TRANSLATING A LOWER-DIVISION DATA SCIENCE COURSE: LESSONS LEARNED AND CHALLENGES ENCOUNTERED

S. Wang[1], D. Harding[2], C. von Vacano[2]

[1]Mills College at Northeastern University (UNITED STATES)
[2]University of California, Berkeley (UNITED STATES)

Abstract

Translating courses and programs in data science education while fostering diverse perspectives can present significant challenges in today's skill-demand-driven landscape. Educators are increasingly realizing the value of adapting and translating successful courses and programs and sharing insights gained in promoting diversity within data science education. In this context, we detail our experiences adapting a lower-division data science course from the University of California, Berkeley, a large public R1 university, to Mills College, a small liberal arts college for female-identified students.

While both institutions share the goal of offering an introductory data science course to students from various majors, they differ markedly in terms of student populations, institutional environments, and class sizes. We outline the key modifications made in course infrastructure and content, and share lessons learned. These lessons, drawn from the balance of structure and flexibility, as well as experiences across different scales and institutional contexts, contribute to discussions on promoting diversity within the data science field. Throughout our adaptation process, we encountered challenges, and, in this paper, we discuss some strategies to overcome them.

Our findings underscore the importance of adapting courses to align with current curricula, student demographics, technological infrastructure, and the resources available to faculty. Additionally, smaller class sizes provide the opportunity to facilitate interactions and design tailored assignments that resonate with students' academic majors, career aspirations, and passion for driving social change.

Keywords: Education, data science, adoption.

1 INTRODUCTION

The emergence of data science in education presents a significant challenge for undergraduate education to meet the needs of a diverse workforce [1]. With the vast generation of data, data science is becoming indispensable across various fields of communities. A sharp increase in students' interest in data science has led to a rising demand for new courses in the field. Educators are increasingly recognizing the value of adapting and transferring successful courses and programs and sharing their insights gained in promoting diversity in data science education. Yet pedagogical methods and curricula cannot be directly transferred from one institution to another. Instead, faculty aiming to incorporate successful practices from different institutions must engage in systematic and deliberate efforts to adopt what has proven effective in one context, integrating it into their existing curriculum and customizing it to meet the interests and needs of their student population.

Numerous data science curricula have been proposed and implemented [1, 2, 3, 4, 5, 6, 7]. The University of California, Berkeley (UC Berkeley), stands out as an early pioneer in developing a comprehensive data science education curriculum. Its initiatives began with a focus on expanding lower-division data science courses, notably the foundational entry-level data science course.

This paper delves into key modifications made, the lessons learned, and the challenges encountered during the process of translating and adapting the Data 8 Foundations of Data Science course (Foundations) [8] from UC Berkeley to Mills College (Mills).

2 INSTITUTION CONTEXT

UC Berkeley is one of California's flagship public R1 research universities with a student population of over 42,000 and over 350 different degree programs. Twenty-two percent of undergraduates are underrepresented minorities, 21 percent of undergraduates are transfer students, and 23 percent of first-year admits are first-generation students. We estimate that almost one-quarter of UC Berkeley's

undergraduates take the Foundations course at some point in their college career. The class size for the Foundations course ranges from 1,300 to 1,900.

Mills was a small, nationally renowned liberal arts women's college with a student population of 960. (In July of 2022, Mills College merged with Northeastern University and became Mills College at Northeastern University.) As one of the most diverse liberal arts colleges in the country with 65% students of color, a Hispanic Serving Institution (HSI), 44% first-gen, and 58% LGBTQ+, Mills had a strong record of academic success with its students and a deep commitment to equity, inclusion, and social justice. The Mills experience was distinguished by small, interactive classes, one-on-one attention from exceptional faculty, a culture of creative experimentation, and cutting-edge interdisciplinary learning opportunities that empower students to make a statement in their careers and communities.

3 ADAPTATIONS

The Foundations course offered at UC Berkeley serves as an entry-level, four-unit semester-long course designed for students across all majors, requiring no prerequisites. It teaches critical concepts and skills in computer programming and statistical inference, in conjunction with hands-on analysis of diverse real-world datasets spanning economic data, document collections, geographical data, and social networks. By integrating data with Python programming skills, statistical inference techniques, modeling approaches, and hypothesis testing methods, the course delves into problem-solving across multiple domains. The format of the course consists of three one-hour lectures and a two-hour lab session each week. A wealth of academic resources for data science, including those pertinent to the Foundations course, can be accessed through [9].

