2023 Annual Conference & Exposition
Baltimore Convention Center, MD | June 25 - 28, 2023
The Harbor of Engineering Education for 130 Years
ASEE
Paper ID #39119

Board 58: WIP: Enhancing Workforce Development of Data Science Skills within Domain-Specific Programs

Dr. Ryan L. Solnosky P.E., Pennsylvania State University
Prof. Rebecca Napolitano
Wesley F. Reinhart, Pennsylvania State University

WIP: Enhancing Workforce Development of Data Science Skills within Domain Specific Programs

Abstract

In 2018, The National Academies of Sciences, Engineering, and Medicine identified a need for undergraduate students to have access to critical data science skills development opportunities. Over the next several decades, the world's reliance on cloud computing and big data will continuously increase, and new data-centric technologies and engineering approaches will be developed. Due to this rapidly developing field, there is a need to track these trends and incorporate the corresponding developments into our current science and engineering curriculum. Besides data science skills already taught in traditional engineering curricula, such as mathematical, computational, and statistical foundations, the National Academies guide discusses that key concepts in developing data acumen include domain-specific considerations and ethical problem-solving.

This work-in-progress (WIP) paper will highlight the foundation of a comprehensive study to explore data science education in two domain-specific programs: material science and engineering and architectural engineering. This project is broken down into the following objectives: 1) facilitate data science education and workforce development for engineering and related topics, 2) provide opportunities for students to participate in practical experiences where they can learn new skills through opportunities in new settings to transform data science education, and 3) expand the data science talent pool by enabling the participation of undergraduate students with diverse backgrounds, experiences, skills, and technical maturity. The paper will focus on the topics, deployment strategies within courses and curricula, establishing data sets, representative examples of work-in-progress efforts and their success.

Keywords: data science education; contextualized learning; modular course materials; workforce development

Introduction

The integration of digital literacy and data has grown exponentially over the last 15 years [1-2] to the point where the US Bureau of Labor Statistics projected careers in Computer Science (CS) fields were the fastest growing [3]. A more recent study by the National Academies of Sciences, Engineering, and Medicine conducted in 2018 [4] emphasized the crucial requirement for the improvement of data science skills. This is noteworthy, considering that North American high schools have experienced a 20-year decline in computer concept opportunities [5]. Students entering college between 2036-2040 are expected to see a strong reliance on cloud computing, big data, and new data-centric engineering approaches [6-7]. The National Academies Guide [8] proposes developing robust ways to educate the future workforce. A key attribute within this guide revolves around concepts that develop data knowledge from domain-specific considerations and ethical problem-solving. Zakaria [9] identified that to be successful in these approaches, it is vital to track and incorporate emerging trends into our current engineering curriculum to provide the proper context. The current state of data science education in traditional engineering curricula often involves the teaching of mathematical, computational, and statistical foundations through non-contextualized classes. This approach, however, has been shown to limit student engagement, motivation, and lead to shallow learning among disinterested students [10]. To address these issues, this study aims to answer the following research questions: (1) Does participating in engineering-specific, contextualized programming courses significantly improve students' self-reported understanding of data analysis tasks? And (2) Is there a difference in the average performance of students in the architectural engineering and materials science departments for different data science tasks?

Why Contextualized Approaches

Educational studies that have examined dedicated CS courses that all majors take have shown to limitedly engage and motivate non-CS students [11-13]. The limitations of non-contextualized data science deliveries are often because the presented materials fail to emphasize the relevance of the concepts, leading

to difficulty for students in recognizing the significance of the content for their careers in engineering and its practical applications. This has been noted as a hindrance in the appreciation of the material and its perceived usefulness [14]. This approach has been historically driven by curricula aimed at preparing software developers and computational scientists [15-16]. If context is absent, it has been documented that students will try to place context into their learning themselves with varying levels of success or even misinterpretation [17-18]. The arguments favoring contextualized approaches have shown varying levels of empirical data demonstrating improved student learning [1,19]. One example of a positive impact is from Forte and Guzdail [11], who observed improved motivation and computational thinking when data science skills were put into the context of a given major. According to Yardi [16], appropriately formatted and scoped content can enhance conceptual understanding, problem-solving skills, and reflective learning among other benefits. Other research indicates that both faculty and students are more satisfied with courses that adopt this approach, leading to higher course success rates and increased enrollment [20]. However, there is still a need for further research to fully understand the potential impact of contextualized approaches in education [21].

