Library and Information Services in Astronomy VII: Open Science at the Frontiers of Librarianship
ASP Conference Series, Vol. 492
A. Holl, S. Lesteven, D. Dietrich, and A. Gasperini, eds.

Data Scientist Training for Librarians

Christopher Erdmann

Harvard–Smithsonian Center for Astrophysics, Cambridge, MA, USA.

Abstract. Recent studies suggest that there will be a shortfall in the near future of skilled talent available to help take advantage of big data in organizations. Meanwhile, government initiatives have encouraged the research community to share their data more openly, raising new challenges for researchers. Librarians can assist in this new data-driven environment. Data Scientist Training for Librarians (or Data Savvy Librarians) is an experimental course being offered by the Harvard Library to train librarians to respond to the growing data needs of their communities. In the course, librarians familiarize themselves with the research data lifecycle, working hands-on with the latest tools for extracting, wrangling, storing, analyzing, and visualizing data. By experiencing the research data lifecycle themselves, and becoming data savvy and embracing the data science culture, librarians can begin to imagine how their services might be transformed.

1. Background

An often-cited McKinsey Global Institute study forecasts a significant gap in big data skills within the U.S.: "By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions" (Manyika et al. 2011). According to Gary King, director of Harvard's Institute for Quantitative Social Science, "the march of quantification, made possible by enormous new sources of data, will sweep through academia, business and government. There is no area that is going to be untouched" (Lohr 2012). Enter the "data scientist," a term used by DJ Patil and Jeff Hammerbacher to describe their positions at LinkedIn and Facebook, respectively, where they derived valuable insights from big data to develop innovative solutions for their companies.[1] Much has been written about the need for more data scientists. The origins of "data science" go back as far as John W. Tukey and Peter Naur, and recent champions such as Hal Varian continue to advocate the importance of understanding data and extracting value out of it. Hal Varian explains, "the ability to take data — to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it — that's going to be a hugely important skill in the next decades."[2] These skill sets will need to be widened beyond just data scientists and developed within the greater workforce in order to respond to the greater deluge of

[1]Wikipedia — DATA Science `http://en.wikipedia.org/wiki/Data_science`

[2]Press, G. 2013, Forbes Technology, A Very Short History Of Data Science `http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science`

data (Miller 2014). Also, a cultural shift is necessary within teams or organizations to understand the importance of data throughout its lifecycle (Schultz 2014).

Big data is having a transformational effect on industries and professions. For instance, a recent survey of journalists by the European Journalism Centre (EJC) highlighted a desire by the community to learn data driven journalism and a recognition of its growing importance.[3] This same desire and recognition was reflected in the registration responses received for Data Scientist Training for Librarians (DST4L) where librarians highlighted the need to learn new data-related skills, understand the research data lifecycle, and in the end, "help facilitate the huge data needs of our patrons."[4] Librarians recognized that their newfound skills could also be used to improve library workflows. In addition to the registration responses, library schools are also noticing the decline of traditional employment areas, the rise of new, data-related positions and the opportunity to update library school curriculums.[5]

One place in the research data lifecycle where librarians can make an impact is in the discovery, understanding, and cleaning of data for analytic use. "The process of iterative data exploration and transformation that enables analysis," or data wrangling, is referenced frequently by data scientists as having the biggest impact on the data science process, taking up to 80 percent of their time (Kandel et al. 2011). An analyst interviewed by Heer and Kandel noted, "I spend more than half of my time integrating, cleansing, and transforming data without doing any actual analysis. Most of the time I'm lucky if I get to do any 'analysis' at all!" (Heer and Kandel 2012). In addition to the innovative approach to data wrangling proposed by Heer et al, it is clear that scientists can also benefit from the assistance of data savvy librarians to help reduce the possible number of iterations within these initial stages of research. For instance, librarians are already adept at finding novel data sources, providing subject background, cataloging records, and offering metadata advice, to list a handful of library services that data scientists can leverage. Scientists may be unaware of these services, highlighting a possible marketing challenge and opportunity faced by the library community. Experiencing the research data lifecycle firsthand and upgrading to data savvy skills can help librarians improve outreach and services to scientists.