Mills undertook the translation of the Foundations course, adopting it into their own entry-level data science offering titled DATA80A: Data Science for Everyone in the fall of 2021 with an enrollment of 20 students. In the subsequent discussion, we, henceforth referring to Mills, describe our adaptations grouped according to thematic considerations.

3.1 Curriculum

The initial consideration in translating the course centered on its integration within the existing curriculum and institution context. Given that the Foundations course is designed for students across all majors, we took into account factors such as students' workload, as adding this course would increase their academic responsibilities, and their financial capability to afford the course fees.

To maintain alignment between contact hours and credit allocation at Mills, and to incentivize student enrollment, we reduced the lab hours to one hour per week, as opposed to the two hours allocated at UC Berkeley. Furthermore, the course received approval to fulfill a programming requirement for the general BS degree, enhancing its attractiveness to students.

3.2 Technological infrastructure

The Foundations course employs JupyterHub [10], providing students with access to a computing environment and Jupyter notebooks via a web interface. This standardized interface streamlines computing for students, eliminating the need for individual laptop setups and software installations. However, setting up the JupyterHub infrastructure posed significant challenges initially and required a ramp-up period. The initial implementation encountered issues related to versioning, Python library dependencies, security, and authentication.

Despite these challenges, once the infrastructure was established, students found the authentication process, pulling files from GitHub, file access, and Jupyter notebook usage to be mostly smooth and seamless. This positive user experience reaffirmed the decision to implement JupyterHub. Given the Foundations course's heavy reliance on Jupyter notebooks, it was crucial to have the JupyterHub operational well in advance, with sufficient testing conducted to ensure its functionality.

Mill received support from UC Berkeley's IT staff and assistance from 2i2c [11] in setting up the JupyterHub infrastructure. While we had prior experience with Colab [12], a web-based computing environment used in upper-division computer science courses, we recognized that its multi-step process for working with data files and images might be challenging for novices encountering Jupyter notebooks for the first time.

In contrast to web-based computing environments like JupyterHub, Anaconda [13] offered an alternative by allowing students to download and install the software on their individual laptops. Once Anaconda was installed, students could run Jupyter notebooks. We provided detailed instructions on downloading Anaconda and installing essential libraries, such as datascience, a package designed for the Foundations course, and otter-grader [14], an autograder. This served as a backup plan in case JupyterHub experienced downtime, which was rare, and as an alternative for students preferring local computing environments.

Additionally, we conducted a brief lab session during the third week of classes to guide students through the process of running Jupyter notebooks on Anaconda, ensuring they were equipped with the necessary skills for both scenarios.

3.3 Autograding

Otter-grader is a convenient tool for autograding purposes. In addition to streamlining the grading process and saving valuable time and effort, it offered students the benefit of immediate feedback. This real-time feedback not only served as a motivational factor but also facilitated their progress towards devising multi-step solutions.

However, while autograding offered benefits, it also presented certain drawbacks. Some students became overly reliant on the autograding system, which, in turn, hindered their ability to engage in deep thinking. Furthermore, there were instances where it was challenging to autograde students' work without inadvertently providing them with the expected answer.

To address these challenges, we implemented autograding for labs, allowing students to receive prompt feedback on their work, which proved to be effective. For assignments and projects, which required more nuanced and detailed evaluation, we employed a combination of autograding and manual grading. This approach was feasible due to the small class size, enabling us to provide comprehensive feedback while leveraging the efficiency of autograding.

3.4 Content

We maintained the same lab, homework, and project structures as those in the Foundations course. However, our adaptation involved reconfiguring and distributing the course content across the semester to accommodate the reduced lab hours and allow time for problem-solving and discussions. At the beginning of the semester, we gauged students' backgrounds in computing and statistics through a survey, enabling us to pace the course at a level students could follow and actively engage.

