prevalent, being taught in 54 undergraduate courses 69%, and in 18 of our surveyed graduate courses (67%). Surprisingly, given the supposedly advanced nature of the topic, two of the undergraduate instructors mentioned they require knowledge of this topic as a prerequisite for their course, whereas none of the graduate instructors do. Attendees at DataEd'22 furthermore suggested declarativeness and conceptual modeling as topics that should be taught in more courses [8]. These are very broad topics, and as such, further investigations are required to map such suggestions more precisely.

*4.3.2 Motivations behind the syllabus.* Investigating the motivations behind data systems educators' choices for syllabus design in data systems courses (specifically looking at Table 6, which allowed for multiple-selection and qualitative input), it appears that most educators (40%) inherited the course from a predecessor and continued to teach it the same way. From the qualitative data, it is clear that many educators do not have the time to renew the course they are given and many are forced to re-teach using past materials due to institutional constraints. This was similar to the findings of those who used a course textbook and its materials (41%). Despite this, nearly half of data systems educators (48%) found the time to design or redesign their own courses. Choices regarding what to include in their courses were based on intuition and past experiences, and several choices were based on curriculum guidelines (36%). Furthermore, many data systems educators based their course content on industry needs (45%).

## 5 Industry Needs

The final study in this paper investigates the current and future skills required by graduates looking to move into the data systems industry. In this section, we investigate **RQ4**: Which data-related skills are valued for industry roles such as software developers, data engineers, and data scientists?

This last piece of the puzzle allows us then to discuss the alignment between curriculum guidelines, course curricula, and industry requirements, in order to design education in such a way that it prepares students well for their further careers.

### 5.1 Methods

We perform two different analyses: a survey of industry professionals and a job advertisement analysis. The aim was to capture the industry perspective on the data-related skills deemed essential for entry-level positions requiring database knowledge, as well as identify the alignment of topics with the curriculum guidelines and the teacher survey. The job ad analysis then provides real-time data on the current demands of the job market. Combining the survey of industry professionals with a job advertisement analysis provides a more holistic view of the labor market, which can help validate our findings from the survey as well as complement them.

*5.1.1 Job Advertisement Analysis.* Job advertisements can provide an accurate insight into the current state of the job market and the skills that companies seek in recent graduates. Therefore, we used Lightcast Q1 2024 USA Data Set [67] to analyze job postings for database-related roles (Table 8) and all computing occupations (Table 9) in the USA. Lightcast's data is a robust hybrid dataset, derived from authoritative sources such as the U.S. Census Bureau, Bureau of Economic Analysis, and Bureau of Labor Statistics, capturing over 99% of the workforce in the United States. This extensive

[2]https://en.wikipedia.org/wiki/International_Standard_Industrial_Classification

| ID | Question | Style | Options/(Restrictions) |
|---|---|---|---|
| Q1 | Which country are you based in? | Open-Textbox | N/A |
| Q2 | What primary International Standard Industrial Classification (ISIC) category does your company belong to? | Radio-Button | ISIC categories[2] |
| Q3 | What is your company size? | Radio-Button | - Micro: 0-9 employees<br>- Small: 10-49 employees<br>- Medium: 50-249 employees<br>- Large: 250+ |
| Q4 | The most recent total number of employees in my work unit/department is approximately: | Open-Textbox | (numbers only) |
| Q5 | For a strong candidate for **entry-level positions requiring database knowledge**, please indicate the necessary skills by checking the appropriate level: Critical, High, Moderate, or Low. If a topic is not applicable, simply skip it. | Radio-Buttons per topic | - relational theory<br>- relational algebra<br>- data visualization<br>- database optimization (including tuning and performance analysis)<br>- database scalability<br>- cloud computing<br>- distributed database management systems<br>- NoSQL database management systems<br>- logical and physical data independence<br>- database management system components/internals<br>- functions and stored procedures<br>- data modeling<br>- database normalization<br>- object-oriented data models<br>- semi-structured traditional data models (e.g., XML, JSON)<br>- SQL (Programming Language)<br>- JDBC, ODBC, or alternative connection to applications<br>- transaction processing<br>- concurrency control and isolation levels<br>- database back-ups and recovery<br>- this is a sanity check, select "High"<br>- data mining<br>- data privacy and ethics<br>- data security and database access management<br>- data warehousing<br>- data processing pipeline |
| Q6 | What gaps, if any, in database knowledge have you observed in recent graduates you have hired? | Open-Textbox | N/A |
| Q7 | What database-related resources (including certifications/training) do you offer your employees to support them? | Open-Textbox | N/A |
| Q8 | Indicate the five most important **technical** skills you believe will be essential or fundamental in the next decade for **entry-level positions requiring database knowledge**. | Open-Textbox | N/A |
| Q9 | Indicate the five most important **non-technical** skills you believe will be essential or fundamental in the next decade for **entry-level positions requiring database knowledge**. | Open-Textbox | N/A |
| Q10 | Please share any additional important skills, resources, comments, or other information. | Open-Textbox | N/A |

**Table 7: These are the 10 questions of the DataEd industry survey questions, including the presentation style provided, and any relevant options. All questions were required to complete.**

coverage is further enriched with data from online social profiles, resumés, and job postings, providing a comprehensive view of the workforce. By leveraging Lightcast's dataset, our analysis gains a high level of accuracy and depth, offering a detailed understanding of the current job market demands and the specific skills employers are seeking. This combination of official data and real-time job market insights makes Lightcast a reliable source for evaluating the skills landscape in computing and database-related occupations. Our analysis focused on three categories of skills: top distinguishing skills by demand, top defining skills by demand, and top necessary skills by demand. This detailed analysis helps us understand the current demands in the job market, providing insights into the skills required for all computing occupations and database-related roles.

| Type of Skill | Skill | # Postings | Projected Growth |
|---|---|---|---|
| Top Distinguishing Skills | SQL Server Integration Services (SSIS) | 18,576 | 14.2% |
| (Specialized advanced skills) | Data Lakes | 17,067 | 10.7% |
| | Database Design | 13,926 | 2.7% |
| | Relational Database Management Systems | 13,499 | 9.8% |
| Top Defining Skills | SQL (Programming Language) | 115,189 | 6.4% |
| (Core daily tasks) | Data Engineering | 96,334 | 16.1% |
| | Extract Transform Load (ETL) | 71,813 | 9.0% |
| | Data Warehousing | 53,052 | 6.0% |
| | Data Modeling | 45,657 | 19.3% |
| Top Necessary Skills | Data Analysis | 40,134 | 25.8% |
| (Foundational specialized skills) | Scalability | 29,508 | 25.2% |
| | Business Intelligence | 27,893 | 21.0% |
| | Data Quality | 27,862 | 21.7% |

**Table 8: Top skills by demand for database-related occupations**

| Type of Skill | Skill | # Postings | Projected Growth |
|---|---|---|---|
| Top Distinguishing Skills | Computer Science | 7,779,181 | 27% |
| (Specialized advanced skills) | SQL (Programming Language) | 5,306,332 | 18% |
| | Agile Methodology | 5,181,547 | 9% |
| Top Defining Skills | Communication | 10,378,351 | 36% |
| (Core daily tasks) | Management | 7,556,555 | 26% |
| | (Problem Solving | 6,066,410 | 21% |
| Top Necessary Skills | SQL (Programming Language) | 5,306,332 | 18% |
| (Foundational specialized skills) | JavaScript (Programming Language) | 4,117,888 | 14% |
| | Python (Programming Language) | 3,552,581 | 12% |

**Table 9: Top skills by demand for all computing occupations**

The top distinguishing skills are the advanced competencies that set candidates apart in the job market. These skills are highly sought after and often indicate a candidate's specialization and ability to perform unique tasks. The top defining skills are the core competencies required to perform daily tasks in the role. These skills are fundamental and ensure that the candidate can meet the essential job requirements. The top necessary skills are specialized skills that are not only required for those roles but are also relevant across similar jobs. These skills form the building blocks for more advanced tasks.

These categories of skills show current demands in the job market. Table 8 shows the top skills in demand for database-related occupations, the number of job postings requiring each skill, and the projected growth for these skills. The data reveals a strong demand for a mix of specialized advanced skills, core daily tasks, and foundational knowledge in database-related occupations. Skills like SQL and Data Engineering remain critical, while areas like Data Analysis, Scalability, and Data Modeling show high projected growth, indicating evolving priorities in the data landscape. This suggests a job market that values both the ability to handle day-to-day database management tasks and the specialized skills needed to innovate and improve data systems.

*5.1.2 Industry Survey.* The industry survey comprised eleven questions and was distributed to a targeted group of industry professionals. These professionals were identified through the University Industry Boards and partnerships with working group members. Additionally, the survey was shared on LinkedIn and Reddit, specifically on database-related subreddits. Similar to the teachers' survey, an international and diverse range of industry professionals were surveyed to ensure a broad perspective was considered. For the full questionnaire, see Table 7.

Different countries employ various standards for classifying companies based on industry and employee count. To ensure consistency and comparability in our survey, we adopted the classifications provided by the United Nations Statistics Division. For classification of the type of industry, we utilized the International Standard Industrial Classification of All Economic Activities (ISIC) [125]. This classification is the global standard for classifying productive activities. It provides unified categories for collecting and reporting statistics. Since 1948, it has been widely adopted by countries as a basis for national classifications, facilitating international comparisons of economic activity data [125]. Regarding company size, we referred to the guidelines outlined in the Manual on Principal Indicators for Business and Trade Statistics [124]. To streamline the survey, we summarized the 38 categories defined in the teachers' survey, by combining them into broader, high-level topics such as merging indexing, query execution plans, and query optimization into database optimization. This adjustment allowed the participants to focus on overarching themes, as the detailed granularity of specific topics was deemed less essential for industry-wide understanding and application.

