Creating a Data Science Framework: A Model for Academic Research Libraries

Nandita S. Mania, Michelle Cawley, Amanda Henley, Therese Triumph, and Joe M. Williams

Abstract
The University Libraries participated in a pan-campus planning exercise to determine the feasibility of a comprehensive data science (DS) initiative. In addition, Library Leadership set forth a charge to develop a DS framework to explore and articulate how the Libraries could add value to campus DS efforts. This article provides a model of how a large research university library system developed a DS framework by examining current and future services, skills, structures, stakeholders, and put forth recommendations around DS engagement for the University of North Carolina at Chapel Hill community. This framework could be applied to other academic departments and research support units.

Keywords
Data science; libraries; partnership; data science planning; framework; research library

Background
The University of North Carolina at Chapel Hill’s (UNC-CH) strategic framework, Carolina Next, lists eight pillars of focus for the UNC campus; many of which include engagement with and about data (Carolina Next: Innovations for Public Good, n.d.). To note, the second pillar, Strengthening Student Success, mentions the creation of a Data Science (DS) school to help identify opportunities to facilitate experiential and collaborative learning around the use and application of data. During Fall 2019, a pan-campus planning exercise was initiated by Executive Vice Chancellor and Provost, Dr. Robert A. Blouin, to determine the feasibility of a comprehensive DS initiative at UNC-CH (Data Science @ Carolina). Key concepts identified to guide this work included an interdisciplinary approach that strengthens existing cross-campus collaboration, the need for the initiative to represent the entire data life cycle, and for DS efforts to value the human impact DS has on individuals and can contribute to the public good. The vision of the
UNC-CH University Libraries (Libraries) is to set the standard for a public university academic research library in the digital age. As such, the Libraries aims to partner across campus to transform the learning environment and address questions pertinent to community initiatives and priorities. The Libraries, one of the largest research libraries in the South, is an essential part of the University’s educational and research efforts and is comprised of nine libraries on campus, including the Health Sciences Library, Walter Royal Davis Library, Louis Round Wilson Special Collections Library, and the R.B. House Undergraduate Library.

Data services at the Libraries grew from the reference department, with a data services section established in 2008. Since then, services have grown considerably and are offered across multiple libraries. The Libraries offers services related to data management, visualization, impact measurement, data preservation, support of FAIR data practices, and consultations to assist with locating, acquiring, preparing, analyzing, and visualizing data. Additionally, the Libraries offer a robust series of workshops, with topics including bioinformatics, visualization, GIS, Python, R, and Excel. Librarians with expertise in preparing and analyzing data serve as partners and coauthors on research projects, publications, and grants. The Libraries’ engagement around data provides a strong base for growth. Library staff who currently provide data support are well-positioned to anticipate DS needs and engage with campus partners.

The Libraries has placed DS as a strategic priority for the coming years with the aim to bolster current data-related programs and services and create opportunities to address new skills, expertise, and partnerships required to transform the research library in the digital age. The Libraries’ mission is to “…help our users find, evaluate, and use the information to create knowledge.” The evaluation of data falls under this mission and further exemplifies the importance of DS as one of the core strategic priorities. For this reason, Vice Provost for University Libraries and University Librarian, Elaine Westbrooks, along with library staff, participated on the Data Science @ Carolina pan-campus committee. The inclusion of librarians on those committees highlights the relevance and the extent to which libraries support students, faculty, and staff with knowledge of data curation, manipulation, and evaluation as agnostic members of the University. A research library in the digital age requires data-savvy staff that has an explicit understanding of the use, value, and benefits of data-intensive research. In addition, it is anticipated that shifts will occur in how some library roles are becoming DS roles (e.g., data librarians, data analysts, engineers, and archivists) that specialize in data curation, data management, visualization, GIS, open science, machine learning, artificial intelligence, and reproducibility.

In tandem with the pan-campus DS effort, Libraries Leadership set forth a charge to develop a DS framework designed to explore, mobilize, and articulate how the Libraries may add value to campus DS efforts, including how to communicate impact. The framework examined current and future library services, skills, structures, and stakeholders and put forth recommendations for providing DS services, technologies, and spaces for the UNC-CH community. To establish the Libraries DS Committee (LDSC) and envision the design of the framework, the following overarching principles were integrated:
- **One Library**: Library work reflects and achieves an aligned approach across all libraries. Library staffs are connected and unified where staffs seek to eliminate silos to optimize user experience across the entire library system.

- **Alignment**: Investment in data science will be aligned with the Libraries strategic goals and priorities.

- **Universal Design & Open Access**: Environments are created that are usable by all people and aspire to open data principles.

- **Innovation**: Innovation will be achieved by solving problems through encouraging inquiry, leveraging differences, and integrating diverse perspectives.

- **Growth & Development Mindset**: Acknowledging the continuous pursuit of learning, challenges will be seen as opportunities.

To build the LDSC using a One Library approach, a diverse group of staff was brought together from across the libraries with expertise in the fields of DS, programmatic design and evaluation, software development, digital collections, archives, and digital humanities. Upon establishment of the LDSC, how other R1 (Carnegie Classification of Institutions of Higher Education, n.d.) university libraries supported DS on their campuses was investigated. Using multiple data sources, including an environmental scan, published reports, interviews with library counterparts at exemplar institutions, preliminary results from the library staff Skills Matrix Survey, and LDSC experience and brainstorming, a robust framework with time-based recommendations was developed around six categories, including: Communication & Branding, Assessment & Reporting, Human Resources & Employee Engagement, Cultivating Partnerships, Expanding and Creating Services, and Space & Infrastructure.

A literature review was conducted to identify trends in libraries and DS, with a special emphasis on identifying the relationships between DS initiatives and academic libraries. Currently, there is significant variation among university libraries’ involvement with DS, but this is a major growth area with new information becoming available regularly.

