Data Science Drivers: A Report on the National Library of Medicine Workshop

Maryam Zaringhalam, PhD, AAAS Science & Technology Policy Fellow

Lisa Federer, MLIS, Data Science and Open Science Librarian

Mike Huerta, PhD, Associate Director of NLM for Program Development and NLM Coordinator of Data Science and Open Science Initiatives

Executive Summary

This report provides a summary of the discussions and major themes that emerged from the Data Science Drivers Workshop convened by the National Library of Medicine (NLM) on September 20, 2018. Data science experts from different sectors and domain areas were invited to help increase NLM's understanding—and that of the National Institutes of Health, more broadly—of what drives data scientists to work with data from a particular domain. Insights from this workshop are expected to inform efforts of NLM and NIH to incentivize and attract data scientists not currently working with biomedical data to do so. Their participation would both bring fresh approaches and perspectives to accelerate discovery and advance health, as well as present data scientists with significant and interesting challenges with which to enhance progress in data science. Topics of discussion included:

1. **Characterizing biomedical data science**, particularly with respect to developing standardized, but still flexible, training for the next generation of data scientists.
2. **Responsible conduct of biomedical data science research**, including a discussion around ethics, informed consent, and the need to raise awareness around socially responsible use and analysis of data early in a data science training.
1. **Fostering diversity and inclusion in the data science training and workforce development** to ensure the data science workforce reflects the population it seeks to serve and to build better and more efficient teams.
2. **Promoting collaboration between domain experts and data scientists** by encouraging co-equal partnerships built around collaboration and communication.
3. **Funding opportunities**, both to independently attract data scientists to biomedical data science problems and to create opportunities for collaboration between data scientists and biomedical domain experts.

The report concludes with a set of findings regarding repositories and data providers, subject matter experts interested in working with data scientists, trainers and educational institutions,

and funders. These findings provide insight into how these various stakeholders can enhance the biomedical data science workforce by encouraging data scientists to engage with biomedical researchers, datasets, and science.

Background & Purpose

Over the last decade, data science has pervaded many sectors of society and has become an increasingly salient component of biomedical science. New technologies that allow for the generation of vast amounts of high quality digital biomedical data have proliferated, creating new opportunities for data science to contribute to biomedical understanding and improved health.

The potential for data science to advance biomedical research, especially in the context of more open research paradigms, is made clear in the Strategic Plan of the National Library of Medicine (NLM) and the National Institutes of Health (NIH) Strategic Plan for Data Science. The former envisions NLM as becoming a platform for data-driven discovery and data-powered health, while the latter provides a roadmap to modernize and build out the NIH-funded biomedical data science ecosystem.

On September 20, 2018, NLM convened a group of data scientists for a roundtable workshop to gain insights into two broad categories of questions centered on data science training and drivers for data science collaborations. **The overall purpose was to help NLM and NIH better understand how to attract data scientists to work on biomedical data problems**. Eleven experts from academia, industry, and the nonprofit sector, representing a diversity of gender, racial, ethnic, and geographic makeup, were invited to participate. In order to get a fresh perspective on what can attract data scientists to biomedical research questions, most of the participants had no prior experience working with biomedical data and no existing relationship with the NIH as a funding institution. Participants and their affiliations are listed in Appendix A.

The day was organized into discussion sessions, with each driven by a set of questions in a particular area, including: (1) the training and career trajectory of each of the participants and how they choose which domains to work with; (2) the participants' experiences with the institutional and funding environments that incentivize and drive research and collaboration; (3) ideas on how to facilitate collaboration between data science and domain experts, as well as how students can receive training that promotes collaboration. The schedule, including sample discussion questions, can be found in Appendix B.

