Understanding Data Science Instruction in Multiple STEM Disciplines

Caitlin Snyder, Vanderbilt University

Caitlin Snyder is a PhD student in the department of Computer Science at Vanderbilt University. Her research focuses on understanding how students work collaboratively in open-ended learning environments with the end goal of developing semi-automated analysis tools for researchers and teachers.

Mr. Dawit M Asamen, North Carolina A&T State University
Mr. Mohammad Yunus Naseri
Dr. Niroj Aryal
Dr. Gautam Biswas, Vanderbilt University

Gautam Biswas is a Cornelius Vanderbilt Professor of Computer Science, Computer Engineering, and Engineering Management in the EECS Department and a Senior Research Scientist at the Institute for Software Integrated Systems (ISIS) at Vanderbilt University. He has an undergraduate degree in Electrical Engineering from the Indian Institute of Technology (IIT) in Mumbai, India, and M.S. and Ph.D. degrees in Computer Science from Michigan State University in E. Lansing, MI. Prof. Biswas conducts research in Intelligent Systems with primary interests in hybrid modeling, simulation, and analysis of complex embedded systems, and their applications to diagnosis, prognosis, and fault-adaptive control. He is also involved in developing simulation-based environments for learning and instruction. In his research, he has exploited the synergy between computational thinking ideas and STEM learning to develop systems that help students learn science and math concepts by building simulation models. He has also developed innovative educational data mining techniques for studying students' learning behaviors and linking them to metacognitive strategies. Prof. Biswas is a Fellow of the IEEE and the PHM society.

Prof. Abhishek Dubey
Dr. Erin Henrick, Vanderbilt University
Dr. Erin R Hotchkiss, Department of Biological Sciences, Virginia Tech

www.hotchkisslab.com

Dr. Manoj K Jha P.E., North Carolina Agricultural and Technical State University

Dr. Manoj K Jha is an associate professor in the Civil, Architectural, and Environmental Engineering department at the North Carolina A&T State University. His research interests include hydrology and water quality studies for water resources management under land use change and climate change. His educational research interests include critical thinking and active learning.

Dr. Emily C Kern
Dr. Vinod K Lohani, Virginia Polytechnic Institute and State University

Dr. Vinod K. Lohani is a Program Director at the National Science Foundation and his portfolio includes the NSF Research Traineeship (NRT), Innovations in Graduate Education (IGE), and CAREER programs. Dr. Lohani is on leave from Virginia Tech where he is a Professor of Engineering Education. During 2016-19, he served as the Director of education and global initiatives at an interdisciplinary research institute called the Institute for Critical Technology and Applied Science (ICTAS) at Virginia Tech. He is the founding director of an interdisciplinary lab called Learning Enhanced Watershed Assessment System (LEWAS) at VT. He received a Ph.D. in civil engineering from VT. His research interests are in the areas of computer-supported research and learning systems, hydrology, engineering education, and international collaboration. He has served as a PI or co-PI on 30 projects, funded by the National Science Foundation, with a $8.4 million research funding participation from external sources. He directed/co-directed an NSF/Research Experiences for Undergraduates (REU) Site on interdisciplinary water sciences and engineering at VT during 2007-19. This site has 100+ alumni to date. He also led an NSF/Research Experiences for Teachers (RET) site on interdisciplinary water research during 2016-19 with 30+ alumni. He also led an NSF-funded cybersecurity education project and served as a co-PI on two International

Research Experiences for Students (IRES) projects funded by the NSF. He has published over 90 papers in peer-reviewed journals and conferences.

Dr. Landon Todd Marston, Virginia Polytechnic Institute and State University
Dr. Christopher P Vanags, Vanderbilt University
Dr. Kang Xia, Virginia Polytechnic Institute and State University

Kang Xia received her Ph.D. from the University of Wisconsin-Madison (1997), M.S. from Louisiana State University (1993), and B.S. from Beijing Agricultural University (1989). She was a Postdoctoral Researcher at the University of Wisconsin-Madison (1997-1998), an Assistant Professor at Kansas State University (1998-2001), University of Georgia (2002-2005), and Assistant Professor, Dept. of Chemistry, Mississippi State University (2006-2010), an Associate Professor at Mississippi State University (2010-2011) and at Virginia Tech (2011-2016). She also served as Director for Re-search Division and Industrial and Agricultural Services Division, Mississippi State Chemical Laboratory (2006-2011). She is currently a Professor at Virginia Tech (2016-present). She has served as adhoc reviewer for a number of scientific journals and funding agencies. She served as associate editor for the Journal of Environmental Quality and the Soil Science Society of America Journal. She is an expert on method development for analysis of organic chemicals in environmental matrixes and environmental occurrence, fate, and impact of organic chemicals. She has successfully managed and accomplished close to $11 million federal and state funded interdisciplinary environmental projects. She has published 67 peer-reviewed papers, 6 book chapters, and given 126 professional presentations. She holds membership of the American Chemical Society , the Soil Science Society of America, and SigmaXi.

