THIS MANUSCRIPT IS A PRE-PRINT OF: Hill, B. M., Dailey, D., Guy, R. T., Lewis, B., Matsuzaki, M., & Morgan, J. T. (2017). Democratizing Data Science: The Community Data Science Workshops and Classes. In N. Jullien, S. A. Matei, & S. P. Goggins (Eds.), *Big Data Factories: Scientific Collaborative approaches for virtual community data collection, repurposing, recombining, and dissemination.* New York, New York: Springer Nature.
https://doi.org/10.1007/978-3-319-59186-5_9

Democratizing Data Science: The Community Data Science Workshops and Classes

Benjamin Mako Hill, Dharma Dailey[†], Richard T. Guy[†], Ben Lewis[†], Mika Matsuzaki[†], Jonathan T. Morgan[†]

Nearly every published discussion of data science education begins with a reflection on an acute shortage in labor markets of professional data scientists with the skills necessary to extract business value from burgeoning datasets created by online communities like Facebook, Twitter, and LinkedIn. This model of data science—professional data scientists mining online communities for the benefit of their employers—is only one possible vision for the future of the field. What if everybody learned the basic tools of data science? What if the users of online communities—instead of being ignored completely or relegated to the passive roles of data producers to be shaped and nudged—collected and analyzed data about themselves? What if, instead, they used data to understand themselves and communicate with each other? What if data science was treated not as a highly specialized set of skills but as a basic literacy in an increasingly data-driven world?

In this chapter, we describe three years of work and experimentation around a vision of *community data science* that attempts to explore one set of answers to these "what if?" questions. This work has primarily involved designing curriculum for, and then running, five series of 4-day workshops, plus three

[†] Authors contributed equally to this work.

traditional university courses taught to masters students. We have used these workshops and classes to explore the potential of, and challenges around, this vision of democratized data science. We begin by framing our goals and approach in terms of similar and analogous efforts. Next, we present our philosophy and design goals. With this background, we describe the structure of the curriculum we have developed. Finally, we use data from several pre-session, within-session, and post-session surveys to discuss some of the promises and limitations of our approach.

Background

Data Science Education

There is little doubt that, driven by surging interest in the power and potential of "big data" for business, data scientists have found themselves in high demand (Manyika et al., 2011). *Harvard Business Review* has called "data scientist" the "sexiest job in the 21st century" (Davenport & Patil, 2012) and several reports have pointed to massive shortages of data scientists in labor markets. For example, in their widely cited report published by the McKinsey Global Institute, Manyika et al. (2011) suggested that the United States is already facing a massive shortfall of skilled data scientists that will only be aggravated in the coming years. In 2014, Dwoskin (2014) suggested that there were 36,000 advertised, but unfilled, positions for data scientists in more than 6,000 firms.

In response, a whole series of education programs have been created, or rebranded, in what West and Portenoy (2016) have described as a "data science gold rush in higher education." Using a dataset of more than 100 degree-granting programs in related spaces collected by North Carolina State University's Institute for Advanced Analytics,[1] West and Portenoy point to dozens of new programs created in a matter of years in the United States alone.

Although there is no consensus—either in popular accounts or among data scientists educators—on exactly what such programs should cover (Davenport & Patil, 2012; Gellman, 2014; Miller, 2013), there is some agreement that data scientists should be able to collect and integrate datasets and conduct analyses using some combination of programming, statistical modeling, and data mining techniques. Similarly, there is consensus that a critical skill for professional data scientists is the ability to ask and answer to questions of substantive in-

[1] http://analytics.ncsu.edu/?page_id=4184 (https://perma.cc/6MKH-7KVY)

terest and to be able to clearly communicate their results (Davenport & Patil, 2012).

End User and Conversational Programmers

Although not all descriptions of data science involve social media, many of the most widely cited accounts of the rise of data science focus on the massive growth of datasets of online behavior from sites like Facebook, LinkedIn, Google, and Etsy (Dwoskin, 2014; Manyika et al., 2011). The absence of any mention of users of these websites from these discussions of data science is striking. Although left largely implicit, the role of end users in these accounts is to produce data and, ultimately, have their behavior shaped by the output of algorithms. Of course, as evidenced by the quantified self movement (Nafus, 2016; Neff & Nafus, 2016; Wolf, 2010), at least some users of these systems are likely interested in the data created and stored by these systems.

Data analysis is often pointed to as a classic example of *end user programming*—commonly defined as the authoring of code by non-professional programmers (Jones, 1995; Nardi, 1993). Intriguingly, as data science has grown into an established professional practice itself, the potential emerges for *end user data science*. Through web application programming interfaces (APIs) created to facilitate user access to personal data from online communities, the infrastructure already exists to provide users with structured data about themselves and their friends from many of the most widely used social computing systems. That said, this access is almost only ever taken advantage of through apps with preset interfaces and dashboards. What remains missing is widespread access to the knowledge and skills to facilitate *end user data science* using currently available data.

Recent research has suggested that learning to program can be understood as a valuable tool even among users who never engage in programming. A study by Chilana et al. (2015) showed that students from majors like management with no intention to engage in programming of any sort expressed a strong interest in learning to program so that they could speak effectively with programmers they might work with. In a follow-up survey of non-programmers in a large multinational software company, Chilana, Singh, and Guo (2016) found that nearly half of their respondents (42.6%) had invested time in learning to program and that over half of these individuals were what they called “conversational programmers” who were interested only in improving technical conversations and their own marketability. To the extent that it is increasingly common for non-professional data scientists to encounter data scientific analyses, being exposed to the basic tools of data science may be seen as useful for these conversational data scientists with no intent to engage in analysis themselves.

To extend the metaphor to programming one final time, it is worth considering how over the last several decades, computer science educators have explored what curriculum might best serve the goals of teaching non-professional programmers. To cite one example, Mark Guzdial and Andrea Forte have published a series of papers that reported on, in various ways, a attempt to develop, deploy, and evaluate curriculum teaching programming to non-computer science majors (Forte & Guzdial, 2005; Guzdial, 2003; Guzdial & Forte, 2005). The degree to which this type of curriculum might differ from attempts to teach conversational programmers has been described as an open issue by Chilana et al. (2016). We know of no attempts to develop curriculum or explore pedagogical approaches around end user and conversational data science.