Recognizing the importance of hands-on practice for students to grasp the material and build confidence in their coding and problem-solving skills, we augmented the curriculum with two additional homework assignments comprised of practice problems aimed at preparing students for the midterm and final exams. We replaced some homework and lab problems with local datasets that are better aligned with students' interests. Moreover, with the advantage of a smaller class size, we could flexibly adjust the coverage of topics and the course pace to better support student learning.

In addition to the modifications mentioned above, other content changes we made include:

a) Integration of an ethics component one-third way into the semester, featuring a new homework assignment, lab, and a guest lecture.

b) Replacement of the third project with one centered on linear regression, offering students the option to work with a prepared dataset or select one aligned with their interests. We provided resources for datasets and project ideas.

c) Due to time constraints, we omitted topics on probability and classification from the course.

3.5 Teaching Style

At Mills, our class sessions were conducted in an interactive manner, with time allocated for problem-solving, discussions, and Q&A sessions. Leveraging the advantage of a smaller class size, we were able to facilitate individual interactions with students, adjusting our approach to accommodate their learning styles, evaluate their understanding of the material, and identify areas of difficulty.

We emphasized the importance of reinforcing concepts through hands-on coding practice. A particularly effective strategy we employed involved presenting a problem and demonstrating its solution, primarily through Jupyter notebooks. Subsequently, we would present a similar problem, accompanied by a

relevant dataset, allowing students time to think through solutions, share ideas, and collaboratively work towards solving the problem.

Active engagement through hands-on practice during class not only provided students with opportunities to solidify their understanding but also enabled instructors to offer real-time guidance and support. This interactive approach facilitated a dynamic learning environment where students could actively participate and receive immediate feedback on their progress.

3.6 Unchanged Features

We have experienced a number of features particular to the adaptation of this course that have worked well, either unchanged or with some modifications. We believe these features, listed below, can be replicated at other institutions with relative ease.

- **No prerequisites** – open to all students
- **Datasets** – real-world, public, diverse fields
- **Jupyter notebooks** for lectures, homework, labs – adjustments made as needed
- **Jupyter notebook deployment** – Interact links and GitHub worked well for live demos
- **Textbook** – online and free
- **Lecture materials**: slides, notebooks, videos – adjustments made as needed

4 LESSONS LEARNED AND CHALLENGES ENCOUNTERED

An initial evaluation of how students received the adapted Foundations course at Mills in terms of motivations and perceptions of being in data science was reported in [15] in 2023.

4.1 Lessons

We highlight two major areas that merit further exploration for those considering adopting the Foundations course. First, Mills shortened the labs and adjusted the lectures to fit their institution's credit model and style of teaching and to better align with students' backgrounds. Second, smaller class sizes not only facilitated increased interaction with students but also allowed for experimentation in changed content and labs, making it possible to have more flexible assignments and projects. These assignments leveraged the diverse backgrounds, varied interests and mindsets of students by encouraging and supporting diversity in thought and dialogue.

4.2 Challenges

We encountered three sets of challenges throughout the course. The first set relates to content and delivery, where certain topics lacked clarity in explanation and adequate support from datasets. For instance, probability seemed disconnected from future concepts, while topics like histograms and randomization required clearer explanations. Additionally, more practice problems were needed for computation-intensive topics such as histograms, p-values, and confidence intervals.

The second set of challenges pertains to the technical nature of the course, with some students struggling to grasp computing and statistical concepts amidst a fast-paced course. This challenge also reflects the broader institutional context. One potential solution involves offering a short module to familiarize students with key computing and statistical concepts before enrolling in the Foundations course. Another option is to provide an alternative course that focuses on social science and ethical considerations in data science, catering to students less inclined towards intensive computing, possibly satisfying different requirements in the curriculum. This approach would allow for course variations tailored to diverse student populations while preserving the core principles of data science foundations.

The third set of challenges revolves around meeting students' demand for comprehensive resources, including practice problems of varying difficulty levels, supplemented with datasets and notebooks sourced from diverse disciplines to engage and challenge them. While the textbook offers numerous examples, it falls short in providing end-of-chapter problems. Incorporating additional problems would greatly enhance student learning experiences.