The relationship between libraries and DS was explored at the Fall 2015 Coalition for Networked Information (CNI) meeting by participants of the 2013 five-year partnership between the Gordon and Betty Moore and Alfred P. Sloan foundations and New York University, the University of California at Berkeley, and the University of Wisconsin (Steeves et al., 2015). These partnerships highlighted the connections between library information science and DS. Currently, several schools of library science offer a combined Master’s in both sciences.

The literature review highlighted several opportunities and challenges for university libraries with regards to DS. The interdisciplinary nature of DS was a recurring theme, as was the need for DS instruction for students. As artificial intelligence (AI) and DS become more prevalent in the world, the need for individuals to develop a foundational competency in data has become essential. The elements of data literacy, or “data acumen” as Laura Haas termed it in her July 2019 HDSR interview, include “a healthy skepticism about conclusions drawn from data, the ability to recognize good data from bad, an awareness of the need for data fairness and privacy, and familiarity with some of the tools, e.g., elementary exploratory data analysis and visualization” (Haas et al.,
Libraries also connect with students outside the curriculum. On many campuses, libraries are a major employer of student labor and internships. Libraries often hire students and provide them with DS opportunities that, in tandem with DS curricula, provide a great opportunity to increase DS awareness, expertise, etc. These hands-on experiences can further help to develop data-savvy students with marketable skills sought by employers today.

One opportunity for libraries is to partner in teaching students about the ethical issues surrounding data-driven research. Libraries are often already engaged in providing some form of data literacy instruction on campus, along with other core competencies like information and media literacy. Ideally, academics and industry would support a code of ethics that would be integrated into DS curricula (Wing et al., 2018). Data literacy, including areas such as data ethics, could be incorporated into existing literacies instruction or developed as new course-integrated offerings. Libraries could partner with departments to offer instruction on data ethics, along with the information literacy skills that they already teach.

Libraries provide neutral space for departments to meet, and librarians trained in DS provide support to researchers from all disciplines. To support DS, libraries will need to expand services and integrate with all stages of the research and data lifecycle (University of Wisconsin-Madison, 2018). One DS opportunity unique to libraries and other cultural institutions is the creation of machine-actionable collections, or collections as data, to expand the research potential of digitized materials. Some proponents argue that libraries should make machine-actionable collections a core activity (Padilla, 2020).

The library community is charting the path forward for DS, to support both internal operations and the research enterprise. To be successful, workforce development in libraries is needed to ensure library staffs are skilled in DS, machine learning, and AI. To integrate algorithmic methods among the library community, shared development and distribution of methods and training data are needed (Padilla, 2020). It is clear from the literature that there are many opportunities for university libraries to engage with DS efforts.

**Approach**

**Environmental scan**

To gain an understanding of how other university libraries are integrated with campus DS efforts, the LDSC conducted a broad environmental scan of over 60 institutions, collecting information from publicly available online sources to identify trends and establish familiarity with baseline services offered to support campus DS efforts. From this broad scan, 20 schools identified as notable in the Moore–Sloan Data Science Environments (MSDSE) (Evaluation of the Moore-Sloan Data Science Environments, 2019) and the “Academic Data Science Centers in the United States” (Luba Katz, 2018) reports were further examined. In this broad survey of institutions, it was found that although there was a great deal of DS activity taking place across university campuses and libraries, there was room for greater integration of these programs. For example, only rarely was the library listed as a partner or collaborator with campus DS efforts. Evidence of partnerships was often limited to blog posts or announcements about collaborative projects or events.
Ultimately, nine schools were selected to evaluate more closely, including: Arizona State University (Arizona State), University of California at Berkeley (Berkeley), Brown University (Brown), University of Indiana (Indiana), Massachusetts Institute of Technology (MIT), New York University (NYU), University of Rhode Island (Rhode Island), University of Virginia (UVA) and University of Wisconsin at Madison (Wisconsin). These universities were chosen because they are existing or emerging leaders around DS, have a special focus on AI, social sciences, the humanities, or data cataloging, or have a stated connection to campus libraries or library schools. The following information was collected, when available:

- Description of the DS program including strengths.
- DS degrees offered.
- DS program organization and physical location.
- Library data services, activities, and connections.

Overall, no single university encapsulated all these aspects as part of their collaboration between their university library and DS program.

**Interviews with library counterparts at exemplar institutions**

To get a better understanding of academic libraries’ relationship with DS, the LDSC invited library counterparts from the nine universities listed above to participate in interviews designed to elicit how they support DS at their campuses (see Appendix A for interview questions). Within the allocated time frame, interviews were scheduled with six of the nine universities’ key library-affiliated contacts. Interviews were conducted in Fall 2019 by one or more members of the LDSC with library counterparts at Brown, MIT, NYU, Berkeley, UVA, and Wisconsin. Interview questions were based on topical DS areas concerning how libraries work collaboratively with campus colleges and/or institutions to support student learning, teaching, preservation, and faculty research (IRB exempt 19-2916).

The LDSC found that each library’s approach to supporting DS was unique, where each library had between 4 and 8 dedicated DS individuals and several institutions indicated they anticipate hiring additional staff to support campus DS efforts, including non-Library Science Master’s (MLS) positions. Many have formed strong partnerships across their campus, involving graduate education, central IT, research computing, and with their Vice Chancellor for Research and Graduate Education. None were aware of any memorandum of understanding (MoU) to formalize project-based partnerships or joint appointments between departments and the university libraries.

In addition, some interviewees indicated offering a variety of services, some focusing on data within the research lifecycle and others on data management in support of the research lifecycle. Others indicated they developed “toolkits” for digital scholarship workshops (e.g., GIS, web scraping, and text analysis) or library staff training such as that offered by The Carpentries (The Carpentries, n.d.) as part of their mission. Unsurprisingly, it was noted that the type of services offered stemmed from established partnerships and units on campus that were engaged with the library and those who
have goals aligned with the libraries’ mission. Due to the uptick in DS across their campuses, most libraries reported an increase in their DS services (e.g., data curation, text analysis, all forms of managing large data sets, DS instruction, workshops/tutorials/consultations). Predictably, institutions varied with how their library made physical space available for DS. None of the individuals to whom the LDSC spoke indicated they had exactly what they needed or wanted in terms of space for DS activities.