Democratizing Data Science

To the extent that data science is powerful and provides its practitioners with the ability to understand and affect behavior, it can be understood as politically important to make access to these tools more widespread. Although statistics are much less solid than they are in more established fields, there is evidence that data scientists are overwhelmingly white and overwhelmingly male. Though women, minorities, people with disabilities, and veterans are underrepresented in STEM fields generally, they remain most underrepresented in the fields that data science draws upon most strongly: computer science, math, and statistics.[2]

One important approach to reducing inequality in participation used in feminist critiques of computer science is to attempt to remove systematic barriers to participation. Margolis and Fisher (2001) famously use the metaphor of unlocking clubhouses to describe the goal of breaking down these systematic barriers to interested women in computing communities. A second approach involves designing new forms of participation that appeal to wider audiences. Buechley and Hill (2010) use the metaphor of building new clubhouses to evoke the idea that computing can be reimagined to appeal to women uninterested by computing as it is typically framed. Buechley and Hill argue

[2] http://www.nsf.gov/statistics/2015/nsf15311/digest/nsf15311-digest.pdf (https://perma.cc/74E5-T4YJ)

that this approach can broaden participation in computing. Although there are almost certainly many systematic barriers to participation in data science that affect members of underrepresented groups, imagining data science as practiced by the large majority of people uninterested in careers as professional data scientists is the first step on the path of "democratizing" data science in the ways suggested by Buechley and Hill (2010).

There have been a series of efforts to involve users of online communities in data science. The most famous and common techniques are citizen science projects. The citizen science model, made famous by Galaxy Zoo (Raddick et al., 2007), Zooniverse (Simpson, Page, & De Roure, 2014; Smith, Lynn, & Lintott, 2013), and eBird (Sullivan et al., 2009; Wood, Sullivan, Iliff, Fink, & Kelling, 2011) is similar to "crowdsourcing" where participants role is active and intentional but also limited to a handful of typically low-level and repetitive tasks. In citizen science, participants act as sources of distributed labor and human computation (Howe, 2006). Like crowdsourcing, task execution is distributed but the tasks of posing questions and performing analyses remain the exclusive domain of the platform operators and the "real" scientists (Benkler, 2016).

A smaller body of work has explored the potential of involving online communities in participatory data analysis where both task selection and execution are distributed. There are a number of attempts to support data analysis through participatory and social data visualization on the web (e.g., Heer, Viégas, & Wattenberg, 2007; Luther, Counts, Stecher, Hoff, & Johns, 2009; Viegas, Wattenberg, van Ham, Kriss, & McKeon, 2007; Wattenberg & Kriss, 2006). Although powerful, these systems are often restricted to particular datasets provided by researchers or to a set of predefined types of visualizations or analyses. For example, users of these systems are often unable to create new variables in ways that are a basic part of most data scientists' work. Another interesting approach occurred on the Reddit online community through an experimental research process used by Matias (2016). In his study of a large social mobilization in Reddit, Matias discussed initial results and worked with participants to refine models and hypotheses. Although users were deeply involved in the process of hypothesis construction, they still relied on an academic researcher with access to programming and statistical knowledge and skills to carry out tests. Both social visualization systems and Matias's work are limited by their desire to involve users without also asking them to learn new technical skills.

Perhaps the most clear attempt to democratize data science in the way we have articulated is a system by Sayamindu Dasgupta (Dasgupta, 2016; Dasgupta & Hill, 2016, 2017). Deployed in the Scratch programming community (Resnick et al., 2009), Dasgupta's system provides programmatic access to data about activity in the Scratch community to each member. Dasgupta documented the way that Scratch's young users used the system to enthusiastically analyze their own data in ways that were powerful, unanticipated, and empowering. Dasgupta's system is limited both in the analytic tools it makes available and in the depth and scope of data provided. That said, the level of enthusiasm shown by users of the system, and the creativity these users displayed, is deeply inspiring. Like Dasgupta, our goal is to move one step beyond both citizen science and participatory hypothesis testing to give users of online communities the ability to ask and answer their own questions (*end user data science*) and to build the skills to engage with other analysts and analyses (*conversational data science*).

Toward that end, we designed a series of workshops and courses. In designing, teaching, and evaluating this curriculum, we were motivated by three broad questions. First, what are the essential skills for end user and conversational data scientists? Second, what would a curriculum teaching these skills involve? Finally, how would one evaluate attempts to democratize data science? We describe the work we have done in our workshops to explore potential answers to these questions over the rest of this essay.

Philosophy and Pedagogy

The philosophy informing our pedagogical approach is primarily influenced by Margolis and Fisher's (2001) seminal work on breaking down barriers to the participation of women in computing, Lave and Wenger's (1991) theory of legitimate peripheral participation, and Papert's (1980) concept of constructionism. From Margolis and Fisher, we draw a commitment to broadening participation in data science. From Lave and Wenger, we draw a commitment to the idea of authentic learning environments and the ability to learn through apprenticeship-like relationships. From Papert, we draw the idea that knowledge can be constructed through the creation and manipulation of knowledge in a social environment.

Broadening participation

The first pillar of our community data science approach is the goal of broadening participation. We seek to broaden participation along several dimensions including not only the kinds of academic fields or professional backgrounds of participants but also demographic characteristics including gender and race. Many other approaches to teaching data science require existing programming or statistical experience. For example, the Software Carpentry and Data Carpentry workshops seek to attract participants with undergraduate-level programming experience (Wilson, 2014). We target absolute beginners. Indeed, one central criterion for making acceptance decisions for our workshops and classes is that applicants have no previous programming experience. This has an additional benefit of ensuring that participants begin with a similar skill level.

Meaningful participation in STEM requires successful negotiation of cultural, social, and symbolic elements of STEM fields (Joshi, Kvasny, Unnikrishnan, & Trauth, 2016). Therefore we strive to create an inclusive environment that considers several factors known to influence inclusiveness in STEM. For example, signifiers of masculine tech culture such as Star Trek posters have been shown to inhibit participation by women. Conversely, more neutral "ambient" signifiers such as nature posters do not inhibit anyone's participation (Cheryan, Plaut, Davies, & Steele, 2009). Toward this end, we have intentionally hosted all of our workshops and classes outside of the engineering buildings at the University of Washington campus. We have made attempts to recruit and encourage women and people of color to act as mentors, lead sessions, and give lectures. Inclusiveness is also influenced by the kinds of examples one uses and our curriculum emphasizes working with data about people.

