Democratizing Data Science

Effecting positive social change with data science

Sophie Chou*
MIT Media Lab
75 Amherst St.
Cambridge, MA
soph@media.mit.edu

William Li*
MIT CSAIL
32 Vassar St
Cambridge, MA
wli@csail.mit.edu

Ramesh Sridharan*
MIT CSAIL
32 Vassar St
Cambridge, MA
rameshvs@csail.mit.edu

ABSTRACT

The effective translation of data into novel insights, discoveries, and solutions, also known as data science, has enormous potential to bring about positive social change. In this paper, we propose ways to "democratize data science": that is, to allocate the power of data science to society's greatest needs. Two underlying challenges are 1) the misalignment of economic incentives to aspiring data scientists choosing what problems to pursue, and 2) the continued underrepresentation of certain demographics in data science. We suggest paths forward in the domains of data science systems and algorithms, educational initiatives, and social movements that will help realize our vision of democratizing data science.

Keywords

data science, democratization, society

1. INTRODUCTION

Data science is a nascent but rapidly growing discipline. In the past few years, entire industries have formed around applications of "Big Data" [3]. Government, civil society, and the media [2, 4] have begun to embrace data-driven approaches to their work, and job postings and career paths for "data scientists" have become widespread. The availability of large datasets and new tools, systems, and algorithms to analyze them means that data scientists can have an enormous impact on society.

As with any technological advancement, however, data science could benefit different people, groups, and sectors unevenly. Although data science is an interesting area of study and practice with great potential, it does not intrinsically improve social good. Rather, like any technology, it is merely a force multiplier: it can be used to augment the power of people to achieve positive or detrimental ends. However, as the potential for using data science for positive social change is enormous, we believe that there is a need — and a moral obligation — for capable minds and institutions to drive data science as a field toward solutions that effect positive social change.

"Democratizing data science", in our view, means ensuring that our discipline promotes the common good, which is often beyond the narrow commercial and demographic interests of the groups that most frequently use data science today.

Our motivation for writing this paper stems from our experiences and observations in working at technology companies, academic institutions, and civil society organizations: well-capitalized firms, particularly those on Wall Street or in Silicon Valley, are generally far ahead of governments and non-profit organizations in their sophistication and application of data science to achieve their business goals. These advantages lead to better salaries, more prestige, and a wider range of technical challenges in private industry that are difficult for our data scientist classmates and colleagues to turn down. Interestingly, publicly oriented organizations often have many interesting and meaningful datasets and challenges that data science might be able to solve. However, because of their limited resources and lack of data science capacity, they have far fewer opportunities for data scientists to apply their skills.

In this paper, we critically evaluate the data science ecosystem through the lens of "democratizing data science." In this process, we discover that the verb " democratize" has been used in different contexts to mean "make easier", "reduce the cost of", and even "crowdsource," depending on the goals of the groups that seek to popularize these competing meanings. Our definition, we argue, is most closely aligned with the benefit towards society as a whole. We describe how these existing efforts fit into our definition of democratizing data science and the gaps that must be addressed in order to realize this vision.

The remainder of this paper is structured as follows: First, we briefly provide a working definition of "data science" in Section 2, and then survey related work in Section 3. We define our vision of "democratizing data science" and the underlying causes of the undemocratic allocation of resources in Section 4. We survey related work that addresses parts of these concerns in Section 5. Finally, we describe the unaddressed gaps and discuss paths forward to address them in Section 6.

2. WHAT IS "DATA SCIENCE"?

The term "data science" has many different interpretations depending on where it is used. For the purpose of this paper, we loosely define "data science" as the set of activ-

*Sophie Chou, William Li, and Ramesh Sridharan contributed equally to this work.

ities involved in transforming collected data into valuable insights, products, or solutions. These processes typically include (but are not limited to) the following:

1. collecting data from their original sources and ingesting it into storage systems;
2. cleaning and structuring data into machine-readable forms;
3. analyzing and finding interpretable patterns or trends in datasets, often using statistical techniques;
4. visualizing and communicating the results to readers, customers, or other audiences.

This full pipeline of activities is important to keep in mind as we discuss current and future efforts to democratize data science.

3. RELATED WORK

Existing work has largely been framed around the phenomenon of Big Data and its utopian or dystopian implications for society. Although data science does not inherently require massive amounts of data, the use of large datasets creates additional substantial technical challenges. In this paper, we will distinguish the term "Big Data" from data science as being more focused on the acquisition and use of data in large quantities.