Finally, the LDSC found that some DS programs are using Jupyter notebooks (Project Jupyter, n.d.) for their undergraduate program (University of California, Berkeley, 2016); libraries are buying, creating, and/or working with indexing services to make data sets available for their patrons; staffs were provided the opportunity to work in cross-unit teams on small scale library data projects to hone their skills; libraries were developing fellowship programs; library liaisons did not see boundaries between data work and their traditional roles.

**Stakeholder matrix**

An initial step in the LDSCs approach to developing the library’s DS framework was to identify the types of users with whom libraries partner and their potential DS needs. DS needs on campus vary by type of stakeholder and their level of data needs. To understand the needs of campus stakeholders, the LDSC considered individuals based on their roles at the university (students and faculty) and their potential level of involvement with DS (ranging from low to high). Four categories of stakeholders were identified including researchers, instructors, graduate students, and undergraduate students. For each stakeholder category, potential DS needs where the libraries could provide support were identified. For example, undergraduate students in non-data focused disciplines may benefit from instruction around data literacy, including data ethics, while students in data-focused disciplines may benefit from computing resources. Instructors may benefit from data curriculum modules for different disciplines, while researchers may require librarian collaborators on projects requiring data expertise (see Tables 1 and 2 for more examples). The authors anticipate that needs will be met primarily through workshops, Open Labs, consultations, instruction, events (e.g., speaker series; Data Day event to share research), staff expertise, and library resources (e.g., purchased datasets).

The LDSC also developed a survey to identify research and instructional needs of UNC-CH partners and stakeholders that could be used by a future implementation committee. These efforts, in conjunction with focus groups of stakeholders, can be used to develop a strategic approach for creating, expanding, and sunsetting services.

**Assessing staff expertise and skills gap**

The LDSC was charged with assessing the DS skills and talent among existing staff to identify areas where additional capacity would be needed. To assess existing skills and talent, and to test out a method of evaluation, a Qualtrics survey (Appendix B) was developed and distributed, requesting Library department heads to report how many staff in their unit had some DS experience/exposure or expertise in specific areas. Areas assessed related to specific phases of the data lifecycle (e.g., sourcing, management, curation) or specific software used for managing data or data analysis (e.g., R, Python,
GIS). Initial results were visualized with a heat map to provide a quick assessment of where gaps exist. This initial effort was designed to test the utility of the survey instrument and the process for collecting and analyzing information.

Initial results from the pilot skills survey indicated skills and capacity gaps in multiple areas. Areas of strength included data sourcing, data cleaning and preparation, data preservation, and data management. Growth areas included research computing and analysis. Ultimately, areas, where the Libraries will increase capacity, will depend on a more formal assessment of DS needs on campus and strategic decisions by University Library Leadership. These results were consistent with experience, in that staff with a high level of expertise are in high demand and spend a significant amount of time providing instruction, one-to-one consultations, and support to faculty research. Skills gaps identified as critical to supporting campus DS needs will assist in developing a strategy to build capacity among existing staff and in guiding the creation of new positions.

**Establishment of the DS framework**

The LDSC developed a framework to expand DS support to students, faculty, and researchers. The framework was centered on a set of recommendations covering six areas put forward by the LDSC. For each recommendation category, the LDSC defined

| Table 1. Potential data science needs for faculty and post-doctorate researchers and instructors. |
|---------------------------------------------------------------|
| Level of data science involvement for faculty | Potential needs |
| Faculty & post-doctorates (researchers) | Data literate |
| - Awareness of resources & expertise available from the Libraries. |
| - Identifying potential partners/collaborators (including colleagues from other disciplines and library partners). |
| - Librarian consultations for projects requiring data expertise. |
| - Physical space. |
| Data user | Support preparing Data Management Plans. |
| - Support creating documentation. |
| - Computing resources/infrastructure. |
| - Purchased datasets. |
| - Assistance overcoming methodological hurdles. |
| - Assistance with publications. |
| Data-intensive/Data Science Faculty | Application of open research methods. |
| - Long-term storage and preservation. |
| - Handling sensitive data and de-identification. |
| Faculty (instructors) | Non-data focused course instructor |
| - Data literacy curriculum modules for different disciplines. |
| - Course-specific instruction on data literacy. |
| - Access to introductory resources providing data & visualizations. |
| - Examples of classroom work. |
| - Data, including digitized primary source materials. |
| Data science adjacent course instructor | Virtual space for course materials/data. |
| - Pedagogical support. |
| - Virtual sandbox for projects. |
| - Specialized software for data manipulation and analysis. |
| Data-intensive/data science course instructor | Remedial support for students. |
| - Problem sets. |
| - Collaborative space and tools. |
| Level of data science involvement for students | Potential needs |
|-----------------------------------------------|----------------|
| **Graduate students**                         |                |
| Students in non-data focused disciplines     | Instruction around data literacy, including data ethics. |
|                                               | Awareness and acquisition of data for all disciplines (including primary source materials). |
|                                               | Opportunities to learn about projects. |
| Non-data science students using data for coursework or thesis | Communication skills for presenting research (e.g., through posters, presentations, visualizations). |
|                                               | Reputation management (e.g., via ORCID). |
| Students in data-intensive/data science       | Information sharing to identify opportunities and increase visibility (e.g., for projects, employment). |
|                                               | Support preparing data management plans. |
| Students in non-data focused disciplines     | Small curricular modules for recitations. |
|                                               | Physical space. |
| Non-data science students using data for coursework or thesis | Support locating and acquiring datasets. |
|                                               | Research assistance and project ideas. |
|                                               | Information about grant funding opportunities. |
| Students in data-intensive/data science       | Support preparing data management plans. |
|                                               | Information sharing to identify opportunities and increase visibility (e.g., for projects, employment). |
| Students in non-data focused disciplines     | Support preparing data management plans. |
| Non-data science students using data for coursework or thesis | Support preparing data management plans. |
| Students in data-intensive/data science       | Support preparing data management plans. |
| Undergraduate students                        |                |
| Students in non-data focused disciplines     | Communication skills for presenting research (e.g., through posters, presentations, visualizations). |
|                                               | Instruction around data literacy, including data ethics. |
|                                               | Opportunities to hear from fellow students. |
| Non-data science students using data for coursework or thesis | Access to introductory resources providing data & visualizations. |
|                                               | Programming around real-world applications of data science, data ethics, and other topics. |
| Students in data-intensive/data science       | Support locating datasets. |
|                                               | Virtual sandbox space. |
| Students in data-intensive/data science       | Research support including advice on research plan. |
|                                               | Campus research computing resources. |
| Students in non-data focused disciplines     | Access to and support for specialized software for data manipulation and analysis. |
| Students in non-data focused disciplines     | Assistance overcoming methodological hurdles. |
goals and provided specific actions along with corresponding timelines to assist with prioritization (see Tables 3–8).

