Shifting to Data Savvy: The Future of Data Science In Libraries

Matt Burton
Liz Lyon
Chris Erdmann
Bonnie Tijerina

EXECUTIVE SUMMARY

The Data Science in Libraries Project is funded by the Institute for Museum and Library Services (IMLS) and led by Matt Burton and Liz Lyon, School of Computing & Information, University of Pittsburgh; Chris Erdmann, North Carolina State University; and Bonnie Tijerina, Data & Society. The project explores the challenges associated with implementing data science within diverse library environments by examining two specific perspectives framed as 'the skills gap,' i.e. where librarians are perceived to lack the technical skills to be effective in a data-rich research environment; and 'the management gap,' i.e. the ability of library managers to understand and value the benefits of in-house data science skills and to provide organizational and managerial support.

This report primarily presents a synthesis of the discussions, findings, and reflections from an international, two-day workshop held in May 2017 in Pittsburgh, where community members participated in a program with speakers, group discussions, and activities to drill down into the challenges of successfully implementing data science in libraries. Participants came from funding organizations, academic and public libraries, nonprofits, and commercial organizations with most of the discussions focusing on academic libraries and library schools.

The workshop findings are oriented around four distinct facets of a **Multi-Faceted Framework** and commentary is provided in two sections connecting **Structures & Skills** and **Services & Stakeholders**. For each section, there is a scope note, the drivers and barriers are considered, exemplars are provided and recommendations and actions are listed.

The **Structures & Skills** facet highlights the importance of library and information schools (iSchools).

Drivers include:

- changing research needs
- recognition
- collaboration
- supply and demand

Barriers include:

- formal LIS education
- drive-by workshops
- information overload
- the brick wall
- leadership
- incentive structures
- branding

Exemplars include:

- the Carpentry suite
- training programs such as Data Scientist Training for Librarians (DST4L)
- new trends such as reproducibility librarianship

Recommendations include:

- highlighting success stories
- collaborating with leadership institutes
- utilizing physical learning spaces
- advocating for the Carpentries
- leveraging existing educational resources
- repositioning the Masters in Library and Information Science (MILS)

The **Services & Stakeholders** facet recognizes both internal and external data science opportunities.

Drivers include:

- physical space
- facilitating inclusivity
- cost-neutral facilities
- strategic planning
- problem-solving
- stakeholder demand
- cross-disciplinary facilitation
- re-engineering services
- impact metrics

Barriers include:

- the silo effect
- scale
- resources
- credibility and image
- experience
- culture

Exemplars include:

- Library of Congress
- Caltech Library
- University of Illinois Urbana-Champaign
- Carnegie Public Library, Pittsburgh

Recommendations include:

- performing a data audit
- identifying external datasets
- discovering data science requirements
- piloting data services
- quantifying resources to scale up
- re-engineering services
- identifying campus stakeholders
- reaching out to researchers
- showcasing capacity and capability

The workshop identified a number of emerging themes, tensions, challenges, and opportunities. The first theme focuses on culture and embraces both professional culture and organizational culture. Within professional culture, two tensions are highlighted. The first is a **'credentialing tension'** defined as the educational continuum represented by formal versus informal education (multi-year degrees or short courses). When library managers are recruiting individuals, there is a tension between the value and benefits of individuals with a deep disciplinary knowledge (as evidenced by a doctoral qualification, i.e. people who have a Ph.D.), compared to graduates with a broader knowledge (demonstrated by a Masters degree, i.e. people who have an MLIS or similar credential).

Similarly a **'re-engineering tension'** is identified, which can be summarized by the complementary activities of stewardship (roles and functions which focus on policy, data management planning, and advocacy), and science (roles and functions which focus on technical workflows to manipulate and wrangle the data to produce novel insights). Data-savvy librarians occupy a space at the intersection of these competing tensions, and these individuals may offer hybrid or blended skills, e.g. disciplinary knowledge plus experience of curating data as stewards or scientists.

A data-savvy library organization can be characterized by: 1) routine collection and strategic application of quantitative evidence, 2) effective communication and messaging driven by data-rich stories, 3) established professional education programs to extend and expand data savvy skills, and 4) an explicit understanding and articulation of the value and benefits of science & stewardship roles and responsibilities. Two further recommendations were made to ***acquire blended skills and implement a data-savvy 360 model.***

Twenty-two additional recommendations were also made associated with scale and assessment, infrastructure, and ethics and values. A data-savvy roadmap is introduced to help to position the recommendations and actions within short-, medium-, and long-term timeframes. These include ***conducting rigorous assessment, building data partnerships, considering ethics and privacy, offering triage services, and developing libraries as amplifiers.***

Finally, three specific next steps are proposed to take this work to a second phase: 1) holding a series of events framed as a **Data Science in Libraries Community Events**; 2) coordinating information about data savvy training and education initiatives in a **Data Science in Libraries Learning Registry**; and 3) showcasing the **Data Science in Libraries Report** to platforms with library leaders such as the Association of Research Libraries (ARL), and to iSchool Deans.

TABLE OF CONTENTS

INTRODUCTION

Recent 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 methods and products have transformed commerce, healthcare, and government and they will continue to transform other sectors including libraries. ***The Federal Big Data Research and Development Strategic Plan,*** released by the Obama Administration's Big Data and Research Initiative, explicitly identifies curators, librarians, and archivists as core specialists to help to meet a growing demand for analytical talent and capacity across all sectors of the national workforce. The report acknowledges that ***"investments are needed to expand the current pipeline of support to the field of data science;"***[1] as society is increasingly infused with data, librarians will have a crucial role in the future development of the data science ecosystem.

The Skills Gap

A number of universities are making significant investments in research computing and data infrastructure across their campuses[2] and librarians will need to be prepared to assist with and contribute to these efforts. The Association of College & Research Libraries (ACRL) 2015 Environmental Scan[3] 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 a subject—will have a transformational impact on professional roles, associated practices, and the perceived value of libraries and librarians more widely. However, many librarians lack the technical skills to be effective in a data-rich research environment. We call this the **skills gap**.

There is a growing ecosystem of training opportunities for librarians, including training camps and informal workshops, that are oriented toward teaching librarians computational skills and perspectives for working effectively with data. For example, 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).[4] ***"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."***[5] While librarians developing the same skills as scientists and researchers is not entirely realistic, librarians gaining exposure to these skills and developing a spectrum of skills, across a team, or even across the various business units at a university, is a more viable alternative. North Carolina State University (NCSU) started the Data Science and Visualization Institute for Librarians (DSVIL),[6] a week-long course to help librarians ***"develop knowledge, skills, and confidence to communicate effectively with faculty and student researchers about their data."*** Developed by Australian librarians and the Australian National Data Service (ANDS), 23 Research Data Things[7], a popular training program for librarians worldwide, held over the course of months, introduces participants to data-related topics, provides them with greater context, and eases them into a broader spectrum of possibilities and more advanced training activities. Library Carpentry[8] is a volunteer-led continuing education program that has been developed by and for librarians, teaching them skills such as programming in R, data cleaning in OpenRefine, and version control, allowing them to be more effective in their own work and to gain more of a perspective of researchers needs around data. The program emerged to fill a growing gap between current Library Science training programs and the technical skills needed to automate their work and facilitate computational research.

The Management Gap

The ability of librarians to take advantage of ad-hoc, informal, or short-term training and education programs is limited because they operate outside of 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 NCSU Libraries Data Science and Visualization Institute or host a Library Carpentry workshop—has the potential for significant benefits, but only when managers 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."*** On top of this, librarians are faced with a growing number of tasks while "keeping the trains running on time." Job descriptions can be inflexible, making technical professional development difficult, which may discourage the acquisition of new skills. The ability of library managers to understand and to value the benefits of in-house data science skills (e.g. to inform decision making and 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, e.g. the Harvard Leadership Institute for Academic Librarians[9], provide mechanisms for developing future senior-level managers, but could also provide a more specific emphasis on managing data-intensive librarians. Managers need to understand how to vertically and horizontally integrate data-centric practices into their organizations and envision the diverse contexts, opportunities, and benefits in applying data science methods. To be an effective manager in the data-driven era, one will need the vision of where library data services are transitioning to, know how to stay ahead of the data science waves, and have good understanding of and skills in data science. Library managers and administrators also 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 developing librarian teams supporting data-intensive communities on campus and beyond. Yet, before embarking on data science programs, library managers and their teams should address the hard work of defining exactly what set of problems involving data need solving.

