



National Institute of
Environmental Health Sciences



Workshop on Developing a Data Science Competent EHS Workforce

August 14-15, 2018

Table of Contents

Workshop Summary -----1
Agenda -----7
Participant List-----10

Workshop Summary

National Institute of Environmental Health Sciences Workshop on Developing a Data Science Competent Environmental Health Sciences Workforce

August 14 – 15, 2018
NIEHS Building 101, Rodbell Auditorium
Research Triangle Park, NC

Description

On August 14 and 15, 2018, the National Institute of Environmental Health Sciences (NIEHS) convened an interdisciplinary workshop at the NIEHS Main Campus in Research Triangle Park, NC, to explore strategies to develop a data science competent Environmental Health Sciences (EHS) workforce. The workshop brought together experts from relevant research disciplines to examine existing data science and EHS resources (including trainee pipelines, mentors, and research) and identify how these resources can address EHS-specific training goals in data science. Throughout the workshop, participants discussed the challenges and opportunities involved in equipping the EHS workforce with the data science tools, understanding, and expertise to solve data-intensive environmental health research questions. Participants also provided recommendations for next steps to advance data science training in the EHS domain.

Background

With rapidly developing technology and more efficient data collection procedures, environmental health scientists are now collecting vast amounts of data. These data sets, termed “big data”, can be large, complex, multidimensional, diverse, and are often generated using new technologies. They are associated with basic, translational, clinical, social, behavioral, environmental, or informatics research questions. Such data types may include imaging, phenotypic, genotypic, molecular, clinical, behavioral, environmental, and many other types of biological and biomedical data. Data science has emerged from its roots in applied statistics, analytics, and

What is Big Data? Biomedical Big Data is more than just very large data or a large number of data sources. Big Data refers to the complexity, challenges, and new opportunities presented by the combined analysis of data. In biomedical research, these data sources include the diverse, complex, disorganized, massive, and multimodal data being generated by researchers, hospitals, and mobile devices around the world (<https://commonfund.nih.gov/bd2k>).

What is Data Science? Data science is an interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data. (<https://datascience.nih.gov>).

bioinformatics as a new area of research to meet the challenges in sharing, accessing, analyzing, and interpreting big data.

The National Institutes of Health (NIH) made early efforts to address the gap between the needed and existing biomedical data science skills through investments in training and education as part of the [Big Data to Knowledge \(BD2K\) Initiative](#). The programs and Funding Opportunity Announcements released had two main, and somewhat separable, goals: 1) improving big data skills of biomedical scientists; and 2) increasing the number biomedical data scientists. These NIH-wide efforts were not domain-specific and were intended to develop resources which could benefit all NIH institutes.

Workshop Discussion

The charge for the workshop was to develop an overall strategy to build a data science competent EHS workforce. The workshop agenda was organized into three major sessions. The first session examined the current state of data science in the EHS domain as it relates to training. Current limitations for data science training in EHS were identified through the evaluation of representative scientific ‘use cases’ nominated by the NIEHS Division of Extramural Research and Training (DERT) program branches. The second session focused on relating the accomplishments of BD2K to EHS training goals and examined existing big data training resources relevant to the intersection of EHS and data science. The final session examined existing NIEHS training resources relevant to data science. The concluding discussion focused on formulating how to build EHS training in data science and identifying next steps to increase training emphasis in big data.

Highlighted Challenges

Throughout the workshop, participants were asked to identify current gaps and barriers to advancing data science research and training. These include, but are not limited to:

Challenges associated with integration of diverse environmental health data streams. The EHS field spans multiple scientific disciplines, including, but not limited to, toxicology, epidemiology, basic and mechanistic science, environmental science, and clinical science. Accordingly, data variety is a major challenge for the EHS domain. Furthermore, the lack of standardization for many data types in the EHS domain represents a major barrier for data integration. A comprehensive understanding of data harmonization is needed, including development and knowledge of common EHS language standards and data structures as well as awareness of unique issues related to analysis of environmental exposure data, including spatial and temporal elements of exposures, gene-environment interactions, and causal inference. The field needs innovation in how we combine different EHS data streams to bring together data resources in new ways. This requires thinking about data integration early in the research planning process to consider data architecture, data quality, and standards. In addition to field-specific challenges, the general challenges associated with big data, including preprocessing (normalization) and hot spot detection, reproducible research, network methods, and data visualization and presentation, are indeed applicable to the EHS field.

