Core Skills for Biomedical Data Scientists

Maryam Zaringhalam, PhD, AAAS Science & Technology Policy Fellow

Lisa Federer, MLIS, Data Science Training Coordinator

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

Executive Summary

This report provides recommendations for a minimal set of core skills for biomedical data scientists based on analysis that draws on opinions of data scientists, curricula for existing biomedical data science programs, and requirements for biomedical data science jobs. Suggested high-level core skills include:

1. **General biomedical subject matter knowledge**: biomedical data scientists should have a general working knowledge of the principles of biology, bioinformatics, and basic clinical science;
2. **Programming language expertise**: biomedical data scientists should be fluent in at least one programming language (typically R and/or Python);
3. **Predictive analytics, modeling, and machine learning**: while a range of statistical methods may be useful, predictive analytics, modeling, and machine learning emerged as especially important skills in biomedical data science;
4. **Team science and scientific communication**: "soft" skills, like the ability to work well on teams and communicate effectively in both verbal and written venues, may be as important as the more technical skills typically associated with data science.
5. **Responsible data stewardship**: a successful data scientist must be able to implement best practices for data management and stewardship, as well as conduct research in an ethical manner that maintains data security and privacy.

The report further details specific skills and expertise relevant to biomedical data scientists.

Motivation

Training a biomedical data science (BDS) workforce is a central theme in NLM's Strategic Plan for the coming decade. That commitment is echoed in the NIH-wide Big Data to Knowledge (BD2K) initiative, which invested $61 million between FY2014 and FY2017 in training programs for the development and use of biomedical big data science methods and tools. In line with

this commitment, a recent report to the NLM Director recommended working across NIH to identify and develop core skills required of a biomedical data scientist to consistency across the cohort of NIH-trained data scientists. This report provides a set of recommended core skills based on analysis of current BD2K-funded training programs, biomedical data science job ads, and practicing members of the current data science workforce.

Methodology

The Workforce Excellence team took a three-pronged approach to identifying core skills required of a biomedical data scientist (BDS), drawing from:

a) **Responses to a 2017 Kaggle[1] survey[2] of over 16,000 self-identified data scientists working across many industries.** Analysis of the Kaggle survey responses from the current data science workforce provided insights into the current generation of data scientists, including how they were trained and what programming and analysis skills they use.

b) **Data science skills taught in BD2K-funded training programs.** A qualitative content analysis was applied to the descriptions of required courses offered under the 12 BD2K-funded training programs. Each course was coded using qualitative data analysis software, with each skill that was present in the description counted once. The coding schema of data science-related skills was inductively developed and was organized into four major categories: (1) statistics and math skills; (2) computer science; (3) subject knowledge; (4) general skills, like communication and teamwork. The coding schema is detailed in Appendix A.

c) **Desired skills identified from data science-related job ads.** 59 job ads from government (8.5%), academia (42.4%), industry (33.9%), and the nonprofit sector (15.3%) were sampled from websites like Glassdoor, Linkedin, and Ziprecruiter. The content analysis methodology and coding schema utilized in analyzing the training programs were applied to the job descriptions. Because many job ads mentioned the same skill more than once, each occurrence of the skill was coded, therefore weighting important skills that were mentioned multiple times in a single ad.

Analysis of the above data provided insights into the current state of biomedical data science training, as well as a view into data science-related skills likely to be needed to prepare the BDS workforce to succeed in the future. Together, these analyses informed recommendations for core skills necessary for a competitive biomedical data scientist.

