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

Alex P. Albright
Research Fellow, Stanford Law School
559 Nathan Abbott Way
Stanford, CA 94305
1 (917) 751-1910
apa@law.stanford.edu

Sarah M. Levine
Research Fellow, Stanford Law School
559 Nathan Abbott Way
Stanford, CA 94305
1 (201) 658-8669
slevine@law.stanford.edu

ABSTRACT

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 then 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. After surveying these programs, we make a number of recommendations with respect to data science training curricula, non-profit hiring systems, and the data science for social good community's practices.

General Terms

Economics, Reliability, Experimentation, Theory.

Keywords

Data Science, Education, Social Impact.

1. INTRODUCTION

While the overwhelming majority of data scientists are employed in the for-profit sector, there is a growing movement taking advantage of their technological savvy and unique toolkit for the benefit of social good projects and programs. Conventionally trained data-scientists are encouraged more and more to play a pivotal role in data-driven social good projects as team members, consultants, or volunteers. However, this phenomenon assumes that the data scientists' standard toolkit in the for-profit sector translates seamlessly to the realm of social good. We challenge this assumption and argue that while the term "data scientist" has become an amorphous catch-all for programmers, statisticians, bloggers, and other empirically inclined individuals, the skills and methodological knowledge required of a data scientist can and should differ across the for-profit and non-profit sectors. We use this paper as an opportunity to highlight the shortcomings of mainstream data science education and practice when it comes to the non-profit sector and social impact endeavors.

We begin by comparing and contrasting the roles of data scientists in the for-profit and non-profit environments, and identify three key differences. First, while for-profit data scientists often work with in-house data, non-profit data science often involves working with foreign data that merits greater scrutiny and sensitivity in its treatment. Second, while the corporate environment provides control over the quality of "insights" in the form of management, the non-profit environment can lack effective checks and balances on data and analysis quality. Third, in experimental design, for-profit data scientists often have near-omniscient control over the environment containing study variables, whereas real-world data and studies are seldom so fortunate. We conclude that whereas for-profit data science can often afford to be "insights"-driven and results-oriented, non-profit data science must be less content-driven and more process oriented to avoid results, conclusions, and even policies that are built on poor quality data and inappropriate methods.

Next, 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 then compare and contrast the skills delivered by contemporary data science education with those required for meaningful contribution to social impact projects, and find that the former caters strikingly to a for-profit position. For example, we find that there is little to no focus in current data science education on investigating the quality of data or the identification and integrity of experimental variables. The curricula of these courses 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 make several recommendations as to (1) how data science training programs can better prepare their students for roles in organizations doing social good, (2) how non-profit organizations can and must be more targeted in their hiring practices to find data scientists who are adequately suited for their projects, and (3) how the data science for social good community can and must develop best practices and ethical codes akin to those in the academic community.

Conference'10, Month 1–2, 2010, City, State, Country.

2. BACKGROUND

We begin by comparing and contrasting the roles of data scientists in the for-profit and nonprofit environments, and we identify three key differences. First, 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. Working with foreign data, generated outside of the organization seeking to analyze it, has many potential pitfalls. For example, in working with inspections data that spans several years, the researcher must ask several questions that establish a base understanding of the data: Were inspections conducted uniformly for all years across this sample? Was the selection process for establishments subject to inspection the same for all years in this sample? Beyond these questions, which are key for successful analysis of the data, there are a myriad of more granular questions: Are total profits per establishment adjusted for inflation? Do total profits per establishment represent gross or net values? Understanding the true meaning of each data point is rarely straightforward when using foreign data. Moreover, it is increasingly difficult to accurately interpret data when several degrees removed from the people or organization that performed the collection and cleaning. In a corporate environment, it is not atypical to have teams devoted to collecting and managing all company-related data. Data scientists, even if operating in a completely different wing of the organization, will often have access to the very individuals who make decisions about how data is recorded. Clearly, working with "native" data provides many advantages. Understanding how to properly treat certain values or interpret variables is a tremendous asset and time-saver. Having immediate access to the individuals who collected or managed the data is an added bonus, besides the benefit of avoiding bureaucratic barriers, privacy protocols, and other logistical hurdles.

