# Building an Urban Data Science Summer Program at the University of Washington eScience Institute

Ariel Rokem Cecilia Aragon Anthony Arendt
Brittany Fiore-Gartland Bryna Hazelton Joseph Hellerstein
Bernease Herman Bill Howe Ed Lazowska
Micaela Parker Valentina Staneva Sarah Stone
Anissa Tanweer Jacob Vanderplas
The University of Washington eScience Institute
Seattle, WA

## ABSTRACT

During the Summer of 2015, the University of Washington eScience Institute ran an interdisciplinary summer internship program focused on urban informatics, civic engagement, and data-intensive social science. Borrowing elements from the successful Data Science for Social Good (DSSG) programs at the University of Chicago and Georgia Tech, and building on our own previous consulting and "incubation" programs for data-intensive projects in physical, life, and social sciences, we brought together teams of students (graduate, undergraduate, and high school), data scientists, project leads and stakeholders from the University of Washington and local NGOs to design, develop, and deploy new solutions to high-impact problems in the Seattle Metro Area.

In this paper, we describe the inaugural offering of the eScience DSSG and reflect on the process of organizing and structuring the program. The DSSG attracted 144 graduate and undergraduate student applicants from over 10 different fields of study. The final DSSG fellow cohort included 16 students accepted from this pool of applicants. In addition, we included six high school students who joined us from a separate program designed to expose young people to research activities and an undergraduate student who had already started working on one of the projects through another summer research program. We solicited project proposals from research professionals across academic, non-profit, and government institutions. Ultimately, 4 projects were chosen out of 11 submitted proposals: two addressing transportation access for people with limited mobility, one identifying factors affecting whether homeless families find permanent housing, and one deriving new metrics of community well-being from social media data and other relevant data sources. All datasets were sourced from Seattle businesses, foundations, and agencies, with the exception of social media. The teams worked in a shared studio space designed in part for this purpose, and participated in tutorials on relevant tools and technologies, such as GitHub, Python, R, Amazon Web Services, and SQL, as well as topical presentations and discussions related to social good and multi-stakeholder collaborations.

We found that striking a balance between training and software "flow time" is essential, and that determining the right balance between structured and unstructured activities is delicate. The diversity in software and disciplinary experience among participants was initially challenging for tutorial organization and scoping projects. A mix of advanced and introductory material meant that some participants were either lost or bored at any given time. But in the end, this diversity actually helped to improve the scope of the projects. For example, GIS experts added mapping components, software engineering experts designed APIs, and domain experts sanity-checked findings. Overall, the enormous interest implied by the number and diversity of the applicants to our program suggests that similar programs could be operated in many other cities. We intend this paper to facilitate reuse and optimization of the key components of our program.

Bloomberg Data for Good Exchange Conference.
28-Sep-2015, New York City, NY, USA.

## Categories and Subject Descriptors

K.3 [**Social and Professional Topics**]: Education

## Keywords

Data Science, Urban informatics

## 1. INTRODUCTION

Data is ubiquitous, and its potential for advancing social good is becoming increasingly clear. However, the government agencies, non-profits, and NGOs working on pressing social issues often do not have the in-house expertise needed to extract knowledge from large, heterogeneous and noisy datasets. Similarly, students across a variety of disciplines are motivated to work on issues with real social impact, but opportunities for training in data science (and data-intensive science) are limited.

One model for solving both problems is to offer mentored, project-based internships involving both explicit training and hands-on technical work to solve a specific problem in cooperation with a stakeholder organization. The internship model serves as the basis for the Data Science for Social

Good (DSSG) program at the University of Chicago.[1] The program, started in 2013, hosts 12-14 projects each year with teams of 3-4 students. Their success [1] encouraged others to follow this model: a similar program exists at the Georgia Institute of Technology[2], currently in its second year. In this paper, we describe the DSSG program that we organized at the University of Washington eScience Institute during the summer of 2015, including its distinguishing features from earlier programs, and report on lessons learned. Our vision is to bring together academic researchers, stakeholders, and professional data scientists to address critical challenges facing cities — poverty, inequality, safety, health, housing, and efficient use of resources.

