A Perspective on Designing Undergraduate Education in the New Discipline of Data Science

Michael J. Franklin^{1,2} Dan L. Nicolae^{3,4}

¹Department of Computer Science, University of Chicago, Chicago, Illinois, United States of America,

²Department of Electrical Engineering and Computer Science, Berkeley Institute for Data Science, University of California Berkeley, Berkeley, California, United States of America,

³Department of Statistics, University of Chicago, Chicago, Illinois, United States of America,

⁴Data Science Institute, University of Chicago, Chicago, Illinois, United States of America

Published on: Jun 07, 2021

DOI: <https://doi.org/10.1162/99608f92.8d7bff79>

License: [Creative Commons Attribution 4.0 International License \(CC-BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

Adhikari et al. (“[Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley](#),” this issue) provide an insightful and stimulating perspective on UC Berkeley’s pioneering approach to data science education and data science more broadly. The development of a very successful academic program is described jointly with a perspective on data science history, the principles that went into the design of learning goals, the concepts underlying the computational and inferential thinking at the core of the curriculum, and techniques for delivering data science education at scale. At our university, as at many institutions creating data science efforts, cross-campus discussions have explored the myriad options for designing a new data science curriculum. These discussions have centered on questions such as: What are the foundational topics that should be covered in an undergraduate program? What is the appropriate mix of theory and applications? How should a Data Science undergraduate program differ from existing statistics and computer science programs?

Our comments, emphasizing the evolution of the field, the foundations of the discipline, and the unique opportunities and challenges of data science education, outline some of the key places where our emerging vision is leading us to different choices than those of Adhikari et al. Our intent is to contribute to the ongoing discussion on this important topic and our comments should be taken in the context of our deep appreciation for the important and innovative contributions that the team at Berkeley has made and continues to make to data science pedagogy.

What Are the Foundations of Data Science?

Data science curriculum design has been a widely discussed topic, not only in the United States but also around the world. Proposed curricula often reflect the unit where data science is centered on campus—Computer Science, Statistics, other departments (business analytics, biomedical informatics etc.), or university-wide—and range from simply repurposing existing courses in Statistics, Computer Science, and machine learning to efforts (such as that described by Adhikari et al.) that encompass a substantial number of new courses based on a well-articulated philosophy on data science education.

The definition of data science has been debated extensively, including in the *Harvard Data Science Review*, with Irizarry (2020) advocating that it is an umbrella term describing complementary skills. Our campus has chosen to approach data science as a new emerging discipline that, while building on reasoning and methods from statistics and computer science (i.e., inferential and computational thinking), aims to solve its own set of foundational problems, many of which have yet to be enumerated and clearly defined. This recognition of data science as a new and evolving field has guided us in the planning of our educational programs. We highlight the main axes along which we see data science education differing from programs in highly-related areas such as computer science, statistics and applied mathematics.

The Human Context of Data-Driven Decision Making

Increasing evidence of the potential negative effects of data-driven systems on social equality, political discourse, and other major societal concerns have made it clear that in addition to computation and inference, the human context of data-driven decision making must be a key component of data science. Regardless of whether such decisions are made by humans or by algorithms, it is humans and society that will ultimately have to deal with their consequences. This recognition leads to two fundamental requirements for data science education: **data ethics** and **communication**.

The need for integrating ethics into STEM education more generally has received increasing attention in recent years. This need is particularly acute for data science, where the potential for harm due to incorrect, biased, or oblivious analyses and decisions has become apparent as data-driven and algorithmic systems influence and increasingly determine outcomes such as who gets what medical treatment, who can buy a house, whether or not someone is released on bail, and many other life-changing decisions. Furthermore, the magnitude of the impact of data-driven algorithms underlying much of online life (which is increasingly much of life) is just beginning to be fully appreciated; for example, algorithmic news bubbles and echo chambers combined with disinformation are having an increasingly visible impact on political discourse.

While there is widespread agreement that ethics must play a central role in data science education, there is less consensus on how such considerations should be delivered in the curriculum. Ethics classes are a component of many data science programs. Adhikari et al. describe one such course at Berkeley, which is typically taken by students in their junior or senior years. Our Computer Science colleagues at University of Chicago, Blase Ur and Raul Castro Fernandez, have created a technically-oriented course called “Ethics, Fairness, Responsibility, and Privacy in Data Science,” also aimed at juniors and seniors. The course examines the stages of the data science lifecycle (see Section 1.2) through various lenses (statistical, legal, philosophical) but focuses on grounding theory with a deep understanding of the underlying technology. For example, in one project students implement and experiment with a form of differential privacy on a real data set. A goal for the course is stated as: “Through both programming assignments and discussions, students who complete the course will learn how to design systems that are inclusive and respectful of all data subjects” (Ur & Castro Fernandez, 2020).