In addition to the aggregation, based on the teacher survey outcomes, new topics were incorporated to address gaps and ambiguities. Additionally, we ensured that the language used aligned with industry terminology by reviewing current job postings and consulting industry reports. Cross-referencing recognized standards like ISIC classifications and using tools such as Lightcast's workforce data helped maintain accuracy and relevance in our terminology. The topics added from the curriculum guidelines were: data processing pipeline, cloud computing, and database environmental management and impact. To enhance clarity, some examples were provided: "database optimisation" was updated to "database optimisation (including tuning and performance analysis)", and "database management system components" was updated to "database management system components/internals". Similarly, "semi-structured traditional data models (e.g., XML)" was updated to "semi-structured traditional data models (e.g., XML, JSON)".

Analysis of the teachers' survey qualitative data also led to additional topics for the industry survey. The survey results revealed that many courses include topics such as ODBC, JDBC, and other methods for connecting databases to applications, so we have added "JDBC, ODBC, or alternative connection to applications" to the list of topics.

Finally, a sanity check question was included in the list of topics to help identify any erroneous data to ensure respondents were completing and accurately reading the questions.

To ensure accurate tracking and analysis of responses the industry survey was distributed to a wide community using two methods: direct contacts and social media such as Reddit and LinkedIn. By separating the survey into two versions, responses could be tracked from known contacts within the industry versus those from a broader, open-source community. The duplication of the survey facilitated precise tracking of responses and ensured that we could manage and analyze data from each community separately, which allowed us to maintain the integrity of the data and provided a more organized dataset for subsequent analysis. By using separate surveys for different communities, we could more effectively identify and filter out responses that appeared to be generated by trolls or non-serious participants. This separation allowed us to implement targeted checks and controls for each group, improving the overall quality and credibility of the data. Overall, the duplication of the survey was a strategic decision aimed at enhancing the accuracy, relevance, and reliability of the data collected from diverse respondent groups.

For the topics in question 5 on assessing the necessary skills, the scale of Critical, High, Moderate, or Low was chosen as it provides a clear and structured evaluation framework. Responses were aggregated and analyzed to identify patterns and trends across Critical, High, Moderate, and Low ratings for each skill category.

Our job advertisement analysis highlighted the importance of non-technical skills. Therefore, we ended our survey with questions about potential gaps in database knowledge observed in recent graduates they hired, including a question on both technical and non-technical skills. Finally, employers were asked about desired database-related certifications or training they would prefer their employees to pursue within their first three years of employment. This approach allowed for nuanced insights from employers regarding specific challenges encountered in the workforce and expectations for ongoing professional development. The questionnaire was wrapped up with an open question for any additional comments or remarks.

## 5.2 Results

The social media survey initially yielded 12 responses, with 6 being fully completed. After applying a sanity check, 2 responses were excluded, leaving 4 valid responses from this group. The second survey, distributed to direct contacts, received 38 responses, of which 30 were fully completed. All participants in this group passed the sanity check. After combining and cleansing the data, a total of 34 valid responses were used. Our analysis focuses on the combined dataset to provide a comprehensive overview.

Table 10 provides detailed statistics on the countries represented in the survey. The survey responses were spread across six countries, with the majority coming from the United States (11 responses in total, 10 from direct contacts and 1 from social media). Canada and Finland also had significant representation, with 7 and 6 responses from direct contacts, respectively. The United Kingdom had a total of 5 responses (3 from direct contacts and 2 from social media). Other countries represented include the Netherlands (4 responses from direct contacts) and Switzerland (1 response from social media). This diverse spread indicates a range of perspectives from different geographical regions; however, its limitations are further discussed in the limitations section.

| Country | Direct Contacts | Social Media |
|---|---|---|
| United States | 10 | 1 |
| Canada | 7 | None |
| Finland | 6 | None |
| Netherlands | 4 | None |
| United Kingdom | 3 | 2 |
| Switzerland | None | 1 |
| **Total** | 30 | 4 |

**Table 10: Count of participating industry professionals by country.**

The survey responses represent organizations of varying sizes, providing insights into database management practices across different scales of operation. For an overview of the statistics, see Table 12. The majority of responses came from large organizations with over 250 employees, accounting for 19 responses. Such larger enterprises often have more complex and extensive database requirements. Mid-sized organizations, with 50 to 249 employees, contributed 7 responses (6 from direct contacts and 1 from social media), and smaller organizations with 10 to 49 employees were also represented by 7 responses (6 from direct contacts and 1 from social media), indicating that database management is relevant across diverse organizational sizes, though potentially with differing priorities and resource availability. Only one response came from an organization with fewer than 10 employees, highlighting a potential gap in understanding how very small businesses approach database management.

| Topic | Average | Median | Mode |
|---|---|---|---|
| SQL (Programming Language) | 3.2 | 3 | 4 |
| data security and database access management | 2.7 | 3 | 3 |
| data modeling | 2.6 | 3 | 2 |
| JDBC, ODBC, or alternative connection to applications | 2.5 | 2.5 | 4 |
| semi-structured traditional data models (e.g., XML, JSON) | 2.4 | 2 | 4 |
| cloud computing | 2.4 | 2 | 2 |
| functions and stored procedures | 2.3 | 2 | 2 |
| data privacy and ethics | 2.3 | 2 | 2 |
| data processing pipeline | 2.3 | 2.5 | 3 |
| database optimization (including tuning and performance analysis) | 2.2 | 2 | 2 |
| NoSQL database management systems | 2.1 | 2 | 3 |
| database scalability | 2.1 | 2 | 2 |
| database normalization | 2.1 | 2 | 2 |
| concurrency control and isolation levels | 2.1 | 2 | 1 |
| data visualization | 2.0 | 2 | 1 |
| database back-ups and recovery | 2.0 | 2 | 2 |
| distributed database management systems | 2.0 | 2 | 1 |
| transaction processing | 2.0 | 2 | 1 |
| object-oriented data models | 1.9 | 2 | 2 |
| data warehousing | 1.9 | 2 | 2 |
| data mining | 1.9 | 2 | 1 |
| database management system components/internals | 1.9 | 2 | 2 |
| relational theory | 1.9 | 1.5 | 1 |
| logical and physical data independence | 1.7 | 2 | 2 |
| relational algebra | 1.4 | 1 | 1 |

**Table 11: Summary of Responses by Topic**

| Number of Employees | Direct Contacts | Social Media |
|---|---|---|
| 250+ | 17 | 2 |
| 50-249 | 6 | 1 |
| 10-49 | 6 | 1 |
| 0-9 | 1 | None |
| **Total** | 30 | 4 |

**Table 12: Participants based on their companies' number of employees.**

To present the results for the topics in a tabular format (Table 11), numerical values were assigned to the scale: Critical (4), High (3), Moderate (2), and Low (1). We report the topics sorted from most critical to least critical. At a glance, the table shows many ratings between 1.9 and 2.1, giving them a Moderate score. In addition, only three scores have medians of High: SQL, data security, and data modeling.

The responses to the open-ended industry questions consistently indicated that recent graduates often lack practical experience with SQL, with their knowledge being largely theoretical. This limitation hinders their ability to adapt to different database platforms. They also face challenges in problem-solving, particularly when dealing with errors—a common occurrence in database work. Additional gaps were identified in areas such as writing high-performance queries for large databases, basic database administration, and preventing SQL injection. A noticeable deficiency in troubleshooting abilities and depth in areas like database optimization, data normalization, and performance tuning was also observed. These findings suggest a need for more hands-on training and exposure to real-world database scenarios.

The survey emphasized that a blend of technical and non-technical skills will be crucial for entry-level positions requiring database knowledge in the coming decade. Foundational technical skills such as SQL and NoSQL remain essential for querying databases, while knowledge of cloud-based architectures, services, and data integration platforms is increasingly vital as organizations transition to cloud solutions. Practical expertise in data modeling, database optimization, performance tuning, and data security is also highlighted, alongside proficiency in tools for data visualization and data pipeline management. Understanding data security tools and practices, as well as ethical considerations around Personally Identifiable Information, is deemed essential. Emerging areas like machine learning, AI, and big data management are also noted, with skills in vector databases and functional programming gaining importance for developing advanced data models and automation processes.

Non-technical skills are equally critical, emphasizing communication, adaptability, and teamwork, both in remote and in-person settings. The ability to convey data insights and collaborate with others is vital in data roles. Adaptability and continuous learning are necessary to keep pace with rapidly evolving technologies, while problem-solving and analytical thinking skills are essential for applying data to real-world scenarios. Attention to detail, ethical awareness, emotional intelligence, and a positive attitude further enhance a professional's ability to work effectively within teams and maintain high standards in their work.