Cross-domain communication barriers. Effective communication is key to the success of transdisciplinary collaborations. However, overcoming barriers to communication between data scientists and environmental health scientists remains a major challenge. Workshop participants noted a long acclimation period for developing language understanding and effective information exchange between the data scientist and the environmental health domain scientist. There are currently very few “translators” who can facilitate interpersonal communication at this intersection, and opportunities to develop transdisciplinary communication skills are needed. Institutional silos around research and training, including the practices of department-based degree programs versus free-standing (multi-department) degree programs can be barriers to supporting transdisciplinary training.

Challenges around perceptions and credit as well as recruitment and retention of data scientists. Workshop participants noted challenges associated with perceptions and credit associated with the work of data scientists in the current research environment. Data scientists and bioinformaticians can often be perceived as resources rather than independent researchers and generally work in a flat hierarchy system. Substantial behind-the-scenes work is involved in creating and maintaining data resources; however, this work is minimally rewarded. In the current paradigm, the order of authorship for scientific publication matters, with anchor/lead authorship being key to career advancement. The person who generated the data is usually the anchor author, and this is a challenge for bioinformaticians. Ultimately, creating culture change around these issues is important to moving the field forward. We need to change from the traditional model where a scientist is working alone to creating an environment where scientists communicate and work collaboratively. Furthermore, participants noted challenges in recruiting trainees from quantitative disciplines. Growing market demand and higher salaries in other fields contribute to this issue. Stipends vary across disciplines, and biology or health-focused trainee stipends tend to be less than stipends for quantitative programs. Moreover, data science skillsets are valuable across sectors, and keeping these trainees in the environmental health field is a challenge.

Workshop Recommendations

Participants suggested several actions to help overcome current challenges and enhance development of a research workforce poised to capitalize on advances in data science. These include, but are not limited to:

Improve access to EHS data to drive innovation and training opportunities. Revolutions in science are driven by access to data, and there is a need for FAIR¹ (Findable, Accessible, Interoperable, and Reusable) EHS datasets and tools to support data science training. Access to rich, interconnected, publicly-available EHS datasets is crucial. This may require new and creative ways to link datasets and increasing awareness and accessibility of existing toxicology databases. As the next generation of data scientists desire meaning, purpose, and

¹ Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3:160018. doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18) PMID: [26978244](https://pubmed.ncbi.nlm.nih.gov/26978244/)

understanding, we should leverage opportunities to get students involved in addressing socially-conscious EHS questions that are tied to impacts on public health. To do this, the field needs to develop exciting open-access case studies for data science relevant to EHS and connected to health outcomes. Additionally, fostering and development of repositories for computational methods and pipelines is important. We not only need to build these resources, but we need to train students and the community to maintain and maximize data repositories and resources. Currently, students are not extensively trained in data management.

Support creation or adaptation of short courses, online training, and/or core curricula at the intersection of EHS and data science. Participants noted a shortage of EHS-focused data science training opportunities and mentors. While many data science training courses are available online, they generally are not focused on the unique aspects of environmental health data (e.g. biomarkers of exposure, spatial measures), and intensive short courses in targeted EHS topics would be beneficial. Importantly, these training opportunities should involve hands-on experience working with real-world, messy data sets. In this effort, the field should embrace online-based FAIR curriculums to recruit, support, and retain individuals. Further, innovative training venues that go beyond traditional training settings, including data challenges and code-a-thons, could also prove beneficial. However, participants stressed that there is a challenge of scale, where thousands of individuals need training, and the limiting factor is the availability of qualified instructors and mentors. Potential solutions include train-the-trainer models (e.g. Data Carpentry) or recurring short-courses with multiple offerings per year. A further recommendation noted by participants was to consider development of model core curriculums for environmental health data science at a couple of institutions that could be exported more broadly to other institutions. This process begins with a focused effort to enumerate the essential skills that students need and then prune that list further to the skills that are truly essential. Next, those essential skills should be matched to existing courses, which can be adapted accordingly. Domain-specific opportunities should be built into the process. Workshop participants recommended implementing project-based learning where students work together as teams to undertake design-related projects alongside coursework, a practice currently implemented in engineering curriculums. In this discussion several important considerations were mentioned, including the need to have clear goals on who we are trying to train, the need to consider what trainees learn on their own versus what is needed to be included in formal training, and that while cross-training in biomedical and data science is important, trainees must have sufficient depth in one or the other (e.g. 80/20 split). Further, it is important to realize that if something is added (e.g. data science) to an environmental health training program, something has to come out.

