Broadening the Pipeline: Best Practices for Training in Biomedical Data Science

Specific Aims

The complexity of biological problems that we can address, and the speed at which we solve them, is dependent upon our ability to transform biological data into knowledge. Proficiency in biomedical data science is critical to unlock the potential of big data in modern biology. To be effective, training must combine tried-and-true instructional strategies with biomedical data analysis. The technological advances that drive change in biomedical research also demand new approaches to training researchers, physicians, and educators. We propose a new framework for training that will extend the approaches of *Software Carpentry* and *Data Carpentry* (the *Carpentries*) to diverse training challenges within biomedical data science. There are a number of efforts underway to create training materials. However, the impact, effectiveness, and sustainability of these materials are greatly increased when we build communities—people enabled to care, collaborate, and bring their own competency—that can unify around common research and educational challenges.

Central Question: What are the best methods for training in biomedical data science to the diverse community of learners?

We will execute the project through three specific aims:

1. Host a strategic meeting to develop a consensus set of best practices for biomedical data science training, and develop a community-curated white paper.
2. Host a "*CommunityCon*" that actualizes white paper best practices by gathering special interest groups who will collaboratively develop training materials.
3. Assess and evaluate the impacts of the project (including support and assessment of sustainable training materials), disseminate project findings, and plan for future activities.

There is not a universal consensus on curriculum for biomedical data science, but there are a number of practical approaches that can be collected and communicated as a standard. We propose a two-conference series that 1) identifies best practices (process, technology, and pedagogy) for training in biomedical data science; 2) implements those practices in a number of data-intensive biomedical domains; and 3) builds and strengthens communities that can reinforce practices and sustain training materials. A strategy meeting at Cold Spring Harbor Laboratory (Aim 1, fall 2017) will assemble a group of national and international training experts to explore facets of the central question and develop a white paper of recommendations. A larger "*CommunityCon"* on biomedical data science at the University of Arizona (Aim 2, summer 2018) will implement white paper recommendations by gathering researchers, trainers, and educators to develop open and findable, accessible, interoperable, and reusable (FAIR) training materials that address training and educational gaps in biomedical data science. In the following year, we will provide additional mentoring and support to the organizers of workshops and courses that are catalyzed by the *CommunityCon* event.

Additionally, we propose using the *Carpentries* model to scale and develop communities. We will employ an approach that to our knowledge has not been used outside of events organized by the projects' investigators: repurposing hackathons for collaborative creation of training materials. Collaborative development of training materials does take place within individual projects (e.g. CyVerse, Galaxy, ELIXIR, etc.), but the proposed *CommunityCon* is designed to help educators make learning materials sustainable, open, and community-developed. We are committed to ensuring training is 1) effective at welcoming and including women and all groups underrepresented in biomedical science, and 2) responsive to learners by integrating metrics and assessment that define training impact. A dissemination and evaluation plan (Aim 3) ensures sharing of materials and project findings, enabling further community use and development.

SIGNIFICANCE

Scope of the problem and unmet needs Biological science has been transformed into a data-intensive discipline by technologies, such as high-throughput sequencing and automated imaging. This transformation is accompanied by a critical need for training that prepares researchers, physicians, students, and educators to derive insight from Big Data. "We are data rich, but knowledge poor," is a refrain often heard in the field. A recent survey of 704 principal investigators supported by the National Science Foundation's Directorate of Biological Sciences identified the computation and data science needs of biology researchers[1]. The three most common needs were for training—surpassing needs for resources such as high-performance computing and data storage. At the top of the list was training on integration of multiple data types, with 89% of investigators reporting that their institutions did not meet this need. These results mirror a survey by the European Molecular Biology Laboratory (EMBL), in which 60% of respondents reported a need for more training compared to 5% who needed more computing power[2].