**Communication and branding**

The DS framework lays out recommendations around communication and branding designed to raise awareness of library expertise (Table 3). The LDSC determined that establishing a clear and intentional strategy around branding and communication for internal and external purposes will enable the library to showcase expertise and impact to stakeholders. Based on the environmental scan and web site review, it was discovered that few libraries provided clear communication and branding around their DS role, support, and partnerships within their campus; this was true even among institutions with strong DS programs. Additionally, having a strong communication and branding strategy assists in building relationships with donors, campus partners, recruiting new staff, and providing clarity around how library expertise, services, and spaces help DS efforts flourish. Based on these findings, the LDSC felt communication and branding should be done in collaboration with Library Communications and Development units and should occur through multiple channels including the Library website, social media, signage, internal meetings, and events for campus partners and stakeholders.

It was determined that the UNC library website should be a central point of reference that would outline the partnership with the Data Science @ Carolina initiative, illustrate library support of DS efforts in terms of services, staff, spaces, events, and workshops, and provide a directory of library staff working in DS related areas. In addition, to ensure library work in DS is highlighted and to raise awareness of library staff expertise and impact via social media outlets, a communication strategy was recommended. Highlighting how the Libraries partner with researchers and faculty, assists students in acquiring data skills, engages with faculty on curriculum integration, and the ways in which the Libraries contribute to the Carolina Next strategic framework are extremely valuable and impactful stories that should be shared with current donors and prospects. The LDSC asserted that it was equally important to identify ways to use internal communication opportunities to keep staff aware of DS initiatives, highlight staff achievement (e.g., via presentations, projects, publications, grants), and share opportunities.

**Assessment and reporting**

Recommendations around assessment and reporting include establishing a clear assessment and evaluation program to establish pre-defined milestones to gauge progress, identify opportunities for improvement, and verify stakeholder needs are being met (Table 4).
Assessment should also include a plan for identifying resources needed for space and technology as well as areas where the library could further integrate with established partnerships and/or expertise. In addition, an evaluation plan should be created to monitor the progress of each DS element described in the framework, which may include monitoring services, space, skills acquisition, curricula, and partnerships. Assessment and evaluation plans should identify specific metrics for each of these categories, to be measured regularly (e.g., annually). For example, in the area of curriculum integration, the library could identify the array of topics being taught, how many sessions are offered, which sessions are course-integrated versus standalone workshops, and the number of students reached. For sessions in which library staff teach alongside faculty, an evaluation component for library-based workshops/sessions that gauge acquisition of skills and knowledge, could be included.

Reporting mechanisms should also be created to highlight the progress and impact the Libraries is making with Data Science @ Carolina. A robust assessment and evaluation plan should be established for each of the following core areas:

- Instruction and Curriculum Integration.
- Research Support.
- Priority Partnerships.
- Library Spaces.
- Library Computing, including infrastructure.
- Library Skills and Services (e.g., data curation, data storage and management, classes, and workshops).

**Human resources and employee engagement**

The framework addresses human resources and employee engagement in the areas of reskilling, performance management, and professional development, and skills required for new hires (Table 5).

**Table 4.** Assessment & reporting goals and recommendations.

| Goals | Includes recommendations around |
|-------|--------------------------------|
| Showcase library expertise and capacity. | • Assessment and evaluation plan |
| Cultivate data science partnerships. | • UNC data science needs |
| Develop relationships with donors. | • Data science spaces assessment team |
| | • Data science skills assessment |
| | • Annual overview |

**Table 5.** Human resources & employee engagement goals and recommendations.

| Goals | Includes recommendations around |
|-------|--------------------------------|
| Bridge the skills gap within University Libraries to support data science activities on campus. | • Reskilling |
| Incentivize library staff to gain skills relevant to data science. | • Performance management and professional development |
| Provide ample professional development opportunities. | • New hires |
| Develop a tiered approach to providing services around data science. | |
As mentioned above, to illustrate and maintain a One Library approach, the LDSC surveyed all library staff to access their DS skill level. Seeking input from library staff helped galvanize support for, and integration with, the DS framework, while empowering the staff to feel part of the University Libraries’ strategic vision for DS. Reskilling staff will be a critical element to meeting the demands of Data Science @ Carolina and will require ongoing attention. Workshops and courses on DS could be offered in-house or with a campus partner, including the School of Library Science (SILS) and the Odum Institute for Research in Social Science (co-located in the Library).

One recommendation suggested the implementation of a tiered system ranging from no DS skills (Tier 0) to exceptional DS skills (Tier 3), to address professional development opportunities. The system would be tied to staff professional goals, with the intention of upskilling and addressing library service needs and reducing gaps in DS expertise. Reskilling will provide opportunities for career growth among current staff and allow for their participation in national conferences to showcase their contributions to DS as well as the Libraries’ DS program growth and overall impact on the campus.