Major Discussion Themes

Characteristics of Data Science as a Discipline

Data science is a relatively new field that draws from a variety of disciplines, methodologies, and approaches. By convening data scientists who bring their expertise to bear on data from a variety of disciplines and in different sectors of society, NLM sought to better understand data science as a field, with an eye to understanding what skills and knowledge characterize a data scientist. Broadly, data science is conducted at the intersection of subject matter knowledge, computer science, statistics, and data management. Data scientists often collaborate with others who have deep subject matter knowledge, and frequently engage in exploratory research, often combining datasets to answer a particular domain question.

Data science is a rapidly growing discipline

Data science is a rapidly growing field, in part because of the massive amounts of data being generated across domains and sectors. Because of the growing volume of data—and, with it, the increasing opportunities to answer new and exciting questions—the demand for data scientists continues to rise (Henke et al., 2016). Data scientists are also afforded a unique degree of flexibility to move across domains, using data science tools, methodologies, and approaches to answer a wide range of questions with data from diverse sources, disciplines, and sectors.

Data science training methodologies

Given the high demand for data scientists and the wide applicability of the field, there is a growing need to develop training opportunities in data science. Throughout the discussion, participants noted that data science is still not particularly well-defined, even among those who identify as data scientists. There are a variety of reasons that contribute to this lack of clarity. There is no authoritative "Data Science" textbook around which to build a standardized curriculum, but identification and agreement around core skills required of a data scientist might be useful to guide development of curricula.

Despite agreeing on the utility of identifying core skills, participants emphasized that there is no single way of doing data science. Depending on the particular research question or domain at hand, a data scientist may use different methodologies, as well as different tools and programming languages. Importantly, data science tools, methodologies, and approaches—even to the same question—evolve and change quickly. As a result, specific training needs will likely be driven by the questions being addressed, and skills will need to be continually updated.

Despite difficulty in precisely defining the discipline of data science, some standard methodologies have been offered. For example, the Cross Industry Standard for Data Mining

(CRISP-DM) is a standardized data science methodology, though it is not necessarily well known in the data science community. CRISP-DM is an iterative, step-by-step methodology for data mining projects that was developed in 1996 and has become a widely used analytics model.

Publication in data science is competitive and often through conferences

The rapid growth of the data science discipline has made publication, particularly in the area of machine learning, highly competitive. The primary venue for data scientists to share their work is at conferences rather than in journals, reflecting the dynamic nature of data science. The acceptance rate for prestigious conferences is only about 10 percent. Participants also noted that the competitiveness of the publication process can introduce bias, with researchers from more prestigious institutions, those with more seniority, or those with more connections within the field being favored in peer review. These biases disadvantage researchers from less prominent institutions, particularly in geographic regions with less access to computing infrastructures, and those from underrepresented backgrounds. The resulting lack of visibility and access to recognition can prevent researchers from attracting more funding for their work, as well as developing and accessing infrastructures to further enable their research.

Participants pointed out that the volume of papers being submitted for publication also introduces a concern around proper quality control of the articles that do make it to publication. Because reviewers are inundated with submissions, they may not have adequate time to dedicate to ensuring that the research was rigorous, the appropriate experimental controls were used, appropriate data collection and analysis methodologies were used, and the conclusions are reasonable given the limitations of the data and methods. In this regard, publication of data science is similar to that of biomedical science.

Because data science is so quickly evolving, data scientists find it difficult to stay up-to-date on ongoing developments. Participants also noted that social media platforms like Twitter have become a popular way to communicate and learn about recent developments because of the instantaneous nature of the medium.

Ethics in Data Science

Participants emphasized the importance of ethics in the discipline of data science. They noted that there are many ethical implications that come from drawing conclusions based on multiple large and varied datasets for which the data were collected for very different reasons. Data collected from or about people, which is the focus of much data science work, amplify ethical concerns. Consideration and awareness of issues of privacy, bias, and discrimination are extremely important at every step in the research and analytic process. These considerations should include understanding the limits of the data being analyzed—including how and for what purposes the data were collected, what data points are included in the datasets, and implications of inferences made on the basis of combining or analyzing varied datasets—as

well as thinking through the consequences and implications of a particular line of inquiry. Participants further noted that data scientists should be trained early in their careers in the responsible use of data.

