Integrating Data Science into Undergraduate Science and Engineering Courses: Lessons Learned by Instructors in a Multi-University Research Practice Partnership

Yunus Naseri, Caitlin Snyder, Katherine X. Pérez-Rivera, Sambridhi Bhandari, Habtamu Alemu Workneh, Niroj Aryal, Gautam Biswas, Erin C. Henrick, Erin R. Hotchkiss, Manoj K. Jha, Steven Jiang, Emily C. Kern, Vinod K. Lohani, Landon T. Marston, Christopher P. Vanags, and Kang Xia

***Abstract*—Contribution: This paper discusses a research-practice partnership (RPP) where instructors from six undergraduate courses in three universities developed data science modules tailored to the needs of their respective disciplines, academic levels, and pedagogies.**

***Background:* STEM disciplines at universities are incorporating data science topics to meet employer demands for data science-savvy graduates. Integrating these topics into regular course materials can benefit students and instructors. However, instructors encounter challenges in integrating data science instruction into their course schedules.**

***Research Questions:* How did instructors from multiple engineering and science disciplines working in an RPP integrate data science into their undergraduate courses?**

***Methodology:* A multiple case study approach, with each course as a unit of analysis, was used to identify data science topics and integration approaches.**

***Findings:* Instructors designed their modules to meet specific course needs, utilizing them as primary or supplementary learning tools based on their course structure and pedagogy. They selected a subset of discipline-agnostic data science topics, such as generating and interpreting visualizations and conducting basic statistical analyses. Although instructors faced challenges due to varying data science skills of their students, they valued the control they had in integrating data science content into their courses. They were uncertain about whether the modules could be adopted for use by other instructors, specifically by those outside of their discipline, but they all believed the approach for developing and integrating data science could be adapted to student needs in different situations.**

Keywords—Data science integration, data literacy, modular approach, STEM, research-practice partnership.

This research is supported by NSF grants #1915538, #1915487, #1915268, and #2144169.

At Virginia Tech, M. Yunus Naseri (mohammadyunusn@vt.edu), and Landon T. Marston are in the Department of Civil and Environmental Engineering; Vinod. K Lohani is in the Department of Engineering Education; Katherine X. Pérez-Rivera and Erin R. Hotchkiss are in the Department of Biological Sciences; and Kang Xia is in the Department of Environmental Science. At Vanderbilt University, Gautam Biswas and Caitlin Snyder are in the Department of Computer Science; Christopher P. Vanags is in the Department of Earth and Environmental Sciences; and Erin C. Henrick is in the Department of Leadership Policy and Organization; along with Erin Henrick, Emily C. Kern is a part of Partner to Improve. At North Carolina A&T State University, Sambridhi Bhandari and Manoj K. Jha are in the Department of Civil, Architectural and Environmental Engineering; Habtamu Alemu Workneh is in the Department of Applied Science and Technology; Niroj Aryal is in the Department of Natural Resources and Environmental Design; and Steven Jiang is in the Department of Industrial and Systems Engineering.

I. Introduction

DATA SCIENCE LITERACY is becoming increasingly important in undergraduate education across STEM disciplines [1]. In particular, engaging students to the main components of the data science life cycle (data acquisition, pre-processing, visualization, and analysis [2]) allows students to gain a better understanding of the processes used to collect, process, visualize, and analyze data within their area of study. Embedding data science instruction [3], [4] into undergraduate courses can increase student competence and experience with analytical tools [5], [6]. However, instructors face a variety of challenges when integrating data science concepts and applications into their courses, which include already full curricula and supporting students with a wide range of backgrounds and familiarity with data science [7]. While previous research has led to the development of instructional data science materials in specific domains [8], [9], principles for integrating data science instruction across STEM domains are not well-established [7]. This paper presents the process used and lessons learned by instructors who integrated data science into their science and engineering courses while participating in a multi-university research practice partnership (RPP).

In this project, six STEM instructors from three universities collaborated to integrate data science into their courses by developing discipline-specific data science modules (these modules can be found at ds4stem.org). The instructors came from different backgrounds and had differing levels of comfort and experience teaching data science. The courses they taught differed in academic level, student background, and instructional modality. Each instructor had the freedom to design data science modules tailored to their course content, syllabus, and student learning goals. Instructors focused on providing their students data science learning opportunities, leveraging real-world data sets situated within their particular discipline, and developing modules that others may use in the future. Each instructor developed one to three data science modules for their course, implemented them one to three times during the project, and refined them through implementation and discussions in the RPP. A total of 12 modules were developed and implemented in six different STEM courses.

The primary aim of this paper is not to disseminate an undergraduate data science curriculum. Instead, its focus is on sharing insights gained in an RPP and providing resources that can be adopted, adapted, or used as models. Consequently, the paper focuses on presenting instructor

perspectives on integrating data science modules into existing STEM courses. This work has implications for other STEM instructors who are interested in integrating data science into their courses, university leaders who want to foster this type of collaboration, and interested education researchers and policy makers.

This paper aims to answer the following overarching research question (RQ): *How did instructors from multiple engineering and science disciplines working in an RPP integrate data science into their undergraduate courses?* This RQ can be divided into the following sub-research questions: (RQ1) *How did participating college instructors integrate data science topics into their curricula?* (RQ2) *How were instructors' integration approaches influenced by various factors, including course formats, disciplines, and academic levels*? (RQ3) *What data science topics did participating college instructors select to integrate into their curricula*? (RQ4) *What were the instructors' perspectives about their experiences integrating data science topics in their courses and how other instructors could use their data science modules*?

More specifically, the paper discusses how instructors assembled, formatted, and organized data science learning content into modules and produced assessments and assessment rubrics within the RPP through an iterative development and implementation process. The RPP collected data, including modules, course documents, instructor interviews and surveys, and student surveys to inform this process. The larger research project was run in an RPP, and a qualitative multiple case study approach was used in this paper to analyze the data. To answer RQ1 and RQ2, each module was characterized to provide a comprehensive understanding of how various instructors integrated data science into their curricula using factors such as instructor role, module length, activity, and assessment types. To answer RQ3, the assessments used in each module were analyzed to identify the data science topics instructors identified as important and relevant to their courses. To answer RQ4, instructor interviews were analyzed to gain a better understanding of the practical choices they made while integrating data science into their courses.

II. Background

Learning data science concepts and practices is recognized as important in STEM education to prepare students for future careers that require data literacy and data science skills. This is reflected in new course offerings and updates to core courses in different fields [9], such as courses that engage students with real-world datasets [10] and elective and required data science-focused courses like "Concepts in Computing with Data" [11] and "Data Science in Practice" [12]. These data science courses expose students to basic data science concepts, such as data cleanup, visualization, and reporting. While such courses are typically offered by statistics departments, most enrolled students are from other departments [9]. Students recognize that data science familiarity is important across multiple disciplines. Through RPPs, the National Science Foundation (NSF) has also been funding undergraduate and K-12 initiatives that expose students to data science concepts [13].