Understanding Data Science Instruction in Multiple STEM Domains

Abstract

As technology advances, data driven work is becoming increasingly important across all disciplines. Data science is an emerging field that encompasses a large array of topics including data collection, data preprocessing, data visualization, and data analysis using statistical and machine learning methods. As undergraduates enter the workforce in the future, they will need to "benefit from a fundamental awareness of and competence in data science"[9]. This project has formed a research practice partnership that brings together STEM+C instructors and researchers from three universities and an education research and consulting group. We aim to use high frequency monitoring data collected from real-world systems to develop and implement an interdisciplinary approach to enable undergraduate students to develop an understanding of data science concepts through individual STEM disciplines that include engineering, computer science, environmental science, and biology. In this paper, we perform an initial exploratory analysis on how data science topics are introduced into the different courses, with the ultimate goal of understanding how instructional modules and accompanying assessments can be developed for multidisciplinary use. We analyze information collected from instructor interviews and surveys, student surveys, and assessments from five undergraduate courses (243 students) at the three universities to understand aspects of data science curricula that are common across disciplines. Using a qualitative approach, we find commonalities in data science instruction and assessment components across the disciplines. This includes topical content, data sources, pedagogical approaches, and assessment design. Preliminary analyses of instructor interviews also suggest factors that affect the content taught and the assessment material across the five courses. These factors include class size, students' year of study, students' reasons for taking class, and students' background expertise and knowledge. These findings indicate the challenges in developing data modules for multidisciplinary use. We hope that the analysis and reflections on our initial offerings has improved our understanding of these challenges, and how we may address them when designing future data science teaching modules. These are the first steps in a design-based approach to developing data science modules that may be offered across multiple courses.

1. Introduction

As technology advances, familiarity and expertise in data-driven analysis is becoming a necessity for jobs across many disciplines. Data science is an emerging field that encompasses a large array of topics including data collection, data preprocessing, data quality, data visualization, and data analysis using statistical and machine learning methods. A recent National Academy of Sciences report recommends that in order to prepare students for the proliferation of data driven

work “academic institutions should encourage the development of a basic understanding of data science in all undergraduates” [9]. However, it is unclear how to put this into practice, especially across courses in multiple disciplines. Our research practice partnership, defined as “a long-term collaboration aimed at educational improvement and transformation through engagement with research, intentionally organized to connect diverse forms of expertise and to ensure that all partners have a say in the joint work” [19]. This partnership includes STEM (Science, Technology, Engineering, and Math) and CS (Computer Science) instructors and researchers from three universities and an education research and consulting group. We are conducting research and development work to address the absence or lack of clarity in data science instruction across multiple disciplines by using high frequency monitoring data from real-world systems to develop and implement an interdisciplinary approach that enables undergraduate students to develop relevant data science expertise through disciplinary STEM courses. These courses include hydrology and civil engineering, environmental sciences, ecology, engineering statistics, and an interdisciplinary undergraduate course in smart city applications.

In this paper, we perform exploratory analysis towards a design-based research approach with the goal of understanding how different instructors implement data science topics in their courses. Our initial efforts in going through the module development process is motivated by the goals of this project: *creating modules that can be integrated with small modifications into courses across multiple STEM+C disciplines*. As a first step toward this goal, this paper discusses how different instructors implemented data science topics in their courses, by considering the differences in the courses themselves and the level of students who were enrolled in these courses.

Curricula for most of these courses leave little room for accommodating additional material and assessments, Therefore, data science topics have to be taught as components of domain-specific instruction. Clearly, unless this is well thought out, it becomes hard to identify common approaches to teaching and assessing data science topics across these disciplinary courses. To gain a better understanding of such pedagogical approaches that can be applied across different courses, we analyze data collected through instructor interviews and surveys as well as student surveys across five undergraduate courses at the three universities. In addition, we perform preliminary analyses on the assignments associated with data science content that were administered in each course. We adopt a qualitative approach to code and analyze the survey data with the goal of identifying commonalities in data science instruction and assessment components across the disciplines. These include data sources, data science topics, class structure, and assessment design. We also find variations across the five courses, including class size, student year, students’ reasons for taking the class and students’ background expertise and knowledge. Finally, we discuss the implications that these commonalities and variations have on developing multidisciplinary data science modules and the challenges encountered.

2. Background

This project uses real world high frequency monitoring data to develop and implement data science modules. This section describes the two labs that provide this high frequency data.

2.1 Learning Enhanced Watershed Assessment System (LEWAS)

The LEWAS monitors high- frequency (1-3 min.) water and weather data from a small urban stream on the Virginia Tech (VT) campus (watershed: 2.78 km^2) with documented water quality issues [5-7]. Water and weather parameters (flow rate, water temperature, dissolved oxygen, specific conductance, turbidity, pH, rainfall, air temperature, air pressure, humidity) sensed by the LEWAS can be accessed remotely in real-time through OWLS, an open-ended, guided cyber-learning system developed and refined for education and research since 2013 [1-3] . The LEWAS field site drains a highly urbanized residential and commercial area. Students are able to study the quick response times of a small urban watershed using real-time, high-frequency water and weather monitoring equipment [4]. Although the physical location of the LEWAS site promotes hands-on research and education on water quality, the primary design of the LEWAS is to reach a wider audience using an engaging and interactive web-driven platform. The platform, OWLS, broadcasts live data from the field site through an interface that encourages users to visually explore and analyze data collected from the Webb Branch watershed [7].

