Data Science Support at the Academic Library

JEFFREY C. OLIVER
Data Science Specialist, Office of Digital Innovation and Stewardship, University Libraries, University of Arizona, Tucson, AZ, USA

CHRISTINE KOLLEN
Data Curation Librarian, Office of Digital Innovation and Stewardship, University Libraries, University of Arizona, Tucson, AZ, USA

BENJAMIN HICKSON
Geospatial Specialist, Office of Digital Innovation and Stewardship, University Libraries, University of Arizona, Tucson, AZ, USA

FERNANDO RIOS
Research Data Management Specialist, Office of Digital Innovation and Stewardship, University Libraries, University of Arizona, Tucson, AZ, USA

ABSTRACT. Data science is a rapidly growing field with applications across all scientific domains. The demand for support in data science literacy is outpacing available resources at college campuses. The academic library is uniquely positioned to provide training and guidance in a number of areas relevant to data science. The University of Arizona Libraries has built a successful data science support program, focusing on computational literacy, geographic information systems, and reproducible science. Success of the program has largely been due to the strength of library personnel and strategic partnerships with units outside of the library. Academic libraries can support campus data science needs through professional development of current staff and recruitment of new personnel with expertise in data-intensive domains.
INTRODUCTION AND LITERATURE REVIEW

The interdisciplinary field of data science is emerging as one of the fastest growing endeavors on and off college campuses. Data science, which applies concepts from statistics and computer sciences to questions in numerous domains, is driving novel discoveries across disciplines. Such applications include automated linguistic analysis for “just in time” action in online support forums (Kornfield et al., 2018), neuroimage processing for predicting health outcomes (Lancaster, Lorenz, Leech, & Cole, 2018), and relationships between environmental variation and childhood cognitive development (Stingone, Pandey, Claudio, & Pandey, 2017). However, applying such approaches requires specialized skills and resources, including computer programing and data visualization. Such skills are critical for the modern workforce (National Academies of Sciences, Engineering, and Medicine, 2018; Ridsdale et al., 2015). The absence of comprehensive data science support at many academic institutions is creating a growing list of unmet needs (Barone, Williams, & Micklos, 2017; Galanek & Brooks, 2018; Garcia-Milian, Hersey, Vukmirovic, & Duprilot, 2018; Oliver, 2017), preventing many researchers from taking advantage of the big data revolution.

A number of factors are contributing to the gap in skills and resources necessary to capitalize on data science applications. The first is the relative youth of data science; while the field has arguably existed for decades (Donoho, 2017), it has only recently attracted widespread attention across academic disciplines. Many researchers are only now realizing the necessity of skills development (Barone et al., 2017), and they are often in positions where formal training is especially challenging (e.g. post-doctoral or faculty positions). Also contributing to the gap is an absence of training opportunities, often due to a lack of interest in teaching “remedial” computer programing courses (Wilson, 2016). When training is offered, it often suffers from “silbing” observed on many campuses, where IT departments teach to one audience, libraries teach to another, and computer science departments teach to yet another. Efforts are underway to support researchers’ needs, such as Software and Data Carpentry workshops (https://carpentries.org), yet there remain significant deficiencies in resources to support the development of data science literacy on many, if not most, campuses (Galanek & Brooks, 2018).
Supporting data science efforts presents an opportunity for academic library engagement (Martin, 2016; Maxwell, Norton, & Wu, 2018; Wang, 2013). Libraries serve as hubs for college and university campuses, and like the field of data science, are intrinsically interdisciplinary. The audience for an academic library is also arguably larger than that of any other campus unit: the library serves the entire campus, including students, staff, and faculty. Supporting aspects of data science in the library is not new; indeed, libraries have been supporting data collection, management, and sharing best practices for many years (Antell, Foote, Turner, & Shults, 2014). Additionally, some academic libraries have been supporting researchers with training in computational and statistical skills (e.g. North Carolina State University: https://www.lib.ncsu.edu/services/data-visualization, New York University Libraries: https://library.nyu.edu/departments/data-services/, Arizona State University: https://lib.asu.edu/data). With this in mind, now is an opportune time for academic libraries to explicitly address the growing need for data science training.

The current work describes how the University of Arizona Libraries has been meeting data science needs by promoting skills and resources development. A variety of approaches, including internal and external partnerships have been essential in developing a successful program. Also included is a discussion on the challenge of providing support in data science efforts, primarily in the form of personnel expertise.