Despite the advantages of contextualizing data science education to domain-specific knowledge, as documented in the literature [22], college curricula still feature non-contextualized data science course offerings to a significant extent. Part of the barrier to adoption is rooted in [23-24]: 1) number of credits in a program typically is tight so making room is hard for new materials [25], 2) some faculty see non-core discipline topics as general education and do not feel it is needed at a department level [26], 3) if it must be covered in a class, it limits the amount and exposure of other topics, 4) many current faculty are not trained or are experts in data science and do not know the topics to teach them, 5) creating examples and projects is one delivery mechanism but there could be a steep learning curve student will encounter [27], 6) current demands from larger employers who may not all use these techniques, and lastly [28]; 7) Creating new tracks is possible but requires new resources and faculty to teach them. Given these benefits and challenges, many engineering students are still often pushed to take computer science course(s) to compensate for their lack of in-department offerings. This research looks to help overcome several aspects of these barriers in the discipline specific domains of architectural engineering (AE) and material science and engineering (MATSE). Both fields were selected given their renewed emphasis and need for more data skills as their design approaches are changing.

Research Methodology

For this research project, our expansion towards data-centric skills centers on modularized learning of key data science concepts for easy adoption. To better scope and guide the project, the following three objectives drive our design: (1) Facilitate data science education and workforce development for engineering and related topics, (2) Provide opportunities for students to participate in practical experiences where they can learn new skills through opportunities in new settings to transform data science education, (3) Expand the data science talent pool by enabling the participation of undergraduate students with diverse backgrounds, experiences, skills, and technical maturity. The result of this project is to overcome known and document hurdles in traditional non-contextualized data science deliveries by establishing resources to easily adopt into curriculums and to lower the entry burden for faculty unfamiliar with data science from the contextualized perspective of AE and MATSE.

To achieve these broad objectives, a three-year study funded by NSF will develop training program (TP) courses and curate datasets that will be available so that any adopter of the material can include them in their undergraduate instruction. To provide broad impact, a website to house the data science training programs; curated, didactic datasets; and teach-the-teacher resources. These TP resources advance and permit interested faculty members to port any part of this data science education framework into their own institutions. The goal is that this alleviates some burden from professors adopting this into their own institutions. Five scoping principles were established to align the created materials with the project goals. They are:

- Establish the curated training programs by targeting them towards undergraduates at the intersection of data science and engineering,
- Establish materials in a "plug-and-play" fashion for easy adoption into existing curricula without the need to revamp or recreate materials at a fundamental level.
- Establish at least one dataset from each industry and community partners to ensure real-world practical application and exposure to actual data.
- Establish Teach-the-teacher and inter-institutional translation documentation in the form of a webinar, self-reflection materials, best practices documentation, and shared feedback from prior professors who taught the material
- Establish a website that covers the following attributes: a Q&A forum for professors, repository for educational materials, surveys, and example code tailored to AE and MATSE students, repository for community related datasets, and teach-the-teacher and inter-institutional translation documentation.

A Contextualized DS Approach in MATSE and AE

A review of the most prevalent and useful data-centered skills was conducted to ensure that emerging graduates entering the workforce possess the necessary skills for success. When looking at Architectural Engineering (AE) and Material Science and Engineering (MATSE) topics for material creation and inclusion, seven key areas were identified. Five of these topics were directly applicable to both AE and MATSE, they are: intro to data science, data management, machine learning, advanced data science, and data science ethics. An additional two topics were identified that focus only on AE given industry advancements in computational design; they are: Parametric Modeling and Analysis and Robotics.

In the creation of each of these seven courses, their development will follow a standardized process when applicable. When TP materials are finalized, lessons are intended to be integrated in a "plug-and-play" fashion for easy adoption into existing curricula. For example, a professor could pull two lessons from one course, three lessons from another course, and then integrate these within the lessons they had previously taught. Additionally, each TP lesson is intended for a broader audience than just engineering students in higher education—they may be useful for practitioners in engineering fields, educators in additional contexts we have not considered, and extracurricular student organizations. All lessons are being designed assuming a hybrid classroom format.

