Biomedical Data Science Curriculum Initiative:
February 2018 Workshop Report

Introduction & Context

The first workshop of the Biomedical Data Science Curriculum Initiative was held on February 7th and 8th, 2018 at the Department of Biomedical Informatics in Countway Library at Harvard Medical School. Five sessions were held across the two days. Each session had a moderator and up to three discussants who presented brief prepared remarks at the beginning of the session. There were no formal presentations, as the goal of the sessions was to promote conversation among all the working group members. Breakout groups held each afternoon provided opportunities to discuss some issues in greater depth.

Key Themes

Over the two days of the workshop, several key themes emerged:

- *How to balance what needs to be taught, while not being too prescriptive*. Data science is a fast-moving field. A "one size fits all" program isn't appropriate across institutions. Every program has its own unique goals and philosophies. In addition, it is important to embrace the diversity of students.

- *The value of case-based teaching*. Case studies are a useful way to teach skills through real-life problems. One institution is already spearheading an effort to create case studies. The group discussed whether it would be useful to have a panel or workshop at the AMIA meeting to generate additional momentum. Another alternative might be a special issue of JAMIA.

- *Some biomedical knowledge is essential for trainees, especially for those with only a computer science or other non life sciences background*. This foundational knowledge helps biomedical informaticians act as partners on a team, rather than as service providers. Without meaningful context, people run the risk of pursuing irrelevant research.

Day 1: February 7th, 2018

Opening Remarks

Drs. Alexa McCray and Nils Gehlenborg welcomed the group to Harvard and the Department of Biomedical Informatics.

Opening remarks were given by David E. Golan, MD, PhD who is the Dean for Basic Science and Graduate Education at Harvard Medical School, as well as by Valerie Florance, PhD, who is Director of Extramural Programs at the National Library of Medicine.

Dean Golan highlighted that informatics and data science are essential disciplines for medical education. He noted that at Harvard demand is high for data science programs. Dr. Florance discussed that the National Library of Medicine is issuing a new strategic plan with three pillars: NLM as a platform for biomedical discovery and data-powered health, effective dissemination of information across a broad range of users, and a workforce prepared to advance data-driven discovery and health.

Dr. Florance also highlighted several relevant activities ongoing at the NIH. The NIH will be issuing new funding announcements focused on workforce and methods development tools. The agency recently issued an RFI focused on next generation grand challenges. In the next month, the NIH strategic plan for data science will be published.

Session 1: Quantitative & Computational Methodology

Guiding Question: What are the essential quantitative and computational methods that students must master to become successful biomedical data scientists?

- *Moderator*: Lucila Ohno-Machado, University of California
- *Discussants*: Tianxi Cai, Harvard School of Public Health; Harry Hochheiser, University of Pittsburgh; and Peter Park, Harvard Medical School

Key Topics of Discussion

- *We must avoid programs that give trainees shallow knowledge across many areas, but no depth in any particular area*. People must learn to be critical thinkers and to understand why things work in certain ways. Trainees must know how to interpret analyses and understand biases that are in models. It’s important to have a healthy dose of skepticism about results – one approach is to review major academic papers that were later found to be wrong. Trainees need to understand how methods break down. Ideally, students will learn logical ways of thinking that will be transferable as programming languages change over time. Trainees must also know how to do reproducible research.

- *Graduates need to know about both biomedicine and computer science/data science.* Institutions should steep computer scientists in the culture of biomedicine (i.e., experimental design and fundamental results) and biomedical researchers in the culture of computer/data science (i.e., efficiency, appropriate data structures and algorithms, software engineering principles). When trainees with a computer science background take biomedical courses, they will be introduced to the methodology of biomedical research. Clinicians who go on to data science need to know how to structure a problem well enough to explain it to an engineer who will code it. Having some basic coding skills is important.