Finally, we have sought to offer our workshops at times, and at a cost, that makes participation by diverse groups of people possible. For example, we have scheduled our workshops on evenings and on Saturdays to make it possible for participants with full-time jobs to attend. So far, we have been able to offer all of our workshops at no cost to participants. Similarly, we have built our curriculum entirely around tools, APIs, and datasets that can be installed and used for free.

Project-based construction

A second pillar of our approach is a strong emphasis on project-based construction and authenticity. Although we do not entirely eschew more traditional lecture-based pedagogy, the bulk of our workshops and classes involves participants' programming on their own computers. Even during lectures, all participants are encouraged to program using their own computers by repeating the programming constructs being demonstrated by instructors and modifying them in ways that interest them.

The decision to have individuals program on their own computers reflects a strong commitment to creating authentic experiences (Lave & Wenger, 1991). We strongly believe that participants in our workshops and classes should program using the tools that we use in our own work as end user and conversational data scientists. When we teach individuals to use APIs, we have them create API keys and engage directly with real APIs. Although this leads to challenges and unpredictability around setup related to heterogeneity of participants' devices, it also turns data science into something that happens directly on each participant's computer. When the workshops end, participants leave with all the software necessary to continue engaging in data science.

We ensure that less than half of any session is dedicated to more traditional lecture-based teaching. Instead, participants spend the majority of their time in the sessions writing software and analyzing data. We encourage participants to program and analyze data the way we do—by modifying existing code and by searching sites like Stack Overflow for error messages, recipes, and solutions to problems. This approach encourages people to wrestle with many of the real issues brought up by data analysis in ways that make critical engagement a central part of the process (Ratto, 2011). For example, when we teach about APIs, participants deal with questions about the degree to which APIs are owned or controlled by companies.

Learning Communities

A final pillar of our approach is the idea that learning happens through collaborative construction of knowledge in convivial social environments. In ways that are inspired by both Lave and Wenger's (1991) apprenticeships and Papert's (1980) samba schools, we attempt to maximize one-on-one interactions between beginners and more skilled data scientists. Concretely, this involves recruiting a large number of skilled data scientists to serve as "mentors." We try to keep to a four-to-one student-to-mentor ratio. Over the course of running the workshop series five times, we have observed that the mentors who are most reliably effective at helping learners solve their problems often come from non-traditional engineering backgrounds. Most encouragingly, we have found that many of the most effective mentors were originally introduced to data science through previous iterations of the workshops and classes.

Excellent mentors embody a warm environment by helping participants solve the problems they are facing in ways they will be able to replicate and build upon when they are working on their own rather than trying to teach "their way" or the "right way" to do something. A low student-to-mentor ratio enables opportunities for extensive one-on-one coaching. This is especially helpful for beginners since their ability to troubleshoot a problem can be brittle and because troubleshooting can be stressful and frustrating (Estrada & Atwood, 2012).

A sense of belonging is another factor that has been demonstrated to influence the inclusiveness of STEM participation (e.g., Good, Rattan, & Dweck, 2012). Providing lunch—the workshops biggest expense by far—is a time-honored way to foster informal interactions and an important component of how we help to foster social support for participants. During lunch, participants often debrief with each other over the morning workshop while getting to know each other and mentors. For these reasons, we also encourage and support meet-ups and learning sessions outside of the formal workshops and classes.

Community Data Science Workshops

In early 2014, we designed a set of 4-day workshops in Seattle, Washington that aimed to answer the three questions we raised in our background section while attempting to adhere to the philosophy and pedagogy laid out above. For the initial set of workshops, we drew both inspiration and some initial curriculum from the Boston Python Workshops (BPW)[3] and Software Carpentry[4]—two curricula with which we had experience. In particular, we leveraged BPW's detailed Python setup instructions and introductory Python programming curriculum. Additionally, the way we structure our daily schedule and

[3] http://bostonpythonworkshop.com/ (https://perma.cc/5Y36-R9FM)

[4] See Wilson (2014) and http://software-carpentry.org/ (https://perma.cc/23SE-BPHA)

our project-based afternoon sessions was drawn directly from BPW. Although several of us teach at the University of Washington, we sought to arrange these workshops as volunteers outside of a formal classroom setting.

The initial workshops were an enormous success with 115 applicants of whom we were able to admit 52. In response to this demand, we ran the workshops again in late 2014, twice again in 2015, and once in early 2016. Additional workshops are planned in Seattle, twice a year, going forward. As we have been able to recruit more mentors, each workshop has been larger than the previous iteration. Our most recent workshop in early 2016 was attended by 97 participants.

Each time we have run them, the workshops were organized over one Friday evening and three Saturdays. A Friday session before the initial Saturday session ensured all participants (and their computers) were prepared for the following morning. The four sessions were numbered from 0 to 3 in reference to about zero-indexing in the Python programming language. We collected feedback from participants after each day and debriefed instructors after each session and again after each series of workshops has concluded. Based on this process, we iterated on the curriculum and design of the workshops each time we ran them.

Each Saturday session begins with a 2-hour interactive lecture in the morning that builds upon the topics presented in previous sessions. Lectures introduce new concepts and show real examples of carrying out tasks through "live coding." A picture of a lecture is shown in the bottom right panel of Figure 1. We encourage participants to participate in the lecture by actively programming on their own computers. The concepts discussed in each lecture introduce participants to a handful of tools and concepts that are then explored in the afternoon challenges. Each afternoon session is organized around open-ended questions designed to foster structured exploration of the morning's concepts to help participants synthesize and use their new skills.