Second, while the corporate environment provides some control over the quality of data "insights" in the form of management, the non-profit environment can lack effective checks and balances on data and analysis quality. In a corporate workplace, the standards for quality of data and analyses are not uniformly high. However, to a degree, the nature of competition in a profit-driven environment sufficiently regulates the quality of work. In contrast, non-profit organizations, often operating on tight budgets, infrequently have any empirical staff, much less several individuals who can check each other's work. Budget-strapped organizations may be unable to match the salaries of competitive firms, leaving them with researchers whose skills and methods are not the most up-to-date. When nonprofit organizations skirt this problem by outsourcing their empirical analyses, by soliciting volunteers and "citizen hackers," the quality of the work can all too often be haphazard and lack a necessary degree of rigor. Regardless of the intentions of the citizen hacker, they simply do not bear the same emotional or social ties to the organization as an employee would, nor do they have equivalent accountability. Ultimately, the quality of results can be undermined at many locations along the pipeline: the data, the researcher, the analysis, or the tools. Lack of proper checks and balances leaves this nonprofit empirical research open to many pitfalls.

Academia is an interesting outlier in this regard for two distinct reasons. First, the role of the reputation of an academic is critical for maintaining high standards of empirical work. It functions by incentivizing researchers to be thoughtful in their treatment of data, thorough in their evaluations, and specific in their conclusions. At the same time, peer-review requirements ensure that any piece of work is critically evaluated by other experts in the field long before any study becomes public. A particularly thoughtful, thorough, and precise work is then rewarded by a culture of citation. In addition to these practical checks and balances, there exist less tangible imperatives that maintain the standard of work in academia, namely, the university honor code.

This provides a stark comparison to the empiricist in a non-profit organization, who 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 even government oversight, a lack of both formal and informal controls, specifically on the quality of data and analysis, render the nonprofit sector especially prone to poor insights.

Third, in experimental design, for-profit data scientists often have near-omniscient control over the environment containing study variables, whereas real-world data and studies are seldom so fortunate. For example, imagine that a corporate data analyst working for a technology firm would like to perform an A/B test to determine whether users of a website respond more favorably to a button that says, "keep me in the loop" or "send me emails." For one month, the company serves 50% of its online visitors one site with the first button, and serve 50% the other version. In short, the data scientist wants to evaluate which treatment has a more favorable effect on the study group. In contrast, a data scientist consulting for a nonprofit wants to explore which of two anti-smoking policies has had the most positive behavioral effect. State A had the treatment, policy A, applied, and State B had policy B applied. While the analogy is clear - two treatments applied to two samples - the latter empiricist will encounter many more difficulties. First, the application of the treatment is more straightforward in the first case: the button is served. In the case of real-world data, the application of treatment can be hard to gauge: how were the policies implemented? How were they enforced? The corporate data analyst has a clear if not obvious metric of success: the number of times that visitors click the button. For the social researcher, there are a vast number of ways to evaluate the success of the policies on limiting the number of smokers. While the two data scientists are working towards the same goal in the abstract, their paths to a clean evaluation of the respective treatments are not equally direct.

In many cases, like that of the A/B tester, and often in technology companies, corporate data scientists have the opportunity to design experiments and conduct studies in entirely controlled environments. In contrast, a nonprofit data scientist will often be analyzing data from a study or social program with greater uncertainty and a greater predisposition to omitted variable bias. Naturally, the causal pathways in these types of studies are infrequently linear and there can be many steps between intervention and outcome. Data scientists in the corporate environment may have the opportunity to perform many experiments testing virtually unlimited hypotheses, thus refining their research question through iterative perfection. The experiments are inexpensive and easy if the audience is large. A social researcher infrequently has the opportunity to conduct experiments that are relatively cheap and easy to iterate. Therefore, meticulous study design is far more critical before beginning any experiment.