- *The diversity of student backgrounds is a major challenge for teaching*. Some trainees have extensive coding experience, but limited or no biology/medical background and vice versa. It's tough to teach topics like Bayesian statistics to individuals who have never taken calculus. Another challenge is phrasing questions in ways that everyone understands. Some students think courses are too basic, while others feel they are too complex. At one institution, courses teach algorithmic thinking, rather than a specific programming language. This has been an effective approach.

- *Skills required by PhD and Post Doc Trainees*. The participants outlined several competencies that biomedical data science PhDs and Post Doc trainees should possess:

    - *Programming*: Trainees should be able to write a few hundred lines of code and do some scripting
    - *Other computer science related skills*: scientific fit analysis, data cleaning, data integration, machine learning models, data structures, abstractions, Linux Shell, version control, metadata
    - *Statistics*: Bayesian, frequentist, SAS programming, regression analysis, exploratory factor analysis, clustering
    - *AI/Machine Learning*: How to compare tools, evaluate limitations. As machine learning becomes more mature, it will become important to understand how models are calibrated and what they were intended for initially.
    - *Study design and validation*
    - *Operations research*
    - *Infrastructure*
    - *Natural language processing/imaging*

- *The role of dual mentors.* They work well as long as the mentors understand one another's discipline. However dual mentors must be trained so that they are most effective, which can be a challenge.

Session 2: Quantitative & Computational Foundations

Guiding Question: What previous knowledge do students need in the areas of mathematics, statistics, and computer science to succeed in a biomedical data science graduate program?

- *Moderator*: Lydia Kavraki, Rice University
- *Discussants*: Shannon McWeeney, Oregon Health and Science University; Noemie Elhadad, Columbia University; and Peter Szolovits, MIT

Key Topics of Discussion

- *Programs and faculty must be flexible*. Institutions must consider learners, their backgrounds, and goals. There are many different sorts of trainees (e.g., Master's, PhD, Clinical Fellows, Post Docs). All have different learning goals and there are different expectations regarding their educational foundations. Pathways may need to be mapped based on the learners and what they hope to achieve. An open question is whether Master's and PhD students should be taught at the same time in the same classes.

- *The role of aptitude assessments*. The workshop participants offered several examples of different approaches:

    - At one institution, before the program begins, students take two assessments: one for algebra and probability, as well as one for programming and data structures. If they test out of those topics, they go to the core courses of the Master's program. If not, they do one semester of remedial courses.

    - Another institution offers a similar exam and gives recommendations about the areas where students need more education. It's their responsibility to fill the gaps. Since there's not a lot of accountability, it hasn't worked well.

    - Another institution offers a prerequisites exam the summer before the program starts. Students can get a waiver out of remedial courses.

    - At an institution with a PhD program, a preliminary exam is given at the end of the first year which includes computer science, biology, and statistics concepts. Students understand what they will be tested on and they can review exams from prior years. During the first year, students must figure out how to pass this exam. Most students know their own weaknesses. Many do online courses like Coursera or edX.

  The group also discussed challenges associated with aptitude assessments. For example, adding time for remedial coursework can be problematic, e.g., if a Master's program is only a year in duration. Also, finding bandwidth to teach students what they need to know can be problematic. At some institutions, demand for seats in foundational computer science/statistics courses in other departments is very high. It's almost impossible to get graduate students in. On the other hand, many program faculty don't want to teach the (remedial) foundational courses, preferring, instead, to teach the biomedical data science/informatics courses.

- *Trainee fundamentals*. Generally speaking, many thought this list should not be prescriptive since each program has unique goals and philosophies. With that said, the group outlined the following technical fundamentals and other recommended competencies.

    - *Technical fundamentals*.
        - Decision theory and basic statistics
        - Programming skills: Students must able to program well enough to produce the right abstraction that people can use and understand

        - The basics of software engineering: Software testing and the software lifecycle
        - Study design and how to conduct data analysis in a structured way
        - An understanding of medical problems
        - Data: How to obtain it (one of the hardest problems), data cleaning, data visualization, storage and processing of large amounts of data
        - Natural language processing ("the dark matter of medicine")

    - *Other fundamentals*
        - Common sense, communication, collaboration, and empathy: All of these are essential but hard to teach

Session 3: Data Skills

Guiding Question: What do students need to learn about data management, including data description and curation? What skills in identifying and mediating limitations of data (e.g., data quality, biases, incomplete data) do we need to teach?

