RFI Summary: Executive Summary

On February 20, 2013, the NIH issued a Request for Information titled "Training Needs In Response to Big Data to Knowledge (BD2K) Initiative." The response was large, with 103 responses received from individuals, departments, companies, and professional organizations. A theme that emerged throughout is that meeting the challenge of Big Data requires closely-collaborating scientists who are experts in their domain but also cross-trained, so that they can form high-functioning teams. There is a need to develop both human-focused and technical skills, the latter including skills in three basic categories: (1) statistics and math, (2) computation and informatics, and (3) a domain (e.g., biomedical, behavioral, clinical) science.

RFI respondents confirmed that education and training is needed at all professional levels, from undergraduates all the way through senior faculty and clinicians, and at varying levels of detail, with the assumption that all types of academic institutions need to be included. Respondents gave suggestions about how to train individuals for the new Big Data challenges: a consistent theme was the need for hands-on, immersive experiences with data and a deep understanding of the ideas underlying the methodology in order to learn persistent skills that transcend particular data types or tools. Specific suggestions for training include the following:

1) Modify features of existing training programs (e.g. use dual mentors; start with boot camps to teach basics of less familiar area; offer joint classes and assignments that foster teamwork across the two areas)

2) Offer short-term experiences and courses on focused topics (workshops, lectures, lab rotations)

3) Develop novel and technology-enabled learning systems that reach a wide audience (MOOCs, online courses, online/physical hybrids, cognitive tutors)

4) Develop and share new curricula (new courses focused on data science principles and practices, new modules to build into existing courses, use cases that can be incorporated into existing classes)

5) Encourage broad dissemination (materials should be publicly available online under open license and shared through web portal of online learning resources)

RFI Summary: Full Report

On February 20, 2013, the NIH issued a Request for Information titled "Training Needs In Response to Big Data to Knowledge (BD2K) Initiative." The response was large, with 103 responses received from individuals, departments, companies, and professional organizations.

Knowledge and Skills Needed to Utilize Biomedical Big Data

Both technical knowledge and skills and human-based competencies are needed to make discoveries with Big Data. Human-based competencies include working well in teams, communicating, and ethics training. At the highest level, technical knowledge and skills could be classified as coming from three major areas: (1) statistics and mathematics, (2) computational sciences, and (3) domain science. A solid foundation in biomedical or behavior domain science, by individuals or within teams, was considered essential.

Also essential for Big Data teams is knowledge or expertise in fundamental "data skills" to manage and process big data, which may require programming in a variety of languages. Platforms such as R or SAS and SQL-based databases are sufficient to a certain scale but eventually break as the data size increases and data sorting and aggregation become more complex. This is the entry point for what can be called Big Data technologies, such as cloud computing, MapReduce, NoSQL, and parallel databases. Scientific and information visualization is critical to make sense of the data. A common theme in the RFI responses was the need for semantic skills with ontologies and metadata, and it was pointed out that academic library science is a good source of this expertise. Efforts to include all possible types of Big Data are necessary, particularly from the clinical perspective but also from disciplines outside the typical biomedical sciences.

Much of the data analysis itself involves statistics and machine learning methods. High-level analysis includes modeling, statistical inference, decision theory, and evaluation of properties of estimators. A compelling need for data scientists with the skills for reproducible research was observed. Finally, new ways to present and visualize Big Data are important in fostering greater utilization and understanding of the opportunities.

Teamwork and Cross-training

A theme that emerged throughout is that meeting the challenges of Big Data requires teams of closely-collaborating scientists who have been cross-trained in multiple areas. High-functioning teams are critical for success in utilizing Big Data. They are a practical and effective way to find, develop, deploy, and retain the needed combination of skill sets across disciplinary boundaries. Relatively few responsers maintained that a perfect bio-computational hybrid could be trained or was practical. Team members need an "empathy (or at least sympathy) and respect for those in other disciplines." As one respondent described, “Computation-centric researchers cannot just be viewed as human computers, to whom one delivers Excel spreadsheets and from whom one receives ‘results’. Multi-disciplinary teams need to share a common goal and have a vested interest in other team members’ success.”

Another common theme among the responses was the need to cross-train individuals across traditional disciplinary boundaries such as biology and computation. Cross-training is needed so that team members can communicate effectively with each other. Successful cross-training should help individuals to develop and/or strengthen their core expertise in one domain (or more) while learning about, connecting with, and becoming strongly committed to another domain. An effective joint curriculum will strengthen both the data awareness of biomedical researchers and the bioscience awareness of computational researchers; thus, strong teams and cross-training have the same ultimate purpose: to produce successful, expert collaboration.

Audience for Data and Informatics Training and Education

RFI respondents confirmed that education and training are needed at all professional levels, from undergraduates all the way through senior faculty and clinicians, and at varying levels of detail for both basic and clinical scientists:

- Practicing scientists and clinicians may need to update their skills to take better advantage of Big Data resources for basic research and the translation of information for health care.
- Trainees (postdocs, doctoral students, Masters-level students, postbacs) should be exposed to data analysis and informatics either in dedicated courses or woven throughout their training.
- Undergraduates should be exposed to more opportunities to develop quantitative and computational skills to improve the baseline skills of entering graduate students.