## 1.1 The eScience Institute

The University of Washington eScience Institute was founded in 2008 with a mission to support data-intensive discovery across all fields. Core to the mission is the recognition that while research computing infrastructure plays an important role, it is the *intellectual infrastructure* — the expertise in software development, data management, machine learning, and visualization — that needs to be broadly developed and broadly applied.

In 2013, the Institute received a five-year, $32.8M grant jointly with University of California, Berkeley and New York University from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation to launch three interconnected *Data Science Environments* within these universities. The goal of these centers is to develop new models for collaboration centered around data-intensive research, and to develop and maintain the software and organizational infrastructure needed to support this research.

With additional support from the University of Washington and from the Washington Research Foundation, the Institute established a Data Science Studio (DSS) on the UW campus (Figure 1.1). This space provides the physical infrastructure to facilitate collaboration between researchers in different fields surrounding data science, for example by hosting seminars and workshops on data science topics and providing an open space for joint work. It also hosts office hours for students and researchers to seek help from eScience data scientists and research scientists, along with representatives from partner organizations within the university (the University Libraries, UW-IT, the statistical consulting service of UW's Center for Statistics in the Social Sciences) and outside organizations (e.g. Amazon Web Services, Google Cloud Computing). Other opportunities for collaboration are facilitated through the eScience *incubator* program, on which the DSSG program was based. The goal of the incubator is to enable innovative data-intensive scientific research by bringing together data scientists and domain scientists to work on focused, intensive, collaborative projects together in the Studio. In previous iterations of the incubator, projects frequently involved a component of software engineering, cloud and cluster computing, statistics and machine learning, and visualization. The quarter-long projects not only facilitated new research, but also provided training on basic data science techniques. The tools and knowledge returned with the researchers back to their labs and facilitated discovery in these labs and beyond. In some cases, incubator projects led to longer-term interdisciplinary collaborations.

## 1.2 From an Incubator to the Urban DSSG

The DSSG program follows the pattern established in the eScience "incubator" projects. Based on findings from previous iterations of the incubator, an effort was made in this iteration to find projects that were thematically coherent, as a way to encourage interactions between participants. Specifically, the topic chosen was *urban analytics*, with a focus on the Seattle metropolitan area. The choice of urban issues was inspired by the recent inauguration of a University-wide initiative called Urban@UW,[3] and our own experience witnessing an explosion of interest in data-intensive social sciences within an urban context. Urban data science also served as a fertile thematic ground for applications of data science to advancing social good, due to the direct link between research, policy, and social impact. Projects were solicited not only from University departments, but also from project leads in the surrounding community, such as government and non-governmental organizations that analyze urban data with a focus on advancing social good. Finally, in contrast to the previous incubator projects where the team consisted of one eScience lead and one project lead, we solicited applications for full-time student DSSG fellows to form larger teams and encouraged the participation of outside stakeholders.

# 2. THE STUDENTS

## 2.1 DSSG fellows

A solicitation for applications for the DSSG fellowship was announced in the middle of March, and applications were accepted until March 31st. Students were offered a stipend of $6500 for the duration of the program, consistent with other summer internship opportunities of this type. There is clearly a high demand for this kind of program among students: even with the short timeline, we received 144 applications from students all over the US, and some international students. The student selection process emphasized selection of students with excellent technical backgrounds, and/or demonstrated background in social science and interest in social good. The ultimate selection also aimed to meet a diversity of backgrounds and skills, attempting to balance technical skills and social sciences knowledge on each team.

Sixteen students were selected for the program. A majority of the accepted students (12/16) were from the University of Washington. This was partially due to minimal advertising (largely in-house via email) and minimal funding. This program was bootstrapped from available funds and therefore funds for travel and accommodations for non-Seattle residents were not included. Of the students selected, twelve were graduate students (Masters and Ph.D.), and four were undergraduate students. Fields of study were diverse, spanning social sciences and STEM fields, and ranging from sociology, economics, geography and political science, to astronomy, applied math, engineering, computer science and statistics. The sixteen students were joined by an additional undergraduate student who was funded through a selective summer research program for underrepresented students in

[1] http://dssg.io/

[2] http://dssg-atl.io/

[3] http://urban.uw.edu/

**Figure 1: DSSG participants working at the University of Washington Data Science Studio**

computer science to work on one of the projects (Paratransit; see below section 3).