[1] Kaggle is an online community for data scientists, serving as a platform for collaboration, competition, and learning: http://kaggle.com

[2] In August 2017, Kaggle conducted an industry-wide survey to gain a clearer picture of the state of data science and machine learning. A standard set of questions were asked of all respondents, with more specific questions related to work for employed data scientists and questions related to learning for data scientists in training. Methodology and results: https://www.kaggle.com/kaggle/kaggle-survey-2017

Results

Skills Reported by Kaggle Respondents

This study analyzed the Kaggle survey respondents who indicated they worked in either an academic (n = 1,645, 9.8%) or medical-related (n = 252, 1.5%) data science position. The majority of Kaggle survey respondents in these fields (91.1% of academic and 87.7% of medical) began their data science training through university courses, structured online courses (i.e. MOOCs), or on their own (Figure 1). University courses played a larger role in data science training for academic data scientists, while online courses playing a larger role for medical data scientists. As biomedical data science becomes more established as a field, formal university training will likely play a greater role in training competitive candidates to enter the data science workforce. Consequently, university training programs must adequately prepare trainees with a core set of skills to meet the demands of the data science profession.

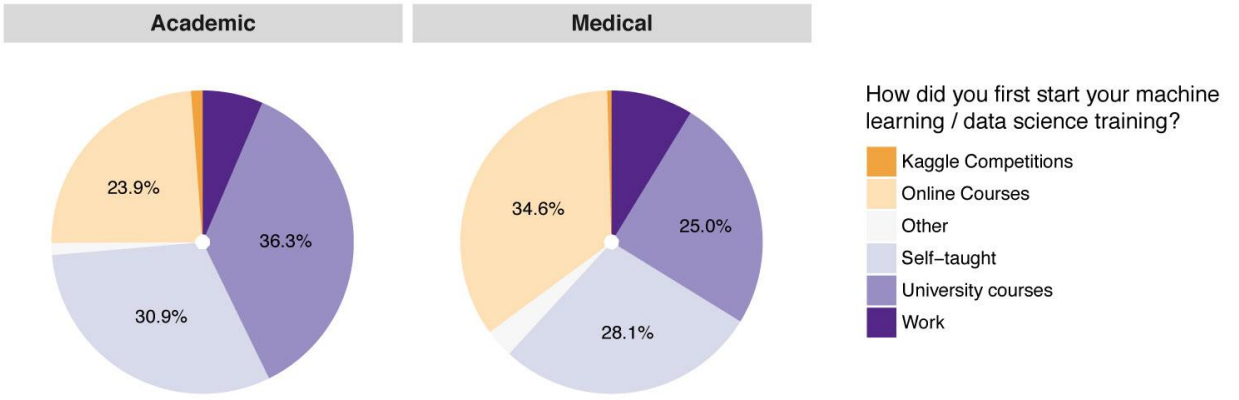

Figure 1. Kaggle survey responses for first experiences with data science training from 1,639 academic data scientists and 228 medical data scientists.

Among biomedical data scientists responding to the Kaggle survey, the skills considered important to the highest percentage of respondents were: (1) data visualization; (2) statistical methods including random forests, logistic regression, and cross-validation; and (3) programming skills including Python, R, and SQL. This survey did not ask respondents about the importance of specific subject matter expertise, likely because the audience for the survey was data scientists in general, not just *biomedical* data scientists. Future research focusing specifically on biomedical data scientists could help elucidate the role of subject matter expertise to their daily work.

Skills Covered in BD2K-Funded Training Programs

Qualitative content analysis of required courses revealed an emphasis on computer science and subject knowledge skills in the twelve BD2K-funded training programs (Figure 2). The exact breakdown of skills varied across programs, with some emphasizing computer science skills while others had a significant focus on subject matter expertise. However, all of the programs provided holistic training that incorporated elements of statistics and math, computer science, and subject matter expertise, as well as more practical general skills, like team science and written communication. Figure 2 shows the proportions of data science-related topics covered in BD2K-funded training programs. Appendix A provides additional information about the percent of programs that teach individual skills.