Afternoon sessions involve independent project work. Participants are given an archive of several simple programs written using only concepts that participants were introduced to in the lectures. After a short exposition and explanation of the sample programs by a session leader, participants are encouraged to modify, build upon, or be inspired by these programs to solve problems of their choosing. Participants work on projects individually, or in groups, with help from more experienced mentors present. This independent project work continues over 3-4 hours. We have experimented with many dif-

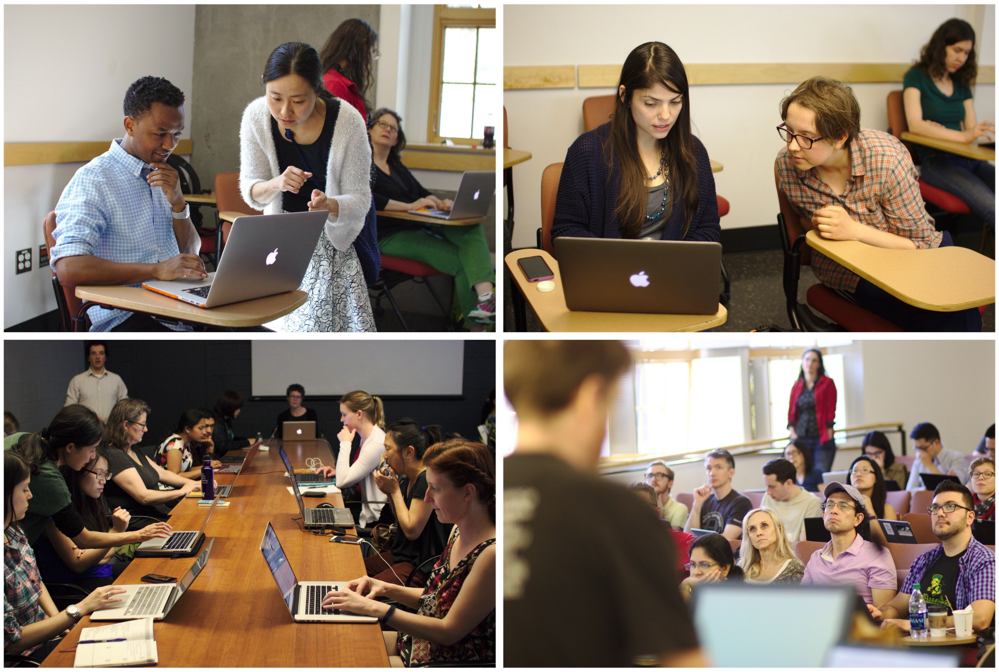

Figure 1: Four photographs from the Community Data Science Workshops held in April and May 2016. The top two panels show mentors working one on one with participants. The bottom left panel shows a breakout afternoon workshop with participants working independently on projects. The bottom right panel shows participants during a morning lecture with mentors standing to the side and ready to help participants when they require assistance.

ferent projects. In general, we have offered participants two or three choices during each afternoon so that participants can choose projects that align with their interests. All but the bottom right panel in Figure 1 show these project-based sessions. The top two panels both show mentors working one on one with participants. All of our curriculum—including sample projects, code, and recordings of lectures, are made available on our website.[5]

Day 0: Setup

In the first Friday session, participants walk through a checklist for installing Python and installing a programmer's text editor. Next, they work through a brief tutorial on the basics of using the command line. After the participants have completed these setup tasks, they are encouraged to work through some simple Python programming exercises. This makes the next morning

[5] http://wiki.communitydata.cc/CDSW (https://perma.cc/G36T-KLG8)

lecture easier by pre-introducing material covered in the Saturday lecture. The evening session is completely self-guided and allows participants to warm up to the concepts presented at their own pace. Mentors are on hand to provide technical assistance, help participants through difficult programming concepts, and verify that each student has completed session goals before they leave.

Day 1: Introduction to Programming

The first Saturday session starts with a reinforcement of how to work in the command line, and then introduces variables, Python's built-in data types including integers, floating point numbers, strings, lists, and dictionaries. As a result, after only one lecture, participants are familiar with all of Python's first-order data structures and all of the data types used in the rest of the workshops. Finally, we introduce conditional logic and loops. As in all of our lectures, we do not use slides. Instead, we demonstrate and discuss concepts while programming example code in an interactive Python interpreter using an iterative trial-and-error method. For example, we demonstrate strings by constructing messages from strings and demonstrate dictionaries by mapping names to ages ("`{Mako: 33, Ben: 24}`"). Throughout the lecture, mentors are distributed throughout the room to be able to answer participants' questions about issues they are having in their code.

The first afternoon project session aims to support participants in engaging in simple data analysis using Python. For example, one session we have designed begins by downloading an archive that includes both code and a dataset drawn from the US Social Security Administration on the popularity of different baby names among US children. Projects like this allow participants to start analyzing real-world data to ask and answer questions of their own design almost immediately. For example, participants often begin by answering a question like, "How many times does *your* name show up in the dataset?" and proceed to more complicated questions (e.g., "Which names are strict subsets of other names?"). These data reveals common challenges in data analysis immediately. For example, the exclusion of names given to fewer than five people of one gender leads directly to insights about missing data, while the binary nature of gender in the dataset leads to insights about how data collection decisions can support or suppress specific conclusions.

Day 2: Web APIs

For the second session, we step back from Python to spend time working with web APIs—web services that allow a program to acquire data from online communities and social media sources. One API we rely upon in the lecture is the PlaceKitten API, which takes a request for an image of a specified size and then returns an image of a kitten of that size. Participants are first shown how to make API requests through a web browser. We then show them how to make the same requests in Python.

Next, we demonstrate how to parse more complex API responses. We have often relied on data drawn from Wikipedia about articles related to Harry Potter as an example because there is a very large amount of data, and it exhibits interesting patterns (e.g., bursts of edits around the release of each film and book). Afternoon sessions on the second full day involve working through and modifying simple programs that pull data from Twitter's API to build a tweet-gathering tool for use in the third session, from the Yelp API to find out about local restaurants, and from the Wikipedia API to answer questions about editing activity and article metadata.

Day 3: Data cleanup and analysis

The final session acts as a capstone highlighting the process of sourcing, cleaning, and using a dataset to ask and answer a question. In the morning lecture, we walk through a program that collects a dataset about every contribution to articles in Wikipedia related to Harry Potter using the Wikipedia API. Using these data, we generate a series of time series plots to answer several questions related to the way that Wikipedia editing on Harry Potter topics has changed over time.