## 2.2 High-school students

In addition to the student DSSG fellows, the program incorporated six high-school students participating in the UW Alliances for Learning and Vision for Underrepresented Americans (ALVA) program. The ALVA program targets students from underrepresented groups for summer internships. These students spend nine weeks on the UW campus, attending classes in a variety of scientific fields including chemistry and mathematics, and participating in scientific research through immersion in a specific research project. Three of the DSSG projects included two ALVA students each. This distribution allowed the cohort of high school students to be exposed to a breadth of different projects while encouraging them to work in pairs to design their own individual research projects around an element of the larger DSSG projects.

The integration of high school students with the interns was initially a challenge. The diversity of skill levels within the cohort of student interns was already quite broad, and the introduction of additional students at a much earlier level of education complicated the design of tutorials and activities. Engagement on the teams themselves was also challenging, in part due to the software-intensive nature of the projects. Realizing that our initial attempts to fully integrate the ALVA students and DSSG interns were not delivering the benefits we anticipated to the ALVA students, we decided to place the ALVA students on a separate track where they could focus on using a single software package to solve a well-defined, introductory, but still highly-relevant component of the broader project. For example, some students began writing simple scripts to clean and format the data, while others focused on building map-based visualizations of the spatial patterns in various datasets. Within this framework the students became more engaged and were able to later re-integrate into the larger projects.

# 3. THE PROJECTS

Applications for projects were solicited in early April and due one month later. Eleven different project proposals were received, and four of these were selected. Selection was based on match to the overall theme (urban data), as well as the specificity and feasibility of the proposed deliverables within the 10-week time frame. Other considerations were preference for proposals which could use similar technical tools (primarily Python, R, SQL, and web technologies) and availability of initial datasets to work with. A more detailed description of the data and results of the projects is beyond the scope of this presentation. Here, we provide an overview of the projects that were selected, their goals, data-sets and some preliminary results of their analysis.

1. **Assessing Community Well-Being Through Open Data and Social Media (Community Well-being)**: The objective of this project was to provide neighborhood community organizers with better understanding of the factors that impact their communities' well-being. Through crowd-sourced community networks that leverage diverse social media and open data sources, neighborhoods can identify emerging issues, see how they compare with other neighborhoods on key factors, and coordinate a community response. While the project's goal was to provide tools that can serve all neighborhoods, the Community Well-being team also actively engaged underserved neighborhoods in designing the program. This project was led by Shelly Farnham, from Third Place Technologies,[4] a local non-profit organization.

2. **King County Metro Paratransit (Paratransit)**: King County Metro Paratransit is an on-demand public transportation program that provides a vital link to mobility for people with disabilities who are unable to use traditional fixed route services, picking up passengers at or near their doorstep and delivering them to their specified destination. Currently, King County Metro paratransit trips cost approximately ten times as much as an equivalent trip using a fixed-route service. To date, little investment and research has been made surrounding the technical complexities of providing ADA paratransit. By analyzing current Metro system information, the project aimed to help dispatchers and

[4] http://thirdplacetechnologies.com/

**Table 1: Roles of different participants**

| Participants | Role | Organization |
|---|---|---|
| DSSG fellows | Full-time research in teams; participation in educational and professional development activities (tutorials, lectures and reading group; see section 4) | Undergraduate and graduate students from different universities (12 from UW); see section 2 |
| High school students | Participated in DSSG research, but also attended separate classes and activities; see section 2.2 | UW ALVA Program (incoming UW freshmen from underrepresented groups) |
| Project leads | Data "owners"; designed project, including questions, scope, and methodology; worked alongside the student interns two days/week, providing domain-specific expertise, and guiding the project at the conceptual level | Partner organizations (e.g Taskar Center, Gates Foundation); see section 3 |
| Stakeholders | Provided feedback and guidance during the course of the projects through periodic meetings, email, etc. | A variety of government organizations, NGOs, corporate stakeholders, etc. |
| Data Scientists | Participated in selecting projects and students; worked alongside students two days a week, providing data science expertise; instructed tutorials in programming, version control, etc. | University of Washington eScience Institute |
| Ethnographer | Participant observer; conducted observations of most program activities and interviews with participants; helped coordinate the planning phase of the program; arranged for external speakers and facilitators (see section 5); see section 6 | UW eScience Institute |
| Program managers | Designed and managed all DSSG activities, from planning through preparation and selection of students and projects, to coordination of the day-to-day operations and activities of the DSSG | UW eScience Institute |

schedulers make informed and more efficient routing decisions that improve the paratransit services offered to passengers while containing the costs of those services. This project was led by Anat Caspi, from the Taskar Center for Accessible Technology[5] in the University of Washington Computer Science & Engineering Department.