In "Critical Questions for Big Data," boyd and Crawford note emerging issues related to the study and application of large social datasets. Of particular relevance is their discussion of how people with data analysis expertise (i.e., data scientists) are "the smallest, and the most privileged: they are also the ones who determine the rules of how Big Data will be used, and who gets to participate" [5]. In addition, the people and organizations that have the capacity to collect and manage data also have power in the data ecosystem [9]. The inequalities brought about by these characteristics of data science are the focus of our paper.

Our motivation extends beyond contributing to an academic debate about the semantics of the phrase "democratizing data science"; we believe that the meaning of the phrase can shape the discourse and future of data science. One analogous example is Yu and Robinson's discussion of the ambiguity of the term "open government data" [12]. Briefly, "open government data" can refer to the processes and systems for machine-readable data, or it can be data about open and transparent government. Conflating the two uses of the term potentially muddies policy discussions about both issues. In a similar fashion, this paper seeks to describe the different meanings of "democratizing data science," which could help bring clarity to discussions about the variety of activities within data science as well as its potential impacts.

4. VISION AND PROBLEM STATEMENT

We envision a world in which the application of data science is more democratic; that is, the tools and outputs of data science are applied to the areas of greatest social need. Organizations focused on issues such as economic inequality, education, environmental protection, and civic engagement often lag far behind in terms of attracting talent and developing technical capacity to obtain the benefits of data science. In other words, there is a substantial under-allocation of human and computational effort toward these important challenges. Addressing this misallocation is an important, high-impact problem that deserves the attention of experts in data science and public policy.

We attribute this misallocation of data science resources and expertise to two key factors: biased economic incentives and misrepresentation of society and societal interests among data scientists.

First, the economic incentives are misaligned: in the absence of external incentives, for-profit companies should and will seek out applications that can increase profits, which are often different from the applications that promote overall social welfare. Such firms can invest more in obtaining data and hiring data scientists, enabling them to use the results of analyses to obtain a competitive advantage and reach markets more successfully. In the quantitative finance industry, better-funded firms can obtain more data about the market with lower latency, hire more data scientists ("quants") to analyze this data, and consequently make better investment decisions. In the technology sector, having access to large datasets and the systems to manage them allows companies to attract data scientists and engineers, which in turn increases the success and adoption of their products while producing even more data.

Second, the lack of diversity in the field of data science perpetuates the problem of misallocation. Human nature dictates that we are attracted to problems close to us; after all, it is difficult to work on problems that are beyond our awareness or understanding. People in demographics who are given less access to the skills and application of data science will be equally denied its benefits. As with computer science, data science faces challenges related to underrepresentation of gender, sexuality, and minorities, especially among groups that are disadvantaged and could benefit most from its applications. Even more so than in computer science, the barrier to entry in data science is high – after all, a data scientist is a multidisciplinary scientist. She must, in addition to writing code, have a strong background in mathematics and statistics, fields which suffer in demographic inequality as well. These issues in the education of data science participate in a negative feedback loop between those who do the analysis and those who most benefit from the results. In turn, this mismatch results in the current undemocratic state of affairs: the under-allocation of data science to problems that affect disadvantaged groups.

5. SOLUTIONS: EXISTING EFFORTS

To our knowledge, no existing initiative is uniquely devoted to making data science more democratic. However, a number of current efforts in the data science ecosystem could make useful contributions to increasing attention to pressing societal needs. We describe these efforts, the niches that they can fill in democratizing data science, and their limitations.

5.1 Data Science Technologies: Systems, Tools, and Algorithms

The rise of data science has led to new technologies designed to make data science easier. Recent advances include cloud computing platforms (e.g., Amazon EC2) to eliminate the need for in-house computer clusters, graphical tools for cleaning and structuring data (e.g., DataWrangler), machine learning algorithms that reduce the human effort required for analysis (e.g., deep learning), and easy-to-use data exploration and visualization software for finding insights (e.g., Tableau). These technologies claim to "democratize data sci-

ence" by reducing the cost and technical expertise required to work with data. As a result, more people and groups, including those focused on important societal issues, can in theory harness the benefits of data science.