### 5.3 Discussion

Direct comparison between the Lightcast results and the industry survey is challenging due to reporting in frequency versus importance. The industry survey aligns with academic course topics, emphasizing exclusively foundational and technical skills. In contrast, Lightcast focuses on resume terminology and job postings, highlighting skills in specific tools and frameworks such as SQL Server Integration Services (SSIS) and Data Lakes. However, we do see that both emphasize foundational skills like SQL and data modeling, underlining their importance in academic courses and practical job applications. On the other hand, there are differences in focus areas: while the industry survey lists topics like data warehousing, NoSQL database management, and data processing pipelines, Lightcast emphasizes more specific skills such as SSIS and Data Lakes. The concept of Data Lakes from Lightcast can be viewed as an extension or modern iteration of Data Warehousing in the industry survey, reflecting evolving industry needs for handling vast and varied data sources. Additionally, Lightcast mentions broader skills like data engineering and business intelligence, which encompass multiple topics from the industry survey, suggesting a shift from individual technical skills to more integrated and comprehensive roles in data management.

The Lightcast dataset primarily reflects the current job market demands as captured through job postings, while the industry survey offers insights into the perceived importance of various skills from professionals' perspectives. The alignment in skills like SQL and data modeling suggests a consensus on foundational skills necessary for database-related roles. However, divergences in areas like database scalability and cloud computing indicate a potential gap between what employers seek and what professionals prioritize. This gap may stem from evolving industry trends, with employers preferring forward-looking skills while professionals focus on more established practices.

Our findings suggest that while there is consensus on certain core skills, a dynamic landscape of emerging skills requires both professionals and employers to adapt continually to remain relevant in the field of database management.

## 6 Discussion

The competencies identified in the different parts of this paper are derived from various sources. Even so, they are all quite specific, as that makes it easier for instructors and industry professionals to rank them. Unfortunately, this makes comparing their prevalence to the Cruickshank et al. [29] framework for data education challenging. Their framework uses over-arching themes such as *domain knowledge*, *problem formulation* and *data management*. Overall, the competencies seem to focus on the data pipeline, something that was only rarely mentioned in the curriculum guidelines, and also not very pressing to most of the database instructors in our teacher survey. This might be explained by our focus on database curricula over data science, although data science students are in the target group of many of the surveyed courses. On the other hand, our industry analysis shows that many of the practical skills by Cruickshank et al. [29] are also coming up in job openings. Lightcast analysis shows topics related to infrastructure building, model building, and model production, as well as data analysis and communicating results. Furthermore, the qualitative answers to the survey show that the industry would value it if students were able to practice SQL and data modeling more in-depth, instead of the primarily theoretical approach. This could help to reduce the challenges in problem-solving and prepare students for writing more advanced queries, taking into account query performance.

The competencies of Aasheim et al. [4] are slightly more specific, but also more data systems-adjacent, including concepts of *visualization techniques* and *analytics techniques*. The competencies are not focused on the data pipeline, although the latter concept shows that it is considered important. Most of the data systems-specific topics in the competencies are also present in at least one curriculum guideline, including visualization, data capture (information extraction), and even ethical considerations. There is also some overlap in the teacher survey, although only on less than half of the competencies. Overlap is present on the topics:

- Data management (topics 10, 12, 31, 37) which is covered in between 48 and 67 percent of surveyed courses.
- Data mining techniques (topics 32, 33, 34, 35), covered in 17 to 30 percent of the surveyed courses.
- Data visualization (topic 4), covered in 44% of courses.
- Modeling/analytics techniques (topic 14, 15, 16, 22), with semi-structured models taught in less than half of the courses but data modeling on average in 80%.
- Data security (topic 37) in 67% of courses.

Although there is some overlap that is surprising, such as the inclusion of visualization and data mining in some courses, the focus on data systems courses over data science courses can explain the lack of overlap in topics. As for the industry preferences, there is overlap in the practical focus of the Aasheim et al. [4] competencies. The qualitative analysis indicates that students lack experience in many areas such as query formulation and performance tuning. The competencies also overlap the industry findings on the topics of non-technical skills, the survey uncovered skills such as communication of insights and collaboration, which maps onto the communication skills by Aasheim et al. [4], while problem-solving and analytical thinking can be mapped to decision making and evaluation. Finally, ethical awareness was a factor present in both the survey and the competencies.

### 6.1 Alignment of curriculum guidelines and course contents

Overall, curriculum guidelines and teachers take similar approaches to the design of data systems courses. Various elements of data management courses can be intuited as important, based on their ubiquitousness in industry or their use as preliminaries for subsequent courses.

In our course analysis, we see that commonly included topics are practical skills such as SQL and modeling, and management topics such as data independence and transactions. These practical skills were also the most common in our curriculum guideline analysis. In our course sample, courses were taught that way as they were inherited from a predecessor or based on a book, without time to update. This means that the course details are unlikely to be up to

date on recent guidelines, although another 36% of our participant mention their course is built on curriculum guidelines.

On the other hand, our analysis shows that the guidelines are overall significantly distinct tools, having different foci and lack of common ground. The guidelines all provide distinct lenses on computer science. Although all included guidelines contain some skills that we categorize as data management-adjacent, IS2020 and SE2014 both have fewer than five relevant topics.

The most popular topics in the guidelines are visualization, analytics, security, modeling, big data, and data management. Comparing this to the most common topic in the courses, we notice that query languages in general, or SQL specifically, are not a separate topic in the curriculum guidelines. Query construction is mentioned in CS2023 and the UK guidelines, but whether these require more explicit attention is a topic for future discussions.

Finally, neither our analysis of course contents nor our analysis of curriculum guidelines focused on non-technical skills. For the curriculum guidelines, these may have been included in the originals, but we exclusively extracted data management skills. Given the importance of these skills according to our industry participants, the inclusion of these skills into data management courses warrants further discussion.

## 6.2 Alignment of curriculum guidelines and industry requirements

Compared to the alignment between curriculum guidelines and course contents, the guidelines and industry preferences are much more in line. Both of these sources indicate a focus on practical skills, such as the data pipeline and forms of data mining and analysis, as well as database operations and day-to-day management. The industry survey does show an evolving demand for more advanced skills such as scalability and cloud computing, which were not explicitly mentioned in the curriculum guidelines.

We are not able to compare the focus of both sources on non-technical skills, as we dropped these elements in our analysis of the curriculum guidelines to focus on data systems-specific skills.

## 6.3 Alignment of course contents and industry needs

Although 45% of our instructors say they based their course on industry needs, Figure 8 shows that there are large discrepancies between the prevalence of topics being taught and the values that our industry participants assign to these topics. Some of the topics with the highest discrepancies are again the more practical, data management-adjacent topics such as visualization, data mining, and privacy and ethics.

Topics with a good match with regard to prevalence and importance are relational theory, relational algebra, data modeling, and data normalization. These are more traditional data management topics and seem to be valued relatively well. These form a stable core for data management courses, even in light of more recent foci on data science topics or scalability and cloud computing.

One noteworthy topic is SQL, which gained the highest importance score from industry participants, as well as being one of the most commonly taught subjects. Nonetheless, the extent to which SQL is taught seems to be lacking in terms of industry needs. The qualitative industry questions show us that new hires are often unprepared for the required query formulation and improvement tasks. They stress that students need more practice with these topics than they are currently getting.

On the other hand, it is important to note that the average and median scores for all of these topics are relatively small, as can also be seen in Table 11. One hypothesis could be that different company types appreciate different skills differently, leading to lower scores overall. As we do not have enough industry results to run a statistical analysis, this is a research question for future work.

The qualitative industry data suggest shifts to distributed, cloud-based architectures, and the skills that are associated with this are not reflected much in our course data. However, this might be due to the skew of our data to undergrad courses, whereas distributed data management might be seen as a graduate course in many universities.

## 6.4 Limitations

First of all, when surveying a multinational audience, several terminology interpretation differences across countries may arise. For example, interpretation of the term "student outcomes" varies significantly: in the USA, it refers to what students are expected to know and be able to do by the time of graduation [5], whereas in the UK, "Graduate Outcomes" is employer information for the students graduated from a higher education course in a specific academic year [54]. Another difference in interpretation arose in the teachers' survey, where the question "What is the general course size in number of students?" was interpreted differently by respondents; some provided the size of a course section, while others reported the total number of students who took the course throughout the academic year. These discrepancies highlight the challenges in obtaining consistent data from a diverse international audience. Finally, the terms "not covered" and "prerequisite" in the teacher survey seem to have been interpreted differently by different participants. The term prerequisite is something that could refer to the academic regulations that are mandatory for the eligibility of another course or merely an expectation that this will have been covered prior to undertaking this course. This ambiguity potentially resulted in some topics being reported as "not covered" rather than "prerequisite".

While the industry survey responses show representation from six countries, there are several limitations to this distribution. First, the majority of responses are concentrated in just a few countries, particularly the United States and the United Kingdom, which may result in a sample that is not fully representative of global perspectives. The relatively low number of responses from other regions, such as Asia, Africa, and South America, limits the generalizability of the findings across a more diverse international context. Additionally, the number of responses from social media is small (4 in total), which could skew the data toward the perspectives of direct contacts. This uneven geographical distribution and sample size may affect the applicability of the results to different cultural, educational, and professional settings around the world.