Those are two examples of how the recognition of the centrality of the human context to data science is generating significant innovation in the creation of courses to address this need. However, there is also recognition that having only specific “ethics” courses that are distinct from the other courses in the curriculum runs the risk of treating human context as an add-on rather than a core component of data science. Thus, ways of suffusing human and societal concerns throughout the curriculum must also be investigated. An innovative approach in this regard is [Harvard’s Embedded Ethics program for Computer Science](#), where philosophers are “embedded” into CS courses to teach modules that explore ethical issues that arise in those courses.

Educating students to understand how to communicate with data is a second important requirement that human context imposes on data science curricula. If the goal of data science is to inform and influence decisions, then a deep understanding of how to present and consume data-driven arguments is necessary. Such skills are sometimes referred to as **data literacy**.

Data visualization and communication skills are two core topics. More than just producing effective graphical representations, it is also important that students learn how to explain the reasoning and interpretation that underlie results, and to understand and be able to communicate the robustness, assumptions, and confidence in the results they produce. An appreciation for the workings of human decision-making processes is also important.

Furthermore, exposure to pitfalls in analysis and learning how to recognize incorrect or specious arguments that are purportedly based on data are other aspects of data literacy that are essential for a complete data science education. Similar to the issues around data ethics and fairness, these topics should be embedded in all classes, but dedicated courses are also required so that their principles and foundations can be rigorously studied.

The Data Science Lifecycle—A Unifying Theme

A distinguishing trait of data science as an emerging discipline, is its appreciation of data, models and analyses as first-class objects that are developed through a series (often iterated) of steps and that persist and evolve over time. Thus, a concept that is central to data science is the unifying framework of the data science lifecycle. As described by Berman et al. (2018, p. 68):

Data never exists in a vacuum. Like a biological organism, data has a life cycle, from birth through an active life to "immortality" or some form of expiration. Also like a living and intelligent organism, it survives in an environment that provides physical support, social context, and existential meaning. The data life cycle is critical to understanding the opportunities and challenges of making the most of digital data.

That is, data science as a discipline involves more than just combining the traditions of computational and inferential thinking. Rather, it is an overarching philosophy and methodology that takes into consideration the entire process of extracting value from data across many facets and over many time scales. Training data scientists involves exposing them to the myriad concepts of working with data. The lifecycle starts with question formulation, experimental design, data discovery, and/or data collection. It progresses through a “data wrangling” process of organizing, cleaning, and transformation at which point the data can be analyzed. Analysis, which is where many of the traditional computational and inferential advances come into play, can then be performed through model building, statistical analysis, machine learning, and so on. Then the results of the analysis must be communicated and explained to aid decision making. Finally, data sets, models, results, and other inputs and products of the process must be curated so that, if appropriate, they can be correctly and

reliably reused and built upon to extract further value and insights in the future. Layered across all of these steps are concerns regarding the human, regulatory, and domain-specific aspects of the data science endeavor.

Of course, the lifecycle is not simply a sequence of discrete, independent steps. Rather, the stages are interrelated, interdependent and often overlapped in time. For example, a recent study by Sambasivan et al. (2021) examines the compounding effects of decisions in several of the lifecycle steps in the context of machine learning pipelines. They call such dependencies “Data Cascades” and they identify, through interviews of practitioners, the complex ways in which various upstream data manipulation steps impact the quality of real-world analysis and prediction. Understanding and appreciation of these issues should not be relegated to a single course, but rather must be a central organizing theme of any data science educational program.

Data Science Research Agenda

A third aspect of data science that we feel has been under-appreciated by many both inside and outside of the field is the need and opportunity to develop a robust, distinct, and intellectually rich research agenda. While the focus of this discussion is on education, we believe strongly that as in all academic disciplines, the interplay between research and pedagogy is fundamental. This interdependency, of course, is the thesis on which the modern research university is based, and we see no reason why data science should be treated differently.

