A modular approach for integrating data science concepts into multiple undergraduate STEM+C courses

Abstract

With increasingly technology-driven workplaces and high data volumes, instructors across STEM+C disciplines are integrating more data science topics into their course learning objectives. However, instructors face significant challenges in integrating additional data science concepts into their already full course schedules. Streamlined instructional modules that are integrated with course content, and cover relevant data science topics, such as data collection, uncertainty in data, visualization, and analysis using statistical and machine learning methods can benefit instructors across multiple disciplines. As part of a cross-university research program, we designed a systematic structural approach–based on shared instructional and assessment principles–to construct modules that are tailored to meet the needs of multiple instructional disciplines, academic levels, and pedagogies. Adopting a research-practice partnership approach, we have collectively developed twelve modules working closely with instructors and their teaching assistants for six undergraduate courses.

We identified and coded primary data science concepts in the modules into five common themes: 1) data acquisition, 2) data quality issues, 3) data use and visualization, 4) advanced machine learning techniques, and 5) miscellaneous topics that may be unique to a particular discipline (e.g., how to analyze data streams collected by a special sensor). These themes were further subdivided to make it easier for the instructors to contextualize the data science concepts in discipline-specific work. In this paper, we present as a case study the design and analysis of four of the modules, primarily so we can compare and contrast pairs of similar courses that were taught at different levels or at different universities. Preliminary analyses show the wide distribution of data science topics that are common among a number of environmental science and engineering courses. We identified commonalities and differences in the integration of data science instruction (through modules) into these courses. This analysis informs the development of a set of key considerations for integrating data science concepts into a variety of STEM + C courses.

1. Introduction

A basic understanding of data science has been suggested as a fundamental component of undergraduate education due to increasingly data-driven work across all domains [1]. Data science topics such as data collection, uncertainty in data, data visualization, and analysis using statistical and machine learning methods are relevant to students across multiple disciplines. Embedding data science instruction into undergraduate courses can lead to increased student comfort level and experience with analytical tools [2]. However, instructors face a variety of

challenges when integrating data science concepts into their courses such as already full course contents and the wide range of students' backgrounds and familiarity with data processing and data analysis tools [3]. While previous research has led to the development of instructional data science materials within specific domains [4], [5], such resources focus on data science instruction embedded in one domain. Principles for integrating data science instruction across a variety of STEM domains are not clear.

As part of a cross-university partnership funded by the NSF's IUSE (Improving Undergraduate STEM Education) program, we have developed 12 modules using an interdisciplinary approach to incorporate data science concepts into undergraduate STEM courses in a systematic and generalizable manner. In this paper, we analyze four modules that integrate data science concepts into courses in a systematic manner, while meeting the different needs of the instructional disciplines, academic levels, and pedagogies.
This study attempts to answer the following research questions:

> (1) What are the similarities and differences in the approach instructors use to integrate data science topics into their curricula across academic levels, disciplines, and universities?
>
> (2) What are the similarities and differences in data science topics covered across academic levels, disciplines, and universities?

We present a systematic module design process that applies across all of our courses, and report the structure and assessments that we have developed for each module. For analysis, we adopt a case study approach to identify the commonalities and differences in integrating data science instruction through our module design into these courses. This analysis informs the development of a set of key considerations for integrating data science concepts into a variety of STEM courses. Our approach is aligned with the emergent and bottom-up characteristic of this research-practice partnership, where each instructor developed their own data science learning objectives and integration approach independent of other instructors in the project. This approach enables us to critically analyze the characteristics and dynamics of each case to understand the similarities and differences between them which, in turn, will help us to gain a more comprehensive understanding of the data science integration process across universities, STEM disciplines, and academic levels.

2. Background Information

Data science education has been recognized as an important part of education for students in all STEM fields. Fairleigh Dickinson University offers the course "Modern Technologies" in its undergraduate engineering department. This course focuses on providing first year students with

real-world datasets that allow them to experience the application of data science in engaging ways [6]. Other universities have also taken approaches to introduce data science into a wider field of undergraduate studies [5]. These approaches include offering elective courses, such as the Data Science course offered at Smith College, to the required course, Concepts in Computing with Data, which is jointly offered to upper level undergraduate students at UC Berkeley and UC Davis [7]. A common theme that arises from these data science oriented courses is that they expose students to the basic concepts of data science, such as data cleanup and data reporting. While the UC Berkeley and UC Davis courses are offered by their statistics departments, it should be noted that a majority of students who enroll in Concepts in Computing with Data were not in the statistics department [5]. This speaks to the recognition by today's students that data science familiarity is important regardless of their program of study. This sentiment is echoed by the National Science Foundation, and is expressed by their funding of this project and the funding of data science initiatives focused on exposing K-12 students to data science concepts [8].

