HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2017 May 10.

Published in final edited form as:

Pac Symp Biocomput. 2016 ; 22: 640–645.

The Training of Next Generation Data Scientists in Biomedicine[1]

Lana X Garmire[2,†], **Stephen Gliske**[3,†], **Quynh C Nguyen**[4,†], **Jonathan H. Chen**[5,†], **Shamim Nemati**[6,†], **John D. Van Horn**[7,†], **Jason H Moore**[8], **Carol Shreffler**[9], and **Michelle Dunn**[10]

[2]Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, 96813, USA.

[3]Department of Neurology, University of Michigan, Ann Arbor, MI 48109-5322, USA

[4]Department of Health, Kinesiology and Recreation, University of Utah, 84112, USA

[5]Department of Medicine, Stanford University, Stanford, CA, 94305, USA

[6]Department of Biomedical Informatics, Emory University, Atlanta, GA, 30322, USA

[7]Mark and Mary Stevens Neuroimaging and Informatics Institute, University of Southern California, Los Angeles, CA, 90032, US

[8]Institute of Biomedical Informatics, University of Pennsylvania, Philadelphia, PA, 19104, USA

[9]National Institute of Environmental Health Sciences, Research Triangle Park, NC, 27709, USA

[10]Office of the Associate Director for Data Science (ADDA), National Institute of Health, Bethesda, MD, 20892, USA

With the booming of new technologies, biomedical science has transformed into digitalized, data intensive science. Massive amount of data need to be analyzed and interpreted, demand a complete pipeline to train next generation data scientists. To meet this need, the transinstitutional Big Data to Knowledge (BD2K) Initiative has been implemented since 2014, complementing other NIH institutional efforts. In this report, we give an overview the BD2K K01 mentored scientist career awards, which have demonstrated early success. We address the specific trainings needed in representative data science areas, in order to make the next generation of data scientists in biomedicine.

1. Biomedical science as data intensive science

There is little doubt that biomedical science has become data intensive science. In the last decades, we have witnessed the booming of new biomedical technologies which generated massive amount of bio-data. In the genomics realm, next generation sequencing (NGS) has produced various types of omics-data. It is now a reality to sequence patients' genomes to

[1]This work is supported by BD2K K01 program

[†]Work partially supported by grant NIH Big Data 2 Knowledge Award K01ES025434 (to LXG), K01ES026839 (to SG), K01ES025433 (to QCN), K01ES026837 (to JHC), K01ES025445 (to SN), U24 ES026465 (to JDV), by the National Institute of Environmental Health Sciences through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative
[†]Work partially supported by grant NIH Big Data 2 Knowledge Award K01ES025434 (to LXG), K01ES026839 (to SG), K01ES025433 (to QCN), K01ES026837 (to JHC), K01ES025445 (to SN), U24 ES026465 (to JDV), by the National Institute of Environmental Health Sciences through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative, work is also partially supported by P20 COBRE GM103457 awarded by NIH/NIGMS, NICHD R01 HD084633, NLM R01 LM012373, and Hawaii Community Foundation Medical Research Grant 14ADVC-64566.

seek personalized medication. In the medical imaging field, petabytes of imaging data are stored, processed and analyzed in institutions[1]. Sensor-based wearable devices monitor daily exercise and other life-style routines, and generate real-time physiological data. With the adoption of Electronic Health Record (EHR) data by hospitals, it is now feasible to access and mine the massive amount of clinical and phenotypic data.

For junior researchers, the timing has never been better to seek a career in data science. Given the global “open-data” movement, many of the data types mentioned above are available publically, significantly saving the time and cost to conduct large-scale data mining and discoveries. We perceive that secondary data analysis would empower a whole new level of knowledge discoveries and hypothesis generation, which will reciprocally benefit other fields of biomedical research. Facility-wise, high-performance-computing (HPC) environments are well set-up in many major research universities; moreover, private sectors such as Google and Amazon offer cloud-computing as an alternative to the localized (thus restrictive) HPC access. Additionally, advancing in mathematical and statistical modeling, machine learning and the new derivatives of deep learning, is playing an increasingly important role in biomedical and healthcare industries.

2. The increasing needs to train the next generation data scientists in biomedicine

Compared to the prolific amount of biomedical data, developing computational methods and algorithms and training data scientists with domain expertise in biomedicine are major limiting factors to understanding the complex interactions in human health and disease. Unlike many other disciplines, data science in biomedicine is very interdisciplinary and requires training in domains including computer science, statistics, mathematics, and biomedicine. This interdisciplinary nature requires that data science in biomedicine be adaptive and involve constant learning and training by all, from undergraduate, graduate, postdoc to faculty levels.