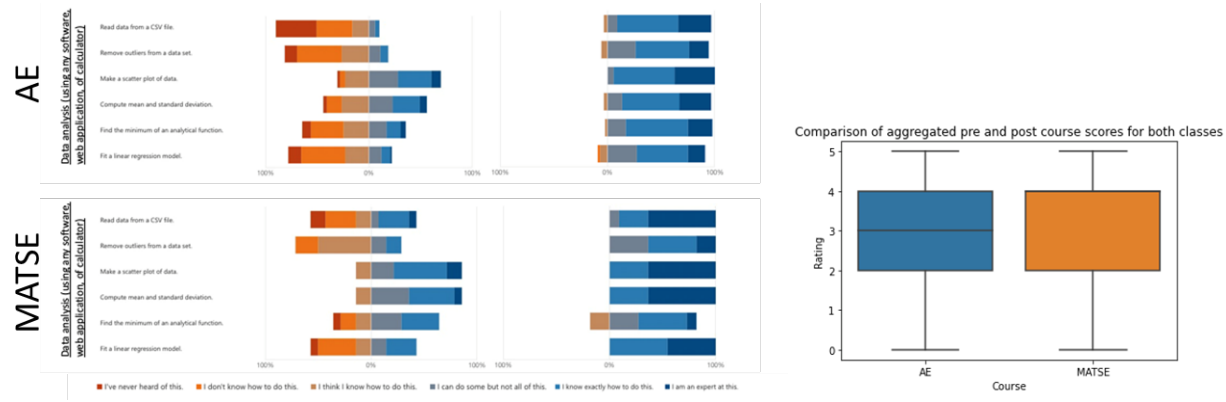
Sample Results

At the time of writing, materials are being developed and generated across the 7 topics. To assess the educational impact of the developed materials, evaluations were conducted to inform the design and measure the benefits of the introduction to data science topics. This section presents representative results from the initial deployment of these materials. The introduction to data science topics include: (i) understanding data with Excel, (ii) data visualization, (iii) understanding documentation, (iv) correlation and regression, (v) indexing, iterating, and logic, (vi) modeling nonlinear relationships, (viii) writing documentation, (ix) interfacing Python with computer-aided design programs, and (x) interfacing Python with complex analysis software. The data was taken from two contextualized undergraduate programming courses, one in architectural engineering AE 240 (n=114) and one in materials science MATSE 297 (n=12). AE 240 is offered for in the spring semesters and is required for all AE students while MATSE is offered in the Fall semesters and is optional for MATSE at this time.

To collect the data for evaluations two things were done, first was a pre and post survey with several data science concept inventory questions along with several student perception questions. The second data source was student performance grades. Surveyed data was collected at the start of the class (before) and at the end of the class (after). Student perception questions covered 6 tasks each with a Likert scales ranging from 0 to 5 where 0 indicates that the student had never heard of the concept, 1 indicates that the student did not know how to do the task, 2 indicates that the student thinks they know how to do this task, 3 indicates

that the student thinks they can do some but not all of this task, 4 indicates that the student believes they know exactly how to do the task, and 5 indicates that the student believes they are an expert at this.

Pre- and post-data for these six items are presented in Figure 1a. Here we can see significant changes from start of class to end of class. Considering this data, we can compare the means of the different tasks between the departments to see if there is a significant difference in the average performance between the two departments. Based on the results (Figure 1a and 1b), it appears that the average score for the tasks in MATSE (3.24) is higher than the average score for the tasks in AE (2.88).



A) pre and post Likert distributions B) aggregated pre & post scores

Figure 1: Student perceptions of their topical skills

However, this alone does not necessarily indicate a significant difference. To determine whether the difference is significant, an ANOVA test was completed comparing the means between the two departments. The F-value and p-value obtained from running an ANOVA test on AE vs MATSE give us insight into the statistical significance of the difference in means between the two groups. Results from the ANOVA test are represented in Table 1. It can be seen in these results that for most of the tasks (reading data from csv, making a scatter plot, computing mean and standard deviation, and doing a linear regression), the p-value is below 0.05, which indicates that there is a significant difference in means between the two departments. However, for two tasks (removing outliers and finding the minimum of an analytical function), the p-value is above 0.05, which indicates that there is not a significant difference in means between the two groups for these two tasks. In summary, the ANOVA test suggests that the difference in means between the two departments is statistically significant for most of the tasks, but not significant for two of the tasks.