OWLS promotes "active learning" through modules that connect the participants to the field site and motivates them to actively participate in learning activities focused on environmental data monitoring. To promote widespread use, OWLS uses an HTML5 web interface to deliver system data in multiple forms and types: visual, environmental, geographical, etc. in a platform-independent manner. It is being constantly updated to make the interface more interactive. For example, a JavaScript based visualization library called Data-Driven documents (D3), has made the Live graph page more interactive. We have developed a number of case studies that are available from the OWLS site and can be easily integrated into courses. An example included in OWLS is illustrated in Fig. 1 which depicts changes in water temperature during a summer thunderstorm with hail from Sept. 28, 2016. During the thunderstorm, the water temperature quickly rose by several degrees as warmer rainwater entered the stream. After a typical storm, the temperature would decrease exponentially back toward its pre-storm level, but this event was quite different. After the rain ended, the water temperature rapidly declined before rising steadily in a roughly linear manner. Finally, a second, smaller storm brought the temperature close to pre-storm levels. This unusual temperature pattern was the result of water from melting ice entering the stream. This example clearly illustrates how the "dynamics" of both temperature and flow in this urban watershed can be captured with high-frequency data.

Figure 1. Case Study Example from the OWLS site

2.2 Smart City Lab

The Smart City Lab at Vanderbilt University(VU) is a multidisciplinary initiative that includes faculty from computer science, civil engineering, earth and environmental sciences and education. This lab works closely with urban communities focusing on three application domains: transportation, emergency response, and building energy management. For this purpose we have built four core modules that help with (a) high- resolution data acquisition and storage, (b) feature selection and model development, (c) model validation and (d) data visualization. Currently, these four modules are domain specific, i.e., they are not reused across the different application projects we have built. It includes: (a) models for identifying non-recurring congestions [11], (b) services to suggest optimal routes for trips [12], (c) services to suggest improved bus routes [13], and (d) modules to simulate different transportation modification strategies [14]. For example, we can simulate the effect of preferential selection of public transportation buses as compared to use of personal cars in the city.

Similarly, the emergency response toolchain includes modules for processing data about all the incidents that have occurred in the city in the past, then developing models to predict the likelihood of incidents in different areas of the city, and finally developing algorithms to suggest stationing and dispatch of emergency responders in anticipation of future incidents [15]. Figure 2a shows a map of the traffic incidents that have occurred in the city of Nashville over the period of a year. The figure clearly shows that a larger number of incidents happen in and around the downtown area, which is more congested, and where a number of large events, such as conventions and music concerts take place. This data is used to build stochastic models that predict incident occurrence, and use them to to reduce the average incident response time as described above, while ensuring that the emergency response vehicles are not driven more than a set mileage per month to reduce the maintenance costs.

The building management application also integrates similar loops of data collection, model-learning and application to building temperature set point control for improving building energy efficiency [16]. A tool chain exists for pulling heating and cooling energy data from buildings on the VU campus through a BACnet connection, and preprocessing the raw data collected before it is stored in local databases on our server for further analysis. Students in the Smart Cities course have worked on a number of projects using this data. Figure 2b shows an example of energy consumption that has been used in one student project. These projects include using AI, Machine Learning, statistical, and visualization methods to study: (1) energy reduction in buildings; (2) effect of occupancy on building energy consumption; and (3) scheduling of loads in office buildings with labs to reduce overall energy consumption in the buildings.

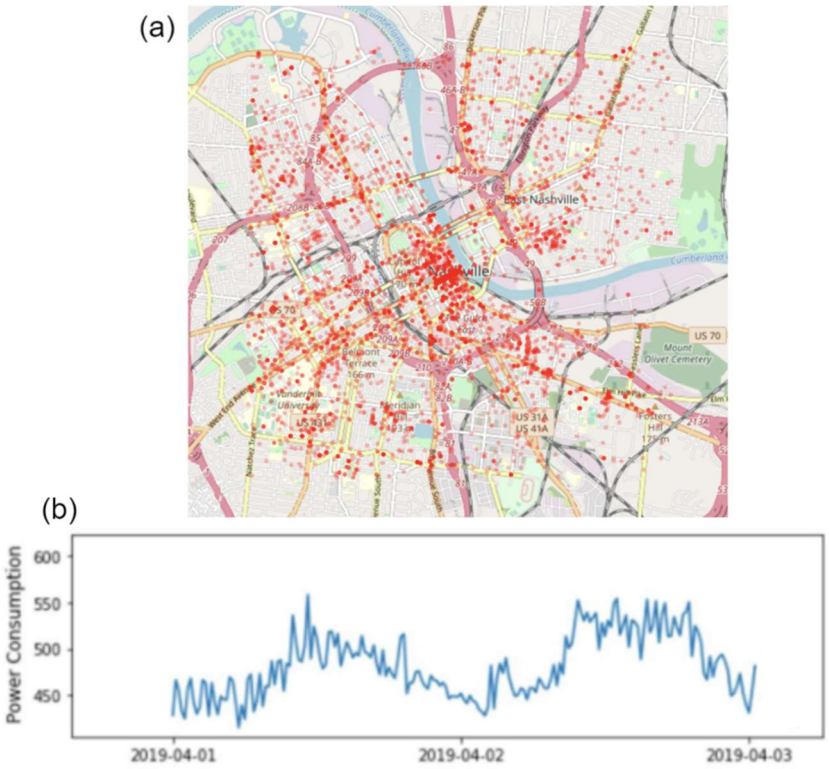

Figure 2: Example of data utilized from Smart City Lab. (a) Traffic accident data (b) Building power consumption

Working with these three application areas we have developed a good understanding of what are the key common capabilities that we can educate the next generation of engineers and scientists on. These core capabilities are: (1) Reliable Infrastructure for data collection and analysis; it is important to educate the students about the importance of privacy as well; (2) Application of machine learning and data analytics across multiple domains; and (3) Distributed application development for deployment of services and applications in an efficient manner.