Supporting Data Science

The University of Arizona Libraries (UAL) has significantly expanded support for data science skills development in recent years. This includes hiring additional library staff, developing digital materials, and forming new partnerships with other intramural and extramural units. The success of these efforts is evident in growing participation in library offerings and has relied on determining which areas of data science are appropriate for the library to support and which areas are better suited for partners outside of the library.

Areas of Emphasis Within the Library

Investment in data sciences at UAL has supported growth in several areas, but three in particular have been especially popular on the University of Arizona campus. These heavily rely on the expertise of library staff, including deep domain knowledge in areas outside the library and information sciences. All three efforts have been successful by rapidly assessing and responding to campus needs and refining support offerings in response to the changing research landscape.
Computational Literacy

The need for computational skills is rapidly growing and is often exceeding the capacity of current training opportunities on college campuses (Wilson, 2016). The University of Arizona provides weekly workshops on the popular programming language R (R Core Team, 2018). These two-hour sessions alternate week-to-week between two different formats: (1) a hands-on instruction session, focusing on a specific topic such as the graphics package ggplot (Wickham, 2016) or an introduction to basic statistical analyses (https://jcoliver.github.io/learn-r); and (2) an open lab, where researchers bring their questions and code to work and get support in a positive, inclusive environment. The former (instruction session) are designed to provide “quick-wins”, whereby researchers can immediately apply what they learn to their own work. However, they also offer the opportunity to teach best practices regarding reproducibility, abstraction, and automation. The latter format (open session) has been especially supportive of the R community of practice at the University of Arizona - workshop participants often help one another to solve challenges in data analysis and visualization. These workshops align with many of the “Ten simple rules for biologists learning to program”, including “begin with the end in mind”, “develop good habits early on”, and “phone a friend” (Carey & Papin, 2018).

The design of R workshops has been an iterative process, with some implementations working better than others. The current workshop formats, alternating between focused instructional sessions and open lab sessions, arose after trying a variety of approaches, with varying success. One approach employed early on was to offer very general workshops, introducing the R language from a very conceptual approach, with topics like variables, data types, and control flow. This approach left many participants wondering how to apply what they learned to their own work. This led to a switch to the topic-specific workshops, which are still general enough for a broad audience, and for the most part designed for novices. Another aspect of the workshops that required refinement was the duration. Initial R workshops were two full-day sessions; this was too long for many researchers, and often had considerably high attrition by the middle of the second day. Furthermore, the two-day workshop had significant duplication with Software Carpentry (https://software-carpentry.org/) workshops being run on the University of Arizona campus. In contrast, attempts at one-hour instruction sessions often failed to provide enough material to be useful. The two-hour sessions provide an ideal compromise between enough time to be useful and short enough duration to allow many full-time researchers to participate.

In addition to formal workshops, UAL’s Data Science Specialist (an author on this work, JCO) provides in-depth consults in R data analyses
and visualization, focusing largely on support in the life sciences (the Data Science Specialist’s own background is in bioinformatics and ecology). The largest audience for these consults is graduate students, who often have little to no formal training in computational skills (Rubinstein & Chor, 2014; Valle & Berdanier, 2012). Balancing the level of involvement requires careful consideration - the library cannot offer a concierge data analyst service to every researcher on campus, but consults can provide an opportunity for true collaboration and deeper learning. Communicating clear expectations and understanding other campus resources that provide support, such as statistical consulting centers, is critical for meaningful, appropriate computational literacy support to the University of Arizona research community.

The University of Arizona Libraries also provides limited support for the Python programming language (Python Software Foundation, 2018). This limited support is largely driven by two factors. First, when computational literacy support programming was being developed, many life science researchers were requesting R support, rather than Python. Second, other campus units support Python programming, thus R programming represented a greater unmet need when the workshops were being developed. However, through conversations with R workshop participants it is increasingly evident that there remains an unmet need for Python skills development, especially in the life sciences. Should a library face the choice of offering R or Python support, assessing researchers’ needs would be invaluable in making an impactful decision.

GEOGRAPHIC INFORMATION SYSTEMS (GIS)

Cartography is often overlooked as one of the earliest forms of data visualization as well as the most prevalent. Visualizing spatial data on a map can fundamentally be thought of as plotting data in a well-defined domain. Spatial data differs from other data types in that its relevance, by definition, is tied to its position in the domain (where it’s located on a map). This can be thought of as the difference between using a table of attributes to generate a chart (non-spatial) and using a chart (the spatial component) to generate or add to a table (spatial). With this regard, most data produced in society or obtained from the natural environments are, or can be, regarded as geospatial data because they have a geographic location component. Data can span from the field of remote sensing with the exponentially increasing stream of satellite and aerial data all the way to social media sources such as Twitter, which includes geographic location of individual tweets.
The ubiquity and enormity of geospatial information which is produced today has spawned an increase in recognition and usage in the potential for spatial analysis across disciplines. An increase in the recognition of geospatial data types and better access to geospatial tools has led to a subsequent increase in the need for support of spatial data. Being able to provide support services specifically targeting geospatial data and software from a centralized campus entity like a library allows research in any college or discipline at the institution to benefit from that support.

