Annual Report on Investments in Training and Education for the NIH Big Data to Knowledge (BD2K) Initiative

Report on investments during FY14-16

Prepared by the BD2K Training Program Management Working Group

March 2017

Executive Summary

Training is a major limiting factor to extracting knowledge from data, and therefore it is a significant part of the Big Data to Knowledge (BD2K) Initiative. One of the primary aims of the BD2K Initiative is to enhance training in the development and use of methods and tools necessary for biomedical Big Data science. To accomplish this aim a number of training and education opportunities have been developed that are designed to increase the number of biomedical data scientists and to improve the data science skills of all biomedical scientists.

Because training needs vary greatly based on an individual's prior background and intended use for data science, the BD2K investments in training are also varied. For those individuals who are primarily biomedical scientists and do not intend to become specialists in data science, BD2K supports courses and educational resources that are meant to enable participants to become conversant in data science and attain skills to utilize data science methods. To support those individuals who wish to become specialists in biomedical data science, BD2K includes training programs for predoctoral students, research rotations for early career scientists, and career development awards for postdocs and more senior researchers. To foster the development of new teams consisting of biomedical scientists and data scientists, BD2K is supporting the QuBBD (Quantitative Approaches to Biomedical Big Data) Program along with the National Science Foundation.

Among the first BD2K awards issued (in September 2014) were research education awards and mentored career development awards, and they are starting to bear fruit. For example, the first BD2K Open Educational Resources, in the form of Massive Open Online Courses, have been released and already boast thousands of graduates; three summer courses were offered and filled to beyond capacity; and individuals supported by career development awards have transitioned to independent research positions. Although the existing awards made in FY14-16 are promising, additional investments are needed to keep up with the fast-changing area of data science.

In order for BD2K-supported resources to have maximal impact, they need to be findable, accessible, interoperable, and reusable (FAIR). To help biomedical scientists find and access data science educational resources, the BD2K Training Coordination Center (TCC) has developed an Educational Resource Discovery Index. To enhance this resource the TCC is working with international partners such as ELIXR that provides an online training portal that gathers life science training materials and training courses from across Europe, and allows you to search it in one website. Through the TCC and the other training awards, BD2K aims to improve the ability of the entire biomedical science community, whether specialists in biomedical science or in data science, to utilize the growing volume and complexity of data.

Overall Goals

Focusing on training was one of the main recommendations from the Data and Informatics Working Group Report to the Advisory Committee to the Director (link to June 2012 report here). It is also one of the major thrust areas of the BD2K Initiative and the NIH Office of the Associate Director for Data Science. Training currently accounts for approximately 15% of the budget for the BD2K Initiative with FY16 expenditures of nearly $17 million for new and continuing awards. The term "Training" is meant to encompass training, education, and workforce development that provides learners, no matter what career level, either foundational knowledge or skills for immediate use.

The two main goals for BD2K training are to improve big data skills in biomedical scientists and to increase the number of people who specialize in biomedical data science.   These two goals include sub-goals as well:

1) To improve big data skills of biomedical scientists
   A. Support training opportunities, both in-person and online
   B. Ensure training opportunities and resources are more readily discovered and accessed
   C. Enhance diversity in the biomedical and biomedical data science workforces
2) To increase the number of biomedical data scientists
   A. Establish biomedical data science as a career path
   B. Foster collaborations between biomedical scientists and data scientists

To accomplish these goals, the currently supported training portfolio is diverse. Funding opportunities were issued (see Appendix A) that were designed to be responsive to the two goals:

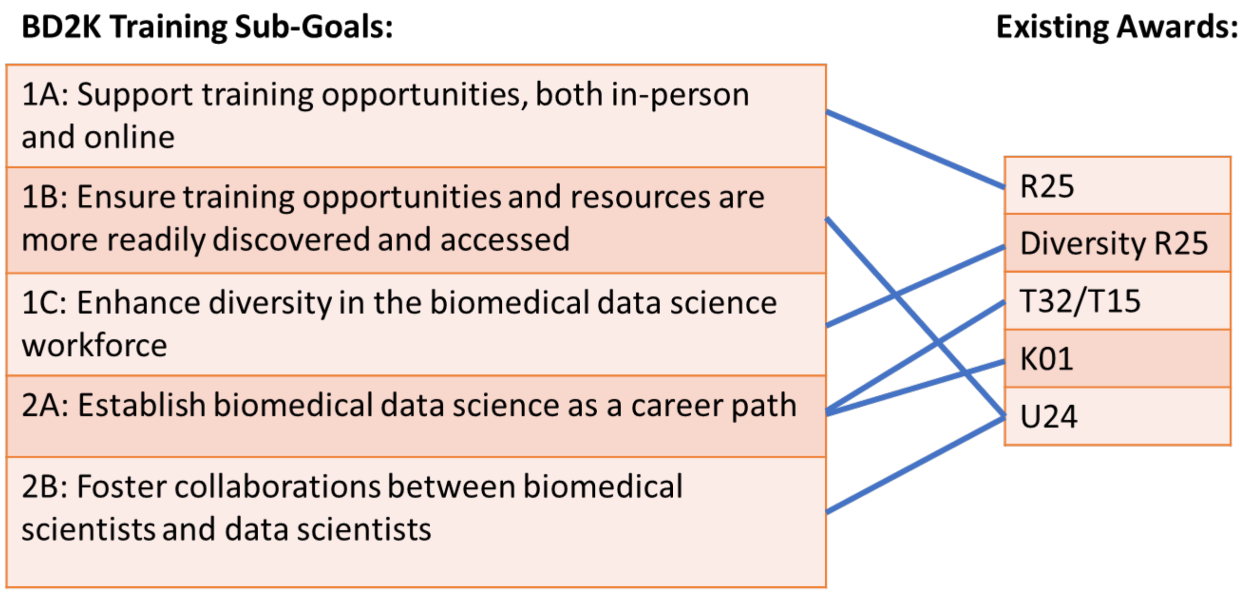
Goal 1: To improve the big data skills of biomedical scientists:

- Goal 1A: Research Education Grants (R25) to support the development of courses and resources in data science, data management and data sharing, annotating and curation (4 funding opportunities were issued to allow for different budgetary categories and structures)
- Goal 1B: Training Coordination Center (TCC) (U24)
- Goal 1C: Research Education Grants (R25) to enhance diversity (denoted dR25 to distinguish it from the other R25s)

Goal 2: To increase the number of biomedical data scientists:

- Goal 2A: Mentored Career Development Award (K01) and Predoctoral Training Award Programs (T32 and T15) (3 funding opportunities to support both new programs and supplements to existing T32s and T15s)
- Goal 2B: Training Coordination Center (U24) to provide support for innovative approaches to fostering collaborations (Note: BD2K also supports the QuBBD (Quantitative Approaches to Biomedical Big Data) Program in collaboration with the National Science Foundation)

Figure 1: Existing Awards Map onto One or More BD2K Training Sub-Goals.


BD2K Training Sub-Goals:
Existing Awards:
1A: Support training opportunities, both in-person and online
1B: Ensure training opportunities and resources are more readily discovered and accessed
1C: Enhance diversity in the biomedical data science workforce
2A: Establish biomedical data science as a career path
2B: Foster collaborations between biomedical scientists and data scientists
R25
Diversity R25
T32/T15
K01
U24

The first set of funding opportunities were issued in FY14 and FY15. During FY16, a portfolio analysis was conducted and based on a continued need and identified gaps in the portfolio several funding opportunities were issued. These include opportunities that were reissued for support for predoctoral training grants (T32), mentored career development awards (K01), open educational resources for skills development (R25), and research education grants to enhance diversity and build capacity at under resourced institutions (dR25). Two new opportunities were also released to support the transition of intramural investigators to independent research careers in the extramural community (K22) and to support the creation or significant expansion of courses in Data Science targeting undergraduate and/or graduate students with the objective of enabling institutions to easily incorporate tailored data science skills into their standard curriculum for biomedical scientists (R25) (see Appendix A). At the time of this report there are no results to report on from the reissued or new opportunities.

Figure 2: Number of Awards by Goal. A total of 70 grants or awards were issued by BD2K in response to the two goals (Appendix B). There are 33 awards providing support

to improve the big data skills of all biomedical scientists (Goal 1) and 37 awards providing support to increase the number of biomedical data scientists (Goal 2). For the purposes of this report, the Training Coordination Center Award was assigned to Goal 1. During FY16 several administrative supplements were issued to supplement and enhance the activities of some of the awards that are not accounted for in Figure 2 but are listed in Appendix B.

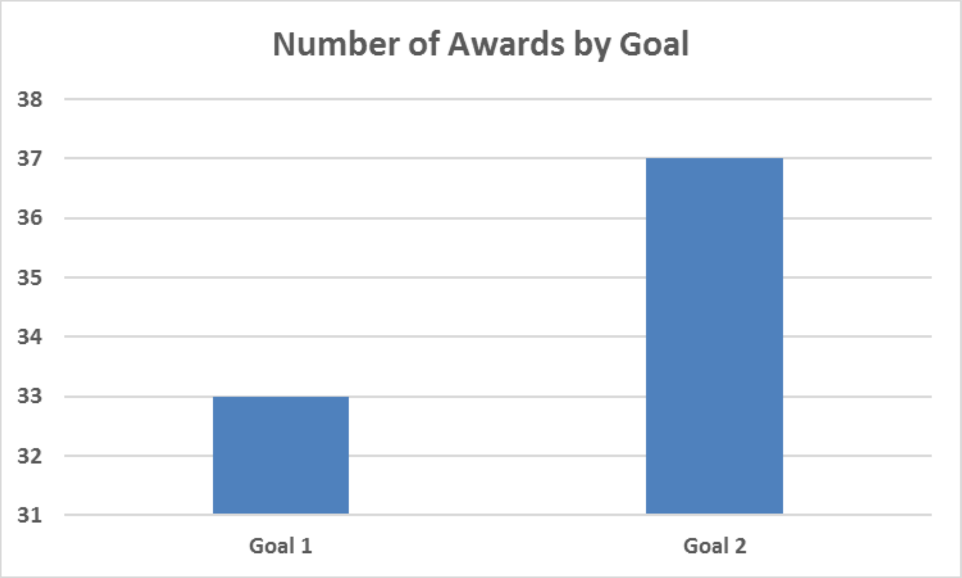
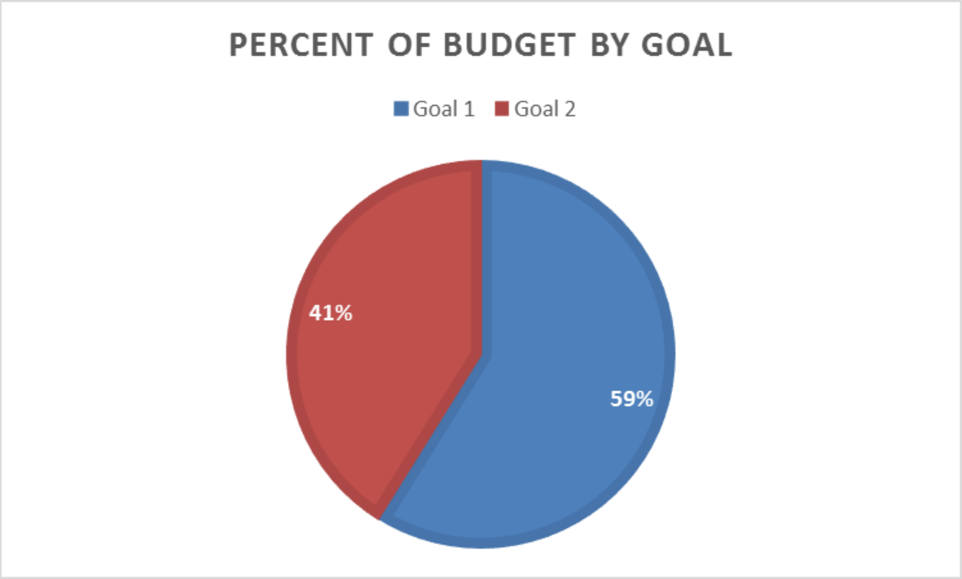


Figure 3: Percent of Budget by Goal. During FY16 the training component of the BD2K Initiative expended 59% of the training budget on Goal 1 and 41% of the training budget on Goal 2.



In addition, each of the BD2K Centers of Excellence has training components. Because the Centers' training is diverse and mainly focused on training about specific tools developed by the Centers, they are not included in this report.