A notable limitation of this study is the number of respondents who started but did not complete both the teachers and industry

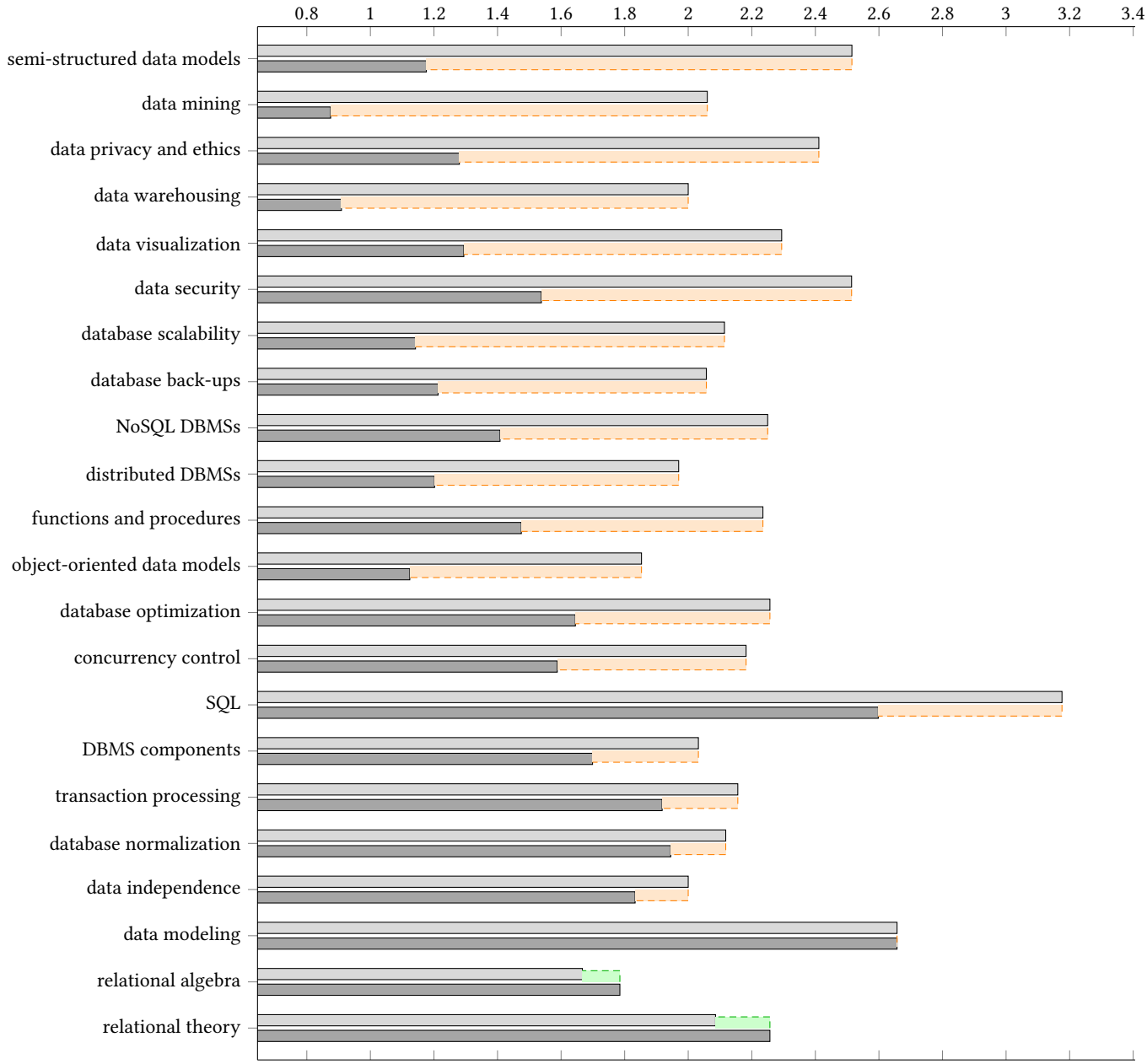


**Figure 8: Topics taught (dark gray) and needed in the industry (light gray); the y-axis is normalized from a 4-point Likert scale and Bloom's taxonomy; the topics are arranged based on the amount of discrepancy between teaching and industry**

surveys. This issue introduces potential biases, as incomplete responses may reflect differing characteristics or opinions compared to those who completed the survey. Only complete answers were considered. Consequently, the final dataset may not fully capture the diversity of perspectives present in the initial sample. The high rate of incomplete responses also reduced the overall sample size, which may affect the statistical power and reliability of the findings. Additionally, incomplete responses suggest potential issues with the survey design or respondent engagement, which should be addressed in future iterations of the survey. To mitigate these concerns, we analyzed the completed responses while acknowledging the limitations related to missing data and its impact on the generalizability of our results.

Additionally, another limitation of this study is the potential misalignment of terminology and language between academic and industry respondents. The terminology used in the survey may differ between academic research contexts and practical industry applications, leading to possible misunderstandings or differing

interpretations of survey questions. This misalignment could affect the consistency and accuracy of responses, as academic respondents might approach questions with a theoretical perspective, while industry professionals might interpret them based on practical experience. Future research should consider this potential discrepancy by refining survey language to ensure that it is comprehensible and relevant to both groups, thereby enhancing the validity of the findings.

Variability in the implementation of data systems topics across different institutions may limit the ability to draw uniform conclusions about what is taught. Differences in course structure, instructor expertise, and institutional priorities can lead to significant variations in how topics are covered. To address this, a more detailed examination of how specific topics are taught and the factors influencing these variations could be done.

The evaluation of which data systems topics are valued in industry roles may be influenced by the subjective opinions and biases of various stakeholders. This subjectivity can lead to variability in the perceived importance of different topics. Future studies should consider employing multiple evaluators or using objective measures to assess the value of different data systems topics to help mitigate this.

Some smaller limitations that have to do with methods and designs are: the lack of sanity check in the teacher survey, although this is mitigated by the participants being recruited within-network. We also did not think to consider whether the focus of a data management course within a computer science program would be different than within a data science or AI program. Including a question on the program that the course resides in would have made us able to distinguish between these.

Overall, a set of studies of this size will always have some shortcomings, but we hope that the practical implementations and recommendations in the next section will provide useful suggestions.

### 6.5 Practical Implications and Recommendations

In general, aligning the data system topics in curriculum guidelines and course contents with industry needs serves higher education students by efficiently preparing them for their future work. However, the discrepancies between industry needs and higher education should not necessarily be interpreted as gaps to fill in data systems education. Higher education does not exclusively aim to train students on the topics which are relevant in industry, but instead aims to educate a new generation of academics. This means a focus on other topics too, to build knowledge foundations and skills such as scientific reporting.

On the other hand, the cycles in which curriculum guidelines and course contents are updated, and simply the fact that it takes several years to graduate, might effectively mean that students are always equipped with an at least somewhat out-of-date skill set when moving to industry. With that in mind, teaching the foundations of data management and the underlying principles (e.g., relational algebra) provides a safeguard and strong base in case the specific query languages and other state-of-the-art elements the instructor teaches do not align with local industry preferences.

Furthermore, our industry analysis points to a lack of practice among students. Data management courses need to touch upon many different topics, meaning that most topics are only in focus during one or two lectures. As a result, students only have superficial practice on these topics, making it harder for them to become fluent in working with them. It would be interesting to study how the practice of foundational data management skills can be expanded by means of follow-up courses, capstones, or even industry internships, as well as to talk to recent graduates to evaluate their experience.

Finally, data management-adjacent topics such as data mining, analytics, and visualization were shown to be important, both in some curriculum guidelines as well as in our industry survey. However, given the large number of topics and lack of practice that we already see in the current setup of data management courses, perhaps as teachers we should decide that the place for these data science topics is in another course.

## 7 Conclusions

Data systems have been a part of effectively all computing curricula for decades, and several curriculum guidelines include data system topics as one of the core topics in computing. Additionally, many positions in industry require data systems knowledge such as database programming, query languages, database design, and data analytics. Consequently, there are many courses in higher education that prepare students for various data-focused roles in their future careers. However, possibly due to the ubiquitous and changing nature of data systems, it has remained largely unclear how accurately curriculum guidelines, data systems teaching in higher education, and the needs of data-focused roles in industry align. To that end, we analyzed the similarities and discrepancies between curriculum guidelines, course contents, and industry needs in order to understand what is potentially missing from guidelines and courses, and which topics are potentially extraneous when educating future data professionals in higher education. Our recommendations and guidelines suggest to keep teaching the foundations, while making space for repeated practice and include non-technical skills practice wherever possible.

## References


[1] 2022. ACADEMIC ACCREDITATION GUIDELINES 2022. https://www.bcs.org/media/1209/accreditation-guidelines.pdf British Curriculum Guidelines. Accessed: 2024-09-17.

[2] 2024. BCS LEVEL 5 DIPLOMA IN IT DATABASE SYSTEMS - SYLLABUS. https://www.bcs.org/media/9227/heq-dip-ds-syllabus.pdf British Curriculum Guidelines. Accessed: 2024-09-17.

[3] Vision 2030. 2021. Human Capability Development Program 2021-2025. https://www.vision2030.gov.sa/media/pgid4z3t/2021-2025-human-capability-development-program-delivery-plan-en.pdf Accessed: 2024-09-18.

[4] Cheryl L Aasheim, Susan Williams, Paige Rutner, and Adrian Gardiner. 2015. Data Analytics vs. Data Science: A Study of Similarities and Differences in Undergraduate Programs Based on Course Descriptions. *Journal of Information Systems Education* 26 (2015).

[5] ABET. [n. d.]. Accreditation Criteria. https://www.abet.org/accreditation/accreditation-criteria. Worldwide organisation. Accessed: 2024-07-05.