The following report is a synthesis of the conversations which took place at the Data Science in Libraries Workshop held at the University of Pittsburgh, May 16-17, 2017. The workshop fostered discussions on a variety of topics between a diverse group of attendees including funders, librarians, technologists, administrators, trainers, researchers, and students. Workshop participants expressed a range of perspectives and opinions on the topic of data science in libraries resulting in a complex, multi-dimensional structure. This report synthesizes those conversations and reduces the dimensions and complexity. The following section finishes setting the stage for the report by providing a rough definition of data science and more importantly, focuses on ***data-savvy librarians***. The report then introduces a multifaceted framework that provides recommendations and a means of thinking through the four facets of data science in libraries. There is a discussion focusing on four areas within libraries: culture, scale and assessment, infrastructure, and ethics and values. This report concludes with a roadmap for data science in libraries that we hope can provide some practical guidance for library administrators, practicing librarians, or others interested in data science in libraries.

[1] https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/bigdatardstrategicplan-nitrd_final-051916.pdf

[2] https://ddd-moore.github.io/datasci-at-R1-institutions/

[3] http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/EnvironmentalScan15.pdf

[4] http://altbibl.io/dst4l/

[5] http://web.archive.org/web/20160318110233/http://library.harvard.edu/02042014-1336/harvard-library-offers-data-scientist-training

[6] https://www.lib.ncsu.edu/datavizinstitute

[7] https://www.ands.org.au/23-things

[8] https://librarycarpentry.github.io/

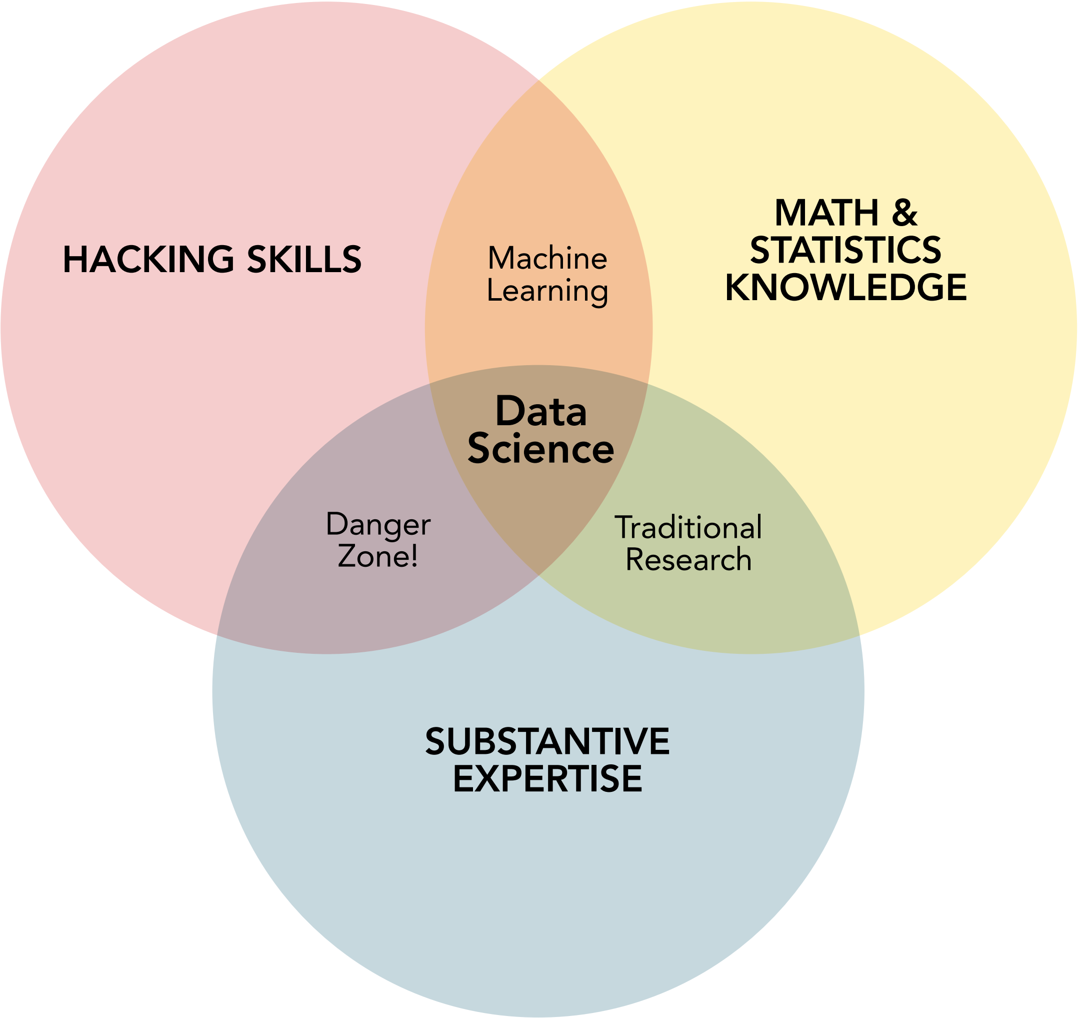
[9] https://www.gse.harvard.edu/ppe/program/leadership-institute-academic-librarians

WHAT IS DATA SCIENCE?

Data science has many definitions, but at the core it is about ***"generating insight from data to inform decision making."***[10] While a full history of the terminology is outside the scope of this report, it is valuable to understand the statistical origins of data science and the arc of its evolution into the ***"sexiest job in the 21st century."***[11] Data science exists on a spectrum and can span work that requires deep statistical and software engineering skills, to work focusing on advocacy, policy development, data management planning, and evidence-based decision making. This report does not provide a ***definitive*** definition of data science, but rather articulates a workable definition, *data savvy*, for the purposes of thinking about data science in libraries.

The term data science was first used by academic statisticians to position the discipline with respect to big data, data analysis, and broader trends. These early discussions emphasized mathematical foundations and the new statistical methods made possible by an abundance of data (Cleveland, 2001; Donoho, 2015). From this academic origin, industrial data science emerged driven by new technology and the ability to extract business value from data.[12] One of the more popular definitions of data science in the industrial sector is Drew Conway's data science Venn diagram.[13]

Figure 1. The Data Science Venn Diagram. Source: Drew Conway



From this perspective, data science is the blending of competencies in computer programming and software engineering (hacking skills), statistics, and a particular domain expertise. This perspective is visible in the many definitions and discussions of data scientists that emphasize the algorithms, machine learning, and statistical techniques. While these areas of expertise are certainly important, the work of data cleaning and preparation is possibly the most important set of skills for data science.[14] Extracting value from data is more than just the rote application of a classification or clustering algorithm; there is a significant amount of "janitor work" involved in any data-centric process.[15] Data preparation, data wrangling, data munging, even linked data workflows (e.g. reconciling data in OpenRefine)—these are all forms of data curation which present an opportunity for librarians to participate in data-centric research.

What is Data Savvy?

A family of data science roles has been identified, which can be characterised by the real-world requirements for actual positions, as described in two related small-scale studies (Lyon & Mattern, 2017; Lyon, Mattern, Acker, & Langmead, 2015). The six roles are: data archivist, data curator, data librarian, data analyst, data engineer, and data journalist. While all of these roles have been framed as data science roles, other framing which comes from the corporate sector has tended to describe only data analyst-type roles as data scientists. However, in reality, there are a wide gamut of roles—**'data savvy**' roles—that orbit within and around the world of data scientists. Data savvy librarians gain familiarity with the datasets, understand technical methods and techniques, and speak multiple disciplinary languages allowing them to work more closely with researchers or the public. Some librarians engage more deeply, becoming technically proficient in data preparation and analysis, allowing them to work with data, automate workflows, and become fully embedded in research teams. In other words, data science exists more or less on a spectrum, depends on an institution's size and mission, and spans work requiring the deep statistical and software engineering skills, to work that focuses on advocacy, policy, communication, and data management. Being data savvy is an essential ingredient of all of these roles.