4.3 Reflections for UC Berkeley

The experiences and adaptations made by Mills carry lessons for UC Berkeley's data science program as well, particularly with regard to best serving students from groups traditionally underrepresented in STEM fields. First, course material on topics related to human contexts and ethics in data science can serve as an entry point for a wider range of students by demonstrating the relevance of technical material to contemporary social problems and public policies. Second, opportunities for more individualized assignments and projects can also motivate students to continue the pursuit of their own interests within data science. Third, explicit and ongoing discussions of the importance of diversity in data science may help to provide students from marginalized groups with a greater sense of belonging and identification with data science as a field.

5 CONCLUSIONS

In this paper we detail our experiences adapting a lower-division data science course from the University of California, Berkeley, a large public R1 university, to Mills College, a small liberal arts college for female-identified students. Our findings underscore the importance of adapting courses and programs to existing curricula, student demographics, technological infrastructure, and the available resources of faculty and staff. Smaller class sizes allow for opportunities to facilitate interactions with students and to craft more individualized assignments that cater to students' majors, career aspirations, and social change motivations. While some students may require additional support to acclimate to the course material, they consistently exhibit a high level of enthusiasm for the field of data science and its applications.

ACKNOWLEDGEMENTS

This project is supported by the National Science Foundation award 1915714.

REFERENCES

[1] National Academies of Sciences Engineering, Medicine, et al., "Data Science for Undergraduates: Opportunities and Options", 2018. https://doi.org/10.17226/25104

[2] UC Berkeley, College of Computing, Data Science, and Society, Curriculum Overview, 2023. Retrieved from https://cdss.berkeley.edu/curriculum-overview.

[3] Ismail Bile Hassan, Thanaa Ghanem, David Jacobson, Simon Jin, Katherine Johnson, Dalia Sulieman, and Wei Wei, "Data Science Curriculum Design: A Case Study," in *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education* (Virtual Event, USA) (*SIGCSE '21*), Association for Computing Machinery, New York, NY, USA, pp. 529–534, 2021. Retrieved from https://doi.org/10.1145/3408877.3432443

[4] Richard D De Veaux, et al., "Curriculum guidelines for undergraduate programs in data science," in *Annu Rev Stat Appl* 4 pp.15–30, 2017.

[5] Yuri Demchenko, Adam Belloum,Wouter Los, Tomasz Wiktorski, Andrea Manieri, Holger Brocks, Jana Becker, Dominic Heutelbeck, Matthias Hemmje, and Steve Brewer, "EDISON Data Science Framework: A Foundation for Building Data Science Profession for Research and Industry," in *2016 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 620–626, 2016. https://doi.org/10.1109/CloudCom.2016.0107

[6] ACM Data Science Task Force, "Computing Competencies for Undergraduate Data Science Curricula," Association for Computing Machinery, New York, NY, USA, 2021.

[7] Aimee Schwab-McCoy, Catherine M. Baker, and Rebecca E. Gasper, "Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years," in *Journal of Statistics and Data Science Education* 29, sup1, S40–S50, 2021. https://doi.org/10.1080/10691898.2020.1851159 arXiv:https://doi.org/10.1080/10691898.2020.1851159

[8] *Foundations of Data Science*, UC Berkeley, 2015. Retrieved from http://data8.org/

[9] *Data Science Resources at Berkeley,* UC Berkeley*,* 2024*.* Retrieved from https://cdss.berkeley.edu/data-science-resources-berkeley

[10] *JupyterHub,* 2024*.* Retrieved from https://jupyter.org/hub

[11] *2i2c, Interactive Computing for Your Community*, 2022. Retrieved from https://2i2c.org/.

[12] *Colab*, 2024. Retrieved from https://colab.research.google.com/

[13] *Anaconda*, 2024. Retrieved from https://www.anaconda.com/

[14] *Otter-Grader,* 2024. Retrieved from https://github.com/ucbds-infra/otter-grader

[15] Susan Wang, Vandana Janeja, David Harding, Claudia von Vacano, and Daniel Lobo, "Adopting Data Science Curricula: A student Centric Evaluation," in *Proceedings of INTED2023 Conference,* pp. 8309-8315, 2023. https://doi.org/10.21125/inted.2023.2276