Data science literacy is typically introduced through two approaches: (1) standalone general and core disciplinary courses, immersive data science degrees, minors, certificates, and massive open online courses (MOOCs); and (2) integration within existing disciplinary courses. In the standalone method, students often struggle to apply skills to their disciplinary context [14], [15] [16], [17]. The integrated approach provides a more sustainable, evidence-based, and efficient method for introducing data science literacy to students [3], [4]. This approach can facilitate both learning and application of data science skills to bridge the data science instructional gap [16] [18], [19]. Moreover, integrated approaches align with learning theory principles that posit the learning process can be facilitated if the topics build on students' previous experience and knowledge [3], [20], [21]. NSF argues that acquiring a basic data literacy education is a requirement for all undergraduates [1]; therefore, both parallel and integrated methods should be promoted at academic institutions since industry needs data literate graduates with different skill levels.

Despite the importance of data literacy education in STEM disciplines, specifically in engineering and science, there is a lack of consensus on what data science is [22] and how to teach it [23]. While there has been research on its integration and education in single (e.g., [5], [12], [24]–[26]) or multiple (e.g., [4], [6], [27]) STEM and non-STEM disciplines, prior efforts were stand-alone parallel approaches that isolated data literacy from disciplinary context. Also, studies are often conducted in single disciplines that leave a gap in developing common principles for data science integration into STEM disciplines [7].

In this study, STEM instructors selected data science topics and their learning objectives and developed them as *modules* using common integration principles and a multi-disciplinary, bottom-up, and discipline-embedded process. In the RPP, data science-focused modules were developed across the STEM courses for the three universities: (1) North Carolina A&T (NC A&T), (2) Vanderbilt University (VU), and (3) Virginia Tech (VT). The developed modules were used by over 800 students.

III. Research Practice Partnership (RPP)

As long-term mutually beneficial formalized collaborations between researchers and practitioners, RPPs are a promising strategy to address complex problems of practice in education, capitalize on the expertise of various professionals, and produce more relevant and useful research [28]. This project developed a four-year partnership between university instructors of six different undergraduate courses, learning science and engineering education researchers, graduate and undergraduate students, and program evaluators, with input from industry professionals, all seeking to learn how to integrate data science across multiple STEM disciplines and courses. The RPP valued the expertise and experience of the instructors with a goal "to build and study solutions at the same time in real world contexts" [29]. It was collaborative, iterative, and based on the realities of implementation as the group engaged in cycles to develop, implement, and refine data science modules for their courses.

The RPP structure is illustrated in Fig. 1. The initial plan was for the partners to develop a set of common data science modules using real-world datasets, implement them across STEM disciplines, and compare student learning across courses and universities. The first in-person workshop held in November 2019 identified some challenges with developing common modules to be implemented in the upcoming semester: variations in student academic levels (sophomores to seniors) and student background knowledge of data science, an ambitious timeline (i.e., six weeks until the beginning of the semester) with tight academic schedules and

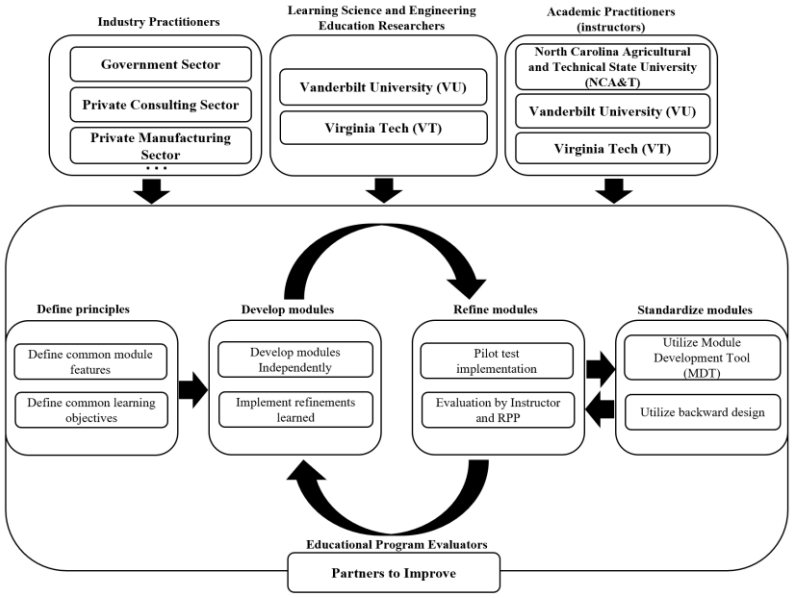

Fig. 1. RPP structure

packed syllabi, and instructors' varying comfort levels teaching data science.

One hallmark of an effective RPP is to flexibly respond to the needs and realities of the practice partners [30]. Given the challenges with creating common modules, the RPP adopted a bottom-up approach where each instructor developed their own data science module to teach in Spring 2020 based on their own interests, needs, and course context. The goal was to standardize the modules after they were developed. Instructors were asked to develop their modules to include assessments, lessons, and activities for their courses around common learning objectives that the team had agreed to adopt across all courses, e.g., use and analyze high-frequency, real-world data and evaluate the efficacy of the data collection system. Each instructor established more specific student learning goals for their courses independent of other instructors in the project. In this way, instructors had individual control over how data science topics were integrated into their courses.

During monthly meetings, the RPP discussed module development and implementation, especially as it related to the onset of the COVID-19 pandemic and the resulting mid-semester shift of courses from in-person to online. Despite these challenges, four instructors implemented their modules in their courses for the Spring 2020 semester, and the RPP surveyed students and instructors, interviewed instructors, and collected information about the implementation of the modules. That summer, the RPP discussed module content, implementation, and the data collected so instructors could refine their modules (and two additional instructors could develop their initial modules for Fall 2020). The discussions at this stage highlighted important differences within the STEM discipline, the difficulty level of the courses, and the different class sizes.

Through this work, the RPP concluded that the original proposal of developing modules that could be used across all courses would not work and decided instead to focus on standardizing how the modules were to be designed and refined moving forward. The RPP adopted the "*backward design*" learning science theory [31] for refining the modules. Backward design starts with identifying student learning goals (what data science topic do you want them to know and be able to apply), and then determining how students would show they had met those goals (i.e., assessing their learning) before planning the learning activities. The RPP used these principles to create a Module Development Tool (MDT) to guide the structure of individual course data science modules (the MDT can be found at ds4stem.org/resources and is discussed in more detail in [7]). The MDT offered a way for instructors to concisely list student learning objectives and then work backward, ensuring assessments and activities provided students pathways to meet those objectives.

The module development process was iterative, with instructors going through two or three cycles of implementing, evaluating, and refining their modules using lessons learned during previous implementation(s). Instructors were asked to reflect on what went well during implementation as well as the challenges they faced in developing and deploying their modules. These reflections were then discussed as a group while brainstorming ideas for refinements that could be applied in the next iteration. The twelve modules developed in the project over three years are shown in Table I. The modules were developed for environmental science, engineering, and biology departments as well as a general university course on Smart Cities.

Between implementations, instructors often made changes to modules based on their observations and a shared module structure. This shared module structure was created as an extension of the MDT. It suggested specific components that should be present in all modules to allow for easier understanding and use of the modules by other instructors. This shared module structure included the following components: (1) a description of the disciplinary topic that a module was developed for; (2) data science learning goals, which covered the key concepts and practices that students need to learn; (3) instructor materials, which included background knowledge acquisition resources and student facing materials; (4) data sources and software used in the module; and (5) student assessments aligned with the learning goals.

