Building on the Shoulders of Bears: Next Steps in Data Science Education

Dustin Tingley

Deputy Vice Provost for Advances in Learning

Professor of Government

Adhikari, DeNero, and Jordan of UC-Berkeley (Bears) provide a thrilling tour of one of the most ambitious curricular experiments in modern history. They provide not just a conceptual tour de force taken from a myriad number of academic fields, but also an articulation of why they designed the series of courses, along with their connector courses, in the way that they did. The efforts described in their paper are not to be underestimated. Shaping an interdisciplinary program from the ground up at any large academic institution is extremely challenging. And omitted from the article is a characterization of the many challenges they were able to overcome in doing so. While they document the impact that they have had at Berkeley, and through the delivery of data8 via the edX platform, omitted from the article is the profound impact they have had on other institutions and people. This has come in the form of other institutions directly drawing on their curriculum and design. But it has also come in the form of catalyzing numerous conversations at other institutions that do not directly draw on their work but are certainly inspired by them.

In light of these remarkable accomplishments, I would like to briefly offer some thoughts that are more in the realm of "what is next". My thoughts here are heavily influenced by both an administrative role to help advance teaching and learning, as well as a practicing "data scientist" operating within the social sciences.

First, there are numerous opportunities to reach learners that extend beyond those who are enrolled in BA, MA and PhD programs. As documented in this journal (Chen 2020), there are substantial opportunities for younger students to begin exploring data science. Many years ago I taught at the high school level and had the opportunity to teach a hybrid course on statistics and game theory. Needless to say, the statistics components were dry and disconnected rehashing of basic statistical concepts. In contrast, the more modern data science turn can expose young learners to concepts around algorithms that are now part of their everyday life. And we can expose young learners to the fact that not all "data" sits nicely in clean spreadsheets. Music is data, text is data, etc.. Exposing students to these ideas does not require detailed math or even programming. Of course, those components can be built into such curricula but they need not be barriers. On the flip side are opportunities for reaching

individuals who will never take a technical data science course but instead want to be literate in data science thinking. These individuals might even work with or manage data scientists. As such this requires different types of content and pedagogy than what team Berkeley has so impressively built up.

This leads to my next observation—well probably more of an opinion. We need to make sure that we are leading with questions or problems, rather than data and algorithms. Indeed, the authors describe impressive ways about how they link their content to making "decisions". And their connector courses pull a lot of weight in this respect. But problem definition—what is the problem we are trying to solve—is too often neglected in industry and academia when it comes to data science. This creates a trap wherein impressive resources are deployed but any insights, inferences, etc. that come out are unable to be used in practice or even in principle.

Connecting to the first observation about broadening out who can be reached with a data science education, part of this problem is that managers, decision-makers, are not posing problems in coherent ways that then data scientists can act upon. This plays out in academia as well. Too often I hear graduate students marveling over how long it took them to put together a massive new data set. And yet when you ask 'what questions will you be able to answer', the birds chirp too often. Indeed, I myself sometimes catch myself in this trap. And the same concern even holds for those developing new algorithms.

Enforcing problem centered approaches to educating aspiring data scientists themselves is also important because it will help them communicate more effectively—and with greater confidence—with managers and decision-makers in this organization. And it might well be that encouraging data scientists to take purely non-data courses will help in this. I see great benefit in such students taking courses in microeconomics so they better understand concepts like compliments and substitutes, thinking on the margin, etc., game theory so they understand the primacy of strategic interaction in many contexts, and ethics courses to understand different notions of freedom and liberalism. As such, establishing better "two-way" communication between data scientists and those who work alongside them or those who draw on their talents is crucial. This focus forms part of the pedagogical bedrock of a series of online courses Harvard is producing (link).

Finally, I think there are several content areas that data science education might do well to beef up. First, more discussion of data quality and the attendant data wrangling/cleaning is needed. These steps are often needed to transform data into a place where it can reasonably be used to answer a question/solve a problem. It is covered in the great Berkeley curriculum. But when in

the trenches, it becomes a *massive* part of day to day life (indeed, a sort of inside joke amongst data scientists inspiring many social media memes). And this goes beyond the mechanical aspects of data wrangling (merges, filtering, reshaping, etc.). It includes thinking about data quality from the perspective of 'what process generated this data'. Are there selection effects such that causal claims might be difficult to support? What is this data representative of? Second, causal thinking is at times incorporated but often is somewhat marginalized in favor of computational or inferential statistics topics. If you are making causal claims based on data with severe endogeneity problems, I do not really care all that much about the computational tools or the size of a test-statistic/out-of-sample performance. Research design and critical thinking becomes crucial (e.g., see Bueno de Mesquita and Fowler 2021). Finally, how can we learn more by reflecting on the data we do not have. This "Dark Data" (Hand 2020) is out there. Much can be done to reflect on what this lack of data implies about the problems we are trying to solve.

References

Bueno de Mesquita, Ethan and Anthony Fowler. 2021. Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis. Princeton, Princeton University Press.

Chen, A. (2020). High School Data Science Review: Why Data Science Education Should Be Reformed. *Harvard Data Science Review*, *2*(4).

Hand, David J. *Dark Data: Why What You Don't Know Matters*. Princeton University Press, 2020.