Collectively, BD2K training reaches an audience of varying experience levels and intentions, from undergraduates to senior faculty and instructors who will be teaching data science. Some of the awards are targeted at particular career levels. For example, the R25 awards to enhance diversity are for undergraduates, the T32/T15 programs are for predoctoral trainees, and the K01 awards are for postdoctoral fellows and beyond. Other awards, such as the U24 TCC and the R25s can accommodate a broad range of experience levels.

Figure 4: Number of Awards by Grant Mechanism (FY16). From FY14-16 a total of 70 awards have been issued in the area of training and education for the BD2K Initiative. Figure 4 displays the number of awards issued for each grant mechanism. They are: dR25: Research Education Grant to support activities to diversifying the workforce and building capacity at under resourced institutions; R25: Research Education Grants to support the development of courses and resources; T32/T15: Training Grants to support the development of predoctoral training programs; K01: Mentored Career Development Award to support mentored career development; U24: Cooperative Agreement to support the development of the BD2K Training Coordination Center.

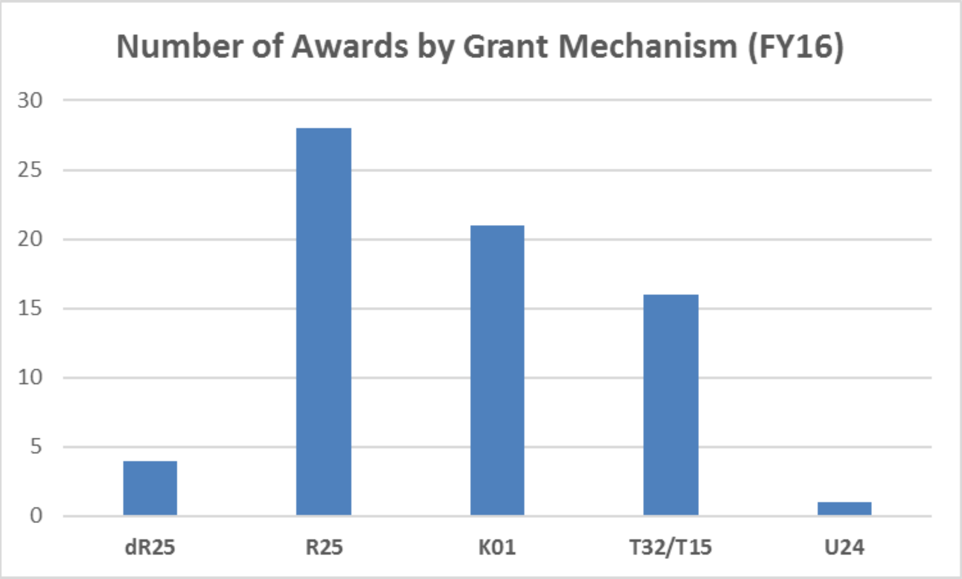


Figure 5: Awards as a percentage of total budget dedicated to training and education for the BD2K Initiative during FY16.

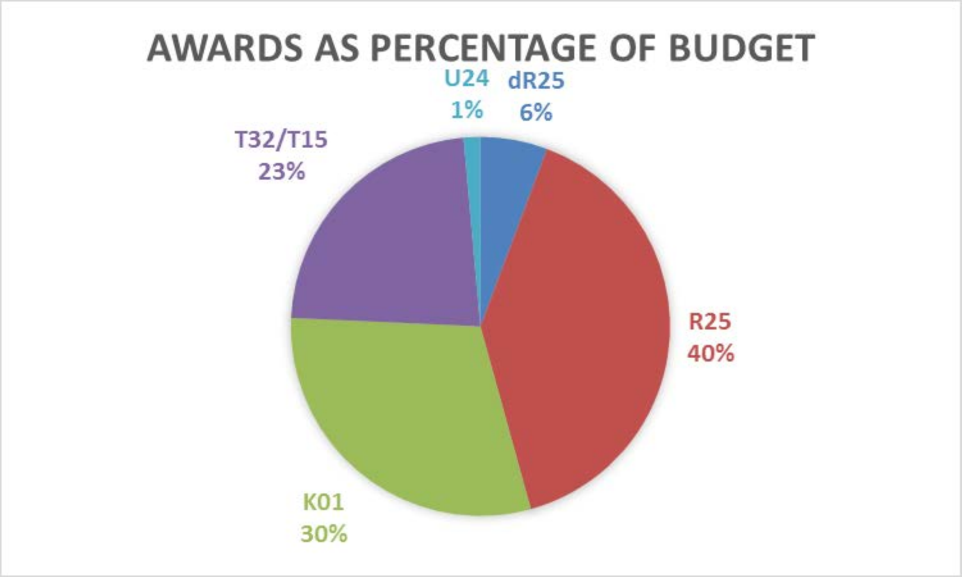
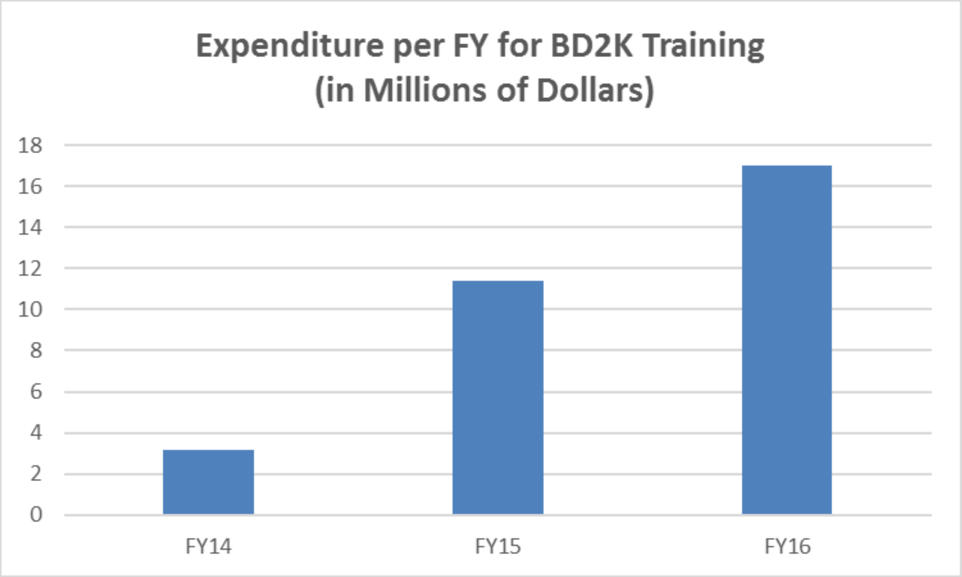


Figure 6: Total Expenditure to Date for BD2K Training Efforts in millions of dollars. The expenditures include new and continuing award costs.



Because BD2K is a trans-NIH program, with funds coming from all NIH Institutes and Centers (ICs) and the NIH Common Fund, IC input has been actively solicited. The funding opportunity announcements (FOAs) were developed by a trans-NIH group of program directors, first called the "BD2K Training Subcommittee" and later referred to as the "BD2K Training Program Management Group". This group has welcomed all-comers, from all ICs and Offices within the NIH Office of the Director, with effort to recruit members through two presentations to the NIH Training Advisory Committee. All programmatic aspects, including FOA development, review attendance, pay plan development, and decisions about award management, have been done through the trans-NIH group that includes 13 ICs, the Common Fund, and the Office of Behavioral and Social Science Research (see Appendix G). Day-to-day management of the awards is distributed across 5 ICs to include as many ICs as possible and to ensure that grantees are treated uniformly.

Figure 7: Number and Distribution of Awards by Managing IC. Note that during FY16 one-year administrative supplements were issued to the NIH Science Education Partnership Award (SEPA) Program that is managed by the NIH OD in the Office of Research Infrastructure Programs (ORIP). The purpose of these supplements was to incorporate data science into curriculum that is aimed to increase student interest in pursuing biomedical research careers and increase understanding of the scientific research progress and/or hold professional development courses for K-12 teachers that would enhance their knowledge of data science. The budgets of these awards are accounted for in previous figures using the R25 grant mechanism but were not counted as individual awards per se as they are supplements to grants that did not originate from the BD2K Initiative.

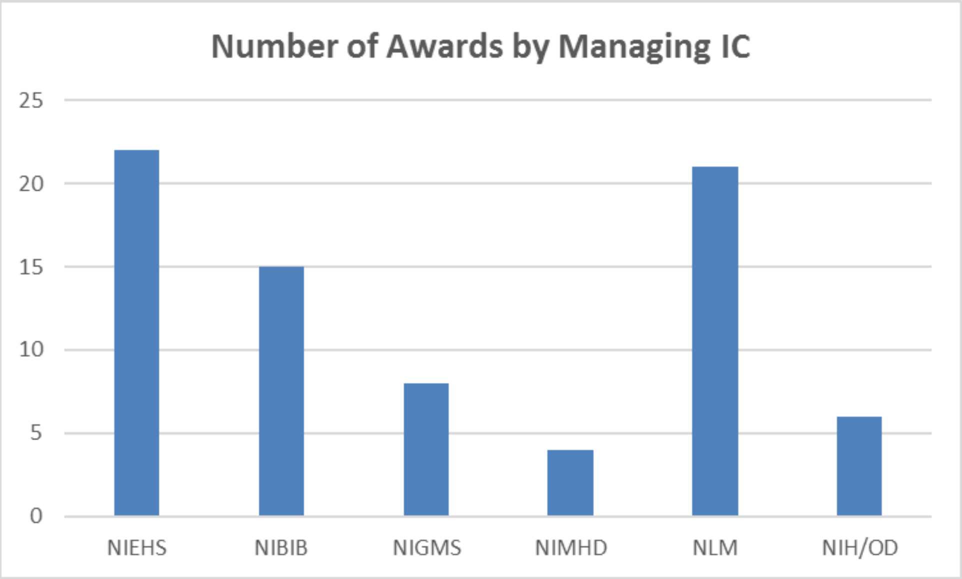
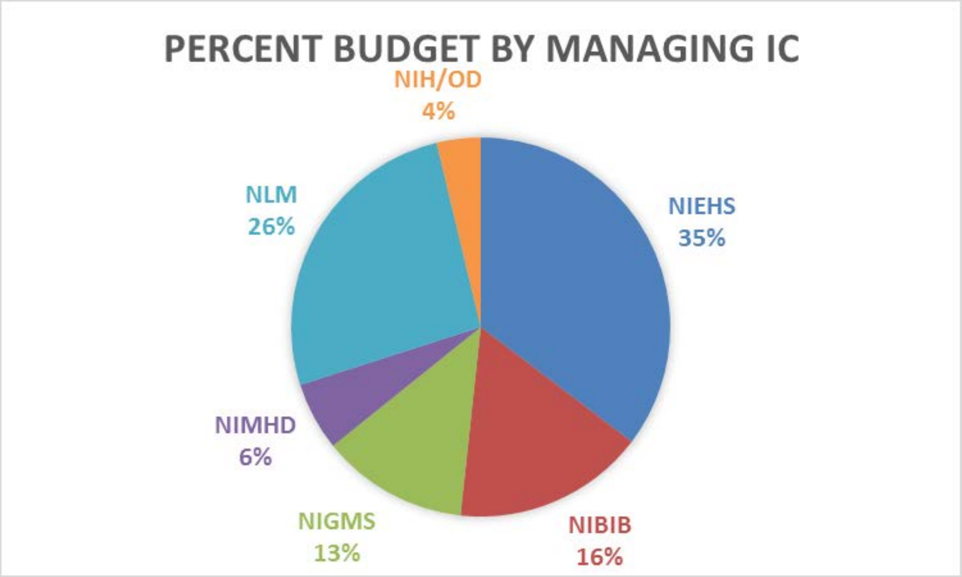


Figure 8: Percent Budget by Managing IC. Distribution of awards across the NIH as a percentage of total budget dedicated to training and education for the BD2K Initiative during FY16.



Goal 1: To Improve Big Data Skills of Biomedical Scientists

To improve the big data skills of biomedical scientists, the BD2K Initiative provides support for the development of training opportunities, both in-person and online; ensures training opportunities and resources are more readily discovered and accessed; and provides support to enhance diversity in the biomedical and biomedical data science workforces.