The afternoon projects for this session focus on the process of data analysis and visualization. For example, we have used a pre-collected set of tweets about earthquakes (collected using code that was crafted, in part, by participants during an afternoon session on the second day) to generate time series in different resolutions and identify earthquakes around the world as they appear in the dataset. Other sessions have focused on gathering geocoded social media data and visualizing these data on a map. By showing participants different ways of interacting with datasets that they have gathered, we are able to contextualize the act of analyzing data and to provide examples of the process of analyzing social media data from start to finish.

Community Data Science Classes

In response to requests from our university, three of us have developed and taught quarter-length, for-credit, masters level courses based on the Community Data Science Workshops. The classes were taught at two different departments at the University of Washington: three times in the Department of Communication in our Communication Leadership program, and once in the Department of Human Centered Design and Engineering. The courses directly incorporate most of the workshop curriculum described above. Unlike most other data science curricula, these classes' central focus is an extended, self-directed project which forms most of each student's grade. Curriculum for these classes are made fully available on our website.[6] Courses were taught to groups of 20 and 30 students with 1 instructor and 1 teaching assistant.

Teaching this material over 10 weeks, instead of 4 days, provided us with more opportunities to iterate on our lesson plans. The practice of sending out anonymous feedback surveys after each class session, carried over from the workshops, helped us adjust the pace and teaching style between sessions. However, other than the addition of more examples of APIs (essentially, the ability to teach more than one of the afternoon session from Day 2), we found that the additional time did not allow us to increase the scope of the material presented. We were challenged to address all core programming concepts thoroughly within the first few weeks of the course so that students would feel confident deploying those concepts in their own work while leaving them with sufficient time to select a dataset, to frame a research question, and to gather, analyze, and report their findings. The nature of the course work changed dramatically at roughly the halfway point: the first half of the quarter provided a crash course in data science programming; the second half focused on supporting students as they applied those lessons to specific datasets and research problems. Students with no previous programming experience needed to absorb a great deal of new knowledge within the first few weeks in order to successfully complete their class project.

The introduction of grades substantially raised the stakes of mastering the material and it risked conflict with our "low stakes" approach in the workshops. Homework assignments were graded on effort, not code quality. Each course culminated in a final project where success depended more on gathering and synthesizing data to tell a story than on the quality of the code written

[6] e.g., https://wiki.communitydata.cc/Community_Data_Science_Course_(Spring_2015) (https://perma.cc/T3S9-ZZRN)

along the way. As an example, a student would receive full credit for an inefficient program or a program with a few missing edge cases but would lose credit for failing to identify a potential source of error like incomplete data. In one rendition of the class, data visualization was worth 25% of the project grade. Points were awarded if a plot represented the data correctly by using sensible color schemes and axes, not based on the students' choice or mastery of plotting technology (Excel was most commonly used).

Instructors teaching the courses did not always experience the same challenges. One course instructor felt that the move to a traditional classroom setting, which meant dramatically increasing the ratio of students to available mentors, reduced opportunities for *ad hoc*, one-on-one support. He attempted to compensate for this by building opportunities for peer support into the class and by grouping students with little or no previous programming experience with others who had some familiarity with programming in other contexts and languages. Another instructor found that the shift to more hours in class meant he could spend more time on average with each student.

There was consensus that while it was not possible to cover substantially more material in 10 weeks than in 3 weekends, it was possible to cover it more thoroughly. The higher student-to-mentor ratio made it more difficult to support struggling students, but the addition of assignments, feedback surveys, a more drawn out schedule, and self-directed projects helped assure that students had the opportunity to master the material. Students were also exposed to some new challenges, chiefly the challenge of finding data relevant to their subject of interest.

Outcomes

As we have developed the workshops and classes, we have devoted time to a discussion of our own goals. Although the organizers share a goal of "democratizing data science," this is an amorphous goal understood differently even within the team that developed the curricula. Through discussion, we established that there were several dimensions on which we feel our efforts should be evaluated. First, we believe that our approach should be evaluated in terms of its ability to support *skill development* among participants. In this first sense, we consider our approach effective only if participants are building skills associated with end user or conversational data science.

Second, given our goals of democratization, we believe that it is important that the curriculum be a successful form of *outreach* in that it should attract large numbers of individuals, especially from groups that are underrepresented in more traditional data science communities. Third and finally, we believe a success criterion for our approach is its ability to support *empowerment*. In this final sense, we believe that it is not enough that learners simply have skills but that they feel able to build on these skills in ways that shift power.

Skill Development

The informal nature of our workshops makes it difficult to systematically ascertain the degree to which participants have learned skills. Some evidence of skill development comes from the opt-in surveys we have run after our sessions. In one typical response to an open-ended question about outcomes, a participant explained that the sessions helped build skills around programming and data analysis:

> It helped me become more comfortable with reading and writing code and taught me how to think more about how to use social media data to answer questions that are not necessarily academic. It also made me more confident to take the lead as the person responsible for writing code in a class project.

Although it is certainly the case that not every participant felt comfortable writing code at the end of the four sessions, many explained that they felt more comfortable in a role of end user or conversational data scientists. For example, one explained that:

> Before the workshop I had no idea what Python can do, what API is for, or what data visualization is. The workshop basically was my entry point to the world of data analysis.

Another participant's feedback is an example of someone who became a more effective and confident conversational data scientist through their experience in the workshops:

> In my work as a librarian where I help clients navigate various sources of information, I feel more comfortable talking about how they can use programming to find or analyze the data they have access to.