Through discussions, our project has identified a number of cross cutting data science concepts, such as data acquisition, quality issues, pre-processing, analysis, and visualization that apply across disciplines. Using these topics as established student learning goals, we have employed a backward design to ensure that individual course data science modules are structured to meet these goals [9]. Project team members then got together to design module development tools for instructors in a way that they could concisely list student learning objectives then work backwards, designing assessments and activities that provided students pathways to meet those objectives. The assessments, lessons, and activities created using these module development tools were then packaged and used for classroom instruction and assessments with accompanying metrics. Overall, this approach adopted by our project promotes module refinement and reuse, and also opportunities for other instructors to adopt these modules as is, or with refinements and modifications that are suited to their individual courses

3. Data Collection and Methods for Data Analysis

We analyze one module each from four different courses. The *Monitoring and Analysis of the Environment* course is a lecture and lab based course which consists of 30-40 senior level students. The module in this course studies methods for identifying errors in measured data using data from the LEWAS [10] dataset presented to students as Excel worksheets. The *Ecology* course is a lecture based course with 90-100 sophomores. The data science module in Ecology focuses on the effects of acid rain on aquatic and terrestrial ecosystems using data from the Hubbard Brook Experimental Forest dataset (hbwater.org). The data set is made available to students in Google Sheets, and students perform their analyses in the same environment. Both these courses are taught by faculty at Virginia Tech (VT). The third course, *Engineering Hydrology* taught at North Carolina A&T (NCA&T), is a lecture and project based course with

30-40 junior level students. The module analyzed for this course covers rainfall-runoff analysis using real-world high-frequency data from the LEWAS dataset, which students analyzed using Excel worksheets. The fourth module was developed for a *Hydrology* lecture-based course with 40-50 senior and graduate level students at VT. This module covers frequency analysis in hydrology using the LEWAS and USGS (data.usgs.gov) datasets. Students used Excel and HEC-SSP (Hydrologic Engineering Center Statistical Software Package) to analyze and draw conclusions from the data.

Our data sources include course summary forms (CSFs), module development tools (MDTs), which create a framework for comparing course-specific modules [3], and the modules themselves. The CSFs consist of details about the courses including semester/year, instructor/institution, course identification code/level/description/modules, student enrollment, teaching mode and pedagogy, data science instruction goals and methods, and software used for instruction. The MDTs cover student learning goals, student assessments, student activities, lesson plans, data sources and software, and project information. From sources, we analyzed the modules according to Table 1.

Table 1. Approach Components

No.	Approach Components	Coding Schemes	Description
1	Instructor role	Central	Instructors' presence is necessary.
		Supplementary	Instructors' presence is not necessary but supplementary.
2	Module length	Single session	The module is implemented over a single classroom session.
		Multiple sessions	The module is implemented over multiple classroom sessions.
3	deployment Mode	In-person	The module is implemented in person in a classroom.
		Online	The module is implemented online and self-paced over a specified period.
4	Student activities	Individual	Students carry out the required activities of the module individually.
		Group	Students carry out the required activities of the modules in groups.
		Individual & group	Students carry out the required activities of the module both individually and in groups.
5	Student assessment	Classwork	Students' learning outcome is assessed only through classwork.
		Homework	Students' learning outcome is assessed only through homework.
		Project & report	Students' learning outcome is assessed through an individual project and its associated report.
		Homework + project & report	Students' learning outcome is assessed through individual homework and a project and its associated report.
		Project & report + oral presentation	Students' learning outcome is assessed through an individual project and its associated report and an oral presentation.
6	Data analysis method	Point-and-click-based	Data analysis is done through point-and-click-based software such as Excel.
		Script-based	Data analysis is done through a script-based programming language such as Python on Google Colab.
7	Publication platform	Institution Learning Management System (LMS)	The instructors have used the official learning management system (e.g., Canvas or Blackboard) of their institutions.