Libraries can help foster the environmental conditions necessary to expand the data science skill sets of scientists and librarians, together. Time is wasted on both sides "doing things badly that could be done well in just a few minutes."[6] Greg Wilson and the team of volunteers behind Software Carpentry lead bootcamps aimed at increasing scientists' computational understanding, while enhancing their habits and routines (Wilson 2014). The bootcamps have been successful at responding to needs that have arisen within the scientific community for practical, hands-on programming and software assistance. They also mirror elements of modern hacker culture, where open source software and open access data are used and promoted (Schrock 2014). Like hackathons, groups of individuals come together to collaborate intensively on software

---

[3]Lorenz, M. 2011, Data Driven Journalism `http://datadrivenjournalism.net/news_and_analysis/training_data_driven_journalism_mind_the_gaps`

[4]Erdmann, C. 2013, DST4L Registration (Responses). Unpublished raw data.

[5]Larsen, R.L. 2014, 9th International Digital Curation Conference, `http://www.dcc.ac.uk/events/idcc14/programme-presentations`

[6]Software Carpentry FAQ `http://software-carpentry.org/faq.html`

projects, and the social learning that takes place around these projects can facilitate rich social networks.[7] Historically, libraries have been centers for the diffusion of knowledge; they can adapt to support training programs like Software Carpentry, and reinforce their role once more in the process.

2. Training

Started at the Harvard–Smithsonian Center for Astrophysics (CfA) John G. Wolbach Library, the DST4L program grew out of a local need to train staff in data-centric services. Early training activities focused on learning digital curation methods and techniques employed by astronomy libraries (Accomazzi et al. 2012). Later, it became evident that the CfA Library staff also needed to gain a broader understanding of the research data lifecycle, to respond to an evolving list of data related needs from the CfA community. Inspired by a CfA Library staff member, the DST4L training program was devised to respond to this need and the decision was made to open the training to the local library community, beyond Harvard University.

Resources used to draft the first two DST4L courses came from a number of places.[8] Hammerbacher and Franklin provided the most beneficial resource: an online curriculum for a course offered at University of California, Berkeley titled "Introduction to Data Science." Beyond these traditional sources, Twitter provided a rich, constant feed of information on data science, with references to online blog stories and tutorials, experts communicating helpful information back and forth to one another, and local meetings where one could meet individuals interested in or presenting on the subject of data science (e.g. OpenVis and Lynn Cherny). Meetups such as the Data Science Group in Boston served as another rich source for networking with local data science experts (e.g. David Dietrich, EMC). Also, local members of the Harvard University and Smithsonian Astrophysical Observatory communities were tapped for feedback and assistance (e.g. Rahul Dave). Many experts from the Boston area contributed to the course, teaching sessions (e.g. Tom Morris from OpenRefine) or giving talks (e.g. James Turk from the Sunlight Foundation), while providing greater context to the material.

DST4L took a hands-on approach to teaching the different aspects of the research data lifecycle (see Figure 1). Students started learning about data sources and how to extract data from them, either through an API or by scraping a website. From there, they moved on to wrangling with the data which involved cleaning, reconciling, and transforming the data into a format more useful for analysis. Next, they covered statistical analysis and natural language processing and finished with data visualization principles. Throughout DST4L, the participants used tools with varying levels of complexity, from Excel to Python. The main technologies used in both courses included Unix Shell, Git, GitHub, Python, iPython, Excel, OpenRefine, SQL, Data Repositories, R, RStudio, Tableau, D3, NoSQL, MongoDB, and Gephi.

---

[7] Wikipedia: Hackathon `http://en.wikipedia.org/wiki/Hackathon`

[8] Patil, D.J. 2012, Data Jujitsu: the art of turning data into product; Segaran, T., Hammerbacher, J. 2009, Beautiful data: the stories behind elegant data solutions; Gray, J., Chambers, L., Bounegru, L. 2012, The data journalism handbook, all from O'Reilly Media, Inc.; EMC Education Services 2012, Data Science

Figure 1. DST4L class picture demonstrating hands-on approach.

Much of the DST4L course material is currently accessible to the outside world. A WordPress site is available with blog entries, written by participants, capturing each session, with accompanying notes, code, data, and anything else used in the sessions[9] (see Fig. 2.). Course details such as the syllabi are also open for perusal. For each class, instructors and students utilized open collaborative writing and note taking tools such as Google Docs and Etherpad. When combined with the WordPress site, these documents could easily be searched through the website or Google for quick reference. iPython Notebook proved to be a powerful tool for instruction, walking through code line by line with the students. This tool is being used more and more within the scholarly workflow and is an example of new forms of scholarly output that librarians should be aware of (see Fig. 3). Live streaming courses proved troublesome, and in the end students preferred meeting physically, especially since the group projects helped to engage them with the resources. The participants in the course were a very diverse group, including librarians from the Federal Reserve Bank, students from Simmons Graduate School of Library and Information Science, and MIT, which helped to enhance the experience for everyone involved.