While these systems, tools, and algorithms are helpful, they are only one part of democratizing access to the benefits of data science. In reality, most of these technologies cannot simply be used by anyone: they require experience with the underlying principles, technical jargon, unwritten know-how, and culture of data science. For instance, the need for careful statistical analysis requires particular experience and understanding; a poorly conducted or misinterpreted statistical analysis can be worse than no analysis at all. This need for familiarity and a firm understanding of the underlying principles serves as a barrier to entry; as a result, such platforms and systems are likely to disproportionately benefit groups that already have significant data science mastery. To truly realize the vision of "democratizing data science", software and algorithms that aim to make data science "accessible to everyone" should also think about how to reach currently underserved individuals and organizations.

5.2 Data Science Education

If data science technologies are failing to level the playing field, then perhaps better education is the solution. Many recent trends seem to hold promise for democratizing access to data science skills: more people are entering studies computer science, which is useful preparation for data science, and employment opportunities for data scientists seem to be growing. However, although overall enrollment in such fields has increased, female students have become even more disproportionately underrepresented, while ethnic minorities such as African Americans and Hispanics have only made very slight gains in representation [10]. As discussed above, this lack of diversity in education leads to a lack of diversity among data scientists themselves, which in turn hinders the application of data science to problems affecting underrepresented groups.

The recent advent of freely available course material, particularly massively online open courses (MOOCs), has the potential to make great strides in democratizing access to data science. Indeed, two leading MOOC providers, edX and Coursera, offer over a dozen data science-related courses between them. While this increases access to data science education, a large fraction of students enrolled in MOOCs are likely industry professionals already working for established companies, and only a small portion take the course to gain knowledge for a new job [7]. While MOOCs have tremendous potential to reach students across national and socioeconomic barriers [6], we believe this potential has not yet been fully realized: true democratization of access to skills in computer science and data science requires further outreach and educational efforts.

As with many skills, expertise in data science is nearly impossible to obtain without practice and experience. However, most data scientists are PhDs who spend years in homogeneous academic institutions, and practical data science training outside of academia often occurs in places with a substantial data science talent pool, leading to a rich-get-richer effect where companies with large reserves of money and talent are able to train and retain strong data scientists.

5.3 Crowd-based Efforts

Many institutions, such as Netflix and Kaggle, have created open machine learning competitions. Notably, the 2014 KDD Cup, organized jointly by SIGKDD and Kaggle, focuses on predicting projects to profile on the charity website DonorsChoose.org. Such competitions could contribute to our vision of democratizing data science on at least two fronts: they offer a relatively low-barrier entry for aspiring data scientists to contribute to real problems, and they can potentially focus the data science community on meaningful causes.

We note that most crowd-based data science competitions are supported by for-profit entities. For these companies, these competitions can be a low-cost way to improve their products or services. Perhaps most famously, the Netflix prize, in which participants were asked to predict customers' star ratings for movies, benefited the company far more than the cost of its $1 million prize [8]. Nonetheless, we are encouraged by the potential of open data science competitions and the willingness of companies like Kaggle to encourage their users to work on problems of that promote overall social welfare.

5.4 Data Science as Public Service

A recent philanthropic effort, Data Science for Social Good (DSSG), matches summer fellows with public-spirited data science problems that a community organization is facing [1]. Such a program highlights opportunities to improve the effectiveness of these community organizations' valuable work. These philanthropic efforts bring attention and prestige to applying data science to socially important causes. They also allow participants to learn more about the important work that their partner community organizations are doing and may inspire them to continue or inform others about that work. Similar efforts that introduce underrepresented groups of people to data science could produce a sustainable pipeline of people who are fundamentally interested in these issues.

6. SOLUTIONS: PATHS FORWARD

While the efforts described in the previous section are all components of democratizing data science, gaps remain. Many important data science challenges in government and civil society remain unaddressed, and demographic underrepresentation issues persist. The shortcomings of existing efforts suggest several paths forward:

6.1 Business Opportunities

Developers of tools, algorithms and systems that make it easier to analyze data need to think about how to reach organizations with low or non-existent data science capacity. For example, they could consider in-person training programs, pro bono work with community groups, and other initiatives that introduce data science to these groups. Ideally, such efforts would ultimately benefit developers as well, since such groups could become customers for services as they become more adept with data science tools.

We also believe that groups less familiar with data science are an underserved, untapped market. In this era of Big Data, many companies are working on products and services that cater to extremely large datasets and emphasize scalability. Many datasets of interest to important societal

issues, however, may not necessarily be big; instead, thinking about what data to obtain and how to collect it, clean and integrate it into machine-readable forms, and provide value to public-spirited causes are arguably more important challenges. Great opportunities exist to serve this "long tail" of the data science market.