		Specialized LMS	The instructors have used a more specialized learning management system (e.g., GitHub Classroom or HydroLearn).

An inductive method [11] has been adopted for coding components of the data science modules into their respective categories. In this process, first, the MDTs for all the developed modules and their associated CSFs were organized, observed, and discretized into data segments. Second, the coding process was started by placing the data segments into categories and subcategories and were labeled with descriptive names/codes. Based on the results of the second step, categories were developed as described in Table 1. We used an iterative approach during the coding process. On many occasions, the developed codes were revised to accommodate new findings about the instructors' approach components across the modules.

The information for some of the approach components like the student assessment, activities, module length, and instructor role has directly come from the MDTs. However, for other approach components such as deployment mode, data analysis method, and publication platform, the information has been integrated from various sources including modules themselves and different parts of MDTs.

The general module framework was created by one of the project faculty with a data science background, who worked with the graduate research assistants (RAs) on the project and a faculty member in education to develop the module structure and the proposed components. The "instructor role" code describes the instructor role only during the deployment of the modules rather than during the module development process. Module development was done primarily by the graduate RAs, who worked closely with the instructors who played a central role in setting the module goals, the instructional material, the data sets, the assessments, and the grading rubric. During the deployment of their modules, as part of their classroom instruction, they played a primary role in guiding their students in completing the tasks and assessments in the module. However, instructor roles were categorized as supplementary if the instructors asked their students to complete the modules' tasks as homework assignments or take a stand-alone module online with no further instruction from the instructor.

The coding scheme for the module length applies whether the module was implemented in-person or online. If a module was implemented in person, the code categories indicate whether an instructor had decided to implement the module in one session or over multiple sessions. However, if a module was implemented online, the code categories indicate whether the instructors had allowed their students to complete the module tasks over multiple equivalent class sessions (e.g., multiple days) or a single.

The data science topics that instructors incorporated into their respective modules were collected from the modules themselves. It was decided that the assessment prompts of the modules were appropriate indicators of what data science topics each module entails and provides accompanying assessments. Each of the modules had multiple questions or prompts that students were asked to complete. Rather than categorizing the module as a whole, we decided to break down the student assessments in each of the modules into the individual questions or prompts students were asked to answer or complete and then code those assessment prompts individually.

After all the assessment prompts from the modules had been collected in one place, they were discretized into logical units. These units were components of each assessment with a unity data science concept that could be categorized in one or another data science subcategories. The discretization process was done to ease the subsequent process of categorization and coding. 36 individual prompts were identified from the four representative modules that were subsequently categorized and coded.

As a next step, each prompt was double-coded into more specific categories (Table 2). After the initial double-coding process, 28 out of 36 prompts matched the broad data science topics. Of the 28 that matched the broad data science topic, 22 matched a specific topic. The team discussed the non-matched prompts and produced consensus codes as a group. This led to having a third coder reviewing the non-matched topics, listening to the discussions of the two initial coders, and finally coding the non-matched prompts into an existing subcategory.

A combination of an emergent and predetermined approach [11] was adopted for the categorization and coding in this section, in part due to the bottom-up organization of this research-practice partnership. Based on this organizational approach, instructors developed their modules for different STEM disciplines, course pedagogies, academic levels, and needs independent of each other. However, using only an emergent approach to coding would have obscured the topical inadequacies of our modules. Therefore, we conducted a literature review on the most common categorization of data science concepts and techniques. Despite the evolving nature of data science as an academic discipline, we found general trends of data science concepts and techniques common across disciplines. These general trends were categorized into six broad categories: (1) data acquisition, (2) data quality issues, (3) data use and visualization, (4) machine learning, (5) data ethics, privacy, and security, and (6) miscellaneous. Table 2 summarizes the coding scheme and gives a description of each of the subcategories under the six broad categories.