3. Framework

3.1 Module Types

Based on the NSF guidelines for "Data Science: the science of planning for, acquisition, management, analysis of, and inference from data" [10], this study focused on ***Data Analysis and Interpretation through Interdisciplinary Learning***. As seen in Figure 3, this includes activities on the basic fundamentals of data acquisition and data quality issues with the major goal of ensuring students understand the genesis and uncertainties of data by introducing them to the data collection process and conditions.

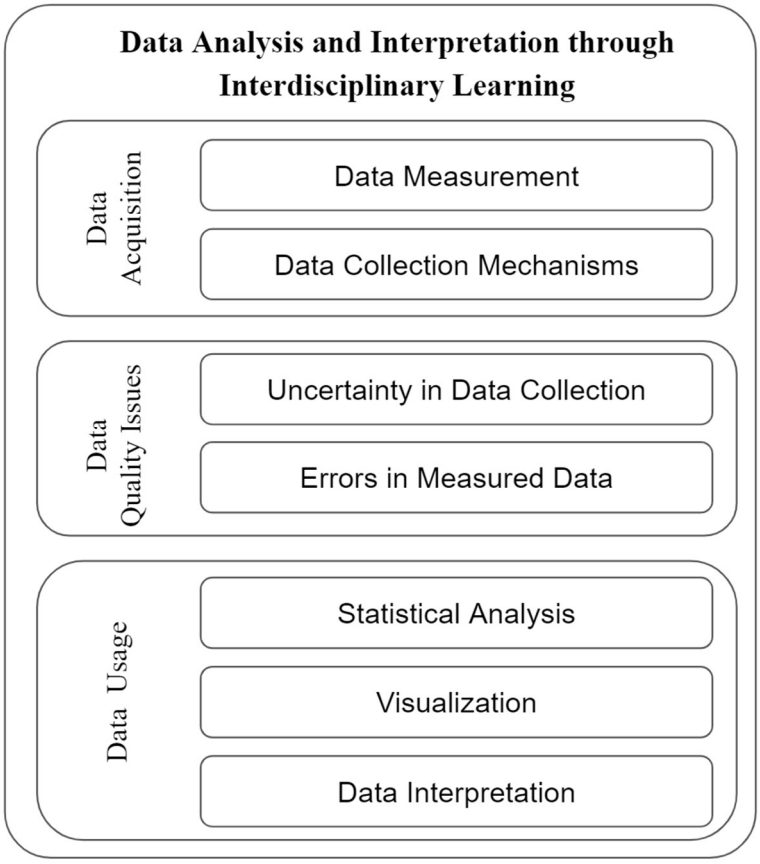

Figure 3: Data Analysis and Interpretation through Interdisciplinary Learning topics (see descriptions below).

- Data Measurement: Students will understand the data being measured, its meaning, and how they are measured. Simply collecting data that represents a measure or a process without an understanding of the meaning often leads to poor assumptions, analysis, error evaluation, and ultimate decision making.
- Data Collection Mechanisms: Students will be introduced to a variety of digital data collection systems and their sampling rates will be explained and demonstrated.

This will include an overview of the variety of sensors in use by both the LEWAS and the Smart City Lab and their scientific principles. An understanding of the technology used to collect the data is imperative to evaluate the quality of the data. The role of embedded computing, e.g., the use of Raspberry Pi's as data collection and local analysis (e.g., for noise removal) devices will be studied. Students will need to understand the tradeoff between frequency of monitoring and power requirements, and learn how to calibrate and measure the accuracy of sensors.

- Uncertainty in Data Collection: Students will learn the possible reasons why data may be erroneous and the uncertainty bounds around a data value is critical in Data Science.
- Errors in Measured Data: Students will learn about key methods needed to deal with experimental design, measurements, and statistics and to minimize error propagation.
- Statistical Analysis: Students will develop an appreciation for data preparation and transformation, an understanding of the data requirements for the various algorithms and learn to match algorithms to specific problem needs. Specific topics that will be covered include the basics of statistical inference, testing statistical hypotheses, and building confidence intervals to report results, distance and similarity measurement that becomes the basis for regression, classification and clustering algorithms.
- Visualization: Students will be introduced to spatial and temporal representations of data and learn key processing methods necessary to create these data analysis representations for analysis.
- Data Interpretation: Students will learn how to interpret statistical analysis results and data visualizations. Presentation of results will be an additional key component covered in this module.

In addition, some of the more advanced courses, such as an interdisciplinary Smart Cities course at VU, used a number of machine learning algorithms that ranged from the use of support vector regression and deep learning methods to analyze transportation and building energy data.

3.2 Module Development Tool

The introduction of data science modules into very different multi-disciplinary courses requires modules to be *integrated* into the courses instead of replacing discipline-specific material with general data science modules. The creation of data science modules that can be easily integrated into a variety of courses is a difficult task. In the initial phase of our module development process, instructors individually created and taught modules that were specific to their courses. The identification of commonalities across course-specific data science instruction will inform the next step of the development process: creating multidisciplinary modules. To support the initial phase of the module development processes, we developed a module development tool that creates a framework for comparing course-specific modules. This tool covers the following components:

- Student learning goals: These are discipline-neutral learning goals that cover the key concepts and abilities that students should learn based on this module.
- Student assessments: This covers how students are assessed based on the learning goals before and after the module.
- Student activities: This describes what students will do in and out of class during the course of the module. These descriptions include whether students work in groups and what background knowledge is necessary. Finally, the activity description included what student level these activities are appropriate for and how the module activities can be made more challenging or simpler.
- Lesson plans: This covers the details of how many class sessions this module will cover, the instructor's role and the materials necessary for instructor preparation.
- Data sources and software: This describes the data sources and software used in the module.
- Project information: This covers how this module supports students use and analysis of high-frequency real-time data. It also explains how this module supports students' evaluations of the efficacy of data collection systems.