Issues of ethics can also arise in the context of collaborations if collaborators have different expectations and practices around ethics. Participants highlighted the importance of establishing a shared set of expectations at the start of a collaboration around considerations like how the data will be used and the conclusions disseminated. While conversations about ethics and responsible conduct of data science can be challenging, the participants noted that making assumptions about collaborators' motivations and intentions can create more difficult points of friction later down the line.

Informed consent

Informed consent was one of the primary ethical issues that participants discussed. Data science research is often more exploratory than hypothesis-driven, with data scientists gaining access to datasets and using them to pursue novel research and analytic questions. When those datasets are collected from human subjects without the appropriate privacy protections in place, research conclusions may have unintended consequences for the subjects. Informed consent is designed to protect subjects by informing them about what their data will and will not be used for. Placing constraints on how data can be used in turn places constraints on what lines of research can reasonably be pursued.

Developing appropriate informed consent measures is not a trivial task. Data are increasingly collected from a variety of sources that are integral to people's everyday lives, such as wearable technologies, smartphones, search engines, and social media. Determining ways to properly consent individuals for the use of their data presents a range of logistical problems. Terms of use policies are extremely long and full of technical legal language, so individuals often skim through them without understanding what types of data use they are consenting to. In addition, determining the scope for consent is difficult. On one extreme, individuals may be asked to consent to their data being used for any purpose at all. On the other extreme, potential researchers may be asked to specifically state and qualify their research questions, which restricts future avenues of inquiry.

Participants also raised concerns around the use of geospatial data. The ability to track an individual's movements with fine resolution can reveal an individual's personal identity and ultimately infringe on their right to privacy. One participant noted that researchers and individuals often assume that laws exist to protect individuals' right to privacy, with adequate restrictions in place to constrain the proper use of data. In reality, however, very few regulations protecting privacy exist. While geospatial data may be used to inform decisions on a range of important topics—like determining how best to deliver medical support to an area affected by disaster or understand how an epidemic has spread across a region—in the wrong

hands, geospatial data may be used for unethical purposes, such as to discriminate against historically marginalized groups living in a particular geographic area.

Despite these challenges, workshop participants did note that the issue of informed consent provides an opportunity to better engage with and educate the public to increase data literacy, or what one participant termed "data readiness." Data readiness can cover a range of topics—from how individuals can engage with their own data to make better informed decisions to understanding what protections the law truly provides them around how their data may be used.

Accounting for bias in data collection

Data collection requires setting inclusion and exclusion criteria. These decisions may introduce bias, particularly when research involves people, which can widen existing participation disparities. For instance, medical genomics has historically focused on people of European descent. Therefore, a substantial portion of the global population is excluded from benefiting from advances based on that foundation. Generalized conclusions cannot be drawn from biased datasets, and meaningful biomarkers in historically underserved populations may go undiscovered, thus widening the gap in health disparities. Participants noted that it is quite challenging to de-bias existing datasets and generalize to diverse populations. NIH is currently working towards increasing representation of women and minorities in biomedical datasets with the launch of the *All of Us* research initiative to more equitably uncover paths to precision medicine. NLM's National Network of Libraries of Medicine serves as the initiative's partner, providing high quality information about precision medicine to communities underrepresented in biomedical research.

The conclusions drawn from a particular dataset are necessarily limited by the decisions made during the course of data collection. Participants noted that data scientists must take the limitations of the data into consideration in designing and conducting the analysis, as well as ensuring that they report conclusions with the appropriate caveats and qualifications. Participants also urged caution when using proxy variables to measure difficult-to-quantify variables of interest, like using FitBit activity as a proxy for exercise to determine insurance premiums.

Participants emphasized the need for data scientists to make clear what the intended goals of their research are and what the potential unintended consequences might be. What kinds of data should be included in the study and from whom should they be collected? How can the limitations of the data, their implications for conclusions and their applications be clearly communicated? To promote transparency and allow others to independently assess bias, participants suggested making available that are used to arrive at a particular conclusion.

