REPORT OF

Workshop on Enhancing Training for Biomedical Big Data

Big Data to Knowledge (BD2K) Initiative

July 29-30, 2013

Big Data to Knowledge (BD2K) is a new NIH initiative that aims to enable scientists to effectively manage and utilize the large, complex data sets (Big Data[1]) that are already being generated and whose number and value will only increase in the future. The BD2K initiative is based on a set of recommendations presented on June 12, 2012 by the Data and Informatics Working Group (DIWG) to the Advisory Committee to the Director, NIH. The DIWG report can be found at *http://acd.od.nih.gov/diwg.htm*.

One of the DIWG recommendations was to "Build Capacity by Training the Workforce in the Relevant Quantitative Sciences such as Bioinformatics, Biomathematics, Biostatistics, and Clinical Informatics." The NIH organized the "Workshop on Enhancing Training for Biomedical Big Data" as one approach to obtaining input from the biomedical[2] data science community on priorities for training needs and activities. The Workshop was co-chaired by Karen Bandeen-Roche (Professor and Chair of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University) and Isaac Kohane (Professor of Pediatrics and Chair of the Bioinformatics Program, Boston Children's Hospital and Dana-Farber/Harvard Cancer Center).

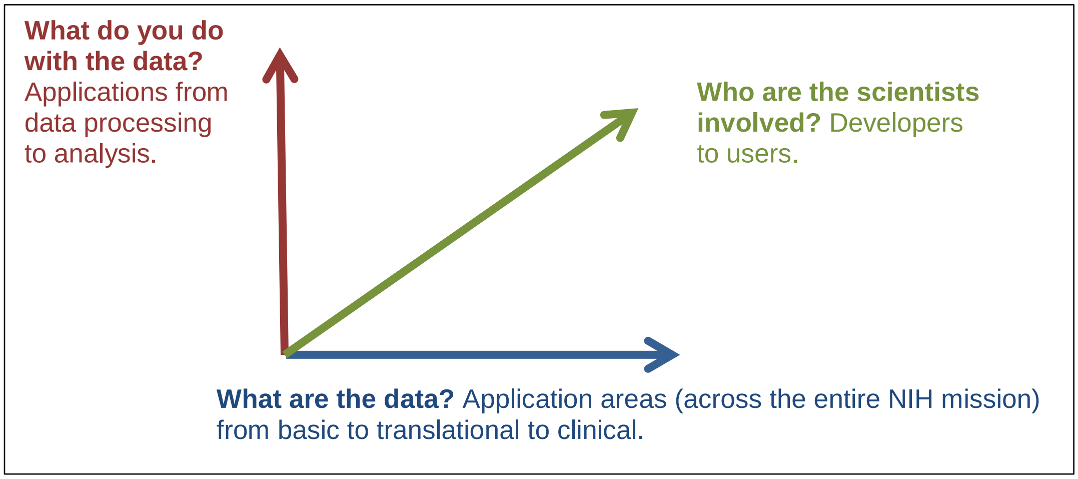
PURPOSE AND GOALS OF THE WORKSHOP

Michelle Dunn presented the purposes of the workshop as the following: (1) to identify the knowledge, skills, and resources that the biomedical research enterprise needs to organize, process, manage, analyze, and visualize large, complex data sets, and (2) to make recommendations on specific objectives for the NIH in the area of training for the utilization of Big Data and their priorities. The full rationale for the workshop, as distributed in advance to the participants, can be found in Appendix I, the agenda in Appendix II, the roster of participants in Appendix III, and a list of members of the NIH BD2K Training Working Group in Appendix IV. An archived videocast of the workshop can be found at *https://videocast.nih.gov/PastEvents.asp*.

[1] "Big Data" is meant to capture the opportunities and address the challenges facing all biomedical researchers in releasing, accessing, managing, analyzing, and integrating datasets of diverse data types.
Such data types may include imaging, phenotypic, molecular (including -omics), clinical, behavioral, environmental, and many other types of biological and biomedical data. They may also include data generated for other purposes. The datasets are increasingly larger and more complex, and they exceed the abilities of currently-used approaches to manage and analyze them. Biomedical Big Data primarily emanate from three sources: 1) a few groups that produce very large amounts of data, usually as part of projects specifically funded to produce important resources for the research community; 2) individual investigators who produce large datasets for their own projects, which might be broadly useful to the research community; and 3) an even greater number of investigators who each produce small datasets whose value can be amplified by aggregating or integrating them with other data.

[2] In this document, the term "biomedical" will be used in the broadest sense to include biological, biomedical, behavioral, social, environmental, and clinical studies that relate to understanding health and disease.

Dr. Dunn described the issues that need to be addressed in “Big Data” training in four “dimensions”: (1) data that span the NIH mission; (2) applications that span the pipeline from data acquisition and processing to data analysis; (3) scientists, from developers to users; and (4) career stage, from students to professionals.


What do you do with the data? Applications from data processing to analysis.
Who are the scientists involved? Developers to users.
What are the data? Application areas (across the entire NIH mission) from basic to translational to clinical.

She noted that the workshop participants embodied many disciplines and scientific areas (including informatics, computational biology, biostatistics, biology, genomics, mathematics, computer science, and education), but that not all relevant disciplines were able to be represented. Therefore, she asked the participants to think broadly, beyond their specific scientific expertise or disease of interest, and to focus in the workshop on the overall needs and priorities for BD2K training. Finally, Dr. Dunn stated that the outcome of the workshop deliberations would be used by NIH staff to develop training initiatives designed to prepare and empower the biomedical research workforce to take full advantage of Big Data for research into the understanding of human biology and improving human health.