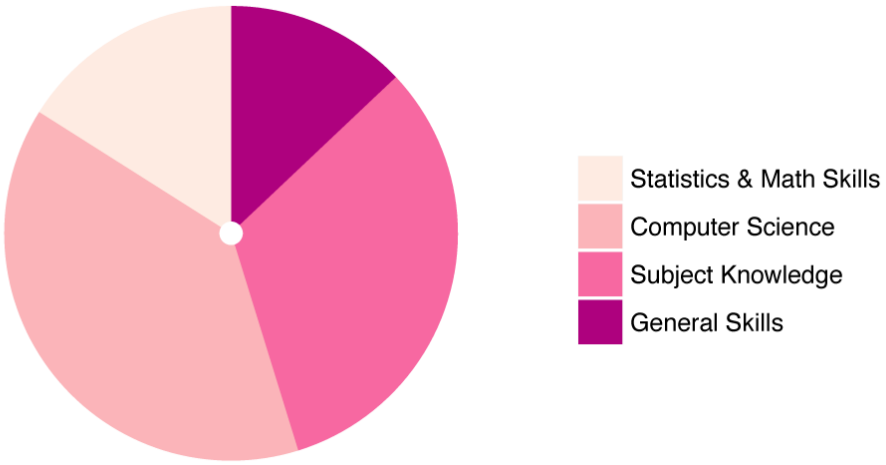

Figure 2. Proportions of data science-related topics covered by BD2K-funded training programs.

The above analysis was restricted to descriptions of required courses, so skills acquired through electives or thesis research were not included. For instance, thesis research builds general skills like collaboration and communication as well as subject matter expertise required for pursuing a research project.

Skills Included in Data Science Job Ads

Job ads typically required a variety of skills drawing on all three areas of expertise (statistics, computer science, and subject matter knowledge), with particular emphasis on computer science skills. Jobs in academia differed from those in other industries in placing additional emphasis on skills like teaching, mentoring, grant writing, and publishing. Across all industries, applicants were widely expected to be able to work well in teams and communicate effectively both verbally and in writing. The emphasis on general skills seen across industries suggests that training and experience in various data science skills is necessary but not sufficient for success; the most desirable candidates for data science jobs will also have skills and expertise that make them effective members of teams and enable them to communicate their findings.

Figure 3 shows proportions of skills required for jobs across various industries, and Appendix A provides additional information about the percent of jobs ads that include individual skills.