Survey results emphasized the need for new hires that possess expertise in DS. University Libraries’ staff with deep DS expertise have limited capacity to be embedded in research projects due to heavy commitments for instruction and consultations. In the short term, expertise can only be increased by hiring new staff, as upskilling of the current staff will be a lengthier process. New staff members will allow the Libraries to quickly expand DS engagement, while current staff continues to grow DS skills.

**Partnerships and stakeholders**

The LDSC developed goals and recommendations around partnerships. Robust partnerships with other campus units will be critical to the University Libraries’ successful involvement with Data Science @ Carolina (Table 6).

The LDSC identified campus units that share the University Libraries’ goals and offer services complementary to that of the Libraries. For each campus unit, the LDSC considered primary activities, any unique data-related services offered, overlap with University Libraries services, and the existence of or potential for a data-focused intensive partnership. Several prospective partners were identified, along with potential outcomes focused on curricular and research integration.

Certain partners could enable the creation of a data literacy program focused on the undergraduate curriculum. Other partnerships could:

- lead to an expansion of the Libraries’ current teaching associations;
- ensure that DS collaborations are integrated at the point of need;

| Goals                                                                 | Includes recommendations around                                                                 |
|----------------------------------------------------------------------|------------------------------------------------------------------------------------------------|
| Cultivate data science partnerships around research and curriculum integration. | - Library priorities and partnerships |
| Transform existing services into an immersive model where library staff are integrated into research projects. | - Concierge roles |
|                                                                       | - Mechanisms for collaboration                                                               |
|                                                                       | - Formal communications channels                                                              |
|                                                                       | - Joint-funded positions with campus partners                                                |
• guarantee that researchers are able to obtain needed data from the Libraries’
digital and data collections; and
• allow for an increase in the Libraries’ participation in research projects.

Lastly, partnerships could serve as a unifying force for the various disparate initiatives,
experiments, and labs that pop up on campus.

The LDSC also recognized the importance of communication with partners and recom-
mended that DS concierge roles be established to foster relationship building and
provide a formal channel for communication between the Libraries and partners.
Memoranda of Understanding (MoU) were recommended as a tool for communication
among partners, and joint or co-funded positions were identified as a mechanism for
solidifying relationships.

Creating and expanding services
The framework laid out recommendations around expanding existing services as well as
developing new ones (Table 7). The LDSC noted high demand exists for DS services
currently offered by the Libraries and noted that increased demand is likely in all
instances. Specifically, a significant increase in demand is expected around services
related to predictive analytics using AI and machine learning, text analytics, data mining
(particularly on library resources), and providing datasets for humanists from digitized
primary source materials.

Most of these current data-related services are at capacity, and redundancy in staff
skills is low. Reskilling existing staff, strategic hires, and significant investment and
cooperation with campus partners (e.g., joint appointments) are all expected to play a
role in meeting the increased demand for data-related services.

In addition, there were several other data-related services with growing or expected
demand, which are not currently offered by the University Libraries or are only sup-
ported at a minimal level. These growth areas include computing and infrastructure
support services, open science/research reproducibility, and other types of analysis (e.g.,
image analysis, Genome-wide association studies (GWAS), and meta-analyses). Further,
the Libraries anticipates the need to establish a data acquisition strategy and service,
and increase instructional support around data literacy and data ethics. Again, to meet
the anticipated demands for these new services, significant investment and collaboration
with campus partners will be essential.

Space and infrastructure
The LDSC identified physical space needs and organizational and technological infra-
structure considerations in support of DS in the framework (Table 8). Broadly speaking,
LDSC recommendations for both space and infrastructure were to cultivate commu-
nities, support interdisciplinarity, and catalyze new partnerships around DS. Through the

| Table 7. Creating and expanding services goals and recommendations. |
|---------------------------------|-----------------|
| Goals                          | Includes recomm | |
| Meet Carolina’s needs around data science. | • Services to expand |
| Cultivate data science partnerships.    | • Services to create |
Libraries’ own experiences and interviews conducted, the LDSC found that providing a variety of physical spaces would be warranted when supporting DS and other research using digital methods. Space needs may include places to hold meetings and consultations, support idea incubation, host presentations, facilitate teamwork, and provide access to needed technology. A single space can often be configured to meet and support multiple needs and uses.

Library spaces could be effectively configured and programmed to encourage the use of library services and promote collaboration. To fully support these goals, bringing together a space assessment team that includes key partners and service providers could begin by reviewing or creating a formal space use policy. This group could help assess and inventory existing spaces and technologies in the University Libraries and identify hardware and software needs. A space use policy could help ensure unique library resources are made available as broadly and equitably as possible. To fully support and promote collaboration, this type of policy should also provide some level of access for users who are not directly affiliated with a DS program.

The computing and technological infrastructure required to support DS are complex and expensive. Most likely, there are several groups on any campus already providing technical support to users. Therefore, a recommended first step would be to appoint a technical infrastructure review team to document the existing ecosystem on campus. In addition to fostering a better understanding of the landscape, this group could also identify the most strategic ways for the Libraries to participate within this system and identify which partnerships to pursue or strengthen. For example, the LDSC recommended strengthening existing work with campus Research Computing operations as well as pursuing other, new campus partnerships related to licensing, preservation, and tool development to integrate commercial infrastructure into the research lifecycle more fully. In each case, the LDSC focused on ways the Libraries could add value to existing infrastructure by leveraging core strengths.

In time, the technical infrastructure review team could turn their focus to recommending or developing internal and external consultation services. The team could also develop a business plan to manage and control the growing cost of infrastructure over time and could address long-term issues such as ongoing maintenance and planned sunsetting of projects.

In addition to technological infrastructure, the framework addressed issues of library organizational infrastructure—ways to address possible organizational barriers and incentivize DS work through good change management practices. Recommendations center mainly around fostering a culture of communication, increasing professional development funding and opportunities for staff, and working with managers and supervisors to help establish and assess new staff performance goals, duties, and expectations related to DS.