Since data science draws from computer science, statistics and other disciplines, there are many points of interaction and overlap among the research topics already being pursued in those disciplines and a data science research agenda (see Wing, 2020, for example). However, we believe that the opportunity for data science research is to identify those questions that currently are not receiving adequate attention because they do not squarely fit within the confines of an existing domain or that are not even being asked because they fall between the cracks at the boundaries of the existing research communities. Thus, we see research opportunities at the intersections and interstices of existing fields. Questions such as the new rules of the data economy, developing a formal foundation for data integration, how to truly integrate the computational and inferential perspectives of data, and research into the properties of the data life cycle (a la Sambasivan et al. 2021) are just a few examples of the rich and varied landscape for data science research. As we continue to develop data science educational programs, we must also push forward on defining the contours of a robust research agenda and community for the field. To not do so, will hinder the standing of data science in the academy and slow innovation in the field, thereby shortchanging the students whom we aim to educate.

The Opportunities and Challenges in Data Science Education

Having discussed the main new foundational topics that must be addressed in a data science curriculum, we now take a step back and consider some of the salient attributes of data science that require new thinking and new solutions for undergraduate education.

The Interdisciplinary Nature of Data Science

Data science is an inherently interdisciplinary field. As noted by Adhikari et al., statistics has long been interdisciplinary in the sense that practitioners “learn to embed themselves in teams along with domain experts.” In recent years, computer science has increasingly expanded its purview, encompassing topics that span traditional disciplinary boundaries from physical and biological sciences to social sciences, arts, and humanities. Data science, which is emerging from these two fields, is particularly well-positioned to serve as an intellectual hub across domains. Furthermore, the field is emerging at a time when many universities are rethinking the organizational barriers that exist due to departmental and other traditional structures, making this an ideal time for a new model of interdisciplinary research. Many see the opportunity for data science to be a catalyst for new thinking in interdisciplinary approaches.

The article describes a novel approach used at Berkeley, namely, the development of “connector courses” that provide bridges from the material covered in the introductory data science course to a wide variety of other fields. The connector courses range in scale, some are full-blown courses, while others have fewer hours than a full course. It is impressive that this approach was implementable in the context of existing majors and courses at a large university, but there are other possible approaches as well. Some universities have taken the path of creating hybrid majors. This approach is called “Data Science + X” (or DS+X) where “X” is typically a more traditional existing major. A third alternative, which is the one we have been following at University of Chicago, is to design a data science major in a way that makes it more amenable to being combined with existing majors for students who wish to double major. This solution requires thoughtful considerations on size (number of courses) and flexibility (allowing diverse paths for completion of the degree), while preserving the essential mathematical, statistical, computational, scientific, and human context foundations in the curriculum.

While the above alternatives will work better or worse at different types and sizes of institutions, the choice begs a larger question of how a comprehensive data science program should engage with the greater university. A program that exists only at the undergraduate and professional Masters levels will have difficulty flourishing at a research institution, making it difficult to develop the “core ideas” of the field. Thus, in coordination with pedagogical development at these levels, a robust research program at the Ph.D. and research Masters levels will provide the intellectual framework in which a discipline of data science will be developed. As with the undergraduate and professional education programs, this research program will necessarily be interdisciplinary, raising additional challenges in terms of faculty recruitment and departmental organization.

The Breadth of Data Science

We cannot argue for a more comprehensive curriculum that includes other core ideas of data science without acknowledging the following tension in data science education (similarly to what has been observed in other interdisciplinary domains): should data science majors be required to know as much statistics and probability as statistics majors and computing as computer science majors while absorbing the other foundational issues in

data science and engaging in real-world applications? It is obvious (for many) that it is impossible to pack three disciplines into one major and that a data science major is different from a Computer Science-Statistics double major. The history of statistics and computer science education is informative for moving out of this conundrum; during their first years, curricula were largely based on mathematics and applied mathematics (with some engineering for computer science). The disciplines have evolved and created distinct identities reflected in their educational programs. We anticipate a similar (but accelerated) evolution for data science. Adhikari et al. start their article by arguing that a curriculum should be developed based on “the grand conceptual achievements of a field” while “conveying the core ideas”; we agree and we believe that we will achieve a commonly accepted notion of the data science curriculum after we clearly define the core ideas and foundations of data science.

