National Advancement of Data Science Education

Anthony Suen[‡*], Alan Liang[‡], Amal Bhatnagar[‡]

Data science is a burgeoning field that is quickly being adopted across all domains and sectors. Data Science education initiatives have been growing rapidly, but also largely simultaneous and in an uncoordinated manner. Programs are often developed and implemented in silos, often leading to duplications 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 difficulties in setting up data science curricula, specifically pedagogical and infrastructural barriers. Additionally, data science education in many universities has been focused on graduate programs, with very few undergraduate ones in place.

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 workshop 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 July 16th-July 19th 2018, the Division of Data Sciences at the UC Berkeley hosted a Workshop on Undergraduate Data Science Pedagogy and Practice. As a pioneer in undergraduate data science education, UC Berkeley has developed a comprehensive set of open-license and open-source resources that range from teaching materials and curriculum to its system infrastructure and share them at the workshop. With support from the NSF, the Berkeley organizers put out a call for interested professors and instructors across the country 41 instructors from a variety of higher-education institutions that hugely vary in terms of size, location, and experience with data science education.

The workshop sought to 1) Examine the *Foundations of Data Science (Data 8)* course, and discuss the course content and pedagogical methods employed behind this innovative and accessible curriculum, 2) Build a national community of practice around undergraduate Data Science education, 3) Showcase the infrastructural platform that Berkeley has developed for its courses that involve data science, and empower participants to use its many open-source components and 4) Engage workshop participants to identify challenges and needs for moving Data Science education forward at their institutions participants collaborated to develop a common vision for expanding and moving forward the cause of Data Science Pedagogy nationally. These four items were broken down core topics including *1)* **Integrating and Serving across Disciplines using Computation** *can be used and applied in various disciplinary fields past the scope of that from a traditional computer science or statistics faculty 2)* **Educational Cyber-Infrastructure** *looks into the technical resources and human support needed for running data science courses, the financial and technical barriers many institutions face from implementing such an infrastructure, and the possible solution of generating a regional or nationwide educational cyber-infrastructure for sustainable and scalable benefits. 3)* **Integrating Modular DS content into teaching including Ethical Content** *properly integrate discussions and critical thinking about ethics into data science courses, and how such an integration could result in new courses that will better complement the overall scope of the educational field.*

1) Integrating and Serving across Disciplines using Computation

Create a space for Data Science to exist as a cross-campus endeavor and engage faculty in different departments.

Within higher education, 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 possibility of data science in almost every area of scientific endeavor. 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.

The first step toward creating a successful Data Science program within an institution comes from the successful implementation of an introductory-level data science course, as utilization existing introductory computer science and statistics courses in place of a foundational data science course slows students learning and limits the audience. Additionally, having a course that focuses on relevant data science techniques allows students to focus on industry-level and research-level concepts that would be useful to them in their professional lives. Such a course also allows students to explore the realm of data science at an introductory level, so they can understand the basic concepts without getting lost in the more complex ones. This introductory data science course

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

specifically helps students start to ask questions and think like a data scientist. By having a campus-wide introductory data science course available to students of all academic disciplines, students are exposed to a growing field relevant to their disciplines and majors.

The university community can benefit by data science bridging multiple disciplines and introducing these programming techniques and other quantitative tools from statistics and mathematics within the context of rich applications drawn from a variety of disciplines. 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. Exposure to such subjects will allow students to become clearer thinkers and problem solvers in their home disciplines through their knowledge and training in computational work with real data. Data science is not just a service course, as the emerging challenges of data usage in applied disciplines can set the agenda for data science within computer science and statistics. In undergraduate education, a broad student population learning data science can benefit students operating outside of the specific data science domain, while the data science field can attract bright minds with interesting questions.

As for its implementation, Computer Science and Statistics fields may be a traditional base but the core foundational facilitators must 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. A full implementation will involve enriching the teaching of the entry-level course, injecting computational content into existing courses, and helping to reframe methods courses within the disciplines so they can build off of integrative interdisciplinary general training in data science.