### Table 8. Space & infrastructure goals and recommendations.

| Goals                                                                 | Includes recommendations around                                      |
|----------------------------------------------------------------------|---------------------------------------------------------------------|
| Cultivate community, interdisciplinarity, and catalyze new partnerships around data science. | • Data science labs                                                   |
| Establish a group to review the Libraries’ current technical infrastructure. | • Formal space use policy                                            |
| Create space use policies.                                           | • Technical infrastructure review team                             |
|                                                                      | • Infrastructure partnerships                                      |
|                                                                      | • Technical consultation services program                         |
|                                                                      | • Use policy for libraries’ technical infrastructure             |
|                                                                      | • Business plan to manage infrastructure costs                    |

---
Engaging in the framework

Although the UNC Libraries has not yet increased staffing to support the DS initiative, the Libraries has been working to advance the DS strategic initiative by reskilling staff, generating collections as data, participating in research partnerships, and supporting DS related instruction.

Dedicating resources to training staff will be critical to the success of DS efforts. Providing opportunities and encouraging staff to develop data skills by attending conferences and workshops as part of their professional development is one way to foster skills awareness and acquisition. As an example, a consortial application for funding was secured from the National Network of the National Libraries of Medicine—Southeastern/Atlantic Region to provide Library Carpentry (Library Carpentry, n.d.) training for staff across the Triangle Research Libraries Network (Triangle Research Libraries Network, n.d.) that includes UNC-CH, Duke University, North Carolina Central University, and North Carolina State University. This is an excellent example of how libraries can work together to build capacity around areas such as DS that are essential to the profession.

Given the Libraries’ distinctive collections and expertise, the Libraries are uniquely positioned to support the application of DS in the humanities. For example, the Libraries’ project On the Books: Jim Crow and Algorithms of Resistance created a text corpus of North Carolina statutes optimized for computational use and applied machine learning to identify Jim Crow laws. The project recently received additional funding from the Association of Research Libraries and is the University Libraries’ second collection as a data project, with DocSouth Data (Documenting the American South, n.d.) being the first. UNC-CH envisions encouraging applications of DS across all disciplines and these two data projects are examples of how the Libraries is advancing the adoption of DS methods in the humanities.

Librarians continue to participate in research partnerships, working with scholars as co-authors on research papers. Many scholars are subject matter experts but need assistance with integrating data into their research. The Libraries currently offer services in GIS, visualization, impact measurement, and text and data analysis. Examples of current librarian research partnerships centered on data include integrating data analysis into research methodology, creating data visualizations for faculty publications, and partnering with campus units to demonstrate impact through visualizations, including the use of bibliometric analysis. Additionally, the LDSC has also seen how library space and expertise can be used strategically, with partners such as the Global Women’s Health collaboration, which is a multi-disciplinary team located at the Libraries’ Health Sciences Library. Through this partnership, the library plans to identify ways to integrate data expertise into the data-intensive study that will include the development of algorithms, systematic reviews, and metadata applications.

One of the goals included in the University’s strategic framework is “ensuring that every graduate of UNC-Chapel Hill is data literate” (Carolina Next: Innovations for Public Good, n.d.). The Libraries have been supporting this goal by increasing data-focused instruction through workshops on topics including bioinformatics, visualization, GIS, Python, R, and Excel. The workshops increase the data literacy of students and offer library staff an opportunity to hone their skills. Library staffs have been
encouraged to attend workshops, and some are advancing their skills enough to assist with, and ultimately lead instruction. Librarians are participating in the Ithaka S+R research study Teaching with Data in the Social Sciences (Ithaka S+R, 2020) to gain a better understanding of the support needed by faculty teaching with data. Lastly, the University’s new general education curriculum includes a mandatory course for undergraduates called Ideas, Information, and Inquiry, that includes a data literacy component. Librarians are in communication with campus stakeholders on how to support these courses.

**Conclusions**

The UNC-CH initiated a pan campus LDSC that included librarians in the Fall of 2019 to investigate the feasibility of establishing a DS school. The Libraries saw an opportunity to conduct a review of its current support of DS on campus and provide recommendations and a framework to Library Leadership on how they can enhance support of student learning and faculty research in the area of DS. As a result of this review, a framework was developed for providing DS services, technologies, and spaces for the UNC-CH community. The result is a robust framework with time-based recommendations developed around six categories, including Communication & Branding, Assessment & Reporting, Human Resources & Employee Engagement, Cultivating Partnerships, Expanding and Creating Services, and Space & Infrastructure. The completed framework is published in “University Libraries: Data Science Framework” (Mani et al., 2020).

DS programs are large, interdisciplinary undertakings that impact many aspects of an organization and campus. They are often pan-university in nature and may pose questions to campus around design and support of physical spaces, technical infrastructure, research support, and partnerships, or pedagogy and the ways curricula interweave into student learning. Traditionally, libraries have experience and expertise in supporting large, interdisciplinary research endeavors, so there are many clear ways for them to engage with DS programs.

First, libraries can play a critical role in the design, teaching, and promotion of data literacy. This instruction could be developed as course-integrated modules or stand-alone content and could build off existing information literacy programs. Data literacy competencies can also be gained through student employment or other experiential learning opportunities the libraries provide (e.g., DS consultations, hosting workshops, or teaching one-shot data literacy lessons).

In addition to instruction, libraries also possess some of the infrastructure needed to support a DS program. Libraries typically have an organizational structure in place (e.g., individual subject specialist librarians, subject teams, specialized support units, etc.) for supporting researchers across multiple disciplines. This service model and expertise can inform, or perhaps even be expanded or adapted to include support for new DS initiatives.

Because of their unique storage and access needs, libraries often maintain their own technology infrastructure within a campus community. Libraries should seek ways to partner with other, broader campus infrastructure programs and support, such as high performance or research computing. When pursuing these partnerships, libraries should
seek ways to leverage their core strengths to add value to the existing infrastructure. Core strengths may include library collections, including datasets, expertise in finding, licensing, and purchasing datasets and other resources, and specialized skills in using and transforming data.