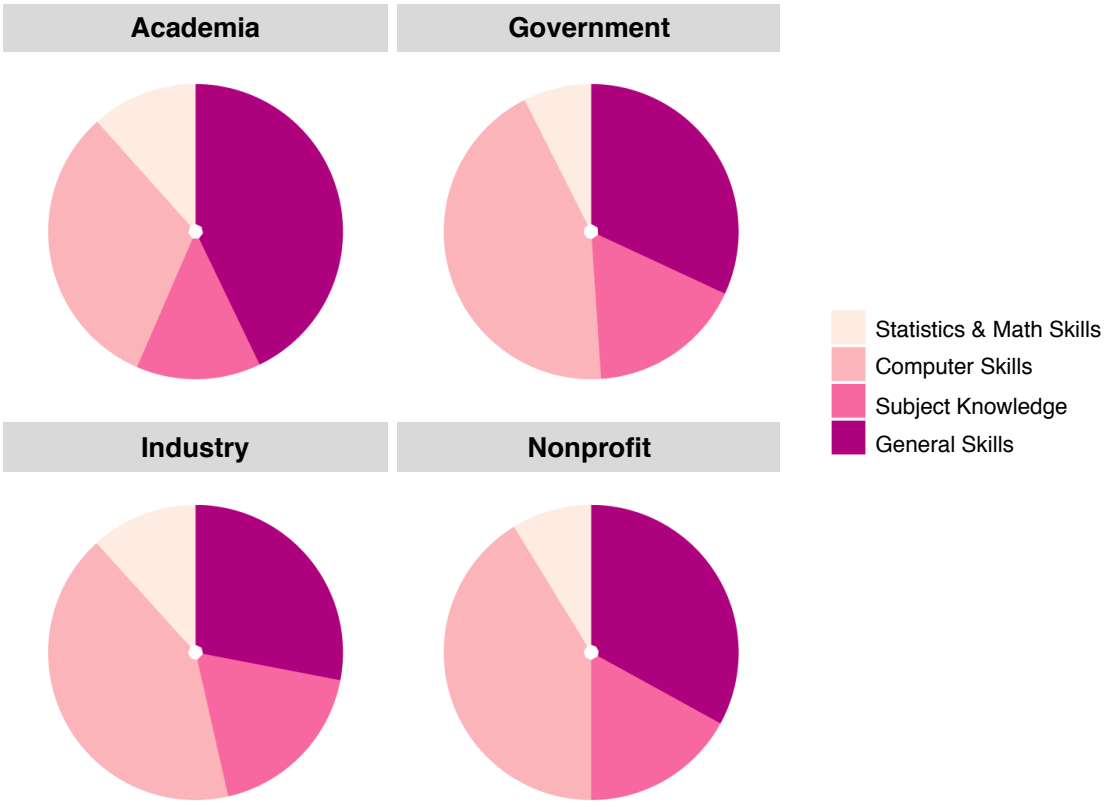

Figure 3. Proportions of data science-related topics required or preferred in survey of 59 BDS job ads.

Recommendations for Core Skills

1. Subject Matter Knowledge

The findings of this study suggest that being a biomedical data scientist requires at least a basic understanding and level of comfort with general biomedical subject matter knowledge. Given that many biomedical data scientists will work as a member of a team, the depth of subject matter knowledge that a specialist scientist in the field would bring is likely not required. Instead, **biomedical data science trainees may find it more useful to acquire a general working knowledge of the principles of biology, bioinformatics, and basic clinical science**, rather than an in-depth understanding of a specific science like immunology or neuroscience. The curricula of many of the programs considered here generally take this approach, giving students the option to specialize with electives and focusing required courses on more general topics that are likely to be widely applicable across a range of specific biomedical data science applications.

2. Programming Language Expertise

The findings of this study suggest that **a biomedical data scientist should be fluent in at least one programming language**. Which programming language is most useful for a biomedical data scientist is less clear; while the job ad analysis indicated a slight preference for R (49.2%) followed closely by Python (45.8%), biomedical respondents to the Kaggle survey indicated a slight preference for Python (27%) over R (25%). R is designed with statisticians in mind and is well-suited to data visualization applications. Python on the other hand is a general-purpose language that is well-suited to developing applications. Nevertheless, a deep understanding of how one language works can equip trainees with the logic structures and conceptual vocabulary underlying that language. Trainees can then apply that basic understanding to learning another language, if necessary. Most of the BD2K-funded programs explicitly mention programming language skills in their curricula, while others imply training in languages geared towards specific applications, such as bioinformatics or machine learning.

3. Predictive Analytics, Modeling, and Machine Learning

While the exact statistical methods applicable to any given data science problem may vary, **predictive analytics and modeling emerged as dominant statistical methods**. Machine learning approaches, and the software and tools that support them, are also widely relevant. These types of statistical techniques have many potential applications in biomedical data science and are likely relevant regardless of the particular disciplinary or topic focus of a specific data science problem.

Although most job ads do not specifically mention them, statistical methods that are foundational to predictive analytics, modeling, and machine learning are also important topics for preparing biomedical data scientists. For example, though few job ads mention regression analysis or probability, understanding these topics is a prerequisite for effectively applying machine learning or developing predictive models.

4. Team Science and Scientific Communication

Definitions of data science often emphasize expertise in computer science, statistics, and subject matter knowledge, but the findings of this study suggest that **"soft" skills, like the ability to work well on teams and communicate effectively in both verbal and written venues, may be as important as the more technical skills typically associated with data science**. For more than 90% of the job ads included in this analysis, the successful applicant will be expected to work in a team setting. For data scientists interested in working in an academic setting, experience with leading and working in teams, presenting research findings, and publishing in peer-reviewed articles are especially important.