Students in both courses were encouraged to demonstrate what they learned in the course either through projects describing their learning experiences or participating in a hackathon. Data stories from the first course are available via the DST4L website. Not much remains from the hackathon at the end of the second course except for a topic modeling visualization by Sands Fish,[10] a Massachussetts Beer/Data Map,[11] and a repo on GitHub from Jeremy Guillette and other members of the CfA Library called DST4L Mapathon.[12] In addition to these activities, participants had an opportunity to share

---

and Big Data Analytics; Hammerbacher, J., Franklin, M. 2012, CS 194-16: Introduction to Data Science, University of California, Berkeley `http://datascienc.es/`

[9]Data Scientist Training for Librarians Website `http://altbibl.io/dst4l/`

[10]`https://www.flickr.com/photos/sandsfish/12076246955/`

[11]`http://goo.gl/Ly6dxU`

[12]`https://github.com/jaguillette/dst4l_mapathon`

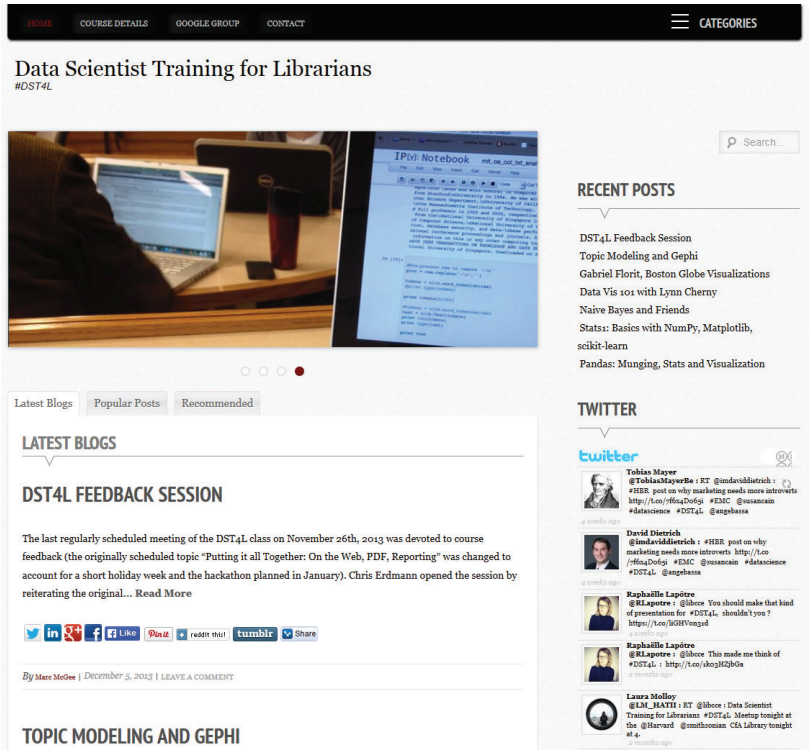

Figure 2. The DST4L website.

their feedback and present to the library community.[13] Two quotes from past participants capture their thoughts on the program. Vernica Downy said at a talk, paraphrased via John Overholt's tweets, "Data scientist training really changed how I think about my job... I'm an intrepid data explorer, and we need more of that in librarianship."[14] Vernica also referenced the challenges catalogers face, but once again captured in a tweet from John Overholt, she explained that "catalogers with the right tools can be more powerful than ever." Jeremy Guillette said via a feedback session, "in preparing to enter the field [of librarianship] for a long career, I'm going to keep on seeing this stuff."[15]

Following both courses, participants reported that they gained a better understanding of the research data lifecycle, which was the main goal of the course. As a result of DST4L, reference librarians have assisted patrons with advanced questions involv-

[13]Data Scientist Training for Librarians: Summing Up `http://altbibl.io/dst4l/summing-up/`; Data Scientist Training for Librarians: Feedback Session `http://altbibl.io/dst4l/dst4l-feedback-session/`