Goal 1A: Support training opportunities, both in-person and online

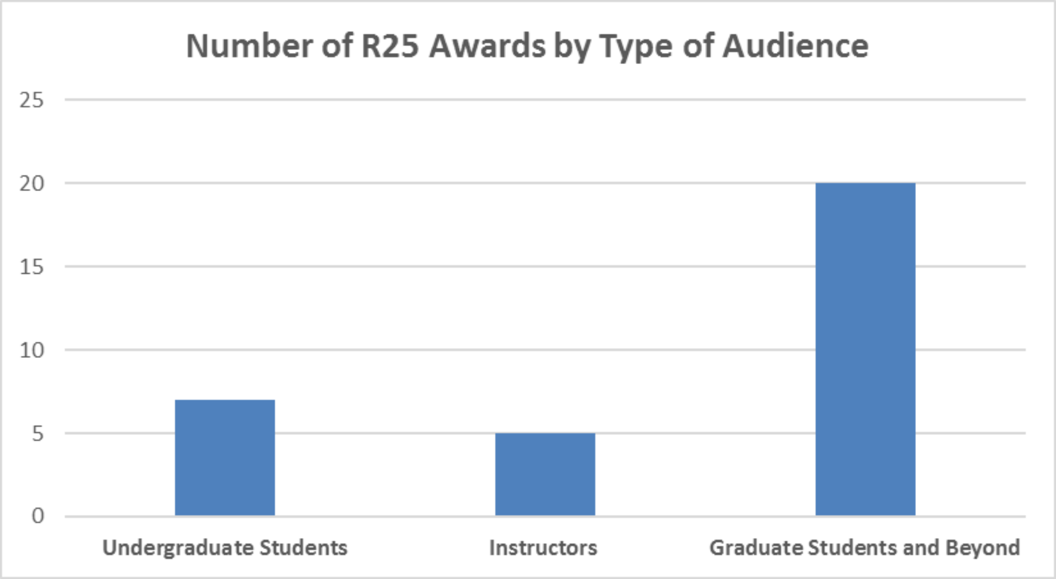
The BD2K open educational resources and short courses aim to introduce biomedical scientists to data management and data science. These programs were designed to be flexible, to allow for innovation and for specialization to a particular audience, domain science, or data science. Although a few of the funded programs are confined to a narrow audience or scientific area, most are for general audiences and multiple data types. During FY14-FY16, a total of 32 R25 programs were funded, and these programs have a broad reach geographically. More detail about the audience, scientific breadth, and reach follows.

Audience:

The R25 awards address a variety of educational levels (undergrads to senior faculty) and intended usages (end users or instructors).

- Five programs focus on helping instructors of advanced undergrad/early grad courses. Examples include:
- A train-the-trainers course in biomedical data science for instructors, who will collectively develop a curriculum for undergraduates at non-research-intensive colleges
- Curricular materials (slides, assessments, reading suggestions) that can be used and adapted by other instructors
- A Toolkit to help librarians teach data management to biomedical scientists
- Seven programs target undergrads directly, through summer programs (including didactic and research experiences) that aim to recruit data science students into biomedical science or to expose biomedical students to data management and data science. Four of the seven programs have a primary goal of enhancing diversity through partnering with BD2K Centers.
- The remaining 16 programs focus directly on the graduate student or the more advanced learner; although the data management and data science material is introductory, because it is new, learners are just as likely to be senior faculty as graduate students.

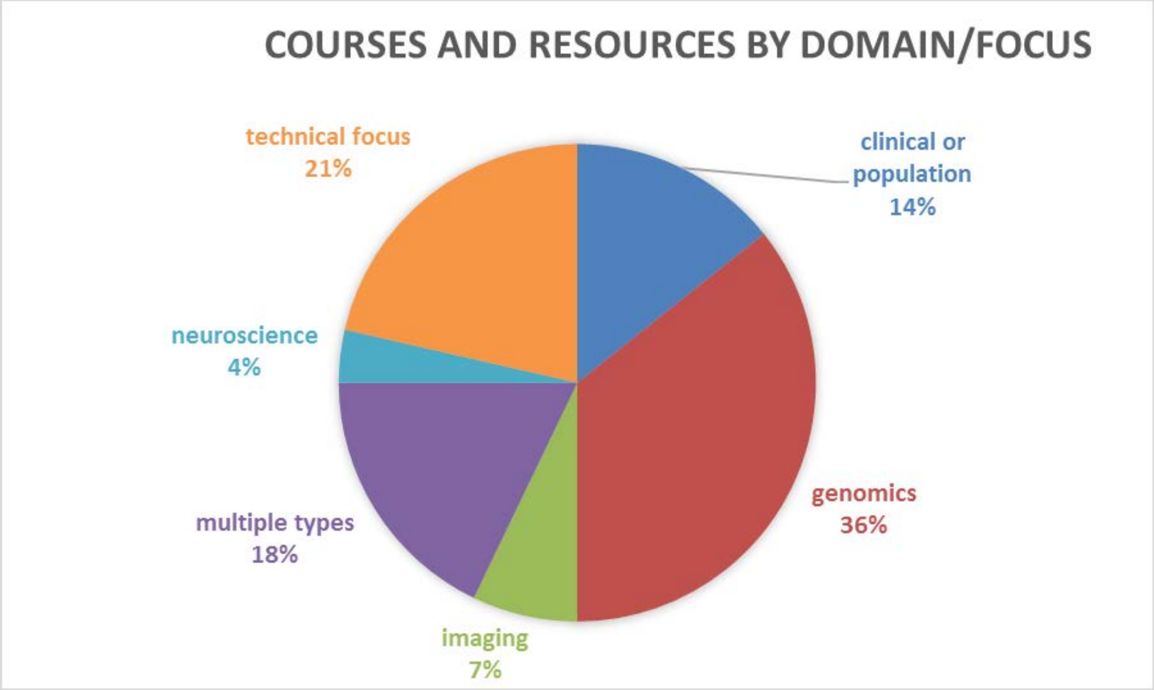
Figure 9: Number of R25 Awards by Type of Audience



Scientific Breadth:

Although the majority of the programs aim for a general biomedical audience, some of them focus on a particular data type. Over a quarter of the programs focus on genomics, and approximately another quarter focus on one of imaging, clinical, or population data. Several courses or resources utilize multiple types of data types, one course is focused on utilization of neuronal data sets and several other courses and resources have an emphasis on the technical aspects of data science and do not focus specifically on a particular type of data.

Figure 10: Courses and Resources by Scientific Domain or Focus



Areas with few applications and hence few funded applications are the clinical and population sciences. Within these areas, this is notable dearth of activity in mHealth. This was noted in the reissuance of the R25 for open education resources (reissued in FY16). The new FOA included language encouraging applications that would educate participants on using, analyzing and/or integrating clinical or mobile health data among other topics. BD2K continues to work closely with relevant Institutes, Centers and Offices, such as NCATS and OBSSR, to ensure that new funding opportunity announcements contain language to encourage applications in all domain areas relevant to biomedical, behavioral, clinical, and social science research focused on health and that new opportunities are widely advertised to the appropriate communities.

- Data Management and Data Science: Collectively, the awards span a broad range of topics, including data management, data exploration, data representation, computing, data modeling, and data visualization. Each of these topics can be further broken down in the following way:
    - Data management: reuse of data, data standards, locating and accessing data and tools, organizing and curating data through ontologies and use of metadata
    - Computing: distributed or parallel computing, workflows, programming, algorithms, optimization, and natural language processing
    - Data representation: data structures and databases
    - Data exploration: data munging and preparation, exploratory data analysis
    - Data Modeling: probability, stochastic modeling, introductory statistics, advanced statistics (e.g. multiple testing, dimension reduction), machine learning, experimental design, Bayesian methods, reproducible research, network models
    - Data visualization and communication

Although collectively the R25 awards span the range above, they are often covered with little depth, and some of the topics are only covered by one Open Educational Resource, and these tend to be the more technical (e.g. optimization and stochastic modeling) or specialized [e.g. ethical, legal and social implications (ELSI) and team science] topics. Appendix C and Appendix D shows which of the topics given above are included in each of the R25 awards.

Reach:

- Eleven programs with in-person components are spread evenly between East coast, West coast, and the middle of the country. Based on the planned enrollments, the 11 programs will serve over 250 participants in the summer of 2016.
- The three programs that were funded in FY14 held courses in the summer of 2015. They collectively reached undergraduates, PhD students, and faculty from over 30 different universities.
- Demand for the 2015 in-person courses was high, with reported acceptance rates for the programs with limited slots as being 35% and 14%. Another program gave support to a limited number of students but opened a large auditorium for the course.
- Fourteen of the awards are online Open Educational Resources. These reach a large number of students and instructors, providing a great value per student.
- For example, about 5,000 students completed the first 8 courses in Rafa Irizarry’s series of biomedical Data Science MOOCs in the first offering,

amounting to about $40 per student based on an NIH investment of $200K (year 1 direct costs).

Although BD2K is supporting the development and discovery of training resources, continued support in the area is needed for a number of reasons: gaps in content coverage exist (e.g. methods for mHealth data, algorithms and optimization methods, advanced statistics, network models); demand for in-person courses continues to exceed supply; different ways of explaining material resonate with different learners; and materials need to be updated to take into account new science and new developments in the understanding of learning.

At the August 2016 meeting of the BD2K Multi-Council Working Group, Council members suggested that BD2K should focus on training and education not only for *current* scientists but also on *future* scientists – that is students in elementary, junior and high school. In response to this call, the BD2K training program management group reached out to the NIH Science Education Partnership Award (SEPA) Program. The SEPA Program funds innovative K-12 STEM and Informal Science Education (ISE) educational projects. SEPA projects create partnerships among biomedical and clinical researchers and K-12 teachers and schools, museums and science centers, media experts, and other educational organizations. SEPA K-12 resources target state and national K-12 standards for STEM teaching and learning and are rigorously evaluated for effectiveness. BD2K offered to fund one-year administrative supplements through the SEPA Program to: 1) Incorporate data science into curriculum that is aimed to increase student interest in pursuing biomedical research careers and increase understanding of the scientific research process; and/or 2) Hold professional development courses for K-12 teachers that enhance their knowledge of data science. For an interim report of these programs please see Appendix E.

Goal 1B: Ensure training opportunities and resources are more readily discovered and accessed

Training opportunities need to be available, and biomedical scientists need to find and access the ones that best fit their needs. To this end, BD2K supports the development of training opportunities and their dissemination to large numbers of learners, as well as infrastructure for discovering them.

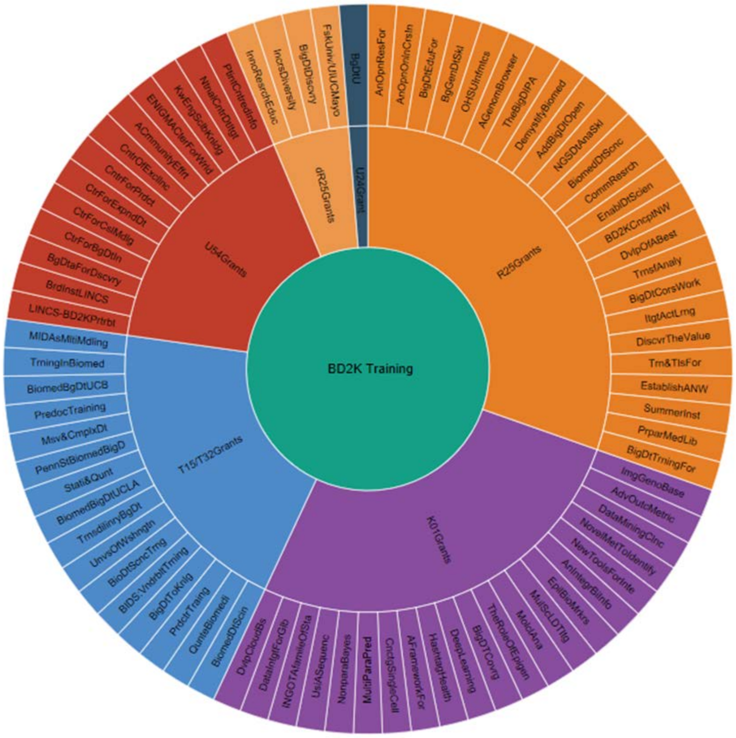
To help biomedical scientists find, access, and choose training opportunities, the Training Coordination Center has created an **E**ducational **R**esource **D**iscovery **I**ndex (**ER**u**DI**te). ERuDIte is a discovery index that organizes pointers to educational content, utilizing metadata describing the educational resource. Utilizing and extending common metadata is being pursued through an international collaboration between the TCC and ELIXIR, a European-based federation of organizations that build infrastructure for the life sciences. Please see link to access ERuDIte: here.

ERuDIte, when combined with a knowledge map that shows how Big Data skills relate to one another, may form the basis of a personalized learning system for biomedical scientists to efficiently acquire new skills to tackle Big Data.