[10] http://web.archive.org/web/20160304151135/http://christianlauersen.net/2016/01/11/librarians-as-data-scientists/

[11] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[12] https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/

[13] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

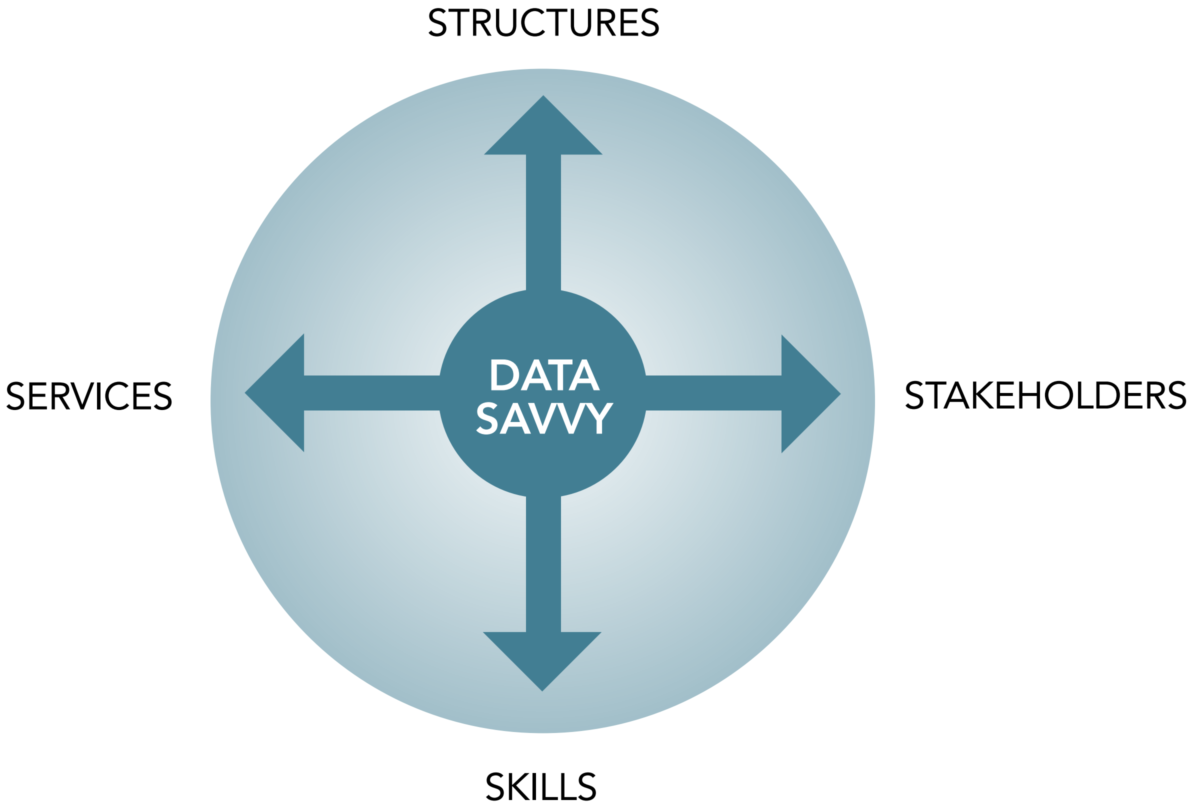
[14] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/

[15] https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html

A MULTI-FACETED FRAMEWORK

This section explores data science in libraries and is oriented around four distinct facets, illustrated in Figure 2 below, that help organize thinking and action around data science in libraries. We examine the relationships between each of the four facets, the particular concerns and challenges as identified by community representatives, and the specific recommendations and actions emerging from these diverse contexts. The commentary is arranged in two main sections connecting the corresponding axes of **structures & skills** and **services & stakeholders.**

Figure 2. Four Facets of Data Science in Libraries



Facets: Structures & Skills

Structures and skills form an axis that connects the individual expertise of data-savvy librarians (the **skills gap**) with organizational structures that can effectively leverage data science (the **management gap**). This axis highlights the need for training at all levels of the profession, from the acquisition of skills by students and particularly early/mid career librarians, to management and visioning by library leadership. It is crucial to understand that these facets are complementary; you cannot train librarians to be data savvy if your organization is not suitably structured with appropriate, equitable incentives and operational models.

Scope

The ability to use data and analytic methods is not only changing library roles, but is leading library and information schools (iSchools) to re-evaluate established practices and approaches to the education of information professionals. A skills gap exists at all levels of librarianship, from the formal training received in library school, to the practices of mid-career librarians, to the leadership of library directors. Addressing the skills gap and helping librarians become data savvy involves both formal and informal training. Formal training of data savvy information professionals should begin in library and information schools, but this may require a repositioning of existing traditional curricula including computer programming and statistical training. Informal training is also crucially important, as many mid-career librarians want to better serve their communities by learning data-centric skills. However, for such educational programs to be effective, the organizational and managerial structures of libraries need to be reconfigured to cultivate and sustain data-savvy librarians. Libraries can foster a space that is permeable and supports greater inclusion of data-savvy expertise and perspectives.

Library leadership must guide their staff and organization to understand the rapidly evolving data-driven landscape and work strategically to ensure that libraries remain a vital resource for their communities. There is also a '**management gap**,' since library managers and administrators must now grapple with how to structure a library to support data-savvy librarians and services. Professional development programs exist for library managers, such as the Leadership Institute for Academic Librarians, but less emphasis is given to managing data-intensive librarians. Managers need to be aware of new roles, and need guidance 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. However there are factors which both facilitate and inhibit implementation.

Drivers & Barriers

There are positive and negative pressures to training data-savvy librarians and maintaining a data-savvy library. Drivers present opportunities for libraries open to the possibility of change, whereas barriers address some of the specific problems implementing such changes to take advantage of new growth areas.

Drivers

- **Changing Research Needs:** Due to the increasing computational nature of research, librarians have an opportunity to respond to a growing need from the research community to learn about new data-centric methods that can improve research productivity and generate new insight. Meeting the researcher's emerging needs can open the door to a librarian's understanding of the underlying discipline-based research in groups and labs and respond to a perceived disconnect with the library community. The research community, in turn, can be a strong advocate for the library. However, this should not only be limited to researchers, as libraries are now grappling with their own research questions from internal systems and collections data.

- **Recognition:** The emergence of research data management requirements from funders, and other data policy-related issues, have led to libraries increasingly being seen as a place to go for help with research data management questions as a collaborator, consultant, facilitator, and/or project manager. There is an opportunity in libraries to grow other data-related services and expertise that leverage existing relationships and awareness of data-centric capabilities

- **Collaboration:** Libraries must think about strategic collaborations or joint initiatives at the local level starting with administrators, IT, the research office, schools, departments, and to the wider community level, including industry, publishers, vendors. Stakeholders outside the library are actively seeking partnership to overcome data-related challenges, so libraries can be seen as a valued partner and collaborator, instead of being seen solely as a service provider.

- **Supply and Demand:** There is a recognized shortage of data-savvy workers across multiple sectors[16] and the enrollments for degrees in data and computer science are surging.[17] When LIS students and librarians learn data skills, a whole new world of job offerings opens to them. Information professionals with library training can address this national and international need.

Barriers

- **Formal LIS Education:** While data skills, like programming and statistics, are a requirement in some MLIS programs, these requirements are not evenly distributed. It can be difficult to "reboot" LIS curricula when LIS faculty do not have expertise and lack broader agreement about what new expertise is fundamental to the profession. MLIS curricula are also constrained by American Library Association (ALA) accreditation standards which can sometimes slow progress towards incorporating data-savvy curriculums.

- **Drive-By Workshops:** The typical three-hour to multi-day workshop is often not enough to yield substantive expertise in data savviness. These drive-by workshops do not have enough depth because of time constraints, nor do they have the structure to do the necessary follow-up to ensure the knowledge "sticks" (Feldon et al., 2017). Embedding a local data-savvy trainer that links training programs with organizational goals and projects might be helpful.[18] Workshop programs such as Library Carpentry are also maturing.