BACKGROUND RELEVANT TO THE WORKSHOP DISCUSSION

In June 2012, three reports were presented to the NIH Director by working groups of the NIH Advisory Committee to the Director (ACD), all of which are relevant to training and careers in the area of Big Data. Dr. Sally Rockey presented the recommendations of, and NIH follow up to, two working groups: the ACD Working Group on the Biomedical Workforce and the ACD Working Group on Diversity in the Biomedical Research Workforce. A summary of Dr. Rockey’s presentation can be found in Appendix V. Dr. Mark Guyer presented the recommendations of and follow up to the ACD Working Group on Data and Informatics.

ACD Working Group on Data and Bioinformatics (DIWG)

The DIWG made five recommendations:

1. promote data sharing through central and federated catalogues;

2. support development, implementation, evaluation, maintenance, and dissemination of informatics methods and applications;
3. build capacity by training the workforce in the relevant quantitative sciences;
4. develop an NIH-wide "on-campus" IT strategic plan; and
5. provide a serious, substantial, and sustained funding commitment to enable recommendations 1-4.

The NIH's initial response to the DIWG report has three components:

- *Appointment of an Associate Director for Data Science (ADDS)* to lead the trans-NIH effort in data science, including the development of a long-term strategic plan. The ADDS will also be the primary NIH focus for coordination with data science activities beyond the NIH. Dr. Eric Green, Director, National Human Genome Research Institute, is currently the Acting Director, and a search is underway for a permanent ADDS.

- *Creating a Scientific Data Council* as a high-level internal NIH committee that, working with the ADDS, will provide oversight for trans-NIH data science activities. The Council will be chaired by the ADDS, and its members will be senior leaders from across the NIH.

- *Implementation of the Big Data to Knowledge (BD2K) Initiative (http://bd2k.nih.gov)* as the programmatic arm of the trans-NIH activities in biomedical Big Data. The overarching goal of the BD2K Initiative is to enable, by the end of this decade, a quantum leap in the ability of the biomedical research enterprise to maximize the value of the growing volume and complexity of biomedical data*.* There is wide-spread support for BD2K across NIH, with at least 24 NIH institutes/centers/offices participating in the initiative.

The DIWG report discussed a number of major problems in the use of Big Data:

- locating the data;
- getting access to the data;
- organizing, managing, and processing the data;
- developing new methods for analyzing data; and
- finding trained researchers who can utilize the data effectively.

The DIWG also noted that cultural changes at NIH are needed, including new approaches to data sharing and recognition that extracting the value of Big Data will require significant resources, i.e. data handling can no longer be considered to be free.

BD2K has identified four programmatic approaches to addressing the DIWG's recommendations:

I. Facilitating Broad Use of Biomedical Big Data;
II. Developing and Disseminating Analysis Methods/Software for Biomedical Big Data;
III. Enhancing Training for Biomedical Big Data; and
IV. Establishing Centers of Excellence for Biomedical Big Data.

Initial BD2K activities are focused on planning by means of workshops (of which this is the first) and obtaining community input through Requests for Information. The first BD2K Funding Opportunity Announcement, for BD2K Centers of Excellence, was just published (http://grants.nih.gov/grants/guide/rfa-files/RFA-HG-13-009.html). NIH has set aside $27M in

FY2014 for BD2K and has plans to scale up funding up to approximately $100M per year by FY2016.

Summary of Request for Information (RFI): Training Needs in Response to Big Data to Knowledge (BD2K) Initiative

In preparation for the BD2K training workshop, the NIH issued a Request for Information (http://grants.nih.gov/grants/guide/notice-files/NOT-HG-13-003.html, see Appendix VI) in which the community was invited to comment on 1) the skills and knowledge needed by a BD2K workforce, 2) the characteristics and content of plans for cross-training at all career levels, and 3) how to develop a diverse BD2K workforce. More than 100 responses were received. The NIH BD2K Training Working Group analyzed the responses, and the results were sent to the participants in advance of the workshop. Richard Baird presented a brief summary of the responses.

- Who to Train: The BD2K workforce will need both quantitative (statistical and computational) expertise and biomedical domain expertise, taken together as "data science" expertise. Examples of biomedical fields that already incorporate varying amounts and mixtures of quantitative expertise are bioinformatics, computational biology, biomedical informatics, biostatistics, and quantitative biology. Both basic and clinical researchers at all career levels need to receive training.

- When to Train: Training is needed at all career stages: exposure courses for undergraduates, cross-training for graduate students and postdoctoral fellows, training as needed for researchers at all levels to facilitate their work, refresher courses or certificates in specific competencies for mid-level researchers, and relevant continuing medical education courses for clinical professionals.

- What to Train: Both long- and short-term training is needed, and efforts should be guided by the competency level required for the technical knowledge and skills to be gained. The technical knowledge and skills needed include: (1) computational and informatics skills; (2) mathematics and statistics expertise; and (3) domain science knowledge.

- How to Train: Several ways to cross-train biomedical and quantitative scientists were suggested, including through (1) new or expansion of existing long-term research training programs (which can incorporate activities such as boot camps, joint and team coursework, delayed laboratory rotations, dual or team mentoring, clinical and industrial externships, and team challenges); (2) short-term courses and hands-on immersive experiments (which can span short courses, certificate programs, immersive workshops, summer institutes, clinical immersion and shadowing, and continuing medical education opportunities); (3) curricula for biomedical Big Data; (4) technology-enabled learning systems and environments (e.g., web-based courses and Massive Open Online Courses (MOOCs) to offer training to a much larger audience; and (5) a training laboratory that has tools and resources for self-directed learning and exploration.