- *Moderator*: Alexa McCray, Harvard Medical School
- *Discussants*: Steven Horng, Harvard Medical School; John Gennari, University of Washington; and Javed Mostafa, University of North Carolina

Key Topics of Discussion

- *All clinical data is biased*. It's important to know how data were collected, by whom, and for what purpose. Trainees must understand the importance of ontologies and their structures (e.g., ICD-9, SNOMED codes). They must also be cognizant that bias is inherent in all data and that data sets are usually noisy and incomplete. Data wrangling and curation are vital, but they aren't incentivized. It's important for trainees to understand how to structure data requests – for example, what types of requests are easy to service and what types are difficult or impossible to service.

- *Documenting code is essential*. One participant recounted how he was working on a project with the Alzheimer's Disease Neuroimaging Initiative (ADNI) Data Set which is well-curated and documented. The team needed a data selection that showed the patient's MRI scan at the time of diagnosis. To save time, the team wanted to leverage data selection work done by another group at the same institution. Many problems arose because the other team hadn't documented the data well. Different scan technologies were coded in different ways, but there was no documentation about how the coding was done/what it meant.

- *Necessary data skills*. The group outlined various data skills that biomedical data science trainees should possess.

    - Structural data standards, interchange data standards, how to create ontologies, data management, modeling, security, infrastructure, and how to build a platform.

  - On the execution side: scalable data infrastructure, automated workflow management, interfaces
  - Data integration: This is a big issue in health. Trainees must learn how to incrementally add to data sets, how to combine static and streaming data sets, and how to reconcile data from different sources.
  - Encryption, data de-identification, how to construct good data use agreements (DUAs)
  - Consent when collecting data
  - Data governance
  - Data acumen: What are good questions to ask of the data set?
  - Reproducibility: Data science can't be a science until we get this right. Programs need to teach and deploy better techniques.

- *Teaching data skills*. The group discussed different approaches for educating trainees about data skills.

  - *Capstone courses*. One institution has a capstone course where students identify a real problem and translate it to a computational data science issue. There is a strong data visualization component to the project. The capstone is a service learning project. It's valuable because students talk with people invested in the outcome of the analysis.

  - *Research project focused curriculum*. One institution with a PhD program doesn't follow a fixed curriculum. Students participate in small research projects that they pick. When they run into problems, that's when they learn. They fail in controlled ways and discover what they need to know. Hands on experience drives them to seek knowledge.

  - *Case studies*. One participant suggested teaching via case study. In a given case study, multiple people could comment on different approaches from a data science perspective and their strengths/weaknesses.

  - *"Lab notebooks."* The tradition of "lab notebooks" would be useful in informatics. How can institutions develop a culture that values that sort of documentation? Lab notebooks in data science might include data wrangling, data curation, data integration problems/procedures, and more. One institution has used R markdown documents successfully for this purpose.

- *The lack of available data sets is challenging*. Two sources are MIMIC and ADNI. The MIMIC project also has a GitHub repository with data cleaning code. In general, incentives aren't there to share citations or cyber repositories. Incentives must exist before sharing will occur. Incentives are also lacking for data cleaning.