3. **Open Sidewalk Graph for Accessible Trip Planning (Open Sidewalks)**: This project addressed the information challenge to design an open source software toolkit and set of algorithms to help those with limited mobility plan a commute. By developing city-wide sidewalk accessibility analytics and applying routing algorithms, the project assembled disconnected sidewalk segments into a coherent graph, which could provide rapid and convenient routing for those with limited mobility. Future algorithm development based on this work will be designed to avoid steep hills, uncrossable intersections, stairs, or construction that blocks sidewalks. Efforts were made to maintain a flexible design that can accommodate datasets for cities other than Seattle. This project was led by Nick Bolten, University of Washington Electrical Engineering Department.

4. **Predictors of Permanent Housing for Homeless Families in King, Snohomish, and Pierce County (Permanent Housing)**: The main objectives of the Permanent Housing project were to identify the barriers preventing homeless families from finding housing, as well as the trends and factors that affect a family's length of stay in a homeless shelter. The research will be used to improve decision making and prioritize resources to help homeless families find permanent housing and reduce their length of stay in a shelter. This project was led by Neil Roche and Anjana Sundaram, from the Bill & Melinda Gates Foundation.[6] The Gates Foundation and Building Changes[7] are co-leading an effort to cut family homelessness in the region in half by the year 2020. An important aspect of this effort is the measurement and analysis of data about homeless families, collected in shelters in the three counties surrounding the Seattle metropolitan area and analyzed by the DSSG student interns.

All projects included a field trip experience in which interns could interact with project stakeholders, gain better understanding of the problems and the collected data, and discuss deliverables. Community Well-being participants conducted surveys in the International District, which was one of the targeted neighborhoods for the project. Paratransit team met with representatives of the Paratransit IT team at King County Metro and learned about Access bus operations, current hurdles and opportunities for improvement. Open Sidewalks team visited Seattle Department of Transportation to meet with their GIS Lead and also met with an individual with a limited mobility to get insight into a wheelchair user's experience of commuting in Seattle. The Permanent Housing team had several meetings at the Bill & Melinda Gates Foundation with the data officers of the involved counties and NGOs, allowing the student interns to ask detailed questions about the data and better understand the context of their analyses. This direct interaction provided DSSG interns with the invaluable opportunity to refine their projects and deliv-

[5] http://tcat.cs.washington.edu/

[6] http://www.gatesfoundation.org/

[7] http://www.buildingchanges.org/

erables to the needs of the stakeholders, and was an overall interesting and informative experience for all participating parties.

# 4. TUTORIALS

During the course of the program, many tutorials were given on a variety of data science topics. Data scientists and research scientists from the UW eScience Institute, UW graduate students, and industry professionals volunteered their time to teach the DSSG interns the skills and best practices for becoming successful data scientists. We aimed to include tutorials which cover all stages of the data science workflow. In the tutorials, we tried to emphasize skills and tools that would be useful in a variety of contexts and scenarios, while the project teams worked with their Data Scientists and Project Leads to further develop the specific skillsets and toolkits they needed to complete their projects (see table 2.2 for the different roles). Many tutorials were concentrated in the first few weeks of the program to try to get students up to speed on important technical skills. This front-loading provided more time to devote to project work as the program advanced. We list tutorials in table 4 and provide more details below.

**Table 2: Topics and technologies for DSSG tutorials**

| **Category** | **Tutorials** |
|---|---|
| Data Management | SQL, ArcGIS, Socrata |
| General Programming | Python, R |
| Improving Workflow | Git, GitHub, Reproducibility, Design |
| Cloud Services | Amazon, Google |
| Analytics | GraphLab, Machine Learning, Twitter Data Analysis |
| Data Visualization | Tableau, D3 |
| Scientific Communication | Social Media, Blogging |

### Data Management

*SQL & SQLShare:* A tutorial on SQL introduced the DSSG interns to writing basic SQL queries, connecting to a database, and extracting summary statistics from city datasets. Students also learned about SQLShare: a Database-as-a-Service environment developed at UW aiming to increase uptake of relational database technology in the sciences.