Limitations to current training approaches There is no shortage of learning material for those in need of cross-training in biology, computing, and data analysis. However, the persistence of the "training gap" indicates the need for more effective training. Although online data science and computing courses reach many thousands each year, they often attract those with computational, rather than biomedical, backgrounds—thus limiting opportunities for cross training. Online materials (e.g. Massive Open Online Courses - MOOCs) help fill the gap but fall short[3], often because they: 1) vary widely in scope, rigor, and approach; 2) persist even when content is in error or obsolete; 3) make it difficult for learners to interact and ask questions, especially when offered asynchronously. Often, online courses are more effective at communicating facts than understanding. Traditional academic courses offer success stories, but effective training developed at one institution is often irreproducible at other, often less well-off, institutions. As a result, solutions and innovation in teaching methods are difficult to spread and scale. Biomedical data science lacks a large pool of educators who can organize around a set of practices and shared curricula. Online training and isolated academic courses can scale up the number of learners, but communities transcend isolated institutions or MOOCs, not only increasing the number of educators and learners, but also improving the quality of learning and sustainability of content.

Adopting and extending the Carpentry training model Two organizations—*Software Carpentry* and *Data Carpentry* (the *Carpentries*)—deliver effective and scalable training that can serve as a model for training in biomedical data science. The *Carpentries* are community-driven organizations dedicated to teaching best practices in computing and data analysis to researchers[4]. *Software Carpentry* promotes reproducibility by teaching automated workflows, version control, and software testing. *Data Carpentry* addresses data literacy and domain-specific understanding in biology. In contrast to training materials that stem from a single source or a single institution, the *Carpentries* solve the scalability problem by cultivating a volunteer pool of more than 800 certified *Carpentry* scientist-instructors. *Software Carpentry* has delivered over 500 workshops in more than 30 countries and has reached more than 16,000 learners. Surveys of more than 5,000 *Carpentry* learners indicate that our audience is more than 40% graduate students—a key target population for training. The *Carpentries* make training more effective by borrowing the technologies of software development (e.g. GitHub, an online collaboration tool) to create content through a crowdsourcing approach. As a result, lessons are openly reviewed and iteratively developed—capturing the consensus of community knowledge. Over 400 persons have contributed to lessons in the last three years. Instructors—a significant portion of whom are graduate students and postdoctoral fellows—are coached in the methods of evidence-based teaching, allowing them to combine science domain expertise with techniques for peer-training. We believe the *Carpentry* model is extensible to many areas of learning within biomedical data science. The recently formed *Library Carpentry*[5] project is an example of community members (in this case, library scientists) independently emulating this model to address their own challenges. The *Carpentries'* approaches will be the starting point for generalizing training that can be adopted and scaled by communities of biomedical data science educators.

INNOVATION

Project goal and specific aims Our goal is to promote an infrastructure that supports educators with the best resources and approaches to training in biomedical data science. Given the broadness of the topic, it is both impossible and undesirable to develop a single consensus curriculum. Instead, we will provide a framework that assembles the practices—technological, pedagogical, and process—needed for effective training. The project will be supported by a community infrastructure—groups of people who care about, maintain, and teach content. Nearly 20% of the NIH BD2K budget goes into training[6], and sustaining these and other training investments requires community effort. Training infrastructure is the catalyst that enables grassroots communities to coalesce around the data sets and educational challenges of biomedical science. This is the essence of community development with the *Carpentries*. Educators who are embedded in communities of practice are best positioned to target the immediate training needs of researchers, physicians, students, and educators who work with biomedical data. To capture the project's impact, we will formally assess and evaluate the project and disseminate results. We will execute the project through three specific aims:

1. Host a strategic meeting to develop a consensus set of best practices for biomedical data science training, and develop a community-curated white paper.
2. Host a "*CommunityCon*" that actualizes white paper best practices by gathering special interest groups who will collaboratively develop training materials.
3. Assess and evaluate the impacts of the project (including support and assessment of sustainable training materials), disseminate project findings, and plan for future activities.

