

EDITOR'S SUMMARY

The new field of data science involves advanced knowledge in statistics and computer science, combined with copious amounts of data. A report from the Big Data and Research Initiative under the Obama Administration, *The Federal Big Data Research and Development Strategic Plan*, calls attention to the roles that librarians will play in the future of data science. However, there are skills and management gaps librarians face that inhibit their ability to move forward in data science. A number of educational programs are now offered to remedy this problem, such as the Data and Visualization Institute for Librarians from North Carolina State University, the volunteer-led Library Carpentry program, and most recently, the Data Sciences in Libraries Project, funded by the IMLS. This project aims to get librarians and library managers together to discuss the world of data science and create a roadmap for strategic planning.

KEYWORDS

data science
skills
librarians
managers
training
strategic planning
domain knowledge

Data Science in Libraries

by Matt Burton and Liz Lyon

Advances in statistics and computer science, combined with an abundance of data, have given rise to a new professional ecosystem called *data science*. Data science has many definitions, but at the core it is about “generating insight from data to inform decision making.” Data science methods and products have transformed commerce, health care and government, and they will continue to transform other sectors. *The Federal Big Data Research and Development Strategic Plan*, recently released by the Obama Administration’s Big Data and Research Initiative, explicitly identifies curators, librarians and archivists as core specialists to help to meet growing demand for analytical talent and capacity across all sectors of the national workforce. As society is increasingly infused with data, librarians will have a crucial role in the future development of the data science ecosystem across multiple sectors; the report acknowledges that “investments are needed to expand the current pipeline of support to the field of data science.” [1, p.29]

Matt Burton is a visiting assistant professor in the School of Information Sciences at the University of Pittsburgh and post-doctoral researcher in the department of digital scholarship services at the University of Pittsburgh Libraries. He can be reached at mburton@pitt.edu.

Liz Lyon is co-chair, Information Culture & Data Stewardship Program, interim Doreen E. Boyce Chair and visiting professor in the School of Information Sciences, University of Pittsburgh. She can be reached at elyon@pitt.edu.

The Skills Gap

Universities are making significant investments in data infrastructure across their campuses, and librarians will need to be prepared to contribute to these efforts. The ACRL 2015 Environmental Scan identified a need for more advanced data curation services, urging librarians to have a deeper knowledge of domain research practices to enable them to help researchers with data management, sharing and preservation. Librarians’ blended domain knowledge, that is, technical skills combined with contextual understanding of the domain, will have a transformational impact on professional roles, associated practices and the perceived value of libraries more widely. However, many librarians lack the technical skills to be effective in a data-rich research environment. We call this the *skills gap*.

Fortunately, there are many continuing education programs, training camps and informal workshops oriented toward teaching librarians technical skills. Out of an increasingly urgent need for technical expertise at the Harvard/Smithsonian Center for Astrophysics, Chris Erdmann created data science training for librarians (DST4L). “The main objectives [of DST4L] are for participants to learn to extract, analyze and present data using the most-up-to-date techniques...our staff should have the same skills as the scientists and researchers who patronize their libraries, so they can understand their data needs better and build services that respond better to their needs.” [2] North Carolina State University has launched

the [Data and Visualization Institute for Librarians](#), a week-long course to help librarians “develop knowledge, skills, and confidence to communicate effectively with faculty and student researchers about their data.” The program focuses on deep technical skills, such as data analysis, visualization and sharing, and also deals with advanced topics such as statistical analysis or version control with GitHub. [Library Carpentry](#), a volunteer-led continuing education program, has been teaching librarians the computing skills necessary for data-intensive research. The program emerged to fill an ever growing gap between the technical skills needed to automate their work and facilitate computational research and their previous training.

The Management Gap

The effectiveness of ad-hoc, informal or short-term training and education programs is limited because they operate outside traditional institutional support structures, professional development and incentive systems. The incentive structures for mid-career librarians can be misaligned or opposed to the development of technical skills. Investing in professional development, like creating opportunities for librarians to attend the NSCU Data and Visualization Institute, or hosting a Library Carpentry workshop, has the potential for significant benefits, but only when administrators create opportunities for the application of the newly acquired skills. Practicing librarians may be blocked by constraining or regressive organizational structures and limitations on personnel; “I learn new skills, but I still need to do my old job.” Job descriptions can be inflexible, making technical professional development difficult, which may discourage the acquisition of new technical skills. The ability of library managers to

understand and to value the benefits of in-house data science skills (for example, to inform decision-making and to enhance services) is critical while organizational and managerial support is essential if technical skills are to be acquired, effectively applied and have an impact. We call this the *management gap*.

Despite the wealth of training programs dedicated to librarians acquiring data skills, there are few focused on cultivating tech-savvy managers or on operationally managing data-intensive teams. Current management programs such as the Harvard Leadership Institute for Academic Librarians provide mechanisms for developing future senior-level managers, but could provide a more specific emphasis on managing data-intensive librarians. Managers need to be aware of new data-science roles [3] and the specific requirements for these positions [4]. They also need to understand how to integrate them into their organizations and envision the diverse contexts, opportunities and benefits in applying data science methods. Library managers and administrators need supporting frameworks and toolkits to leverage data science capability in their strategic planning and decision-making, in the cost-effective operational management of library services and in the development of librarian teams supporting data-intensive communities on campus and beyond.

The Data Science in Libraries Project

The IMLS-funded Data Science in Libraries project seeks to foster a dialog addressing these mutually constitutive skill and management gaps. The overarching goal of this project is to bring library practitioners, educators, managers/administrators and the data science communities into conversation, to foster a multi-directional flow of information,

knowledge and collaborative opportunities. There are many stakeholders at the intersection of data science and libraries. We seek to bring key players together through a structured workshop discussion, authoring a Roadmap and other community-building activities. The Roadmap provides a unifying vision, an outline for strategic planning and operational guidance in implementing data science in libraries. It will include case-studies and best practices and be a rich resource for both library managers and practitioners.

Our related goals are to foster lively community debate, to develop and broaden the data-science-in-libraries community and to lay the foundations for continued discussions, expanding networks and sustained

collaborations. Current data science efforts will benefit from a more coordinated approach, where critical mass advantages can be realized and network effects may be catalyzed. Finally, it is important to note that we are still at the beginning of this data science journey. There is much work still to do to raise awareness, to get success stories out to wider public and to build a cohesive international community of enthusiastic data science library practitioners and engaged library managers. We hope that our IMLS-funded project will be a step in the right direction and will generate some stimulating and provocative outcomes. For more information visit <http://datascienceinlibraries.org/> and sign up for our mailing list. ■

Resources Mentioned in the Article

- [1] Big Data Senior Steering Group, Networking and Information Technology Research and Development Program, National Science and Technology Council, Executive Office of the President.(May 2016). *The federal big data research and development strategic plan*. Retrieved from www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf
- [2] Harvard Library offers data scientist training. (September 27, 2013). Retrieved from <http://library.harvard.edu/02042014-1336/harvard-library-offers-data-scientist-training>
- [3] Lyon, L., & Brenner, A. (2015). Bridging the data talent gap: Positioning the iSchool as an agent for change. *International Journal of Digital Curation*, 10(1), pp. 111-112.
- [4] Lyon, L., & Mattern, E. (In press). Education for real-world data science roles (Part 2): A translational approach to curriculum development. *Journal of Digital Curation*.