To provide more insight into these results, data was examined using an unpaired sample t-test using pre- and post- class to self-reported understanding of the concepts. For the different tasks respectively, the p values are recorded in Table 2. As all p-values are all less than a 0.05 significant level there is evidence of a significant difference between the two groups. Thus, these p-values indicate that the difference in the scores before and after the class is significant and not just due to random chance. Seeing these appreciable changes in the responses before and after the classes, each class individually was examined to understand if there were significant differences between the scores at the beginning and end of each class between the two departments (Table 2). The before class t-test results indicate that there is not a significant difference in the AE and MATSE for the tasks where the students were asked to plot a scatter plot or find a minimum. However, there was a statistically significant difference between the AE and MATSE scores in terms of reading data from a csv, removing outliers, computing basic statistics, and calculating linear regressions. For the after-class t-test results, indicates that there is not a significant difference in the AE and MATSE for the tasks where the students were asked to read csv data, remove outliers, plot a scatter plot, or find a minimum. However, there was a statistically significant difference between the AE and MATSE scores in terms of computing basic statistics and calculating linear regressions.

Table 1: MATSE and AE ANOVA Test Results

Task	F-Value	P-value	Significant?
Csv reading	5.831	0.016	Yes
Outlier detection	1.568	0.212	No
making a scatter plot	3.804	0.052	Yes
computing statistical values	5.937	0.015	Yes
fitting a linear regression	0.404	0.526	Yes
Finding the minimum of an analytical function	0.019	0.889	No

Note: The F-value represents the ratio of between-group variability to within-group variability, and the p-value represents the probability of obtaining the observed F-value if the null hypothesis is true, that there is no difference in means between the two groups.

Table 2: MATSE and AE unpaired t-test results

Task	p-Value For both classes before versus after implementation	p-value for AE vs MATSE before	p-value for AE vs MATSE After
Csv reading	4.40e-62	0.015	0.095
Outlier detection	7.52e-44	0.040	0.963
making a scatter plot	2.48e-20	0.060	0.074
computing statistical values	7.65e-20	0.014	0.008
fitting a linear regression	3.54e-35	0.007	0.006
Finding the minimum of an analytical function	3.38e-36	0.263	0.334

To visually compare the distribution of the scores to see a shift in the distribution of the "before" and "after" scores for each discipline a box plots of the scores can be created (Figure xxx A).

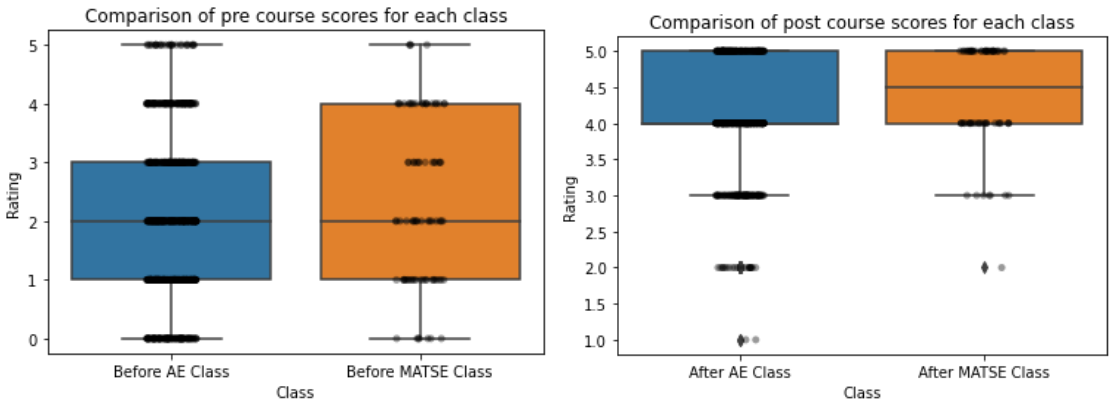


Figure 2: Pre- to post- discipline performance

Conclusions and Continued Work

Consider our first research question: “Does participating in engineering-specific, contextualized programming courses result in significant improvements in students' self-reported understanding of data analysis tasks?” Based on the results from the statistical analysis above, we can state that the overall difference between the scores (before and after the class) is significant and not just due to random chance. Consider our second research question: “Is there a difference in the average performance of students in the architectural engineering and materials science departments for different tasks?” The statistical testing has shown there is not a significant difference in the AE and MATSE for the tasks where the students were asked to plot a scatter plot, remove outliers, find a minimum, or read csv data. However, there was a