These elements - data origin, checks and balances, study design - 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. Most importantly, while both roles are undeniably critical and can learn from the practices of their counterparts in other sectors, data science should not be treated as a "one size fits all" solution to any empirical question.

Recruiting data scientists as volunteers or casual contributors is increasingly popular. In Harvard Business Review, Claudia Perlich went so far as to propose a "year-round virtual marketplace (perhaps modeled after DonorsChoose) where data scientists can find NGOs whose needs are well-matched to the skills and time they can donate." [1] However, this very notion requires that the organization requesting analytical assistance and the data scientist herself have a baseline understanding of the skills required to meet the challenges of the data. This raises the question: 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?

3. INVESTIGATION

As Data Science has grown in popularity, programs designed to mold individuals to fit this in-demand role have quickly and substantively evolved. Not only are there an increasing number of formal higher education degrees offered in Data Science, but there is also a steady emergence of online courses and bootcamps focusing on the same topic. Given Data Science's adolescence as a discipline, the field is continuously shaped by existing data scientists as well as by these very curricula. These developing programs are refining the domain, scope, and approaches of Data Science, which is why they are important to consider when establishing the platonic ideal of a well-trained data scientist.

Instead of attempting to survey all Data Science programs in order to investigate the ways data scientists are being created, or molded from other related disciplines, we focus on five particular programs of interest. These five programs are a cross-section of different types of programs; we consider well-known bootcamps (Metis and Galvanize), part-time classroom-style courses (General Assembly), online courses (Coursera, University of Washington Course), and one Master's program (Data Science@Berkeley).[1] In surveying these five programs, we posit that we are considering a representative sample of the overarching data science education. We attempt to model the formation of a typical data scientist, and to identify the common threads throughout these varying styles of education. Our evaluation is admittedly cursory: we use online materials from these courses and are guided by feedback from former students. However, we believe this first step is sufficient to identify themes and areas for improvement.

In considering an array of popular Data Science programs, we are able to identify common elements among their diverse approaches and we take close note of their respective technical focuses. In particular, 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.[2] We then use a simple table, Table 1 on the following page, to illustrate whether or not each program features the elements in this aforementioned framework.

Before elaborating on the elements that are most lacking in these curricula, it is worth discussing the diverse domains of knowledge and training that come together in Data Science. A commonly cited resource to explain the often nebulous concept of "Data Science" is Conway's Venn Diagram, which breaks data science up into three components of equal size: Hacking Skills, Mathematical and Statistical Knowledge, and Substantive Experience. [2] 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. This investigation yields Table 2, presented on the following page.

In comparing Tables 1 and 2, it is immediately evident that Data Science programs are not entirely embracing the "Substantive Experience" element of Data Science. Even Conway was aware of the bias against this piece of the field; he admits that it is:

"In the third critical piece—substance…where my thoughts on data science diverge from most of what has already been written on the topic. To me, data plus math and statistics only gets you machine learning, which is great if that is what you are interested in, but not if you are doing data science. Science is about discovery and building knowledge, which requires some motivating questions about the world and hypotheses that can be brought to data and tested with statistical methods." [3]