FOUR EXAMPLES OF BIG DATA CHALLENGES AND COMPETENCIES NEEDED TO MAKE FULL USE OF THE DATA

Four types of Big Data research problems were presented as examples of the opportunities offered, the challenges presented, and the competencies needed.

Electronic Health Records (Daniel Masys): Dr. Masys described Big Data as having two main characteristics: it exceeds the capacity of unaided human cognition for its comprehension, and it strains current technology capacity and is therefore CPU-bound, bandwidth-limited, and/or storage-limited. Electronic Health Records (EHRs) may contain many data types, including quantitative clinical measurements; textual lab reports; narratives; images and signals used to construct images; DNA sequence (and increasingly in the foreseeable future, gene expression, proteome, and metabolome values); complex physiological signal data; as well as billing data, demographics and other coded name-value-pair data, consents and other legal instruments, and e-mail and other forms of patient-provider and provider-provider communications. As a collection of data, important characteristics of the EHR are time sensitivity, the inclusion of both objective and subjective information, an inherent structure understood by users but with the data not always recorded in a structured format, and its confidential nature. Also, the EHR data have primary and secondary uses. Dr. Masys used the Electronic Medical Records and Genomics (eMERGE) Network as an example to describe the challenges of extracting phenotypic data from EHR for use in genotype-phenotype studies. In his view, competencies needed for the use of EHR in this type of correlation research include expertise in human physiology and disease pathophysiology, molecular biology and molecular genetics, clinical documentation rules and business practices, data modeling and database design, natural language processing, use of controlled vocabularies and ontologies, and biostatistics, particularly of methods for association testing using noisy high dimensionality data. Finally, Dr. Masys noted that this type of research typically requires an interdisciplinary team of five to seven people in which each team member has knowledge that spans more than one area.

Imaging (Ron Kikinis): The amount of imaging data being generated is increasing from gigabytes to terabytes, is becoming more complex, and has more modalities and applications than ever, including both research and clinical. Challenges to using imaging data include the large number of subjects, the length of time (up to years) needed for analysis, and quality assurance. Logistics are challenging in several ways, including standardizing imaging equipment and protocols, getting the patient to the scanner, getting the data to the data center, and post-processing that requires an automated pipeline and large computational resources. As an example, Dr. Kikinis discussed a multi-center COPD genetic epidemiology study that has 21 clinical sites, three image analysis centers, two imaging platforms, involves two contrast mechanisms per visit and two visits per subject, has four processing pipelines, and has collected information on 10,000 subjects. The total analysis takes 320,000 CPU hours per run of the processing pipeline. In his view, competencies needed in imaging research include expertise in medical image computing, medical informatics, image acquisition, and domain science. Cross-training takes three to five years and is apprentice-style. An additional challenge to the use of imaging data is the need for robust, user-friendly tools, the development of which is hindered because tool creation applications generally do not fare well in the NIH peer review system. Dr. Kikinis pointed out that the medical imaging community is a compact one, making it a well-suited as a test bed.

Genomics (Michael Boehnke): The decreasing cost of sequencing a genome (now less than $10,000 per genome) allows many more genomes to be sequenced, generating much more data to process and analyze. Many current research projects are generating hundreds of terabytes of data. Challenges in using these data include processing raw sequence image files into useable data, aligning sequence reads to the human reference sequence, building error models to allow accurate variant calling, identifying and accounting for DNA sample contamination, imputing dense genotypes from a reference set of sequenced genomes to genomes with less dense genotype data, testing disease-genetic variant association in sequenced and imputed data, and combining data and results across studies. Other challenges include the large amounts of CPU time to analyze the data and memory to store the data; data storage is a major challenge because multiple copies of the data are required (since different software requires different versions of the data) and processed data sets may be as large as the raw data set. In Dr. Boehnke's opinion, dealing with genomics data requires knowledge in more than one scientific discipline, an aptitude to be actively engaged with the data in order to understand its context and identify problems, the ability to work in teams and communicate with experts in other disciplines, and creativity and flexibility to deal with a rapidly changing landscape. He also emphasized that producing well trained, cross-disciplinary scientists takes longer than training single-discipline scientists and that training needs may differ for developers, creative users, and general users of these methods.

Integrations of Large and Small Datasets (Mark Musen): Dr. Musen noted that there are a number of challenges to effective data integration. Many databases are not robust, attracting individuals to develop standards is difficult (standards development is not exciting but is essential to data integration), and obtaining support for the development of standards is difficult. He then provided several examples that have developed effective approaches in both biological and clinical arenas. In Dr. Musen's opinion, trainees need to understand the processes and frameworks needed for data integration. There is a spectrum of data integration, from the very 'heavy' integration tools needed for using data warehouses like i2b2 to 'light' serendipitous mashups that support discovery of associations. In the latter case, data are integrated on a 'just in time' basis, while in the former, data are integrated on a 'just in case' basis. To integrate data, however, people need to find them, use standard metadata descriptions, use standard ontologies to create value sets, and represent the data in frameworks at the right level of granularity (warehouse to mashup). The next generation of investigators will need to understand how to: model biomedical domains to create new ontologies and new metadata specifications, evaluate the appropriateness of an ontology for a given data-integration task, search for data sets using relevant ontologies, and apply semantic technology at different locations on the data-integration spectrum (from data warehousing to mashups).

SUMMARY OF THE WORKSHOP DELIBERATIONS AND RECOMMENDATIONS

The discussions ranged broadly over many issues relevant to data science, extracting knowledge from Big Data, and training; many specific ideas and suggestions were offered. There were, however, a number of themes that consistently ran through the different sessions and topics.