[6] Willem Aerts, George Fletcher, and Daphne Miedema. 2024. A Feasibility Study on Automated SQL Exercise Generation with ChatGPT-3.5. In *Proceedings of the 3rd International Workshop on Data Systems Education: Bridging education practice with education research*. 13–19. https://doi.org/10.1145/3663649.3664368

[7] Alireza Ahadi, Julia Prior, Vahid Behbood, and Raymond Lister. 2016. Students semantic mistakes in writing seven different types of SQL queries. In *Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE)*.

272–277. https://doi.org/10.1145/2899415.2899464
[8] Efthimia Aivaloglou, George Fletcher, Michael Liut, and Daphne Miedema. 2023. Report on the First International Workshop on Data Systems Education (DataEd '22). *SIGMOD Rec.* 51, 4 (jan 2023), 49–53. https://doi.org/10.1145/3582302.3582314
[9] Vangel Ajanovski. 2020. Tools for Analysis of Curricula Evolution Across Computer Science Curriculum Guidelines. In *Proceedings of the 2020 ACM Conference on Innovation and Technology in Computer Science Education* (Trondheim, Norway) *(ITiCSE '20)*. Association for Computing Machinery, New York, NY, USA, 539–540. https://doi.org/10.1145/3341525.3393995
[10] Abdussalam Alawini, Peilin Rao, Leyao Zhou, Lujia Kang, and Ping-Che Ho. 2022. Teaching Data Models with TriQL. In *1st International Workshop on Data Systems Education*. 16–21. https://doi.org/10.1145/3531072.3535320
[11] Abdul Rahman R Alazmi and Abdul Aziz R Alazmi. 2012. Data mining and visualization of large databases. *International Journal of Computer Science and Security* 6, 5 (2012), 295–314.
[12] Ridha Alkhabaz, Zepei Li, Sophia Yang, and Abdussalam Alawini. 2023. 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*. ACM, Seattle WA USA, 30–36. https://doi.org/10.1145/3596673.3596976
[13] Jatin Ambasana, Sameer Sahasrabudhe, and Sridhar Iyer. 2023. SQL-Wordle: Gamification of SQL Programming Exercises. In *Proceedings of the ACM Conference on Global Computing Education Vol 2*. 190–190. https://doi.org/10.1145/3617650.3624949
[14] Mark Ardis, David Budgen, Gregory W. Hislop, Jeff Offutt, Mark Sebern, and Willem Visser. 2015. SE 2014: Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering. *Computer* 48, 11 (2015), 106–109. https://doi.org/10.1109/MC.2015.345
[15] Australian Computer Society Inc. 2021. ACS Core Body of Knowledge V3.2. https://www.acs.org.au/cpd-education/acs-accreditation-program.html Accessed: 2024-11-13.
[16] Erik Barendsen and Jos Tolboom. 2016. Advies examenprogramma informatica havo/vwo: Inhoud en invoering. (Feb. 2016). https://www.slo.nl/@4491/advies-0/
[17] Brett A. Becker. 2019. A Survey of Introductory Programming Courses in Ireland. In *Proceedings of the 2019 ACM Conference on Innovation and Technology in Computer Science Education*. ACM, Aberdeen Scotland Uk, 58–64. https://doi.org/10.1145/3304221.3319752
[18] Benjamin Samuel Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1964. *Taxonomy of educational objectives*. Vol. 2. Longmans, Green New York.
[19] Douglas B Bock and Susan E Yager. 2002. Improving entity relationship modeling accuracy with novice data modelers. *Journal of Computer Information Systems* 42, 2 (2002), 69–75. https://doi.org/10.1080/08874417.2002.11647489
[20] Miloš Bogdanović, Aleksandar Stanimirović, Nikola Davidović, and Leonid Stoimenov. 2008. The development and usage of a relational database design tool for educational purposes. In *Informing Science & IT Education Conference (InSITE'08)*. Citeseer, 251–258.
[21] Børne- og Undervisningsministeriet. 2024. Vejledning til Informatik C, hhx, htx, stx, hf. https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf23/aug/vejledninger/240807-vejledning-til-informatik-c--hhx--htx--stx--hf.pdf Accessed: 2024-11-13.
[22] Stephen Cass. 2022. SQL Should Be Your Second Language. *IEEE Spectrum* 59, 10 (2022), 20–21. https://doi.org/10.1109/MSPEC.2022.9915547
[23] Peter Pin-Shan Chen. 1976. The entity-relationship model—toward a unified view of data. *ACM Transactions on Database Systems (TODS)* 1, 1 (1976), 9–36. https://doi.org/10.1145/320434.320440
[24] 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
[25] Michael A Chilton, Roger McHaney, and Bongsug Chae. 2006. Data modeling education: The changing technology. *Journal of Information Systems Education* 17, 1 (2006), 17.
[26] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data* (San Francisco, California, USA) *(SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 2201–2206. https://doi.org/10.1145/2882903.2912574
[27] Thomas Connolly and Carolyn Begg. 2015. *Database Systems (6th. ed.)*. Pearson.
[28] Thomas M Connolly, Carolyn E Begg, et al. 2006. A constructivist-based approach to teaching database analysis and design. *Journal of Information Systems Education* 17, 1 (2006), 43.
[29] Iain J Cruickshank, Nathaniel D Bastian, Jean R.S. Blair, Christa M Chewar, and Edward Sobiesk. 2024. Seeing the Whole Elephant - A Comprehensive Framework for Data Education. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*. ACM, Portland OR USA, 248–254. https://doi.org/10.1145/3626252.3630922
[30] Kathryn Cunningham, Miranda C. Parker, and Jonathan Zhang. 2023. The Landscape of Computer Science Education Courses: A Syllabi Analysis. In *Koli '23: Koli Calling*. https://doi.org/10.1145/3631802.3631831
[31] Jonathan Danaparamita and Wolfgang Gatterbauer. 2011. QueryViz: Helping Users Understand SQL Queries and Their Patterns. In *Proceedings of the 14th International Conference on Extending Database Technology* (Uppsala, Sweden) *(EDBT/ICDT '11)*. Association for Computing Machinery, New York, NY, USA, 558–561. https://doi.org/10.1145/1951365.1951440
[32] Christopher John Date. 2003. *An introduction to database systems (8th. ed.)*. Pearson Education.
[33] Karen Collins Davis. 2022. Instructional Design for Teaching Relational Query Optimization to Undergraduates. In *1st International Workshop on Data Systems Education*. 44–50. https://doi.org/10.1145/3531072.3535325
[34] Sociedade Brazileira de Computação. [n. d.]. Referenciais de Formação para os Cursos de Graduação em Computação. https://www.sbc.org.br/documentos-da-sbc/summary/131-curriculos-de-referencia/1165-referenciais-de-formacao-para-cursos-de-graduacao-em-computacao-outubro-2017 Brazilian Curriculum Guidelines. Accessed: 2024-04-11.
[35] María de los Angeles Constantino-González and Daniel D Suthers. 2000. A coached collaborative learning environment for entity-relationship modeling. In *International Conference on Intelligent Tutoring Systems*. Springer, 324–333. https://doi.org/10.1007/3-540-45108-0_36
[36] République Tunisienne Ministère de l'Éducation Direction Generale du Cycle Preparatoire & de l'Enseignement Secondaire. 2008. Programmes d'Informatique Enseignement secondaire. http://www.edunet.tn/ressources/pedagogie/programmes/nouveaux_programme2011/secondaire/info.pdf Tunisian Curriculum Guidelines. Accessed: 2024-05-08.
[37] Suzanne Wagner Dietrich and Susan Urban. 2005. *An advanced course in database systems: beyond relational databases*. Pearson Education.
[38] Kavisha Duggal, Anukool Srivastav, and Satvinder Kaur. 2014. Gamified approach to database normalization. *International Journal of Computer Applications* 93, 4 (2014). https://doi.org/10.5120/16207-5505
[39] Mustafa Eid. 2012. A Learning System For Entity Relationship Modeling.. In *PACIS*. 152.
[40] Ramez Elmasri and Shamkant B. Navathe. 2016. *Fundamentals of Database Systems (7th. ed.)*. Pearson.
[41] EU Labour Force Survey (EU-LFS). 2023. *Employment Rates of Recent Graduates*. Accessed: 2024-07-06.
[42] Quality Indicators for Learning and Teaching (QILT). 2022. *2022 Graduate Outcomes Survey - National Report*. https://www.qilt.edu.au/docs/default-source/default-document-library/2022-gos-national-report.pdf?sfvrsn=c5d342c8_2 Australian Graduate Outcomes. Accessed: 2024-07-06.
[43] Office for National Statistics (ONS). 2023. *Graduate Labour Markets*. https://explore-education-statistics.service.gov.uk/find-statistics/graduate-labour-markets UK Graduate Outcomes. Accessed: 2024-07-06.
[44] ACM Data Science Task Force. 2021. *Computing competencies for undergraduate data science curricula*. Association for Computing Machinery, New York, NY, USA. ISBN: 9781450390606.
[45] The Joint ACM/AIS IS2020 Task Force. 2020. A Competency Model for Undergraduate Programs in Information Systems. https://www.acm.org/binaries/content/assets/education/curricula-recommendations/is2020.pdf
[46] Nadime Francis, Alastair Green, Paolo Guagliardo, Leonid Libkin, Tobias Lindaaker, Victor Marsault, Stefan Plantikow, Mats Rydberg, Petra Selmer, and Andrés Taylor. 2018. Cypher: An evolving query language for property graphs. In *Proceedings of the 2018 international conference on management of data*. 1433–1445. https://doi.org/10.1145/3183713.3190657
[47] Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. 2009. *Database systems: the complete book*. Pearson Education.
[48] Jason Gorman, Sebastian Gsell, and Chris Mayfield. 2014. Learning relational algebra by snapping blocks. In *Proceedings of the 45th ACM technical symposium on Computer science education*. 73–78. https://doi.org/10.1145/2538862.2538961
[49] Nataša Grgurina, Jos Spronk, Hans de Vries, and Martin Klein Tank. 2024. Kerndoelen digitale geletterdheid. https://www.slo.nl/publicaties/@23474/conceptkerndoelen-digitale-geletterdheid Dutch Curriculum Guidelines.
[50] Minzhe Guo, Kai Qian, and Li Yang. 2016. Hands-on labs for learning mobile and NoSQL database security. In *2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 2. IEEE, 606–607. https://doi.org/10.1109/COMPSAC.2016.126
[51] Lynne Hall and Adrian Gordon. 1998. A virtual learning environment for entity relationship modelling. In *Proceedings of the twenty-ninth SIGCSE technical symposium on Computer science education*. 345–349. https://doi.org/10.1145/273133.274327
[52] Erika Hernández-Rubio, Marco Antonio Rodríguez-Torres, Humberto Vázquez-Santiago, and Amilcar Meneses-Viveros. 2023. Learning System for Relational Algebra. In *International Conference on Human-Computer Interaction*. Springer, 54–63. https://doi.org/10.1007/978-3-031-34411-4_5
[53] Higher Education Statistics Agency (HESA). 2024. *Graduate Outcomes 2021/22: Summary Statistics: Data for 2022*. UK Graduate Outcomes. Accessed: 2024-07-06.