TABLE I
PARTICIPATING COURSES AND PRODUCED MODULES

Course Name	Department	University	Academic Level	Pedagogy	Average Number of Students	Modules
Monitoring and Analysis of the Environment (*MAE*)	Environmental Science	VT	Senior	Lecture & Lab	30-40	(1)Errors in measured data (MAE1)
Engineering Hydrology (*EH*)	Civil and Environmental Engineering	NC A&T	Junior	Lecture & Project	30-40	(2) Time-series Analysis of Precipitation data (EH1) (3) Rainfall-runoff analysis using real-world, high-frequency data (EH2)
Hydrology (*HYDRO*)	Civil and Environmental Engineering	VT	Senior & graduate	Lecture	40-50	(4) Frequency analysis in hydrology (HYDRO1)
Smart Cities (*SC*)	University Course	VU	Senior & graduate	Lecture & Project	20-30	(5) Confidence Interval (SC1) (6) Clustering (SC2) (7) Supervised learning (SC3)
Ecology (*ECO*)	Biological Sciences	VT	Sophomore	Lecture	65-100	(8) Introduction to data science: Visualization and Interpretation (ECO1), (9) Ecology is data! (ECO2), (10) Effect of acid rain on aquatic and terrestrial ecosystems (ECO3)
Engineering Statistics (*ES*)	Industrial Engineering	NC A&T	Junior	Lecture & Lab	30-40	(11) Basic statistics (ES1), (12) Hypothesis testing (ES2)

In addition to the instructors, this project also included an Industry Advisory Panel that consisted of experts from both private and public sectors (see Fig. 1). These individuals represented companies and organizations that manufactured sampling equipment and performed data monitoring as part of their business. This group also showed interest in hiring graduates with data science skills. The nine advisors interacted with the project leaders and advised them on best practices in the industry to ensure that curricular materials were relevant to industry standards and students who graduated had desirable and relevant skills.

Another objective of the industry advisory panel was to connect directly with students through online discussions, which were moderated by the research team. The panel answered questions about a wide range of topics posed by the moderator focusing on career pathways, industry practices concerning sensors, and the use of real-world data to inform decision making. Questions were also posed by student attendees who were especially interested in the work climate of different organizations, the data science skill sets required to succeed in the industry, and the trends that the advisors could see on the horizon.

Although this paper focuses on the instructor perspective of integrating data science into their undergraduate courses, the RPP also collected data to measure student learning. Students were asked in post-surveys about their perceptions of their learning around various data science topics. Analyses of student learning are ongoing but beyond the scope of this paper. However, to demonstrate the success of the approach on student learning, a broad summary of the student survey data is included in the *Results and Discussion* section.

IV. Data Collection and Methods for Data Analysis

A. Data Collection

The RPP collected data from the six participating courses and their associated learning modules. This included the data science modules and course materials, instructor surveys and interviews, and student surveys. Student modules and course information included student academic level (freshman, sophomore, etc.), course pedagogy, instruction style (online, in-person, or hybrid), data science instruction goals and methods, course description, software used, and module(s) developed for use in the course. In this work, the course was used as the unit of analysis as the differences across courses were larger and more indicative of variations in disciplinary approaches and content than differences across institutions.

The course and module data were supplemented with data gathered from one-on-one instructor interviews. The semi-structured interviews [32] were conducted remotely to gather instructor perspectives and observations about module development and implementation. The interview protocol was developed by a combined team of graduate research/teaching assistants (GR/TAs), education researchers, and external evaluators of the RPP. The first section of the protocol validated data collected by GR/TAs about courses and modules by member-checking [32] with instructors. The second section of the protocol included ten items across four themes: (1) how each module supported students to use and analyze high-frequency, real-world data; (2) what students learned after experiencing one or more modules; (3) how the data science modules were implemented; and (4) key considerations applied for integrating data science concepts into their disciplines. These items targeted the overall goal of the RPP, i.e., developing reusable, shareable data science modules integrated into STEM undergraduate courses.

B. Methods for Data Analysis

1) RQ1-2: Integration Approach:

The RPP, interested in better understanding how instructors developed and implemented modules and the commonalities across modules, adopted an emergent method [32] for coding components of the modules and their implementation. In this process, the elements in the MDTs for all modules and their associated course summary forms were categorized. Categories included instructor role (primary or supplementary), module length (single or multiple sessions), implementation mode (in-class or out-of-class), activity type (individual, group, or a combination), assessment type (classwork, homework, project, and oral presentation, or a combination), data analysis method (point-

and-click or programming based) and sharing method (institutional or specialized learning management systems (LMS).

The instructor's role refers to their role during module implementation. Instructors played a *primary* role if they guided their students in completing the tasks and assessments, in contrast to a supplementary role if students completed a stand-alone module online with no further instruction. If instructors implemented a module online, the module length was defined by whether students completed the module over multiple class sessions (multiple days) or a single day. An iterative approach was used during the coding process to allow for the developed codes to be revised to accommodate new findings about the instructors' module development and implementation approach.

2) RQ3: Data Science Topics:

Data science topics integrated into the modules were determined by each instructor based on the specific needs of their course. The data science topics instructors incorporated were identified by analyzing each module's assessment. Rather than analyzing the module as a whole, the individual assessment prompts were separated and categorized individually. One hundred individual assessment units (each unit an assessment question or part of one) were identified from the modules assessments.

A combination of emergent and predetermined approaches [32] was adopted to categorize and code the units, allowing the use of topics from the data science life cycle while remaining open to any new topics that might be encountered during the coding process. The assessment prompts were initially coded by a GR/TA into broad data science topics. The predetermined part of the coding scheme included some data science life-cycle phases, such as data acquisition, data pre-processing, data analysis, and visualization that are common across any data-driven effort irrespective of discipline. During the coding process, new discipline-specific categories emerged, e.g., machine learning and real-world applications, resulting in the final coding scheme seen in Table II. Each assessment unit was coded by two GR/TAs to ensure validity. Units with unmatching codes were discussed between two to three GR/TAs to produce a consensus.

After classifying the module assessment prompts into broad data science categories, the GR/TAs then followed the same method to further categorize the prompts into more specific subtopics. Table II includes both the broad topic and subtopic codes used in the analysis as well as an example prompt for each broad category.

This project started too early to be guided by the ACM Data Science Task Force's Computing Competencies for Undergraduate Data Science Curricula [33] report. However, the broad data science topics and subtopics that emerged from the classification in this study align with some of the knowledge areas and sub-domains of ACM's report. For example, the broad topic and subtopic of Data Acquisition and Data Access generally align with the knowledge area and sub-domain of Data Acquisition, Management, and Governance (GD) and Data Acquisition of ACM Data Science Task Force's report. It is noteworthy to mention that the goal of this RPP was not to develop independent data science curricula, (the focus of the ACM's report) but to embed data science into fully-fledged undergraduate STEM curricula to engage students to data science principles and methods. As such, the depth and breadth of data science topics and subtopics in this study are limited compared to the broader knowledge areas and sub-domains of the ACM's report.