Day 1 Breakout Group Reports

Group 1: *What are the critical competencies that must be taught in biomedical data science graduate programs?*

Attendees: Noemie Elhadad, Lucila Ohno-Machado, Shannon McWeeney, Robert McDougal, Wendy Chapman, Alexa McCray, Adam Wright, Lydia Kavraki, Meghan Dierks

This group generated the following chart of critical competencies:

Competencies	**Slider (0 to 5 - critical)** **The slider indicates the level or amount of training required in the topic**
Data	
Data security, privacy, protection and integrity (governance)	
Provenance, metadata, indexing, DOIs etc.	
Designing a database (data management)	
Fluency in use of databases, querying, data management	
Knowledge representation: Ontologies / Data Standards etc.	
Data wrangling (discovery, ingestion, transformation, etc.)	
Data integration	
Methods	
Use and understanding of unsupervised and supervised learning methodology (this includes statistical methods for classification)	
Development of new or significant enhancement of existing learning methods	Can go to zero in certain cases
Evaluation of Learning Models (performance, tuning, etc.)	
Statistical, mathematical and computational theory and techniques to measure uncertainty	
Visualization and Data Exploration	
Study design	
Evaluation and error analysis	
Understanding complexity - algorithmic thinking	

Matching methods to questions/problems - when not to use certain methods	
Extract meaning from unstructured / heterogeneous data (structure vs unstructured) e.g. text, audio, images, waveforms, ontological knowledge	0.5 in certain cases
Temporal methods and scaling methods	Can go to zero in certain cases
Software engineering best practices (code agnostic)	
Programming	
Interpretation and Dissemination (FAIR – data should be Findable, Accessible, Interoperable, and Re-usable)	
Model Deployment in context/real setting	
Interpretation of Results and actionability	
Communication of results (scientific writing, presentations/visualization)	
Sociotechnical issues	0.5 in certain cases
Organizational Behavior	0.5 in certain cases
Research / Scientific / Operational Skills	
Search (human searching for information) (literature, databases, data)	
Responsible conduct of research / Ethics	
Critical Thinking (assessing evidence, stopping criteria, next steps, limitations, understanding explicit and implicit assumptions)	
Resource allocation	0.5 in certain cases
Project management	
Teamwork/Team Science/Collaboration	
Domain knowledge (need at least one of)	
Biology / Genetics / omics	
Clinical / Medicine / Nursing / Pharmacy etc.	

Public Health / Consumer health	
Biomedical imaging and biomedical signals	

Group 2: How to teach problem solving and algorithmic thinking with a focus on biomedical data science? Style of teaching and learning: apprenticeship learning vs. classroom teaching?

Participants: Radhika Khetani, Cindy Gadd, Steve Horng, John Gennari, Brian Chapman, Ted Feldman, Alexander Diehl, Brian Dixon, Nils Gehlenborg

The group discussed the following topics:

- *PhD vs. Master's*. To include these students in the same classroom, they must have similar goals and be subjected to a similar admissions process.

- *Case studies*. The overarching goal is to teach skills through real-world problems. Cases should be structured in separate stages to reveal more about the complexity of the case over time. In addition, they should have an explicit learning objective or objectives. An open question is how many class sessions would it take to cover one case. Thought should be given to how case studies are organized and stored. The group consensus was that case based learning is effective and cases are useful for both MS and PhD students.

    - Examples of possible case study topics

        - Imaging informatics with a real MRI
        - Public health: lab reporting and public health requirements.
        - Analysis of co-location studies with fluoroscopy to determine statistical significance.
        - Air quality data from EPA, asthma reporting from clinics. Teaches data integration, data wrangling.
        - Predicting suicide attempts from medical record data.
        - Longitudinal case: could be used to illustrate the whole analysis and data wrangling pipeline and reproducibility

            - Requires an appropriate data set

- *Capstone courses*. These are typically used solely in Master's and undergraduate courses. Capstones demonstrate integrated knowledge. Unlike a thesis or dissertation, there is no expectation of novelty, though some students are able to accomplish work that is publishable.

- *Apprenticeship learning*. This applies to both PhD and Master's trainees. They may be research internships or industry, skill-based internships.

