One Size Does Not Fit All: The Shortcomings of the Mainstream Data Scientist Working for Social Good

Alex Albright & Sarah Levine, Stanford Law School Research Fellows
apa@law.stanford.edu & slevine@law.stanford.edu

Data Science: One Size Does Not Fit All

Data scientists are increasingly called on to contribute their analytical skills outside of the corporate sector in pursuit of meaningful insights for nonprofit organizations and social good projects.

We challenge the assumption that the skills and methods necessary for successful data analysis come in a "one size fits all" package for both the nonprofit and for-profit sectors.

By comparing and contrasting the key elements of data science in both domains, we identify the skills critical for the successful application of data science to social good projects. We analyze five well-known data science programs and bootcamps in order to evaluate their success in providing training that transfers smoothly to social impact projects.

We compare and contrast the roles of data scientists in the for-profit and nonprofit environments, and identify three key differences:

- While for-profit data scientists often work with in-house data, nonprofit data science often involves working with foreign data that merits greater scrutiny and sensitivity in its treatment.
    - Understanding the true meaning of each datum is rarely straightforward with foreign data.
    - Having immediate access to the individuals who collected or managed the data is an added bonus, besides avoiding bureaucratic barriers, privacy protocols, and other logistical hurdles.
- While the corporate environment provides control over the quality of data "insights" in the form of management, the non-profit environment can lack checks and balances.
    - To a degree, competition in a profit-driven environment, in addition to the presence of several technical staff to check each other's work, regulates quality of analysis.
    - In academia, the academic's reputation maintains high standards of empirical work. Peer-review ensures critical evaluation by experts before a publication. A thorough and precise work is rewarded by being cited. The university honor code is a less tangible imperative that maintains the standard of work in academia.
    - The empiricist in a non-profit organization is often the only of her kind in the organization, and operates with neither the critical review of profit-minded management nor the impetus of peer review and an academic code of ethics. While nonprofits are scrutinized by stakeholders, funders, and government oversight, a lack of formal and informal controls on data quality and analysis, analogous to a business or academia, render nonprofits prone to poor insights.
- In experimental design, for-profit data scientists can have near-omniscient control over the environment containing study variables, whereas real-world data and studies are seldom so fortunate.
    - Data scientists in the corporate environment may have the opportunity to perform experiments testing virtually unlimited hypotheses, refining their research questions through iteration. Experiments are inexpensive and easy if the audience is large.
    - A social researcher infrequently can conduct experiments that are relatively cheap and easy to iterate. Therefore, meticulous study design is far more critical before beginning any experiment, and experimentation is a less viable option in general.

These elements are not cut and dry in either sector, and there may be considerable overlap in the nature of the work that data scientists do in any type of organization. However,

the skills and methods required for data science in a corporate setting are sufficiently different from those in a nonprofit environment to merit further discussion and invoke questions about how data scientists are being educated and prepared for each scenario.

Recruiting data scientists as volunteers or casual contributors is increasingly popular. However, this very notion requires that the organization and the data scientist herself have a baseline understanding of the skills required to meet the challenges of the data. This raises the further questions: **How can data scientists become better prepared for the challenges that face them in nonprofit work? How can organizations be better prepared to receive this assistance?**

A Theoretical Framework for Holistic Data Science Education

We survey popular data science curricula across bootcamps, online courses, and master's degree programs in order to generalize the baseline knowledge of emerging data scientists. We compare and contrast the skills delivered by contemporary data science education (we hone in on five particular programs) with those required for meaningful contribution to social impact projects, and find that the former caters **strikingly to a for-profit position.**

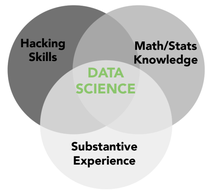
1. We compare how well-known programs stack up when it comes to the steps necessary in social impact projects. To do this, we create a theoretical framework for the successful implementation of such a data science project. The framework that we choose is sequentially as follows: *Question Conceptualization, Research Design, Data Selection, Data Collection, Data Investigation, Data Wrangling, Analysis, Interpretation of Results, and Communication of Results*. We use Table 1 to illustrate whether or not our five programs of interest feature the elements in this framework:

Table 1: The Elements and Non-Elements of Data Science Curricula

	Metis	Galvanize	General Assembly	Coursera	Data Science@Berkeley
Question Conceptualization	■				■
Research Design			■		■
Data Selection					
Data Collection	■	■	■	■	■
Data Investigation					
Data Wrangling	■	■	■	■	■
Analysis	■	■	■	■	■
Interpretation of Results			■		■
Communicating Results	■	■	■	■	■

Note: The above table clarifies which elements are present in each of the five selected Data Science programs. A green cell means that an element is covered in a program, while a gray cell means that an element is not covered in a program.