- **Information Overload:** The proliferance of tools and guidance on data science leaves the librarian community overloaded and overwhelmed. Librarians struggle to identify exactly what it is they need to focus on; there are so many tools and learning resources, the multitude of which can be daunting at first. This is a classic problem of resource discoverability; there are so many data science resources out there, but what is good? Librarians ***should*** be able to solve this problem, ***it is what they do!***

- **The Brick Wall:** Practicing librarians may be blocked by constraining or regressive organizational structures and limitations on role responsibilities; ***"I learn new skills, but I still need to do my old job."*** Library staff face competing demands and a proliferation of new and changing services. Job descriptions can be inflexible, making technical professional development difficult. Meaningful and substantive changes are necessary at both the individual and organizational levels.

- **Leadership:** Without specific direction or clear articulation of the end goals of becoming data savvy, the application of data science can be steered in many directions. The challenge is to identify shared, mutual goals for the common good. The descriptions and requirements of library roles can be rewritten to foster a data-savvy culture. However, without an agreed framework to work from or alignment to institutional goals and policies, a number of fears can intervene to impact progress towards integrating data science into library culture. Leveraging data science without an articulated purpose can leave staff in limbo. Equally challenging is if staff are not given the proper resources, (time and money), even if their direction is clear. Workshop participants highlighted the fact that when ***"Google is just down the street, the graduate students leave before finishing their degrees,"*** which creates a brain drain on the profession.

- **Incentive Structures:** Management needs to fully embrace data savvy learning opportunities and understand the effort needed to keep current. Many managers lack an understanding of data science fundamentals and the strategic, organizational approaches necessary to implement it (balancing traditional versus forward looking services), and ultimately, tie it to their library mission and goals. Finally, management can benefit from more use cases and stories as they consider strategic directions. A coordinator plus trainer role, that participates in community data science projects, gathers requirements, develops learning opportunities, and communicates stories to management, can catalyse strategic development.

- **Branding**: While libraries have a long history of being a trustworthy institution and source of information, they struggle with branding. Data science-based services have the potential for convincing researchers, politicians—***"even your mum"*** as one participant mentioned—that the library is a valuable resource in the digital landscape. Libraries must think like startups to go viral and increase awareness of data science services to attract new customers.

Exemplars

Software Carpentry[19] began in 2010 with the aim to teach research programmers better practices in software development. Now The Carpentries, it has expanded to include introductory computing skills and serves as an umbrella for a variety of training programs that help the research community to learn new tools and methods: ***'Software, Data, and Library Carpentry are nonprofit, volunteer organisations that develop training material, train instructors, organise workshops to teach computing and data skills, and support the development of communities of practice.'***[20] The Carpentries has streamlined its approach, reducing tool overload anxiety, bridging the gap between disciplines and roles, and creating community-developed and shared learning materials. Software Carpentry inspired the development of a new branch of The Carpentries, Library Carpentry, which is being supported by IMLS.[21] The Carpentries offers a unique space where multiple communities, from researchers to librarians, can come together, break down barriers, and learn new tools and methods in a lean, iterative fashion. The Carpentries sits at the doorstep of libraries with a curriculum already in place, a worldwide network to learn from, and certificate-based instructor training programs.

Vicky Steeves is the Librarian for Research Data Management and Reproducibility with a dual appointment at NYU Libraries and the Center for Data Science,[22] and is an exemplar of data savviness in action. For example, she is part of the team developing ReproZip, a tool for bundling up all of the software and data dependencies of computational research into an easily distributable and reproducible package.[23] Steeves is a practitioner of ***reproducibility librarianship*** (Vicky Steeves, 2017), and holds a position with the explicit aim of helping researchers with the reproducibility and replication of their research, a *critical* component of their work but one where they need support to do so effectively. Supporting reproducible research requires all the skills of a data-savvy librarian, from understanding digital preservation policies and standards, to unpacking R scripts, to building tools and infrastructure like ReproZip.

Lauren Di Monte's empowered, collaborative work with the research community at NC State, particularly with graduate students on training programs such as ***Web Scraping with Python*** and ***Scientific Computing with Python and Raspberry Pi,*** serve as examples of "win-win" scenarios. Graduate students and the research community benefit by gaining teaching experience, mentorship, and advice via librarians, adding interesting projects to their CVs, receiving monetary compensation and/or recognition, and connecting with students and faculty outside their group or lab. Librarians gain by learning new topics, developing community-driven services, sharing the costs, and discovering new project directions.

Juliane Schneider is the Team Lead for eagle-i (www.eagle-i.net), an open-access Resource Description Framework (RDF) discovery tool for biomedical resources. She has parlayed a traditional cataloging career into a series of jobs supporting data discovery and best practices, from designing databases for EBSCO to metadata consulting/data curation for researchers at Harvard and UC San Diego. She is dedicated to furthering effective instruction in computational skills and data approaches for librarians, being a certified Software Carpentry instructor, and has taught several Library Carpentry workshops. She is also the maintainer for the Library Carpentry regular expression (regex) and OpenRefine lessons.

Training programs such as Data Scientist Training for Librarians (DST4L), and Data Science & Visualization Institute for Librarians (DSVIL), equip librarians to be data savvy. DST4L allows librarians to gain hands-on experience with the research data lifecycle, build practical experience with tools, forward a culture of experimentation, and grow a community of data-savvy librarians. Librarians immerse themselves in learning about data science and visualization in collaboration with academic peers at NCSU Libraries' DSVIL. As such, it provides a comforting environment for skills training and module development, a relatively low barrier to entry for librarians (and their institutions), and presents an opportunity to build a common skills development program with the research community.

The Australian National Data Service (ANDS) led a program called 23 Research Data Things[24] aimed at increasing and strengthening capability and confidence in supporting data services. The 23 training modules, based on the work of the Research Data Alliance's Libraries for Research Data Interest Group, allowed roughly 1.5K librarians to dive into key topics in research data, learn and network with a community of their peers, grow their level of understanding, and gain the confidence to connect with their research communities. The program helped to bridge the disconnect between librarians and researchers, and paved the way for Carpentry-inspired workshops.

Recommendations & Actions

Scan

1. **Highlight Success Stories:** identify library leaders as ambassadors or advisors, to speak about their experiences with data science in libraries (e.g. Mendeley Advisor Program[25]).
    a. Based on the EDISON[26] project's work, develop a suite of materials, (including personas, job descriptions, lifecycle maps, project descriptions, stories, reference material), to assist in planning data science in libraries.

Structures

2. **Collaborate with Leadership Institutes:** such as the Harvard Leadership Institute for Academic Librarians. Learn from programs like EMC's Data Science and Big Data Analytics for Business[27] to establish a (modular) 'data science essentials for library leaders' workshop that includes brief dives into data science topics, hands-on experiences with tools and methods, success stories/case studies, and ample time for discussion, networking, and collaborative opportunities.
3. **Utilize Physical Learning Spaces**: in the library to create a "Data science learning space" where librarians, researchers, and students can come together to learn data science in physical proximity to each other (e.g. a local MOOC Study Group like the Mozilla Study Groups[28] or Harrison Dekker's work at Berkeley's D-Lab[29] and University of Rhode Island Libraries AI Lab[30]).

Skills

4. **Advocate for The Carpentries** (Software, Data, Library, High Performance Computing): to address the computational skills gap in the research, teaching, and learning community, using the library as a hub for Carpentry training programs and activities.[31]
    a. Partner with The Carpentries (leveraging their experience and assessment data) to develop a 1-5 slide pitch for library leaders and staff to use in building the case for data science in libraries at the local level and community-wide.
    b. Better understand the effectiveness of bootcamp-like training programs (see Feldon et al., 2017; Word, 2017) by leveraging local assessment expertise together with The Carpentries expertise (http://www.datacarpentry.org/assessment/), especially Dr. Kari L. Jordan [32] , Deputy Director of Assessment at Data Carpentry.
5. **Leverage Existing Educational Resources:** take advantage of existing open and self-directed data science educational resources, from Open Data Science Masters[33] to DataCamp[34], instead of reinventing new resources solely for libraries. Furthermore, a lightweight metadata-standard for training materials, like CodeMeta,[35] along with a centralized clearinghouse (e.g. The Journal of Open Source Education[36]), could improve discoverability of open educational materials.
6. **Reposition the MLIS:** modernize library and information school programs to be more data-centric and oriented towards the changing research and societal needs. Plug into the recent IMLS-organized ***Positioning Library and Information Science Graduate Programs for 21st Century Practice***[37] where this conversation has already started.