The creation of a DS framework for the UNC Libraries is a model in how bringing together a diverse library team resulted in the creation of a comprehensive and mission-focused roadmap. For those seeking to engage in formulating a DS framework, the following are highly recommended:

1. Mobilize a diverse team with varied experience related to strategic planning, research and data, instruction, information technology, assessment, space planning, and subject specialization.
2. Establish a clear charge with a defined scope and timeframe.
3. Engage internal and external stakeholders that include library staff, campus partners, and experts external to the organization to gain a holistic understanding.
4. Conduct a library-wide DS skills assessment, and if applicable, a needs assessment with campus partners.
5. Create the framework to include potential timeframes around each milestone, to help establish resources and support needed.
6. Create an evaluation plan to monitor progress.
7. Seek to have a dedicated implementation team to enact the recommendations and monitor success.

The interdisciplinary nature of DS and the extent to which academic research libraries are integrated into their campuses’ research and instruction realms further exemplify how libraries are integral partners in DS ecosystems. Libraries can use expertise as it relates to information identification, synthesis, and evaluation to enhance the lens by which questions or problems can be investigated. The ubiquitous nature of data necessitates the re-envisioning of how libraries evolve to partner and provide expertise, information, access, and space to further data inquiry as it impacts societal needs.

Acknowledgments
The authors wish to thank additional members of the LDSC: Lorin Bruckner, Jason Casden, Adam Dodd, Matt Jansen, Jamie McGarty, Morgan McKeegan, Sarah Morris, and Jessica Venlet, as well as library colleagues from external institutions whose expertise and contributions helped formulate this library data science framework.

ORCID
Nandita S. Mani http://orcid.org/0000-0002-0955-1066
Michelle Cawley http://orcid.org/0000-0002-7178-328X
Amanda Henley http://orcid.org/0000-0002-5269-2630
Therese Triumph http://orcid.org/0000-0001-5494-4574
Joe M. Williams http://orcid.org/0000-0001-5113-1041
References

Carnegie Classification of Institutions of Higher Education. (n.d.). Basic classification description. https://carnegieclassifications.iu.edu/classification_descriptions/basic.php

Carolina Next: Innovations for Public Good. (n.d.). Planning our future. https://carolinanext.unc.edu/

Documenting the American South. (n.d.). DocSouth Data. https://docsouth.unc.edu/docsouthdata/

Evaluation of the Moore-Sloan Data Science Environments. (2019). Final report. http://msdse.org/files/MSDSE_Eval_Final_Report_Feb_2019_v2.pdf

Haas, L., Hero, A., & Lue, R. A. (2019). Highlights of the National Academies Report on “Undergraduate Data Science: Opportunities and Options.” Harvard Data Science Review, 1(1). https://doi.org/10.1162/99608f92.38f16b68

Ithaka S+R. (2020, May 14). Launching two projects on supporting data work. https://sr.ithaka.org/blog/launching-two-projects-on-supporting-data-work/

Library Carpentry. (n.d.). Library carpentry. https://librarycarpentry.org/index.html

Luba Katz, P. (2018). Academic Data Science Centers in the United States: A study of 20 universities. Zenodo. https://doi.org/10.5281/zenodo.3887010

Mani, N., Cawley, M., Henley, A., Triumph, T., Williams, J., Jansen, M., Dodd, A., Casden, J., Mc Keehan, M., Venlet, J., Bruckner, L., Morris, S., & Mc Garty, J. (2020). University libraries: Data science framework. Carolina Digital Repository. https://doi.org/10.17615/5c78-3t83

Padilla, T. (2020, August 26). Responsible operations: Data science, machine learning, and AI in libraries. OCLC. Project Jupyter. (n.d.). Jupyter. https://www.jupyter.org

Steeves, V., Mitchell, E., & Muijenburg, J. (2015, December 11). Organizational implications of data science environments in education, research, and research management in libraries. OSF. https://osf.io/fqcr7/

The Carpentries. (n.d.). The carpentries. https://carpentries.org/index.html

Triangle Research Libraries Network. (n.d.). Triangle Research Libraries Network. https://trln.org/

University of California, Berkeley. (2016). Data science planning initiative – Faculty advisory board report. DataSciencePlanningInitiative-FacultyAdvisoryBoardFinalReport.pdf

University of Wisconsin-Madison. (2018). Wisconsin in the information age a report by the Working Group on Computing at the University of Wisconsin-Madison. https://www.cs.wisc.edu/wp-content/uploads/sites/871/2019/01/Report-by-the-Task-Force-on-Computing-at-UW-Madison-Google-Docs-1.pdf

Wing, J., Janeja, V., Kloefkorn, T., & Erickson, L. (2018). Data Science Leadership Summit summary report. Columbia University. https://doi.org/10.13140/RG.2.2.13710.61764

Appendix A: Interview questions for library counterparts

IRB exempt 19-2916

Purpose: Understand how exemplar institutions have partnered with their data science programs.

Audience: Library counterparts at nine exemplar institutions.

Method: Interviews via teleconference.

Interview questions

1. What partnerships exist between the Library and other units on campus around data science?

Follow-up where applicable:
   - How did you go about cultivating the partnerships?
   - Do you have any formal structures or frameworks that you use, such as joint appointments or MOUs?

(continued)
Appendix B: Skills matrix survey

The survey below is an example of how to evaluate existing experience and proficiency in skills associated with the practice of data science. The directions and survey were designed for response from department heads.