Table 2. Coding Scheme for Data Science Topics

NO.	Broad Data Science Topic	Specific Data Science Topic	Description

1	Data Acquisition	Data Measurement	Concerned with data measurement frequency; includes such topics as spatial and temporal data resolution
		Data Collection Mechanisms including Sensors	Concerned with different methods of data collection, including different sensor types and their characteristics
		Data Access	Concerned with how students can access data from online repositories and data streaming websites such as that of the U.S Geological Survey website
2	Data Quality Issues	Uncertainty in Data Collection	Concerned with quality impacts of methods on the measured data; includes such topics as impacts of temporal frequencies of collected data on modeling results
		Errors in Measured Data	Concerned with post data collection quality checks; includes such topics as variability and outlier detection in the measured data
3	Data Use and Visualization	Visualization	Concerned with any data visualization including raw and processed data visualization; includes such topics as time-series data visualization
		Statistical Analysis	Concerned with any kind of data analysis including the use of both statistical and deterministic models; includes such topics as finding measure of central tendency of a dataset
		Data Interpretation	Concerned with post data analysis interpretation; includes such topics as offering explanation to the results obtained from a statistical analysis
4	Machine Learning	Supervised Methods	Concerned with supervised algorithms
		Unsupervised Methods	Concerned with unsupervised algorithms
5	Data Ethics, Privacy, and Security	Data Ethics	Concerned with ethical issues in the use of data and algorithms
		Data Privacy	Concerned with data privacy issues including rules and regulations
		Data Security	Concerned with data security issues including cybersecurity
6	Miscellaneous	Real-world Application	Involves prompts that assess the students on relating the results of their statistical and/or machine learning analysis to a real-world situation; for example, selecting an appropriate design for a hydraulic structure for which students must refer to what they did in the data analysis phase
		Check Model Assumptions	Involves prompts that assess the students on recognizing the assumptions of statistical and/or machine learning models they used at the data

			analysis phase
		Data Presentation	Involves prompts that assess students on communicating their analysis results and/or another disciplinary concept through data beyond what data visualization prompts had assessed

The combination of emergent and predetermined coding approach [11] allowed us to add new categories and/or subcategories to the predetermined data science topics through emergent design. For example, the subcategories Data Access in the Data Acquisition category and all the ones in the Miscellaneous category emerged (i.e., were added) through emergent design. Moreover, our coding approach allowed us to detect inadequacies as well as the distribution of different data science topics across disciplines and academic levels. For example, we did not find any prompts aligned with the subcategories Data Collection Mechanisms including Sensors in Data Acquisition and the ones in Data Ethics, Privacy, and Security.

4. Results and Discussion

4.1. Module Development and integration approaches

Instructors assumed central instructional roles in three out of the four modules. The only module in which the instructor did not have a central instructional role comes from the senior/graduate Hydrology class called Frequency Analysis in Hydrology. This module has been designed as a stand-alone instructional tool with instructional videos and recorded lectures along with other self-explanatory components, such as learning activities and exercises. Moreover, this module has been published on an LMS that provides many features and scaffolding to the students to navigate the module without any external help. The rest of the modules discussed in this study in which the instructors have assumed central instructional roles are not stand-alone modules or published on such LMS as that of Frequency Analysis in Hydrology.

Instructors in this study showed a common predisposition to assume central instructional roles during the deployment of their respective modules irrespective of whether their classes were consisting of the majority upper- or lowerclassmen. For the modules in which instructors assumed central roles, not much context for the exercises was provided. In other words, such modules were dependent on the instructors' necessary information to fill the context gap to allow students to comprehend the broad purpose of the module. For example, the role of the instructor for the modules implemented in the sophomore level Ecology class included providing pre-exercise lectures, being available as students completed the exercise within the modules, and facilitating post-exercise discussions. For the one module in which the instructor assumed a supplementary instructional role (i.e., Frequency Analysis in Hydrology), the module is

considered stand-alone since it includes all the required text and lecture materials that help students to have a complete sense of the overall purpose of the module and the exercises with which they engaged.

Out of the four representative modules, only the module Frequency Analysis in Hydrology - the same module with a supplementary role for the instructor - has been designed to be implemented online. The rest of the modules were implemented synchronously, in person, or remotely on zoom, and instructors assumed central instructional roles. These modules were designed to be implemented in person with the presence of the instructor. However, due to the COVID 19 situation that canceled in-person classrooms, some of these modules like the module developed in the sophomore level class Ecology called Effects of Acid Rain on Aquatic and Terrestrial Ecosystems were implemented remotely through synchronous and/or asynchronous online sessions. For the three modules that were designed to be implemented in person, the presence of the instructors is necessary for seamlessly incorporating them into course contents. Some of these modules have components (e.g., assignments) that students have taken online. However, some others have been completely implemented in a classroom context.