Geospatial data support at UAL spans two broad, overlapping categories: desktop or online platforms specializing in GIS and GIS-focused application of scripting languages. While there is a sustained demand at the University of Arizona for comprehensive and heavy desktop software platforms such as ArcGIS (Environmental Systems Research Institute, 2018) and QGIS (QGIS Development Team, 2018), there is a prominent increase in desire to use web-based platforms such as ArcGIS Online (Environmental Systems Research Institute, 2018). Because of the simplicity of access, web-based GIS has allowed for faster and broader adoptions of geospatial analysis and cartography which has allowed entire classrooms to incorporate GIS into their curriculum with little instruction and a low overhead of software dependency. The use of online GIS platforms allows many users in the campus research environment to accomplish their cartographic and analysis needs without needing to embrace a fully comprehensive desktop GIS.

Given the increasing size of datasets, need for reproducibility, and efficiency provided through computational programming, many users in the research community have expressed a desire for GIS support of the R and Python scripting languages. This is especially true for GIS applications in the environmental sciences (Valle & Berdanier, 2012). Knowledge of the geospatial application of programming languages is also critical to leverage the computation resources necessary to address big-data research scenarios. Cloud computing and high performance/throughput computing resources provided by the National Sciences Foundation such as CyVerse (http://cyverse.org, Merchant et al., 2016) and XCEDE (Towns et al., 2010) enable researchers to address larger resources, but there is a discrepancy between usage and access. Such a deficit is often due to limited training opportunities for novice to intermediate users (Wilson, 2016; Ye, Brown, & Harding, 2013). Providing educational resources and events centered around programmatic tools and open-source software allows researchers to familiarize themselves with the tools and skills needed to use these systems.

Efforts to promote and assist in broadening the skillsets of researchers in geospatial analysis have focused on offering a variety of rotating introductory workshops in one- to two-hour sessions. Providing researchers
with software, data structure, and geospatial concepts in short format workshops has been more successful for the novice user base in large part because users are being introduced to languages and software for the first time and the overhead in teaching necessary background concepts can be substantial in geospatial materials. Concepts are made more digestible by partitioning lessons into focused content. Based on attendance numbers, there is a growing interest in workshops teaching open-source software, such as QGIS, GRASS (GRASS Development Team, 2017), and SAGA (Conrad et al., 2015). This is likely due to the fact that the established formal curriculum for GIS at the University of Arizona is centered on the use of the ArcGIS platform, with less attention given to open source options which students are curious about. Furthermore, because ArcGIS is not supported on the macOS or Linux operating systems, many open source options are platform-independent and thus are accessible to a broader audience of researchers on desktop machines and in the cloud.

At UAL, supporting more advanced analysis has typically fallen to one-on-one consultation due the customization and complexity of the particular application to the research question. Offering and scheduling one-on-ones is contingent on the availability of expertise, availability of meeting spaces, and the convenience for the patron. In an effort to provide more informal support, weekly open drop-in hours in a centrally located laboratory on campus have been successful in engaging researchers. Attendance at these sessions has also facilitated collaboration between geospatial researchers on campus in disparate fields.

While support for geospatial data analysis at UAL is continuing to grow, outreach and support strategies are evolving to match the needs of the research body on campus. In many cases, there is still a lack of understanding at the University of Arizona campus that UAL offers geospatial data support services. While an up-to-date website enumerating UAL options (http://libguides.library.arizona.edu/GIS) continues to be a successful avenue to discover what UAL has to offer in terms of geospatial data support, developing a network in the broader communities of practice has been just as effective. This includes ensuring that other service points at UAL are informed of geospatial data offerings as well as initiating and participating in frequent communication with other campus units that are providing geospatial data support.