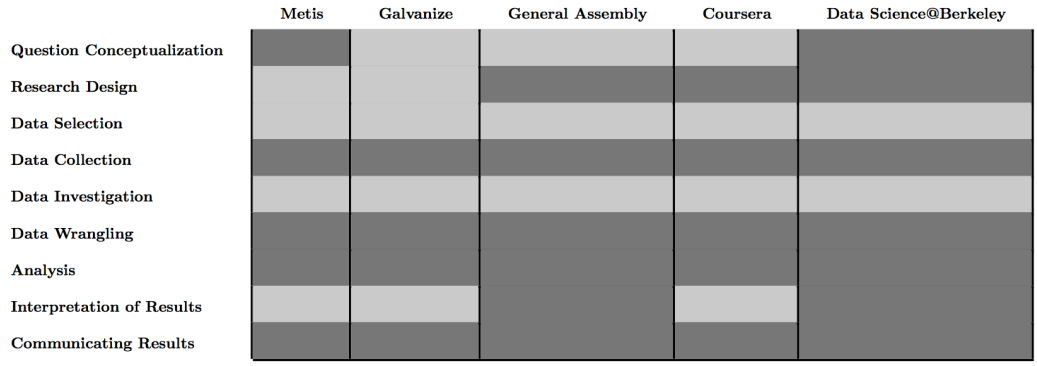
The curricula of the programs we surveyed can read like a laundry list of tools with no structure to rein them in - as though they lack the specific critical element that Conway describes. The focus on hacking and knowledge of Python, pandas, Git, matplotlib, MapReduce, NoSQL, R, SQL, D3, Javascript, 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. 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.

Having provided a theoretical framework for educating data scientists qualified to embark on social good projects, we now explore the particulars that these programs are lacking. Many of these elements are directly related to components of data science in nonprofit and for-profit sectors, as identified in the previous section.

[1] See section following footnotes for program information.

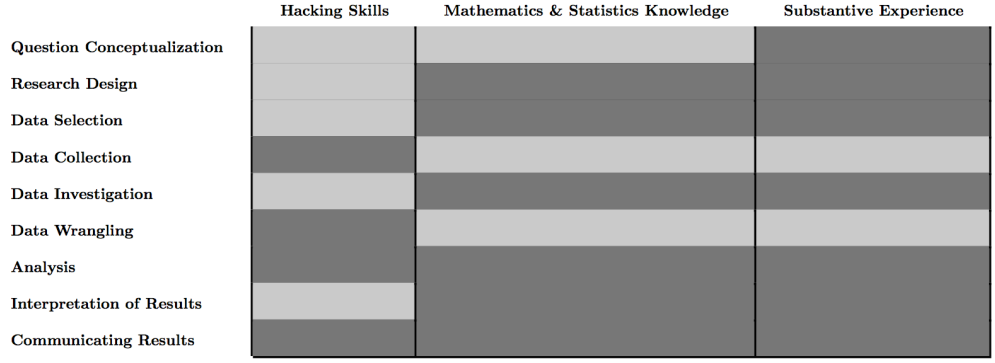
[2] There are a plethora of frameworks that one could specify. We present one that we believe covers all the necessary steps for approaching a social good problem rigorously. In this framework, it is worth noting the difference between Data Selection, Collection, Investigation, and Wrangling. Collection and wrangling are the two elements that are most familiar to data scientists. However, the Data Selection element refers to compiling data sources and determining usability of data. Meanwhile, data interpretation refers to the steps one takes in order to better understand chosen variables and how transformations of the variables could be useful.

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



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

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



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 dark cell means that an element is in a domain, while a light cell means that an element is not in a domain.

3.1 Question Conceptualization and Research Design

Consider first the lack of question conceptualization. The scientific method requires questions and hypothesis testing, not just the admittedly complex, technical application of models. However, the curricula herein mentioned often do not apparently push students to think about questions and approaches, as they assign a focused research question, and distribute data suited to solving that problem.[3] These practices do a disservice to students as they do not allow for flexibility in scoping out data sources and questions, which is a skill continuously practiced in the social impact sphere. While it may make logistical sense for capstone projects to provide all students with a given question and dataset in order to make them comparable and more easily gradable, there is a cost to not spending more course time on generating good research questions and surveying available data.

The emphasis of only two of three (of the five) programs on theory-based research design is disturbing in the context of social impact projects. While in some contexts, a simple prediction might be a suitable insight, a conceptual understanding of a phenomenon requires heeding the warning of Professor Gary King, "If we start with a dependent variable and try to search for all possible (or all 'big' or all 'important') explanatory variables, we shall continually lose leverage over the problem." [4] Moreover, an empirical investigator must understand that "[t]he usefulness of a particular model specification depends entirely on what causal or forecasting goals one pursues...thus, our theoretical reason for a model is our best guide to specification." [5] In this sense, research design and question conceptualization go hand in hand - study design is inextricably dependent on the motivating research question.