In terms of student activity types, three out of four modules have incorporated both individual and group activities. However, the senior/graduate level module that was implemented online (i.e., Frequency Analysis in Hydrology) has only incorporated individual student activities. In terms of methods used for assessing student learning outcomes, the sophomore level module from the course Ecology used classwork besides homework; however, the upperclassmen modules used other methods such as project and report or a combination of project and report and homework assignment or presentation. Both modules coming from the engineering discipline (i.e., civil engineering) have a project and report or a combination of project and report with a homework assignment or oral presentation as methods of assessing student learning outcomes. This suggests that classwork at the group level may provide more suitable scaffolding for the lower-level undergraduates compared to project and report and/or individual homework assignments [12]. Moreover, disciplinary tradition as well as whether or not courses include a laboratory or discussion section may play a role in helping instructors select their method of learning outcome assessment [13].

There is an association between the mode of deployment and the types of student activities instructors incorporated into the respective modules. The instructor who implemented their modules in an online mode tended to incorporate individual student activities into their module. However, those instructors who implemented their modules in-person have also incorporated group activities. This implies that instructors who designed their module for in-person deployment found group activities more practicable compared to the instructor who designed their module for online deployment. Also, there is a general tendency toward incorporating

individual student activities in all four modules. This suggests that this is related to the ease of implementation of individual student activities compared to that of group activities.

All the four modules analyzed in this study used point-and-click-based software such as Excel and/or HEC-SSP, a software for statistical analysis of hydrologic data developed by the U.S. Army Corps of Engineers, for data analysis. This may indicate that the choice of data analysis method is associated with the kind of STEM discipline that a module comes from as well as the academic level of students. For upper-level undergraduate modules developed in technology and mathematics/statistics courses, instructors might find it practicable to use script-based programming languages like Python as a tool for data analysis [14]. However, using such a data analysis method for lower-level undergraduates and/or undergraduates in disciplines such as environmental science and civil engineering might not be suitable [2]. In fact, in many previous studies, it has been claimed that using a script-based programming language for data analysis can be intimidating to students and often beyond the scope of what content-based lecture courses can support [2].

The instructors in three out of the four modules decided to publish their module through their institutions' LMS, Canvas for the modules developed at VT (i.e., Errors in measured data and Effects of Acid Rain on Aquatic and Terrestrial Ecosystems) and Blackboard for the module developed at NCA&T (i.e. Rainfall-runoff analysis using real-world high-frequency data). The only module that used a specialized LMS (called hydrolearn.org in this case) is the module developed in the senior/graduate class Hydrology at VT and designed to be implemented online (i.e., Frequency Analysis in Hydrology). This suggests the instructors' choice of platform for the publication of their modules is guided by the specific features a platform provides that can facilitate instructors' workflow.

Finally, except for the module Errors in Measured Data developed in the senior class Monitoring and Analysis of the Environment which was designed to be implemented over a single typical class session, instructors designed their modules to be implemented over multiple typical class sessions. However, for both the module Effects of Acid Rain on Aquatic and Terrestrial Ecosystems and Rainfall-runoff analysis using real-world high-frequency data instructors decided to only dedicate a portion of the time of their class sessions each time they deployed their modules. For the online module Frequency Analysis in Hydrology, the instructor estimates that on average it takes 15 to 20 hours for a student to complete the module at a self-paced manner. This estimated time is equivalent to multiple typical class sessions. As such, the module was categorized as a multiple-session module.

4.2. Data Science Topics Categories

Analyses of the broad data science categories across all the four modules found that 30 prompts out of 36 come from Data Use and Visualization and two from each of the board categories Data

Quality Issues, Data Acquisition, and Miscellaneous. No modules were found to be aligned with the last two broad categories of Machine learning and Data Ethics, Privacy, and Security. The number of prompts belonging to each of the broad categories is not equal but variable, from no prompts aligned in the broad categories of Machine Learning and Data Ethics, Privacy, Security to 30 out of 36 prompts aligned in the broad category of Data Use and Visualization.