Reproducible Science

As a cornerstone of the scientific process, the ability to verify research claims by repeating the process using one or all of the same methods, materials, data, and software, is of the utmost importance to not only maintain the credibility and reliability of research, but also to enable
others to build on the work. Although there is no universally accepted definition of the term “reproducibility” (Barba, 2018), it is used here to mean the ability to examine and verify the claims supported by the application of tools and methods from data science as defined in the introduction section. This of course implies that the researcher has provided enough detailed information such as how the analysis was conducted – the code, any needed computer or software settings, protocols used etc. There have been several studies in a wide variety of disciplines that have demonstrated how a large number of results in published studies cannot be reproduced (Aarts et al., 2015; Camerer et al., 2016; Ioannidis, 2005).

Data science and reproducibility go hand-in-hand. That is, the proper application of data science principles and tools enables reproducibility and adopting practices that encourage reproducibility (e.g., well-documented data collection, following data management best-practices) enables good data science. For example, to move towards research reproducibility, it is important to verify that the data science tools selected are applicable to the data and the problem at hand and to remain mindful of pitfalls such as p-hacking (selective reporting of data to obtain statistical significance) and HARKing (Kerr, 1998). When it comes to good data science, especially in the context of computational analyses, following best practices for the organization of data files and code, using interoperable file formats, following best practices for working with tabular data, and using literate programming (Knuth, 1984) all enable good data science. The final point, literate programming, which combines computer code and explanatory text in a single “living document,” has become increasingly accessible through tools such as Jupyter Notebooks (Kluyver et al., 2016) and the knitr (Xie, 2014) and rmarkdown (Allaire et al., 2018) R packages.

Although many academic libraries (including UAL) already have expertise to support one or more aspects of data-intensive and computational research, the knowledge is usually distributed amongst data management, geospatial, digital humanities, and data science specialists (Sayre & Riegelman, 2018). The idea of personnel dedicated to reproducibility is still quite novel, but the combination of reproducibility and data science support is a natural one (Steeves, 2017).

At UAL, support for reproducibility has taken the form of integrating best practices for data management, promotion of scripting/software to automate workflows, promotion of tools that support reproducible research (e.g., Jupyter notebooks), and advocating for open research practices into workshops and lectures. Target audiences have ranged from undergraduate students to faculty, with the exact nature of the content naturally varying depending on the audience. As an example of how research reproducibility is beginning to be integrated into instruction, a two-part workshop for postdocs dealing with topics such as data management
basics, data management planning, and funder/publisher policies was revamped to include material (including hands-on activities) on best practices for data archiving, curation, and publication (https://osf.io/td5jr). When targeting reproducibility topics to undergraduates, the approach was to introduce the concepts of open science, transparency, and reproducibility from the point of view of being a responsible researcher, focusing less on publication and compliance with data and code sharing policies and more on hands-on skills for enabling reproducibility. These included an introduction to doing analyses using the R programming language and a hands-on activity for how to use the Open Science Framework (OSF) to organize research (Spies, 2013).

Hands-on activities with the OSF have been especially popular in workshops across all audiences, with many attendees mentioning the potential utility of the platform for their own work. The UAL has also hosted workshops from the Center for Open Science (the developers of the OSF) to speak on reproducible research and the OSF. Although the number of University of Arizona-affiliated users on the OSF has increased by 40% likely as a result of these activities, combined with the fact that projects on the OSF are private by default, it is difficult to quantify if the initial enthusiasm from new users translates to sustained use. Nevertheless, examining the publicly visible projects affiliated with the University of Arizona shows there are signs that researchers are finding the OSF a useful vehicle for sharing their finished work in a more reproducible way.

The design of reproducibility workshops has been an iterative process, and interspersing short presentations along with hands-on exercises has been the most effective format. Participants are more interested in attending workshops that meet a practical need, such as how to write a data management plan, how to find data in a data repository, or how to use the Open Science Framework. It is much easier for participants to apply knowledge and skills if they have the opportunity to try out different tools and best practices.

In addition, the UAL sponsors programs during International Open Access Week (http://www.openaccessweek.org/) and Love Data Week (http://lovedataweek.org/). Open Access Week programs have included issues around reproducibility, computational science, and visualizing “big data”, while the Love Data Week 2018 program engaged researchers to share how data facilitates telling a story about their research.

Although efforts to promote and enable reproducible research by the UAL are still in their infancy, continuing and expanding such efforts in conjunction with campus and external partners is an ongoing goal. These include workshops focused specifically on “doing” reproducible research using established tools, exploring new tools such as Code Ocean (https://codeocean.com/), and partnering with the University of Arizona’s
Research Computing office to develop training for applying data management and reproducibility best practices specific to high performance computing resources.