[14]Overholt, J. 2014, Data Scientist Training for Librarians Tells All `https://storify.com/libcce/data-scientist-training-for-librarians-tells-all`

[15]Rubin, L. 2013, Data Scientist Training for Librarians Tells All, DST4L Presentation Panel Discussion `http://www.youtube.com/watch?v=U5ZYM085bNo&t=1m21s`

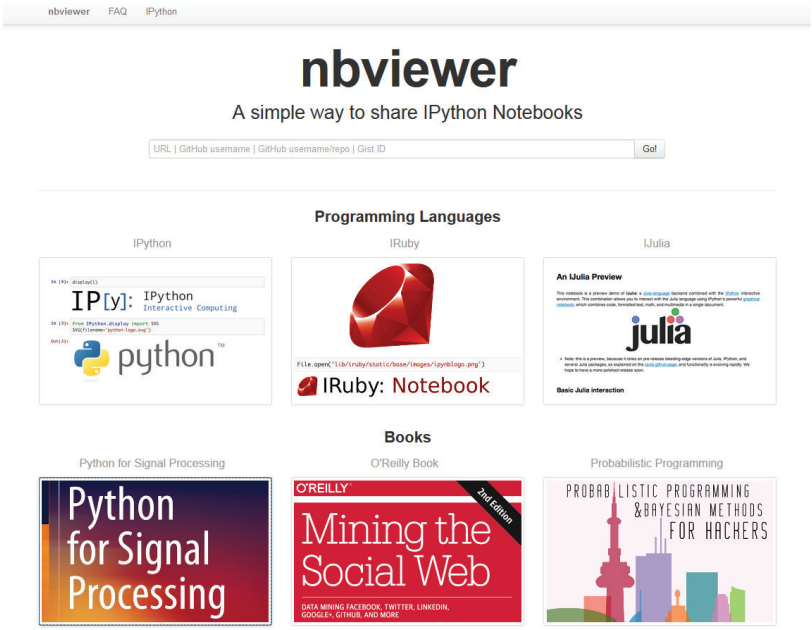

Figure 3. nbviewer is a simple way to share IPython Notebooks.

ing data analysis and catalogers have streamlined OCLC reporting processes. In other cases, librarians have advanced far beyond expectations. For instance, one participant from the course is now assisting the NASA ADS with visualization tools, and another is creating interfaces and visualizations for a controlled vocabulary of terms.

3. Conclusion

The DST4L program hits on a number of goals. First, allow librarians to experience the research data lifecycle so that they can start thinking about how they might modify or offer new services. Second, train librarians to be data savvy. Third, address the culture within libraries and change the "library mindset" through abstract thinking, continuous learning, hacking, and other approaches. Fourth, grow a community of data savvy librarians that can act as a support network not only for other librarians but also for the research communities they support. Through these goals, the DST4L program aims to foster the services and environments that libraries will need in order to respond to the changing data needs of their communities.

Acknowledgments. Thanks to the CfA Library for hosting and supporting DST4L, to the Harvard Library and Arcadia Fund for staffing and funding support, and to the many volunteers, instructors, speakers and participants that made the course possible.

References

Accomazzi, A., Henneken, E., Erdmann, C., Rots, A., 2012, Proc. SPIE, 8448, 4480K-84480K-10, http://dx.doi.org/10.1117/12.927262

Gray, J., Chambers, L., Bounegru, L., 2012, The data journalism handbook, O'Reilly Media, Inc.

Heer, J., Kandel, S., 2012, XRDS, 19, (1), 50 http://doi.acm.org/10.1145/2331042.2331058

Kandel, S., Heer, J., Plaisant, C. et al. 2011, Information Visualization Journal, 10, (4), 271. http://idl.cs.washington.edu/papers/data-wrangling

Lohr, S., 2012, The New York Times, 2, 12 http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H., 2011, Big data: The next frontier for innovation, competition, and productivity, The McKinsey Global Institute

Miller, S., 2014, Journal of Organization Design, 3, (1), 26 http://dx.doi.org/10.7146/jod.9823

Schrock, A.R., 2014, InterActions: UCLA Journal of Education and Information Studies, 10, 1 https://escholarship.org/uc/item/0js1n1qg

Schultz, J.R., 2014, Performance Improvement, 53, (5), 20 http://dx.doi.org/10.1002/pfi.21411

Wilson, G., 2014, F1000Research, 3:62 http://dx.doi.org/10.12688/f1000research.3-62.v1