The distribution of prompts is highly skewed towards the broad category Data Use and Visualization, and the subcategories Data Interpretation and Statistical Analysis within this broad category (Table 3). The distribution of prompts across other broad categories is sparse with no prompts categorized within the last two broad categories which might indicate the topical inadequacies of the modules given the importance and utility in the fields of science and engineering from which the four modules come.

Table 3. Count of Prompts in each Data Science Category/Subcategory

No.	Broad Data Science Topic Categories	Data Science Topic Subcategories	Count of prompts
1	Data Use and Visualization	Data Interpretation	18
		Statistical Analysis	9
		Visualization	3
2	Data Acquisition	Data Access	1
		Data Measurement	1
		Data Collection Mechanisms including Sensors	0
3	Data Quality Issues	Errors in Measured Data	1
		Uncertainty in Data Collection	1
4	Miscellaneous	Real-world Application	1
		Data Presentation	1
		Check Model Assumptions	0
5	Machine Learning	Supervised Methods	0
		Unsupervised Methods	0
6	Ethics, Privacy, and Security	Ethical Issues	0
		Data Privacy	0
		Data Security	0

Irrespective of academic levels, disciplines, and universities, the greatest number of prompts in each module belong to the broad category Data Use and Visualization (Figure 1). 19 out of 20 prompts from the three modules Effects of acid rain on aquatic and terrestrial ecosystems, Rainfall-runoff analysis using real-world high frequency data, and Errors in measured data come

from the category Data Use and Visualization. The only one prompt from the module Errors in measured data that is not categorized into the Data Use and Visualization is about data presentation which is close to the subcategories Data Interpretation and Visualization from the Data Use and Visualization but with a focus on oral communication. The one prompt in the module Frequency Analysis in Hydrology that is categorized in the Miscellaneous category introduces a real-world case study that requires the students to conduct a series of tasks that are mostly aligned with the Data Use and Visualization category. One reason for the prevalence of the data science category Data Use and Visualization is the fact that it involves basic data wrangling, analysis, and visualization techniques that anyone handling any type and quantity of data must deal with, such as using a histogram to visualize a quantitative dataset and interpreting its distribution. The non-existence of more advanced topics such as machine learning in the four modules we analyzed implies that the use of such advanced data science techniques might exist in highly specialized undergraduate courses and that in other undergraduate courses with less data analytics focus, instructors tend to use less specialized data science techniques, such as the ones categorized in Data Use and Visualization.