Supporting Equity, Diversity, and Inclusion in Data Science

Throughout the workshop, participants emphasized the importance of equity, diversity, and inclusion in the discipline of data science as a key to not only ensuring equitable progress in the field, but also mitigating bias and enhancing innovation by having a diversity of perspectives tackling a range of research and analytic questions. The conversation around diversity included recruiting and retaining individuals from different races, ethnicities, genders, and regions of the country, particularly those from historically underserved or underrepresented populations and areas.

Infrastructure

Data science requires sufficient technical infrastructure, such as storage systems for processing and storing big data, high performance computing systems, and high speed servers for computable notebooks. Such infrastructure does not exist universally and is difficult to obtain access to by under-resourced institutions. Equitable access to such infrastructure, however, is a key ingredient in promoting equity and diversity in the field of data science. One participant pointed to Canada's efforts to provide equitable access to computational infrastructure as a model for promoting equity and diversity. In 2017, Compute Canada and the Pacific Institute for the Mathematical Sciences partnered to launch a pilot that would provide access to JupyterHubs—a computing and data analysis platform—to universities across Canada.

Implicit bias training

Workshop participants highlighted the need for training on implicit or unconscious bias to support inclusion of underrepresented groups within data science. Such training can raise awareness about the ways that bias enters researchers' everyday thinking and perceptions of others, as well as encouraging researchers to look out for and interrogate their biases.

Participants cautioned, however, that diversity training is only helpful for people who already are primed on its importance and are willing to listen and question their own unconscious biases. Institutions and workplaces must therefore work to weave the importance of diversity into the culture. One suggestion was to promote inclusion riders, or contractual obligations tied to funding or promotion to leadership positions, for instance, that would guarantee equitable hiring practices. There is, however, a difference between complying with diversity-promoting mechanisms and actively fostering inclusion through workplace culture. In other words, inclusion riders may get people from underrepresented backgrounds through the door, but an inclusive environment does not exist if their contributions are not adequately listened to and incorporated in the research practice.

Participants also raised concerns around ageism, as mid- and late-career stage people have gone back to school to get training in data science and find themselves unable to find work.

Participants suggested that this pool of data scientists with non-traditional backgrounds might be good candidates for NIH to attract and recruit as it strengthens its data science workforce.

Fostering Collaboration with Data Scientists

Given that workshop participants have engaged in extensive collaborations and have consulted with different stakeholders, NLM leveraged their expertise to better understand how biomedical researchers can more effectively engage with data scientists—both as active collaborators and as providers of data. Collaboration would have bidirectional benefits, with data science providing insight to biomedicine and biomedicine presenting novel challenges to data science.

Best practices for effective teamwork

Participants discussed effective models for collaboration between data scientists and subject matter experts. Successful models included those offered by the National Cancer Institute's Collaboration and Team Science Field Guide, and the Research-Practice Partnership (RPP) Development Workshops, which were designed to support and build partnership capacity in the *CS for All* community. Both models emphasize trust within a team, effective communication skills, and the importance of mentorship within the group. Participants emphasized the need for "connectors" within a team, or individuals who ensure that channels of communication remain open and build trust among team members. Connectors make the collaboration feel like a safe space for exploring ideas, setting priorities, and airing concerns. In interdisciplinary collaborations, a connector may act as a translator between parties, setting a common language for the collaboration. This role is particularly important in ensuring that all parties understand the strengths and weakness of the data, the methodology for analysis, and the ultimate conclusions.