The Data Science students having classes from a field of application will create a space to apply methods learned in their core classes, and opportunities to learn theoretical methods that Data Science may be lacking. Having other departments engage with a Data Science major or minor will give those departments a way to grow their offerings and create possibilities for curricular innovation. The Berkeley model provided a number of illustrations of cross-campus collaboration including introductory Data Science course is accompanied by a series of smaller, applied "Connector" courses which give students a flavor of how data science may be carried out within a given domain. The Data Science student teams have also supported the creation of data science content for inserting in other type of (usually non data science) courses in self-contained "Modules" that can illustration aspects of data science to a different audience. The Berkeley Data Science program has also undertaken a series of summer "bootcamp" style workshops open to all faculty to encourage faculty to engage and innovate their curriculum. Finally, at a research university such as Berkeley, there has been considerable success due to graduate students or postdocs, who may be leading the way for faculty, in adapting new data science methods to different disciplines.

2) Common Educational Cyber-Infrastructure

Implementation of a data science course in a scalable way 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 manage when the number of students increases. Additionally, as the number of courses that require such infrastructure increases, local computation would become too time consuming. For many small institutions and universities, this proves a difficult task that can be a barrier to innovation in curriculum and course delivery. As a result, development of regional or national cloud based computing solutions that can serve array individual educational institutions is needed.

Fund and pilot a regional or national data science hub for education that will expand access and encourage innovation in data science education.

Educational cyber-infrastructure is different than research cyber-infrastructure due to differences in its goals, resource needs, deployment timelines, cost and pricing of models, and broad access mandate. Educational infrastructure is deployed for a relatively low resource use by a large number of relatively unsophisticated users. Making the infrastructure accessible means making it easy to use both by instructors and students, and potentially integrating it into existing campus Learning Management Systems (LMS), eg Canvas. For institutions teaching data science courses, the infrastructure is crucial for creating and deploying data science homework and lab assignments. Having this educational cyber-infrastructure is more efficient than local infrastructure, as instructors can teach students from all around the world and the system holds all the necessary material. It also makes teaching data management and analysis and allowing the ability to have instructors illustrate the visualization of data easier.

However, the adoption costs of cyber-infrastructure is high and problematic, especially for smaller institutions. While the component cost for hardware and software are going down through virtualization (cloud), human talent is hard to acquire. For many institutions, the ability to setup the necessary support systems for JupyterHub or other infrastructure is beyond the expertise of a single course instructor. Even qualified instructors may not have the capacity to take on such a task, as their time is required for equally important tasks of planning lesson outlines and curriculum. Institutional IT staff members would also be required to go through additional training if they were assigned the task, and the trainings required would vary across institutions to better fit the differing needs and implementations of the data science courses. Thus overall startup costs are expensive, and the long term sustainability for maintaining a educational cyber-infrastructure would come with too many question marks for many institutions faculty to make implementation a priority.

Autograding 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 see 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 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. 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.

Creating a national educational cyber-infrastructure allowing participation from all institutions and universities can solve the

problem of high individual institution startup costs in infrastructure. We believe that the best way to accomplish this is to would be to work with the existing regional Big Data Hubs, which may have access to cloud resources, and host partners and expertise. To maximize learning within the pilot, 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.

The successful formation of a national educational cyberinfrastructure will allow for data science courses to be supported at institutions and universities under a cost efficient structure. Stakeholders would no longer need to implement their own system, but instead could go through training and onboarding for a national system that will be easy to use and consistent between institutions. If such a process could be undertaken, the ability to host data science courses for undergraduate university-level student will be readily accessible to schools.

3) Guidelines for Creating and Incorporating Modular DS 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.*

Implementation and integrating the new course to fit in the overall academic curriculum is key for a seamless student experience. 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 institution curriculum will go a long way in communicating the relevance of the study to students taking the course. This process will need time for planning and preparation before the actual steps for integration can start. In addition, a useful step in this process would be to form arrangements with faculty from different departments to see if there exists a possibility of connector courses or incorporation of data science into other subjects. Connector courses are supplemental courses which build on the introductory data science course by using similar statistical and computational techniques, but in different disciplines, such as business, economics, and geography. Finally, there are many places where the class will be fitting in as a prerequisite, or satisfying a requirement, for different campus departments, and these will have to be finessed with each department. It may be necessary to navigate between faculty offering related courses, using other programming languages, and departments which operate in areas similar to that of the subject. In order to alleviate the burden of redistributing finances and increase funding, faculty might have to reallocate their time to develop and adopt new curriculum. Hires for these positions could come from graduate students, institution volunteers, and even renowned academics.

