Accelerating the Advancement of Data Science Education

Eric Van Dusen[‡*], Anthony Suen[‡], Alan Liang[‡], Amal Bhatnagar[‡]

CONTENTS

Abstract—We outline a synthesis of strategies created in collaboration with 35+ colleges and universities on how to advance undergraduate data science education on a national scale. The four core pillars of this strategy include the integration of data science education across all domains, establishing adoptable and scalable cyberinfrastructure, applying data science to non-traditional domains, and incorporating ethical content into data science curricula. The paper analyzes UC Berkeley's method of accelerating the national advancement of data science education in undergraduate institutions and examines the recent innovations in autograders for assignments which helps scale such programs. The conversation of ethical practices with data science are key to mitigate social issues arising from computing, such as incorporating anti-bias algorithms. Following these steps will form the basis of a scalable data science education system that prepares undergraduate students with analytical skills for a data-centric world.

Index Terms—data science education, autograding, undergraduate institutions

Introduction

Data science is a burgeoning field that is quickly being adopted across all domains and sectors. Undergraduate data science education initiatives have been growing rapidly, but also largely in an uncoordinated manner. Programs are often developed and implemented in silos, leading to duplication of efforts and differences in pedagogical approaches and course quality. Furthermore, while a number of curriculum guidelines for degrees in data science have been proposed, opportunities for engaging in pedagogical exchanges and sharing resources remain rare. Without a common knowledge base of resources and platform for undergraduate support, many institutions have encountered pedagogical and infrastructural barriers in setting up Python centric data science curricula across campus.

Stakeholders around the country previously identified the necessity of gathering data science enthusiasts and discussing its implementation in institutions. The importance of defining data science curriculum guidelines has been the subject of numerous workshops and meetings such as the "Workshop on Theoretical Foundations of Data Science," "The Park City Math Institute 2016 Summer Undergraduate Faculty Program," and "Envisioning the Data Science Discipline: The Undergraduate Perspective." While these workshops produced comprehensive guidelines on the structure of the programs, the actual content and teaching modalities remain unclear.

On June 24-27, 2019, the Division of Data Sciences at the UC Berkeley hosted the 2nd National Workshop on Data Science Education which brought together nearly 70 faculty from a diverse range of higher-education institutions on at different stages of data science education. As a pioneer in undergraduate data science education, UC Berkeley shared its comprehensive set of open-license and open-source resources that range from teaching materials to cloud infrastructure at the workshop.

The workshop sought to build a national community of practice around undergraduate Data Science education by focusing around four primary areas:

1) Examining the *Foundations of Data Science (Data 8)* course - How does the content to pedagogical methods of Data 8 integrates various disciplines through the use of computational and inferential thinking.

2) Showcase the infrastructural platform that Berkeley has developed for its courses, and empower participants to use its many open-source components to overcome the *financial and technical barriers.*

3) Applying Modular Data Science education content into various disciplinary fields past the scope of that from a traditional computer science or statistics courses provide.

4) Implementing Ethical Content into *data science courses and examines how such integration could work.*

Establishing Foundational Course that Serves the Entire Campus

Data science can touch all different genres and disciplines of academia. The proliferation of affordable computational capacity, migration of publishing channels to the internet, advanced sensing technology, and other data collection methods has led to the

* *Corresponding author: ericvd@berkeley.edu*
‡ *University of California, Berkeley*

possibility of data science in almost every area of scientific endeavor. Applications of data science have created opportunities to teach students programming, statistical techniques, and other computational methodologies earlier in their academic careers to expand the academic possibilities within their chosen area of focus. Current examples of data science being integrated into other disciplines include randomized controlled trials in Development Economics published through open data repositories and the integration between law and data technologies.

An essential step toward creating a successful campus wide Data Science program is creating a campus wide introductory-level data science course that utilizes the Python programming language available to students of all academic disciplines. Using existing introductory computer science and statistics courses in place of a foundational data science course slows students' learning and limits the audience. Such a course also allows students to explore the realm of data science at an introductory level, so they can understand the basic concepts using custom made Data Science "Tables" library without getting lost in the more complex syntax of Pandas and transition into statistics and computer science, as many students do not have prior coding experience. In 2015, UC Berkeley launched Data 8 *Foundations of Data Science* for its undergraduate students. With the course centered upon students learning inferential thinking, computational thinking, and real-world relevance, students learn how to apply such statistical or computer programming techniques onto non-traditional fields through economic and geographical data and social issues. Using the Python datascience package, students work with real-world datasets to ask questions and find answers.