Respondents also advocated for promoting data science at the high school level to engage interest and cultivate the foundational knowledge of students early in the pipeline. Although NIH rarely enters the K12 domain (the development and distribution of NIH curriculum units is an exception), online courses and modules in introductory subjects would be accessible to K12 students. Another respondent pointed out that managers of the research enterprises in academia and industry need to understand the existing inherent potential of Big Data, as well as the means to achieve functioning collaborative teams.

Some respondents described the difficulty of finding and retaining skilled individuals who often can be better compensated in industry. However, innovative programs and flexible environments could also attract individuals with cutting-edge skills to new intellectual challenges. It is important for NIH to address not only attracting Big Data talent to biomedical research, but also keeping that talent by encouraging the development of rewarding career paths that may accommodate rapidly shifting priorities, responsibilities, and opportunities.

Respondents were supportive of efforts to help increase diversity in the workforce and gave some concrete suggestions to achieve this goal:

- Encourage major public and private institutions to partner with minority serving institutions such as HBCUs and HSIs
- Provide access to educational and technical resources necessary to teach students Big Data skills, such as hardware and software or access to cloud computing
- Engage high school students in data, e.g. encourage data management and analysis components in science fair projects or in classroom projects by working with the National Association of Biology Teachers.

Solving the problems of Big Data will a much larger and diverse workforce, so education and training for all interested groups – regardless of professional level, demographic, or educational background – is important.

How Training Should be Achieved

Because the target audience is diverse both scientifically and by career stage, a variety of approaches is needed to deliver the knowledge and skills needed to utilize biomedical Big Data. Suggested approaches include development of (1) training programs for long-term training, (2) short-term experiences, and (3) innovative use of technology to reach a broad scale. Also emphasized was the need for curriculum development and sharing, as well as the development of resources for teaching and learning.

(1) Modify the features of existing training programs to produce cross-trained individuals who work well in Big Data teams while having a deep knowledge in multiple areas. Features include

- Dual mentors, one in each domain,
- Boot camps to teach entering graduate students the basics of the less familiar area, and
- Joint classes and assignments that foster teamwork across the two areas.

A respondent suggested that NIH require that all existing NIH-supported training programs ensure students obtain a set of minimal competencies in computational and statistical methods for biological research and that the availability and success of such efforts be a proposal evaluation criterion.

(2) Develop short-term experiences and courses on focused topics necessary for Big Data, aimed at all career levels. Examples include the following:

- Short-term workshops (including lectures, lab rotations, and shadowing) to allow one group of researchers (e.g., biologists, other basic scientists, clinicians, informaticians) to learn more about the field of a different group of researchers and to help break down some of the communication barriers by developing a common vocabulary.
- Summer intensive programs with topics of general interest (e.g. data mining) along with breakouts that provide specific information (e.g. particular databases or tools). In person workshop attendance could be optimized by having participants participate in some online training sessions before coming together as a group.

- Summer programs that aim to recruit undergraduate students, giving intellectually motivated students early exposure to Big Data challenges, similar to the SIBS program in biostatistics.

(3) Develop novel or technology-enabled learning systems or environments to reach a wide audience including trainees and researchers. Distance learning has the potential to engage non-traditional sources of talent; in this case, it could be used to expose biomedical scientists to data science and vice versa.

- Examples include MOOCs, online courses, online/physical hybrids, and cognitive tutors.
- Web-based and flexible educational offerings, including certification programs, are being rapidly created and disseminated, particularly in the quantitative and computational sciences.

All three of these approaches to knowledge acquisition require curricula (for new courses, new modules, and use cases). The curricula should have the following content and characteristics:

- Cover the intersection of data science and biomedical science that fuse topics from informatics and the computational and quantitative areas with health, behavioral, and clinical topics, such as health service research.
- Include mentored research projects, which requires access to faculty with appropriate expertise and datasets that can answer important research questions in health and healthcare.
- Include modules (e.g. CME courses) for clinicians (as well as medical students and residents) that focus on skills needed to utilize large datasets, as the importance of Big Data in everyday clinical situations grows.

RFI respondents stressed the need not only to develop but also to share new curricula. Broad dissemination of curricula should be encouraged, including making them publicly available online under open license and shared through a web portal of online learning resources. Sharing is a theme that applies not just to curricula but also to training itself through online lectures, to data sources, and even to virtual machines on the cloud pre-loaded with data and tools so that students and instructors can skip the difficulties of installation and configuration.

The themes that emerged from the RFI responses centered around inclusiveness. Respondents urged NIH to push for sharing of resources such as curricula and courses through web-based portals. They identified knowledge and skills, whether quantitative, computational, or domain, that are needed at all professional levels. Finally, they described the necessity of teams, inclusive of informatics experts, computational scientists, quantitative scientists, as well as the biomedical scientists who are domain experts.