On this topic, it is critical to note that not all research designs are equally appropriate or rigorous in all contexts. In particular, consider that the Metis curriculum explains that "[i]n preparation

[3] This is also true for DrivenData and Kaggle competitions. The question and raw materials for these specific competitions are givens while the real focus lies within the methods of statistical analysis.

for Project 2, students start to learn one of the most important tools a data scientist uses: the iterative design process." [6] While iterative design is an incredibly useful tool, it is only applicable in specific contexts. Most importantly, highlighting iteration as a key tool for data science detracts from the far more important topic: careful research design based in theory before experimentation. 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.

3.2 Data Selection and Data Investigation

There is scarcely ever a single correct answer about what collection of data should be used to answer a question. The best path forward involves carefully weighing the options and acknowledging any uncertainty in a data-based decision. Moreover, there is the possibility in the social sector that data is sparse and poor in quality. D.J. Patil, the current US First Chief Data Scientist, explained in 2012 that, "[M]ore than anything, what data scientists do is make discoveries while swimming in data." [7] Patil's description, likely geared toward a corporate audience, describes a scenario of abundant data from which "discoveries" can be extracted. In the case of social good projects, data can be scant and unreliable, and therefore merits a careful wading, rather than unscrupulous, exploratory "swimming." For these reasons in particular, data topics beyond Data Collection and Wrangling must be addressed in Data Science curricula.

A brief example of the need for thoughtful Data Selection and investigation can be found in an article about University of Chicago's Data Science for Social Good Fellowship, which describes challenges in a project evaluating graduation rates, "[T]here have been issues parsing through data, as many schools report disciplinary issues, truancies, and other factors differently," due to limited human resources for data collection at the schools. [8] Schools report similar outcomes differently, which is an issue when aggregating across sources rather than acquiring all data from one central source. In this example, the data scientist discovered that there is heterogeneity in school measures, which is a common issue that is dangerous when disregarded in analysis. Data scientists who do not have formal training in inspecting the source and nature of data may implement bad practices and derive meaningless results by not addressing concepts such as heterogeneity and, therefore, not devising suitable models for the data.[4]

One telling detail on this topic is that General Assembly's program features a section on "explor[ing] and visualiz[ing] data" in the very first lesson, titled "Unit 1: The Basics." [9] We posit that data exploration and visualization should appear only secondarily to a fundamental understanding of study design, data treatment, and interpretation of results. Presenting data visualization as a core component of data science - as opposed to a single method of communicating information derived from data - may be putting the proverbial cart before the horse. Similarly, Galvanize's program describes the first week of the curriculum as, "Exploratory Data Analysis and Software Engineering Best Practices." [10] While this is promising, "engineering" best practices should certainly be secondary to statistical fundamentals and best data practices. However, buried in their FAQ page, they do state, "Through working with messy, real-world data sets, students gain experience across the data science stack — data munging, exploration, modeling, validation, visualization, and communication." [12] Ultimately, it's unclear how much hands-on experience with scrutinizing data students will receive.

In the last three weeks of Metis's Data Science bootcamp, students work on a capstone project. One of the necessary steps is to "[c]hoose data sources that can be used to address" the problem of interest. [13] The program explains that students have been "slowly developing" the project over the past 8 weeks of their training. However, the absence of this essential step from the first 8 weeks of the curriculum indicates an assumption that this component is straightforward and does not merit technical expertise, unlike the laundry list of hacking skills promised in promotional materials. This step is included as one of many finishing touches on the Data Science education when, in reality, this process is the very crux of what data scientists must do for successful contributions to social good projects.