6.2 Education and Outreach Programs

In spite of the growing enrollment of students in computer science and the rise of MOOCs focused on data science, inequalities and barriers to access persist. In computer science, a number of organizations and initiatives seek to address this issue using targeted outreach programs, including Girls Who Code and CODE2040, which focus on introducing women and underrepresented minorities, respectively, to computer science. In addition to teaching technical skills, these programs often provide a safe, comfortable environment for participants to learn from each other, meet role models, and envision themselves as computer scientists. Similar initiatives exist for mathematics, engineering, and technology. Since an interest and familiarity with computer science, programming, the scientific method, and mathematics are prerequisite skills for data science, its democratization stands to benefit from such programs.

Data science presents unique challenges beyond computer science: effective analysis of data often requires a strong background in not only programming, but also mathematics and statistics. While basic programming skills are accessible to students even at the elementary school level, the advanced mathematics and statistics needed for thorough analyses of data are more difficult to fully understand without sufficient background. Despite these challenges, we believe that the right curriculum structure and teaching methods could improve access to valuable data science skillsets.

A related approach is to target domain experts, such as journalists, who are already interested in important issues in society but who currently lack data science skills. Courses or other educational experiences targeted toward interest groups with a stake in promoting social good (such as journalists) could be yet another way to democratize data science skills.

6.3 Data Science Academic Research

Data science is a growing area in academia: across the United States, institutes for data science are being launched, and researchers in computer science, information, journalism, and business schools are focusing on data science and its applications. The rise of computational social science and digital humanities are also arguably the by-product of newly accessible datasets in these fields.

We believe that democratizing data science could be an important, challenging, and rich area of academic research. First, from a practical standpoint, problems related to overall social welfare often rely on freely available public datasets, which, unlike proprietary company data, are accessible to academic researchers. As well, such research is inherently interdisciplinary: research designed to democratize data science methods cuts across different sub-areas in computer science, while applying data science to society's biggest problems will require experts in other disciplines and the actual problem's stakeholders. Perhaps most importantly, focusing on "real-world problems", as others have argued [11], can help ensure that data science research has meaningful societal impact. In short, excellent, early-stage opportunities exist for academic researchers to develop methods and applications of data science that help solve some of society's most important problems.

7. CONCLUSION

In this paper, we asked, what *should* "democratizing data science" mean, and how can this vision be realized? In pursuit of a better definition, we argue that democratization should ultimately be about ensuring that we realize the potential of data science to solve problems that promote social good. This is the answer that is most vital in the era of Big Data, and new approaches are needed to level the playing field in terms of demographics, in addition to initiatives that incentivize data scientists to focus on solving problems with greatest impact. Without these changes, democratic data science cannot be realized. The purpose of our writing is to raise awareness of the caveats in a rapidly growing and promising field. While we have proposed several paths forward, we recognize that we do not have all the answers, and look to the data science community to find and propose solutions going forward. We believe that the democratization of data science, in the end, will arise from a combination of private, public, and community efforts, led by the individual choices that we make as data scientists.

8. REFERENCES

[1] Data Science for Social Good. `http://www.dssg.io/`. Accessed: 2014-07-22.

[2] FiveThirtyEight. `http://fivethirtyeight.com/`. Accessed: 2014-07-24.

[3] McKinsey Global Institute — Big Data: The next frontier of innovation, competition, and productivity. `http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation`. Accessed: 2014-07-24.

[4] The Upshot - New York Times. `http://www.nytimes.com/upshot/`. Accessed: 2014-07-24.

[5] d. boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, 15(5):662–679, 2012.

[6] D. Koller and A. Ng. Log on and learn: The promise of access in online education, 2012.

[7] S. Kolowich. Who takes MOOCs. *Inside Higher Ed*, 5:2012, 2012.

[8] S. Lohr. A $1 Million Research Bargain for Netflix, and Maybe a Model for Others. `http://www.nytimes.com/2009/09/22/technology/internet/22netflix.html`. Accessed: 2014-07-22.

[9] L. Manovich. Trending: the promises and the challenges of big social data. *Debates in the Digital Humanities*, 2011.

[10] N. S. B. (US). *Science & engineering indicators*. National Science Board, 2012.

[11] K. Wagstaff. Machine learning that matters. In *ICML*, 2012.

[12] H. Yu and D. G. Robinson. The new ambiguity of "open government". *UCLA Law Review*, 59:178–230, 2012.