Foster partnerships between health researchers and quantitative scientists. Communication, teamwork, and partnerships are critical for promoting data science workforce development in EHS. Environmental health work is inherently interdisciplinary, and training for data science in EHS requires a team approach (e.g. team science). Participants noted that synergistic academic research is key to transdisciplinary training. As we need to tackle both research problems and training collaboratively, cross-field awareness is key. Scientists need to develop awareness and appreciation of skillsets across domains, including expectation management – what can be

done and what can't be done. This goes both ways between biomedical domain scientists and data scientists; individuals working at this intersection must be conversant in both. Further, interpersonal communication between data scientists and EHS domain scientists is highly important; we need "translators" who are conversant across disciplines. NIEHS should foster partnerships between health researchers (e.g. at medical schools, schools of public health) with data/quantitative scientists (e.g. at engineering schools). These partnerships could be accelerated through venues including cross-disciplinary conferences and workshops (e.g. innovation labs) or through hands-on transdisciplinary research opportunities. When thinking about transdisciplinary research, teams should be brought together at the beginning of a project to design the research, and this idea can be applied to training.

Support data science training across career stages and knowledge levels. Paralleling the diversity of science in the EHS field, participants noted a need to support a heterogeneity of training opportunities at the intersection of data science and EHS. Current training and workforce development needs span all levels of the data science knowledge hierarchy, as goals and content for training vary according to expertise and organizational role. Levels and content in this hierarchy include (1) basic data literacy for nearly everyone, (2) basic training in numeracy for data-driven decision makers (manager-level individuals in data-driven organizations), and (3) specialized training for expert data scientists capable of designing and implementing complex analytics (researchers and analytic teams at complex organizations). Likewise, training is needed not just for trainees and junior scientist but across the spectrum of career stages. Participants highlighted data science training needs for several specific groups and purposes including, but not limited to: re-tooling for existing trained EHS scientists to make better partners for data science experts, training mid-career scientists with backgrounds in other disciplines, developing of mentors and instructors, promoting early mentored career development (with protected time), creating masters-level health data scientists, as well as enhancing education of trainees spanning high school, undergraduate, graduate, and postdoctoral levels. Regarding developing data science career paths for environmental health, workshop participants noted that there is no singular niche for EHS data scientists. It is not realistic to form one person with all EHS domain expertise and all data science expertise. As evidenced by the experiences of career development awardees at the workshop, the learning path to becoming a data scientist is rarely linear. Personalized learning paths should be customized for individuals and should accommodate different strategies for different needs. For enhancing data science training and career development in EHS, workshop participants recommended cross-disciplinary mentorship (dual or multiple mentorship with complementary skill sets) across computer science and informatics, statistics and mathematics, and biomedical science. Importantly, these training efforts developed should encourage participants from diverse backgrounds.

Outreach to a broader spectrum of stakeholders within and outside environmental health. Workshop participants noted a need for outreach to a broader spectrum of stakeholders within and outside of environmental health. Suggested venues included engaging with professional societies in quantitative fields such as the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). Additionally, participants recommended

engaging with industry. An example cited was learning from the spatial and geographic data and tools utilized by the oil and gas industry. Furthermore, participants encouraged seeking feedback from current EHS trainees and recommended outreach activities such as focus group meetings for trainees at scientific conferences such as the Society of Toxicology (SOT) Annual Meeting or NIEHS grantee meetings.

Moving Forward

While we are just at the beginning of the process of developing a data science competent workforce for EHS, the field has a good foundation to build upon to advance data science training. Programs, including existing quantitative and EHS programs, have started bridging the gap between biomedical science and data/computer science. NIEHS can adapt and extend existing models and strategies developed through BD2K and other quantitative programs. However, the field of data science moves fast, and as the field evolves, training goals are a moving target. Approaches need to deal with the changing needs of the workforce and adapt to new technologies. Trainees, as well as more seasoned investigators, are eager to gain data science skills and partnerships, and we need to provide opportunities. Moving forward, individuals and teams with data science expertise are in a unique position to make an impact on the EHS field.