Promotion and Advertising

A constant consideration, not unique to data science services, is that of finding the appropriate media for effectively communicating library services with the audiences that would benefit from those services. In many cases, once a critical mass has been reached, word of mouth has proven very effective in increasing participation in drop-in sessions. Colleague recommendations have driven many new users to library services, including R workshops and consultations, GIS support, and assistance with data management practices. However, reaching that critical mass can be challenging. Library websites provide some details on events and services, but require the audience to already be looking for those services on the library web page. At UAL, the best means of self-promotion has been e-mail messaging. Joining mailing lists serving key communities on campus (e.g. graduate students, GIS users, individual departments) has allowed broadcasting services to broad, yet appropriate audiences. Working with coordinators of specific majors, graduate programs, or departments has also been effective for forwarding announcements for workshops or training events. While this can add some administrative burden to library staff (i.e. sending an e-mail to each of fifteen different program coordinators), it has worked better than impersonal, en masse e-mail messages. Finally, it has been critical to maintain constant communication with liaison colleagues at UAL, who are then able to promote data science support to their constituent academic units.

Partnerships Outside the Library

Recognizing that no one academic unit could provide comprehensive data science support, UAL has formed several partnerships with the intent of improving data science literacy. The partnerships vary in the type and amount of resources provided by the library and include instructional space, marketing and promotion, and expertise of library staff. In terms of space, UAL has several contemporary learning spaces well-equipped for workshops involving in-person and remote instruction on data science tools. In collaboration with University Information Technology Services (UITS), UAL has hosted workshops on machine learning, big data analyses with Hadoop, and an introduction to computing with graphical processing units (GPUs). Workshop registrations for these popular topics routinely reach capacity within hours of being announced. Partnerships such as
these highlight a growing interest in library spaces a critical contributions to data science education. As “neutral ground,” outside any one department’s or unit’s physical space, the library serves a natural nexus for interdisciplinary collaboration for data science education (Moore-Sloan Data Science Environments, 2018). Library spaces are also increasingly important for “hack”-style events, which are growing in impact and popularity in data science education (Huppenkothen et al., 2018); UAL hosts three such annual events, the Women’s Hackathon (https://womenshackathon.arizona.edu/), Hack Arizona (https://new.library.arizona.edu/events/hack-arizona), and the Research Bazaar (https://resbazaz.github.io/resbaz2018/tucson/).

To address campus-wide data science literacy, UAL partners with BIO5 (http://www.bio5.org) and the National Science Foundation’s cyberinfrastructure project, CyVerse (Merchant et al., 2016) to sponsor Software and Data Carpentry workshops at the University of Arizona campus. Briefly, Carpentry workshops are intensive, two-day workshops that provide a hands-on introduction to assorted data science approaches. Rather than duplicating efforts in lesson development, all Carpentries workshops are taught from open instructional material available under a Creative Commons Attribution license. Another strength of Carpentries workshops is they are led by volunteers who have undergone the Carpentries Instructor Training program (https://carpentries.github.io/instructor-training/), which focuses on the science of learning. UAL has recently increased the number of badged Carpenter instructors, and one of the authors of this work (JCO) is part of the Carpentries Instructor Training team.

Several units across the University of Arizona offer support in data sciences, yet “siloing” hinders communication to audiences outside the core constituencies of each respective units. To address these communication gaps, and strategically develop new data science opportunities, UAL partnered with CyVerse and the Data Science Institute (https://datascience.arizona.edu) to form the Data Science Resources and Training (DSRT) group. A grass-roots organization, the DSRT steering committee advocates for data science education and oversees a clearinghouse of training events and resources at the University of Arizona (https://datascience.arizona.edu/dsrt). The clearinghouse uses a controlled vocabulary of data science topics, allowing researchers to identify those opportunities relevant to their work. The diversity of resources enumerated on the DSRT clearinghouse website illustrates the importance of interdisciplinary collaboration in developing a data-literate campus. As more data science centers and institutes are established at institutions of higher learning, the academic library has a valuable opportunity to collaborate towards improved information sharing and elevating campus data science training resources.
Two of the authors (CK and JCO) are members of the Research Computing Governance Committee (RCGC, http://rcgc.arizona.edu/), the campus governing committee that guides the development of central research computing resources and determines how to best serve the needs of researchers across campus. Subcommittees include research data management and curation, big data, campus networking, high performance computing, and research computing policies. UAL involvement with RCGC has provided an opportunity to be involved with researcher’s computing issues and helped establish and reinforce partnerships with other campus units.