To date, ERuDIte contains over 255 videos and educational resources that have been generated through the support of the BD2K Initiative. An additional 10,000 training resources have been collated from open resources collected across the internet including YouTube, Coursera, VideoLectures, Kahn Academy, and ELIXIR. Additional content is uploaded at regular intervals. It is expected that content from the International Neuroinformatics Coordinating Facility (ICNF – link for more information) will be uploaded soon as well as content from the H3Africa Initiative (link for more information).

For an interactive map of all BD2K Training activities with active links to the direct resources (Massive Online Open Content (MOOCS) or other online content), description of the in-person courses, predoctoral training programs and mentored career development awards please see the following sunburst: here.

Figure 11: Interactive sunburst clickable image map that can be used to explore awards supported through the BD2K Initiative. Produced by the BD2K Training Coordination Center.



Educational content may also come from other sources that the BD2K Initiative did not support. Some of the BD2K Centers, which each have a training component, are producing educational content such as TED-like talks. Curriculum for data science courses may come from BD2K T32/T15 training programs, which were given the opportunity to apply for $20K in funds to develop and share curriculum of new courses. The number of educational resources supported by BD2K, or even NIH, is dwarfed by the number supported elsewhere, whether by the National Science Foundation, the Department of Education, foundations, universities, or private industry. Educational resources from all of these sources has been included as content in ERuDIte.

Another opportunity that the TCC has developed is weekly Fundamental in Data Science webinars. This is a virtual lecture series on the data science underlying modern biomedical research. Beginning in September 2016, the webinar series consists of weekly presentations from experts across the country covering the basics of data management, representation, computation, statistical inference, data modeling, and other topics relevant to "big data" in biomedicine. The webinar series provides essential training suitable for individuals at an introductory overview level. All video presentations from the seminar series are streamed for live viewing, recorded, and posted online for future viewing and reference. These videos are also indexed within ERuDIte. This webinar series is a collaboration between the TCC, the NIH Office of the Associate Director for Data Science, and BD2K Centers Coordination Center (BD2KCCC). For more information see here.

Goal 1C: Enhance diversity in the biomedical and biomedical data science workforces

The dR25 programs are a main component of BD2K's diversity efforts. Four undergraduate programs aim to enhance diversity in the biomedical workforce through partnerships between the BD2K Centers of Excellence and low-resourced institutions. The partnerships support the development of curriculum and research experiences for undergraduates and faculty from low-resourced institutions. For example, in partnership with Harvard University and the University of Puerto Rico, Río Piedras (UPR-RP) the students at the UPR-RP take the Massive Online and Open Course (MOOC) developed by Rafael Irizarry from Harvard. This MOOC was one of the first Open Educational Resources funded by the BD2K Initiative. Collectively, these four programs reach 134 students directly over the course of 5 years. However, the number of students touched by the improved curriculum and the strengthened faculty is far greater. For example, with administrative supplement support provided to California State University at Fullerton (PI: AJ McEligot), an introductory video on big data science was developed with the goal of introducing students to big data and how this knowledge can be applied to provide solutions to critical questions in bioscience and health. The video has been incorporated into classroom instruction with most students agreeing or strongly agreeing that they understand better the respective topic being covered. (Video link: here.) This video, targeted towards undergraduate students in multiple departments has the potential to reach over 400 students per year. In addition, the PIs of these four programs met at the BD2K All Hands Meeting and have worked together to share experiences and resources. An accepted publication in Ethnicity and Disease on Enhancing Diversity in Biomedical Data Science demonstrates these collaborative interactions (Authors: JE Canner, AJ McEligot, M-E Perez, L Qian, X Zhang). This publication suggests that these four programs provide models for what various types of institutions might consider to enhance diversity in biomedical big data science.

In addition to the four programs funded explicitly for diversity, other R25s make serious efforts to recruit and train underrepresented minorities. For example, in the first offering, the short course from Oregon Health Sciences University trained 9 participants that were underrepresented in health-related research out of the 17 total participants. Although the R25 programs form the core of BD2K's diversity efforts, underrepresented participants in health-related research can be supported by BD2K through the T32/T15 training programs, which must have plans to recruit and retain a diverse pool of students.

Goal 2: To Increase the Number of Biomedical Data Scientists

To increase the number of biomedical data scientists, trainees need to gain the appropriate skills, want to work in biomedical science, and have an appropriate place to work. Because all of these trainees have self-selected toward biomedical science, BD2K's focus is on helping trainees get the appropriate skills initially and use those skills in the long run. Trainees may be 1) predoctoral students, who gain foundations in biomedical data science through T32/T15 training programs, 2) postdocs/faculty, who are trained in either biomedical science or data science and recognize the need to complement their existing knowledge and skills through K01 career development awards, or 3) students or faculty who need specialized training that is unavailable locally but attainable through a Research Rotation.

Retaining trainees is both important and a challenge, due to the demand for data science skills across sectors. Although retention is being addressed primarily through providing support, BD2K is also interested in facilitating conversations with university leaders to discuss career paths for open science and data science. Support for the further development of data scientists into biomedical data scientists is given through K01 awards. Before blending biomedical and data science knowledge, data scientists may be able to immediately contribute to addressing biomedical problems through working in teams with biomedical scientists. Such teams are being fostered through joint NSF/BD2K support in the QuBBD program, for building collaborations between data scientists and biomedical scientists.

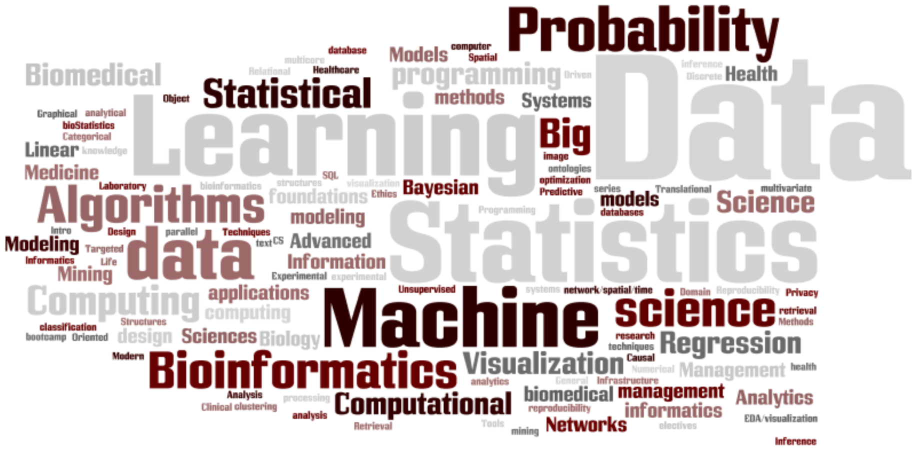
Goal 2A: Establish biomedical data science as a career path

T32/T15 Training Programs:

Predoctoral trainees are supported through new T32 programs and supplements to existing NLM T15 programs: 6 were funded in FY14, and another 10 were funded in FY15. Each training grant could request up to 6 predoctoral training slots per year, however, since many of the departments that the programs reside in are new, not all programs requested the maximum of 6 slots. To date, these training grants have supported 85 unique individuals. Though it is extremely early in the program, out of these 85 individuals, 2 have been successful at acquiring NIH predoctoral fellowship support through the Ruth L. Kirchstein National Research Service Award (NRSA).

To enhance the trainees interactions with each other, BD2K supported trainee networking events at the 2016 American Statistical Association Conference on Statistical Practice (link here) with approximately 35 students in attendance and at the annual BD2K All Hands Meeting of BD2K-funded investigators with approximately 20 students in attendance.

Because the field is nascent, there is still not full agreement as to what the core competencies are. However, some of the common areas of most BD2K training programs are: modern statistics (e.g. handling multiple comparisons, high dimensional data analysis, spatial-temporal correlation), computational techniques (e.g. using cloud and parallel computing, optimization, and algorithms), and machine learning. Most programs are creating new courses in order to integrate these topics together and focus class time on the ones that need to be learned didactically. Below is a word cloud of the core courses of the programs; a listing of these courses is given in Appendix F.

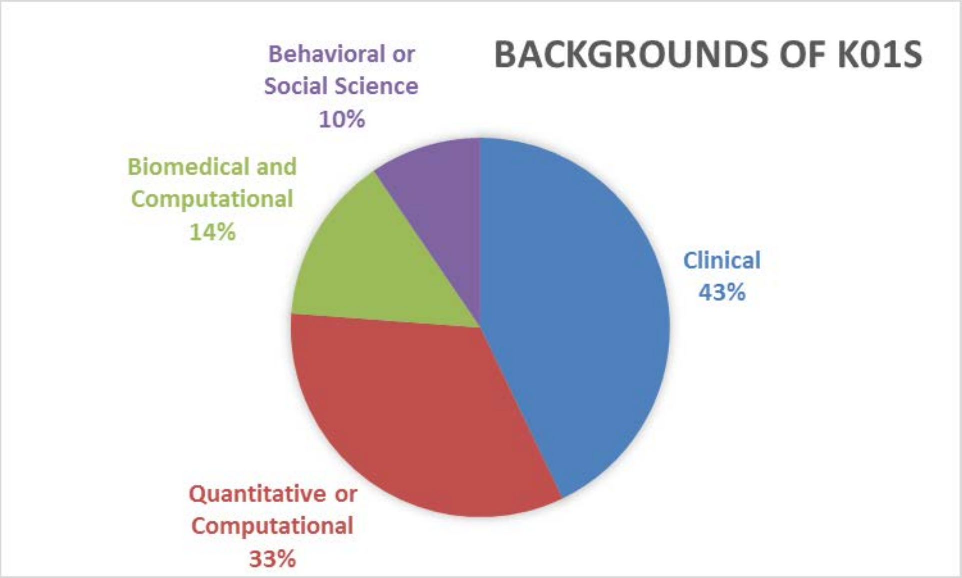


K01 Career Development:

BD2K is supporting 21 postdocs and faculty with mentored career development awards (K01). The PIs come from diverse backgrounds:

- 9 are physicians, with specialties in hematology/oncology, neurology, neuroradiology, surgery, urologic surgery, pulmonary and critical care medicine, and internal medicine
- 7 have primarily quantitative or computational backgrounds, with degrees in Electrical Engineering and Computer Science, Physics, Nuclear Physics, and Biomedical Engineering
- 3 have backgrounds in fields that blend the biomedical and computational sciences (molecular genetics, bioinformatics and computational biochemistry)
- 2 are behavioral or social scientists (Social Epidemiology, Quantitative Psychology)

Figure 12: Backgrounds of the Funded Mentored Career Development Awardees.



The group of K01 awardees is diverse not just by scientific background but also demographically and geographically:

- Nine out of 21 are female.
- They work at 18 unique institutions.

Through a sustained period of research career development and training, the K01 awardees will gain the knowledge and skills necessary to launch independent research careers. Awardees have at least two mentors, and the mentors were required to be active investigators in at least one of the areas involved in Big Data Science (computer science or informatics, statistics and mathematics, and biomedical sciences) and their experiences must encompass two of these areas and complement each other. The ultimate goal of this opportunity is to provide support for the awardee to become competitive for new research project grant (e.g., R01) funding in the area of Big Data Science.

To date, the K01 awardees have been extremely productive. Four have moved to different academic institutions for tenure track jobs or promotions. Two have moved from an academic institution into an industrial setting, continuing with their research project but without the support of the K01. Fifteen out of twenty-one (71%) are actively publishing on their projects with an average publication rate of 4 publications per person with a range of 1-11 publications. As of February 2017, two have received R01 funding from the NIH and another three have active R01 applications pending review.

Data Science Rotation for Advancing Discovery – RoAD-Trip Research Rotations

In addition to creating this index of educational resources, the TCC supports fellowship support to junior biomedical scientists wishing to acquire new knowledge and collaborations with more senior-level data scientists. This opportunity is called "Data Science Rotation for Advancing Discovery" (RoAD-Trip). The TCC has developed a match-making process that provides the opportunity for participants to conduct a focused joint project with an initial rotation of at least 2 weeks within a 5 month period where the participant can be immersed within the mentors lab. Most of the trainees are expected to be graduate students, but they could also be postdocs/junior faculty. This opportunity provides up to $4000 for the participant to help defray travel and living expenses and an honorarium to the mentor of $1000. In late 2016, 23 applicant fellows and 16 mentors applied for matching, with 10 applicant fellows and 8 mentors selected. The fellows career levels included masters level, postdoctoral and Assistant Professors with projects related to behavioral, clinical and biological big data questions.