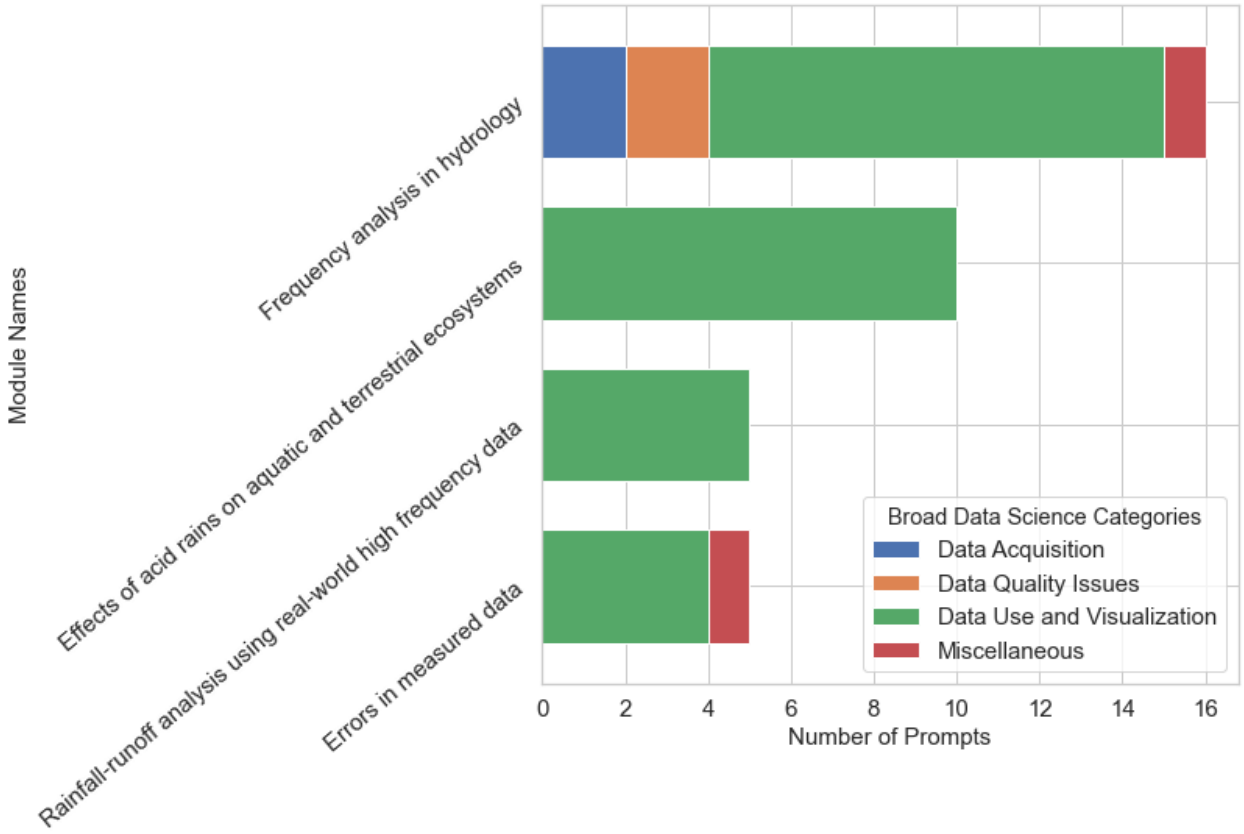

Figure 1. Count of prompts across modules and broad data science categories

It is only the stand-alone online module Frequency Analysis in Hydrology from the senior/graduate level class Hydrology course that involves prompts aligned with the categories Data Acquisition and Data Quality Issues. The discussion of data quality issues in this module is

likely influenced by the traditional focus on the quality of collected hydrologic/hydraulic data and the difficulty in maintaining data collection systems for collecting such data in hydrology and water resources engineering. Similarly, the existence of data acquisition topics such as data access and data measurement in only this module might be as a result of demonstrating such techniques as data visualization, statistical analysis, and data and/or analysis interpretation with actual hydrologic time series rather than dummy data. That's why accessing readily available data through online portals (such as that of the USGS's portal) and data repositories and how such data has been captured using a multitude of sensors are discussed in this module.

5. Limitations and future research

As an initial step on this topic, there are some limitations to our study. One such limitation is in how we defined the approaches used by instructors when they developed and integrated data science instructional materials into their STEM courses. We believe both the components of the approach as well as the categories within each of these components could be made more comprehensive. For instance, the components of the approach could become more complete by including information about the interaction between modules and the courses in which each of the modules has been developed. Moreover, the document data about each of the modules and courses could be coupled together with post-course instructor interview data to create a more precise context as to the decisions instructors made during both the development and deployment of the modules. Furthermore, the categorizations within each component can be made more flexible by adding more categories to preserve the uniqueness of situations in each of the cases.

Another limitation is how data science topics were extracted from each of the modules. Currently, the topics were identified using only the assessment prompts from each of the modules. This approach to the extraction of data science topics might oversimplify the topical context of each module to the wide variability between individual modules developed through a bottom-up approach in which different instructors developed their own teaching modules independent of each other. This variety is reflected in how instructors have chosen to assess student learning outcomes in different modules. Therefore, using a more holistic approach to the assessment of data science topics which is not only looking at the module assessment prompts but the entire module, as well as information about the course in which the module has been developed along with the opinion of the instructor of the course, can provide a more descriptive topical context discussed in each of the modules.

6. Conclusion

The research-practice partnership in this study has a four-phase bottom-up organizational structure of 1) development of principles and expectations of the project, 2) development and deployment of modules, 3) refinement of the modules, and 4) adapting modules for

multidisciplinary use. The initial phase of the partnership produced a systematic modular framework based on shared instructional and assessment principles that was flexible enough to allow instructors to construct data science modules that are tailored to meet their disciplinary, academic level, and pedagogical requirements and needs. This framework allowed instructors from three different universities (i.e., VT, NCA&T, and VU) to develop and integrate 12 modules, including the 4 modules discussed in this study, into their respective courses.