Recognizing such needs, National Institute of Health (NIH) spearheaded The trans-NIH Big Data to Knowledge (BD2K) Initiative in 2014. The mission of the BD2K initiative is to support training in and research and development of innovative and transformative new approaches and tools, in order to maximize and accelerate the utility of the Big Data being generated. Since its inception, training has been one of the major thrust areas of the BD2K program. The term “training” is meant to include training, education, and workforce development that provides learners, no matter what career level, either foundational knowledge or skills for immediate use. Training currently accounts for 15% of the BD2K budget. There are two main goals for training: (1) to increase the number of people trained in developing the tools, methods, and technology to maximize the information which can be obtained by biomedical Big Data, and (2) to elevate the data science competencies of all biomedical scientists.

3. Funding mechanisms of National Institute of Health to train next generation data scientists

To accomplish these goals, a diverse set of grants and grants types have been developed (see the complete report: https://datascience.nih.gov/bd2k/funded-programs/enhancing-training). The work being showcased in this paper relates to the goal of increasing the number of biomedical data scientists. Although the establishment of biomedical data science as a career requires a complete career pipeline, from undergraduate training on up, the focus here is on the latter end of the pipeline, at the postdoc and junior faculty level. To support junior faculty, the NIH developed the K01 Career Development program. K01s in Biomedical Big Data Science are designed to facilitate the career transition of research oriented interdisciplinary investigators who are significantly altering their research focus. Candidates can enter the mentored experience from any of the three major scientific areas of Big Data Science: (1) Computer science or informatics; (2) Statistics and Mathematics; or (3) Biomedical Science. At the end of the program, awardees are expected to have competence in all three areas, as well as depth in one area. Competence is gained through course work as well as through a mentorship from a team that includes all of the expertise listed above. In 2014 and 2015, BD2K awarded 21 K01 projects. The PIs come from diverse backgrounds, including: (1) 9 physicians with specialties in hematology/oncology, neurology, neuroradiology, surgery, urologic surgery, pulmonary and critical care medicine, and internal medicine; (2) 7 PhDs with primarily quantitative or computational backgrounds, with degrees in Electrical Engineering and Computer Science, Physics, Nuclear Physics, and Biomedical Engineering; (3) 3 interdisciplinary scientists with backgrounds in fields that blend the biomedical and computational sciences (Molecular Genetics, Bioinformatics and Computational Biochemistry); and (4) 2 behavioral or social scientists (Social Epidemiology, Quantitative Psychology). These awardees represent 18 unique institutions from 12 states, among whom 9 awardees are female. .The expectation of the program is that the K01 awardees will be, by the end of the project period, competitive for new research grants (e.g. R01) in the area of Big Data Science. Many K01 awardees have moved on to faculty positions, and some have already obtained competitive NIH grants (e.g. R01s).

4. Areas of biomedical data science demanding new workforce

Data science in biomedicine includes, but is not limited to, the categories: translational bioinformatics and computational biology, clinical informatics, consumer health informatics and public health informatics[2]. Maximal success can be obtained by the biomedical data scientists trained in not only the technical aspects of data science (computer science, signal processing, math, statistics, etc.), but also the specific area of biomedicine of application. This, in part, sets apart the biomedical data scientists from general data scientists. Below we focus on a few representative categories funded by the current BD2K K01 program.

4.1. Translational Bioinformatics

Large national and international consortia and data repositories have formed, significantly increasing the sample sizes and discovery powers for many diseases. Training in translational bioinformatics needs to rapidly adapt to the global environment by emphasizing

broad, interdisciplinary training in computer science, statistics, bioinformatics, and biology. Good suggestions on bioinformatics training courses have been made earlier[3]. Here we put more focus on multi-omics areas, beyond single-omics data analysis and pipeline construction. At the input data level, the trainees will be expected to deal with missing values and normalizing data within and across various technical platforms. The trainees should be able to creatively transform data, by taking advantage of prior biological knowledge such as pathway or network information[4,5]. The trainees should have courses in statistics to thoroughly understand issues such as sample size, power, multiple hypothesis testing, classification (unsupervised learning), and generalized regression techniques (supervised learning)[6,7]. Training in multi-omics data integration (from the same population cohort) and meta-omics data integration (from heterogeneous populations) will be paramount to derive meaningful discoveries on molecular subtypes of diseases[8]. The trainees will also learn about omics-clinical/phenotypic data integration, using methods such as correlational and survival analysis.