Participants also discussed potential points of friction within a given collaboration. They noted that data scientists often feel as though they are being hired onto the team as a service. Conversely, domain scientists may feel that they have been brought into a collaboration because data scientists view their (valuable) datasets as commodities that can be freely used. Compiling a dataset generally represents extensive effort on the part of the domain scientist, and this is not seen as commensurate with the work of the data scientist who has not been involved since the beginning of the project. When one party is perceived as a subordinate of the other rather than an equal partner, the collaboration suffers. Such a dynamic hinders the kind of effective, two-way conversation that enriches the collaboration. Another point of friction arises when team members fail to set a definition for success at the start of the collaboration. For instance, biomedical researchers may define success as publication in a high-impact biomedical journal, while their data scientist counterparts may define success as a presentation at a data science conference or development of a patentable technology.

Good data facilitate collaboration

Collaboration between subject matter experts and data scientists is facilitated by high quality data. Data scientists must be able to understand how the data were collected and where the gaps in a particular dataset may exist. Cleaned reference datasets that are easy to interrogate from well known databases with proven data sources, such as Kaggle and the UCI Data Repository are of particular interest to data scientists. Participants did note, however, that there is no one measure or set of criteria by which a particular dataset is considered clean or good enough. Instead, judgements about the quality of a dataset are context- and user-dependent, based, for example, on the data scientist's level of subject matter expertise and the ultimate goal of analysis.

There are, however, ways for data scientists to get a sense of the quality of a dataset. Datasets that have already been used extensively by others in their field and which have appeared in much of the published literature are more likely to be high quality. Data scientists also tend to use and re-use datasets that have been benchmarked for specific tasks. Workshop participants noted that a common publication type is to publish a more efficient methodology for performing an established task on a specific dataset commonly used in evaluating tools for that task. Participants also cited red flags to look for in a dataset, including signs that privacy and safety were not fully considered.

Participants noted that finding new, high quality datasets and sources remains a challenge. There is so much data available in any given field that it is difficult to sort through and determine quality and usability. They noted in particular that data scientists in industry may be wary of using untested data that has not been discussed widely in the data science literature. They are particularly hesitant to use government data due to the potential of violating privacy , particularly in the context of health data. Collaborating with nonprofits and government agencies on how to use and analyze the data has assuaged some of those concerns. One participant also noted that companies may serve as a third party intermediary, helping to connect data scientists with government data.

Awareness of funding opportunities

Participants were asked how they learned about funding opportunities, with their responses varying depending on what sector the participants worked in. Those from academia reported they learned of opportunities mostly by email through offices of university provosts, department heads, and colleagues. In industry, funding opportunities are distributed at conferences, through discipline-specific listservs, or through competitive internal processes to receive funding. The idea of funding through a mechanism that allowed for multiple co-equal principal investigators appealed to participants since it would allow for a more balanced and equitable collaboration between data scientists and subject experts. In addition, participants from industry were interested in opportunities for funding that did not require affiliation with a university.

Participants with no prior connections to the NIH noted that they were not aware of how to find funding opportunities at the NIH, and were not familiar with The NIH Guide to Grants and Contracts. They discussed making communication to the data science community a priority, emphasizing that NIH is expanding its investments in biomedical data science research. They noted that the National Science Foundation (NSF) actively conducts outreach to universities and investigators, attends conferences to promote funding opportunities, and posts opportunities to a centralized website. While NIH conducts similar outreach, it may not be targeting venues popular among data scientists.

Facilitating Biomedical Data Science

Based on discussions and feedback from participants, we present the following ways that biomedical data science would be facilitated by attracting the expertise of data scientists to biomedical problems. Their expertise would be leveraged to better understand disease and health, while biomedical science and data would be leveraged to present interesting research and analytic challenges to data scientists. Ideally, a variety of stakeholders would be engaged in such facilitation, with specific potential actions identified for them respectively, as described below.

For repositories and data providers promoting data use and reuse

Provide raw data in a non-proprietary format. To facilitate ease of use, data providers could provide raw data in formats that do not require licensing or special software to process and use, such as comma-separated value (CSV) or plain text files.

Foster trust in a given dataset. Providers can also foster the data science community's trust in a given dataset by including details about how the data were collected, including adequate information to reassure data scientists that the data are complete, clean, and responsibly collected. Even if this ideal is difficult to attain for all datasets in a repository, having a few such datasets would be useful as samples for data scientists to explore (identifying them as especially high quality).