Goal 2B: Foster collaborations between biomedical scientists and data scientists

NSF/NIH Quantitative Biomedical Big Data (QuBBD) Program:

Some problems will require a new model of leadership. Particularly when very diverse skills need to be brought to bear on the problem, teams of individuals with complementary expertise will be needed. Such teams are being fostered through a partnership between NSF and NIH called the QuBBD program. This program consists of a series of Innovation Labs and the funding of planning grants.

Innovation Labs are week-long mentored workshops that catalyze interdisciplinary teams and speed up the process of developing the team's research program. BD2K ran a pilot Innovation Lab in July 2015. By the end of the week, 12 new interdisciplinary teams were prepared to submit grant applications together. The teams submitted applications for small planning grants, along with newly-formed teams that did not go through the Innovation lab.

Because an established team can do much work virtually, the teams fostered by the QuBBD program draw in a wide range of talent, many of whom are from schools not otherwise represented within BD2K. These teams include some data scientists who are in physically isolated locations along with biomedical scientists who are unable to find data science collaborators due to the high demand.

The Summer 2015 Innovation Lab was successful. Evidence of this success is that the teams nurtured in the Innovation Lab had a much higher success rate than the other teams. The key to success was the participation of mentors who are from a variety of backgrounds and are leaders in their respective fields. The mentors guide teams through the iterative ideation process, offering feedback on research ideas throughout the week.

During FY16 the theme of the Innovation Lab involved addressing the data science needs arising from the use of wearable or ambient sensors to study health and disease. Thirty-one participants (13 women and 18 men) were invited to the 2016 Innovation Lab focusing on wearables and environmental sensors. The group represented an approximate split in biomedical and quantitative backgrounds (14 biomedical and 17 quantitative). After the week-long workshop to ideate and form new interdisciplinary teams with guidance from the five mentors who participated, 8 new teams formed around 8 new project ideas. During FY17 the theme of the Innovation Lab is on quantitative approaches to biomedical data science challenges in our understanding of the microbiome. Link to more information here.

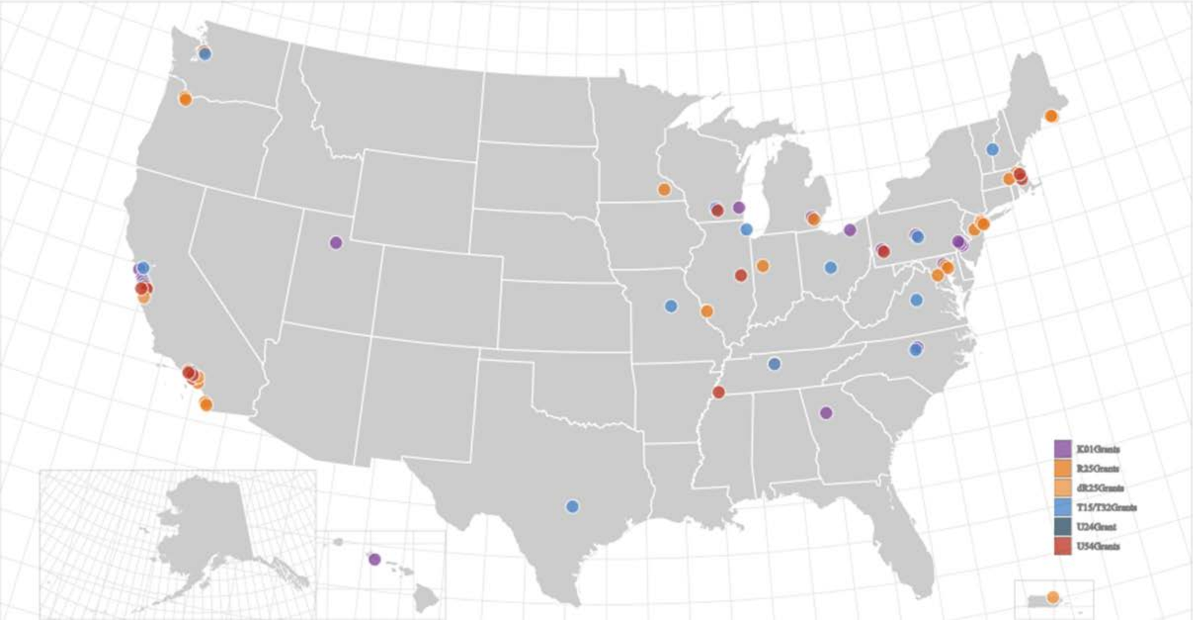
Summary

The BD2K programs in training described in this document are early efforts to respond to and address a need identified by the Advisory Committee to the Director's Data and Informatics Working Group. They cover both biomedical data science specialists, as well as specialists in other biomedical areas. They also cover the educational pipeline from undergraduates to faculty.

These programs are "early efforts" because there is still much work to be done to meet both goals. To quickly increase the base data science skills of a wide variety of biomedical scientists, early resources focused on broad, generally applicable topics. Later resources might be on more specialized topics. Likewise, to quickly jump start the increase in the number of biomedical data scientists, some of the training programs are modifications of existing programs, building on existing infrastructure and courses. As the community converges on the core competencies of the field, biomedical data science training programs will likely evolve and may end up bearing little resemblance to the programs initially funded.

The BD2K Initiative is in its infancy, and the training programs, along with other BD2K programs, are contributing to the development of the field of biomedical data science in the US and across the world. Although the awarded grants are confined to US institutions, the reach extends far beyond through the development of open educational resources and contributions to the global conversations about the field of biomedical data science. In addition, international collaborations surrounding the discovery of open educational resources for biomedical data science have begun.   The BD2K program aims to improve the ability of the biomedical workforce to use Big Data both today and tomorrow, both in the US and across the world.

Figure 13: Geographic Distribution of BD2K Training Awards



Appendix A: List of Funding Opportunity Announcements for the BD2K Initiative in the area of Training

FY Released	Title	Number / Hyperlink
FY16	Notice of NIH/BD2K Participation in the Joint NSF/NIH Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD)	NOT-EB-16-008
FY16	BD2K Research Education Curriculum Development: Data Science Overview for Biomedical Scientists (R25)	RFA-ES-16-011
FY16	NIH Big Data to Knowledge (BD2K) Enhancing Diversity in Biomedical Data Science (R25)	RFA-MD-16-002
FY16	BD2K Open Educational Resources for Skills Development in Biomedical Big Data Science (R25)	RFA-HG-16-016
FY16	NIH Big Data to Knowledge (BD2K) Mentored Career Development Award in Biomedical Big Data Science for Clinicians and Doctorally Prepared Scientists (K01)	RFA-ES-16-002
FY16	NIH Big Data to Knowledge (BD2K) Mentored Career Development Award in Biomedical Big Data Science for Intramural Investigators (K22)	RFA-ES-16-003
FY16	NIH Big Data to Knowledge (BD2K) Predoctoral Training in Biomedical Big Data Science (T32)	RFA-LM-16-002
FY15	NIH Big Data to Knowledge (BD2K) Enhancing Diversity in Biomedical Data Science (R25)	RFA-MD-15-005
FY15	NIH Big Data to Knowledge (BD2K) Biomedical Data Science Training Coordination Center (U24)	RFA-ES-15-004
FY15	NIH Big Data to Knowledge (BD2K) Initiative Research Education: Massive Open Online Course (MOOC) on Data Management for Biomedical Big Data (R25)	RFA-LM-15-001
FY15	NIH Big Data to Knowledge (BD2K) Initiative Research Education: Open Educational Resources for Sharing, Annotating and Curating Biomedical Big Data (R25)	RFA-LM-15-002
FY14	Predoctoral Training in Biomedical Big Data Science (T32)	RFA-HG-14-004

FY14	Revisions to Add Biomedical Big Data Training to Active Institutional Training Grants (T32)	RFA-HG-14-005
FY14	Revisions to Add Biomedical Big Data Training to Active NLM Institutional Training Grants in Biomedical Informatics (T15)	RFA-HG-14-006
FY14	Mentored Career Development Award in Biomedical Big Data Science for Clinicians and Doctorally Prepared Scientists (K01)	RFA-HG-14-007
FY14	Courses for Skills Development in Biomedical Big Data Science (R25)	RFA-HG-14-008
FY14	Open Educational Resources for Biomedical Big Data (R25)	RFA-HG-14-009

Appendix B: List of Funded Awards and Grants FY14-16

Type	Actv	Project	PI Name(s) All	Title	RFA/PA link	RePORTER Proj Info
1	U24	ES026465-01	VAN HORN, JOHN DARRELL	Big Data U: Empowering Modern Biomedicine via Personalized Training	RFA-ES-15-004	RePORTER Proj Info
3	U24	ES026465-01S2	VAN HORN, JOHN DARRELL	Innovation Labs Scoping Workshops	PA-14-077	RePORTER Proj Info
3	U24	ES026465-02S2	VAN HORN, JOHN DARRELL	BD2K TCC International Interactions and Frameworks Big Data Training Standards	PA-14-077	RePORTER Proj Info
3	U24	ES026465-01S1	VAN HORN, JOHN DARRELL	Innovation Labs: An Intensive Big Data Biomedicine Project Development Program	PA-14-077	RePORTER Proj Info
3	U24	ES026465-02S1	VAN HORN, JOHN DARRELL	Promoting Institutional Communities for Open Data Science	PA-14-077	RePORTER Proj Info
1	T32	LM012409-01	ALTMAN, RUSS BIAGIO	Biomedical Data Science Graduate Training at Stanford	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012204-01A1	AMOS, CHRISTOPHER I	Quantitative Biomedical Sciences at Dartmouth	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012414-01A1	DANIELS, MICHAEL J (contact); DHILLON, INDERJIT ; MEYERS, LAUREN ANCEL	PREDOCTORAL TRAINING IN BIOMEDICAL BIG DATA SCIENCE	RFA-HG-14-004	RePORTER Proj Info
1	T32	CA201159-01	KOSOROK, MICHAEL R (contact); FOREST, MARK GREGORY	Big Data to Knowledge Training Program	RFA-HG-14-004	RePORTER Proj Info
3	T32	LM012420-02S1	KOSOROK, MICHAEL R (contact); FOREST, MARK GREGORY	Big Data to Knowledge Training Program	PA-14-077	RePORTER Proj Info
1	T32	LM012412-01	MALIN, BRADLEY A (contact); BLUME, JEFFREY D; GADD, CYNTHIA S	BIDS: Vanderbilt Training Program in BIg Biomedical Data Science	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012413-01A1	NEWTON, MICHAEL A (contact); DEWEY, COLIN NOEL; GOULD, MICHAEL N	Bio-Data Science Training Program	RFA-HG-14-004	RePORTER Proj Info
1	T32	CA206089-01A1	NOBLE, WILLIAM STAFFORD (contact); DANIEL, THOMAS L; FAIRHALL, ADRIENNE L; WITTEN, DANIELA	University of Washington PhD Training in Big Data for Genomics and Neuroscience	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012416-01	PAPIN, JASON (contact); BROWN, DONALD E; LOUGHRAN, THOMAS PATRICK; SKADRON, KEVIN	Transdisciplinary Big Data Science Training at UVa	RFA-HG-14-004	RePORTER Proj Info
1	T32	CA201160-01	PELLEGRINI, MATTEO	Biomedical Big Data Training Grant	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012411-01A1	QUACKENBUSH, JOHN	Statistical and Quantitative Training in Big Data Health Science	RFA-HG-14-004	RePORTER Proj Info