statistically significant difference between the AE and MATSE scores in terms of computing basic statistics and calculating linear regressions. These results suggest that there may be differences in the way that the two departments approach and understand these tasks. To further investigate these results, it may be necessary to conduct additional tests or collect more data to get a better understanding of these differences. Conclusions we can draw from our current work in progress are that the change in the students' self-reported understanding of the concepts between the start and end of the class is more pronounced for the computing basic statistics and calculating linear regressions tasks compared to other tasks. Additionally, the results suggest that the material taught in both departments was able to positively impact the students' understanding of the concepts, as evidenced by the statistically significant difference in the scores before and after the class.

Overall, these conclusions and areas for future work can help to improve the education and understanding of data analysis tasks in both departments. Based on these results further work can be done to examine: (1) How does the teaching style of the instructors in the architectural engineering and materials science departments' impact students' self-reported understanding of different tasks? (2) What are the factors that contribute to the significant difference in students' self-reported understanding of computing basic statistics and calculating linear regressions between the architectural engineering and materials science departments? (3) How can the curriculum in the architectural engineering and materials science departments be modified to improve students' self-reported understanding of different tasks?

The limitations of this study include its reliance on self-reported data, which may be biased, and a limited sample size of only two departments, leading to results that may not generalize. Furthermore, the data collected was from a single instance and may not accurately reflect the overall performance. The study only looked at self-reported understanding of data analysis tasks and did not measure actual performance. Another study limitation is that there was not a control study done for programming courses offered in traditional computer science departments to compare contextualized vs non-contextualized impacts. Future studies should consider incorporating performance-based assessments, controlling for factors such as prior knowledge, experience, or motivation, and measuring the impact of factors beyond the class on student performance.

Funding

This material is based upon work supported by the National Science Foundation under Grant IIS-2123343. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

References

[1] Grgurina, N., & Yeni, S. (2021). Computational thinking in context across curriculum: students' and teachers' perspectives. In Informatics in Schools. Rethinking Computing Education: 14th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP 2021, Virtual Event, November 3–5, 2021, Proceedings 14 (pp. 3-15). Springer International Publishing.

[2] Bocconi, S., Chioccariello, A., Dettori, G., Ferrari, A., Engelhardt, K., Kampylis, P., & Punie, Y. (2016). Developing computational thinking in compulsory education. European Commission, JRC Science for Policy Report, 68.

[3] "Computers Home Page," http://stats.bls.gov/k12/computers.htm, accessed July 2010.

[4] National Academies of Sciences, Engineering, and Medicine. Data Science for Undergraduates: Opportunities and Options. [Internet]. Washington, DC: The National Academies Press; 2018. Available from: https://doi.org/10.17226/25104.

[5] Radev, D., & Levin, L. (2009). Engaging High School Students in Interdisciplinary Studies. Computing Research News, 21(4).

[6] Chang, H. C., Wang, C. Y., & Hawamdeh, S. (2019). Emerging trends in data analytics and knowledge management job market: extending KSA framework. Journal of Knowledge Management, 23(4), 664-686.

[7] Irizarry, R. A. (2020). The role of academia in data science education. Harvard Data Science Review, 2(1).