In the classes where students each worked on projects over several weeks,

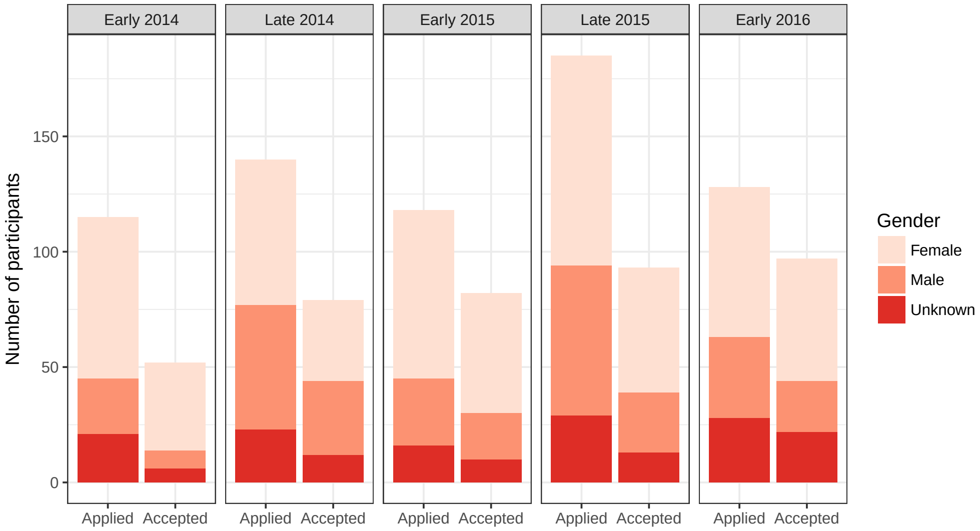

Figure 2: Numbers of admitted participants at each workshop by inferred gender.

more concrete evidence of skill development included the products they were able to create at the end of the class. For example, one student published a detailed report that attempted to understand the relationship between the release of television shows on Netflix and activity on associated Wikipedia articles. The student collected and compared a dataset of Wikipedia editing activity on articles associated with television shows released on Netflix with a similar dataset about broadcast television shows. Using these data, she provided evidence of a strong correlation between episode release dates and editing activity on Wikipedia.[7] There was also evidence of skill development among the academics who attended the workshops. At least one participant emailed us to say that they used skills developed in the class to collect and analyze data from the Twitter API that ultimately led to a published paper.

Outreach

The workshops have consistently attracted a large number of participants. Over the 5 series, 686 people applied to the workshops in Seattle, and 403 were accepted (see Figure 2). In each case, we were constrained by the size of the instructional spaces we had access to and the number of mentors we had

[7] The student, Nyssa Achtyes, published her analysis on a website titled *Long Term User Engagement of Netflix and Non-Netflix shows*: https://nyssadatascience.wordpress.com/ (https://perma.cc/Z9HK-ZVA3)

been able to recruit. Our curriculum has been adapted and taught outside of Seattle as well. For example, a group at the University of Waterloo's Women in Computer Science group has taught a series of workshops that relies heavily on our curriculum.

One of the most striking aspects of our workshops, so far, has been that our participants seem to come from more diverse backgrounds than in typical data science communities. For example, in every workshop and class, participants have been mostly women. This surprised us since we did not make targeted efforts to include (or exclude) a particular gender. To quantify the gender of participants, we analyzed the first names of the participants using the US Census and Social Security data to assign a probable gender to each name. Results are shown in Figure 2 that show that a majority of both applicants and participants were female for each of the five sessions. There was also a fairly high proportion of women among our mentors—especially in later sessions when most mentors were returning participants.

We saw diversity along other dimensions as well. Because we targeted programming neophytes, a large portion of our attendees came from traditionally less technical departments within our university and from outside the university as well. For example, we attracted participants working for both local government and a large number of local non-profits. The workshops were also attended by social media users including bloggers and participants in Wikipedia who were interested in building the skills to analyze data from their own communities.

Empowerment

Perhaps the most important—but difficult to measure—determination of whether our curricula have contributed to the democratization of data science is the degree to which participants felt empowered afterward. Although skill development might include the ability to understand or conduct data analysis, we feel that empowerment goes one step further and suggest that skills can affect and change the power structure in which participants find themselves—at least in relation to data and data analysis. Although empowerment is difficult to measure, opt-in post-workshop surveys of participants suggested that at least some participants felt that exposure to data science was empowering. For example, one former student told us:

> It [ultimately] gave me the confidence to accept a job teaching CS

> at a local CC, which led to me applying to the CS PhD program at [the University of Washington] (and getting in!). So, I guess it contributed to completely changing my life.

Another student reported a similar sense in which the program had led to a shift from a career in administration to one in software engineering:

> Well, I went to Hackbright Academy largely because its curriculum centers on Python. And now I'm a software engineer in San Francisco. So... pretty rad, huh?

One thing we encourage participants to do is to return to future workshops as mentors. Many participants, including two of the current organizers, have returned to become new mentors. This is both a good opportunity for the participants to continue engaging in data science and a sign of empowerment. In our most recent workshops, a majority of mentors were former participants.

Participants often did not continue to engage in data science after the workshop when they felt they did not have projects where they could use and improve their knowledge and skills. Participants who continued to engage in data science often had specific projects or pursued resources like Coursera, CodeAcademy, Data Science Dojo, and classes at the University of Washington. In terms of empowerment, assisting participants at this transitional stage—from the workshop to real-world settings—should be considered an integral part of any community data science curriculum and reflects an area we hope to focus on in future curriculum development.

Limitations

We believe that the community data science approach can benefit participants who seek to gain a working knowledge of programming and data science literacy. The first and most fundamental limitation is that we are trying to cover both data literacy and introductory programming simultaneously. Even individuals who are relatively comfortable exploring, aggregating, and describing data using software tools like spreadsheets often struggle to perform familiar, basic data manipulations using Python. Currently, our workshops and courses emphasize programming but it is unclear that we have the right mix. We could certainly defer more programming concepts, or exclude them altogether, in favor of teaching participants how to use widely available software tools that accomplish the same task.

We could also choose to cover additional programming concepts, such as object orientation, that are useful for working with many common data science libraries. Of course, these decisions—to skip over a basic programming concept or to teach participants a non-programming alternative—would impose new constraints on what we can cover within the workshop as well as what participants will be able to accomplish afterward.

Furthermore, it is not yet clear to us what measures of success we should use to evaluate our approach. Participants seek out our workshops for a variety of reasons, arriving with vastly different types of experience. Some have more practical, immediate, opportunities to continue honing skills than others. Ultimately, success for any individual participant might be best evaluated based on that individual's goals and preparation as well as what they did with what they learned afterward than on direct measures of their performance or engagement during the sessions.