Group 3: Norms and Practices. Includes reproducibility, collaboration with other fields, communication, career development, entrepreneurship

Participants: Larry Hunter, Harry Hochheiser, Javed Mostafa, Olga Vitek, Michael Krauthammer

This group discussed the following topics:

- *Competencies relevant to the culture of science*. These include:
    - Scientific good practices. Doing significant work rigorously.
    - Be able to communicate clearly orally and in writing; Documenting work properly.
    - Being able to crisply define and scope projects
    - Managing a scientific project: planning, budgeting, monitoring, adapting.
    - Working with an inter- or trans-disciplinary team: can you explain your work to a statistician, computer scientist, a biologist and a clinician?
    - Being able to admit errors, being accountable, treating others' errors appropriately
    - Knowing when to give up on a line of inquiry, project, paper, and/or collaboration
    - Integrate the different aspects of one's training

- *Aspects of a good data science culture*. These include:
    - Being a good data citizen:
        - Documenting the work as we do it, metadata & process recording. What is an appropriate "scientific notebook" for a data scientist?
        - Ensuring openness, transparency, access
        - Consideration of licensing, protecting human data
        - Properly structuring complex data collections (file system is not a database; naming conventions are important; version control, change management, dependencies)
        - Appropriate use of metadata and vocabularies.
        - Understanding and ensuring an appropriate IT infrastructure (over time)
    - Being a good data/software user:
        - Acknowledging sources
        - Data quality orientation: sanity checks, manual examination; validation activities
        - Thinking about the suitability of a data set for addressing a scientific question
    - Knowing how to demonstrate one's competence to various audiences:
        - Contributions to open projects: GitHub commits
        - Certifications
        - Descriptions of projects at multiple levels
        - Social media/altmetrics that demonstrate your competencies (not just LinkedIn, but
        - StackOverflow, Biostars, etc.)

- *Lifelong development*. Trainees must know how to follow trends and keep up-to-date. They need to know how to gain new skills quickly.

- *Academic entrepreneurship*. It's important for data informaticians to know how to get support and resources for projects. This includes selling ideas and plans to principal investigators, collaborators, a study section, supervisors, and investors.

- *Ways to educate students*. Trainees must be educated in a variety of ways including in classes, during retreats, teach through example, start students in supervised situations and then enable them to work more independently over time, and seminar series.

Day 2: February 8th, 2018

Session 4: Biomedical Skills

Guiding Question: What is the critical knowledge about biology and medicine that biomedical data scientists must be familiar with? What do we need to teach them to allow them to develop meaningful questions to be answered with biomedical data?

- *Moderator*: Maha Farhat, Harvard Medical School
- *Discussants*: Michael Krauthammer, Yale University; and Chirag Patel, Harvard Medical School

Key Topics of Discussion

- *Students must be taught to ask the right biomedical questions*. At one institution, in the Introductory Course in Data Science, students develop small projects with data sets. They develop a hypothesis, explore how it is contextualized in the literature, and use data to explore the hypothesis. At another institution's Public Health Master's Program, students explore why data are collected and why the data are important.

- *Students need enough knowledge to work with a constrained data set*. One institution shows the commonalities between data standards and analytics for biological and medical data. Faculty walk through data elements and bring them into the clinical context.

- *Biomedical knowledge is required for students to feel like partners in the research*, rather than service providers. It's not so much about a body of knowledge as a feeling of belonging. One institution puts its PhD students through a basic molecular biology course. After that course, they have enough information to have opinions about experimental design.

- *Required clinical knowledge*. The group discussed several areas where data science trainees need clinical knowledge.

    - The infrastructure and history of U.S. clinical institutions – e.g., why billing codes are the way they are. One institution's qualifying exam includes questions related to clinical institutions and practices.
    - The information flow from a patient perspective. Students need a holistic view of the system, which is a challenge.
    - Where data come from and its imperfections. For example, EHR data isn't created for research purposes. Students can't assume that data will be correct and usable when it first comes through the door.
    - People need meaningful context, otherwise they risk pursuing irrelevant research. If trainees are analyzing healthcare systems, they must understand payment policies and what the purpose of the data collection is.