- The opportunity for extraction of knowledge from Big Data is often greatest at the intersection of at least two disciplines, and training programs should be designed to develop the ability to work at intersections.
- Multi-disciplinary approaches are critical to taking advantage of Big Data to advance biomedical science and knowledge. While some individuals with skills and expertise in several disciplines will be able to operate on their own as independent investigators, most of the relevant work will be done in well-integrated, multi-disciplinary teams.
- Training programs should be oriented toward providing trainees with the skills to work effectively in Team Science. This will often involve offering the opportunity to develop in-depth expertise in at least two scientific disciplines. To extract knowledge from data, it will be particularly useful if at least one of those disciplines is a quantitative discipline.
- Dual mentoring should be encouraged.

- There is no one right way to implement Big Data training, and it will be important for NIH to allow enough flexibility in its support for this type of training. Flexibility is needed to encourage innovative approaches and to allow training programs at different institutions to take best advantage of the particular talents and expertise available locally. The majority of the workshop attendees did not think it was a good idea to require that all NIH-supported training grants have a required data science component beyond the teaching of the principles of data science.
- The training experience can be enhanced if the trainees have access to large data sets, of multiple types, including -omics, imaging, and clinical data. NIH was encouraged to explore the idea of developing and providing such training sets as a resource.
- Training in quantitative science and experimental design will be increasingly important to clinical researchers and also clinicians. It would be desirable to add such training into medical school curricula, but that will not be easy. It might be easier to add such training to the pre-medical experience. It was also suggested that incorporating questions or problems that require knowledge of quantitative sciences as part of medical board exams would have a strong influence on training curricula.
- The principles of reproducible research should be stressed.
- There are training needs across the full spectrum of scientists, from technology- and tool-developers, to technology- and tool- users, to those who need to be conversant with the challenges and solutions related to big data.
- Realistic goals and limitations must be recognized for short-term training of non-experts, where training should equip the learner to understand enough about the quantitative analysis and tools available to collaborate with expert users or developers in data acquisition and analysis.
- The jobs necessary for Big Data science may not correspond to traditional scientific, particularly academic, jobs. Training individuals to participate across the full spectrum of scientific roles is desirable. In addition, an appropriate career path must be available when those individuals finish their training.
- A diverse workforce should be a major goal of data science training efforts.

In addition to these overarching themes, there were many interesting points made in the individual sessions of the workshop, although not all of these were consistent with one another.

Knowledge and Skills Needed

Workshop participants discussed the knowledge and skills that biomedical data science teams and trainees should have as well as strategies for fostering their development.

- It is important for trainees to develop both quantitative (computational and statistical) and domain expertise.
- Working in a scientific team requires learning how to participate in active collaborations, including being respectful of the contributions of those with complementary expertise, being committed to working together, and fostering open communication.
- Faculty trainers must be close collaborators, supportive of the team approach, and have an awareness of and appreciation for all the areas that make up the team.
- Data sharing is critical and is becoming the norm. The concepts of managing and sharing data should be introduced early and continue throughout the training experience.
- Online modules, patterned after the human subjects training modules, could be developed and made widely available to teach principles of data sharing.

- The need to develop multidisciplinary capabilities is likely to lengthen the training period, in contrast to current interest in shortening it.

Long-term Training and Career Award Programs

There are a number of interesting approaches that have worked well in particular training programs, including the following:

- Organizing trainees with complementary expertise to work as a team to solve a specific problem.
- Holding boot camps at the beginning of training programs to introduce trainees to disciplines outside of their experience, e.g. quantitative skills to those with a biology background or biological approaches to those with a quantitative background.
- Pairing advanced graduate students with early-stage PhD students to solve a difficult data analysis problem.
- Having graduate students pursue rotations later in the didactic part of their training so that they have a firmer grasp of the principles of data science as they experience different laboratories.

In terms of environment, it is important to have access to the infrastructure needed to manipulate and analyze large data sets. Partnerships between large institutions and smaller institutions, including community colleges, liberal arts colleges, and minority-serving institutions, can be an effective approach to improving access to training. They are also effective in increasing the flow of students from those institutions into graduate training programs.

Short-term Training Programs

Short-term training opportunities should be made available for both basic and clinical scientists at all career levels to provide ongoing training and career enrichment to both new and established investigators. Short-term training can be used to recruit new people into particular research fields or to allow people to bring new fields into their research programs.
Short-term training experiences can take many forms and serve different audiences and purposes, including the following:

- Workshops
    - To bring trainees from different institutions/programs together
    - For postdoctoral fellows or established scientists to learn new techniques, knowledge, skills, in new or familiar areas of science
    - For experienced investigators to come together several times a year for a couple of days to solve interdisciplinary problems
- Boot camps and summer institutes
- Case study workshops
    - For discussing new solutions to difficult data science or unsolved problems
- Modular training
    - To provide graduate students or postdoctoral fellows an in-depth review of a particular scientific discipline
- Continuing medical education
    - To provide clinicians with information about the complexities of electronic health records

- Team challenges
    - In which groups of students compete to find the best solution for a research design
- Code-a-Thon-like intensive experiences
    - In which individuals with varying expertise come together for a short period to solve a problem that is stated in advance (http://www.health2con.com/devchallenge/code-a-thons/)

Innovative Training Technology

The workshop participants agreed that innovative technology could be used to enhance the experience of face-to-face training and to extend training offerings to a larger audience. Examples of innovative uses of technology include the following:

- Massive Open Online Courses (MOOCs).
    - Advantages include broad reach, scalability, and flexibility.
    - Disadvantages include a lack of physical connection and the inability of teachers to adjust to individual students (although they can adjust based on problems the whole class is having).
    - For which a number of issues must be addressed, such as evaluation of success, the initial expense, and updating/access once the course ends (in a rapidly changing field, updating is particularly important)
- Web-based videos and syllabi.
- Online learning tools.
- Technology-driven personalized content.
- Community-controlled online platforms for information sharing.