Training programs in biomedical data science should ensure that trainees have ample opportunities to collaborate in cross-disciplinary settings in order to prepare them for jobs that

will likely involve a significant amount of team science. Biomedical data science trainees could also benefit from training in scientific communication skills as well as real-world opportunities to gain experience with presenting and writing about their work.

5. Responsible data stewardship

Fundamental to data science, is, of course, data. **A successful data scientist must be able to implement best practices for data management and stewardship**, and may be expected to provide leadership in such practices at his or her institution. Responsible data stewardship also underpins scientific reproducibility and enables data sharing, both areas of recent attention from the NIH and other stakeholders in the biomedical research community. Especially in contexts that involve human subjects' data, data scientists must maintain a commitment to ensuring data privacy and upholding high ethical standards. Though responsible conduct of research is a required component for all NIH training, some of the ethical issues and best practices confronting data scientists are unique, so training programs should consider how to prepare the biomedical data science workforce to meet emerging challenges in conducting ethical research and being good stewards of data.

Appendix A - Skills Coding Schema

Program percent: the percent of university programs that include one or more required courses covering the topic.

Coding	Percent of Programs with Coding	Percent of Jobs with Coding
Education level	n/a	n/a
Bachelor's degree required	n/a	23.7%
Master's degree required	n/a	49.2%
MD required	n/a	11.9%
PhD required	n/a	50.8%
Master's degree preferred	n/a	13.6%
PhD preferred	n/a	15.3%
Statistics and math skills (general)	58.3%	57.6%
Experimental design	50%	30.5%
Multivariate analysis	8.3%	5.1%
Predictive analytics and modelling	41.7%	45.8%
Probability	25%	1.7%
Regression analysis	33.3%	8.5%
Sampling	n/a	1.7%
Time series analysis	25%	3.4%
Computer science (general)	58.3%	37.3%
Algorithms	66.7%	27.1%
Big data analytics	50%	42.4%
Cloud based computing	16.7%	15.3%
Computing systems and architecture	8.3%	10.2%
Data mining	41.7%	28.8%

Data security and privacy	16.7%	6.8%
Data structures and database design	50%	42.4%
Machine learning	41.7%	52.5%
Scalable and parallel computing	25%	22.0%
Software development	25%	32.2%
Visualization	16.7%	45.8%
Text analytics (general)	8.3%	8.5%
Computational linguistics	n/a	n/a
Natural language processing	16.7%	10.2%
Programming (general)	58.3%	32.2%
C or C++	n/a	23.7%
Java	8.3%	25.0%
Javascript	n/a	5.1%
Other programming language	n/a	13.6%
Perl	n/a	15.3%
Python	16.7%	45.8%
R	16.7%	49.2%
SQL	n/a	20.3%
Software and operating systems (general)	8.3%	15.3%
Linux or unix	16.7%	16.9%
MATLAB	8.3%	8.5%
Other specific software	n/a	10.2%
SAS	8.3%	22.0%
Stata	8.3%	13.6%
Subject Matter Knowledge (general)	58.3%	42.4%
Bioinformatics	66.7%	55.9%
Clinical science	50%	18.6%
Epidemiology	16.7%	11.9%

Health informatics	33.3%	20.3%
Healthcare systems	n/a	10.2%
Imaging	8.3%	11.9%
Immunology	8.3%	11.9%
Neuroscience	8.3%	3.4%
Omics or genetics	50%	39.0%
Ontologies and standards	41.7%	10.2%
Precision medicine or personalized medicine	n/a	6.8%
Veterinary or animal science	n/a	n/a
General skills or knowledge (general)	n/a	n/a
Data management	33.3%	55.9%
Grant writing and funding	25%	25.4%
Leadership or supervisory skills	n/a	47.5%
Presentation and verbal comm skills	41.7%	74.6%
Publication record	n/a	15.3%
Reproducibility	16.7%	11.9%
Research experience	25%	33.9%
Responsible conduct of research and ethics	58.3%	8.5%
Teaching	8.3%	23.7%
Team science and team work	25%	93.2%
Written communication skills	33.3%	74.6%