In order to successfully adopt a data science 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 will be to share teaching resources such as use cases (dataset and accompanying analyses), open source textbooks or modules, as well as programs used to facilitate data science education. The platform will be inclusive, with contribution and usage open to anyone in the community. There will be a dissemination of use cases, including exercises, activities, and examples sorted by topic/domain that simplify inclusion of relevant and useful examples in new or existing courses. This repository would include canonical examples, such as the Iris and Mauna Loa CO2 data sets commonly used to illustrate classification and time series analysis, and other examples from local industry or research projects. The design of the courses and the planning of the material and activities is key, as highlighted by UC Berkeley's Data Science Pedagogy and Practice. 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 support by collaboration 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. Modularizing textbooks into a catalog of chapters can be independently maintained to satisfy different pedagogical scenarios or requirements.

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 (National Academies of Sciences, Engineering, and Medicine 2018). Some data science textbooks already address such issues as misleading data visualizations, p-hacking, web scraping, and data privacy (Baumer, Kaplan, and Horton 2017).

Databases and algorithms are socio technical objects; they emerge and evolve in tandem with the societies in which they operate (Latour 1990). Understanding data science in this way and recognizing its social implications requires a different kind of critical thinking than is taught in data science courses. Issues such as computational agency (Tufekci 2015), the politics of data classification and statistical inference (Bowker and Star 2000; Desrosières 1998), and the perpetuation of social injustice through algorithmic decision making (Eubanks 2018; Noble 2018; O'Neill 2016) 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

In summary the conference participants set up a course of work to develop data science education and a pathway forwards. The specific proposals are:

1) *Create a space for Data Science to exist as a cross-campus endeavor and engage faculty in different departments.*
2) *Fund and pilot a regional or national data science hub for education that will expand access and encourage innovation in data science education.*
3) *Centralized platform of resources for enhancing collaborating around teaching data science*

Our three-pronged strategy involving creating a foundational course, necessary and scalable infrastructure, and modularized content with feasible replicability pivots institutions to establish sustainable data science curricula. Having an open-source platform would democratize access to resources for creating such data science curricula and course content.

We envision a world where students serve as clear thinkers who learn ethical data-driven techniques regardless of their domain of expertise and can manipulate data to find better solutions to problems. Institutions would integrate data science techniques on campus and collaborate with other facilities across the country on a centralized platform with resources. They would adopt these resources and personalize them on their own curriculums to help their students. A national data science hub for education would bring together these institutions and innovate the data science education. Universities would encourage students to use such data-driven methodologies not just in an institutional setting but also in their professional careers afterwards. We believe our methodology will guide our commitment to work together, structure our cross-campus collaboration, and target grant writing to support these initiatives.

Works Cited

Baumer, Benjamin S., Daniel T. Kaplan, and Nicholas J. Horton. 2017. *Modern Data Science with R*. Chapman & Hall. *http://mdsr-book.github.io*

Bowker, Geoffrey C. and Susan Leigh Star. 2000. *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.

Desrosières, Alain. 1998. *The Politics of Large Numbers: A History of Statistical Reasoning*. Cambridge: Harvard University Press.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin's.

Hacking, Ian. 1996. Normal People. Pp. 59-71 in David Olson and Nancy Torrance, eds., *Modes of Thought: Explorations in Culture and Cognitions*. Cambridge: Cambridge University Press.

Hicks, Marie. 2017. *Programmed Inequality: How Britain Discarded Women Technologists and Lost its Edge in Computing*. Cambridge: MIT Press.

Latour, Bruno. 1990. Technology is society made durable. *The Sociological Review* 38(1, supplement): 103-131.

Light, Jennifer S. 1999. When computers were women. *Technology and Culture* 40(3): 455-483. *https://www.jstor.org/stable/25147356*

MacKenzie, Donald A. 1981. *Statistics in Britain: 1865-1930; The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press.

National Academies of Sciences, Engineering, and Medicine. 2018. *Data Science for Undergraduates: Opportunities and Options*. Washington, DC: The National Academies Press. *https://doi.org/10.17226/25104*

Noble, Safiya Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

O'Neill, Cathy. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.

Tufekci, Zeynep. 2015. Algorithmic harms beyond Facebook and Google: Emergent challenges of computational agency. *Colorado Technology Law Journal* 13(2): 203-218.