Data Sharing

There was a considerable amount of discussion of the need for data sharing in the context of optimal use of Big Data. One approach is to create a data center: large, controlled-access online data sets, together with analytic tools and a (potentially distributed) computing environment, which would provide a sandbox for training and education unique to BD2K. Although there were differing opinions on the value of this approach and not all training will require such online data sets, a small number of widely accessible resources would enable acquisition of key competencies across a large number of trainees and researchers in a cost-efficient manner. Access control and privacy issues, which can be taught via large online data sets, are as integral to training as analytics.

Curriculum Development

Curriculum development was considered in two contexts: as an integral part of an institutional training program and as a standalone curriculum. Broad sharing of curricula across institutions, especially community colleges, small institutions, and minority-serving institutions, was considered to be an important opportunity.

- Sharing outside the group for which the curriculum was designed will require that it be made publicly available and kept up-to-date.

- A tiered curriculum for groups needing different levels of knowledge, e.g. a curriculum to cross-train quantitative and non-quantitative students or members of a research team who bring different expertise to help solve the research problem, should be considered.

Prioritization

In the final session of the workshop, each participant offered his or her single highest priority. Many priorities were identified, in which the common themes were interdisciplinary and team-based training, diversity, openly-accessible data, and flexibility in approach to encourage multiple approaches to training.

APPENDIX I

Rationale for the Workshop

Big Data to Knowledge (BD2K), a new NIH initiative, aims to enable scientists to effectively manage and utilize the large, complex data sets (Big Data) that are already being generated and whose number and value will only increase in the future. The BD2K initiative is based on a set of recommendations on data and informatics from a working group to the Advisory Committee to the Director, NIH (see http://www.nih.gov/news/health/dec2012/od-07.htm).

The NIH seeks to increase the ability of the scientific workforce to utilize biomedical Big Data. Big Data creates challenges to the data pipeline, from acquisition and processing of the data to analysis and visualization. Utilization and analysis of this data will require new knowledge and skills beyond those traditionally employed in biomedical research. Furthermore, such abilities will be required at all levels, from students through established faculty, in a diverse and sustainable workforce. The workshop will, therefore, consider both a refocus of traditional training programs toward being cross-disciplinary, and the development of focused, short-term training programs that are potentially technology-enabled, web-based, or otherwise widely accessible to investigators at all levels.

The workshop will (a) identify the knowledge and skills needed by individuals and by collaborating teams to work productively with biomedical Big Data, and (b) discuss new resources and programs for educating and training both students and practicing scientists with the necessary knowledge and skills. The workshop will address the long- and short-term training needs of professionals and trainees with the purposes of increasing the number of (1) informaticians and computational and quantitative scientists who wish to apply their skills and knowledge in the biomedical, behavioral, and clinical sciences and (2) biomedical, behavioral, and clinical scientists who have the requisite knowledge and skills to effectively access, organize, analyze, and integrate large and complex data sets.

APPENDIX II

Agenda

Big Data to Knowledge (BD2K)
Workshop on Enhancing Training for Biomedical Big Data

29-30 July 2013
Terrace Level Conference Room
5635 Fishers Lane, Rockville, MD
Workshop Co-chairs: Karen Bandeen-Roche and Isaac Kohane

Workshop Goals:

1) Identify the knowledge, skills, and resources needed by biomedical research to organize, process, manage, and utilize large, complex data sets, and
2) Recommend and prioritize specific objectives for the NIH in training for Big Data.

This information will be used by NIH staff to develop short- and long-term training initiatives that prepare and empower the community to maximize the use of Big Data for research aimed at understanding human biology and improving human health.

Monday, 29 July

10:00	Welcome, Introductions, and Overview of BD2K Initiative	Mark Guyer
10:30	Purpose of the Workshop	Michelle Dunn
10:45	Summary of *Request for Information* Responses	Richard Baird
11:15	Intersection of BD2K with Director's Workforce and Diversity Initiatives	Sally Rockey
11:45	Discussion of the Goal and Vision for BD2K Training	K. Bandeen-Roche, Z. Kohane
12:15	Lunch – not provided	
1:15	Data Challenges and Competencies Needed (10 min presentation + 10 min discussion)	
•	Electronic Health Records	Dan Masys
•	Imaging	Ron Kikinis
•	Genomics	Mike Boehnke
•	Integration of Large or Small Datasets	Mark Musen
2:45	Discussion of BD2K Knowledge and Skills	Participants

- What are the necessary knowledge and skills that a Big Data team must include?
- How do the knowledge and skills needed vary according to the individual's:
    - Primary relationship to Big Data?
        - needing to be conversant
        - applying routine methods and tools
        - leading novel applications
        - developing new methods and tools

  - Primary training as basic, clinical, or quantitative scientists?
- How do we allow institutions adequate flexibility and still achieve the BD2K goals?

3:15 Break: Refreshments will not be provided

3:30 BD2K Characteristics of Long-term Training and Career Award Programs Participants

Approach

- What type of person should long-term training aim to produce?
- How should individuals be cross-trained?
- How could the curriculum and other program components be modified or developed so that a cross-trained student would not have a longer time from matriculation to graduation?
- The generation of new methods and software are essential for biomedical Big Data. Since computational and quantitative skills are broadly applicable, how should training programs encourage deployment or specialization of these skills in the biomedical field?
- What are the essential elements (e.g. courses, laboratory, clinical, or research rotations in industry, health care organizations, or government labs with big data) of a training program for a cross-trained student?

Environment

- What kind of an environment would be effective for BD2K-supported training?
- What would be a critical mass of students for a viable interdisciplinary program?
- What training program characteristics foster interaction between students trained in different disciplines, so that they learn from one another?