A possible solution is to design flexible educational programs that have a multitude of tracks allowing students with diverse interests and backgrounds to pursue a degree in Data Science. A theory track would allow students to study advanced topics in mathematics of data science and machine learning, while a computational track would provide an in-depth understanding of data structures, networks and distributed systems, and algorithms. Domain-specific tracks would allow comprehensive explorations of how data science functions in scientific applications in physical and biological science, or how it is engaged in solving important societal issues in humanities and social sciences. The balance of mathematics, computation, and applications can be tailored to the various tracks in a way that reflects the student population and allows the recruitment and retention of students. From a practical perspective, early and continued mentoring of students and plans allowing the transition between tracks are essential.

Experiential Learning

As many argued before (e.g., Blei & Smyth, 2017), data science aims to solve real-world problems and address domain-specific questions, thus making experiential learning a pillar of any comprehensive educational program. Whether the students are engaged in capstone projects at the end of their education, or practicum courses are at the center of their programs (Kolaczyk et al., 2021), the opportunity to design meaningful experiences should be a dominant discussion point in any conversation of data science education.

The development of an experiential learning program needs careful planning and could be very different at small universities versus large institutions. Ideally, students work in teams, on impactful projects, and under the mentorship of experienced data scientists (faculty, research staff, post-doctoral fellows). Implementation of a successful practicum requires “mutually beneficial, long-term partnerships with entities that have ‘real’ data science projects” (Uminsky, 2021), advising resources, and the development of models that allow adoption of this in a variety of institutions, ranging from community colleges to private universities.

Another important set of topics is centered on working with real data sets; data that are messy, need curation and annotation, or are too large for a single computer are common in applications—motivating the importance

of a dedicated course on data engineering. More generally, a complete data science project requires at the minimum expertise in data collection and study design, data provenance and ownership, data cleaning and annotation, analysis and inference, dissemination, and curation.

Closing Thoughts

Data science will be at the core of solving many of the great challenges of the next century, ranging from climate change to inequality and disinformation, and fulfilling this role will require educating the next generation of scientists, entrepreneurs, policymakers, teachers, and so on. While the focus of the discussion here is on undergraduate education, we should continue to have critical debates on doctoral education (where the interplay between research and education is essential and the balance between theory and practice can be different than in undergraduate programs), the importance of data literacy in high school (and how the mathematics curriculum can be enhanced by adding data science), and the need for postgraduate education (for such a rapidly evolving field). We also need to ceaselessly advocate for a sustained and substantial investment in data science education.

Disclosure Statement

Michael J. Franklin and Dan L. Nicolae have no financial or non-financial disclosures to share for this article.

References

- Adhikari, A., DeNero, J., & Jordan, M. I. (2021). Interleaving computational and inferential thinking in an undergraduate data science curriculum. *Harvard Data Science Review*, 3(2). <https://doi.org/10.1162/99608f92.cb0fa8d2>
- Berman, F., Rutenbar, R. A., Hailpern, B., Christensen, H., Davidson, S., Estrin, D., Franklin, M. J., Martonosi, M., Raghavan, P., Stodden, V., & Szalay, S. A. (2018). Realizing the potential of data science. *Communications of the ACM*, 61(4), 67–72. <https://doi.org/10.1145/3188721>
- Blei, D. M., & Smyth, P. (2017). Science and data science. *PNAS*, 114(33), 8689–8692. <https://doi.org/10.1073/pnas.1702076114>
- Irizarry, R. A. (2020). The role of academia in data science education. *Harvard Data Science Review*, 2(1). <https://doi.org/10.1162/99608f92.dd363929>
- Kolaczyk, E. D., Wright, H., & Yajima, M. (2021). Statistics practicum: Placing “practice” at the center of data science education. *Harvard Data Science Review*, 3(1). <https://doi.org/10.1162/99608f92.2d65fc70>

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Article 39). <https://doi.org/10.1145/3411764.3445518>

Uminsky, D. (2021). From outlier to normalcy: Community partnering as a pathway toward modularity, sustainability and diversity when centering ‘practice’ in data science education. *Harvard Data Science Review*, 3(1). <https://doi.org/10.1162/99608f92.80579436>

Ur, B., & Castro Fernandez, R. (2020). CMSC 25900 Ethics, fairness, responsibility, and privacy in data science. <https://www.classes.cs.uchicago.edu/archive/2020/spring/25900-1/index.html>

Wing, J. (2020) Ten research challenge areas in data science. *Harvard Data Science Review*, 2(3). <https://doi.org/10.1162/99608f92.c6577b1f>

©2021 Michael J. Franklin and Dan L. Nicolae. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](https://creativecommons.org/licenses/by/4.0/), except where otherwise indicated with respect to particular material included in the article.