4. Data Collection and Analysis

Five different courses were analyzed during the Spring 2020 semester: (1) Monitoring and Analysis of the Environment; (2) Ecology; (3) Data Science Methods for Smart Cities Applications; (4) Engineering Hydrology; and (5) Engineering Statistics. The courses were taught at Virginia Tech (VT), North Carolina Agricultural and Technical State University (NC A&T), and Vanderbilt University (VU).

For each of the five courses, we collected instructor post semester interviews and surveys, student pre- and post-semester data science perception surveys, and student data science competency assessment assignments. The student data science competency assessments were course-specific, but some of the assignment questions were shared between courses within one university or across two universities. In one case, the Engineering Statistics course at NC A&T used building energy data from VU to study statistical analysis methods that applied to real-world data. The data set was a simpler version of the data set used for a VU Smart Cities course project. The student surveys were voluntary and did not affect their grades for the course. The pre- and post-semester surveys included multiple choice, open-ended and Likert-scale questions that asked about students' perspectives on the importance of data science in their future jobs, previous data science training and their confidence in their data science abilities. Due to the low post-survey response rates that may be attributed to the situation created by the COVID-19 pandemic, in this paper, we only utilized data from the pre-surveys for each of the five classes: Monitoring and Analysis of the Environment (n = 31), Ecology (n = 69), Smart Cities (n = 21), Engineering Hydrology (n = 28), and Engineering Statistics (n = 39).

Specifically, we analyze the responses to three open-ended survey questions : "*Why did you decide to take this class?*", "*What data science training have you received to date?*" and "*What comes to mind when you hear the term "data science?*" The open-ended questions were coded using higher-level codes that were developed to allow comparison between answers. The coding scheme that captures student's reasons for taking the course includes the following: (1) interest in the course material, (2) interest in data science, (3) prior experience with related course material, (4) structure of course, (5) whether the course was a requirement; and last (6) how would the course support career development and future use of data science. For example, the coding scheme used for student background in data science included the following codes: none, personal experience, previous coursework, exposure to specific software/tools, research experience, internship or job, and other. For each survey question, we calculated the percentage of answers that contained each code with respect to the total pre-survey responses for each class. Note that not all students answered every survey question and some students' answers fell under multiple codes, resulting in total percentages that were less or more than 100%.

In addition to student surveys, we collected instructor surveys that included 19 items asking, "*How much did you teach the following data science concepts in your class?*" Instructors could select one of the following four options: not taught, a little, some, and a lot. The topics were collected from our team's original identification of data science concepts and the Wittenberg data science learning goals rubric [17]. Additionally, instructors participated individually in semi-structured interviews lasting 30-60 minutes through Zoom, approximately two weeks after the end of the semester, when grades had been submitted, and instructors had received student feedback in the form of course evaluations. The interviews were recorded and transcribed for analysis. The interviews included questions on class structure, instructors' data analytics learning goals, software and data sets usage, student performance and ability. Using the student and instructor survey answers, instructor interview data and data module descriptions, we qualitatively compared courses according to the following categories: (1) general course information (2) student information, and (3) course data science components.

5. Results

5.1 General Course Information

We can see differences in the basic information for each course in Table 1. The courses occur in a variety of disciplines ranging from engineering to natural science courses. Similarly, the number of students in each course ranges from small (24 students) to large (84 students). The courses share the commonality of a lecture-based course structure, although some courses have an additional element such as a group project or lab. The instructor's role in a small lab-focused course was quite different from that of a large lecture course. Module design must take into account these differences in the instructor role across courses.

Table 1. General Course Information from Year 1 of our Project

Course	Student description	Course Structure Design	Student Total
Monitoring and Analysis of the Environment	Environmental Science majors at VT, typically taken as a senior	Industry-preparation course with hands-on lab work including collecting and using environmental sensors and data	35
Ecology	Sophomore level course taken by a majority of pre-med majors at VT	Lecture-based course supplemented with discussion with neighboring students in class and short activities outside of class	69
Smart Cities	General university course taken by juniors, seniors, graduate students at VU	Lecture-based in the first half of the semester with a transition into group project work for the final half. Unlike the other courses, there were multiple instructors	24
Engineering Hydrology	Junior level course at NC A&T	Lecture-based course supplemented with individual projects	31
Engineering Statistics	Sophomore-level course at NC A&T that is taken by electrical and mechanical engineering students	Lecture-based course	84

5.2 Student Information

We calculated the following proportions of student survey answers for each course: the reasons students took the course, their previous data science experience and their interpretations of the term "data science." Table 2 presents the percentage of student reported reasons for taking the course with respect to total survey responses for each course. Recall that student answers to these open-ended questions were converted into the categories described above. As such, individual answers may be counted in more than one category. In all courses, with the exception of Smart Cities, approximately 80% or more of the student answers in each course indicated that students had chosen to take the course because it was a requirement. Students also rarely expressed an interest in data science as the reason for taking the course. Designing multi-disciplinary modules that utilize the high frequency real world data from the LEWAS and Smart City lab may benefit these students because it introduces students to data science concepts using situated learning [18]. Students have the opportunity to connect classroom learning to experiences in corresponding physical lab environments.