[54] Higher Education Statistics Agency. [n. d.]. Definitions: Graduates. https://www.hesa.ac.uk/support/definitions/graduates. Accessed: 2024-07-05.

[55] Jeffrey A Hoffer, Venkataraman Ramesh, and Heikki Topi. 2011. *Modern database management.* Upper Saddle River, NJ: Prentice Hall.

[56] Jeffrey A Hoffer, Heikki Topi, and Venkataraman Ramesh. 2014. *Essentials of Database Management.* Pearson Education.

[57] Tauqeer Hussain. 2016. Teaching entity-relationship models effectively. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI).* IEEE, 264–269. https://doi.org/10.1109/CSCI.2016.0058

[58] Jozef Hvoreckỳ, Martin Drlík, and Michal Munk. 2010. Enhancing database querying skills by choosing a more appropriate interface. In *IEEE EDUCON 2010 Conference.* IEEE, 1897–1905. https://doi.org/10.1109/EDUCON.2010.5492434

[59] Musa J Jafar, Jeffry Stephen Babb, and Amjad Abdullat. 2017. Emergence of data analytics in the information systems curriculum. *Information Systems Education Journal* 15, 5 (2017), 22.

[60] Daniel A. Keim. 1996. Databases and Visualization. *ACM SIGMOD Record* 25, 2 (1996). https://doi.org/10.1145/235968.280349

[61] Michael Kifer, Arthur J Bernstein, and Philip M Lewis. 2005. *Database systems: an application-oriented approach.* Pearson Education.

[62] Paul J Kovacs and Jeanne M Baugh. 2009. Merging object-oriented programming, database design, requirements analysis, and web technologies in an active learning environment. *Information Systems Education Journal* 7, 52 (2009), 1–8.

[63] David Kroenke and David J. Auer. 2016. *Database Processing: Fundamentals, Design, and Implementation (14th. ed.).* Pearson Education.

[64] Amruth N. Kumar, Rajendra K. Raj, Sherif G. Aly, Monica D. Anderson, Brett A. Becker, Richard L. Blumenthal, Eric Eaton, Susan L. Epstein, Michael Goldweber, Pankaj Jalote, Douglas Lea, Michael Oudshoorn, Marcelo Pias, Susan Reiser, Christian Servin, Rahul Simha, Titus Winters, and Qiao Xiang. 2024. *Computer Science Curricula 2023.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3664191

[65] Sam Lau, Sean Kross, Eugene Wu, and Philip J. Guo. 2023. 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* (Seattle, WA, USA) *(DataEd '23).* Association for Computing Machinery, New York, NY, USA, 50–55. https://doi.org/10.1145/3596673.3596972

[66] Zepei Li, Sophia Yang, Kathryn Cunningham, and Abdussalam Alawini. 2023. Assessing Student Learning Across Various Database Query Languages. In *2023 IEEE Frontiers in Education Conference (FIE).* 1–9. https://doi.org/10.1109/FIE58773.2023.10343409

[67] Lightcast. [n. d.]. About Lightcast Data. https://lightcast.io/about/data. Accessed: 2024-07-05.

[68] Siyuan Liu, Sourav S Bhowmick, Wanlu Zhang, Shu Wang, Wanyi Huang, and Shafiq Joty. 2019. Neuron: Query execution plan meets natural language processing for augmenting DB education. In *Proceedings of the 2019 International Conference on Management of Data.* 1953–1956. https://doi.org/10.1145/3299869.3320213

[69] Matias Lopez, Sebastian Ferrada, and Aidan Hogan. 2024. ERDoc: A Web Interface for Entity-Relation Modelling. In *Proceedings of the 3rd International Workshop on Data Systems Education: Bridging Education Practice with Education Research* (Santiago, AA, Chile) *(DataEd '24).* Association for Computing Machinery, New York, NY, USA, 7–12. https://doi.org/10.1145/3663649.3664372

[70] Jiebing Ma, Sourav S Bhowmick, Lester Tay, and Byron Choi. 2024. SIERRA: A Counterfactual Thinking-based Visual Interface for Property Graph Query Construction. In *Companion of the 2024 International Conference on Management of Data.* 440–443. https://doi.org/10.1145/3626246.3654729

[71] Francesco Maiorana. 2014. Teaching web programming-an approach rooted in database principles. In *International Conference on Computer Supported Education*, Vol. 2. SCITEPRESS, 49–56. https://doi.org/10.5220/0004849300490056

[72] Brandeis Marshall and Susan Geier. 2020. Cross-Disciplinary Faculty Development in Data Science Principles for Classroom Integration. In *Proceedings of the 51st ACM Technical Symposium on Computer Science Education.* ACM, Portland OR USA, 1207–1213. https://doi.org/10.1145/3328778.3366801

[73] Raina Mason and Simon. 2017. Introductory Programming Courses in Australasia in 2016. In *Proceedings of the Nineteenth Australasian Computing Education Conference.* ACM, Geelong VIC Australia, 81–89. https://doi.org/10.1145/3013499.3013512

[74] Raina Mason, Simon, Brett A. Becker, Tom Crick, and James H. Davenport. 2024. A Global Survey of Introductory Programming Courses. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (, Portland, OR, USA,) *(SIGCSE 2024).* Association for Computing Machinery, New York, NY, USA, 799–805. https://doi.org/10.1145/3626252.3630761

[75] Victor M. Matos and Rebecca Grasser. 2002. A Simpler (and Better) SQL Approach to Relational Division. *Journal of Information Systems Education* 13, 2 (2002), 85–88. https://aisel.aisnet.org/jise/vol13/iss2/2

[76] Yoshitatsu Matsuda, Takayuki Sekiya, and Kazunori Yamaguchi. 2018. Curriculum Analysis of Computer Science Departments by Simplified, Supervised LDA. *Journal of Information Processing* 26 (2018), 497–508. https://doi.org/10.2197/ipsjjip.26.497

[77] Gonzalo Méndez, Xavier Ochoa, and Katherine Chiluiza. 2014. Techniques for data-driven curriculum analysis. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (Indianapolis, Indiana, USA) *(LAK '14).* Association for Computing Machinery, New York, NY, USA, 148–157. https://doi.org/10.1145/2567574.2567591

[78] Xiangrui Meng, Joseph Bradley, Burak Yavuz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, DB Tsai, Manish Amde, Sean Owen, et al. 2016. Mllib: Machine learning in apache spark. *Journal of Machine Learning Research* 17, 34 (2016), 1–7.

[79] 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).* IEEE, 1–9. https://doi.org/10.1109/VL/HCC51201.2021.9576431

[80] Daphne Miedema, George Fletcher, and Efthimia Aivaloglou. 2022. Expert Perspectives on Student Errors in SQL. *ACM Transactions on Computing Education (TOCE)* (2022). https://doi.org/10.1145/3551392

[81] Daphne Miedema, Toni Taipalus, and Efthimia Aivaloglou. 2023. Students' Perceptions on Engaging Database Domains and Structures. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1.* 122–128. https://doi.org/10.1145/3545945.3569727

[82] Koby Mike. 2020. Data Science Education: Curriculum and pedagogy. In *Proceedings of the 2020 ACM Conference on International Computing Education Research.* ACM, Virtual Event New Zealand, 324–325. https://doi.org/10.1145/3372782.3407110

[83] Ministère de l'Éducation nationale et de la Jeunesse. 2019. Programme de numérique et sciences informatiques de terminale générale. https://cache.media.education.gouv.fr/file/SPE8_MENJ_25_7_2019/93/3/spe247_annexe_1158933.pdf French Curriculum Guidelines. Accessed: 2024-11-13.