3) RQ4: Instructor Perspectives:

The interview data collected by the RPP were categorized by (1) challenges, (2) module development and implementation, and (3) future module use. In the challenges category, instructor and student challenges were analyzed. Instructor challenges included how instructors dealt with a diversity of student backgrounds in data science, support for instructors, integration of modules into existing course

TABLE II
DATA SCIENCE TOPICS AND THEIR DESCRIPTIONS

Broad Topic Category	Subtopic Category and Counts	Description (Subtopic Category)	Example Prompt (Subtopic Category)
Data Acquisition	Data Access (3); Data Measurement and Collection (1); Uncertainty in Data Collection (1)	Methods of data collection include sensor types (Data Measurement & Collection), accessing data from online repositories and websites (Data Access), and measurement frequency including spatial and temporal data resolution (Uncertainty in Data Collection)	*Identify, download, store, and categorize real-time and historical time-series-based water resources datasets from websites.* (Data Access) (EH1)
Data Pre-processing	Errors in Measured Data (1);	Data cleaning techniques such as data type conversion (Data Cleaning) and post data collection quality checks including measured data variability and outlier detection (Errors in Measured Data)	*"How does the time interval that the hydrologic data were collected impact the plotted hydrographs?"* (HYDRO1)
Data Use and Visualization	Interpretation & Communication (46); Statistical Analyses (20)	Data visualization and post data analysis interpretation including topics such as explaining results obtained from statistical analysis (Interpretation & Communication) and data analysis including the use of both statistical and deterministic models (Statistical Analysis)	*"Graph the average benzene level of each site with standard deviation"* (Interpretation & Communication) (MAE1)
Machine Learning	Supervised Methods (16); Unsupervised Methods (9)	Supervised algorithms and methods of analysis such as support vector machines (Supervised Methods) and unsupervised algorithms and methods of analysis such as clustering (Unsupervised Methods)	"*In the cell below, create and fit a KMeans clustering with K=3 for the IRIS data. Print the labels and the centroids that KMeans produces.*" (Unsupervised Methods) (SC2)
Miscellaneous	Real-world Application (2); Checking Model Assumptions (1)	Relation of the results of analyses to real-world situations (Real-world Application) and recognition of assumptions of models students use (Check Model Assumptions)	"*In the google colab assignment, check model assumptions before conducting hypothesis testing*". (Checking Model Assumptions) (ES2)

curriculum, finding quality/relevant data, and time constraints. Student challenges included difficulty comprehending specific data science concepts, support for students, and instructors' perspectives of students' struggles.

For module development and implementation, data source selection, data set/source science topic selection, data analysis tool selection, and implementation methods were analyzed. The data set/source selection focused on the benefits derived by data selected for instructors and students. The data science topics selection focused on the instructors' prior experience with data science topics they incorporated into their modules. The data analysis tool selection focused on which data analysis tool instructors selected for their module(s), any alternative tool(s) that could be used, and whether the students were required to have any prerequisite knowledge about the tool. Also, implementation decisions, such as instructor role in module implementation, module length, implementation mode (e.g., either synchronous online, self-paced online, or in-person in-class), student activities, assessment methods, module organization, and module publication or sharing, were analyzed.

Lastly, future module use, adaptations for general use of the modules, and future module development and implementation were analyzed. Adaptation for general use included topics such as benefits to other instructors, use in the same or different disciplines, any necessary modifications to the module, and support for other instructors. The topics covered included: online vs in-person implementation, data analysis tools, implementation and assessments, and the uniqueness of the module compared to similar educational resources.

V. Results and Discussion

A. RQ1 - 2: Integration Approach

Module development and implementation methods varied within the RPP based on instructors' needs and preferences. Some instructors designed their modules to be single-unit learning tools that were implemented as homework assignments; others designed their modules as a series of tasks that were implemented over many class sessions. A third group of instructors developed their modules to be independent learning tools made available on public platforms for self-paced learning. The different approaches are summarized in Table III.

Instructors chose their module datasets based on availability, relevance to the disciplinary topic, quality of the data, features of the data source platform, previous familiarity with the data and/or data source, and the requirements of the RPP (e.g., using high-frequency and/or real-world data). For instance, the instructor of *SC* chose standard easily-accessible machine learning data sets [34], [35] to help students develop their understanding of the fundamentals of each algorithm before they applied these algorithms to real-world, noisy, high-frequency data as part of their class projects. The *ECO* instructor used high-quality, discipline-relevant, and real-world long-term datasets that students could easily access using available interactive platforms to make understanding data easier.

Independent of the academic level, instructors of *MAE*, *ECO*, *EH*, and *HYDRO* used point-and-click-based software, such as Excel, Google Sheets, and HEC-SSP [36], for data analysis. These tools were used because of student and instructor familiarity and ease of access. Instructors of *ES* and *SC* – with backgrounds in statistics and computer science, respectively – used the script-based language Python through the Google Colab platform. Colab was chosen due to its convenience of access and use as its cloud-based and collaborative features improved access and allowed instructors to manage collaborative group assignments. It can be hypothesized that tool choice will depend on students' academic levels and disciplines. For example, freshmen or sophomores from biology, environmental science, and civil engineering courses may not be prepared to use script-based tools like Google Colab. In contrast, courses, like *SC* with more experienced students and flexible content may benefit from such advanced script-based data tools.

The instructors of the *ES* and *SC* courses used the specialized LMS GitHub Classroom [37] for sharing their Google Colab modules with students. The instructor of *HYDRO* published their module on a specialized LMS called

TABLE III
APPROACH COMPONENTS

Module	Disciplinary Topic	Data Source	Tool	Sharing Platform	Activities	Assessment
MAE	Benzene contamination spread analysis	LEWAS Lab	Excel	Canvas	Group & Individual	Homework
EH1; *EH2*	Rainfall Time-Series Analysis; Rainfall-Runoff Analysis	USGS; LEWAS Lab	Excel	Blackboard	Individual; Group & Individual	Project & Report; Project, Report & Presentation
HYDRO1	Flood and drought frequency analysis	LEWAS Lab & USGS	Excel & HEC SSP	Hydro Learn	Individual	Project, Report & Presentation
SC1; *SC2*; *SC3*	Statistical analysis and data visualization; Clustering; Regression and classification	UCI MLR; Smart Cities Lab	Google Cola	GitHub Classroom	Individual	Homework
ECO1; *ECO2*; *ECO3*	Ecological data visualization and interpretation; Ecological data pattern recognition and associated phenomena; Analysis of effects of environmental conditions on organ-isms and ecological processes	Virginia Tech StREAM Lab; Ocean Observatories Initiative; Hubbard Brook Watershed Ecosystem Record	Google Sheets	Canvas	Individual; Group & Individual	Homework; Classwork
ES1; *ES2*	Descriptive Statistics and visualization of high-frequency data; Data pre-processing and hypothesis testing	Smart City Lab; LEWAS Lab	Google Colab	GitHub Classroom	Individual	Homework

Hydrolearn, which is a hydrology and water resources public educational platform [38]. This allowed students to use it outside of the classroom. These specialized LMS provide helpful features, such as an easy-to-navigate learning environment for students and auto-grading features for instructors. In addition, instructors could review intermediate work by the students, especially on the projects, and provide contextualized feedback. The other three instructors from environmental science, biology, and civil engineering disciplines used Excel/Google Sheets and university-supported LMS, Canvas and Blackboard, to share their modules with students.

Irrespective of discipline and academic level, all modules had individual student activities. The individual activities included outside classroom work, such as homework as in the *SC* and *ES* courses. However, students sometimes worked on the module in groups in class. Collaborative assignments and projects were often leveraged to address differences in student skills and experiences, as seen in the modules of *ECO* and *SC* courses. For example, the instructor of *SC* formed student groups by pairing students with less experience in data science with students who had more experience.