Policy

- Should NIH encourage common core elements in all BD2K-supported training programs?
- Should ALL training programs incorporate some elements of Big Data knowledge and skills into their curriculum?
- What should be the outcome of BD2K training programs and how should they be evaluated?

5:45 Brief Summary of Recommendations Z. Kohane, K. Bandeen-Roche

6:00 Adjourn until Tuesday, 30 July 8:30am

Tuesday, 30 July

8:30 Distillation of Day 1 K. Bandeen-Roche, Z. Kohane

9:00 Characteristics of BD2K Programs for Short-term Training Participants

- Who should the target audience be—undergraduates, faculty at undergraduate institutions, graduate students, postdoctoral fellows, new and experienced investigators, clinicians?
- What can short-term training accomplish? What concepts and skills can be conveyed via in this format? How would the success of such a program be evaluated? What are the metrics of success?

10:00 Characteristics of BD2K Programs for Innovative Training Technology Participants

- What innovative uses of technology could help 1) large numbers of students become familiar with basic core knowledge, or 2) established investigators acquire updated skills or an appreciation of new skills?
- How can online material be made interactive and adaptive to personalize delivery based on the learner's prior knowledge?
- How can NIH promote the development of training technologies specialized to biomedical Big Data?

11:00 Characteristics of BD2K Programs for Curriculum Participants

- Should NIH support curriculum development to encourage integrated, intersecting curricula?

11:30 Break (Working Lunch) -- Refreshments will not be provided

12:00 General Discussion Participants

- Are there particular challenges to keeping content updated? How can sharing be encouraged?
- How should success of the programs be evaluated? How can this activity be used to increase the number of students in research who are from underrepresented groups or less research-intensive institutions?
- What other training modalities should be considered (e.g. working groups, internships, etc.)?
- Of all the activities discussed, how would you prioritize them?
- Additional advice?

1:00 Summary of Workshop K. Bandeen-Roche, Z. Kohane

1:30 Adjourn

APPENDIX III

BD2K Workshop on Enhancing Training
Invited Guests

Kristine Alpi, MLS, MPH, AHIP
Director, William Rand Kenan, Jr. Library of Veterinary Medicine,
North Carolina State University
2 Broughton Drive
Raleigh, NC 27695
kmalpi@ncsu.edu
919-513-6219

Karen Bandeen-Roche, PhD
Professor and Chair, Bloomberg School of Public Health,
Johns Hopkins University
615 North Wolfe Street
Baltimore, MD 21205
kbandeen@jhsph.edu
410-955-3067

Mike Boehnke, PhD
Richard G. Cornell Distinguished University Professor of Biostatistics
University of Michigan
1415 Washington Heights
Ann Arbor, Michigan 48109-2029
boehnke@umich.edu
734-936-1001

Alex Bui, PhD
Professor of Radiology and Engineering
University of California, Los Angeles
924 Westwood Boulevard Suite 420
Los Angeles, CA 90095
buia@mii.ucla.edu
310-794-3540

Brian Caffo, PhD
Professor of Biostatistics
Johns Hopkins University
615 North Wolfe Street E3610
Baltimore, MD 21205
bcaffo@jhsph.edu
410-955-3504

Carlos Castillo-Chavez, PhD
Director, Mathematical, Computational and Modeling Sciences Center
Arizona State University
PO Box 871904 Tempe, AZ 85287
ccchavez@asu.edu
480-965-2115

Elissa Chessler, PhD
Associate Professor
Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
Elissa.chessler@jax.org
207-288-6453

Mark Cohen, PhD
Professor-in-Residence
University of California, Los Angeles
760 Westwood Plaza 17-369
Los Angeles, CA 90095
mscohen@ucla.edu
310-980-7453

Josh Denny, MD, MS
Associate Professor of Bioinformatics and Medicine
Vanderbilt University
2209 Garland Avenue 448
Nashville, TN 37212
Josh.denny@vanderbilt.edu
615-936-1556

Patricia Dombrowski, MA
Director, Life Science Informatics
Bellevue College
3000 Landerholm Circle
Bellevue, WA 98007
Patricia.dombrowski@bellevuecollege.edu
425-564-3164

APPENDIX III

BD2K Workshop on Enhancing Training
Invited Guests

Ary Goldberger, MD
Director, Margret & H.A. Rey Institute for Nonlinear Dynamics in Medicine
Harvard University/Beth Israel Deaconess Medical Center
330 Brookline Avenue Gz-435
Boston, MA 02215
agoldber@caregroup.harvard.edu
617-667-4267

Betz Halloran, DSc, MD
Professor of Biostatistics
University of Washington
1100 Fairview Avenue North
PO Box 19024
Seattle, WA 98109
betz@u.washington.edu
206-667-2722

Frank Harrell, PhD
Chair, Department of Biostatistics
Vanderbilt University
S-2323 Medical Center North
Nashville, TN 37232
f.harrell@vanderbilt.edu
615-322-2001

Larry Hunter, PhD
Director of the Center for Computational Biology and of the Computational Bioscience Program,
University of Colorado
12801 East 17th Avenue
MS 8303, RC1-North
Aurora, CO 80045
Larry.hunter@ucdenver.edu
303-724-3574

Robert Kass, PhD
Professor of Statistics,
Carnegie Mellon University
4400 Fifth Avenue Suite 115
Pittsburgh, PA 15213
kass@stat.cmu.edu
412-268-8723

Ron Kikinis, MD
Professor
Brigham and Women's Hospital
1249 Boylston Street 352
Boston, MA 02215
kikinis@bwh.harvard.edu
617-732-7389