3.3 Interpretation of Results

In the aforementioned Data Science curricula, there is a quick transition between the completion of analyses and the presentation of results. However, there is little to no discussion of the interpretation of results, the presentation of all inevitable uncertainties, clarifying all caveats to results, and presenting further explanatory theories along with next steps. Data scientists, who are too often trained to systematically produce "insights" without critically investigating the reliability of those results and the implications stemming from them, are poorly positioned to contribute to social good projects. Meaningful interpretation of results is rooted in the formation of a research question, the design of the study, and the selection and investigation of data: the very arenas in which the programs we surveyed are most lacking. Data scientists in training must be encouraged to take time to mull over results, rather than being pushed to be results-driven, which ultimately undermines the value of those very results. It is important to remember Robert Luskin's words that, "[t]he interpretation of statistical results is art as well as science." [14] Data science programs, and future data scientists themselves, must be more conscious of these artistic elements in data science, in addition to the components of science previously discussed. There is a rarely a single correct model or an obviously conclusive result and teaching Data Science as an iterative experimental process of impulsively testing models on data detracts from the likelihood of fit models and meaningful results.

3.4 Summary

Data Science@Berkeley's robust syllabus places the most emphasis on the three components we emphasize. The course description states, "This course introduces students to experimentation in the social sciences… Key to this area of

[4] In fact, there is a DrivenData competition that seeks to use "data from social media to narrow the search for health code violations in Boston." With access to historical hygiene violation records and Yelp consumer reviews, the competitors try to determine which words, ratings, etc. could predict health violations. However, it is worth noting that the outcome measure one is trying to predict here is not totally objective, given that inspector heterogeneity is pervasive. There is no discussion of variation in inspector strictness in this measure, which is necessary to understand at the data interpretation phase in order to accurately explain the caveats and limitations of an empirical approach. [11]

inquiry is the insight that correlation does not necessarily imply causality. In this course, we learn how to use experiments to establish causal effects, and how to be appropriately skeptical of findings from observational data." [15] Building critical analysis, specifically skepticism, into a data scientist's repertoire is crucial for avoiding superficial interpretations and erroneous conclusions. Moreover, the course also mentions ethical considerations in Data Science, "from collection, to storage, processing, analysis and use including, privacy, surveillance, security, classification, discrimination, decisional-autonomy, and duties to warn or act." [16] These two elements, the importance of both ethics and healthy skepticism in approaching any task as a data scientist, merit further emphasis in all other programs.

Despite these commendable programmatic features of the Berkeley program, in sum, the curricula of these courses illustrate that data scientists are being molded primarily as for-profit workers by default. Indeed, many programs acknowledge their corporate-orientation, as Metis' website even greets visitors with a video in which an instructor states, "What a data scientist is is … a person that uses the scientific method, that has been used on a lot of scientific data in nature, on data from *businesses*" (emphasis supplied). However, 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.

4. CONCLUSION

Given the multi-dimensional theoretical nature of Data Science spanning hacking skills, mathematics and statistics knowledge, and substantive experience, there can be no doubt that current data scientists are rarely trained comprehensively in a manner that is aligned with the theoretical vision of the field. We conclude that 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 that are built on poor quality data and inappropriate methods. We suggest that education programs for Data Science incorporate more elements of substantive experience into their curricula and that data scientists with intentions of contributing to social good projects approach these topics with vigor. Specifically, all programs would benefit from the incorporation of the elements we identified in certain programs like that of Data Science@Berkeley, which successfully incorporates topics of thoughtful data selection, careful study design, and other ethical considerations with data.

Nonprofit organizations can and should be more targeted in their hiring practices for data scientists. Considering the fundamental distinctions between for-profit and nonprofit work in the Data Science universe, nonprofits should be cautious before using the same hiring criteria as for-profit organizations. This may require further self-reflection by nonprofits, and should initiate non-profits to set their own requirements separate from the for-profit channel. In other words, nonprofits cannot afford to follow the for-profit model of Data Science. 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 necessary check and balance on the quality of analysis, while simultaneously providing more opportunities for sharing methods and identifying common pitfalls.