Through Data 8, core foundational facilitators now know to find connections to other departments and stakeholders. In fielding an entry-level data science course and fitting it into the curriculum, it's crucial to engage with faculty across a variety of disciplines in an inclusive, supportive way and a spirit of partnership. Berkeley has launched and taught many different connector courses that use Data 8 as a prerequisite. The Data Science students having classes from a non-traditional field allows students to find ways to apply their learning from Data 8 onto other domains.

Setting Campus Wide Educational Cyber-Infrastructure

Implementation of a data science course like Data 8 across the entire campus requires universities and institutions to develop capacity in on-demand cyber-infrastructure to support their educational goals. Local computation is not ideal, as it is harder to scale when the number of courses and students increases. For many institutions, the ability to set up the necessary support systems for JupyterHub or other infrastructure is beyond the expertise of a single course instructor, who already has to distribute their finite time in planning lesson outlines and curriculum. Institutional IT staff members would have to obtain additional training, which would vary across institutions to better fit the differing needs and implementations of the data science courses and can be too costly. For many small institutions and universities, this proves to be a major barrier in course delivery. The development of regional or national cloud-based computing solutions that can serve individual educational institutions is needed.

Universities must invest resources into developing data science educational infrastructure like JupyterHub, a platform not many universities have, that differs from research cyber-infrastructure. The two have different goals, resource needs, deployment timelines, cost and pricing of models, and broad access mandates. Data science educational infrastructure is deployed for relatively low resource use by a large number of relatively unsophisticated users. Making the data science infrastructure accessible requires establishing three components. At UC Berkeley, the core components include setting up a campus wide JupyterHub, integration with existing campus Learning Management Systems (LMS), e.g. Canvas (https://www.instructure.com/), and utilizing autograder technology.

Autograding technology is essential to the scalability of data science education and alleviates substantial work for large classes at UC Berkeley, such as *Data 8: Foundations of Data Science* and *Data 8X*, its massive open online course, or MOOC, version, which sees more than 1,500 students per semester and 75,000 students enrolled respectively. Currently, UC Berkeley uses various grading systems even within its own data science courses. *Data 8* utilizes ok.py, a Berkeley developed solution that has a plethora of features for large and diverse computer science and data science classes. However, this comes with a complexity cost for instructors who only need a subset of these features and sysadmins operating an okpy server installation [Suen18]. On the other hand, Data 100, the upper division core data science course, utilizes *nbgrader*, an open source grading solution built for Jupyter Notebooks. On Data 8X, the newly developed *gofer grader* is used to solely address the needs of a MOOC course and retains similar aspects from Data 8's grading system. The *gofer grader* is relatively new and has run into issues relatively frequently. Yet, it asynchronously supports hundreds of students' grading concurrently

To mitigate high individual institutional infrastructure startup costs, a national educational cyber-infrastructure strategy with industry and universities collaboration is required. Options include leveraging the existing four regional Big Data Innovation Hubs, which can provide access to cloud resources, partners and expertise or increase utilization of currently free industry platforms like Google Colab and Azure Notebooks. To maximize learning within any pilot program, local staff at a given institution would need to be trained and partake in the beta testing of such a system to document problems and best practices. Successful implementation of data science courses across certain locations might lead to partnerships across and within institutions, allowing for successful techniques to be communicated across all partners and similar curriculum modeling to exist for consistency.

All of this infrastructure is crucial for creating, deploying, and grading data science homework and lab assignments. Having this educational cyber-infrastructure is more efficient than local infrastructure, as instructors can teach students for many, the system holds all the necessary material, simplifies data management and analysis, and visualizes data for instructors. Before Berkeley launched its integrated system, the teaching faculty found it difficult to efficiently scale courses at the rate of their increasing interest. Berkeley's adoption of JupyterHub has allowed more than 1,600 students to enroll in Data 8 for its Spring 2019 iteration, a historic milestone that would not have been possible absent Berkeley's educational cyber-infrastructure.