*ArcGIS:* Several of the projects required working with geospatial data, so we offered tutorials on ArcGIS, some of which were geared specifically toward the high school students.

*Socrata:* Socrata specializes in building tools that allow different organizations to make their data openly available. For example, Socrata has built the infrastructure behind data portals of several cities, including Seattle [8], New York [9] and others. A presentation by Socrata described tools available for acquiring open city data and some upcoming functionalities of their API.

### General Programming

*Software Carpentry Workshop:* During the second week of the program we allowed students to participate in a Software Carpentry workshop that was held at the Data Science Studio [4]. These two-day "boot camps" teach scientists hands-on skills and best practices for programming and operating a modern programming environment. Topics include automation with the Unix shell, version control with Git, and programming with R or Python (taught in two parallel sessions).

Software Carpentry participants could choose only one language to learn, and as we wanted to expose students to both languages, we decided to host individual Python and R programming tutorials.

*Python :* Python tutorials were distributed through several sessions. They assumed no programming background and began with basic variable and control flow examples in an IPython notebook. The interactivity of IPython Notebooks and the gentle learning curve of the Python language was especially appealing to the high school students. More advanced students learned more from later sessions, which concentrated on Object Oriented Programming, array and data processing with Numpy and Pandas, and machine learning with scikit-learn.

*R:* The R tutorial was offered later in the session after students had already been exposed to introductory programming concepts, allowing a stronger focus on the distinctive (and idiosyncratic) aspects of the R syntax, the strengths and weaknesses of R itself as a language and programming environment, and common pitfalls encountered by beginning users. The students worked with R through the RStudio development environment.

### Improving the Workflow

*Git:* As all interns were expected to develop their code using modern practices including version control, we introduced them to Git and GitHub during the first week of the program. Although the learning curve of these technologies is rather steep, over the course of the program students became remarkably comfortable with the workflow and appeared to gain a strong appreciation for the value of these systems in managing group development.

*Reproducibility Tutorial and Workshop:* We introduced students to the importance of reproducibility as a guiding principle in science, and taught them best practices for making their research reproducible. The students participated in a half-day reproducibility workshop in which they practiced simple concepts that make research easier to interpret and reproduce by external collaborators. The students also discussed issues they have encountered using GitHub in project organization.

*Design:* A tutorial on the role of design in software engineering reminded DSSG fellows that the majority of development effort is spent on understanding the requirements of the problem and designing the structure of the solution as opposed

[8] `https://data.seattle.gov/`

[9] https://nycopendata.socrata.com/

to typing lines of code.

### Cloud Services

The DSSG fellows were exposed to various cloud computing and storage services provided by Amazon and Google, and were presented some preliminary steps of getting started using those resources.

### Analytics

*GraphLab tutorial:* A representative from Dato [10] gave a two-session tutorial on analyzing data at scale with the *GraphLab Create* platform: the first part described general steps in building classification algorithms, and the second part demonstrated how to use deep learning within GraphLab to classify images.

*Machine Learning:* A DSSG intern volunteered to give a guided tour of common machine learning algorithms through examples in an IPython notebook.

*Twitter Data Analysis:* Shelly Farnham, who was the project lead for the community well-being project (see section 3) provided insights into the workflow for extracting social media data to make inferences about social phenomena.

### Data Visualization

*Tableau and D3:* We had two tutorials focused on interactive visualization: one on Tableau and one on D3. Thus, we introduced two different approaches to visualization, emphasizing that Tableau delivers value by supporting exploratory analysis without programming, while D3 provides a library with which one can design sophisticated and customized interactive visualizations and integrate them into broader web applications. Both Tableau and D3 were adopted in some of the summer projects.

### Scientific Communication

*COMPASS Social Media Training:* This tutorial emphasized the importance of engaging with the public to communicate scientific results and provided tips for using social media effectively.