Conclusion

In their highly cited critique around the discourse of big data, danah boyd and Kate Crawford argue that limited access to big data analytic tools is creating new digital divides. The world, they suggest, is divided into the "Big Data rich" and the "Big Data poor" (boyd & Crawford, 2012). The issues boyd and Crawford raise about access to data are formidable and substantive. We see the community data model as one of very few attempts to address these issues directly. However, by framing big data equity as simply an access issue, boyd and Crawford may understate the problem. In ways that Dasgupta and Hill (2017) have shown, nonprofessional data scientists do not ask the same questions that professional data scientists ask. Democratized data science is not only a broader distribution of knowledge, skills, and power, it has the potential to support the development of new types of data science.

We believe that what we have developed in our workshops and classes is a proof of concept. That said, we feel confident in our demonstration that there is broad demand for data science skills outside of traditional engineering circles and among groups, like women, that the fields most closely associated with data science have historically struggled to engage. We hope that we have also provided one vision of what a democratized data science curriculum might look like. A more democratized data science is possible—potentially even with broad societal effects. We encourage you to join us in the process of understanding what it might look like, and what it might be able to accomplish.

Acknowledgments

The Community Data Science workshops were made possible through the generous actions of dozens of mentors who volunteered to spend their weekends teaching strangers data science. Without them, nothing we've described here would have been possible. Our work was also supported by the Department of Communication at the University of Washington which provided physical facilities and other resources. Finally, this work was supported by a Data Science Environments project award from the Gordon and Betty Moore Foundation (Award #2013-10-29) and the Alfred P. Sloan Foundation (Award #3835) to the University of Washington eScience Institute. eScience supports data-driven discovery at the University of Washington in many ways and provided financial and other forms of support for the workshops and for this chapter.

References

Benkler, Y. (2016). Peer production and cooperation. In J. M. Bauer & M. Latzer (Eds.), *Handbook on the Economics of the Internet*. Cheltenham, UK: Edward Elgar.

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878

Buechley, L., & Hill, B. M. (2010). LilyPad in the wild: How hardware's long tail is supporting new engineering and design communities. In *Proceedings of the 8th ACM Conference on Designing Interactive Systems (DIS '10)* (pp. 199–207). doi:10.1145/1858171.1858206

Cheryan, S., Plaut, V. C., Davies, P. G., & Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *Journal of Personality and Social Psychology*, *97*(6), 1045–1060. doi:10.1037/a0016239

Chilana, P. K., Alcock, C., Dembla, S., Ho, A., Hurst, A., Armstrong, B., & Guo, P. J. (2015, October). Perceptions of non-CS majors in intro programming: The rise of the conversational programmer. In *Proceedings of the 2015 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* (pp. 251–259). doi:10.1109/VLHCC.2015.7357224

Chilana, P. K., Singh, R., & Guo, P. J. (2016). Understanding conversational programmers: A perspective from the software industry. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 1462–1472). doi:10.1145/2858036.2858323

Dasgupta, S. (2016). *Children as data scientists: Explorations in creating, thinking, and learning* (Ph.D. Dissertation, Massachusetts Institute of Technology, Cambridge, Massachusetts).

Dasgupta, S., & Hill, B. M. (2016). Learning with data: Designing for community introspection and exploration. In *Workshop on Human-Centered Data Science*. Position Paper. Computer supported cooperative work and social computing, San Francisco, California.

Dasgupta, S., & Hill, B. M. (2017). Scratch Community Blocks: Supporting Children as Data Scientists. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. doi:10.1145/3025453.3025847

Davenport, T. H., & Patil, D. J. (2012, October). Data scientist: The sexiest job of the 21st century. *Harvard Business Review*. Retrieved July 11, 2016, from https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Dwoskin, E. (2014, August 8). Big data's high-priests of algorithms; 'Data scientists' meld statistics and software for find lucrative high-tech jobs. *Wall Street Journal (Online): Tech*. Retrieved July 11, 2016, from http://search.proquest.com/newsstand/docview/1552020409/abstract/D70B27FC5DA74D5APQ/1

Estrada, T., & Atwood, S. A. (2012). Factors that affect student frustration level in introductory laboratory experiences. *2012 ASEE Annual Conference & Exposition*. American Society for Engineering Education, 25.629.1–25.629.7. Retrieved from https://peer.asee.org/21386

Forte, A., & Guzdial, M. (2005, May). Motivation and nonmajors in computer science: Identifying discrete audiences for introductory courses. *IEEE Transactions on Education*, *48*(2), 248–253. doi:10.1109/TE.2004.842924

Gellman, L. (2014, November 6). Business education: Big data gets master treatment—some business schools offer one-year analytics programs, catering to shift in students' ambitions. *Wall Street Journal*, B.7. Retrieved July 11, 2016, from http://search.proquest.com/newsstand/docview/1620527411/abstract/B21739238EE74F26PQ/1

Good, C., Rattan, A., & Dweck, C. S. (2012). Why do women opt out? Sense of belonging and women's representation in mathematics. *Journal of Personality and Social Psychology*, *102*(4), 700–717. doi:10.1037/a0026659

Guzdial, M. (2003). A media computation course for non-majors. In *Proceedings of the 8th Annual Conference on Innovation and Technology in Com-*

puter Science Education (ITiCSE '03) (pp. 104–108). doi:10.1145/961511.961542

Guzdial, M., & Forte, A. (2005). Design process for a non-majors computing course. In *Proceedings of the 36th SIGCSE Technical Symposium on Computer Science Education (SIGCSE '05)* (pp. 361–365). doi:10.1145/1047344.1047468

Heer, J., Viégas, F. B., & Wattenberg, M. (2007). Voyagers and voyeurs: Supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)* (pp. 1029–1038). doi:10.1145/1240624.1240781

Howe, J. (2006). The rise of crowdsourcing. *Wired Magazine*, *14*(6), 1–4.