The UAL also partners with the Office of Research, Development and Innovation (RDI, https://research.arizona.edu/) in a number of different capacities to facilitate campus data literacy. Each semester, UAL teaches a data management workshop, at the request of RDI Responsible Conduct of Research Program and has taught a research data management session as part of grant proposals workshops offered by RDI Research Development Services (RDS). In addition, RDS refers researchers to the UAL for assistance writing data management plans as part of grant proposals. The Public Access Working Group composed of representatives from the UAL and RDI developed a website to provide guidance on complying with policies for public access for federally funded research (http://new.library.arizona.edu/research/funder-requirements) and made campus presentations related to federal agency changes as a result of the 2013 Office of Science and Technology Policy’s “Increasing Access to the Results of Federally Funded Scientific Research” Memorandum (https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf). These partnerships with RDI illustrate how academic libraries’ work with research offices can help support a culture of open science across campus.

Personnel Needs

Developing data science support at an academic library is predicated on having library staff qualified to provide support. Generally, there are two principal means of expanding expertise within departmental units: training existing personnel and hiring staff anew. Both present challenges, including financial and cultural considerations.

The first option for developing expertise at a library is through training of current staff within a culture of professional development. This is being made increasingly easier through the growing number of opportunities for librarians to develop data science skills. For novice audiences, the Library Carpentry workshops (http://librarycarpentry.org/, Baker et al., 2016) provide an introduction to data and analytical skills. These face-to-face
workshops are generally two-day events that include hands-on lessons on topics such as the Unix shell, version control with Git, or the Python programming language. The University of Arizona hosted Library Carpentry workshops in August 2017 and 2018 (https://jcoliver.github.io/2017-08-09-lc-tucson, https://jcoliver.github.io/2018-08-06-lc-tucson/). These workshops have been well received at the University of Arizona and inspired librarians at UAL to contribute to lesson development in the 2018 Library Carpentry Mozilla Global Sprint (http://librarycarpentry.org/blog/2018/05/16/our-latest-sprint/). For those librarians who possess basic computational competencies, the companion Software and Data Carpentry workshops (http://carpentries.org) provide a deeper dive into research computing skills, but are still designed for primarily novice audiences. The intent of all three Carpentries workshops are not necessarily to make participants proficient practitioners, but to create a community of learners empowered find the answers to data and coding questions on their own (Wilson, 2016).

The short format of Carpentries-style workshops provide a starting point for computational skills development, but there are few additional options that provide a deeper dive into data science topics. However, one outstanding option for in-depth skills development is the Data Science and Visualization Institute for Librarians (DSVIL) at North Carolina State University (https://www.lib.ncsu.edu/data-science-and-visualization-institute). This week-long course focuses on library-related applications of data science. In part an extension of the Data Science Training for Librarians course (http://www.dst4l.info/), the DSVIL curriculum includes data visualization, web scraping, and network analysis. Burton, Lyon, Erdmann, and Tijernia (2018) provide considerable discussion about the opportunities and challenges of skills development of library staff.

In some cases, current library staff may not have the time for developing deep enough skills to provide thorough data science support (Burton et al., 2018). In such cases, hiring additional staff can provide significant value to the type of data science support an academic library can provide. Master in Library Science (MLS) programs focused on data science provide one training route for data literate library and information science professionals. For example, the Indiana University School of Informatics, Computing, and Engineering includes a data science specialization in the MLS degree program (https://www.sice.indiana.edu/graduate/degrees/information-library-science/dual-degrees/data-science-mls.html). With coursework including machine learning and cloud computing, programs such as this provide the specialized training that would allow library personnel to collaborate with researchers in largely underexplored modes.

Alternatively, libraries could hire specialists with education and expertise outside of the library sciences (Maxwell et al., 2018). Such an approach
has the potential to create “credentialing tension” (Burton et al., 2018), but may be especially effective for developing support in more science-oriented areas such as data analysis and visualization. Hiring non-MLS professionals into library positions may also require an institutional cultural shift, as the work of data scientists is often considered the domain of other units on campus (Tenopir et al., 2015). However, the expertise in quantitative skills of data-intensive research afforded by personnel without formal library training can be invaluable in supporting data science through the academic library.