The assessment method, instructor role, and module length varied. Instructors used formative and/or summative assessment methods [39] to assess student learning in their modules, including homework (*SC*, *ES, ECO*), projects (*HYDRO*, *EH*), and in-class discussions (*MAE*, *ECO*). Homework and in-class discussions were often used as formative assessments, whereas projects were used as summative assessments. In terms of role, the *HYDRO* instructor included all the teaching material (e.g., lecture recordings, required text, exercises, assessments) in the module as they intended to publish the module on a public platform to be used as an independent learning tool. As such, this instructor played a supplementary role in the implementation phase. In contrast, other modules were developed as complementary tools to the lectures since instructors had to provide context and background on module exercises, describe learning goals, and moderate post-exercise discussions. Module implementation (i.e., length) spanned from one to four weeks. The implementations included entire class sessions that were dedicated to the module, parts of class sessions on specific days (e.g., *ECO3* and *EH2*), and self-paced (*HYDRO*).

Overall to answer RQ1 and RQ2, instructors successfully designed their modules to meet their course needs. Instructors chose data sets based on variables such as availability, relevance to the disciplinary topic, quality of the data, features of the data source platform, and previous familiarity with the data. Similarly, data science tool choice depended on student and instructor familiarity. Instructors used the modules as both primary and supplementary learning tools depending on the role they adopted during module implementation. Overall, these results offer suggestions about which course elements instructors should consider when integrating data science into courses and highlight the need for a flexible approach.

B. RQ3: Data Science Topics

Instructors integrated data science learning materials into their courses by considering suitable data science topics for integration in their respective syllabi, the suitability of the topics to their students' academic level, the importance and commonality of the topics to the discipline, and the availability of relevant data for the topics chosen. For example, the *HYDRO* instructor integrated data science into the instruction of flood and drought frequency because this topic is part of the learning outcome of any upper-level undergraduate civil engineering hydrology course, the calculation of frequency curves is suitable for one or more of the data science life-cycle phases, and long-term data for calculation of frequency curves is publicly available from the United States Geological Survey [40].

Analyses of the broad data science topics across all 12 modules showed that 66 of 100 student assessment units were categorized as Data Use & Visualization, 25 as Machine Learning, and five or less as Data Acquisition, Miscellaneous, or Data Pre-Processing. Assessment units from the category Data Use & Visualization were the most common and spanned all the modules (see Fig. 2). Almost all the questions from the *ES*, *EH*, *MAE*, and *ECO* modules come from this category. The prevalence of Data Use & Visualization in the modules could be attributed to its links to basic data wrangling, analysis, and generating and interpreting visualizations that are intuitive and common (e.g., using a histogram to visualize a quantitative dataset and interpret its distribution), and provide for easy interpretation of discipline-specific topics. In contrast, relatively few assessment units were categorized in Data Acquisition, Miscellaneous, and Data Pre-Processing (the HYDRO module is the exception, with four out of 16 assessment units categorized as Data Acquisition and Data Pre-Processing).

Assessment units that focused on machine learning were the second largest category with 25 instances. All of them came from two modules in *SC*, where the instructors with a computer science background covered more advanced data science topics. In contrast, instructors for the other five courses focused on data science concepts more closely linked to discipline-specific topics. Instructors used simpler and specialized data science techniques for multiple reasons: (1) lack of time to provide instruction on more advanced data science skills in lecture-only courses; (2) the need to simplify

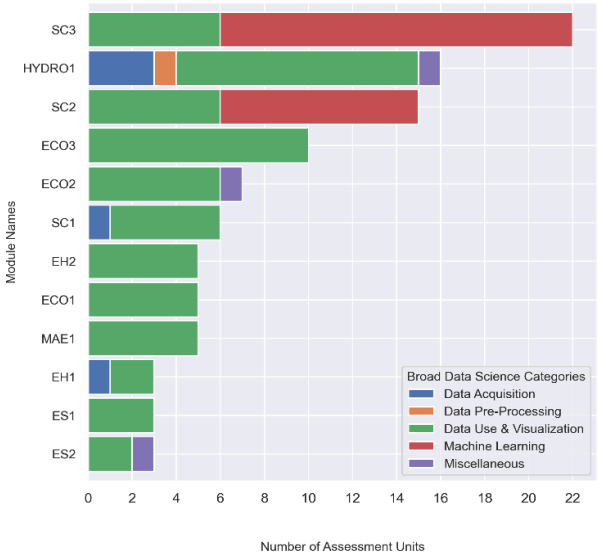

Fig. 2. Count of assessment units across modules and broad data science categories

data science concepts in order to cover vast amounts of fundamental disciplinary material; and (3) the structure of lecture-only courses with no additional meetings times (e.g., lab or discussion sections) for project-based studies that involve the collection, pre-processing, and analysis of data.

Interpreting and generating visualizations as well as conducting statistical analyses constituted most subtopics coming from the broad category of Data Use & Visualization covered by all modules (see Table II). Supervised and unsupervised machine learning methods, from the broad category of Machine Learning, constituted the second highest number of subtopics and came from two modules: *SC2* and *SC3*. Data Access from the broad category of Data Acquisition was another subtopic that was included in many modules. The subtopics in the Miscellaneous category came from three modules and were more about real-world applications. The remaining subtopics in the Data Acquisition and Data Pre-processing broad categories each constituted only one assessment question.

To answer RQ3, instructors chose a subset of data science topics/subtopics that were more discipline-agnostic, such as generating and interpreting visualizations and conducting different kinds of basic statistical analyses. Although data pre-processing includes much of any data science effort, instructors focused less on this topic for many reasons, such as the instructors' choice to use clean datasets, lack of their students' familiarity with data cleaning techniques, lack of time given disciplinary content pressures in the course, and mostly likely because the focus of the RPP, to give students a basic exposure to data science principles. In summary, instructors were mindful when choosing data science topics/subtopics for their modules about such practical considerations as balancing the depth and breadth of the topics they chose given their courses' goals and time availability, the real-world applications of the topics, and their students' interests and background related to the topics.

C. RQ4: Instructor Perspectives

1) Student Learning:

In general, instructors reported high student interest and comprehension related to their module(s) contents. For instance, the *ECO* instructor stated, "*There was a general interest [from students] in the role of data science in environmental sciences ... there were a large number of students who did well.*" As another example, the *HYDRO* instructor mentioned, "*Generally no issues reported by the students; sometimes students would say there is a lot of work.*" Also, the *SC* instructor mentioned, "*Eventually the materials were understandable to students [despite their lack of a Computer Science background] since they did well on the projects.*"

Students' self-assessment on different data science topics measured through four-point Likert scale questions given in pre/post course surveys indicated higher levels of students' perceived data science understanding and skills after participating in a course that integrated data science when comparing pre- and post-survey answers. Further discussion of student survey data is beyond the scope of this paper and will be shared in detail in a subsequent paper.

2) Wide Variability in Student Experience:

A difficulty instructors faced was the wide variability in student data science skills, specifically for the courses where students came from diverse disciplinary and academic backgrounds (e.g., *SC* and *ECO*). For example, the *ECO* instructor stated, "*Some materials might be easy enough for some students [with more quantitative backgrounds] that they might say 'why am I doing this in a college-level course?' ... while some other students might be unable to complete introductory tasks independently. So, figuring out how to accommodate students from such a wide range of skill sets within the scope of a lecture-only class is an ongoing challenge.*" Instructors who implemented their module(s) in classes where students were mostly from the same disciplinary background or academic level did not face the same challenge. For example, the *ES* instructor stated that "*None of the students had a statistics course before, this was their first statistics course so everybody was on an even ground ... for some students, it is easier to implement classroom knowledge to practice, but for others, it takes a longer time; so there is variability but it is not that huge*."