[8] National Academies of Sciences, Engineering, and Medicine. Envisioning The Data Science Discipline: The Undergraduate Perspective: Interim Report. [Internet]. Washington, DC: The National Academies Press; 2018. Available from: doi: https://doi.org/10.17226/24886
[9] Zakaria, M. S. (2022). Data science education programmes in Middle Eastern institutions: A survey study. IFLA Journal, 03400352221113362.
[10] Wastl, J., Porter, S., Draux, H., Fane, B., & Hook, D. (2020). Contextualizing sustainable development research. Digit. Sci.
[11] Forte, A., & Guzdial, M. (2005). Motivation and nonmajors in computer science: identifying discrete audiences for introductory courses. IEEE Transactions on Education, 48(2), 248-253.
[12] Bonfert-Taylor, P., & Ray, L., & Pauls, S., & Loeb, L., & Sankey, L., & Busch, J., & Hickey, T. (2022, August), Infusing Data Science into the Undergraduate STEM Curriculum Paper presented at 2022 ASEE Annual Conference & Exposition, Minneapolis, MN. https://peer.asee.org/41919
[13] Mahmoud, Q. H. (2005). Revitalizing computing science education. Computer, 38(5), 100-99.
[14] Hoegh, A., & Moskal, B. M. (2009, October). Examining science and engineering students' attitudes toward computer science. In 2009 39th IEEE frontiers in education conference (pp. 1-6). IEEE.
[15] Guzdial M. Does contextualized computing education help? ACM Inroads. 2010 Dec 1;1(4):4–6.
[16] Yardi, S. and Bruckman, A. 2007. What is computing?: bridging the gap between teenagers' perceptions and graduate students' experiences. In Proceedings of the Third international Workshop on Computing Education Research (Atlanta, Georgia, USA, September 15 - 16, 2007). ICER '07. ACM, New York, NY, 39-50. DOI=http://doi.acm.org/10.1145/1288580.1288586
[17] Jonassen, D. H. (2000). Revisiting activity theory as a framework for designing student-centered learning environments. Theoretical foundations of learning environments, 89, 121.
[18] Cooper, S., & Cunningham, S. (2010). Teaching computer science in context. Acm Inroads, 1(1), 5-8.
[19] Simon, B., Kinnunen, P., Porter, L., & Zazkis, D. (2010, June). Experience report: CS1 for majors with media computation. In Proceedings of the fifteenth annual conference on Innovation and technology in computer science education (pp. 214-218).
[20] Mohammadi A, Grosskopf K, Killingsworth J. An Experiential Online Training Approach for Underrepresented Engineering and Technology Students. Education Sciences. 2020 Mar;10(3):46.
[21] Kay, J. S. (2011, March). Contextualized approaches to introductory computer science: the key to making computer science relevant or simply bait and switch?. In Proceedings of the 42nd ACM technical symposium on Computer science education (pp. 177-182).
[22] Nagashima, T. (2018). Contextualized Instruction in Data Science and its Effect on Transfer of Learning. In EC-TEL (Doctoral Consortium).
[23] Robinson, L., Ragnedda, M., & Schulz, J. (2020). Digital inequalities: contextualizing problems and solutions. Journal of Information, Communication and Ethics in Society, 18(3), 323-327.
[24] Giannakos, M. N., Pappas, I. O., Jaccheri, L., & Sampson, D. G. (2017). Understanding student retention in computer science education: The role of environment, gains, barriers and usefulness. Education and Information Technologies, 22, 2365-2382.
[25] Snyder, C., & Asamen, D. M., & Naseri, M. Y., & Aryal, N., & Biswas, G., & Dubey, A., & Henrick, E., & Hotchkiss, E. R., & Jha, M. K., & Jiang, S. X., & Kern, E. C., & Lohani, V. K., & Marston, L. T., & Vanags, C. P., & Xia, K. (2021, July), Understanding Data Science Instruction in Multiple STEM Disciplines Paper presented at 2021 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://peer.asee.org/37955
[26] DeClue, T. (2009). A theory of attrition in computer science education which explores the effect of learning theory, gender, and context. Journal of Computing Sciences in Colleges, 24(5), 115-121.
[27] Suthar, K., & Mitchell , T., & Hartwig, A. C., & Wang, J., & Mao, S., & Parson, L., & Zeng, P., & Liu, B., & He, P. (2021, July), Real Data and Application-based Interactive Modules for Data Science Education in Engineering Paper presented at 2021 ASEE Virtual Annual Conference Content Access, Virtual Conference. https://peer.asee.org/37640
[28] Zhang, V. and Neimeth. C. “Changing Data Science”, InfoWorld, April 2017 https://www.infoworld.com/article/3190008/big-data/3-reasons-why-data-scientist-remains-the-top-job-in-america.html