Harvard Data Science Review • Issue 3.2, Spring 2021

Scaling Up Data Science for the Social Sciences

Brian Kim^{1,2}

¹Joint Program in Survey Methodology, College of Behavioral and Social Sciences, University of Maryland, College Park, Maryland, United States of America,

²Department of Electrical and Computer Engineering, A. Kames Clark School of Engineering, University of Maryland, College Park, Maryland, United States of America

Published on: Jun 07, 2021

DOI: <https://doi.org/10.1162/99608f92.d3f14ea4>

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

The effort to build the data science curriculum at the University of California, Berkeley described in [“Interleaving Computational and Inferential Thinking: Data Science for Undergraduates at Berkeley”](#) by Adhikari et al. (2021) is incredibly impressive. By far the most impressive part of the endeavor is not just that they have been able to generate an enormous amount of demand for their classes, but that they have stepped up to the task of meeting that demand. Their strategy of using the computational aspects of data science to help build intuition behind the inferential aspects has helped them expand and reach students beyond just the so-called ‘hard’ science majors.

As Adhikari et al. mention, the applications of data science are an inseparable part of the data science curriculum. However, it is easy to fall into the trap of thinking that this means we should only consider how data scientists apply their methods. Indeed, we should be careful not to think of data science as the realm of data scientists only. As computational tools and resources become more prevalent and data sources increase the quantity and quality of human data, the social scientists of tomorrow must also be trained in the skillful and thoughtful use of data science techniques. Twitter and other social media offer a rich source of data unlike anything seen before. Algorithms and predictive models are increasingly used to make decisions not only at big tech companies but also in federal and local governments (Kreuter et al., 2019). An integral part of teaching data science has been the inclusion of ethics, bias, and fairness in the curriculum, and this is highlighted even more within the social sciences, where these discussions are already taking place outside of the data science context.

Data Science for Social Scientists

I emulate the vignette technique employed by Adhikari et al. here to demonstrate this need. One study discussed in a graduate class titled “Machine Learning for Social Science” that I have taught is described in “To Predict and Serve?” by Lum and Isaac (2016). In this article, the authors describe how using drug arrest data for predictive policing can lead to biased results by comparing it to actual drug use from a survey of drug use. The difference in data sets is startling, with stark differences in the racial makeup and geographic location of individuals within the drug arrest and drug use cases.

This is an example used to highlight possible issues that might arise when using machine learning models to make predictions with biased data. It is used to introduce the concept of ‘garbage in, garbage out.’ But it is also used to dig a bit deeper into that. We had used cross-validation to be able to generalize before—why does it not save us here? What are the mechanisms by which this type of bias might have arisen? This is a relatively obvious case in which we might have reason to suspect the data might be biased in some way, but how might we detect such cases when it isn’t so easy to see? Perhaps most importantly, what can be done to ensure that we are able to utilize powerful data science tools while making sure we aren’t committing these types of mistakes? These are all questions that social scientists must be able to understand and answer because these are the types of questions that arise every day in our increasingly data-driven world. As data science methods become more

and more ubiquitous in the world and how decisions are made, understanding those methods is becoming integral to understanding the people that they affect. It is clear that this concept of placing the problem in context is important for data scientists to understand, but just as important is the need for social scientists to understand the mechanism by which these biases might arise.

The Path Forward

What, then, needs to be done? It is important for social science departments to make sure that data science education is both accessible for their students as well as relevant to the field they would apply it in. One model to achieve these goals is outlined by Adhikari et al. The Berkeley Data 8 class has been wildly successful, attracting many students who will go on to other fields and feeding into connector classes to facilitate the inherently interdisciplinary study of data science. Providing a large, general education–level course to provide the basics with connected, field-specific courses would allow for both the accessibility of material and relevance to the field. With the buy-in of many different units across the campus, such a large, wide-reaching general education course might be possible. However, this model may not be portable to other institutions, and it is possible that students without the background that Berkeley students have may need a more tailored class.

Because of this, building up data science for the social sciences will prove challenging. More traditional academic courses are relatively straightforward to scale up. They might require, for example, simply finding additional faculty to teach the courses, or require a few additional graduate students to act as assistants. Data science courses as imagined by Adhikari et al. require a much more complex ecosystem of support staff. The use of Jupyter Notebooks within a cloud computing environment with auto-grading—along with graduate and undergraduate student assistants—is what allows such synergy between computational and inferential thinking, but the financial and human burden to achieve this is clear.

As part of the effort to modernize and keep up with the growing interest in data science, we at the University of Maryland introduced an undergraduate course with no prerequisites titled Data Science for the Social Sciences. The pilot course was with a set maximum of 25 students for two semesters in Fall 2019 and Spring 2020. Students from various departments in the College of Behavioral and Social Sciences, such as Economics, Government and Politics, and Psychology, took the course. We adapted the material in the Data 8 class to make sure it was relevant and directly applicable to the social science majors, but followed the same philosophical approach. This involved not only using the same Jupyter Notebook method for delivering course material but also setting up a cloud computing environment for students to work in without needing to install anything on their own computers.

The class was a success, with many students indicating that they were happy to have learned these skills. Some even mentioned that the course material would be directly applicable to the job they would be starting after graduation. Each semester, the class quickly reached maximum enrollment and filled up into the waiting list, indicating a clear demand that was difficult to meet with the resources that we had. This is the challenge that

we and many others will face as many more social science students want and need some level of data science training.

Conclusion

Building up the data science capacity in the social sciences is a challenging task. It is also a necessary one. We are at a critical point in guiding how data science is taught, and the social sciences must not be left behind in this endeavor.

Disclosure Statement

Brian Kim has 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>
- Kreuter, F., Ghani, R., & Lane, J. (2019). Change through data: A data analytics training program for government employees. *Harvard Data Science Review*, 1(2). <https://doi.org/10.1162/99608f92.ed353ae3>
- Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, 13(5), 14–19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>

©2021 Brian Kim. This article is licensed under a Creative Commons Attribution (CC BY 4.0) [International license](#), except where otherwise indicated with respect to particular material included in the article.