Include specific use case examples. Including examples of use cases can give data scientists a sense of what they can do with the data, particularly if they are unfamiliar with the given data type or scientific domain (the high quality sample datasets referred to above would be useful for this). Data providers and repositories may also consider developing or pointing to tools that assist with the use of that data to make it easier for data scientists—particularly those lacking subject matter expertise—to engage with the dataset and develop ideas for new hypotheses to test and models to build.

Conduct analyses on datasets using existing tools to establish benchmarks for comparison. Data scientists often want to test tools they have developed against a known gold standard to prove that their solution performs better or faster. Rather than conducting that analysis themselves, data scientists are likely to use datasets for which the benchmark has already been set. Along with describing specific use cases, documenting how a dataset has previously been used to demonstrate performance of a tool or method can attract data scientists who want to compare their tool.

For subject matter experts seeking collaborations with data scientists

Make the collaboration an interactive and communications-based process. Subject matter experts, particularly experimentalists, understand the caveats of data collection and the limitations of a particular dataset. Effectively communicating those limitations to their data science collaborators is essential to ensuring the appropriate analysis methodology is applied and the resulting conclusions are reasonable. It is useful for subject matter experts to be engaged with data scientists throughout the entire analysis process, working to understand why a particular methodology is appropriate and computationally rigorous.

Set expectations for success in advance. To avoid friction downstream of a collaboration, it is useful for subject matter experts and their data scientist counterparts to set expectations for success, whether in the form of a patent, publication, conference presentation, or some other output, at the start of the collaboration. Setting expectations up front can also facilitate generation of a timeline for the collaboration with a concrete end goal in sight.

Hold back some adequate data for model validation. To test data science models and conclusions, it is useful for subject matter experts to hold back some of their data from the analysis to use in those tests. Data scientists are accustomed to having holdout data they can apply to evaluate whether and how well a particular model works.

For trainers and educational institutions

Define and develop core skills required of data scientists. The definition of data science remains broad, even among those who consider themselves data scientists, so there is significant ambiguity about what is required of data scientists. The benefit of such ambiguity is that it allows data scientists to choose the training path that is right for them.

Ethics and responsible conduct of research must be incorporated in data science training. Data science is a quickly growing field and benefits from the generation of large streams of complex and diverse data that can be harnessed to answer a host of questions. To ensure that data are used for the public good rather than to discriminate or reinforce inequity, it is important that data scientists be trained early in their careers on ethics and the responsible

conduct of research, especially focusing concerns unique to data science. The details of such training may depend upon the training path taken.

Promote equity, diversity, and inclusion. There is a body of literature on the benefits of diversity for building more effective and efficient teams. To ensure data science has maximal impact—and is done in a manner that serves the greatest public good—equity, diversity, and inclusion are essential for recruiting and retaining the best talent in the data science workforce.

For funders

Facilitate awareness of biomedical data science and NIH interest in data science. Data scientists can work with data from any discipline and their research interests need not be driven by a particular biomedical context, such as a given disease or organ system. Currently, for data scientists to see what research opportunities (funding, collaboration, etc.) exist or what interests NIH has in data science, they must visit the websites of all NIH Institutes and Centers. This problem could be solved if NIH were to set up a web portal featuring links to data science-related items from across all of NIH, such as research highlights, intramural labs conducting data-intensive research, pertinent funding opportunity announcements, lectures, training opportunities, and a curated set of links to data science related extramural projects. This "Data Science @ NIH" landing page would also serve to highlight NIH's activities and interests in a central area for others to discover and explore.

Create and communicate opportunities for joint leadership of research and training. Funders should communicate the availability of funding opportunities that allow for data scientists and subject matter experts to serve as equal partners to lead research projects or training efforts. Joint leadership opportunities will send an important signal to data scientists that the funder recognizes that data scientists can make intellectual contributions to research or training, and that data science is not merely a technical service. Active encouragement of such opportunities will amplify that signal. While this kind of joint leadership is routinely offered in NIH funding mechanisms, it appears to not be widely known in the data science community.