- *Teaching methods*. The group felt it would be helpful to develop foundational materials like readings, videos, and lectures on different topics as a group. One institution has a course taught by four faculty that exposes students to biology, chemistry, laboratory, and public health. It focuses on the nature of the data collected, stored, managed, used, and shared by these core biomedical disciplines.

  - Work is needed to determine what information is best delivered through a didactic approach versus experiential learning, such as shadowing a clinician in the ICU. Some participants were concerned, however, about the logistics challenges associated with shadowing (i.e., vaccinations, forms, etc.). In addition, shadowing is becoming more difficult. Trainees used to be "housed" in clinical areas in hospitals and clinics and this is no longer the case in many institutions. It's harder to give trainees exposure to clinical environments. One institution has a course on electronic health records that includes field trips to clinicians. There is value in embedding students in the environment where the data come from.
  - Participants thought it could be useful to have a case-driven entry level course that everyone takes regardless of their background. A topic map showing the relationship among topics would be great for students.

Session 5: Professional Skills

Guiding Question: What do biomedical data scientists need to know about ethical issues, such as responsible use and generation of biomedical data? What do we need to teach them about reproducibility? What skills are needed to make them successful communicators and collaborators?

- *Moderator*: Cynthia Gadd, Vanderbilt University
- *Discussants*: Larry Hunter, University of Colorado; Heather Mattie, Harvard School of Public Health; and Brian Dixon, Indiana University

Key Topics of Discussion

- *Required professional skills*. The group discussed a variety of skills that data science trainees should acquire:
  - Acting in an ethically responsible way. This includes being honest, acting with integrity, and exhibiting appropriate conduct.
  - Knowing how to recognize the limits of one's own knowledge and engaging in self-reflection.
  - Considering the broad significance of one's professional actions
  - Understanding issues like gender and race in society
    - Trainees should know how to respond if they are the victims of unethical behavior or witness it occurring to others.
    - They should also recognize the impact of these issues on science, such as the effect of genetic diversity on research.
  - Understanding reproducibility and the consequences when research isn't reproducible
  - Being an effective communicator. This includes how to communicate findings to peers, policy makers and the public.
  - Being an effective collaborator. This includes listening to collaborators and understanding what they want to achieve.

  - Producing accurate and effective visualizations of data. Trainees should be shown examples of good and bad data visualizations.
  - Negotiation skills
  - Using statistics in an ethical manner. Relevant topics include misleading statistical tests and analyzing big data in ways that hide faults. Faculty must encourage the discovery of errors. There's no shame in admitting and fixing an error. Lying to cover up errors is unacceptable and is far worse than committing the errors, which, in fact, can serve as a learning tool.

- *Teaching professional skills*. An overarching theme was that trainees need courses on professional skills and that faculty should weave professional skills throughout program activities. Professional skills must be addressed both inside and outside the classroom. The meeting attendees noted that ethics is an issue in all research fields. It would be productive to partner with others and draw on work that's already been done. For example, ACM has conducted workshops on ethics as it relates to AI and machine learning. The workshop participants discussed a variety of approaches to teaching professional skills to data science trainees. These included:

  - *Ethics notebooks*. One institution recommends that trainees maintain an ethics notebook to track issues that are of interest to them.
  - *Ethics courses*. One institution offers a course that identifies bright lines that shouldn't be violated, how to navigate authorship disputes, where research money comes from, deep questions/big social issues like AI and genetic engineering. Another institution uses a variety of approaches: annual HIPAA training, CITI training, a one-hour course for all trainees on reproducibility, seminars on data management, archiving, data management plans, and programs on how to teach effectively.
  - *Teaching technical approaches to solving ethical problems*. This topic tends to be very popular with students.
  - *Interaction with other health science disciplines*. One institution encourages inter-profession communication through workshops where health sciences students form roundtables and do a case study together.
  - *A professional "oath*." It would be helpful for data science trainees to take an oath that they promise to "do no harm with data." The National Academy of Sciences study, Envisioning the Data Science Discipline: The Undergraduate Perspective, offers a similar oath in Chapter 5.
  - *Case studies*. One institution gives all students a copy of the National Academy of Sciences "On Being a Scientist" which includes case studies. It shows that ethical issues aren't always black and white. It would be helpful to have case studies that were more specific to the biomedical data science field.