Facets: Services & Stakeholders

This section presents a complementary perspective to structures (managing data science people: data librarians, data stewards and data science teams), and skills (competencies, education and training), by focusing on data science services and stakeholders. We begin by considering the range of applications for data science methods, tools, and workflows in a library context, and the communities and individuals who will use them.

Scope

Libraries create, collect, store, use, re-use, and preserve data as integral elements within their service workflows. For the purposes of this section, we can usefully differentiate between: a) data generated internally (in-house data), such as library system transactions, workflow data, legacy data collections, patron records, borrowing and circulation data, library monograph and journal expenditure data, footfall data, biblio/webmetric data etc.; and b) data created, collected, accessed, or used externally (in other departments or organizations) by library patrons such as researchers, student learners, administrators, the public /citizens, e.g. research datasets, learner progress data, government data, but which may potentially act as a substrate for innovative data science services provided by librarians.

Internal

For library directors and operational managers, there is a requirement for timely and insightful management information or business intelligence derived from library data. One example comes from considering Collections as Data[38], and analyzing these data to provide valuable insight into selection, purchasing, circulation, location, development, preservation, and disposal decisions for stock items.

External

Within academia, data science has pervaded almost every discipline and data-savvy skills are used in many professions to increase efficiencies and gain insight. This means researchers and students on campuses need support as they learn these skills and use them in educational programs, research projects, or personal explorations. Keeping up with the evolving needs of the user communities it serves is an important component of what academic libraries do, so equipping librarians with data-savvy skills increases their ability to support this data-focused research. There are also opportunities to support research data curation through cleaning, manipulating, and wrangling datasets using tools such as OpenRefine and analyzing research data using tools such as Python, Jupyter, R, and RStudio, and with interpreting data by creating visualizations to illustrate trends and anomalies using tools such as Tableau and D3. While many libraries are not yet ready to take on a large data science role on campus, there are many opportunities to partner and contribute to an emerging data science support network.

In addition to working with scholarly datasets, librarians now have the opportunity to add value to research data and civic or government or administrative data through new roles as data curators, providing a range of research data management services, such as reviewing data management plans, preserving datasets for the long-term in institutional repositories, tracking the re-use of datasets with persistent identifiers through citation and download metrics, and supporting the creation, interpretation, and use of civic data by the general public.

However, the ability of libraries and librarians to implement these new types of data function and service are influenced by a number of factors.

Drivers & Barriers

Drivers

- **Physical Space:** the library acts as an increasingly shared and well-trafficked space[39] on campus; it can be viewed as the literal heart of campus, both intellectually and geographically. It can be viewed as a place for exposure to innovations and creativity and provides proximity to people, community partners, and collaborators. The library can play a significant role in building a community around data science, and can catalyze new data partnerships (e.g. NCSU Libraries Peer Scholars Program[40]).

- **Facilitating Inclusivity:** the library is representative of the broader community, creating a welcoming environment where people feel comfortable asking questions and working with tools and technology.

- **Cost-Neutral Facilities:** libraries subscribe to research tools and services and curate open resources for researchers.

- **Strategic Planning:** in order to make informed decisions in annual planning, senior library managers need to have tangible evidence to validate operational decision making, future investments, and staff deployments. This evidence can be gathered from data collection, data analysis, and subsequent insight. If longitudinal data analysis is available—i.e. data trends are monitored over time—then some degree of prediction may also be achievable.

- **Problem-solving:** data science approaches may shed light on otherwise hard-to-see or misunderstood problems in the library. For example, analysis of library footfall data can shed light on difficult resourcing challenges, when and if to staff information desks, how to deploy library shelving, and how to identify optimal opening and closing times using tools such as Suma.[41]

- **Stakeholder Demand:** researchers are increasingly using data science in their fields. There is a growing demand for training programs demonstrated by the rise of The Carpentries. Librarians could be able to better support researchers in their work. Data science is used in classes and many students will use data science skills in the world of work, so delivering services to early career researchers and students will benefit them.

- **Cross-Disciplinary Facilitation:** librarians can act as facilitators bringing multi-disciplinary teams together around a common data set.

- **Re-engineering Services:** transitioning from transactional and hybrid service delivery models to an immersive model, where librarians are integrated within disciplinary research laboratory teams and provide a range of data science services e.g. data analysis, data visualization, scripting and coding, etc. Pilot programs may offer a great opportunity for staff to trial these approaches, collect data on feedback and outcomes, and work with management to potentially iterate on or implement the findings.

- **Impact Metrics:** institutional administrators are collecting metrics, performance indicators, and other evidence to demonstrate impact, to raise profile, and to influence national and international rankings.

Barriers

- **Silo Effect:** the library may function as a silo (and be viewed as such by its data stakeholders) with librarians wholly located remotely from their patrons, e.g. the research community. This physical disconnect can act as a primary barrier to building community partnerships around data; alternatively the library may be viewed as a *fortress*[42] with librarians being hidden amongst the dusty shelves, presenting an incorrect image of the library to its 21st-century users!

- **Scale:** many library research data management services are provided on a one-to-one or individual consultancy basis led by demand. Extending these types of service to production-level operations which are routinely available across whole departments or schools, presents additional challenges of scaling-up services at a time when staff resources (in particular time and attention) are limited.

- **Resources:** libraries are often making decisions about what to stop doing or what to move to a lower priority in order to fit in new services with limited bandwidth. Libraries need to be realistic and ask what data science roles library staff can fill with local existing resources. This work requires time and possibly embeddedness and could require a need to react quickly. That may be hard to scale.

- **Credibility and Image:** the library may not be seen as the place to go for data science support. Finding the right campus partners who will want to work in concert with other departments to look holistically at the user needs of those working with data may be challenging.

- **Experience:** Librarians are often not doing research themselves, so they do not have first-hand knowledge of the research lifecycle in practice.

- **Culture:** While some organizations, such as the University of Southampton, have trialed the embedded librarian concept, there were challenges and one particular perspective was that this model was perceived to have "broken rules." However the librarians liked leaving the library and engaging more closely with researchers in departments. One possible answer may be via inter-business units secondments and innovative data stewardship programs like TU Delft's.[43]

Exemplars

The Library of Congress (LoC) Lab Technical Pilot provides an excellent exemplar working with in-house data which was unavailable to the public, and which demonstrates the feasibility of this approach. The LoC used a third-party platform to analyze collections data, developing Python scripts and format conversion workflows to dig deeper into the data. The pilot showed that technical platforms, legal issues, and cost planning were all required for a successful data lab.[44]

Another exemplar is the ongoing work at CalTech where they are seeking to integrate Jupyter notebooks into library collections. The Jupyter integration into library workflows was outlined at the Coalition for Networked Information Fall 2017 Project Briefings talk, "From Stock to Flows." [45]

At the University of Illinois at Urbana-Champaign (UIUC), there is a joint project with the Hathi Trust Research Center to facilitate textual analysis of large-scale computational datasets in order to help researchers understand how to use datasets through collaborative workshops. The Hathi Trust Digging Deeper Project is focusing on developing a curriculum and training librarians to ***"assist librarians in getting started with the tools, services, and related research methodologies of the HathiTrust Research Center (HTRC)."***[46]

At the Carnegie (Public) Library, Pittsburgh (CLP), data librarians have been helping the public to use/re-use open government/civic data, which are accessible from the Western Pennsylvania Regional Data Center (WPRDC[47]). The Library has run a data day with the aims of changing the public perception from open data to public data, and enhancing public data literacy. Additionally, CLP runs a data-centric speaker series with a community bridging focus that has brought speakers like data scientist and author Cathy O'Neil or information designer and data artist Giorgia Lupi.

Recommendations & Actions

Scan

7. **Perform a Data Audit:** for your library using the Data Asset Framework methodology[48] to identify the intra-library (internal) legacy data, e.g. transactions/collections/operations, which are of most value.

8. **Identify External Datasets:** e.g. research data, open government data, where there is scope to add value through diverse library-mediated services, tools, and infrastructure.

9. **Discover Data Science Requirements:** at each stage of the research data lifecycle and build and deploy services to support them.

Services

10. **Pilot Data Services:** plan, run, and evaluate a range of pilot data services at different scales to assess feasibility, impact, benefits, and return on investment.

11. **Quantify Resources to Scale Up:** for data science services from pilot to production level operations, by learning from pilots, assessments, and experiences in other libraries.