Project values A set of easily communicated principles define our aims. Training and learning are very much dependent on interacting with and "growing" people. The following values inform our goal and shape the development of our products:

- **FAIR Principles:** Adopt and borrow open science approaches and technologies to produce training materials that are findable, accessible, interoperable, and reusable (FAIR)[7].
- **Community = Sustainability:** The utility of training is determined by learners. Training must serve the research and career needs of learners, including students in the 9–16 (high school through undergraduate) educational pipeline.
- **Inclusiveness:** Training should include learners of diverse backgrounds and abilities.

Conference Plan

Project staff and organization The project staff and their organizations are strongly vested in the biomedical and data science communities. *Software* and *Data Carpentry* have successfully built international communities of trainers and learners and are aligned with and are collaborators of BD2K[8]. Project PIs encompass depth and breadth in education, community building, and biomedicine. They are also affiliated with large projects (*CyVerse*, *Galaxy*) that advance the causes of life science computing. PIs Micklos, Williams, Teal, Antin, and Clements provide overall leadership—including strategy, budgetary decisions, and reporting; PI Williams will be responsible for day-to-day project activities. As described in the FOA, PIs will coordinate with the project scientist assigned to this project by BD2K to shape conference topics. Organizing committees for the strategy meeting and *CommunityCon* will bring additional expertise in organizing meetings, hackathons, and training—as well as a network of collaborators that will help achieve representation of community expertise and diversity. Consultant McClatchy will work with PI Clements on mentoring and follow-up with post-*CommunityCon* workshops. The remaining project staff includes an independent project evaluator and a multimedia designer for website development and broadcasting of conference activities. As a demonstration of our commitment to inclusiveness, project leadership and staff represent gender, racial, and ethnic diversity.

Aim 1: Host a strategic meeting to develop a consensus set of best practices for biomedical data science training and develop a community-curated white paper.

In fall of 2017 we will assemble a group of experts to generate recommendations that reflect state-of-the-art practices in biomedical data science training and community building. A 2.5-day meeting of eight sessions for 23 invited attendees will be held at the Cold Spring Harbor Laboratory Banbury Center, a "think-tank" retreat with a reputation for gathering scientists to develop high-impact policy and new fields of study. Each session will address a topic and will be chaired by a thought leader on the subject. Sessions will be limited to ~1.5 hours—60 minutes of talks and 30 minutes of discussion and synthesis moderated by the chair. Session chairs will organize and write a white paper, a discoverable set of recommendations that emanate from the sessions. A science journalist will record proceedings and provide post-meeting editorial support to the chairs. We suggest the following tentative topics and, where possible, have identified meeting attendees who will chair (also see letters of support).

Potential Topic Question	Chair/Writing Committee Member
What is *Biomedical* Data Science?	Mark Gerstein, Yale University
What are "Core Competencies" for Biological Data Science?	Mark Pauley, University of Nebraska, Omaha
What are the Barriers and Bottlenecks in Data Science Training?	Vicky Schneider, EMBL-ABR
The Technologies of Open Science and their Impact for Learning	Jaime Whitacre, Project Jupyter, UC Berkley
How has Distributed, Open Training Succeeded in Building Communities?	Tracy Teal, Data Carpentry
How Should Learning in Biomedical Data Science be Assessed?	Ross Nehm, Stony Brook University
How do we Open a Diverse STEM Pipeline?	Susan Singer, Carleton College
What is the Future of Biomedical Data Science?	Brian Chapman, University of Utah

Overall meeting features Banbury meetings are small gatherings with speaking roles for most attendees. There are never evening sessions, so participants have ample time for informal discussion. Importantly, efforts are made to invite attendees who are relevant to the topics, but who would be new or unknown to the organizers' circle of collaborators. Diversity and gender representation are also explicit concerns. Travel support and subsistence will be provided to all attendees who will be housed on the Banbury campus. Banbury Center staff will coordinate abstract submissions, travel arrangements, and any other administrative and logistical details.