Create multiple pathways for discovering funding opportunities. To encourage data scientists with little or no existing relationship to NIH to apply for funding opportunities, it would be useful for NIH to make these discoverable through a range of pathways, such as: (1) reach data scientists where they are by sharing opportunities through discipline-specific listservs, conferences, or societies; (2) consider utilizing librarians as a resource for disseminating information about funding opportunities; (3) create a "Data Science @ NIH" portal where data scientists can go to find information about funding opportunities and more.

Create opportunities for scientists from non-traditional backgrounds. Not all data scientists come from traditional academic backgrounds; they may not hold a doctoral degree or be affiliated with an academic institution. To attract more data scientists to work in a given discipline, funders may wish to consider how to best attract these non-traditional scientists.

Appendix A. Participants

Name	Position and Institution
Patti Brennan, Ph.D., RN	Director National Library of Medicine, NIH
Rumman Chowdhury, Ph.D.	Artificial Intelligence Lead Accenture
Lisa Federer, MLIS	Data Science and Open Science Librarian National Library of Medicine, NIH
Nathan Hodas, Ph.D.	Senior Research Scientist Pacific Northwest National Laboratory
Mike Huerta, Ph.D.	Coordinator Data & Open Science Initiatives Associate Director for Program Development National Library of Medicine, NIH
Michael Li, Ph.D.	Analytics & Data Science and CEO Data Incubator
Alexa McCray, Ph.D.	Professor of Medicine Harvard Medical School
Brandeis Marshall, Ph.D.	Chair of Computer & Information Sciences Associate Professor Spelman College
Lea Shanley, Ph.D.	Co-Executive Director South Big Data Innovation Hub
Jerry Sheehan	Deputy Director National Library of Medicine, NIH
Rachael Tatman, Ph.D.	Data Scientist Kaggle
Tracy Teal, Ph.D.	Executive Director Data Carpentry
Kristin Tolle, Ph.D.	Director of Data Science Initiatives Microsoft
Daisy Zhe Wang, Ph.D.	Associate Professor College of Engineering, University of Florida
Stephen J. Wright, Ph.D.	Professor of Computer Sciences University of Wisconsin-Madison

Appendix B. Workshop Agenda

9:00 – 9:30 Welcome & Overview

- Patti Brennan: NLM welcome and the NLM vision
- Alexa McCray: Workshop welcome and workshop purpose statement

9:30 – 10:30 Introductions and Participants' Data Science Research

10:30 – 10:45 Break

10:45 – 12:00 Discussion – Individual Themes

- Career and science
    - What was your career trajectory with regard to data science?
    - What key research questions/problems are you interested in addressing/solving?
- Choice of data domains
    - What drives your choices to work with the domains of data that you do?
    - Have you worked with different data domains over time?
    - What drove those changes?
    - Do you perceive any special (compared to other domains) barriers to working with biomedical data?
- Understanding data domains
    - What do you need to know about a particular domain before considering working with its data?
    - How do you gain subject matter understanding in the domains of data with which you work?

12:00 – 1:15 Working Lunch – NLM Staff on NIH Funding & Available Biomedical Data

1:15 – 2:30 Discussion – Institutional and Funding Themes

- Funders & funding
    - How diverse are your sources of funding?
    - What are advantages and disadvantages of having diverse funding sources?
    - How do particular funders encourage or discourage multidisciplinary collaborative research?
- Institutional
    - How useful is your institution in telling you about funding opportunities?
    - How does your institution encourage or discourage participation in multidisciplinary collaborative research?