Isaac Kohane, MD, PhD
Professor of Pediatrics, Chair of the Informatics Program
Boston's Children's Hospital, Dana-Farber/Harvard Cancer Center
300 Longwood Avenue
Boston, MA 02115
Isaac_kohane@harvard.edu
617-919-2184

Andrew Laine, DSc
Professor and Department Chair, Biomedical Engineering
Columbia University
351 Engineering Terrace
1210 Amsterdam Avenue
Mail Code 8904
New York, NY 10027
laine@columbia.edu
212-854-6539

Elaine Larson, PhD
Professor of Epidemiology, Associate Dean of Research, School of Nursing
Columbia University
Georgian Building
617 West 168th Street 246
New York, NY 10032
Ell23@columbia.edu
212-305-0722

APPENDIX III

BD2K Workshop on Enhancing Training
Invited Guests

Gary Marchionini, PhD
Professor, School of Information and Library Science
University of North Carolina
Manning Hall 103
Chapel Hill, NC 27599
gary@ils.unc.edu
919-962-8363

Dan Masys, MD
Affiliate Professor, Biomedical and Health Informatics
University of Washington
850 Republican Street, Building C
Seattle, WA 98109-4714
dmasys@uw.edu
360-797-3260

Mark Musen, MD, PhD
Professor, Co-Director, Biomedical Informatics Training Program
Stanford Center for Biomedical Informatics Research
Medical School Office Building
1265 Welch Road
Stanford, CA 94305
musen@stanford.edu
650-725-3390

Mike Newton, PhD
Professor of Statistics, Biostatistics and Medical Informatics
University of Wisconsin
1245A Medical Sciences Center
1300 University Avenue
Madison, WI 53792
newton@biostat.wisc.edu
608-262-0086

Lucila Ohno-Machado, PhD
Professor, Founding Chief, Division of Biomedical Informatics, Associate Dean for Informatics and Technology,
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093
machado@ucsd.edu
858-822-4931

Sastry Pantula, PhD
Division Director, National Science Foundation
4201 Wilson Boulevard 1025 N
Arlington, VA 22230
spantula@nsf.gov
703-292-9032

Giovanni Parmigiani, PhD
Chair and Professor, Biostatistics and Computational Biology, Dana-Farber/Harvard Cancer Center
450 Brookline Avenue
Boston, MA 02215
gp@jimmy.harvard.edu
617-632-3012

Steve Salzberg, PhD
Professor, Director, Departments of Medicine, Biostatistics, and Computer Science, Center for Computational Biology, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University
733 North Broadway
Miller Research Building 459
Baltimore, MD 21205
salzberg@jhu.edu
410-614-6112

APPENDIX III

BD2K Workshop on Enhancing Training
Invited Guests

Latanya Sweeney, PhD
Professor of Government and Technology in Residence
Harvard University
CGIS 1737 Cambridge Street Knafel 310
Cambridge, MA 02138
latanya@gov.harvard.edu
617-496-3629

Peter Szolovits, PhD
Professor, Computer Science and Engineering
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA 02139
psz@mit.edu
617-253-3476

Pablo Tamayo, PhD
Manager, Cancer Genome Informatics
Broad Institute of MIT and Harvard University
7 Cambridge Center
Cambridge, MA 02142
Tamayo@broadinstitute.org
617-714-7469

APPENDIX III

BD2K Workshop on Enhancing Training
NIH Participants

Richard Baird, PhD
Division Director, National Institute of Biomedical Imaging and Bioengineering
bairdri@mail.nih.gov

Vivien Bonazzi, PhD
Program Director, National Human Genome Research Institute
bonazziv@mail.nih.gov

Quan Chen, PhD
Health Scientist Administrator,
National Institute of Allergy and Infectious Diseases
Chenqn2@naiad@nih.gov

Sandra Colombini-Hatch, MD
Medical Officer, National Heart, Lung, and Blood Institute
hatchs@nhlbi@nih.gov

Jennifer Couch, PhD
Branch Chief, National Cancer Institute
Couchj@mail.nih.gov

Leslie Derr, PhD
Health Scientist Administrator,
Office of the Director
derrl@mail.nih.gov

Nancy Desmond, PhD
Office Director and Associate Director,
National Institute of Mental Health
ndesmond@mail.nih.gov

Michelle Dunn, PhD
Program Director, National Cancer Institute
Dunnm3@mail.nih.gov

Valerie Florance, PhD
Division Director, National Library of Medicine
florancev@mail.nih.gov

Nick Gaiano, PhD
Scientific Review Officer, Center for Scientific Review
gaianonr@mail.nih.gov

Jose Galvez, MD
Program Director, National Cancer Institute
galvezjj@mail.nih.gov

Maria Giovanni, PhD
Assistant Director for Microbial Genomics, National Institute of Allergy and Infectious Diseases
Mg37u@nih.gov

Bettie Graham, PhD
Division Director, National Human Genome Research Institute
graham@odder.nhgri.nih.gov

Eric Green, MD, PhD
Director, National Human Genome Research Institute
Acting Associate Director for Data Science
egreen@nhgri.nih.gov

Susan Gregurick, PhD
Division Director, National Institute of General Medical Sciences
susan.gregurick@nih.gov

Mark Guyer, PhD
Deputy Director, National Human Genome Research Institute
guyerm@exchange.nih.gov

Lynda Hardy, PhD, RN
Program Director, National Institute of Nursing Research
hardylr@mail.nih.gov

Ming Lei, PhD
Branch Chief, National Cancer Institute
leim@mail.nih.gov

APPENDIX III

BD2K Workshop on Enhancing Training
NIH Participants

Peter Lyster, PhD
Program Director, National Institute of General Medical Sciences
lysterp@nigms.nih.gov