Agenda



Workshop on Developing a Data Science Competent EHS Workforce

August 14-15, 2018

NIEHS Building 101, Rodbell Auditorium, Research Triangle Park, N.C.

AGENDA

Tuesday, August 14

8:30 – 8:45 a.m. **Opening Remarks and Purpose of the Workshop**
Carol Shreffler, Ph.D., NIEHS
Linda Birnbaum, Ph.D., Director, NIEHS and NTP

SESSION ONE: Understanding the Current State of Data Science in the EHS Domain As It Relates to Training

Goal: Through the evaluation of representative use cases, identify current limitations and rate-limiting steps for data science training in EHS. Presentations should include a high-level description of the persons or things involved, a sequence of actions and process flow, key competencies, and key obstacles. Discussion will entail a deep dive into use cases to identify relevant issues and challenges involved in solving data-intensive EHS research questions.

Moderator: *Charles Schmitt, Ph.D., NIEHS Office of Data Science*

8:45 – 8:50 a.m. **Introduction to Session One**
Charles Schmitt, Ph.D., NIEHS Office of Data Science

8:50 – 9:20 a.m. **Use Case 1: Tools, Skills, and Training for Quantitative Trait Locus Mapping**
Daniel Gatti, Ph.D., The Jackson Laboratory

9:20 – 9:50 a.m. **Use Case 2: Targeted Learning for Exposome Science**
Alan Hubbard, Ph.D., University of California, Berkeley

9:50 – 10:05 a.m. **Break**

10:05 – 10:35 a.m. **Use Case 3: Mapping What Matters: The Promise of Geospatial Health Informatics**
Marie Lynn Miranda, Ph.D., Rice University

10:35 – 11:05 a.m. **Use Case 4: Maximizing the Impact of Exposure Data in Children's Environmental Health Research**
Susan Teitelbaum, Ph.D., Icahn School of Medicine at Mount Sinai

11:05 a.m. – noon **Panel Discussion**

Noon – 1:00 p.m. **Lunch**

1:00 – 1:15 p.m. **Overview of Big Data Science Training**
Carol Shreffler, Ph.D., NIEHS

SESSION TWO: Relating BD2K Accomplishments to EHS Training Goals

Goal: Examine existing training resources relevant to the intersection of EHS and data science, including existing data science training resources (e.g., BD2K).

Moderator: *Amy Herring, Sc.D., Duke University*

- 1:15 – 1:20 p.m.** **Introduction to Session Two**
Amy Herring, Sc.D., Duke University
- 1:20 – 1:50 p.m.** **Applying the FAIR Principles to Online Data Science Training Resources**
John Van Horn, Ph.D., University of Southern California
- 1:50 – 2:20 p.m.** **BD2K Training at the Harvard T.H. Chan School of Public Health**
John Quackenbush, Ph.D., Harvard T.H. Chan School of Public Health and Dana-Farber Cancer Institute
- 2:20 – 2:35 p.m.** **Break**
- 2:35 – 3:05 p.m.** **The Physician Data Scientist: An Unexpected Journey**
Jonathan Chen, M.D., Ph.D., Stanford University
- 3:05 – 3:35 p.m.** **The Training of Next Generation Data Scientists**
Lana Garmire, Ph.D., University of Hawaii
- 3:35 – 4:05 p.m.** **NLM Training in Bioinformatics – Biomedical Data Science Training for Diverse Backgrounds and Career Paths**
Mark Craven, Ph.D., and Chris Bradfield, Ph.D., University of Wisconsin-Madison
- 4:05 – 5:00 p.m.** **Panel Discussion**
- 5:00 p.m.** **Adjourn Day One**

Wednesday, August 15

- 8:30 – 8:35 a.m.** **Day Two Opening Remarks: Chris Duncan, Ph.D., NIEHS**

SESSION THREE: Building EHS Training in Big Data Science

Goal: Examine existing NIEHS training resources (e.g., trainee pipelines) relevant to data science and discuss the next steps to increase training emphasis in big data. This may include both training for quantitative big data scientist career paths and needed big data skills for biomedical scientists in other environmental health programs. Identify the limitations and rate-limiting steps in data science training in EHS and make recommendations for priority areas.