Instructors who did face the challenge of teaching to different student skill levels used methods, such as explicitly teaching the fundamentals of tools and techniques, repeating certain exercises in different contexts (e.g., using different datasets) multiple times over the semester, group work, and encouraging students to reach out to GR/TA(s), instructor, and more importantly, one another for problem-solving support. A few instructors stated the importance of GR/TAs and graduate students who *"have experience using data analytics tools can help me create learning content, especially with the LEWAS data*." Two instructors reported needing additional support in the form of tutorials for themselves or their students if they were to switch from their preferred data science tool, e.g., Excel to R or Python. The *ES* instructor reported, "*I do anticipate students not being familiar with the technology I use in the class [e.g., Google Colab and GitHub Classroom] which is why I provide step-by-step instructions."* The *SC* instructor pointed to the fact that *"...in-class examples and also more documentation... written in general rather than specific computer/data science language."* may help students.

While instructors emphasized the need for student scaffolding, they did not report the need for additional support for their own learning when using their preferred data analysis tools. However, changing the tools (e.g., from point-and-click software to programming languages) might require additional resources for both instructors and students. For example, transitioning from spreadsheet software to Python or R will require adjustments in teaching approach and material preparation. However, providing full autonomy to select data analytics tools and techniques suited to the instructors' skill sets, needs, and course requirements led to minimal need for outside support.

3) Making Space in Existing Curricula:

When asked whether they had to remove any disciplinary topic to make room for data science modules, instructors generally expressed finding it relatively easy to integrate data science content with minimal to no removal of disciplinary topics from their course syllabi. Some instructors reported having to remove minimal content to make room for tutorials about tools that were new to students such as Google Colab

and GitHub Classroom. This contrasts with previous research that found a lack of space in curricula was a major barrier to integrating data science topics [41], [42]. Instructors explained the reasons for this ease to their use of specific disciplinary topics to teach data science. For example, the *HYDRO* instructor mentioned, "*…usually topics such as frequency analysis are common in hydrology and deal with analyzing a lot of dat*a." They also pointed to the pedagogical approach as impacting the ease of integration. For instance, the *EH* instructor said, *"I used to have the project in my class anyway. But, now I adapted it in a way that can focus more on data science."* Also, the *ECO* instructor mentioned, "*A lot of it was quite easy because of the way I teach ecology using data, graphs, etc. instead of [only asking students to] memorize concepts/terms"*. The *HYDRO* instructor pointed to their flipped classroom structure as a benefit, saying it gave them the "*ability to cover both the teaching materials and the hands-on exercises during the semester"*. Overall, the minimal need for removing course content to integrate data science may be explained by the fact that instructors themselves oversaw developing and implementing the data science modules, were willing to participate in the RPP, and their courses' curricula and teaching methods were aligned with data science topics. They had a better understanding of the needs of their students and their courses.

4) Online Versus In-person Courses:

Instructors who taught in-person courses reported leveraging in-class demonstrations on accessing data and statistical analyses, collecting field data rather than using already-collected data, and adding small-group collaborative exercises. However, the instructors in this project had to switch their planned courses to online due to COVID-19. They reported that online courses provided fewer opportunities for interaction between them and their students and a lack of opportunity for hands-on exercises and demonstrations such as field data collection and data access. One instructor helped mitigate this challenge by going out and filming themselves collecting data for the students to review. Another challenge was fewer opportunities for collaborative work in groups of students. For example, the *ECO* instructor mentioned, "*For in-person implementation, I plan to hold in-class collaborative exercises in small groups to be on-call for the students when they encounter some issues and allow students to also learn from their peers since students have different skill sets. This was not possible in the way the lectures were run remotely*." The instructor mitigated this challenge by leveraging online collaboration tools such virtual breakout rooms to facilitate group work, discussion boards, and implementing virtual office hours and online Q&A sessions.

5) Usefulness in Other Disciplines:

Some instructors believed that their modules are content specific and cannot be used in other disciplines. For example, the *MAE* instructor said, "*... perhaps biology or hydrology, but they are all under environmental science ... I don't know how computer science students would use my module because they are completely talking about other concepts.*" However, other instructors believe their modules can be adapted for other disciplines and courses saying other instructors can "*decide what they want to change based on their needs."* These instructors pointed to the structure of the modules as being of interest to other disciplines, for example, the *ES* instructor said, "… *the structure of the modules I developed is the most beneficial tool for other instructors to use."* In some of these modules, instructors also used general data science instruction before moving into domain-specific instruction. For example, in *SC,* the instructor used discipline-agnostic data to teach students data science concepts before using domain-specific real-world data in projects. Such module structures can be easily adapted to other courses across disciplines.

Additionally, the modules were developed based on real-world disciplinary data, include hands-on exercises, and introduce useful resources. The *MAE* instructor mentioned, "*The module is based on real-life cases that students will encounter in their careers.*" Another reason instructors give for the usefulness of their modules to other instructors was the introduction of resources and platforms that other instructors might not be aware of. For example, the *ECO* instructor stated, *"... there are lots of readings, examples, and online data visualization interfaces that may be of interest to other instructors and students find engaging."*

6) Scaffolding Needs:

For scaffolding needs, instructors' answers depended on whether other instructors planned to modify the modules or use them as is. If other instructors from similar disciplines use the same datasets and data analysis tools (e.g., Excel/Google spreadsheets), instructors thought little extra support would be necessary. However, if other instructors use other data (or generate their own data) and/or analysis tools, members of the RPP suggested a need for lab equipment and tutorials since some tools used are not as universal as spreadsheets. Also, instructors believed other instructors who use their modules would benefit from the self-explanatory rubrics, clear learning objectives, and reasons why those learning objectives have been incorporated into modules. For instance, the *ECO* instructor mentioned that "*...including an explanatory rubric that shows what we were expecting and why we included these kinds of assignments can be helpful. Because we created these assignments, we know our expectations and reasons for incorporating the assignments, but it might be difficult for other instructors to understand our motivations for all module content."* All the modules available on our website (ds4stem.org) have these components.

7) Data Analysis Tools:

When asked about whether their module could be used with data analysis software that was different from the original module design, all instructors answered positively. Three out of four instructors who used point-and-click tools such as Excel or Google spreadsheets answered that other software and even programming-based tools such as R or Python could be used. Some felt, however, that students might not have the experience or willingness to use more advanced tools and that instructors might need more class time and tutorials to instruct on programming-based tools: "*Students are not comfortable using R, so time will be needed to teach students how to use it."* One instructor did not believe using programming-based tools is viable for their

module due to the lecture-only content-based nature of their course. The *ES* and *SC* instructors who used Python expressed their openness to using other tools, such as R.