12. **Re-engineer Services:** explore re-engineering existing research data services by embedding data librarians within multi-disciplinary research teams in departments.

Stakeholders

13. **Identify Campus Stakeholders:** initiate conversations about who is doing what. Find partners on campus and build or be part of a data science support network.

14. **Outreach to Researchers:** work alongside passionate researchers to provide proven services and training (e.g. The Carpentries) that are effective for busy researchers to improve or extend data management, software management, code authoring, and publishing.

15. **Showcase Capacity and Capability:** showcase library data science capacity and capability through targeted messaging and project partnerships to influence funders' perceptions of libraries and to raise expectations for future research collaborations.

[16] http://www.bhef.com/publications/investing-americas-data-science-and-analytics-talent

[17] https://cra.org/cra-releases-report-surge-computer-science-enrollments/

[18] https://projects.ncsu.edu/mckimmon/cpe/opd/dl/PS3A.pdf

[19] https://software-carpentry.org/

[20] https://software-carpentry.org/blog/2017/09/merger.html

[21] https://www.imls.gov/grants/awarded/RE-85-17-0121-17

[22] http://library.nyu.edu/people/victoria-steeves/

[23] https://www.reprozip.org/

[24] http://www.ands.org.au/partners-and-communities/23-research-data-things

[25] https://community.mendeley.com/

[26] http://edison-project.eu/

[27] https://education.emc.com/guest/campaign/business_transformation/datascience_business_transformation.aspx

[28] https://science.mozilla.org/programs/studygroups

[29] http://dlab.berkeley.edu/

[30] https://today.uri.edu/news/uri-to-launch-artificial-intelligence-lab/

[31] https://www.cni.org/topics/teaching-learning/software-carpentry-in-the-library-partnering-to-give-researchers-needed-technical-skills

[32] https://twitter.com/DrKariLJordan

[33] http://datasciencemasters.org/

[34] https://www.datacamp.com/

[35] https://github.com/codemeta/codemeta

[36] http://jose.theoj.org/

[37] https://www.imls.gov/news-events/events/positioning-library-and-information-science-graduate-programs-21st-century

[38] https://collectionsasdata.github.io/

[39] https://qz.com/672706/why-the-internet-hasnt-killed-the-library-yet/

[40] https://www.lib.ncsu.edu/news/students-teach-students-in-the-peer-scholars-program

[41] https://www.lib.ncsu.edu/projects/suma

[42] https://christianlauersen.net/2015/08/11/the-fall-of-the-library-fortress/

[43] https://openworking.wordpress.com/2017/08/29/data-stewardship-addressing-disciplinary-data-management-needs/

[44] http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf

[45] https://www.cni.org/topics/digital-curation/from-stock-to-flows

[46] http://teach.htrc.illinois.edu/

[47] http://www.wprdc.org/

[48] http://www.data-audit.eu/

DISCUSSION & REFLECTIONS

Four common themes have emerged from the DSinL workshop and associated activities, which are further unpacked in this section by exploring tensions, challenges, and opportunities.

Culture

The cultural implications of data savviness can be considered in two ways: a) professional culture (people, roles, credentials); and b) organizational culture (leadership, management, communication).

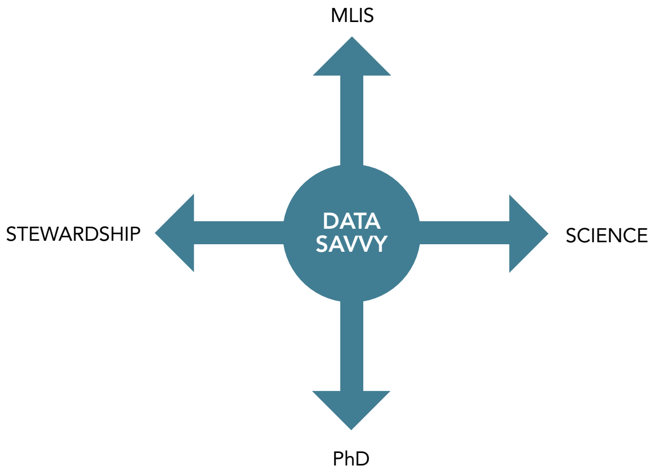
Professional Culture

One of the most controversial topics raised in discussion with workshop participants was the relative value of different qualifications or credentials as educational trajectories toward data science roles in libraries. This conversation encompassed the value of formal versus informal education (multi-year degrees or short courses) and the benefits from recruiting individuals with a deep disciplinary knowledge (as evidenced by a doctoral qualification, i.e. people who have a Ph.D.), compared to graduates with a broader knowledge (demonstrated by a Master's degree, i.e. people who have an MLIS or similar credential). Programs like the Council on Library and Information Resources (CLIR) Postdoctoral Fellowship program (Waraksa in Maclachlan et al., 2015) have inspired debate by institutionalizing a pathway for Ph.Ds into academic libraries.[49] We term this educational continuum: the '**credentialing tension**.'

In parallel, a family of data science roles and competencies (Lyon & Mattern, 2017; Lyon et al., 2015; Semeler, Pinto, & Rozados, 2017), have been grouped into two broad categories based on real-world requirements: stewardship-oriented roles which focus more on policy, data management planning, and advocacy (data librarian, data archivist, data curator); and science-oriented roles which focus on more technical workflows to manipulate and wrangle the data to produce novel insights (data analyst, data engineer). There are tensions between stewardship and science activities in libraries, with the former being much more established as research data management or research data services than is the latter (Cox, Kennan, Lyon, & Pinfield, 2017). We have termed this the '**re-engineering tension**.' However, both stewardship and science represent valid and beneficial 'data curation' activities, using the following definition from the UK Digital Curation Centre: '*curation is adding value to data.*'[50]

We propose that data-savvy librarians occupy a space at the intersection of these competing tensions (see Figure 3). Furthermore, these data-savvy individuals may offer hybrid or blended skills, e.g. disciplinary knowledge plus experience of curating data as stewards and/or scientists. The value of blended skills has been highlighted as desirable by a PwC report on investing in America's data science and analytics talent and underpinning the earlier concept of T-shaped information professionals.[51]