1	T32	LM012415-01	RITCHIE, MARYLYN D (contact); HONAVAR, VASANT G; LI, RUNZE	Penn State Biomedical Big Data to Knowledge (B2D2K) Training Program	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012410-01	SHYU, CHI-REN	Massive and Complex Data Analytics Pre-Doctoral Training in One Health	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012203-01	STARREN, JUSTIN B (contact); KLABJAN, DIEGO	Predoctoral Training Program in Biomedical Data Driven Discovery (BD3)	RFA-HG-14-004	RePORTER Proj Info
1	T32	LM012417-01	VANDERLAAN, MARK J (contact); JORDAN, MICHAEL ; NIELSEN, RASMUS	Biomedical Big Data Training Program at UC Berkeley	RFA-HG-14-004	RePORTER Proj Info
3	T15	LM007079-23S1	HRIPCSAK, GEORGE M	Training in Biomedical Informatics at Columbia University	RFA-HG-14-006	RePORTER Proj Info
3	T15	LM011270-04S1	PAYNE, PHILIP R O (contact); CATALYUREK, UMIT V	MIDAs: Multi-modeling and Integrative Data Analytics Training Program	RFA-HG-14-006	RePORTER Proj Info
1	K01	ES025432-01	AVANTS, BRIAN	Imaging genomics bases of pediatric executive functioning	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026834-01	CALLCUT, RACHAEL A	Advancing Outcome Metrics in Trauma Surgery Through Utilization of Big Data	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026837-01	CHEN, JONATHAN HAILIN	Data-Mining Clinical Decision Support from Electronic Health Records	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025437-01	COFFMAN, DONNA LYNN	Novel Methods to Identify Momentary Risk States for Stress & Physical Inactivity	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026835-01	FARHAT, MAHA	New Tools for the interpretation of Pathogen Genomic Data with a focus on Mycobacterium tuberculosis	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025434-01	GARMIRE, LANA X	An Integrative Bioinformatics Approach to Study Single Cancer Cell Heterogeneity	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026839-01A1	GLISKE, STEPHEN V	Epileptic biomarkers and big data: identifying brain regions to resect in patients with refractory epilepsy	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026832-01	ITAKURA, HARUKA	Multi-scale data integration frameworks to improve cancer outcomes	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026838-01	JOHNSON, MICHAEL HIROSHI	Molecular Analysis and Precision Medicine in Renal Cell Carcinoma	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025431-01	LANDAU, DAN	The role of epigenetic heterogeneity in CLL evolution	RFA-HG-14-007	RePORTER Proj Info
3	K01	ES025431-04S1	LANDAU, DAN	The role of epigenetic heterogeneity in CLL evolution Admin supplement	PA-14-077	RePORTER Proj Info
1	K01	ES026841-01	LEE, GEORGE	Big data convergence of pathology and omics for disease prognosis	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025445-01A1	NEMATI, SHAMIM	Deep Learning and Streaming Analytics for Prediction of Adverse Events in the ICU	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025433-01	NGUYEN, QUYNH	HashtagHealth: A Social Media Big Data Resource for Neighborhood Effects Research	RFA-HG-14-007	RePORTER Proj Info

3	K01	ES025433-03S1	NGUYEN, QUYNH	HashtagHealth: A Social Media Big Data Resource for Neighborhood Effects Research	PA-14-077	RePORTER Proj Info
1	K01	ES025438-01	NSOESIE, ELAINE O	A Framework for Integrating Multiple Data Sources for Modeling and Forecasting of Infectious Diseases	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026842-01	PAGUIRIGAN, AMY	Connecting Single Cell Heterogeneity to Clinical Descriptors of Clonal Evolution in Acute Myeloid Leukemia	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026833-01	PARK, SOOJIN	Multiparametric Prediction of Vasospasm after Subarachnoid Hemorrhage	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025442-01	PEARSON, JOHN	Nonparametric Bayes Methods for Big Data in Neuroscience	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES025435-01	PROKOP, JEREMY WILLIAM	Using a Sequence-to-Structure-to-Function Approach to Functionally Characterize Protein Coding Missense Mutations in the Human and Rat Genomes	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026840-01	SCHMITT, JAMES E	INGOT: a family of statistical computing algorithms for hypothesis-driven imaging genomic and longitudinal neuroimaging analysis	RFA-HG-14-007	RePORTER Proj Info
1	K01	ES026836-01	VAN PANHUIS, WILLEM GIJSBERT	Data integration for global population health through dynamic models	RFA-HG-14-007	RePORTER Proj Info
3	K01	ES026836-02S1	VAN PANHUIS, WILLEM GIJSBERT	Data integration for global population health through dynamic models	PA-14-077	RePORTER Proj Info
1	K01	ES025436-01	WAGENAAR, JOOST B	Developing Cloud-based tools for Big Neural Data	RFA-HG-14-007	RePORTER Proj Info
1	R25	EB022365-01	CHUANG, JEFFREY HSU-MIN	Big Genomic Data Skills Training for Professors	RFA-HG-14-008	RePORTER Proj Info
3	R25	EB022365-02S1	CHUANG, JEFFREY HSU-MIN	Big Genomic Data Skills Training for Professors	PA-14-077	RePORTER Proj Info
1	R25	EB020379-01	DORR, DAVID A (contact); HAENDEL, MELISSA A; MCWEENEY, SHANNON K	OHSU Informatics Analytics BD2K Skill Course	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB022366-01	FOWLKES, CHARLESS (contact); DIGMAN, MICHELLE	The Big DIPA: Data Image Processing and Analysis	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB022364-01	HOFFMANN, ALEXANDER (contact); PAPP, JEANETTE CHRISTINE	NGS Data Analysis Skills for the Biosciences Pipeline	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB020393-01A1	KOVATCH, PATRICIA (contact); CLAUDIO, LUZ ; SHARP, ANDREW JAMES	Community Research Education and Engagement for Data Science (CREEDS)	RFA-HG-14-008	RePORTER Proj Info
3	R25	EB020393-02S1	KOVATCH, PATRICIA (contact); CLAUDIO, LUZ ; SHARP, ANDREW JAMES	Community Research Education and Engagement for Data Science (CREEDS)	PA-14-077	RePORTER Proj Info

1	R25	EB022363-01	MUKHERJEE, BHRAMAR (contact); JOHNSON, TIMOTHY D; MOZAFARI, BARZAN ; NGUYEN, LONG	Transforming Analytical Learning in the Era of Big Data	RFA-HG-14-008	RePORTER Proj Info
3	R25	EB022363-02S1	MUKHERJEE, BHRAMAR (contact); JOHNSON, TIMOTHY D	Administrative Supplement Request for Transforming Analytical Learning in the Era of Big Data	PA-14-077	RePORTER Proj Info
1	R25	EB023928-01	OWZAR, KOUROS (contact); CHAN, CLIBURN C	A hands-on, integrative next-generation sequencing course: design, experiment, and analysis	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB020381-01	PATHAK, JYOTISHMAN (contact); CHUTE, CHRISTOPHER G; NEUHAUSER, CLAUDIA M	Big Data Coursework for Computational Medicine	RFA-HG-14-008	RePORTER Proj Info
3	R25	EB020381-04S1	PATHAK, JYOTISHMAN (contact); CHUTE, CHRISTOPHER G	Big Data Coursework for Computational Medicine	PA-14-077	RePORTER Proj Info
1	R25	EB020389-01A1	RECHT, MICHAEL P (contact); ALIFERIS, CONSTANTIN F; BRAITHWAITE, RONALD SCOTT	Discovering the Value of Imaging: A Collaborative Training Program in Biomedical Big Data and Comparative Effectiveness Research for the Field of Radiology	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB023930-01	SAMORE, MATTHEW H	Curriculum in Biomedical Big Data: Skill Development and Hands-On Training	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB022367-01	SHAW, JOSEPH R	Establishing a Network of Skilled BD2K Practitioners: The Summer Workshop on Population-Scale Genomic Studies of Environmental Stress	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB020380-01	SHOJAIE, ALI (contact); WITTEN, DANIELA	Summer Institute for Statistics of Big Data	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB023929-01	VITEK, OLGA	Summer School: Big Data and Statistics for Bench Scientists	RFA-HG-14-008	RePORTER Proj Info
1	R25	EB022368-01	ZHANG, MIN	Big Data Training for Translational Omics Research	RFA-HG-14-008	RePORTER Proj Info
3	R25	EB022368-02S1	ZHANG, MIN	Administrative Supplement to: Big Data Training for Translational Omics Research	PA-14-077	RePORTER Proj Info
1	R25	GM114821-01	AMARO, ROMMIE E (contact); ALTINTAS DE CALLAFON, ILKAY	AN OPEN RESOURCE FOR COLLABORATIVE BIOMEDICAL BIG DATA TRAINING	RFA-HG-14-009	RePORTER Proj Info
1	R25	GM114827-01A1	BOHLAND, JASON W (contact); EDEN, URI TZVI; KRAMER, MARK ALAN	An open, online course in neuronal data analysis for the practicing neuroscientist	RFA-HG-14-009	RePORTER Proj Info
1	R25	EB020378-01	CAFFO, BRIAN SCOTT	Big Data education for the masses: MOOCs, modules, & intelligent tutoring systems	RFA-HG-14-009	RePORTER Proj Info
1	R25	GM123516-01	CHURCHILL, GARY A	Curriculum Development and Training for Systems Genetics	RFA-HG-14-009	RePORTER Proj Info
1	R25	GM119157-01	ELGIN, SARAH C R	A Genome Browser On-Ramp to Engage Biologists with Big Data	RFA-HG-14-009	RePORTER Proj Info

1	R25	GM114820-01	HERSH, WILLIAM R (contact); HAENDEL, MELISSA A; MCWEENEY, SHANNON K	Adding Big Data Open Educational Resources to the ONC Health IT Curriculum	RFA-HG-14-009	RePORTER Proj Info
1	R25	GM114818-01	IRIZARRY, RAFAEL ANGEL	Biomedical Data Science Online Curriculum on HarvardX	RFA-HG-14-009	RePORTER Proj Info
1	R25	GM114822-01	LEE, CHRISTOPHER	The BD2K Concept Network: open sharing of active-learning and tools online	RFA-HG-14-009	RePORTER Proj Info
3	R25	GM114822-03S1	LEE, CHRISTOPHER	The BD2K Concept Network: open sharing of active-learning and tools online	PA-14-077	RePORTER Proj Info
1	R25	GM114819-01	PEVZNER, PAVEL A	Integrated Active Learning Framework for Biomedical BD2K	RFA-HG-14-009	RePORTER Proj Info
3	R25	GM114819-03S1	PEVZNER, PAVEL A	Development of a new MOOC Programming for Biologists	PA-14-077	RePORTER Proj Info
1	R25	LM012285-01	HADDAD, BASSEM R (contact); GUSEV, YURIY ; MCGARVEY, PETER	Demystifying Biomedical Big Data: A User's Guide	RFA-LM-15-001	RePORTER Proj Info
3	R25	LM012285-01S1	HADDAD, BASSEM R (contact); GUSEV, YURIY ; MCGARVEY, PETER	Demystifying Biomedical Big Data: A User's Guide	PA-14-077	RePORTER Proj Info
1	R25	LM012284-01	MARTIN, ELAINE R	Development of a Best Practices in Research Data Management Massive Open Online Course (MOOC)	RFA-LM-15-001	RePORTER Proj Info
1	R25	LM012286-01	LAWSON, CATHERINE L	Enabling Data Science in Biology	RFA-LM-15-002	RePORTER Proj Info
3	R25	LM012286-01S1	LAWSON, CATHERINE L	Administrative Supplement to Enabling Data Science in Biology	PA-14-077	RePORTER Proj Info
1	R25	LM012288-01	SEYMOUR, ANNE (contact); LEHMANN, HAROLD P	Training & Tools for Informationists to Facilitate Sharing of Next Generation Sequencing Data	RFA-LM-15-002	RePORTER Proj Info
3	R25	LM012288-01S1	SEYMOUR, ANNE (contact); LEHMANN, HAROLD P	Training & tools for Informationists to facilitate sharing of Next Generation Sequencing data.	PA-14-077	RePORTER Proj Info
1	R25	LM012283-01	SURKIS, ALISA (contact); READ, KEVIN	Preparing Medical Librarians to Understand and Teach Research Data Management	RFA-LM-15-002	RePORTER Proj Info
1	R25	MD010391-01	CANNER, JUDITH ELENA	Innovative Research Education and Articulation in the Preparation of Under-represented and First Generation Students for Careers in Biomedical Big Data Science	RFA-MD-15-005	RePORTER Proj Info
1	R25	MD010399-01	GARCIA-ARRARAS, JOSE E (contact); ORDONEZ, PATRICIA ; PÉREZ, MARIA-EGLEE	Increasing Diversity in Interdisciplinary BD2K (IDI-BD2K)	RFA-MD-15-005	RePORTER Proj Info
1	R25	MD010397-01	MCELIGOT, ARCHANA J	Big Data Discovery and Diversity through Research Education Advancement and Partnerships (BD3-REAP)	RFA-MD-15-005	RePORTER Proj Info
3	R25	MD010397-02S1	MCELIGOT, ARCHANA J	Big Data Discovery and Diversity through Research Education	PA-14-077	RePORTER Proj Info