Session formula and responsibilities Each chair/writing committee member will:

- *Assist in crafting a session:* The chair will assist the project leadership in recommending additional attendees (1-2 others, excluding themselves) to speak at that session—endeavoring to invite experts from whom participants will benefit hearing. Speakers recommended by the session chairs will be passed to the conference organizers, who will develop a balanced and diverse list of attendees (i.e. gender, ethnicity, and broad representation of biomedical domains).
- *Contribute to the white paper:* Each session chair will contribute to a section of the white paper relevant to their session. Each white paper topic will include: 1) a statement of the problem and its background; 2) a set of recommendations (i.e. processes, pedagogical approaches, technologies) and/or open questions to the community; and 3) an evaluation of the project values in the context of the recommendations.

Publication and editorial process We anticipate some topics may be combined or differently conceived as a result of the meeting, but each committee member will have a writing assignment. During the

meeting, notes will be compiled by the science journalist. After the meeting, the journalist will also provide limited assistance to each of the chairs and help the project leadership with the final edit of the document. The publication will then be posted on the project website, a pre-print server, and GitHub for community comment. We expect the white paper to be a living document, so we have also budgeted for development of a modest website and dissemination at appropriate conferences and events.

Aim 2: Host a *CommunityCon* that actualizes white paper best practices by gathering special interest groups who will collaboratively develop training materials.

A four-day *CommunityCon* for training in biomedical data science will be hosted by the University of Arizona in the summer of 2018. *CommunityCon* will be a platform for educators who can catalyze development of learning materials in critical areas of biomedical data science and broaden participation of diverse learners. The *CommunityCon* will build and enrich communities of special interest groups (SIGs)—educators who share specific audiences and/or challenges in biomedical data science training. This event will not only create opportunities for developing new curriculum and learning best practices for data science training, but will also create an atmosphere in which participants can share knowledge and learn from each other. Educators who are developing course-based research projects in bioinformatics for undergraduates represent a SIG organized around an audience. Researchers who are developing a training curriculum on analysis of fMRI imaging data are an example of a challenge-focused SIG. *CommunityCon* will be a working conference where attendees participate in hackathons devoted to the creating learning materials based on the principles of the white paper, building community networks, and building the skills needed to be better data science educators.

Conference site and organization The University of Arizona is a large Research I university with ample space, facilities, and logistical capabilities for hosting a cost-effective *CommunityCon* for the 100-150 attendees we anticipate (see facilities description). The organizing team will coordinate advertising and conference logistics—including a conference code of conduct, invitation/transportation for plenary speakers, audio-visual needs, and childcare provisions. Within the framework described below, most of the topics and activities will be community-selected and driven.

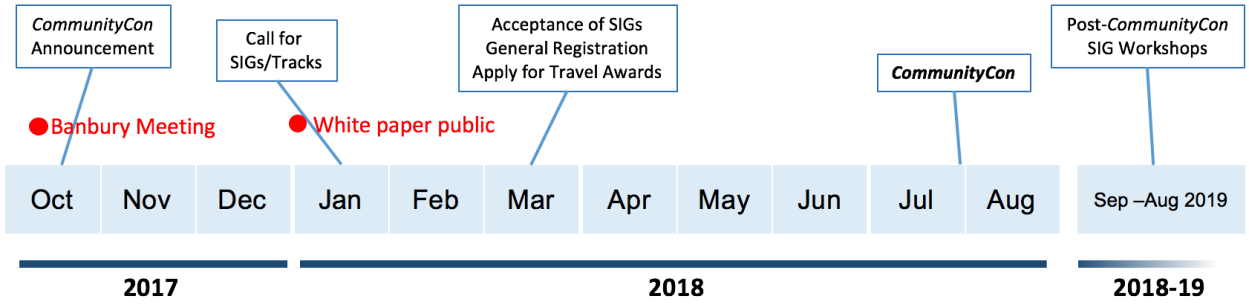
Pre-conference advertising and attendee application process In late 2017, an advertising campaign, including emails to lists from *Software/Data Carpentry*, *CyVerse*, *GOBLET*, NIH awardees, professional societies and mailing lists will announce the conference and deadlines. *CommunityCon* is a working meeting, so applicants must demonstrate ability to contribute and disseminate. Attendee contributions will include delivering a training session, organizing a special interest group, or contributing to a curriculum/lesson hackathon. Attendees will also need to describe how they intend to disseminate what they learn and produce at *CommunityCon* in a way that impacts biomedical research or students in the biomedical research pipeline. Applications will fall into two categories:

Special Interest Group (SIG) Organizers: These attendees (individuals, or as a group of up to three) apply by submitting a plan for a SIG that will hold either a hackathon or a skills session (see descriptions of these session below). The organizing team will shepherd SIG proposals to ensure fit with the conference objectives. To encourage submissions, registration costs are waived and one $500 stipend will be issued per SIG to defray their costs to organize SIG materials. Attendees may organize only one SIG. All applications will be scored by a rubric that assesses participant ability and interest in serving stakeholders in the biomedical data science community. As the number of applicants exceeds meeting capacity, the scoring rubric will help us ensure we serve the largest and most diverse participant audience. We also will broadcast plenary sessions/main talks through Livestream, post recordings to YouTube, and can provide virtual attendance options through Adobe Connect. Finally, all SIG applicants will be asked to participate in our assessment activities.

General Attendees: These are the majority of attendees who have some role in training and education. When applying, these attendees indicate their alignment or interest in SIGs and/or training sessions.

Qualified applicants will pay $100 registration fee collected through Eventbrite (an online event service) to defray conference costs and minimize no-shows. We will assist these attendees with recommendations for low-cost housing and other logistical and travel needs. We will offer a limited number of waivers and travel awards to encourage student participation and diversity.

Major project milestones timeline



Special Interest Groups (SIGs) and conference activities The working activities of the conference are organized by SIGs, which may form around, for example, high-school educators, researchers who analyze fMRI data, or archivists organizing health records at biomedical facilities. Proposals for SIGs will be posted on the project website, and community comments will influence the organizing committee's acceptance decisions. A key objective will be to select non-overlapping tracks that cover a range of biomedical topics and learner constituencies. The organizing committee may also propose SIGs if we do not receive sufficient submissions, or identify important topics that not addressed by any submission. Proposals by project staff will be transparently developed to avoid conflicts of interest and will welcome community input. After final decisions on SIG topics are made in spring 2018, general registration will open, and attendees may join one or more activities, according to their interests and schedule. There will be three major conference activities: talks, and SIG learning material hackathons and skills sessions. While we anticipate several likely topics and sessions, we will be most responsive to the needs of community members by allowing them to shape topics within these frameworks.

Talks: The conference will begin with a welcome talk that presents project aims and values, and summarizes white paper findings. Short talks on Day 2 will introduce the SIGs and the activities they have planned at the conference. Each day, a plenary session features topics centered on meeting the needs of training biomedical data scientists; participants from the Banbury meeting are potential speakers. The final day of the conference will feature lightning talks for summarizing each hackathon and skills session. A concluding talk summarizes the conference and suggests future directions.

SIG hackathon tracks: These 3.5 day sessions create or improve training materials on focused content areas, and follow the format of "learning materials" hackathons developed by *Data Carpentry*. Hackathon products will be deployed in workshops or developed into courses. During the challenge of defining a curriculum, identifying example problems, and crafting lessons, participants will follow (and refine) recommendations of the white paper. An example hackathon goal could be a set of lessons for an open-source software of high value to a community, but that is underutilized due to poor documentation. A hackathon on "metadata best practices" could introduce the value of a data commons to groups of users that would not otherwise adopt it. Like any hackathon, we expect the created products to be starting points for community adoption and refinement. We will provide mentoring and co-teaching support to up to five post-conference workshops for SIG hackathons that develop viable training materials. These post-*CommunityCon* workshops must be taught before the end of the grant term. Project leadership will review workshop content and agenda and mentor organizers through planning and implementation, including co-teaching support.