Table 2. Percentage of student reported reasons for taking the course with respect to total survey responses for each course

	All Classes	Monitoring and Analysis of the Environment	Ecology	Smart Cities	Engineering Hydrology	Engineering Statistics
Requirement	72%	81%	79%	0%	79%	87%
Interest in Course Subject	22%	10%	31%	57%	10%	5%
Career or Future Use	10%	26%	1%	33%	7%	5%
Structure of Course	8%	16%	1%	33%	4%	3%
Prior experience with Subject	5%	3%	0%	38%	0%	0%
Interest in Data Science	5%	0%	1%	38%	0%	0%
Other	1%	6%	0%	0%	0%	0%

Table 3, below, shows the percentages of reported previous data science experience with respect to total survey responses for each course. Students' previous experience with data science varied across courses. With the exception of the Engineering Hydrology and Engineering Statistics students, the majority of the student answers indicated some data science experience, with 20% or fewer of the student answers indicating no data experience. More than half the student answers in Monitoring and Analysis of the Environment, Ecology and Smart Cities indicated they had gained experience in Data Science through previous coursework (77%, 57% and 52% respectively). Fifty-seven percent of Engineering Hydrology and 36% of Engineering Statistics students reported having no experience in data science. These results indicate that data science modules will have to be developed across multiple difficulty levels taking into account the varying prior knowledge of the students. Modules integrated in the Smart Cities course may be able to go more in-depth into data science topics, considering that only 10% of students have no previous experience and nearly half reporting exposure to specific tools/software related to Data Science algorithms. In contrast, while 77% of students in the Monitoring and Analysis of the Environment class indicated experience with data science in previous coursework, only 3% had exposure to specific tools/software. The low percentages of exposure to specific tools/software for non-Smart Cities

courses implies that modules must introduce students to appropriate data science tools, and their use in analysis and problem solving contexts.

Table 3. Percentages of student reported previous data science experience with respect to total survey responses for each course

	All Courses	**Monitoring and Analysis of the Environment**	**Ecology**	**Smart Cities**	**Engineering Hydrology**	**Engineering Statistics**
Previous Coursework	48%	77%	57%	52%	7%	36%
None	28%	19%	20%	10%	57%	36%
Exposure to Specific Tools/Software	13%	3%	13%	48%	0%	10%
Personal Experience	9%	10%	13%	24%	0%	0%
Research Experience	8%	10%	6%	19%	4%	5%
Internship or Job	3%	0%	1%	14%	7%	0%
Other	2%	0%	1%	0%	0%	8%

5.3 Data Science Components

In order to find out which modules would be most applicable to multiple courses, we analyzed the instructor survey responses to see how much individual data science topics were covered in the courses, as seen in Table 4. We have responses from four of the instructors. All topics listed in the table were taught at minimum “a little” in at least one course. The instructors indicated that the most commonly taught topics in these courses related to applying data science methods, visualizing data, and uncertainty in data. The least taught topics related to industry use of data analytics, and analytics and mining systems. This information informs initial module development through the identification of common data science topics that were taught across courses. This will additionally inform the implementation portion of the module development process by identifying which courses can integrate the same multi-disciplinary models in their courses. It also informs the partnership about modules that instructors may need to refine and add to their syllabus to present data science topics and applications of data science in more systematic ways through our RPP process.

Table 4. Instructor responses to data science topics

Data Science Topic	**Instructor Responses Total Score (max value = 12) and individual instructor scores in Individual scores: 3=A Lot, 2=Some, 1=A Little, 0=Not Taught.**
Apply data science concepts and methods to solve problems in real-world contexts	10 (3, 3, 2, 2)
Apply data analytic methods to datasets	9 (3, 3, 2, 1)
Create visualizations of data	9 (3, 2, 2, 2)
Uncertainty in data	9 (3, 3, 2, 1)
Statistical inferences of error in measurement	8 (3, 2, 2, 1)
Data management	7 (3, 3, 1, 0)
Measuring data	6 (3, 2, 1, 0)
Sensors	6 (2, 2, 1, 1)
Statistical analyses with professional statistical software	6 (2, 2, 2, 0)
Analyze high frequency real-time systems	5 (2, 1, 1, 1)
Digital data collection systems	5 (2, 1, 1, 1)
Evaluate the efficacy of the data collection system	4 (2, 2, 0, 0)
Data collection mechanisms using connected sensor networks	4 (2, 1, 1, 0)
Build and assess data-based models	4 (3, 1, 0, 0)
Analytics and mining systems	3 (3, 0, 0, 0)
Industry use of data analytics	3 (2, 1, 0, 0)

As seen in Table 5, we also identified data sources used and data software utilized for each of the courses. All courses used Microsoft EXCEL or Google Sheets. The Smart Cities course introduced students to Python libraries for Machine Learning algorithms, and students used Google Colab to complete their assignments. Taking into account the data experience results in Section 5.2, this may be due to the experience level of the students (and their instructors' familiarity with data science tools). In terms of module development design, the common use of EXCEL or Google Sheets across courses suggests that students may need support in using spreadsheets, especially in using advanced functionalities that would be more appropriate for data science applications and learning of data science concepts.