*Blogging:* We requested that the students write a blog on their experience in the DSSG [11]. To help them get started, we offered two tutorials: one on successful blogging practices and another one on writing a blog using the GitHub Pages platform. Although GitHub could be considered a more difficult platform for blogging than some others, it allowed students to practice using Git and GitHub, and kept a description of their project in close proximity to the code itself.

## 5. SPEAKERS AND DISCUSSIONS

Although we placed a strong emphasis on providing opportunities to develop technical skills, we also incorporated numerous presentations and discussions on topics specific to the role of data science in urban, social, and global contexts.

Guest speakers and facilitators from the governmental, nonprofit, and academic sectors engaged with DSSG participants on variety of topics: key principles for working with multiple stakeholders on data intensive collaborations; the city of Seattle's approach to data-driven accountability and the backend of Seattle's online Performance Dashboard; the importance of maintaining a historical perspective and thinking about macro-level changes when working on projects for social good; ways that data can be leveraged to create more equitable education systems; and innovative methods for using mobile phone data to infer socioeconomic information.

Aside from these presentations, we also convened a weekly reading group aimed at providing a platform to explore and discuss opportunities and challenges specific to the application of data science methods on social data. The format was similar to a journal club, and discussions were facilitated by experienced researchers drawn from various academic units at the University of Washington. These included a team of several advanced graduate students from the Sociology Department, a DSSG intern who had recently received his doctorate in geography, and professors from the Sociology, Geography, and Civil & Environmental Engineering Departments. The facilitators assigned one or two readings and used the discussion sessions to draw out relevant themes, some of which included: epistemological assumptions embedded in data science analyses; the difficulty of ensuring representativeness with organic data that has not been generated for the purpose of research; the importance of domain knowledge in understanding social phenomena; the rarity of communication and collaboration across sectors and disciplines among people using data science to answer social questions; and innovative methods for using novel combinations of data sources in data-poor contexts.

## 6. ETHNOGRAPHY

The eScience Institute works closely with a team of ethnographers who are studying cultural and organizational dimensions of data science in the academy. This ethnographic research was incorporated into the design of the Data Science Environment (DSE) grant that drives much of the eScience Institute's work. The ethnographers working on the DSE embed themselves in the places people are using and learning data science methods in order to understand how different communities make sense of and value data, and what is organizationally required to support data intensive practices and collaborations. They focus on forming insights that lead to better tools, educational programs, and institutional initiatives for data science. This summer, one of these ethnographers (an author of this paper) conducted fieldwork at the DSSG. She interviewed nearly every individual who participated in or helped to organize the DSSG, and observed team meetings, informal work time, tutorials, talks, field trips, and social activities. As a participant observer, she also contributed to the program by helping to coordinate the planning phase of the program and arranging the visits of many of the guest speakers and facilitators. The analysis

[10] `https://dato.com/`

[11] `http://uwescience.github.io/DSSG2015/`

of her ethnographic data will be conducted following the conclusion of the program and considered alongside other ethnographic data the team has collected across various field sites.

## 7. FEEDBACK AND EVALUATION

We solicited feedback from the DSSG fellows periodically throughout the program via online surveys with free-text responses.

Issues that were often raised relate to the use of space in the Data Science Studio. The open space, while it facilitates collaboration between the teams, can also pose challenges for those who prefer a quieter and less active work environment.

Another issue we encountered, one that often arises in groups with diverse backgrounds and skills, is that it was difficult to design training events to meet the needs of all participants simultaneously. While some of the interns found the tutorials to be helpful, others with a stronger technical background could, at times, find themselves bored. In addition, some students relayed that they would have preferred working on their projects to spending time in tutorials. The density of tutorials decreased in later weeks, partially in response to this feedback, and partially because the program was purposely frontloaded with tutorials on skills that students were likely to use on their projects right away.

An additional, more extensive survey was conducted at the end of the program. In this survey, several of the DSSG fellows commented that the experience had changed their views on the importance of reproducibility in their own research ("I have never given much thought to using programming tools to facilitate reproducibility, but it's a very useful practice"; "I will make my own work much more reproducible"; "I always understood why reproducible research is important, but now I have first-hand experience in it"). Many students expressed an interest to stay involved in data science focused on social good, whether through continued work on these projects, or in other projects. Other comments related to the high level and quality of the interaction with stakeholders ("...really unique relative to other internships!").