2:30 – 2:45 Break

2:45 – 4:00 Discussion – Collaboration and Training Themes

- Collaboration from your point of view
    - What are your typical collaborative arrangements (e.g., roles played by you and others, within or across institutions, exchange of post-docs, etc.)?
    - What are major challenges to your collaborative research?
    - How do you define success in a collaborative project?
- Making the match
    - Are indicators of success for you typically the same as for domain collaborators? If or when not, how is the mismatch addressed?
    - What do you think is most misunderstood about data science or data scientists by domain subject matter experts with whom you collaborate? How is that clarified?
    - What do you want your collaborators to know about your expertise before you begin working with them?
- Training
    - Have you worked with collaborators in different disciplines to co-mentor students?
    - Are you aware of gaps in your institution's training of data scientists?

4:00 – 4:30 Highlights and Final Question

- Highlights from the day's discussions
- What might NLM and NIH specifically do to attract data scientists who are currently not working with biomedical data to do so?

Appendix C. Relevant Reports

Throughout the workshop, participants shared relevant reports and resources with the group, which have been organized and shared in the following section.

Relevant NIH Reports

- National Library of Medicine Strategic Plan
- NIH Strategic Plan for Data Science

Data Science Training

Reports

- Data Science for Undergraduates: Opportunities and Options, National Academy of Sciences, 2018
- Data Science Leadership Summit Report, co-funded by the National Science Foundation, the Gordon and Betty Moore Foundation, and the Alfred P. Sloan Foundation, March 2018
- The Digital Competence Framework for Citizens, European Commission, 2017

Methodologies

- CRISP-DM: Cross Industry Standard for Data

Approaches to training

- Reboot undergraduate courses for reproducibility, *Nature*, September 2018
- Hack weeks as a model for data science education and collaboration, *PNAS*, July 2018
- Leveraging analytics 1.0 for the analytics 2.0 revolution, O'Reilly, November 2016
- Let's Make Gender Diversity in Data Science a Priority Right from the Start, *PLOS Biology*, July 2015

Data science books

- Python Data Science Handbook, Jake VanderPlas
- R for Data Science, Hadley Wickham and Garrett Grolemund
- Seeing Theory: Visualizing Probability and Statistics, Brown University
- 2012 Coursera lectures: Neural Networks for Machine Learning by Geoffrey E. Hinton
- MOOCs and advice by fast.ai
- Ethics in Data Science
    - Doing Good Data Science, Mike Loukides, Hilary Mason, and DJ Patil, July 2018
    - Ethics and Data Science, DJ Patil, Hilary Mason, and Mike Loukides, July 2018

    - Technology and the Virtues, Shannon Vallor, September 2016

Open Science

- Open Science by Design, National Academy of Sciences, 2018
- Making Open Science a Reality, OECD, October 2015
- Increasing Access to the Results of Federally Funded Scientific Research, Office of Science & Technology Policy, 2013

Collaborations

Frameworks for successful collaborations

- Collaboration and Team Science: A Field Guide, National Cancer Institute, May 2018
- Open and Inclusive Collaboration in Science: A Framework, OECD, March 2018

Funding opportunities

- NIH Guide to Grants and Contracts
    - NIH RePORTER, for searching successfully funded grants
- Transdisciplinary Research in Principles of Data Science (TRIPODS): NSF funding program to bring together the statistics, mathematics, and theoretical computer science communities to develop the theoretical foundations of data science through integrated research and training activities
- Partnerships between Science and Engineering Fields and the NSF TRIPODS Institutes (TRIPODS + X): NSF program to fund cross-disciplinary collaborations involving machine learning. It's embryonic but a possible model for "changing the culture" in machine learning to encourage more cross-disciplinarity via well designed funding programs.

Datasets

Guidelines for enhancing discoverability of datasets

- Google Dataset Search, Datasheets for Datasets: recommendations for how to describe datasets
- NIH Data Commons Use Cases and Personas

Datasets to use

- The Data Incubator links to several data sources for Cool Data Science Projects
- Examples of Available Biomedical Datasets from NIH
- NIH Data Sharing Repositories
- Clinicaltrials.gov