SIG skills sessions: Analogous to professional development, these sessions develop skills for building communities and delivering training to learners in the biomedical data science pipeline. Training on the design of a course-based research curriculum, adding assessment to a training program, using GitHub in the classroom, or how to teach R to first-year medical students are good examples of qualified topics for these sessions. Training may take the form of a "Technology Hour," a one-day session, or a 3.5-day track. Proposals for teaching a specific skill (e.g. basic R, Python, or ImageJ) for personal enrichment will not be accepted.

Aim 3: Assess and evaluate the impacts of the project (including support and assessment of sustainable training materials); disseminate project findings; and plan for future activities.

***CommunityCon* assessment and mentoring** A series of three attitudinal surveys will measure the effectiveness and impact of the event. Surveys will be administered pre-workshop, immediately post-workshop, and followed up approximately 9-12 months after *CommunityCon*. Post-surveys will collect feedback on the conference and follow-up surveys will document long-term impact and attendee dissemination efforts. In year two, the organizers of the post-*CommunityCon* workshops will receive extensive support from project staff, including virtual assistance to organize workshops and co-teaching support. Impact of these workshop materials will be gauged by attitudinal surveys.

White paper for practice To understand the impact of the white paper, a set of relevant metrics will be collected. These include pre-print downloads/views, publications, citations, community comments, GitHub issues, and sharing on social media. To the extent possible, we will seek community feedback on their efforts to make use of white paper recommendations.

Dissemination plan Through the project website, we will host a narrative about the project aims and link to a community-curated version of the white paper through GitHub. The website will link feedback from community members who make use of training materials to the white paper. We will also disseminate white paper findings and *CommunityCon* outcomes at appropriate conferences, BD2K events, and training registries (such as BD2K Training Coordination Center). Following FAIR principles, learning materials developed at *CommunityCon* will be publicly available through our website and GitHub repos (ideally through CCBY licenses).

Community impacts and future conferences We will solicit feedback from community members who make use of white paper recommendations or lesson materials initiated at *CommunityCon* on the website. At the end of Year 2, project leadership will have a final 1.5-day meeting at Cold Spring Harbor to discuss the assessment outcomes, white paper updates, and the potential for planning future conferences. This small meeting will be timed to search for co-sponsorships of a future *CommunityCon*, including application to the 2018 submission due date associated with this FOA. We will also support the attendance of five community members (SIG leaders and post-*CommunityCon* workshop organizers) who have been the most active in developing and implementing training activities.

References

1. CyVerse. Presentation of user needs assessment. Unpublished; 2016. http://de.cyverse.org/dl/d/F7772370-F02E-4E11-A2E3-DD2A3E590124/CyVerse_Advisory_Board_needs_slides_2016.pptx

2. EMBL. BRAEMBL community survey report - 2013.; 2013. http://braembl.org.au/news/braembl-community-survey-report-2013.

3. Gütl C, Rizzardini R, Chang V, Morales M. Attrition in MOOC: Lessons learned from drop-out students. Communications in Computer and Information Science. 2014 [accessed 2016 Dec 12]:37-48.

4. Teal T, Cranston K, Lapp H, White E, Wilson G, Ram K, Pawlik A. Data Carpentry: Workshops to increase data literacy for researchers. International Journal of Digital Curation. 2015;10(1):135-153.

5. Library Carpentry Project. Librarycarpentry.github.io. 2016 [accessed 2016 Dec 12]. https://librarycarpentry.github.io/about/

6. AIBS Council of Member Societies and Organizations. Addressing biological informatics workforce needs. Reston, VA; 2015. https://www.aibs.org/public-policy/resources/AIBS_2015_Council_Report.pdf

7. Wilkinson M, Dumontier M, Aalbersberg I, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos L, Bourne P et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 2016 [accessed 2016 Dec 15]: 1-9.

8. Dunn M. Inaugural activities of the NIH Data Science Training Center. Input | Output. 2015 [accessed 2016 Dec 12]. https://datascience.nih.gov/blog/inaugural-activities