				Advancement and Partnerships (BD3-REAP)		
1	R25	MD010396-01	QIAN, LEI	Fisk University/UIUC-Mayo KnowENG BD2K Center R25 Partnership	RFA-MD-15-005	RePORTER Proj Info
3	R25	OD016519-03S1	HODGES, GEORGIA	Stimulating Young Scientists To Engage, Motivate and Synthesize (SYSTEMS)	PAR-10-206	RePORTER Proj Info
3	R25	OD016513-03S1	IMONDI, RALPH (contact); SANTSCHI, LINDA ANITA	NeuroLab M3: Discovery-based explorations of scientific models, model organisms,	PA-14-077	RePORTER Proj Info
3	R25	OD010494-03S1	MARKOWITZ, DINA GROSSMAN	Medicines and Me: Understanding and Using Medicines Safely	PAR-10-206	RePORTER Proj Info
3	R25	OD016511-03S1	MICKLOS, DAVID ANDREW	Barcode Long Island: Exploring Biodiversity in a Unique Urban Landscape	PAR-10-206	RePORTER Proj Info
3	R25	OD016542-03S1	WOOD, CHARLES ARTHUR	Pandem-Sim: Saving the World with Biology	PAR-10-206	RePORTER Proj Info
3	R25	OD016490-03S1	WYSS, J MICHAEL	Science Education Enabling Careers (SEEC)	PAR-10-206	RePORTER Proj Info

Appendix C: Topics in BD2K R25 Courses, by PI

Open Educational Resources (OER) – Online:

Amaro (modules)

- biomedical big data
- data access
- data standards / interoperability / ontologies
- workflows
- biostatistics

Caffo (two MOOC series)

- The human brain
- Imaging modalities and data types
- Quantitative neuroanatomy
- Neuroimaging pipelines
- fMRI
- Quantitative clinical neuroimaging
- Introduction to genomic technologies
- Sequence alignment and processing
- Sequence assembly
- Normalization and preprocessing
- Statistical analysis of genomic data

Hersh (curricular materials)

- Biomedical, Clinical, and Translational Research data life cycle
- Introduction to Big
- Ethical Issues
- Terminology
- Computing Concepts
- Clinical Data and Standards
- Basic Research Data Standards
- Public Health and Big Data
- Team Science
- Secondary Use (Reuse) of clinical data
- Limitations of Reuse of Clinical Data
- Information Retrieval
- Version control and identifiers
- Data annotation and curation
- Data and tools
- Ontologies 101
- Data modeling
- Data metadata and provenance
- Semantic data interoperability
- Semantic Web data
- Context-based selection of data
- Translating the Question
- Implications of Provenance
- Data tells a story
- Algorithms
- Basic statistics
- Visualization
- Reproducibility

Irizarry (MOOC series)

- Basic programming
- Python
- R
- Software engineering
- Inference
- Models
- Clustering and prediction
- Smoothing
- Algorithms
- Data Structures
- Large data
- Optimization
- Visualization
- Networks
- Communication

Lee

- N/A; this is a resource not an online course

Pevzner (MOOC)

- Algorithms
- Statistics
- Machine learning
- Data analysis

Bohland (15 MOOC modules)

- Introduction to MATLAB
- File and code management, metadata, web services / APIs, database systems, parallel computing approaches
- Visualization
- basic statistics
- Spectral analysis
- Cross-frequency coupling
- Dimensionality reduction
- point processes
- generalized linear modeling
- Interspike interval distributions
- Field-field coherence and spike-field coherence
- state-space modeling
- false discovery rate
- Linear discriminant analysis, support vector machines, cross-validation
- filtering
- network construction, correlation, partial correlation, network measures
- Heat maps, hierarchical clustering, multidimensional scaling, spectral graph partitioning, Mantel test

Elgin

- N/A; this is a resource, not an online course

Churchill (curricular materials)

- Data analysis
- Statistics
- Modeling

Courses – in person:

Dorr (4 days, undergrad)

- Overview of big data and its uses; teamwork in big data science
- Data types, formats, and definitions in research
- Finding and accessing datasets
- Data curation, ontologies, and metadata
- Managing data: introduction to data warehouses and tools
- Data wrangling and Methods 101: asking a simple question of data

Pathak (5 days, grad+)

- data and knowledge representation standards;
- information extraction and natural language processing;
- visualization analytics;
- data mining and predictive modeling;
- privacy and ethics; and
- applications in comparative effectiveness research and population health research and improvement.

Shojaie (13 days, grad+)

- Accessing Biomedical Big Data
- Data Visualization
- Supervised Methods for Statistical Machine Learning
- Unsupervised Methods for Statistical Machine Learning
- Reproducible Research for Biomedical Big Data

Chuang (5 days, instructors)

- Introduction to Biology Big Data Resources
- Genomics Data Processing
- Evaluation of Data Processing with Annotation Challenge
- Integration of Heterogeneous Genomic Data with Mutation Calling Challenge

Fowlkes (5 days, grad+)

- Introduction to BIG DATA: challenges and applications
- BIG DATA Image Acquisition (ex: uSPIM)
- Physics & Mathematical Frameworks
- Basics of Image Processing
- Machine learning frameworks
- Computer Vision
- Data Mining
- Data Visualization
- High-performance computing
- Dissemination – research access; data warehousing (archiving)

Hoffman (40 days, undergrad)

- Introduction to UNIX command-line
- Next-generation Sequencing Analyses: a Primer for Biologists
- Galaxy Platform for NGS Data Analysis
- Introduction to R and Bioconductor
- Python
- Short read mapping – QC, alignment to reference and quantification
- Informatics for RNA-sequence Analysis
- Statistical Genomics
- Systems Biology and Network Analysis Methods
- Analysis of ChIP-seq data
- Variant-Calling with GATK

Kovatch (10 days, grad students)

- Introduction to UNIX
- Introduction to Computing and Data

- Introduction to Scripting
- Introduction to Python
- Overview of the Human Genome and Genetic Variation
- Genome Technologies
- UCSC Genome Browser
- Galaxy and Galaxy Toolkit
- Introduction to Next Generation Sequencing
- Genomic pipeline tools
- Genomics in the Clinic
- Analysis of rare variant/exome datasets
- Isoform-level analysis of RNA-seq datasets

Mukherjee (20 days, 3 hours per day didactic, undergrads)

- Data Acquisition, Database Management
- Common computing platform, Linux environment
- Data Structure
- Data Visualization
- Probability and Statistical Inference
- Cloud, Parallel and Distributed Computing
- Optimization
- Sampling Methods: Markov chain Monte Carlo
- Medical Informatics/Computing
- Matrix Computation
- Bias and Confounding, Missing Data, Causal Inference with Electronic Health Records
- Machine Learning, Graphical Models, Sparse Learning with Matrices, Social Network Analysis, Imaging
- Case studies in Big Data Using Python
- Computational Genomics

Recht (10 days, researchers)

- Principles of Big Data Analytics
- Applications of Big Data Analytics
- Decision Analysis
- Cost Effective Analysis
- Evidence Synthesis

Shaw (7 days, researchers)

- Introduction to Population-Scale Genomics Studies of Environmental Stress.
- Environmental Genomics
- Philosophy of Genome Science

- Experimental Design
- Sequence Data Workflow
- Comparative Transcriptomics and Population Genomics
- Sequencing Technology, its Strengths and its Limitation
- De novo Genome Assembly and Analysis in a Time of High Throughput Genomics
- RNA-Seq Alignment to Individualized Genomes
- Statistical Considerations for Analyzing Genome-Scale Big Data
- Genomes as an Environmental Indicator
- Exploring Genome Sequence Variation
- Population Genomics (Why it Differs from Population Genetics?)
- Introduction to R
- Visualization of Sequence Data for Quality Assurance
- Visualizing Complex Data
- Navigating the Command Line
- RNA-Seq Data Analysis Workshop Using Tophat and R Bioconductor
- Gene Set Enrichment Analysis
- Biological Inference Using Pathway and Network Analysis
- Exploring Genome Sequence Variation

Zhang (10 days, researchers)

- Identifying public data resources
- Data processing
- Statistical issues
- Data visualization
- Functional analysis
- R programming
- Sequence management and visualization (Galaxy, UCSC browser)
- Comprehensive Packages (e.g. GenePattern)
- Pathway analysis
- Assessing Variant Impact on Proteins
- Computational Challenges (management, storage, and parallel computing)
- Bayesian statistics
- Ethical Issues
- Team Science

Samore (12 weeks summer program, undergraduates, graduates, medical students, basic and clinical faculty)

- Computation
- Representation
- Visualization
- Analysis

Owzar (6 week summer course, advanced undergraduates, graduates and postdoctoral fellows)

- Case studies used: Human Microbiome Project and Cancer Genome Atlas
- Analyzing next generation sequencing data

Vitek (2 week, graduate students and postdoctoral fellows)

- Mass spec and nuclear magnetic resonance
- Modular format
- R programming
- Process mass spec, metabolomic, and proteomic data
- Statistical methods for designing experiments
- Hands-on sessions

Appendix D: Summary of R25 Courses and Open Educational Resources

DATA SCENCE DOMAIN:					DATA MANAGEMENT			COMPUTING						DATA REPRESENTATION		DATA MINING / EXPLORATION	DATA MODELING								DATA VISUALIZATION	OTHER		
FOA	CONTACT PI	PROJECT TITLE	INSTITUTION	AUDIENCE	reuse of data	locating data/tools	organizing/curating data	computing environments	workflows	software engineering	algorithms	optimization	natural language processing	data structure	databases	data mining /exploration	probability	intro biostat	advanced statistics	machine learning	experimental design	bayesian methods	reproducible research	network models	data visualization	ethics	team science	privacy and clinical issues
HG14-009	AMARO	AN OPEN RESOURCE FOR COLLABORATIVE BIOMEDICAL BIG DATA TRAINING	UNIVERSITY OF CALIFORNIA SAN DIEGO	Graduate Students and Beyond	1	1	1	1	1							1		1					1					
HG14-009	BOHLAND	An open, online course in neuronal data analysis for the practicing neuroscientist	BOSTON UNIVERSITY (CHARLES RIVER CAMPUS)	Graduate Students and Beyond				1	1	1				1	1		1	1	1	1					1			
HG14-009	CAFFO	Big Data education for the masses: MOOCs, modules, & intelligent tutoring systems	JOHNS HOPKINS UNIVERSITY	Graduate Students and Beyond					1							1		1	1	1	1	1	1					
HG14-008	CHUANG	Big Genomic Data Skills Training for Professors	JACKSON LABORATORY	Instructors		1			1																			
HG14-009	CHURCHILL	Curriculum Development and Training for Systems Genetics	JACKSON LABORATORY	Instructors															1						1			
HG14-008	DORR	OHSU Informatics Analytics BD2K Skill Course	OREGON HEALTH & SCIENCE UNIVERSITY	Undergraduate Students		1	1							1		1											1	
HG14-009	ELGIN	A Genome Browser On-Ramp to Engage Biologists with Big Data	WASHINGTON UNIVERSITY	Instructors					1																			
HG14-008	FOWLKES	The Big DIPA: Data Image Processing and Analysis	UNIVERSITY OF CALIFORNIA-IRVINE	Graduate Students and Beyond		1	1	1												1					1			
LM15-001	HADDAD	Demystifying Biomedical Big Data: A User's Guide	GEORGETOWN UNIVERSITY	Graduate Students and Beyond	1	1	1																					
HG14-009	HERSH	Adding Big Data Open Educational Resources to the ONC Health IT Curriculum	OREGON HEALTH & SCIENCE UNIVERSITY	Graduate Students and Beyond	1	1	1		1							1		1					1		1	1	1	1
HG14-008	HOFFMANN	NGS Data Analysis Skills for the Biosciences Pipeline	UNIVERSITY OF CALIFORNIA LOS ANGELES	Undergraduate Students				1		1														1				
HG14-009	IRIZARRY	Biomedical Data Science Online Curriculum on HarvardX	HARVARD SCHOOL OF PUBLIC HEALTH	Graduate Students and Beyond				1		1	1	1		1		1		1	1	1	1	1	1		1			
HG14-008	KOVATCH	Community Research Education and Engagement for Data Science (CREEDS)	ICAHN SCHOOL OF MEDICINE AT MOUNT SINAI	Graduate Students and Beyond				1	1	1																		
LM15-002	LAWSON	Enabling Data Science in Biology	RUTGERS, THE STATE UNIV OF N.J.	Graduate Students and Beyond	1	1	1																					
HG14-009	LEE	The BD2K Concept Network: open sharing of active-learning and tools online	UNIVERSITY OF CALIFORNIA LOS ANGELES	Instructors																								
LM15-001	MARTIN	Development of a Best Practices in Research Data Management Massive Open Online Course (MOOC)	UNIV OF MASSACHUSETTS MED SCH WORCESTER	Graduate Students and Beyond	1	1	1																					
HG14-008	MUKHERJEE	Transforming Analytical Learning in the Era of Big Data	UNIVERSITY OF MICHIGAN	Undergraduate Students		1		1				1		1		1	1	1		1		1		1				
HG14-008	OWZAR	A hands-on, integrative next-generation sequencing course: design, experiment, and analysis	DUKE UNIVERSITY	Graduate Students and Beyond				1	1										1		1		1		1			
HG14-008	PATHAK	Big Data Coursework for Computational Medicine	MAYO CLINIC ROCHESTER	Graduate Students and Beyond	1								1	1						1					1	1		
HG14-009	PEVZNER	Integrated Active Learning Framework for Biomedical BD2K	UNIVERSITY OF CALIFORNIA SAN DIEGO	Graduate Students and Beyond							1							1		1								
HG14-008	RECHT	Discovering the Value of Imaging: A Collaborative Training Program in Biomedical Big Data and Comparative Effectiveness Research for the Field of Radiology	NEW YORK UNIVERSITY SCHOOL OF MEDICINE	Graduate Students and Beyond															1	1	1	1						
HG14-008	SAMORE	Curriculum in Biomedical Big Data: Skill Development and Hands-On Training	UNIVERSITY OF UTAH	Graduate Students and Beyond				1	1							1					1				1			
LM15-002	SEYMOUR	Training & Tools for Informationists to Facilitate Sharing of Next Generation Sequencing Data	JOHNS HOPKINS UNIVERSITY	Graduate Students and Beyond	1	1	1												1				1			1		
HG14-008	SHAW	Establishing a Network of Skilled BD2K Practitioners: The Summer Workshop on Population-Scale Genomic Studies of Environmental Stress	MOUNT DESERT ISLAND BIOLOGICAL LAB	Graduate Students and Beyond				1		1														1	1			
HG14-008	SHOJAIE	Summer Institute for Statistics of Big Data	UNIVERSITY OF WASHINGTON	Graduate Students and Beyond		1														1			1		1			
LM15-002	SURKIS	Preparing Medical Librarians to Understand and Teach Research Data Management	NEW YORK UNIVERSITY SCHOOL OF MEDICINE	Instructors	1	1	1																					
HG14-008	VITEK	Summer School: Big Data and Statistics for Bench Scientists	NORTHEASTERN UNIVERSITY	Graduate Students and Beyond										1				1	1		1		1					
HG14-008	ZHANG	Big Data Training for Translational Omics Research	PURDUE UNIVERSITY	Graduate Students and Beyond	1	1		1		1								1				1		1	1	1	1	
TOTAL NUMBER OF RESOURCES RESPONSIVE TO SPECIFIC DATA SCIENCE DOMAINS					**9**	**13**	**9**	**11**	**9**	**6**	**2**	**2**	**1**	**6**	**1**	**7**	**2**	**9**	**8**	**9**	**6**	**5**	**8**	**4**	**11**	**4**	**3**	**1**