[84] Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. 2020. Gamification and SQL: an empirical study on student performance in a database course. *ACM Transactions on Computing Education (TOCE)* 21, 1 (2020), 1–29. https://doi.org/10.1145/3427597

[85] Ellen Murphy, Tom Crick, and James H. Davenport. 2017. An Analysis of Introductory Programming Courses at UK Universities. *The Art, Science, and Engineering of Programming* 1, 2 (April 2017), 18. https://doi.org/10.22152/programming-journal.org/2017/1/18

[86] Pamela Neely. 2007. Mastery level learning and the art of database design. *AMCIS 2007 Proceedings* (2007), 1.

[87] Andy Nguyen, Lesley Gardner, and Don Sheridan. 2020. Data analytics in higher education: An integrated view. *Journal of Information Systems Education* 31, 1 (2020), 61.

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

[89] Joint Task Force on Cybersecurity Education. 2018. *Cybersecurity Curricula 2017: Curriculum Guidelines for Post-Secondary Degree Programs in Cybersecurity.* Association for Computing Machinery, New York, NY, USA.

[90] Task Group on Information Technology Curricula. 2017. *Information Technology Curricula 2017.* Association for Computing Machinery (ACM). Accessed: 2024-07-06.

[91] Beatriz Pérez. 2021. Enhancing the learning of database access programming using continuous integration and aspect oriented programming. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET).* IEEE, 221–230. https://doi.org/10.1109/ICSE-SEET52601.2021.00032

[92] Rubén Pérez-Mercado, Antonio Balderas, Andrés Muñoz, Juan Francisco Cabrera, Manuel Palomo-Duarte, and Juan Manuel Dodero. 2023. ChatbotSQL: Conversational agent to support relational database query language learning. *SoftwareX* 22 (2023), 101346. https://doi.org/10.1016/j.softx.2023.101346

[93] Wagner De A Perin, Davidson Cury, Crediné S De Menezes, and Camila Z De Aguiar. 2023. Active Learning with Concept Maps: Enhancing Understanding of Entity-Relationship Diagrams in Database Modeling. In *2023 IEEE Frontiers in Education Conference (FIE).* IEEE, 1–9. https://doi.org/10.1109/FIE58773.2023.10343471

[94] Putsadee Pornphol and Suphamit Chittayasothorn. 2024. Using LLM Artificial Intelligence Systems as Complex SQL Programming Assistants. In *2024 12th International Conference on Information and Education Technology (ICIET).* 477–481. https://doi.org/10.1109/ICIET60671.2024.10542806

[95] Hidayah Rahmalan, Sharifah Sakinah Syed Ahmad, and Lilly Suriani Affendey. 2020. Investigation on designing a fun and interactive learning approach for Database Programming subject according to students' preferences. In *Journal of Physics: Conference Series*, Vol. 1529. IOP Publishing, 022076. https://doi.org/10.1088/1742-6596/1529/2/022076

[96] Raghu Ramakrishnan and Johannes Gehrke. 2002. *Database management systems.* McGraw-Hill, Inc.

[97] Rami Rashkovits and Ilana Lavy. 2021. Mapping Common Errors in Entity Relationship Diagram Design of Novice Designers. *International Journal of*

*Database Management Systems* 13, 1 (2021), 1–19. https://doi.org/10.5121/ijdms.2021.13101

[98] Phyllis Reisner. 1981. Human factors studies of database query languages: A survey and assessment. *ACM Computing Surveys (CSUR)* 13, 1 (1981), 13–31. https://doi.org/10.1145/356835.356837

[99] Greg Riccardi. 2000. *Principles of database systems with Internet and Java applications.* Addison-Wesley Longman Publishing Co., Inc. https://doi.org/10.5555/556924

[100] James Rumbaugh, Ivar Jacobson, and Grady Booch. 2004. *Unified Modeling Language Reference Manual, The (2nd Edition).* Pearson Higher Education.

[101] Mariam Salloum, Daniel Jeske, Wenxiu Ma, Vagelis Papalexakis, Christian Shelton, Vassilis Tsotras, and Shuheng Zhou. 2021. Developing an Interdisciplinary Data Science Program. In *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education.* ACM, Virtual Event USA, 509–515. https://doi.org/10.1145/3408877.3432454

[102] Johannes Schildgen. 2014. SQL island: An adventure game to learn the database language SQL. In *Proceedings of the 8th European Conference on Games Based Learning (ECGBL 2014).* 137–138.

[103] Johannes Schildgen. 2020. MonstER Park–The Entity-Relationship-Diagram Learning Game. ER Forum.

[104] Johannes Schildgen and Jessica Rosin. 2022. Game-based Learning of SQL Injections. In *1st International Workshop on Data Systems Education.* 22–25. https://doi.org/10.1145/3531072.3535321

[105] Takayuki Sekiya, Yoshitatsu Matsuda, and Kazunori Yamaguchi. 2015. Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge* (Poughkeepsie, New York) *(LAK '15).* Association for Computing Machinery, New York, NY, USA, 330–339. https://doi.org/10.1145/2723576.2723594

[106] Shin-Shing Shin. 2020. Structured query language learning: Concept map-based instruction based on cognitive load theory. *IEEE Access* 8 (2020), 100095–100110. https://doi.org/10.1109/ACCESS.2020.2997934

[107] Shin-Shing Shin. 2022. Teaching Method for Entity–Relationship Models Based on Semantic Network Theory. *IEEE Access* 10 (2022), 94908–94923. https://doi.org/10.1109/ACCESS.2022.3206028

[108] John B. Smelcer. 1995. User errors in database query composition. *International Journal of Human-Computer Studies* 42, 4 (1995), 353–381. https://doi.org/10.1006/ijhc.1995.1017

[109] Hector Suarez and Hooper Kincannon. 2017. SSETGami: Secure software education through gamification. *KSU Proceedings on Cybersecurity Education, Research and Practice* (2017).

[110] Pramuditha Suraweera and Antonija Mitrovic. 2002. KERMIT: A constraint-based tutor for database modeling. In *Intelligent Tutoring Systems: 6th International Conference (ITS).* Springer, 377–387. https://doi.org/10.1007/3-540-47987-2_41

[111] Toni Taipalus. 2020. The effects of database complexity on SQL query formulation. *Journal of Systems and Software* 165 (2020), 110576. https://doi.org/10.1016/j.jss.2020.110576

[112] Toni Taipalus. 2020. Explaining Causes behind SQL Query Formulation Errors. *Proceedings - Frontiers in Education Conference, FIE* 2020-Octob (2020). https://doi.org/10.1109/FIE44824.2020.9274114

[113] Toni Taipalus. 2023. Query Execution Plans and Semantic Errors: Usability and Educational Opportunities. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23).* Association for Computing Machinery, New York, NY, USA, Article 239, 6 pages. https://doi.org/10.1145/3544549.3585794

[114] 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

[115] Toni Taipalus and Hilkka Grahn. 2023. Framework for SQL Error Message Design: A Data-Driven Approach. *ACM Transactions on Software Engineering and Methodology* 33, 1 (2023), 1–50. https://doi.org/10.1145/3607180

[116] Toni Taipalus and Hilkka Grahn. 2024. Building Blocks Towards More Effective SQL Error Messages. In *Proceedings of the 2024 Innovation and Technology in Computer Science Education V.1 (ITiCSE 2024).* ACM.

[117] Toni Taipalus, Daphne Miedema, and Efthimia Aivaloglou. 2023. Engaging Databases for Data Systems Education. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023).* Association for Computing Machinery, New York, NY, USA, 334–340. https://doi.org/10.1145/3587102.3588804

[118] Toni Taipalus and Ville Seppänen. 2020. SQL education: A systematic mapping study and future research agenda. *ACM Transactions on Computing Education (TOCE)* 20, 3 (2020), 1–33. https://doi.org/10.1145/3398377

[119] Toni Taipalus, Mikko Siponen, and Tero Vartiainen. 2018. Errors and Complications in SQL Query Formulation. *ACM Transactions on Computing Education* 18, 3 (2018). https://doi.org/10.1145/3231712

[120] Jess Tan, Desmond Yeo, Rachael Neoh, Huey-Eng Chua, and Sourav S Bhowmick. 2022. MOCHA: a tool for visualizing impact of operator choices in query execution plans for database education. *Proceedings of the VLDB Endowment* 15, 12 (2022), 3602–3605. https://doi.org/10.14778/3554821.3554854

[121] The Joint Task Force on Computing Curricula. 2015. *Curriculum Guidelines for Undergraduate Degree Programs in Software Engineering.* Technical Report. New York, NY, USA. https://doi.org/10.1145/2965631

[122] Jos Tolboom, Jenneke Krüger, and Nataša Grgurina. 2014. Informatica in de bovenbouw havo/vwo. https://www.slo.nl/zoeken/@4251/informatica/ Dutch Curriculum Guidelines.

[123] Heikki Topi, Kate M. Kaiser, Janice C. Sipior, Joseph S. Valacich, J. F. Nunamaker, G. J. de Vreede, and Ryan Wright. 2010. *Curriculum Guidelines for Undergraduate Degree Programs in Information Systems.* Technical Report. New York, NY, USA. https://doi.org/10.1145/2593310

[124] United Nations Statistics Division. 2021. Manual on Principal Indicators for Business and Trade Statistics. https://unstats.un.org/unsd/business-stat/UNCEBTS/Manual_on_Principal_Indicators_Final_White-Cover_Version.pdf. Accessed: 2024-07-05.