Creating and Incorporating Modular Data Science Content

There are two main concerns when modularizing data science content: *Having just one introductory data science class is not enough to warrant an entire data science curricula, and creating a sustainable model that supports the data science curricula is challenging for newly adopting institutions.*

Implementing and integrating the new course to fit in the overall academic curriculum is critical for seamless student experience in data science. UC Berkeley's Division of Data Sciences has also supported the creation of data science content for inserting in other types of (usually non-data science) courses in self-contained "Modules" that can showcase aspects of data science to a different audience. Some examples of modules that students can take include Linguistics 110: *Introduction to Phonetics and Phonology,* Sociology 130 AC: *Neighborhood Mapping,* and Econ 101B: *Macroeconomics*. Developing and implementing such modules allow students to experience data-driven techniques and scientific computing through Python.

Because data science serves functions in a vast array of interdisciplinary fields of study, the ability to modify the introductory course and tailor it to fit in with the current institutional curriculum will go a long way in communicating the relevance of the field to students taking the course. This process will need time for planning and preparation before the actual steps for integration can start. In addition, faculty across different departments should collaborate to explore the possibility of connector courses or incorporation of data science in each others' subjects. Connector courses are supplemental courses which build on the introductory data science course by using similar statistical and computational techniques across different disciplines, such as business, biology, and geography. Berkeley has offered 27 different connector courses since their launch in 2015. To alleviate the burden of redistributing finances and to increase funding, faculty might have to reallocate their time to develop and adopt a new curriculum. To mitigate increasing startup costs, Berkeley has hired graduate students and even undergraduate students who previously excelled at that class to assist in teaching efforts. Incorporating on-campus talent, such as previous students, creates a robust data science culture on campus that is easy to spread among the student population.

To successfully adopt a data science modules curricula, we propose creating a platform to share teaching resources that is available to anyone in the community. Such a platform could be modeled on the popular Data8 public organization (https://github.com/data-8) and the site hosting Data Carpentry lessons (https://datacarpentry.org/lessons/). The principal functions of this platform are to share teaching resources such as use cases (datasets and accompanying analyses), open source textbooks or modules, and programs used to facilitate data science education. National Workshop on Data Science Education proves that the design of the courses and the planning of the material and activities is key. Berkeley's Data 8's success in reaching up to 1,500 students within its first few iterations attests to the importance of curriculum innovation and pedagogical methods. Having staff with technical skills to support the computer infrastructure and collaborative support with nearby/sister institutions who can share best practices and resources makes this model even more successful. Developing collaborative, modularized open-source teaching materials, such as the books used in Data 8 and Data 100, allows other institutions to more easily implement curricula for themselves.

Recently, Berkeley has been sharing such resources with institutions interested in adopting a data science curriculum. By sharing access to textbooks, lecture and lab materials, and similar resources, about 15 domestic and 10 international institutions have adopted Data 8 or a similar course or program. Most questions potential partnering institutions had regarded logistics, course topics, and infrastructure, which were resolved once given access to shared resources. Such partnering institutions range from community colleges to Ivy League universities indicating the widespread approval of Data 8's goals, implementation, and adaptability. Berkeley's cross-campus collaboration proves that transparency and communication is key to start and scale undergraduate data science programs across the world and increase Python literacy.

Integrating The Teaching of Ethics Into Data Science Courses

As data come to structure more and more aspects of our lives, the potential impact of data science on individuals and societies looms ever larger. For this reason, it is critical that data scientists understand the social worlds from which their data are drawn and in which their science intervenes. They must be trained to recognize the ethical implications of their work and act accordingly. The ethics of data science are social, individual, and contextual rather than linear. Ethical content can be incorporated into data science curricula both by integrating ethical topics into existing data science courses and by including ethically-focused courses to data science degree programs. The first approach may be better suited to the ethical questions that individual data scientists encounter in their daily work, while the second may be better suited to the broader issues raised by the growing role of data and algorithms in society as a whole. For example, ethical questions arise at every step of the data science life cycle. Where data science courses teach professional competencies of statistics, computer science, and various content areas, they can also introduce students to the ethical standards of research and practice in those domains [NASEMS18]. Some data science textbooks already address such issues as misleading data visualizations, p-hacking, web scraping, and data privacy [Baumer17].