8) Comparison to Other Learning Materials:

Finally, instructors were asked about how they thought their modules compared to other similar learning materials developed and published elsewhere. Instructors answered that their modules are tailored to specific disciplinary contexts and class pedagogies, which makes them unique. For instance, the *ECO* instructor mentioned, *"I think my modules are different in that they are short activities designed for lecture-based classes with limited class time, as opposed to multi-week lab-based courses."* Another instructor stated *"... this is the only module I know that combines stormwater monitoring and data science. This is very different from other modules ... it signifies the importance of monitoring stormwater for better stormwater management."* The *ES* instructor said *"We don't use a lot of coding to learn statistics. There is a lot of hand calculation. I think these modules provide more resources for students. They will not only know how to do the calculations manually but also how software generates the results."*

Overall, to answer RQ4, instructors faced common difficulties in the wide variability in student data science skills, specifically in courses where students came from diverse disciplinary and academic backgrounds and challenges during online courses, such as the lack of opportunity for hands-on data collection and student collaboration. However, instructors reported successful strategies for targeting these challenges including the development of training tutorials, leveraging teaching assistants, and generating videos to demonstrate the collection of field data. Instructors pointed to their control over the data science curriculum as a positive component of this project. As instructors developed their own data science modules situated within their own course needs and goals, they were able to integrate the data science content more easily within their courses, thus minimizing difficulties with fitting content into their existing courses and allowing them to more easily make adjustments. Furthermore, instructors were conflicted on whether their modules could be adapted "as is" for use by other instructors, particularly outside their discipline. However, instructors pointed to the module structure as a possible starting point for those who may want to develop their modules (just like the instructors in this study did), as well as the usefulness of real-world data. Instructors also highlighted the necessity of developing student scaffolding, particularly when introducing new data science tools and suggested the use of tutorials or support in the form of teaching assistants. In addition, most instructors felt that this approach led to modules that could be adapted to be more complex or simple, depending on student needs, by changing the data science tools that were used. Finally, instructors felt their modules differed from already existing learning materials, particularly because they were discipline specific.

VI. Limitations and Future Work

This study does not include an in-depth analysis of student learning outcomes, as the primary focus of the paper was to present instructor perspectives on the integration process. While the identification of data science topics within the modules was primarily based on the analysis of student assessments, this approach may have overlooked certain data science topics that were taught but not assessed, or those that were evaluated through quizzes or exams not considered as part of the module analyses presented here.

Future work will extend the scope beyond this study by assessing student learning to determine the efficacy of instructional approaches and assess the long-term impacts on student learning and data science integration into STEM disciplines. This future work will include extending preliminary analyses [7], [43] and dive deeper into student perspectives and learning during the course of these classes that integrated data science. Furthermore, future work will include the analysis of additional resources, such as instructor slides and exam questions, to better identify the full range of data science topics covered. Additionally, long-term impact assessments will be conducted to further validate the effectiveness of this educational intervention. Lastly, future work will explicitly focus on the further development of the relationships and collaboration established across instructors, industry professionals, and universities during this partnership to build a network of professionals engaged in longer-term efforts to develop the knowledge necessary to integrate data science topics into higher education and develop students with the data science capabilities needed in STEM careers.

VII. Conclusion

The RPP in this work developed data science modules for a variety of STEM courses through an iterative process that emphasized the needs of the instructors by tailoring the modules to meet the disciplinary, academic level, and pedagogical requirements of each course. Throughout the RPP, instructors developed and organized data science modules that were integrated into their existing courses, shared resources (e.g., data sets) with each other, reflected on the multiple implementations of their data science module and made changes accordingly, and developed a shared module structure. This process resulted in instructors from three universities developing and integrating 12 modules into their respective courses.

Instructors embed data science topics into their courses using disciplinary topics and real-world data with hands-on exercises. Effective data science instruction includes anchoring content into real-world case studies with hands-on exercises [12], [44]. Moreover, instructor background, disciplinary traditions, student backgrounds, and the needs and goals of a course influenced how data science was perceived, topics included, and how integration happened. Instructors showed a common interest in integrating data communication (i.e., interpreting and generating visualizations) and statistical analyses. This finding has implications for those who want to integrate data science into their own course: instructors should be mindful of such points as balancing the depth and breadth of the data science content given their courses' goals and time availability, the content's real-world application, and their students' interests and background about the content.

Instructors faced the challenge of large differences in student backgrounds on data science content and identified ways of dealing with this challenge including development of tutorials, leveraging teaching assistants, and grouping students based on experience levels. Instructors were able to identify and use relevant authentic, real-world data in their

courses. Some previous studies indicate this as a barrier to data science teaching and learning [42], [45], suggesting the need for shared resources, like those that are shared on the website of this study (ds4stem.org). Lastly, no instructor reported major issues with the lack of space in curricula for the integration of data science content despite this being noted as a common barrier elsewhere [46], [47], suggesting that the approach to integrating data science within discipline specific courses may help mitigate this issue when instructors have full control over the content being integrated into their courses. Also, this approach may result in more long-term sustainable integration because such an integration method helps embed data science concepts into the disciplinary context rather than additional topics to be covered in an already busy course. Also, such an integrated approach bypasses the problem students can have in transferring their data science skills to their disciplinary context [16], [17].

Overall, this paper demonstrated the effectiveness of the RPP in developing structured data science modules that can be integrated into multiple STEM courses and engages more than 800 students to data science instruction, of which 80% were from under-represented groups in STEM: women, Black, Hispanic/Latino, English as a second language, or first-generation college student. Additionally, instructors developed and implemented modules not only during the timeline of this NSF project but continue to do so, suggesting the long-term viability of this approach. Furthermore, it was also shown how different instructors can work together and benefit each other in module and assessment development. This suggests that universities that wish to encourage faculty to integrate data science into their STEM curricula could benefit from creating a space for instructors across disciplines to meet and discuss the design and implementation of curricula aimed at integrating data science into their courses.

Substantive contributions statement:

A portion of the data used in this manuscript was previously presented at the 2022 ASEE Annual Conference & Exposition [43]. Specifically, data from four out of the 12 modules analyzed in the current study were included in the conference paper. However, the research questions, objectives, and analyses in this manuscript are distinct from those in the conference paper. The present study incorporates data from eight additional data science modules and includes qualitative data from one-on-one interviews with participating instructors, which were not part of the conference paper. These additional data sources and analyses provide a more comprehensive understanding of the integration of data science modules into undergraduate STEM courses and offer new insights into the instructors' perspectives and experiences throughout the process. Thus, this manuscript presents a substantive and novel contribution beyond the preliminary findings reported in the conference paper.

REFERENCES

[1] National Academies of Sciences, Engineering, and Medicine, *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press, 2018.

[2] J. M. Wing, "The data life cycle," *Harvard Data Science Review*, vol. 1, no. 1, p. 6, 2019.

[3] A. Bundy, "Australian and New Zealand Information Literacy Framework," Australian and New Zealand Institute for Information Literacy, Adelaide, 2004.

[4] D. Deb, R. M. Smith, and M. Fuad, "Infusing Data Science Across Disciplines," in *Proc. 2019 ACM Conf. Innovation and Technology in Computer Science Education*, 2019.

[5] C. Dichev and D. Dicheva, "Towards Data Science Literacy," *Procedia Computer Science*, vol. 108, pp. 2151–2160, 2017.

[6] R. J. Brunner and E. J. Kim, "Teaching Data Science," *Procedia Computer Science*, vol. 80, pp. 1947–1956, 2016.

[7] C. Snyder et al., "Understanding Data Science Instruction in Multiple STEM Disciplines," in *2021 ASEE Virtual Annual Conf. Content Access*, 2021.

[8] J. M. Durden, J. Y. Luo, H. Alexander, A. M. Flanagan, and L. Grossmann, "Integrating 'Big Data' into Aquatic Ecology: Challenges and Opportunities," *Limnology and Oceanography Bulletin*, vol. 26, no. 4, pp. 101–108, Nov. 2017.