Upshot

The opportunities to engage with researchers in the realm of data science are rapidly growing and should be embraced by academic libraries to maintain relevance in research and academic support. Given interest in “big data” approaches, the only question for academic libraries is “how”. Assessing campus data science literacy needs can guide the mode and type of personnel development, either training existing staff or hiring new staff. The pace of change in data science also requires agility to try new services, sometimes even when those services disrupt the model of the “traditional” library. Support from library administration for data science “skunkworks”, like those advocated for digital scholarship (Nowviskie, 2013), affords freedom to capitalize on new opportunities as they arise. The expanding data science ecosystem offers a new mode of engagement with campus researchers that would solidify the academic library’s position as a campus hub for the data-driven twenty-first century.

REFERENCES

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J. …, Open Sci Collaboration. (2015). Estimating the reproducibility of psychological science. Science, 349, aa4716. doi:10.1126/science.aac4716
Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Chang, W. (2018). Rmarkdown: Dynamic documents for R. Retrieved from https://CRAN.R-project.org/package=rmarkdown
Antell, K., Foote, J. B., Turner, J., & Shults, B. (2014). Dealing with data: Science librarians’ participation in data management at association of research libraries institutions. College and Research Libraries, 75, 557–574. doi:10.5860/crl.75.4.557
Baker, J., Moore, C., Priego, E., Alegre, R., Cope, J., Price, L., … Wilson, G. (2016). Library carpentry: Software skills training for library professionals. Liber Quarterly, 26, 141–162. doi:10.18352/lq.10176
Barba, L. A. (2018). Terminologies for reproducible research. CoRR. Retrieved from http://arxiv.org/abs/1802.03311
Barone, L., Williams, J., & Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computational Biology, 13*, e1005755. doi:10.1371/journal.pcbi.1005755

Burton, M., Lyon, L., Erdmann, C., & Tijernia, B. (2018). Shifting to data savvy: The future of data science in libraries. University of Pittsburgh. Retrieved from http://d-scholarship.pitt.edu/id/eprint/33891

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*, 1435–1436. doi:10.1126/science.aaf0918

Carey, M. A., & Papin, J. A. (2018). Ten simple rules for biologists learning to program. *Plos Computational Biology, 14*, e1005871. doi:10.1371/journal.pcbi.1005871.

Conrad, O., Bechtel, B., Bock, M., Dietrich, H., Fischer, E., Gerlitz, L., … Böhner, J. (2015). System for automated geoscientific analyses (SAGA) v. 2.1.4. *Geoscientific Model Development, 8*, 1991–2007. doi:10.5194/gmd-8-1991-2015

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics, 26*, 745–766. doi:10.1080/10618600.2017.1384734

Environmental Systems Research Institute. (2018). ArcGIS. Retrieved from https://www.esri.com/en-us/store/arcgis-desktop

Galanek, J. D., & Brooks, D. C. (2018). Supporting faculty research with information technology. Research report. Educause Center for Analysis and Research.

Garcia-Milian, R., Hersey, D., Vukmirovic, M., & Duprilot, F. (2018). Data challenges of biomedical researchers in the age of omics. *Peer Journal, 6*, e5553. doi:10.7717/peerj.5553

GRASS Development Team. (2017). Geographic resources analysis support system (GRASS GIS) software, version 7.2, Open Source. Geospatial Foundation. Retrieved from http://grass.osgeo.org

Huppenkothen, D., Arendt, A., Hogg, D. W., Ram, K., VanderPlas, J. T., & Rokem, A. (2018). Hack weeks as a model for data science education and collaboration. *Proceedings of the National Academy of Sciences, 115*, 8872–8877. doi:10.1073/pnas.1717196115

Ioannidis, J. P. A. (2005). Why most published research findings are false. *Plos Medicine, 2*, 696–701. doi:10.1371/journal.pmed.0020124

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217. doi:10.1207/s15327957pspr0203_4

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter notebooks – A publishing format for reproducible computational workflows. Paper presented at the Positioning and Power in Academic Publishing: Players, Agents and Agendas, 87. doi: 10.3233/978-1-61499-649-1-87. Retrieved from http://ebooks.iospress.nl/publication/42900

Knuth, D. E. (1984). Literate programming. *Computer Journal, 27*, 97–111. doi:10.1093/comjnl/27.2.97

Kornfield, R., Sarma, P. K., Shah, V., Dhavan, McTavish, F., Landucci, G., Pe-Romashko, K., & Gustafson, D. H. (2018). Detecting recovery problems just
in time: Application of automated linguistic analysis and supervised machine learning to an online substance abuse forum. *Journal of Medical Internet Research*, 20, e10136. doi:10.2196/10136

Lancaster, J., Lorenz, R., Leech, R., & Cole, J. H. (2018). Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in Aging Neuroscience*, 10, 28. doi:10.3389/fnagi.2018.00028