When developing and integrating data science learning objectives into their courses, instructors must answer questions about what data science topics to include and how to include them into their curricula. The results of this study suggest that the answers to both questions depend on the disciplinary requirements and learning goals of instructors' courses as well as the academic levels of their students. For example, if an instructor wants to develop for and integrate data science learning objectives for a lower-level non-technical undergraduate course, they might only need to incorporate such topics as the ones categorized in the Data Use and Visualization broad category in this study. Also, during deployment, they might need to provide more scaffolding to their students by, for example, using point-and-clicks software instead of using a script-based programming language for data analysis and group based classwork instead of projects as a method of student learning outcome assessment. However, with increasing academic level and technicality of their students and courses, instructors might need more advanced topics such as the ones categorized in Data Acquisition, Data Quality Issues, as well as Machine Learning broad categories and might not need a high level of scaffolding during the deployment of their modules.

Acknowledgments

This research is supported by NSF grants #1029711, #1915487, and #1915268.

References

[1] Science, Computer, Telecommunications Board, and National Academies of Sciences, Engineering, and Medicine. "Data Science for Undergraduates: Opportunities and Options." (2018).

[2] K. J. Farrell, & Carey, C. C. (2018). Power, pitfalls, and potential for integrating computational literacy into undergraduate ecology courses. Ecology and evolution, 8(16), 7744-7751.

[3] C. Snyder, Asamen, D. M., Naseri, M. Y., Aryal, N., Biswas, G., Dubey, A., ... & Xia, K. (2021). Understanding Data Science Instruction in Multiple STEM Disciplines.

[4] J. M. Durden, Luo, J. Y., Alexander, H., Flanagan, A. M., & Grossmann, L. (2017). Integrating "big data" into aquatic ecology: Challenges and opportunities. Limnology and Oceanography Bulletin, 26(4), 101-108.

[5] J. Hardin, Hoerl, R., Horton, N. J., Nolan, D., Baumer, B., Hall-Holt, O., ... & Ward, M. D. (2015). Data science in statistics curricula: Preparing students to "think with data". The American Statistician, 69(4), 343-353.

[6] A. R. Rao, Y. Desai, and K. Mishra, "Data science education through education data: an end-to-end perspective," 2019 2019: IEEE, doi: 10.1109/isecon.2019.8881970. [Online]. Available: https://dx.doi.org/10.1109/isecon.2019.8881970 10.1080/00031305.2015.1077729.

[7] U. Berkeley. "Online Master's in Data Science." Berkeley School of Information. https://ischoolonline.berkeley.edu/data-science/ (accessed 2022).

[8] F. Maina, Smit, J., Serwadda, A., "Professional Development for Rural Stem Teachers on Data Science and Cybersecurity: A University and School Districts' Partnership," *Australian and International Journal of Rural Education,* vol. 31, no. 1, pp. 30-41, 2021.

Coburn, C. E., Penuel, W. R., & Farrell, C. C. (2021). Fostering educational improvement with research-practice partnerships. Phi Delta Kappan, 102(7), 14-19.

[9] G. Wiggins, McTighe, J., *Understanding by Design*. 1998, p. 11

[10] P. Delgoshaei & V. K. Lohani (2014). Design and application of a real-time water quality monitoring lab in sustainability education. *International Journal of Engineering Education, 30*(2), 505-519.

[11] J. W. Creswell, *Research Design: Qualitative, Quantitative and Mixed Methods Approaches* (4th ed.). Thousand Oaks, CA: Sage, 2014 pp. 183-213.

[12] B. Baumer, (2015). A data science course for undergraduates: Thinking with data. The American Statistician, 69(4), 334-342.

[13] R. D. De Veaux, Agarwal, M., Averett, M., Baumer, B. S., Bray, A., Bressoud, T. C., ... & Ye, P. (2017). Curriculum guidelines for undergraduate programs in data science. Annual Review of Statistics and Its Application, 4, 15-30.

[14] D. Asamoah, Doran, D., & Schiller, S. (2015). Teaching the foundations of data science: An interdisciplinary approach. arXiv preprint arXiv:1512.04456.