Appendix E: Interim Report from the Science Education Partnership Award (SEPA) Program

Science Education Partnership Award (SEPA) BD2K Administrative Supplement Interim Report – March 2017

Grant	PI Name(s) All	Institution	Target Audience(s)	Deliverable(s)
3 R25 OD010494-03S1	MARKOWITZ, DINA GROSSMAN	UNIVERSITY OF ROCHESTER	High school students and teachers. Formal (classroom) and teacher professional development	Our "Big Data" curriculum module involves students in investigating how data (small data sets and "big data") is used to study the effectiveness of calcium dietary supplements in preventing osteoporosis. We anticipate that students (and teachers) will gain an understanding of the following concepts: (1) Data for a large research study could be acquired from a large scale health project that recruits and collects data from a large number of research subjects. (2) Data for a large research study could be acquired from meta-analysis that combines data from multiple smaller studies. (3) o Big data research typically involves "mining" data bases (sets) to look for trends, patterns, associations, or correlations among variables. (4) o The science of big data research is expanding to include new technologies to collect data and new ways to link existing data bases.
3 R25 OD016490-03S1	WYSS, J MICHAEL	UNIVERSITY OF ALABAMA AT BIRMINGHAM	Middle and High School teachers and students. The participants in SEEC Big Data include 25 teachers from the SEEC TPD training programs in Biology (BioTeach) and Chemistry (PhysChemTeach). The students involved include about 300 of the students of these teachers, most of whom will be in honors or Pre-AP/AP courses. The teachers and students are all from Birmingham City Schools (98% minority, 80% free or reduced lunch program, the largest school system in Alabama).	Grand Challenge Competitions in which classrooms will compete for top prizes for accuracy of data entry, creativity of hypotheses and final analysis for the data. Students will also be able to win prizes for their participation in the extraction of data and entering data in the Web Portal. The education results will be presented at the 2017 SEPA meeting and we expect to publish the obesity results in a subsequent research paper. The educational program will also be published after we have sufficient assessment data for a research paper.
3 R25 OD016511-03S1	MICKLOS, DAVID ANDREW	COLD SPRING HARBOR LABORATORY	High school teachers and students.	This supplement aims to provide teachers and their students with the tools and knowledge to participate in all aspects of the process, from study design, through sample collection and biochemistry, to data and metadata wrangling and analysis. The state of the art microbiome analysis workflow is QIIME. For the project a Jupyter notebook supporting QIIME (shared by Greg Caporaso, Northern Arizona University) was modified for high school use. The Jupyter notebook includes interactive tutorials that introduce key computing concepts and syntax while guiding students through the QIIME workflow. The notebook can be accessed and modified by users remotely from a dedicated Jupyter hub deployed at the DNALC.
3 R25 OD016513-03S1	IMONDI, RALPH (contact); SANTSCHI, LINDA ANITA	COASTAL MARINE BIOLABS	Secondary and post-secondary educators and students. Educational Setting: Formal (predominantly as part of research classes) and ISE (informal science education/outside the classroom)	Creation and deployment of a customized, student/classroom-focused student data interface to the Cell Image Library-Cell Centered Database (CIL-CCDB). Supported, in part, by the National Institute of General Medical Sciences (NIGMS), CIL-CCDB is a public repository of reviewed and annotated images, videos, and animations of cells obtained from a variety of organisms. The parent database is intended to advance research, education, and training, with the overarching goal of improving human health
3 R25 OD016519-03S1	HODGES, GEORGIA	UNIVERSITY OF GEORGIA	Elementary learners and teachers in grades three through five.	The Systems project team utilized the funding from BD2K to create a digital extension that integrates the concept of analysis and use of big data in an immersive learning environment that addresses diabetes and obesity in the US. In addition to creating the tool, our team developed a paper version of the digital extension that we have tested with third through fifth graders (n=64) to guide the development of the digital learning environment.
3 R25 OD016542-03S1	WOOD, CHARLES ARTHUR	WHEELING JESUIT UNIVERSITY	High school students, Grades 9-12, high school biology teachers, and high school computer science teachers Formal: Testing and use will take place in schools.	Pandem-Data deliverables will be disseminated on the Pandem-Data website (www. pandemsim.com/data) including: (1) A curriculum package of lesson and activities that explores and uses Big Data within the context of infectious disease. (2) User-friendly guides for building infectious disease models using Excel, FRED, and GLEAMviz modeling software. (3) A suite of classroom activities that incorporates Big Data and realistic models of disease spread to foster critical thinking skills for data science and infectious disease. (4) Teacher professional development materials to help teachers introduce Big Data concepts and implement simulation modeling activities with students. (5) Development of a case study using the developmental evaluation approach.

Appendix F: Core Courses in BD2K Training Programs, by Institution

In FY15, BD2K received 24 distinct applications for Biomedical Big Data training programs and 4 resubmissions. Of these, BD2K funded 6 applications (Columbia, Northwestern, Ohio State, UCLA, UNC, UW) that scored well in review and where each of the trainees in the program would satisfy requirements in the 3 areas. In FY16, we received 29 applications, all of which were new T32 programs. BD2K funded 10 of these applications, all of which scored well in review and were in line with the expectations set forth in the FOA.

Columbia

- Reproducibility
- Health data science (probability, network/spatial/time series modeling, clustering and classification)
- CS applications in Healthcare
- Algorithms
- Probability/Statistics
- Medicine
- Machine learning
- EDA/visualization

Dartmouth

- Bioinformatics
- Algorithms
- Probability and statistics
- Regression and multivariate statistics
- Epidemiology (experimental design)

Harvard

- Probability
- Statistics
- Methods (Linear Models and Categorical Data Analysis)
- Regression and Statistical Learning
- Data structures and Algorithms
- Health data science
- Computing foundations

Northwestern (core + selectives, competencies that can be fulfilled by a variety of courses)

- Core: Predictive Analytics, Data Mining, Advanced Computing

- Selectives: Domain knowledge, statistics, programming, ontologies, databases, text analytics
- Programming bootcamp

Ohio State University (in addition to the biomedical informatics core curriculum)

- Data Science methods in Biomedicine (data management) and 2 out of the following 3 courses
- Visualization and Machine Learning
- Advanced Modeling (HMM, spatial, network, graphical, dynamic)
- Advanced Computing (HPC)

Pennsylvania State University

- Machine Learning and Predictive Modeling
- Data Management
- Data Privacy
- Big Data Analytics
- Biomedical and Life Sciences
- Data Science electives

Stanford University (quarter system):

- Intro to Comp Bio and Bioinformatics
- Modeling Biomedical Systems
- Numerical Methods
- Statistical Inference
- Algorithms
- Object Oriented Systems Design
- Ethics
- Probability
- Data Driven Medicine
- Translational Bioinformatics
- Intro to BMI research methods
- Machine learning
- Discrete data analysis
- Unsupervised learning

University of California, Berkeley (in addition to home department requirements)

- R, Python, Parallel Computing
- Data Management
- Statistical Learning
- Data mining
- Statistical Computing
- Causal Inference

- Targeted Learning

University of California, Los Angeles (in addition to bioinformatics core curriculum)

- 4 courses in at least 2 of 3 different areas (computer science, biostatistics, informatics)
- 2 seminars (multicore computing, parallel programming, visualization)

University of Missouri (online modules from MS in Data Science prescribed to fill gaps in student knowledge in core competencies)

- Modern data analytical techniques
- Statistical analysis
- Visualization
- Relational database
- No SQL systems
- Information retrieval

University of North Carolina (in addition to home department's core curriculum)

- 6-wk modules with data science applications
- Additional courses in complementary departments

University of Texas Austin (courses in addition to home department requirements)

- Intro to Bio for Data Science (new)
- Tools and Techniques of Computational Sciences
- Statistical Models for Big Data (new)

University of Virginia

- Computational foundations
- Data Mining, Machine learning, Information Retrieval, Linear models, Bayesian methods, or statistics
- Networks, bioinformatics, image processing, or big data in health
- Experimental design and reproducibility

University of Washington

- Probability/Statistics
- Machine learning
- Data science (visualization, management, or a new general course)
- 3 application-area courses

University of Wisconsin (in addition to home department requirements)

- 2 statistics (e.g. statistical inference and Bayesian inference)
- 2 CS (e.g. bioinformatics and optimization)
- 2 biomedical (e.g. genomic science, biology of the mind)

Vanderbilt

- Clinical Informatics
- Bioinformatics
- Biomedical Data Science Laboratory
- Data Structures
- Algorithms
- Machine Learning
- Big Data Infrastructure
- Statistical computing
- Regression and Modeling
- Statistical Inference
- General biomedical science

Appendix G: BD2K Training Program Management Working Group

Team Members Include:

Richard Baird (NIBIB)
David Banks (NINR)
Regina Bures (NICHD)
Quan Chen (NIAID)
Sandra Colombini-Hatch (NHLBI)
Genevieve deAlmedia-Morris (NIDA)
Leslie Derr (OD)
Michelle Dunn (OD/ADDS) (Co-Chair)
Lisa Federer (OD/ORS)
Valerie Florance (NLM)
Bettie Graham (NHGRI) (Co-Chair)
Ming Lei (NCI)
Susan Lim (NCI)
Veerasamy Ravichandran (NIGMS)
Erica Rosemond (NCATS) (Co-Chair)
Cathrine Sasek (NIDA)
Carol Shreffler (NIEHS)
Erica Spotts (OD)
Jane Ye (NLM)
Xinzhi Zhang (NIMHD)