We asked students to mention examples of skills that they had learned over the course of the summer. Students who had indicated that they had not used programming in their research before the summer, or used it only minimally emphasized acquiring these skills, while more experienced students mentioned specific advanced programming skills, such as machine learning techniques that they had learned, or programming with specific libraries (Pandas was often used and often mentioned), for specific APIs, specific analysis techniques, and so forth. Many DSSG fellows mentioned the value of learning how to use version control with git and Github.

## 8. SUMMARY AND REFLECTIONS

An urban-themed program like the UW eScience Institute's Data Science for Social Good simultaneously addresses several challenges in the realm of data science.

First and foremost, it serves as a context in which data science methods can be leveraged to address social issues in cities and other communities. The data scientists provide the expertise with the tools and techniques of data management, analytics, and collaborative software engineering, and visualization. The partner organizations (government, non-profit, academic) provide deep understanding of the problem to be solved and keep the project on the critical path. The students provide a crucial diversity in background and perspectives and skills to ensure creative solutions (in addition to bringing passion and effort). In the best case, this combination can be harnessed to achieve significant impact on social good.

A second problem that this type of program solves is the need for relevant, practical, and rigorous training in data science. Learning objectives are achieved by providing student interns both with explicit training through tutorials and lectures, but also with the collaborative, team-based project experience that professional data scientists (and their hiring managers) value highly. This type of training in data science methods complements standard classroom training, in which students are given crafted, simplified examples with answers that facilitate easy grading rather than deep learning.

The risk with real-world data sources is that a promising project can be scuttled by bad data. On the other hand, the experience of performing the data "janitorial" work that is cited as the most time-consuming component of practical data science projects [2, 3], is invaluable.

The significant interest both from students and from project leads in this program suggests that its design could be translated to many different contexts. The experiences that we describe, and the processes we have in place, could serve as a template from which to design other programs.

We would like to highlight the importance of the role of data scientists in this project. This is a new role within academic institutions: a role for individuals who focus on advancing the use of advanced techniques and technologies for data-intensive science, and the methodologies for creating useful and reusable software. The data scientists served here as mentors and facilitators of the interactions between project leads and the interns, helping to identify the technical tools relevant to each project, and helping to establish processes for collaborative work in each project team. The data scientists also were the instructors of several of the tutorials. While it is not necessary to have data scientists on staff to run a program such as this, it does help to have individuals that can take on the role of technical mentors, and facilitators of working process.

Effective program management also played a crucial role in the success of our effort. The flexible and open nature of the program required a management structure that was nimble and adaptable to change, and capable of directing resources to address problems as they arose. The eScience Institute has two Ph.D. scientists who job-share the Program Management role (both among our authors), and they played an instrumental role in all aspects of this program: during project creation, the managers worked to assign data scientists to the team that best linked with their specific skills, and to assign the high-school students to the most appropriate group. As the project progressed, the managers issued weekly schedules to the entire team, facilitated weekly project updates in a group setting, and networked with individuals on specific topics. We found that this communication within and across teams, at least on a weekly basis, helped to minimize duplication of efforts, increased cross-pollination of ideas, and helped to identify bottlenecks and propose solutions.

## 9. ACKNOWLEDGMENTS

We would like to thank our collaborators from the Taskar Center for Accessible Technology, University of Washington Computer Science & Engineering, from Third Place Technologies and from the Bill & Melinda Gates Foundation. Many thanks to the DSSG student interns for their participation and hard work, and to the various mentors, speakers, and facilitators who graciously volunteered their time to contribute to the program. This project was funded by generous grants from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation, with the Data Science Studio facility funded through the leadership of the Washington Research Foundation.

## 10. REFERENCES

[1] N. R. Council. *Training Students to Extract Value from Big Data: Summary of a Workshop*. The National Academies Press, Washington, DC, 2014.

[2] T. Dasu and T. Johnson. *Exploratory Data Mining and Data Cleaning*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2003.

[3] S. Lohr. For big-data scientists, 'janitor work' is key hurdle to insights. *New York Times*, August 17 2014.

[4] G. Wilson. Software carpentry. *Comput. Sci. Eng.*, 8:66, 2006.