[125] United Nations Statistics Division. 2024. International Standard Industrial Classification of All Economic Activities (ISIC). https://unstats.un.org/unsd/classifications/Econ/ISIC.cshtml. Accessed: 2024-07-05.

[126] King Saud University. [n. d.]. https://cs.ksu.edu.sa/sites/ccis.ksu.edu.sa/files/users/user984/guides/CS_Bachelor_English_2023.pdf Saudi Curriculum Guidelines. Accessed: 2024-05-08.

[127] Anikó Vágner. 2014. Let's learn database programming in an active way. *Teaching Mathematics and Computer Science* 12 (2014). Issue 2.

[128] Hu Wang, Hui Li, Sourav S Bhowmick, and Baochao Xu. 2023. ARENA: Alternative Relational Query Plan Exploration for Database Education. In *Companion of the 2023 International Conference on Management of Data.* 107–110. https://doi.org/10.1145/3555041.3589713

[129] Weiguo Wang, Sourav S Bhowmick, Hui Li, Shafiq Joty, Siyuan Liu, and Peng Chen. 2021. Towards enhancing database education: Natural language generation meets query execution plans. In *Proceedings of the 2021 International Conference on Management of Data.* 1933–1945. https://doi.org/10.1145/3448016.3452822

[130] Patricia Ward. 2008. *Database Management System (2nd. ed.).* Pearson Education.

[131] Richard T Watson. 2008. *Data management, databases and organizations.* John Wiley & Sons.

[132] Lizette Weilbach, Marié Hattingh, and Komla Pillay. 2021. Using Design Patterns to Teach Conceptual Entity Relationship (ER) Data Modelling. In *International Conference on Innovative Technologies and Learning.* Springer, 228–238. https://doi.org/10.1007/978-3-030-91540-7_25

[133] Anthony KL Wong, Michael Morgan, and Matthew Butler. 2012. Designing a Technology Enhanced Collaborative Space for Learning Entity-Relationship Modeling. In *2012 IEEE 12th International Conference on Advanced Learning Technologies.* IEEE, 213–217. https://doi.org/10.1109/ICALT.2012.240

[134] Sophia Yang, Zepei Li, Geoffrey L. Herman, Kathryn Cunningham, and Abdussalam Alawini. 2023. Uncovering Patterns of SQL Errors in Student Assignments: A Comparative Analysis of Different Assignment Types. In *2023 IEEE Frontiers in Education Conference (FIE).* 01–09. https://doi.org/10.1109/FIE58773.2023.10343207

# Appendix

## A Data systems topics covered in curriculum guidelines

We analyzed all undergraduate level curriculum guidelines published by ACM, and marked whether the topics therein were directly or tangentially related to data systems topics. In the tables below, we list topics proposed in each curriculum guideline and how those topics are related to data systems. Topics directly related to data systems are presented in **bold text**, topics tangentially related are presented as plain text, and topics unrelated are presented as gray.

### A.1 CS2023

*Artificial Intelligence:* Fundamental Issues; Search; Fundamental Knowledge Representation and Reasoning; Machine Learning; Applications and Societal Impact; Logical Representation and Reasoning; Probabilistic Representation and Reasoning; Planning; Agents and Cognitive Systems; Natural Language Processing; Robotics; Perception and Computer Vision.

***Data Management:*** **The Role of Data and the Data Life Cycle**; **Core Database System Concepts**; **Data Modeling**; **Relational Databases**; **Query Construction**; **Query Processing**; **DBMS Internals**; **NoSQL Systems**; **Data Security and Privacy**; **Data Analytics**; **Distributed Databases/Cloud Computing**; **Semi-structured and Unstructured Databases**; **Society, Ethics, and the Profession**.

*Human-Computer Interaction:* Understanding the User Individual goals and interactions with others; Accountability and Responsibility in Design; Accessibility and Inclusive Design; Evaluating the Design; System Design; Society, Ethics, and the Profession.

*Software Engineering:* Teamwork; Tools and Environments; Product Requirements; Software Design; Software Construction; Software Verification and Validation; Refactoring and Code Evolution; Software Reliability; Formal Methods.

*Security:* Foundational Security; Society, Ethics, and the Profession; Secure Coding; Cryptography; Security Analysis, Design, and Engineering; Digital Forensics; Security Governance.

*Society, Ethics, and the Profession:* Social Context; Methods for Ethical Analysis; Professional Ethics; Intellectual Property; Privacy and Civil Liberties; Communication; Sustainability; Computing History; Economies of Computing; Security Policies, Laws and Computer Crimes; Diversity, Equity, Inclusion, and Accessibility.

*Specialized Platform Development:* Common Aspects/Shared Concerns; Web Platforms; Mobile Platforms; Robot Platforms; Embedded Platforms; Game Platforms; Interactive Computing Platforms; SEP/Mobile; SEP/Web; SEP/Game; SEP/Robotics; SEP/Interactive.

### A.2 CCDS2021

*Analysis and Presentation:* Foundational considerations; Visualization; User-centered design; Interaction design; Interface design and development.

*Artificial Intelligence:* General; Knowledge representation and reasoning – logic based; Knowledge representation and reasoning – probability based; Planning and search strategies.

***Big Data Systems:*** **Problems of scale**; **Big data computing architectures**; **Parallel computing frameworks**; **Distributed data storage**; **Parallel programming**; **Techniques for Big Data applications**; **Cloud computing**; **Complexity theory**; **Software support for Big Data applications**.

*Computing and Computer Fundamentals:* Basic computer architecture; Storage systems fundamentals; Operating system basics; File systems; Networks; The web and web programming; Compilers and interpreters.

***Data Acquisition, Management, and Governance:*** **Data acquisition**; **Information extraction**; **Working with various types of data**; **Data integration**; **Data reduction and compression**; **Data transformation**; **Data cleaning**; **Data privacy and security**.

*Data Mining:* Proximity measurement; **Data preparation**; **Information extraction**; Cluster analysis; Classification and regression; Pattern mining; Outlier detection; **Time series data**; Mining web data; **Information retrieval**.

*Data Privacy, Security, Integrity, and Analysis for Security:* Data privacy; Data security; Data integrity; Analysis for security.

*Machine learning:* General; Supervised learning; Unsupervised learning; Mixed methods; Deep learning.

*Professionalism:* Continuing professional development; Communication; Teamwork; Economic considerations; Privacy and confidentiality; Ethical considerations; Legal considerations; Intellectual property; On automation.

*Programming, data structures and algorithms:* Algorithmic thinking and problem solving; Programming; Data structures; Algorithms; Basic complexity analysis; Numerical computing.

### A.3 IS2020

*Foundations:* Foundations of Information Systems.

*Data:* Data and Information Management (including Databases); Data and Business Analytics (including Data Mining, AI, BI); Data and Information Visualization.

*Technology:* IT Infrastructure (including Networking, Cloud); Secure Computing; Emerging Technologies (IOT, blockchain).

*Development:* Systems Analysis and Design; Application Development and Programming; Object-Oriented Paradigm; Web Development; Mobile Development; User Interface Design.

*Organizational Domain:* Ethics, use and implications for society; IS Management and Strategy; Digital Innovation; Business Process Management.

*Integration:* IS Project Management; IS Practicum.

***Information Management:*** **Perspectives and impact**; **Data-information concepts**; **Data modeling**; **Database query languages**; **Data organization architecture**; **Special-purpose databases**; **Managing the database environment**.

***Integrated Systems Technology:*** Perspectives and impact; **Data mapping and exchange**; Intersystem communication protocols; Integrative programming; Scripting techniques; Defensible integration.

*System Paradigms:* Perspectives and impact; Requirements; System architecture; Acquisition and sourcing; Testing and quality assurance; **Integration and deployment**; System governance; **Operational activities**; Operational domains; **Performance analysis**.

*Software Fundamentals:* Perspectives and impact; Concepts and techniques; Problem-solving strategies; Program development; Fundamental data structures; Algorithm principles and development; Modern app programming practices.

## A.4 CSEC2017

*Data Security:* Cryptography; Digital Forensics; Data Integrity and Authentication; Access Control; Secure Communication Protocols; Cryptanalysis; Data Privacy; Information Storage Security.

*System Security:* System Thinking; System Management; System Access; System Control; System Retirement; System Testing; Common System Architectures.

*Human Security:* Identity Management; Social Engineering; Personal Compliance with Cybersecurity Rules/Policy/Ethical Norms; Awareness and Understanding; Social and Behavioral Privacy; Personal Data Privacy and Security; Usable Security and Privacy.

*Organizational Security:* Risk Management; Security Governance and Policy; Analytical Tools; Systems Administration; Cybersecurity Planning; Business Continuity, Disaster Recovery, and Incident Management; Security Program Management; Personnel Security; Security Operations.

*Societal Security:* Cybercrime; Cyber Law; Cyber Ethics; Cyber Policy; Privacy.

*Software Design:* History and overview; Relevant tools, standards, and/or engineering constraints; Programming constructs and paradigms; Problem-solving strategies; Data structures; Recursion; Object-oriented design; Software testing and quality; Data modeling; Database systems; Event-driven and concurrent programming; Using application programming interfaces; Data mining; Data visualization.

## A.5 SE2014

*Computing essentials:* Computer science foundations; Construction technologies; Construction tools.

*Software modeling and analysis:* Modeling foundations; **Types of models**; Analysis fundamentals.

*Software design:* Design concepts; Design strategies; Architectural design; Human-computer interaction design; Detailed design; Design evaluation.