[9] J. Hardin et al., "Data Science in Statistics Curricula: Preparing Students to Think with Data?," *The American Statistician*, vol. 69, no. 4, pp. 343–353, Oct. 2015.

[10] A. R. Rao, Y. Desai, and K. Mishra, "Data science education through education data: an end-to-end perspective," in *Proc. IEEE Integrated STEM Education Conf. (ISEC)*, Mar. 2019, pp. 300-307.

[11] U. Berkeley, "Online Master's in Data Science," Berkeley School of Information. [Online]. Available: https://ischoolonline.berkeley.edu/data-science/

[12] S. C. Hicks and R. A. Irizarry, "A guide to teaching data science," *The American Statistician*, vol. 72, no. 4, pp. 382-391, 2018.

[13] F. Maina, J. Smit, and A. Serwadda, "Professional Development for Rural Stem Teachers on Data Science and Cybersecurity: A University and School District Partnership," *AIJRE*, vol. 31, no. 1, pp. 30-41, Mar. 2021.

[14] National Research Council, *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: National Academies Press, 2000.

[15] S. Britton, "Are students able to transfer mathematical knowledge," in *Proc. Second Int. Conf. Teaching of Mathematics*, 2002.

[16] S. Hester, S. Buxner, L. Elfring, and L. Nagy, "Integrating Quantitative Thinking into an Introductory Biology Course Improves Students' Mathematical Reasoning in Biological Contexts," *LSE*, vol. 13, no. 1, pp. 54–64, Mar. 2014.

[17] J. L. Klug, C. C. Carey, D. C. Richardson, and R. Darner Gougis, "Analysis of high-frequency and long-term data in undergraduate ecology classes improves quantitative literacy," *Ecosphere*, vol. 8, no. 3, p. e01733, Mar. 2017.

[18] M. J. Oudshoorn, K. J. Titus, and W. K. Suchan, "Building a New Data Science Program Based on an Existing Computer Science Program," in *Proc. 2020 IEEE Frontiers in Education Conf. (FIE)*, Oct. 2020, pp. 1–5.

[19] B. Marshall and S. Geier, "Cross-Disciplinary Faculty Development in Data Science Principles for Classroom

Integration," in *Proc. 51st ACM Technical Symp. Computer Science Education*, 2020.
[20] K. Hunt, "The challenges of integrating data literacy into the curriculum in an undergraduate institution," *IQ*, vol. 28, no. 2, p. 12, Aug. 2005.
[21] A. Janiak and R. Rudek, "Experience-based approach to scheduling problems with the learning effect," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 39, no. 2, pp. 344–357, Mar. 2009.
[22] T. Koltay, "Data literacy: in search of a name and identity," *Journal of Documentation*, vol. 71, no. 2, pp. 401-415, 2015.
[23] S. Kross and P. J. Guo, "Practitioners teaching data science in industry and academia: Expectations, workflows, and challenges," in *Proc. 2019 CHI Conf. Human Factors in Computing Systems*, 2019, pp. 1–14.
[24] B. Baumer, "A Data Science Course for Undergraduates: Thinking With Data," *The American Statistician*, vol. 69, no. 4, pp. 334–342, Nov. 2015.
[25] S. C. Hicks and R. A. Irizarry, "A Guide to Teaching Data Science," *The American Statistician*, vol. 72, no. 4, pp. 382–391, Oct. 2018.
[26] K. Mike, G. Hartal, and O. Hazzan, "Widening the shrinking pipeline: The case of data science," in *Proc. 2021 IEEE Global Engineering Education Conf. (EDUCON)*, 2021.
[27] D. Schuff, "Data Science for All: A University-Wide Course in Data Literacy," in *Analytics and Data Science*, A. Deokar, A. Gupta, L. Iyer, and M. Jones, Eds. Cham: Springer, 2018, pp. 281-297.
[28] C. E. Coburn and W. R. Penuel, "Research-Practice Partnerships in Education, Outcomes, Dynamics, and Open Questions," *Educational Researcher*, vol. 45, no. 1, pp. 48–54, Jan. 2016.
[29] C. E. Coburn, W. R. Penuel, and K. E. Geil, "Practice partnerships: A strategy for leveraging research for educational improvement in school districts," William T. Grant Foundation, 2013.
[30] E. C. Henrick, P. Cobb, W. R. Penuel, K. Jackson, and T. Clark, "Assessing Research-Practice Partnerships: Five Dimensions of Effectiveness," William T. Grant Foundation, 2017.
[31] G. P. Wiggins and J. McTighe, *Understanding by design*. Alexandria, VA: ASCD, 2005.
[32] J. W. Creswell and J. D. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage publications, 2017.
[33] A. Danyluk et al., "Computing competencies for undergraduate data science programs: An ACM task force final report," in *Proc. 52nd ACM Technical Symp. Computer Science Education*, 2021.
[34] C. L. Blake and C. Merz, "UCI repository of machine learning databases," University of California, Irvine, Dept. of Information and Computer Sciences, 1998. [Online]. Available: https://archive.ics.uci.edu/ml/index.php
[35] Census Bureau (US), "Statistical abstract of the United States," U.S. Census Bureau, 2009. [Online]. Available: https://www.census.gov/library/publications/time-series/statistical_abstracts.html
[36] "HEC-SSP." https://www.hec.usace.army.mil/software/hec-ssp/ (accessed Dec. 31, 2022).
[37] "GitHub Classroom." https://classroom.github.com/ (accessed Dec. 31, 2022).
[38] "HydroLearn." https://www.hydrolearn.org/ (accessed Dec. 31, 2022).
[39] D. D. Dixson and F. C. Worrell, "Formative and Summative Assessment in the Classroom," *Theory Into Practice*, vol. 55, no. 2, pp. 153–159, Apr. 2016.
[40] "USGS Current Water Data for the Nation." https://waterdata.usgs.gov/nwis/rt (accessed Dec. 31, 2022).
[41] J. Bonnell, M. Ogihara, Y. Yesha, and I. Bojanova, "Challenges and Issues in Data Science Education," *Computer*, vol. 55, no. 2, pp. 63–66, Feb. 2022.
[42] C. A. Strasser and S. E. Hampton, "The fractured lab notebook: undergraduates and ecological data management training in the United States," *Ecosphere*, vol. 3, no. 12, p. art116, Dec. 2012.
[43] M. Y. Naseri et al., "A modular approach for integrating data science concepts into multiple undergraduate STEM+ C courses," in *Proc. 2022 American Society Engineering Education Annual Meeting*, 2022.
[44] A. Adhikari, J. DeNero, and M. I. Jordan, "Interleaving computational and inferential thinking: Data science for undergraduates at Berkeley," *Harvard Data Science Review*, vol. 3, no. 2, 2021.
[45] T. A. Langen et al., "Using large public datasets in the undergraduate ecology classroom," *Frontiers in Ecology and the Environment*, vol. 12, no. 6, pp. 362–363, Aug. 2014.
[46] N. Emery et al., "Training Data: How can we best prepare instructors to teach data science in undergraduate biology and environmental science courses?," Cold Spring Harbor Laboratory, 2021.
[47] M. Haynes et al., "Integrating Data Science into a General Education Information Technology Course," in *Proc. 20th Annual SIG Conf. Information Technology Education*, 2019.