5. REFERENCES

[1] Claudia Perlich. 2014. Recruiting Data Scientists to Do Social Good. *Harvard Business Review*. (Aug. 2014). Retrieved August 8, 2015 from https://hbr.org/2014/08/recruiting-data-scientists-to-do-social-good

[2] Drew Conway. 2010. The Data Science Venn Diagram. Retrieved August 8, 2015 from http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

[3] Ibid.

[4] Gary King. 1991. "Truth" Is Stranger than Prediction, More Questionable than Causal Inference. *American Journal of Political Science* 35: 1047-53 (1048). Retrieved August 8, 2015 from http://gking.harvard.edu/files/truth.pdf

[5] Ibid, 1050.

[6] Metis. Data Science, Curriculum. *Data Science Bootcamp*. Retrieved August 8, 2015 from http://www.thisismetis.com/documents/Data-Science-Curriculum.pdf

[7] Thomas Davenport and D.J. Patil. Data Scientist: The Sexiest Job of the 21st Century. *Harvard Business Review*. (Oct. 2012). Retrieved August 8, 2015 from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[8] Hustad, Karis. 2015. How UChicago Is Training Data-Crunching, Do-Gooders Of The Future. *Chicago Inno*. (Aug, 2015). Retrieved August 8, 2015 from http://chicagoinno.streetwise.co/2015/08/05/civic-data-uchicago-data-science-for-social-good-fellowship-hits-third-year/

[9] General Assembly. General Assembly Course Curriculum, Data Science. 1-12. Retrieved August 8, 2015 from https://ga-core-production-herokuapp-com.global.ssl.fastly.net/assets/course_applications/data-science/Data_Science_Course_-_GA-b6dfc3f4b846a87f7577b2d589d63d67.pdf

[10] Daniel E. Ho. Fudging the Nudge: Information Disclosure and Restaurant Grading. *The Yale Law Journal* 122:574. Retrieved August 8, 2015 from http://www.yalelawjournal.org/pdf/1120_ogjao8vy.pdf

[11] Galvanize. Week-to-Week. *Data Science*. Retrieved August 8, 2015 from http://www.galvanize.com/courses/data-science/#.VcYfip1Viko

[12] Galvanize. "What Will I Learn In The Data Science Program?" Retrieved August 8, 2015 from https://galvanize.zendesk.com/hc/en-us/articles/203851409-What-will-I-learn-in-the-data-science-program

[13] Metis. Data Science, Curriculum. *Data Science Bootcamp*. Retrieved August 8, 2015 from http://www.thisismetis.com/documents/Data-Science-Curriculum.pdf

[14] Robert Luskin. 1991. Abusus Non Tollit Usum: Standardized Coefficients, Correlations, and R2s. *American Journal of Political Science* 35: 1030-44 {Nov. 1991). (1044).

[15] Data Science@Berkeley. Online Master of Information and Data Science, Field Experiments. *UC Berkeley School of Information*. Retrieved August 8, 2015 from http://datascience.berkeley.edu/academics/curriculum/course-schedule/

[16] Data Science@Berkeley. Online Master of Information and Data Science, Legal, Policy, and Ethical Considerations for Data Scientists. *UC Berkeley School of Information*. Retrieved August 8, 2015 from http://datascience.berkeley.edu/academics/curriculum/course-schedule/

5.1 Programs Evaluated

Coursera. *University of Washington, Introduction to Data Science by Bill Howe*. Retrieved August 8, 2015 from https://www.coursera.org/course/datasci

Data Science@Berkeley. Online Master of Information and Data Science. *UC Berkeley School of Information*. Retrieved August 8, 2015 from http://datascience.berkeley.edu/

Metis. *Data Science Bootcamp*. Retrieved August 8, 2015 from http://www.thisismetis.com/data-science

Galvanize. *Data Science*. Retrieved August 8, 2015 from http://www.galvanize.com/courses/data-science/#.VcZ9_p1Vikp

General Assembly. Data Science. Retrieved August 8, 2015 from https://ga-core-production-herokuapp-com.global.ssl.fastly.net/education/data-science