Day 2 Breakout Group Reports

Group 1: What are the biomedical skills that all curricula should include?

Participants: Alexa McCray, Michael Krauthammer, Harry Hochheiser, Ted Feldman, Alexander Diehl, Brian Chapman, Radhika Khetani, Maha Farhat, Heather Mattie

The group developed a list of biomedical skills that should be included in all biomedical data science curricula:

- Biology
    - Central dogma
    - Omics
        - The -omes (e.g. genome, proteome, transcriptome)
        - Biological study designs and measures associated with each ome
    - Critical analysis of biological literature
    - Molecular biology
    - Technologies
        - Sequencing
        - Microarrays
        - Mass spectrometry
    - Study design
    - Validation of results
- Medicine
    - Human anatomy and physiology
    - Pathophysiology / nature of disease
    - Interactions with the healthcare system
    - Pharmacology / how medicines are discovered and used
    - Process of care
        - Screening
        - Diagnosis
        - Treatment
    - Roles in the healthcare team
    - EHRs
    - Health systems
        - US
        - International
    - Role of the FDA and government oversight in healthcare
    - Health services
        - Quality measurement
        - Quality improvement
        - Patient safety
        - Cost / economics

    - Clinical informatics
        - Text
        - Codes/Structured data
        - Images
- Public health
    - Epidemiology
    - Study designs
    - Bias and confounding
- Translational topics
    - Pharmacogenomics
    - Genomic medicine
    - Precision medicine
    - Genetic testing
    - Sequencing of patients

Other topics of discussion included:

- *Determining topics and data types that will be relevant in the future*. The breakout group participants felt it was essential to focus on new developments and to build collaborations. Examining developments in basic science (e.g., in disciplines like biology, chemistry, physics labs, clinical areas, etc.) and asking basic scientists to give seminars were recommended approaches. Other suggested actions included interdisciplinary seminars and student presentations, tracking policies and regulations that might drive change, tracking student interests, and monitoring science related social media and Twitter feeds.

- *Appropriate didactic approaches*. To teach biomedical skills to trainees, the group discussed integration across courses, small classes, case studies, modular micro-courses, team learning with accountability, and critical reading of the literature (e.g., journal papers to illustrate concepts, books like Digital Doctor). Other methods that the breakout participants talked about included using data from public repositories and potential outcomes of interest (e.g., explain your "23 & Me" test results, or explain some other pertinent test results or potential artifacts in a result). For each topic, levels should be defined (i.e., basic, medium, advanced).

- *Next steps before May workshop*. The group recommended collecting existing teaching resources that are available across institutions, as well as generating a list of resources that still need to be created.

Group 2: What are the professional skills that all curricula should include?

Participants: Cynthia Gadd, Lucila Machado-Ohno, Brian Dixon, Larry Hunter, Teri Klein, John Gennari, Shannon McWeeney, Wendy Chapman, Javed Mostafa, Brad Coleman, Noemie Elhadad, Lydia Kavraki, Nils Gehlenborg, Robert McDougal

The group discussed:

- *Broad teachable ideas*. These included stewardship of data over time, power differentials, cultural issues, cultural sensitivities, harassment issues, and supervisory skills. Other professional skills discussed were how to debate and criticize with respect, listening well, how to collaborate and communicate with collaborators, how to respond to problems, research reproducibility, intellectual property-related issues, and funding sources (as well as what funders expect in return).

- *Special issues in professionalism for informaticians/data scientists*. This includes technical solutions to ethical problems, licensing (especially around software and data), statistical and visualization ethics, issues relevant to machine learning, data producer/data analyst relationships and authorship responsibilities.