Figure 3. The Cultural Tensions of Data-Savvy Professionals in Libraries



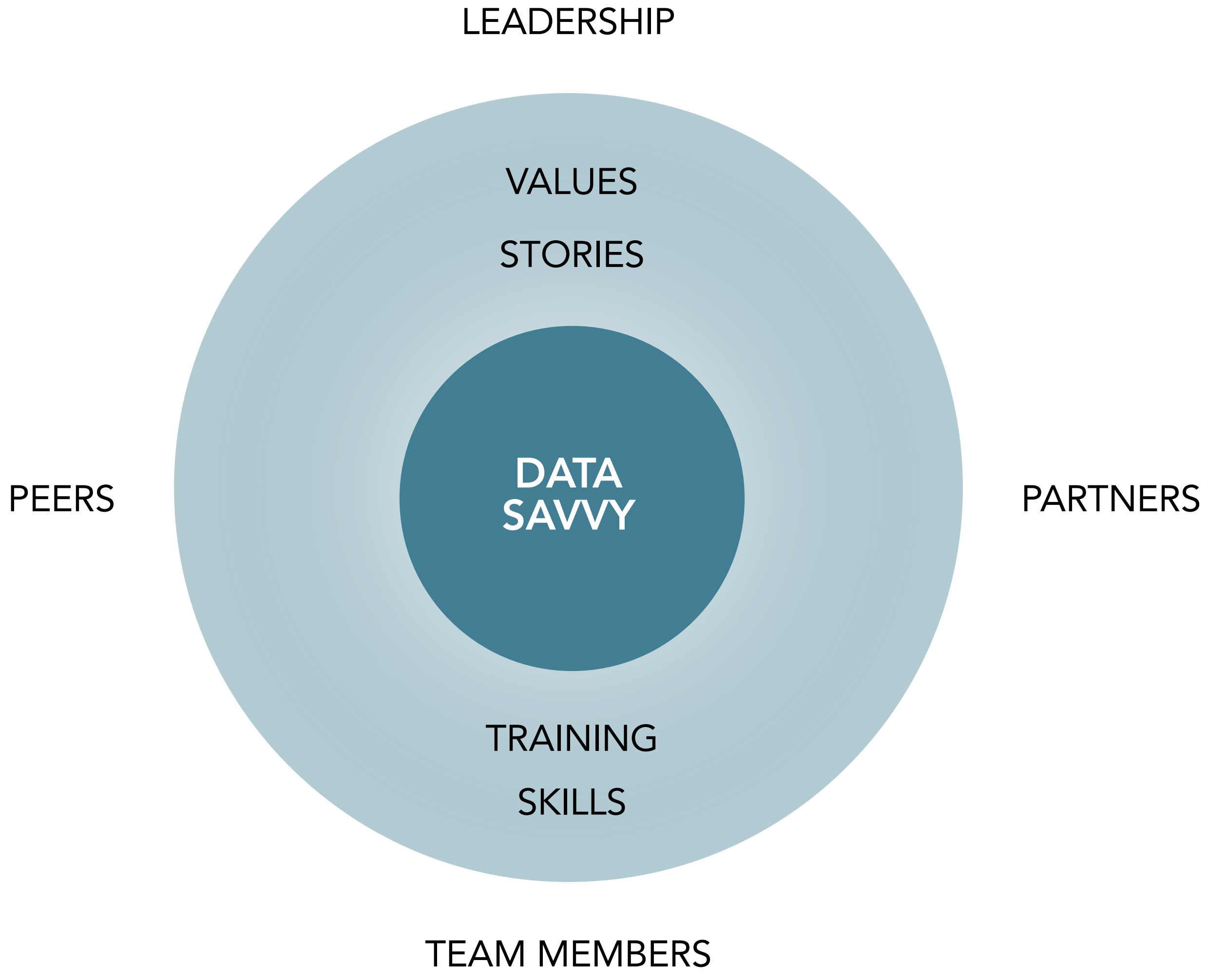
Organizational Culture

The management and nurturing of data-savvy librarians within a supportive organizational culture, provides both challenges and opportunities for stakeholders. Firstly a **data-savvy 360 model** of engagement is presented in Figure 4, where data-savvy roles are: 1) envisioned and proactively implemented by library directors and administrators, 2) supported by a peer-to-peer network both within a library and across the broader library community, 3) integrated within existing teams as key players and enablers, and 4) valued by internal collaborating departments and by external partner organisations in industry and business, as well as by the wider public.

A data-savvy library organization based on the 360 model can be characterized by the: 1) routine collection and strategic application of quantitative evidence, 2) effective communication and messaging driven by data-rich stories, 3) established professional education programs to extend and expand data savvy skills, and 4) an explicit understanding and articulation of the value and benefits of science and stewardship roles and responsibilities.

Clearly while there are challenges for libraries in achieving these (aspirational) goals, there are also exciting opportunities to transition and transform libraries into 21st-century, data-savvy spaces. Realistically there is a continuing requirement for advocacy, ambassadors, or champions and success stories to promulgate the data-savvy concept; the Data-Savvy Roadmap seeks to provide a framework for this process.

Figure 4. The Culture of Data-Savvy 360 Model Organizations



Recommendations:

16. **Acquire Blended Skills:** data science education and professional development programs should aim to facilitate the acquisition of a blended suite of skills, including disciplinary, computational/quantitative/technical, and soft skills. No one person can have all these skills, so developing teams with these complimentary skills will be invaluable.

17. **Implement Data-Savvy 360 Model:** organizations should implement full 360-degree engagement with data-savvy concepts to nurture new data roles and ensure that there is both vertical and horizontal communication with library managers, peers, and team members.

Scale & Assessment

The nature of data science work may be demanding, often requiring a significant degree of hands-on attention and deep disciplinary knowledge, particularly when it comes to cleaning and wrangling with data. The team-based approach to data science is helpful in gaining an understanding of data science projects and librarians can benefit from being embedded in groups or labs at institutions. Unfortunately this approach will inevitably run up against the scaling challenge to meet many diverse needs across research institutions. There are simply not enough librarians. Alternatively, community-based programs such as the Carpentries approach the scaling challenge by collectively determining the baseline computational needs of the research community via modular training material and events. Since the Carpentry programs are guided by members of the research community from various disciplines, their services can be broadly applied at different institutions, labs, groups, and worldwide. To scale data services in libraries, librarians must consider how to manage and bridge their embedded work so that it informs and feeds into broader, community-based approaches. Additionally, librarians must address how their review and assessment at the local level links up at the community level in order to form a more complete understanding around the impact of their data science activities. Taking a rigorous formative (at the start) and summative (at the end) approach to evaluation can be very effective in demonstrating impact.

Recommendation:

18. **Conduct Rigorous Assessment:** ensure that embedded data science-related feedback, assessment, and service evaluation data are collected and shared at the local and institutional level and with community-based training programs (i.e. The Carpentries).

Infrastucture

Supporting data and computational-intensive research requires more than expertise and data savviness. Data science cannot be separated from its infrastructural or cyberinfrastructural (Atkins, 2003) dimensions; extracting value from data implies access to technical systems for the acquisition, curation, analysis, and interpretation of data. The explosion of tools and software for working with big data, driven in large part by industrial open-source software development, creates an opportunity for libraries (though open-source tools can have very high resource costs in implementation and maintenance that are often not recognized when adopted). Supporting data science is a team sport and libraries do not and should not need to develop and maintain data infrastructure alone.

Libraries should work closely with IT organizations and, when present, research computing/high-performance-computing facilities to support data science in their communities. Libraries can bring their existing expertise in data management and their community-centric perspective to complement the technical and infrastructural expertise of IT and research computing. Often these organizations struggle to engage their communities (students, researchers, users, etc.) because of language barriers (jargon) and cultural differences (IT-centricity). Librarians can help bridge the gap between IT professionals (who manage and maintain data infrastructure) with the researchers whose area of expertise may be more narrowly focused on disciplinary concerns.

The National Science Foundation-funded Advanced Cyberinfrastructure Research and Education Facilitators (ACI-REF)[52] is an emerging community of cyberinfrastructure facilitators who come out of high-performance computing with the explicit goal of helping researchers use advanced cyberinfrastructure resources more effectively. This nascent community would benefit greatly from partners in the library who already have deep expertise in facilitating research, managing relationships, and educating researchers about the resources available to them.

Recommendation:

19. **Build Data Partnerships:** librarians should learn from the collaborative relationships built within the Moore-Sloan Data Science Environments[53] and partner with the ACI-REF community and the emerging Campus Research Computing Consortium.[54] Libraries can be effective intermediaries to enable fruitful collaborations like humanities scholars and high-performance computing (Terras et al., 2017).

Ethics & Values

When exploring the needs of researchers doing data science, we learn that there are many unanswered ethical questions and concerns and that no single department is tasked to triage or support researchers in addressing on most university campuses. New and complex datasets raise challenging ethical questions about risk to individuals that are not sufficiently covered by data science training, ethics codes, or Institutional Review Boards. The use of publicly available, corporate, and government data sets in research projects may reveal human practices, behaviors, and interactions in incidental or unintended ways, creating the need for new kinds of ethical support. Researchers and students using these data in their research are navigating issues and making ethical decisions in ways that are not taught in their discipline. Many have only their peers to turn to for difficult questions that could have long-term impacts on their research or reputation.

Research librarians are one set of actors within a support network that university researchers rely on during their research. The values of privacy, ethics, and equitable access to information are core to librarianship, making librarians a unique partner for researchers and for others who are part of the campus support network. At the same time, there is an international conversation about ethical use of data that the field should participate in. There is an opportunity for librarians to find their role in supporting researchers by navigating emerging ethical issues in their research and especially by coordinating efforts with those business units within universities that have responsibility for ethics and research integrity.

Recommendations:

20. **Consider Ethics and Privacy:** in their training programs, librarians should consider the ethical and privacy implications based on the discipline or research area.

21. **Offer Triage Services:** librarians can take on the role of providing triage services for researchers who are unclear where to turn (e.g. a referral map of research support services across the institution). Librarians can offer a network of support and provide background legwork to save researchers time. They should expand this role to include help with privacy and ethical decision-making as it comes up in their research.