A recent trend in incorporating such ethical practices includes incorporating anti-bias algorithms in the workplace. Starting from the beginning of their undergraduate education, UC Berkeley students can take *History 184D: Introduction to Science, Technology, and Society: Human Contexts and Ethics of Data,* which covers the implications of computing, such as algorithmic bias. Additionally, students can take *Computer Science 294: Fairness in Machine Learning,* which spends a semester in resisting racial, political, and physical discrimination. Faculty have also come together to create the Algorithmic Fairness and Opacity Working Group at Berkeley's School of Information that brainstorms methods to improve algorithms' fairness, interpretability, and accountability. Implementing such courses and interdisciplinary groups is key to start the conversation within academic institutions, so students can mitigate such algorithmic bias when they work in industry or academia post-graduation.

Databases and algorithms are socio-technical objects; they emerge and evolve in tandem with the societies in which they operate [Latour90]. Understanding data science in this way and recognizing its social implications requires a different kind of critical thinking that is taught in data science courses. Issues such as computational agency [Tufekci15], the politics of data classification and statistical inference [Bowker08], [Desrosieres11], and the perpetuation of social injustice through algorithmic decision making [Eubanks19], [Noble 18]_, [ONeil18] are well known to scholars in the interdisciplinary field of science and technology studies (STS), who should be invited to participate in the development of data science curricula. STS or other courses in the social sciences and humanities dealing specifically with topics related to data science may be included in data science programs.

Including training in ethical considerations at all levels of society and all steps of the data science workflow in undergraduate data science curricula could play an important role in stimulating change in industry as our students enter the workforce, perhaps encouraging companies to add ethical standards to their mission statements or to hire chief ethics officers to oversee not only day-to-day operations but also the larger social consequences of their work.

Summary & Vision

We envision a world where all students can learn ethical data-driven techniques regardless of their domain and can manipulate data to find better solutions to problems. To do that requires a four part strategy involving creating a campus wide foundational data science course, the modularization of data science course content to integrate it with courses in existing domains, the scalable cloud infrastructure power it all, and the human context and ethics content to reign in misuse of data & artificial intelligence. Integrating Python across different fields exposes students to learning programming in areas they would not have previously expected. These strategies will accelerate the creation of a space for Data Science to exist as a cross-campus endeavor and engage faculty and students in different departments

References

[Baumer17] Baumer, B. S., Kaplan, D. T., & Horton, N. J. (2017). Modern Data Science with R. Retrieved from http://mdsr-book.github.io/

[Bowker08] Bowker, G. C., & Star, S. L. (2008). Sorting things out: Classification and its consequences. Cambridge, MA: MIT Press.

[Desrosieres11] Desrosieres, A. (2011). The politics of large numbers: A history of statistical reasoning. Cambridge, MA: Harvard University Press.

[Eubanks19] Eubanks, V. (2019). AUTOMATING INEQUALITY: How high-tech tools profile, police, and punish the poor. PICADOR.

[Latour90] Latour, B. (1990). Technology is society made durable. The Sociological Review, 38(1), supplement, 103-131. Retrieved from http://www.bruno-latour.fr/sites/default/files/46-TECHNOLOGY-DURABLE-GBpdf.pdf

[NASEMS18] National Academies of Sciences, Engineering, and Medicinemies of Sciences. (2018, May 02). Data Science for Undergraduates: Opportunities and Options. Retrieved from https://doi.org/10.17226/25104

[Noble18] Noble, S. U. (2018). Algorithms of oppression how search engines reinforce racism. New York: New York University Press.

[ONeil18] ONeil, C. (2018). Weapons of math destruction: How big data increases inequality and threatens democracy. London: Penguin Books.

[Suen18] Suen, A., Norén, L., Liang, A., & Tu, A. (2018). Equity, Scalability, and Sustainability of Data Science Infrastructure. Proceedings of the 17th Python in Science Conference. doi:10.25080/majora-4af1f417-002

[Tufekci15] Tufekci, Z. (2015). Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. Colorado Technology Law Journal, 203-218. Retrieved from https://ctlj.colorado.edu/wp-content/uploads/2015/08/Tufekci-final.pdf.