- *Teaching approaches*. The group discussed incorporating professional skills into classes, orientation events, extracurricular events (e.g., seminars, workshops, lectures, events, journal clubs), and faculty led lunches. Learning from peers is another means of gaining professional skills.

- *Resources*. These included, Preparing Future Faculty Program, and the UCSF Mentoring Contract.

Final Plenary Discussion: Next Steps

Prior to the May workshop, WebEx calls will be held during which participants will continue to present information about their programs. As part of the final plenary session, the group discussed how to convey the information from the February and May workshops. Several options were raised, including a special issue with multiple contributors, video interviews with participants at the May event, and a white paper.

Recommended actions from the February workshop included:

- Creating an online pointer to a library of case studies
- Collating the case studies that are currently under development and those that already exist. Based on this information, gaps can be identified.
- Sharing what the applicant pools look like for different programs and the number of people who are admitted
- Gathering information on the gap between what students think they need to know and what program directors think students need to know

Looking ahead to the May workshop, which focuses on teaching and pedagogy, the group made several recommendations. One option could be to include instructional designers to suggest how to distill information to a curricular approach. There was interest in exploring topics such as:

- How to deliver online and in-person education most effectively
- How to address the challenges associated with the infrastructure and resources needed to teach
- How to evaluate how well faculty are teaching and whether a program is teaching the right things

The group asked to include breakout sessions again in May, since they were well-received at the February workshop.

Other Important Points

Over the course of the two-day February workshop, other noteworthy points were raised:

- *The nature of data science programs*. One participant questioned whether data science programs teach about data or about science. Projects are often minimally generalizable. An open question is how to build theory and advance the field.

- *Benefits of a diverse student body*. Although student diversity introduces complexity, often the best mentoring occurs within diverse student groups, rather than only between students and professors.

- *Clinical informatics board certification*. A tension exists between the clinical informatics board certification and modern data science curricula.

- *Biomedical vs. computer science learning*. One institution expects computer science students to perform as well as students with biology or medical backgrounds. Putting someone with a background in computer science in a molecular biology course isn't the same as putting someone with a life sciences background in a computer science course . Programming isn't only

> about knowledge, it's about practices. It's like playing the piano – you don't learn by reading about it, you have to do it.

Appendix 1: List of Links to Documents Discussed During the February Workshop

- Case Studies for Biomedical Data Science (see e.g. https://dbmi.hms.harvard.edu/sites/g/files/mcu491/f/doc/Brad%20Malin%2C%20Vanderbilt%2C%2012.13.17.pdf, slides 18 - 21)
- National Academy of Sciences Report – "Envisioning the Data Science Discipline: The Undergraduate Perspective"
    - Data Science Oath from "Envisioning the Data Science Discipline" report
- Reproducible Research edX course
- Readings on Fairness in Machine Learning
- Weapons of Math Destruction
- Predictive model best practices (EHRs)
- Article related to useful data from public repositories
- NIGMS Training Modules to Enhance Data Reproducibility
- 10 Simple Rules for Responsible Big Data Research
- NIH Office of Research Integrity
- Committee on Publication Ethics

Appendix 2: Workshop Attendees

Tianxi Cai, Brian Chapman, Wendy Chapman, Brad Coleman, Alexander Diehl, Meghan Dierks, Brian Dixon, Noemie Elhadad, Maha Farhat, Theodore Feldman, Valerie Florance, Cindy Gadd, Nils Gehlenborg, John Gennari, Harry Hochheiser, Steven Horng, Larry Hunter, Lydia Kavraki, Radhika Khetani, Teri Klein, Isaac Kohane, Michael Krauthammer, Heather Mattie, Alexa McCray, Robert McDougal, Shannon McWeeney, Javed Mostafa, Lucila Ohno-Machado, Peter Park, Chirag Patel, Peter Szolovits, Olga Vitek, Adam Wright

Rapporteur: Karen McHenry