Jones, C. (1995, September). End user programming. *Computer*, *28*(9), 68–70. doi:10.1109/2.410158

Joshi, K. D., Kvasny, L., Unnikrishnan, P., & Trauth, E. (2016, January). How do black men succeed in IT careers? The effects of capital. In *Proceedings of the 2016 49th Hawaii International Conference on System Sciences (HICSS)* (pp. 4729–4738). doi:10.1109/HICSS.2016.586

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

Luther, K., Counts, S., Stecher, K. B., Hoff, A., & Johns, P. (2009). Pathfinder: An online collaboration environment for citizen scientists. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)* (pp. 239–248). doi:10.1145/1518701.1518741

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011, May). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute. Retrieved July 11, 2016, from http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation

Margolis, J., & Fisher, A. (2001). *Unlocking the clubhouse: Women in computing*. Cambridge, Massachusetts: The MIT Press.

Matias, J. N. (2016). Going dark: Social factors in collective action against platform operators in the Reddit blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)* (pp. 1138–1151). doi:10.1145/2858036.2858391

Miller, C. C. (2013, April 14). The numbers of our lives. *New York Times: ED*, ED.18. Retrieved July 11, 2016, from http://search.proquest.com/newsstand/docview/1326574891/abstract/88A4A39B52A94D3BPQ/2

Nafus, D. (Ed.). (2016). *Quantified: Biosensing technologies in everyday life*. Cambridge, Massachusetts: MIT Press.

Nardi, B. A. (1993). *A small matter of programming: Perspectives on end user computing*. MIT Press.

Neff, G., & Nafus, D. (2016). *Self-tracking*. Cambridge, Massachusetts: MIT Press.

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York, New York: Basic Books.

Papert, S. (1987). Computer criticism vs. technocentric thinking. *Educational Researcher*, *16*(1), 22–30. doi:10.2307/1174251. JSTOR: 1174251

Raddick, J., Lintott, C. J., Schawinski, K., Thomas, D., Nichol, R. C., Andreescu, D., … Slosar, A., et al. (2007). Galaxy Zoo: An experiment in public science participation. In *Bulletin of the American Astronomical Society* (Vol. 38, p. 892).

Ratto, M. (2011, July 1). Critical making: Conceptual and material studies in technology and social life. *The Information Society*, *27*(4), 252–260. doi:10.1080/01972243.2011.583819

Resnick, M., Silverman, B., Kafai, Y., Maloney, J., Monroy-Hernández, A., Rusk, N., … Silver, J. (2009, November). Scratch: Programming for all. *Communications of the ACM*, *52*(11), 60. doi:10.1145/1592761.1592779

Simpson, R., Page, K. R., & De Roure, D. (2014). Zooniverse: Observing the world's largest citizen science platform. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14 Companion)* (pp. 1049–1054). doi:10.1145/2567948.2579215

Smith, A. M., Lynn, S., & Lintott, C. J. (2013). An introduction to the Zooniverse. In *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP '2013)*, Palo Alto, California: AAAI Press.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009, October). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. doi:10.1016/j.biocon.2009.05.006

Viegas, F. B., Wattenberg, M., van Ham, F., Kriss, J., & McKeon, M. (2007, November). ManyEyes: A site for visualization at Internet scale. *IEEE Transactions on Visualization and Computer Graphics*, *13*(6), 1121–1128. doi:10.1109/TVCG.2007.70577

Wattenberg, M., & Kriss, J. (2006, July). Designing for social data analysis. *IEEE Transactions on Visualization and Computer Graphics*, *12*(4), 549–557. doi:10.1109/TVCG.2006.65

West, J., & Portenoy, J. (2016, October). The data gold rush in higher education. In C. Sugimoto, H. R. Ekbia, & M. Mattioli (Eds.), *Big Data is Not a Monolith*. Information Policy. Cambridge, Massachusetts: MIT Press.

Wilson, G. (2014, February 19). Software Carpentry: Lessons learned. *F1000Research*. doi:10.12688/f1000research.3-62.v1

Wolf, G. (2010, April 28). The data-driven life. *The New York Times*. Retrieved August 12, 2016, from http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html

Wood, C., Sullivan, B., Iliff, M., Fink, D., & Kelling, S. (2011, December 20). eBird: Engaging birders in science and conservation. *PLOS Biology*, *9*(12), e1001220. doi:10.1371/journal.pbio.1001220

Author Details

Benjamin Mako Hill

Affiliation University of Washington, Department of Communication

Email makohill@uw.edu

Address Box 353740, Seattle, WA, 98195

Biography Benjamin Mako Hill is a data scientist who studies study collective action in online communities and seeks to understand why some attempts at collaborative production—like Wikipedia and Linux—build large volunteer communities while the vast majority never attract even a second contributor. He is an Assistant Professor of Communication at the University of Washington, a Faculty Associate at the Berkman Klein Center for Internet and Society at Harvard University, and a participant in Wikipedia and a number of other peer production communities.

Dharma Dailey

Affiliation University of Washington, Department of Human Centered Design and Engineering

Email ddailey@uw.edu

Biography Dharma Dailey studies how people get information during crises. She attended the first Community Data Science Workshop as a student and put what she learned into her research! She found the workshop so helpful, she stuck around to help organize more of them. She is is a PhD Candidate in Human-Centered Design and Engineering at the University of Washington.

Mika Matsuzaki

Affiliation University of Washington, Department of Biostatistics

Email m0@uw.edu

Biography Mika Matsuzaki is an epidemiologist. She works at the as a Research Scientist at the University of Washington studying risk factors in vulnerable population with HIV/AIDS.

Jonathan T. Morgan

Affiliation Wikimedia Foundation

Email jmo25@uw.edu

Biography Jonathan Morgan is a Senior Design Researcher at the Wikimedia Foundation. He has a PhD from the University of Washington in the Department of Human Centered Design & Engineering.

Richard T. Guy

Affiliation Microsoft

Email richardtguy84@gmail.com

Biography Tommy Guy is a Data Scientist at Microsoft where he works on large scale experimentation. He enjoys teaching programming and data science wherever they will let him.

Ben Lewis

Affiliation Microsoft

Email benjf5@outlook.com

Biography Ben Lewis is a software engineer who advocates for community involvement in decision making, and seeks to expand access to tools for understanding and shaping the world. He is a graduate of McGill University, an occasional contributor to open source projects, and a participant in Wikipedia.