22. **Develop Libraries as Amplifiers**: libraries can host a series on ethical topics in research or partner with the work of campus organizations such as ethics institutes like Carnegie Mellon's Center for Ethics & Policy [55], technology and society research units (e.g. University of California Berkeley's Center for Science, Technology, Medicine & Society [56]), or cybersecurity initiatives with concerns about privacy.

[49] http://acrlog.org/2006/10/16/clirs-program-a-real-or-imagined-shortage-of-academic-librarians/

[50] http://www.dcc.ac.uk/digital-curation/what-digital-curation

[51] https://en.wikipedia.org/wiki/T-shaped_skills

[52] https://aciref.org

[53] http://msdse.org/files/Creating_Institutional_Change.pdf

[54] https://www.nsf.gov/awardsearch/showAward?AWD_ID=1620695

[55] http://centerforethicsandpolicy.com

[56] http://cstms.berkeley.edu

THE DATA-SAVVY ROADMAP

FACET		
Recommendation & Action		
SCAN		
Highlight success stories Discover data lifecycle requirements	Perform data audit	Identify external datasets
STRUCTURES		
Collaborate with leadership institutes	Utilize physical learning spaces	
SKILLS		
Utilize existing educational resources Ethics and privacy	Advocate for software carpentry Reposition the MLIS	Blended skills
SERVICES		
Share assessment data Libraries as amplifiers	Pilot data services Triage services	Resources to scale-up Re-engineer services
STAKEHOLDERS		
Identify campus stakeholders Outreach to researchers	Showcase capacity and capability 360-degree data-savvy model	Build data partnerships

Roles & Stakeholders

- Leadership
- Librarians
- Educators
- Researchers
- The Public
- Students
- Funders
- Campus
- Industry

NEXT STEPS FOR DATA SCIENCE IN LIBRARIES

We can envision several steps to take Data Science in Libraries to Phase 2. First, it will be important to create a space for sustaining a conversation around data science and libraries in order to continue dialogue among current stakeholders and to grow interest and commitment by the broader library community. This community-building work should include holding additional open workshops or regional events framed as a **Data Science in Libraries Community Events**. Furthermore, we think an annual international conference will also help to spread the message.

Second, it is important to begin working on coordinating and aggregating information about data-savvy training and education initiatives, i.e. a **Data Science in Libraries Learning Registry**. We currently see three clear levels of training that can or are happening. The first is lightweight guides and aggregated resources, for example the Training and Resources Guide for the World Data System[57], but with an expanded scope and audience, or the newly launched Journal of Open Source Education (JOSE).[58] These guides can help libraries and individuals find already created resources and examples that can be tailored to a local setting. The next is medium-weight programming like the Carpentries and other informal professional development workshops or boot camps. These allow individuals and libraries to have a more structured event with well-tested materials and a trained instructor. The third level is the formal master's degree or certification education. This level is for someone who wants a formalized degree that provides in-depth training and expertise alongside professionalization. The next step will be to assemble training materials specifically tailored for data science in libraries. The emphasis is not the development of new materials, but rather the curation of existing materials from a variety of sources.

Finally, we also see a great opportunity to work with management and leadership institutes to create training materials for managing data science in libraries, to showcase our **Data Science in Libraries Report** to Library leaders, e.g. at ARL meetings, and to iSchool Deans as components of re-positioned educational programs. This conversation is only just beginning.

[57] https://www.icsu-wds.org/services/training-resources-guide/training-resources-guide

[58] http://jose.theoj.org

REFERENCES

Atkins, D. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review = Revue Internationale de Statistique*, 69(1), 21–26. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1751-5823.2001.tb00477.x/full

Cox, A. M., Kennan, M. A., Lyon, L., & Pinfield, S. (2017). Developments in research data management in academic libraries: Towards an understanding of research data service maturity. *Journal of the Association for Information Science and Technology*, 68(9), 2182–2200. https://doi.org/10.1002/asi.23781

Donoho, D. (2015). 50 years of Data Science. In *Princeton NJ, Tukey Centennial Workshop*. Retrieved from http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

Feldon, D. F., Jeong, S., Peugh, J., Roksa, J., Maahs-Fladung, C., Shenoy, A., & Oliva, M. (2017). Null effects of boot camps and short-format training for PhD students in life sciences. *Proceedings of the National Academy of Sciences of the United States of America*, 114(37), 9854–9858. https://doi.org/10.1073/pnas.1705783114

Maclachlan, J. C., Waraksa, E. A., & Williford, C. (2015). The Process of Discovery: The CLIR Postdoctoral Fellowship Program and the Future of the Academy. Council on Library and Information Resources. 1755 Massachusetts Avenue NW Suite 500, Washington, DC 20036.

Lyon, L., & Mattern, E. (2017). Education for Real-World Data Science Roles (Part 2): A Translational Approach to Curriculum Development. *International Journal of Digital Curation*, 11(2), 13–26. https://doi.org/10.2218/ijdc.v11i2.417

Lyon, L., Mattern, E., Acker, A., & Langmead, A. (2015). Applying Translational Principles to Data Science Curriculum Development. In *iPres 2015 Proceeding*s. Chapel Hill, North Carolina. Retrieved from http://d-scholarship.pitt.edu/27159/

Semeler, A. R., Pinto, A. L., & Rozados, H. B. F. (2017). Data science in data librarianship: Core competencies of a data librarian. *Journal of Librarianship and Information Science*, 0961000617742465. https://doi.org/10.1177/0961000617742465

Terras, M., Baker, J., Hetherington, J., Beavan, D., Zaltz Austwick, M., Welsh, A., … Farquhar, A. (2017). Enabling complex analysis of large-scale digital collections: humanities research, high-performance computing, and transforming access to British Library digital collections. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx020

Vicky Steeves, N. Y. U. (2017). Reproducibility Librarianship. *Collaborative Librarianship*, 9(2), 4. Retrieved from http://digitalcommons.du.edu/collaborativelibrarianship/vol9/iss2/4/

Word, K. R. (2017, December 11). When Do Workshops Work? Retrieved February 5, 2018, from http://www.datacarpentry.org/blog/reponse-to-null-effects/

ACKNOWLEDGEMENTS

The authors of the report wish to extend their sincere appreciation to all the institutions and contributors who shared their expertise and experience with us during the preparation, editing, and design of this report. Our special thanks go to the participants of the May 2017 workshop which gave the authors access to a broad range of ideas and set the stage for the themes developed in this document.

"Data Science in Libraries" was made possible by the Institute of Museum and Library Services grant number RE-43-16-0149-16.

Institutions

Institute of Museum and Library Services, University of Pittsburgh, Data & Society, North Carolina State University

Workshop participants and reviewers

Matt Burton (University of Pittsburgh), Liz Lyon (University of Pittsburgh), Chris Erdmann (North Carolina State University), Bonnie Tijerina (Data & Society), Tracy Teal (Data Carpentry), Vernica Downey (Harvard University), Andrew Odewahn (O'Reilly Media), Lauren Di Monte (University of Rochester), Honora Eskridge (North Carolina State University), Carl Stahmer (University of California-Davis), Dan Chudnov (Chudnov Consulting), Suchana Seth (Data & Society, Mozilla), Harriet Green (University of Illinois Urbana–Champaign), Alix Keener (University of Michigan), Steve Brewer (EU Edison), Kathryn Unsworth (Australian National Data Service), Jeannette Ekstrøm (Technical University of Denmark), James Baker (Library Carpentry), Josh Greenberg (Sloan Foundation), Amanda Rhinehart (Ohio State University), Ashley Sands (Institute of Museum and Library Services), Gail Clement (California Institute of Technology), Ian Cook (Highmark Health), Keith Webster (Carnegie Mellon University), Eleanor Tutt (Carnegie Library of Pittsburgh), Aaron Brenner (University of Pittsburgh), Tim Dennis (University of California Los Angeles), John Chodacki (California Digital Libraries), Jian Qin (Syracuse), Alistair Croll (Harvard Business School), Jan Brase (SUB Gottingen), Nic Weber (University of Washington), Meredith Schwartz (Library Journal), Renae Barger (University of Pittsburgh), Kornelia Tancheva (University of Pittsburgh), Jessica Kirchner (University of Pittsburgh), Chris Mills (University of Pittsburgh), Lauren Murphy (University of Pittsburgh), Juliane Schneider (Harvard Catalyst)

Cover photo courtesy of Christian Lauersen.