2. Before elaborating on the elements that are most lacking in these curricula, we must discuss the diverse domains of knowledge and training that come together in Data Science. Data Science, while often a nebulous concept, can be concretely visualized using the well-known Conway Venn Diagram, which breaks data science up into three components of equal size: Hacking Skills, Mathematical and Statistical Knowledge, and Substantive Experience:



3. As an exercise to further the discussion of missing components within Data Science education, we categorize which domains of the Venn diagram contain each of the previously mentioned elements from our framework:

Table 2: The Elements and Non-Elements of Data Science Sub-Domains

	Hacking Skills	Mathematics & Statistics Knowledge	Substantive Experience
Question Conceptualization			■
Research Design		■	■
Data Selection		■	■
Data Collection	■		
Data Investigation		■	■
Data Wrangling	■		
Analysis	■	■	■
Interpretation of Results		■	■
Communicating Results	■	■	■

Note: The above table clarifies which of the aforementioned elements fit in each of the three Data Science sub-domains, as defined by Conway's Venn Diagram. A green cell means that an element is in a domain, while a gray cell means that an element is not in a domain.

Bridging the Gap

In comparing Tables 1 and 2, it is evident that Data Science programs are not entirely embracing the *"Substantive Experience"* element of Data Science. Without an emphasis on (or even mention of) substantive knowledge, data science veers dangerously close to a straightforward Bayesian approach, which hopes that simple numbers will reveal the truth of conceptually complex questions without theory supporting any of the underlying ideas.

The curricula of the programs we surveyed can read like a laundry list of tools with no structure to rein them in. The focus on hacking and knowledge of Python, pandas, Git, matplotlib, MapReduce, NoSQL, R, SQL, D3, Javascript, Hadoop, and so forth is attractive to private sector employers. However, these tools do not make for great social impact projects without deeper understanding of data treatment and study design.

One program features a section on "explor[ing] and visualiz[ing] data" in the very first lesson, titled "Unit 1: The Basics." We posit that data exploration and visualization should appear only secondarily to a fundamental understanding of study design, data treatment, and interpretation of results. Similarly, another program describes the first week of the curriculum as, "Exploratory Data Analysis and Software Engineering Best Practices." While this is promising, "engineering" best practices should certainly be secondary to statistical fundamentals and best practices. It is essential to avoid the popular plug-and-chug methodology, in which various models are thrown at a dataset without careful regard for their compatibility, if we seek to maintain even the slightest flavor of science in Data Science.

Unfortunately, the majority of popular programs we survey lack modules on the importance of theoretical foundations and careful research design.

Indeed, many programs acknowledge their corporate-orientation, as Metis' website even greets visitors with a video in which an instructor states:

"[A data scientist is] a person that uses the scientific method, that has been used on a lot of scientific data in nature, on data from businesses."

This focus on corporate data, math and hacking skills ignores the benefits of substantive experience for data science in the social impact realm. In this sense, the shortcomings of current curricula necessitate introspection into the evolving nature of Data Science and a collective decision by the community about how to shape its formation to make the best use of data science across both sectors.

The curricula of these courses clearly illustrate that data scientists are molded to be corporate workers as the default, necessitating a further mechanism to help empirical researchers transition across sectors, even if they bear the same title: "data scientist."

Ultimately, we conclude:

- Whereas for-profit data science can often afford to be "insights"-driven and results-oriented, nonprofit data science must be less content-driven and more process-oriented to avoid results, conclusions, and even policies built on poor quality data and inappropriate methods.
- Nonprofit organizations can and should be more targeted in their hiring practices for data scientists. Nonprofits should be cautious before using the same hiring criteria as for-profit organizations. This may require further self-reflection, and should initiate non-profits to set their own requirements separate from the for-profit channel.
- Lastly, there must be a further discussion of ethics in data science for social good. The community should engage in practical dialogue in order to develop best practices and ethical codes akin to those in the academic community. Data scientists should be held accountable for their insights, especially in the realm of social good projects. Moreover, developing a cultural norm of peer review will provide a check and balance on the quality of analysis, while providing more opportunities for identifying common pitfalls.