Directions provided to respondents

Complete the skills survey below for staff in your department. Refer to the explanations in this table when completing the survey.

| Category          | Explanation                                                                                                                                 |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------|
| Some experience   | Staff with some experience may have either (1) deep expertise in a very limited application of a certain topic or software or (2) has limited or intermittent general experience and the ability to tasks, with occasional help. |
| Expertise         | Staff with expertise have significant experience with a method or software and the ability to quickly learn new tasks or help others troubleshoot. |
| Comments          | Provide comments or additional information as desired. You may wish to note if staff with experience or expertise are primarily public-facing or work on internal library projects only. |
| Total             | Indicate the total number of individuals with some experience or expertise in this cell. It is unlikely that the total number of individuals will be the total of this column (i.e., in many cases one individual will have experience in several areas or software tools). |
| Other             | May include: 3D Printing, ArcGIS or QGIS, Gephi, GitHub, Google Earth, Java script, Jupyter Notebooks, Linux: shell scripting/bash, Nvivo, Solr, Stata/SPSS/SAS, Timeline JS, VosViewer, Wordpress, Other automation tools (specify) |

Appendix B: Skills matrix survey

Follow-up interview questions

1. Have you noticed any changes in the demand for library services due to the data science program/initiative at your institution?
2. If the library can’t provide a data science service that is requested, who do you refer people to?
3. How did you determine the ways in which the library could be integrated into your institution’s data science program (e.g., needs assessment, focus groups, surveys, data gathering, interviews/conversations, etc.)?
4. Have you noticed any changes in the demand for library services due to the data science program/initiative at your institution?
5. How did you determine the ways in which the library could be integrated into your institution’s data science program (e.g., needs assessment, focus groups, surveys, data gathering, interviews/conversations, etc.)?
6. Did partnering with data science initiatives require structural changes to your library’s org chart? If so, what new positions were created? How were existing job descriptions rewritten?
7. Did you establish any physical spaces dedicated specifically to supporting this initiative? Did you partner with any other organizations to design or manage these spaces?
8. In terms of spaces and services, have you found any areas of misalignment between what you planned and actual use?
### Survey

| Skill or Software                                      | Description                                                                                                                                                                                                 | Number of University Libraries Staff with some experience or exposure in this area | Number of University Libraries Staff with expertise in this area | Comments |
|--------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|----------------------|-----------|
| Data Sourcing                                          | Staff who can find and compile appropriate data from external sources to help refine or answer the research question.                                                                                     |                                                                                  |                      |           |
| Data Cleaning and Preparation                          | Staff who can conduct all necessary transformations including merging, reshaping, or other reformattting to make analysis possible.                                                                          |                                                                                  |                      |           |
| Analysis: AI and Machine Learning for Predictive Analytics | Staff who can use machine learning approaches to predict outcomes/events, classify data, identify data/literature sources (e.g., for a systematic review).                         |                                                                                  |                      |           |
| Analysis: GIS                                          | Staff who use specialized tools and methods to work with geospatially referenced data (e.g., coordinate systems, shapefiles and geodatabases, spatial statistics).                                           |                                                                                  |                      |           |
| Analysis: Impact                                       | Staff who can examine the scope and pattern of research collaborations; measure and assess research impact by discipline area; communicate research impact to audiences such as funders or promotion and tenure committees. |                                                                                  |                      |           |
| Analysis: Measurement and Visualization                | Staff who can use statistical methods to infer characteristics of a population or process.                                                                                                                  |                                                                                  |                      |           |
| Analysis: Inference Statistics                         | Staff who use specialized tools and methods to derive meaningful information from unstructured text.                                                                                                          |                                                                                  |                      |           |
| Analysis: Other Cases and Data Types                   | Staff with experience or expertise in other data types or analysis that requires specialized software or methods (e.g., image analysis, Genome wide association studies (GWAS), meta-analyses, music). |                                                                                  |                      |           |
| Analysis: Network Analysis                             | Staff who can identify, measure, and visually represent relationships between groups of entities (people, terms, objects, etc.)                                                                               |                                                                                  |                      |           |
| Visualization and Other Data Presentation               | Staff who discover data insights and communicate findings through techniques that employ our innate ability to distinguish visual patterns in our environment.                                               |                                                                                  |                      |           |
| Data Preservation and Archiving                         | Staff who can assist with using institutional repositories and other data storage systems to arrange, describe, and protect the provenance of data while preserving its integrity and making it available for reuse. |                                                                                  |                      |           |
| Reproducibility                                        | Staff who can document code with metadata to enable scientific replication or reproduction. Tools and methods for sharing code alongside data, with necessary information about compute environment, codebooks, etc. |                                                                                  |                      |           |
| Data Management                                        | Staff who organize and manage data during research; data documentation (e.g., metadata, file format, naming conventions, file organization); version code or files (e.g., git).                                      |                                                                                  |                      |           |

(continued)
| Skill or Software | Description | Number of University Libraries Staff with some experience or exposure in this area | Number of University Libraries Staff with expertise in this area | Comments |
|------------------|-------------|----------------------------------------------------------------------------------|-----------------------------------------------------------------|-----------|
| Computing        | Staff with experience or expertise around research computing infrastructure including specialized platforms to support processor-, memory-, or storage-intensive computing (high performance computing, parallel/distributed processing); data processing pipelines; server operations/management | | | |
| Integrating Data Literacy into Curricula | For example: Identifying learning objectives; creating a rubric for evaluation; updating syllabi; preparing instructional materials; and ethics, including: privacy/confidentiality; algorithm bias; copyright (e.g., legality of web scraping); social impact of data | | | |
| R                | Open-source programming software emphasizing statistical analysis | | | |
| Python           | Open-source programming software | | | |
| SQL              | Relational database management system | | | |
| APIs             | Application programming interface | | | |
| Tableau          | Interactive data visualization software | | | |
| Advanced Excel   | Examples: formulas; pivot tables, power query, etc. | | | |
| Other            | Includes: 3D Design; GitHub; JavaScript; Library Carpentries Instructor; Linux; shell scripting/bash; MS Access; Oxygen; SharePoint; SPSS; VosViewer; Web management; Wordpress; XSLT and Xquery; Library Carpentries | | | |
| Total            | Indicate the total number of individuals in this column. | | | |

*Total number of individuals will likely be less than the total of the column as one individual may have experience or expertise with multiple skills or software.*