Table 5. Data sources and software

Courses	**Data Sources**	**Data Software**

Monitoring and Analysis of the Environment	LEWAS	EXCEL
Ecology	LEWAS and other local sensor data; Ocean Observations Initiative[21]	Google Sheets, NetLogo Predator Prey Simulation Interface[20]
Smart Cities	Smart Cities Lab and a Hydrology database	Python and Google Colab
Engineering Hydrology	US Geological Survey Water Data	EXCEL
Engineering Statistics	Smart Cities Lab Energy data from a building)	EXCEL

We also analyzed three course-specific modules that were developed according to the module development tool described in Section 3. Due to challenges such as full course schedules, different course structures, differences in student backgrounds, etc. our partnership has faced challenges in evolving a common module development process of identifying commonalities. However, by implementing course-specific modules and using the module development tool, we have been able to compare data science instruction in courses more easily. Table 6 presents the comparison of three teaching modules. All three teaching modules require students to perform statistical analysis and data visualization and to use EXCEL spreadsheets for data representation and calculations. While Module 2 does not cover errors in measured data, the other two modules do address this topic. In the next stage of the module development process, we plan to compare how each course-specific module covered topics such as errors in measured data.

Table 7. Example Teaching Modules

Module Tool Topic	**Module 1 (Monitoring and Analysis of the Environment)**	**Module 2 (Engineering Hydrology)**	**Module 3 (Engineering Hydrology)**
Module Topic(s) Covered	Errors in measured data, Statistical Analysis, Visualization	Visualization and Statistical Analysis	Errors in measured data, visualization, statistical analysis and data interpretation
Learning Goals	Graphically present data, statistical analysis (t-test)	Categorize time-series data, basic statistical data analysis and visualization	Analyze and interpret high-frequency data

Student Assessments	Graph data, conduct t-tests, use p-values	Download data, create creation of boxplots, histograms and line graphs, calculating descriptive statistics like mean	Large dataset handling, data comparison, graph data, interpretation of analysis results
Student Activities	In-class group work, class presentations and discussion	Individual work of calculating descriptive statistics and boxplots	Group work consisting of plots, runoff ratios with a class presentation
Lesson Plans	Takes 0.5 hours and the instructor must present an example and facilitate discussion	Takes 1.5 hours and the instructor must demonstrate how to download data and calculate descriptive statistics and create visualization in an example	Takes 1.5 hours and the instructor must introduce the LEWAS and teach the concepts of rainfall runoff analysis
Data Sources/ Software	Source: LEWAS Software: EXCEL, R	Source: US Geological Survey Water Data Software: EXCEL	Source: LEWAS Software: EXCEL

6. Conclusions

This paper presents the results from our initial module development process of identifying course characteristics and commonalities across courses that will inform the development of interdisciplinary modules in year one of a three year NSF IUSE project. The results of a qualitative analysis performed with student and instructor data collected from five courses across three dimensions (general course information, student information, and data science components taught) show a number of commonalities and a variety of differences across courses. We identified a number of differences that affected the methods used for teaching data science modules, the tools used, and student assessments to evaluate the learning of data science concepts. These factors included: class size, the level at which the course was taught (sophomore, junior, and senior), students' reason for taking the course, and students' prior data science experience. Identified commonalities included class structure, data sources, data science topics covered, and assessment design.

The class size differential leads to one design challenge in regard to the instructor's role in delivering the data science module and assessing student work. For example, when data science modules are taught in regular classroom settings with 20 or more students, methods of analysis and tool use can only be presented in lecture and demonstration format, and most assessment has

to be done through homework assignments, both of which reduce the amount of interactivity and feedback that can be provided to the students. In lab courses, the instructor and students can be more hands on, and, therefore, data collection, data cleaning, and data analysis methods can be performed in a more hands on manner. The Smart Cities course was unique -- in the first half of the course students learned data science concepts and algorithms in traditional lecture and homework assignment format. The teaching assistant for this class had to be quite involved in guiding the students in using Machine Learning algorithms, Python libraries, and the problem solving environment (Google Colab) because the students came from different disciplines (computer science, engineering, social sciences, and the humanities). However, in the second half of the course, students worked in interdisciplinary teams on projects that involved real world problems, which covered the technical, social, and humanistic aspects of developing solutions. Since the projects were open-ended, students had to think about the relevant machine learning approaches to use to analyze their data, and how to combine quantitative and qualitative data (from interviews) to analyze problems and propose solutions. As the project goes forward, we will have to consider how to categorize our various courses by level as well as objectives in developing appropriate data science modules that may be applicable across courses.

Additional design challenges we encountered include students' data science experience, student status (i.e., freshmen, sophomores, etc) and their ability to use different data processing and data analysis tools. Design and implementation of data science modules must be tailored to the students' background in both data science concepts and software experience. One such tailoring approach will include categorizing modules according to expertise level and creating tasks that cover the same analysis methods but can be intensified by introducing advanced software or more data. Commonalities across courses can inform our data science module development process by identifying data science topics, such as applying data science methods to solve real world problems, that are most easily integrated into a variety of courses. The design of effective data science modules will require supporting students as they utilize software, such as EXCEL, which many students may not have efficient exposure to. Additionally, using common data sources and software, such as the LEWAS data source and EXCEL, during module design will also create an easier process when implementing our data science modules in courses. An additional fact that needs to be conveyed to our instructors through engaged discussion is that modern data science analysis and applications are very tool-based, as opposed to writing code. This should be exploited in future module and assessment development as we move forward. For example, even if we continue to use EXCEL or Google Sheets, students should be made aware of sophisticated macros they can use to perform more advanced computations and analyses with their data. In the junior and senior level courses, it may make sense to use Matlab machine learning packages or develop interfaces that access Python libraries for performing advanced computations and visualization. This is a challenge we will need to address in subsequent years of our project.