Martin, E. R. (2016). The role of librarians in data science: A call to action. *Journal of eScience Librarianship*, 4, e1092. doi:10.7191/jeslib.2015.1092

Maxwell, D., Norton, H., & Wu, J. (2018). The data science opportunity: Crafting a holistic strategy. *Journal of Library Administration*, 58, 111–127. doi:10.1080/01930826.2017.1412704

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant collaborative: Cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS Biology*, 14, e1002342. doi:10.1371/journal.pbio.1002342

Moore-Sloan Data Science Environments. (2018). Creating institutional change in data science. Retrieved from http://msdse.org/creating_institutional_change.html

National Academies of Sciences, Engineering, and Medicine. (2018). *Data science for undergraduates: Opportunities and options*. Washington, DC: The National Academies Press. doi:10.17226/25104

Retrieved from https://www.nap.edu/catalog/25104/data-science-for-undergraduates-opportunities-and-options

Nowviskie, B. (2013). Skunks in the library: A path to production for scholarly R&D. *Journal of Library Administration*, 53, 53–66. doi:10.1080/01930826.2013.756698

Oliver, J. C. (2017). Bioinformatic training needs at a health sciences campus. *PLoS One*, 12, e0179581. doi:10.1371/journal.pone.0179581.

Python Software Foundation. (2018). Python language reference. Retrieved from https://www.python.org

QGIS Development Team. (2018). QGIS geographic information system. Retrieved from http://qgis.osgeo.org

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., … Wuetherick, A. B. (2015). Strategies and best practices for data literacy education: Knowledge synthesis report. Dalhousie University. Halifax, Nova Scotia, Canada. doi:10.13140/RG.2.1.1922.5044

Rubinstein, A., & Chor, B. (2014). Computational thinking in life science education. *PLoS Computational Biology*, 10, e1003897. doi:10.1371/journal.pcbi.1003897

Sayre, F., & Riegelman, A. (2018). The reproducibility crisis and academic libraries. *College & Research Libraries*, 79, 2–9. doi:10.5860/crl.79.1.2

Spies, J. R. (2013). The open science framework: Improving science by making it open and accessible. Retrieved from https://dl.acm.org/citation.cfm?id=2539283
Steeves, V. (2017). Reproducibility librarianship. *Collaborative Librarianship, 9*, 4. Retrieved from https://digitalcommons.du.edu/collaborativelibrarianship/vol9/iss2/4

Stingone, J. A., Pandey, O. P., Claudio, L., & Pandey, G. (2017). Using machine learning to identify air pollution exposure profiles associated with early cognitive skills among US children. *Environmental Pollution, 230*, 730–740. doi:10.1016/j.envpol.2017.07.023

Tenopir, C., Allard, S., Frame, M., Birch, B., Baird, L., Sandusky, R., … Lundeen, A. (2015). Research data services in academic libraries: Data intensive roles for the future? *Journal of eScience Librarianship, 4*, e1085. doi:10.7191/jeslib.2015.1085

Towns, J., Andrews, P., Boisseau, J., Roskies, R., Brown, J., Boudwin, K., … Wallnau, K. (2010). Project summary: XSEDE: eXtreme science and engineering discovery environment. Retrieved from http://hdl.handle.net/2142/42546

Valle, D., & Berdanier, A. (2012). Computer programming skills for environmental sciences. *Bulletin of the Ecological Society of America, 93*, 373–389. doi:10.1890/0012-9623-93.4.373

Wang, M. (2013). Supporting the research process through expanded library data services. *Program: Electronic Library and Information Systems, 47*, 282–303. doi:10.1108/PROG-04-2012-0010

Wickham, H. (2016). Ggplot2: Elegant graphics for data analysis. New York: Springer-Verlag. Retrieved from http://ggplot2.org

Wilson, G. (2016). Software carpentry: Lessons learned. *F1000Research, 3*, 24. doi:10.12688/f1000research.3-62.v2

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch & R. D. Peng (Eds.), *Implementing reproducible computational research*. Boca Raton, FL: Chapman and Hall/CRC. Retrieved from http://www.crcpress.com/product/isbn/9781466561595

Ye, H., Brown, M., & Harding, J. (2013). GIS for all: Exploring the barriers and opportunities for underexploited GIS applications. Paper presented at the Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings, Nottingham, UK (Vol. 13). doi:10.7275/R5W66j0K. Retrieved from https://scholarworks.umass.edu/foss4g/vol13/iss1/4