Moderator: *Cheryl Walker, Ph.D., Baylor College of Medicine*

- 8:35 – 8:40 a.m.** **Introduction to Session Three**
Cheryl Walker, Ph.D., Baylor College of Medicine
- 8:40 – 9:10 a.m.** **Experiences With Environmental Bioinformatics Training**
Fred Wright, Ph.D., North Carolina State University
- 9:10 – 9:40 a.m.** **Data Science Training in Epidemiology**
Jim Gauderman, Ph.D., University of Southern California
- 9:40 – 9:55 a.m.** **Break**
- 9:55 – 10:25 a.m.** **Building Multidisciplinary Environmental Health Data Science Training Resources**
Chirag Patel, Ph.D., Harvard University
- 10:25 a.m. – 12:20 p.m.** **Panel Discussion on the Training Intersection of Data Science and Environmental Health Sciences**
Additional Panelists:
Wesley Gray, Ph.D., Southern University and A&M College
Ronald Hines, Ph.D., U.S. Environmental Protection Agency
Jeanette Stingone, Ph.D., Icahn School of Medicine at Mount Sinai
- 12:20 – 12:30 p.m.** **Meeting Wrap-Up and Closing Remarks**
Gwen Collman, Ph.D., Director, NIEHS/DERT
- 12:30 p.m.** **Adjourn Meeting**



Participant List

Workshop Organizing Committee

Danielle Carlin, NIEHS
Jennifer Collins, NIEHS
Chris Duncan, NIEHS
Michael Humble, NIEHS
Carol Shreffler, NIEHS

Participants

Janice Allen, NIEHS
Scott Auerbach, NIEHS
David Balshaw, NIEHS
Linda Birnbaum, NIEHS
Abee Boyles, NIEHS
Christopher Bradfield, University of Wisconsin-Madison
Pierre Bushel, NIEHS
Danielle Carlin, NIEHS
Trisha Castranio, NIEHS
Jonathan Chen, Stanford Department of Medicine
Jennifer Collins, NIEHS
Gwen Collman, NIEHS
Mark Craven, University of Wisconsin
Christie Drew, NIEHS
Chris Duncan, NIEHS
June Dunnick, NIEHS
Steve Edwards, RTI International
Lisa Federer, NIH Library
Richard Finnell, Baylor College of Medicine
Suzanne France, NIEHS contractor: MDB, Inc.
Lana Garmire, University of Hawaii Cancer Center
Amanda Garton, NIEHS
Daniel Gatti, The Jackson Laboratory
Jim Gauderman, University of Southern California
Wesley Gray, Southern University Baton Rouge
Alison Harrill, NIEHS
Astrid Haugen, NIEHS
Michelle Heacock, NIEHS
Amy Herring, Duke University
Ron Hines, U.S. Environmental Protection Agency
Jon Hollander, NIEHS
Stephanie Holmgren, NIEHS
Alan Hubbard, University of California, Berkeley
Michael Humble, NIEHS
Bonnie Joubert, NIEHS
Joyce Keith Hargrove, Environmental Health Consultant

Resham Kulkarni, NIEHS
Cindy Lawler, NIEHS
Kelly Lenox, NIEHS
Sarah Luginbuhl, NIEHS
Marie Lynn Miranda, Rice University
Maggie Moakley, NIEHS contractor: MDB, Inc.
Helen Pan, RTI International
Yong-Moon (Mark) Park, NIEHS
Chirag Patel, Harvard University
Kristi Pettibone, NIEHS
John Quackenbush, Dana Farber Cancer Institute, Harvard School of Public Health
Ravi Ravichandran, NIEHS
Trey Saddler, NIEHS
Charles Schmitt, NIEHS
Carol Shreffler, NIEHS
Lesley Skalla, NIEHS contractor: MDB, Inc.
Jeanette Stingone, Icahn School of Medicine at Mount Sinai
Susan Teitelbaum, Icahn School of Medicine at Mount Sinai
Kimberly Thigpen Tart, NIEHS
Brittany Trottier, NIEHS
Steven Tuyishime, NIEHS
John Van Horn, University of Southern California
Kerri Voelker, NIEHS contractor: MDB, Inc.
Cheryl Walker, Baylor College of Medicine
Fred Wright, North Carolina State University
Karen Xu, University of North Carolina at Chapel Hill



Workshop on Developing a Data Science Competent EHS Workforce