Overall, our team of instructors and education researchers have come together to identify a number of these challenges we have discussed in this paper, and there is a realization on how we

need to engage and collaborate to overcome some of these difficulties and structure our efforts to accomplish our project goals. Our evaluators have been critical to guiding us through the process, supporting the data collection and interviews, and making us more aware of how to conduct a successful research practice partnership. The detailed analyses of our first year efforts, and identifying the challenges we have faced, will lead to a better understanding of how to develop and apply data science modules across multiple courses in our project, and eventually provide modules that can be used more generally across many universities and courses.

Acknowledgements

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

References

1. D. Brogan, V.Lohani, & R. Dymond (2014). *Work-in-Progress: The Platform-Independent Remote Monitoring System (PIRMS) for Situating Users in the Field Virtually.* Paper presented at the Proc. 2014 ASEE Annual Conference & Exposition, Indianapolis, IN.
2. D. S. Brogan, W. M McDonald., V. K. Lohani, R. L. Dymond, & A. J. Bradner (2016). Development and Classroom Implementation of an Environmental Data Creation and Sharing Tool. *Advances in Engineering Education (AEE).*, vol. 5, p. n2, 2016.
3. D. S. Brogan, D. Basu, & V. K. Lohani (2016). A Virtual Learning System in Environmental Monitoring *Engineering Education for a Smart Society* (pp. 352-367): Springer.
4. H. A. Clark, W.McDonald, V. Lohani & R. Dymond (2015). Investigating the Response of a Small Urbanized Watershed to Acute Toxicity Events via Analysis of High Frequency Environmental Data. American Journal of Undergraduate Research (AJUR), 12(3), 5-17.
5. 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.
6. P. Delgoshaei., C. Green, & V. Lohani (2010). *Real-time water quality monitoring using labview: Applications in freshman sustainability education.* Paper presented at the American Society for Engineering Education.
7. P. Delgoshaei, & V.K. Lohani (2012). *Implementation of a real-time water quality monitoring lab with applications in sustainability education.* Paper presented at the American Society for Engineering Education.
8. P. Delgoshaei (2012). Design and Implementation of a Real-Time Environmental Monitoring Lab with Applications in Sustainability Education. Doctor of Philosophy dissertation, Department of Engineering Education, Virginia Tech.
9. National Academies of Sciences, Engineering, and Medicine. *Data science for undergraduates: Opportunities and options*. National Academies Press, 2018.
10. National Science Foundation (NSF), 2014. Data Science at NSF, https://www.nsf.gov/attachments/130189/public/Data_Science_at_NSF-Draft_Report_of_StatSNSF_Committee.pdf (accessed Dec. 10, 2018).

11. Fangzhou Sun, Abhishek Dubey, Jules White, *DxNAT - Deep Neural Networks for Explaining Non-Recurring Traffic Congestion*, IEEE BigData 2017 - 3rd Special Session on Intelligent Data Mining, December 11-14, 2017, Boston, MA, USA.
12. Fangzhou Sun, Chinmaya Samal, Jules White and Abhishek Dubey, *Unsupervised Mechanisms for Optimizing On-time Performance of Fixed Schedule Transit Vehicles*, 2017 IEEE International Conference on Smart Computing, May 28-30, 2017, Hong Kong, China.
13. Aparna Oruganti, Fangzhou Sun, Hiba Baroud, Abhishek Dubey, *DelayRadar: A Multivariate Predictive Model for Transit Systems*, IEEE Big Data 2016 Conference Special Session on Intelligent Data Mining, December 5-8, 2016, Washington D.C. USA.
14. Chinmaya Samal, Liyuan Zheng, Fangzhou Sun, Lillian J. Ratliff, Abhishek Dubey, *Towards a Socially Optimal Multi-Modal Routing Platform*, ACM Transactions on Cyber-Physical Systems (TCPS) (Under Review)
15. Ayan Mukhopadhyay, Yevgeniy Vorobeychik, Abhishek Dubey, and Gautam Biswas. 2017. *Prioritized Allocation of Emergency Responders based on a Continuous-Time Incident Prediction Model*. In Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems (AAMAS '17). International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 168-177.
16. A. Naug and G. Biswas, *Data Driven Methods for Energy Reduction in Large Buildings,* 2018 IEEE International Conference on Smart Computing (SMARTCOMP), Taormina, Sicily, Italy, 2018, pp. 131-138. doi:10.1109/SMARTCOMP.2018.00083.
17. https://www.wittenberg.edu/academics/data-science/learning-outcomes
18. Greeno, J., Collins, A., & Resnick, L. (1996). Cognition and Learning. Handbook of Educational Psychology, 15-46: Macmillan Library Reference USA.
19. C. Farrell, W.R. Penuel, C. Coburn, J. Daniel, L. Steup (under review). Research-practice partnerships today: the state of the field.
20. Wilensky, U. (1997). NetLogo Wolf Sheep Predation model. http://ccl.northwestern.edu/netlogo/models/WolfSheepPredation. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.
21. https://datalab.marine.rutgers.edu/explorations/productivity/index.php