Ronald Margolis, PhD
Senior Advisor, National Institute of Diabetes and Digestive and Kidney Diseases
margolisr@mail.nih.gov

Veerasamy Ravichandran, PhD
Program Director, National Institute of General Medical Sciences
ravichanr@nigms.nih.gov

Sally Rockey, PhD
Deputy Director for Extramural Research, National Institutes of Health
rockeysa@od.nih.gov

Erica Rosemond, PhD
Program Officer, National Institute of Mental Health
rosemonde@mail.nih.gov

Cathrine Sasek, PhD
Science Education Coordinator, National Institute on Drug Abuse
csasek@nih.gov

Carol Shreffler, PhD
Program Administrator, National Institute of Environmental Health Sciences
Shreffl1@niehs.nih.gov

Heidi Sofia, MPH, PhD
Program Director, National Human Genome Research Institute
sofiahj@mail.nih.gov

Erica Spotts, PhD
Health Scientist Administrator,
Office of the Director
spottse@mail.nih.gov

Jennifer Sutton, MS
Extramural Program Policy and Evaluation Officer,
Office of the Director
suttonj@mail.nih.gov

APPENDIX IV

NIH BD2K Training Working Group Members

Richard Baird (NIBIB)

Vivien Bonazzi (NHGRI)

Quan Chen (NIAID)

Sandra Colombini-Hatch (NHLBI)

Leslie Derr (OD)

Michelle Dunn (NCI)

Valerie Florance (NLM)

Nick Gaiano (CSR)

Jose Galvez (NCI)

Bettie Graham (NHGRI)

Mark Guyer (NHGRI)

Linda Hardy (NINR)

Ming Lei (NCI)

Veerasamy Ravichandran (NIGMS)

Erica Rosemond (NIMH)

Catherine Sasek (NIDA)

Carol Shreffler (NIEHS)

Heidi Sofia (NHGRI)

Scott Somers (NIGMS)

Erica Spotts (OD)

Jennifer Sutton (OD)

APPENDIX V

ACD Working Groups on Biomedical Workforce **and *ACD Working Group on Diversity in the Biomedical Research Workforce***

Dr. Sally Rockey, NIH Deputy Director for Extramural Research Director, presented an update on the NIH's responses to the reports from these two ACD working groups. The charge to the Biomedical Workforce Working Group (BMW WG) was to (1) develop a model for a sustainable and diverse U.S. biomedical research workforce that can inform decisions about training of the optimal number of people for the appropriate types of positions that will advance science and promote health and (2) recommend actions that NIH should take to support a future sustainable biomedical infrastructure. Dr. Rockey presented data which showed that (1) the number of PhDs in biomedical sciences is increasing while the number of PhDs in chemistry has remained about the same; (2) most doctoral students are supported as research assistants on research grants; (3) the age when biomedical doctoral students get their first non-postdoc or tenure track job is around 36 years compared to about 33 years for those with doctoral degrees in chemistry; (4) the average age of PIs awarded their first R01 or equivalent is 40 years; (5) early in their careers, biomedical scientists earn less than those with degrees in math, physical and social sciences, and engineering, which results in a significant loss in lifetime earnings; and (6) only 2% of the NIH-trained workforce are unemployed, 43% in academic research, and 55% are employed in other science-related activities. The BMW WG recommended (1) for graduate students--shorten and diversify the training and increase financial support; (2) for postdoctoral fellows--increase financial support and training for more than academic careers; (3) for physician scientists--conduct a focused follow-up study; and (4) for staff scientists--encourage study sections to consider them valuable members of the research team. The BMW WG also recommended that NIH gradually reduce the percentage of funds from NIH grants used for salary support and institute a more vigorous evaluation of programs and encourage stronger coordination amongst programs.

NIH has put several efforts in place to respond to the recommendations:

- The eligibility period for postdoctoral students to apply for the K99/R00 will be shortened from five years to four years, effective February 2014
- NIH is in the process of reviewing applications to institutions that responded to a RFA that called for new approaches to broadening the training experiences of pre-and postdoctoral students to reflect the range of career options of trainees (http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-12-022.html)
- All NIH-supported trainees will be required to have an Individual Development Plan (IDP) in place by October 1, 2014 (http://grants.nih.gov/grants/guide/notice-files/NOT-OD-13-093.html)
- Postdoctoral stipends will be increased in FY2014.

NIH will also be encouraging institutions to reduce the length of graduate training; mandating that all NIH Institutes and Centers support F30 and F31 fellowships by April 2014; developing a comprehensive survey on benefit policies and NIH support of faculty salaries; developing a comprehensive tracking system for all trainees; and creating a unit at NIH to assess the biomedical workforce.

Dr. Rockey also discussed the recommendations of the Working Group on Diversity in the Biomedical Research Workforce (WGDBRW**).** She noted that the NIH has been committed to increasing the diversity of the biomedical workforce, and for over 30 years it has supported programs to achieve this goal through institutional and individual programs. However, a paper in Science in August 2011[3] highlighted concerns regarding race, ethnicity, and the awarding of research grants. Even when controlled for institutions, African-American scientists had a lower award rate. On the basis of the recommendations of the WGDBRW, NIH has now developed a comprehensive strategy to redress the problem, which includes the following:

- Establishing a new leadership position, Chief Officer for Scientific Workforce Diversity (the recruitment is underway for a permanent leader).
- Making the effort to increase the pipeline through a new initiative, Building Infrastructure Leading to Diversity (BUILD).
- Developing a National Research Mentoring Network (NRMN).
- Making new efforts to ensure fairness in peer review.

[3] Science (v 333), 19 August 2011; pp 1015-1019.