Two new areas of translational bioinformatics are microbiome and single cell genomics. Both fields have measurement uncertainty. While the microbiome has the unknown variables of microbe numbers and strains, single cell genomics has the unknown variable of noise due to complicated batch effects, cell cycle and stress states, amplification biases etc[9]. In additional to the skills noted above, data visualization and tools to enhance reproducibility should be required for trainees, to enable efficient exploratory analysis and hypothesis generation. Last but not least, the trainees should go through rigorous training in HIPPA compliance to protect the private (including genetic) information of study subjects.

4.2. Clinical Informatics

The generation and dissemination of medical knowledge towards the practice of modern medicine arose in the past century, when there were relatively few effective interventions that the discipline had to offer for patient care. However, such norms now collide with the current reality of an explosive growth in biomedical knowledge[10]. Fortunately, with the new era of biomedical informatics, the clinicians are presented with great opportunities, along with challenges. The meaningful use of electronic health records (EHR)[11] presents the big data opportunity with the widespread routine capture of real-world clinical practice data, further augmented by high volume clinical data streams from claims, registry data, genomics, sensor systems, to patient generated content forms. Such digitized records offer new approaches to generate medical knowledge and to synthesize it into usable tools that can affect real-world clinical practice by assimilating and managing the increasing complexity of medical information. Principled, data-driven approaches are critical to unlocking the potential of large-scale healthcare data sources to impact clinical practice, compared to the otherwise limited and preconceived concepts manually abstracted out of patient chart reviews.

The current clinical practice force needs a paradigm shift. The next generation of data scientists will have the technical capability to generate useful insights from large complex data sources (machine learning, statistical analysis methods). They should have the tenacity to tackle enormous noisy and unstructured data that was not generated for precise research

purposes (data wrangling, software engineering). Training in the appreciation of the applied subject domain is particularly important, in order to transcend the data-information-knowledge-wisdom hierarchy (translational inquiry). For physician scientists, complementary knowledge is needed to bridge the evolving practice of medicine from one that is traditionally apprenticeship, heuristic, pattern based learning to the new approach of using big data analytics creatively to inform decision making. Meanwhile, clinician scientists will need to gain experience on meaningfully informing practice, including recognizing pitfalls and limitations of data science.

4.3. Public Health Informatics

The curriculum for public health graduate students typically includes classes on population health, research methods, ethics of scientific research, and applications in public health. Other courses covering research methods are usually on study design, data analyses, and causal inference. However, training is generally lacking on how to fully utilize larger and nontraditional data sources. Public health investigators are usually trained to implement and analyze health surveys and clinical trials. However, training on processing large unstructured text data is lacking. Clinical text is the most pervasive data type in EHR[11]. Leveraging techniques in data mining, machine learning and natural language processing will enable the extraction of information on patient characteristics and clinical outcomes. Mining EHR allows us to better understand longitudinal patterns in treatment outcomes[12], treatment heterogeneity, and drug interactions. In addition, social media text has been useful for outbreak detection, tracking health conditions, and monitoring social influences on human health[13,14]. New public health training with data science concentration may include additional course in computer science, including database systems, data mining, machine learning, advanced algorithms, and visualization. More specialized training in natural language processing, image processing, high performance computing, and network security would be beneficial, too. These courses would increase expertise in the creation and maintenance of database structures for efficient storage and processing, and also increase the incorporation of large, emerging data sources such as text, images and videos in health research. The addition of training in database management and analytics would further enhance the understanding of drivers of health and disease, by incorporating novel and integrated data sources to account for disease complexities.

4.4 Exemplary Emerging Area of Informatics

In neuroscience, one area in need of data scientists that is only beginning to be recognized involves electroencephalogram (EEG). For example, the US Brain initiative funded many projects focused on acquiring high resolution EEG data, yet little attention has been focused ensuring that there is a sufficiently trained work force to analyze such data. Even with current technology, there is great need for more data scientists related to EEG analysis, both intracranial EEG (e.g., in epilepsy research)[15] and extracranial EEG (e.g., sleep medicine). The training needs for these individuals are similar to other fields: fluent programming skills, a strong understanding of machine learning, statistics and applied mathematics, and an understanding of the application of focus. One training method that has worked quite well for this applications is for students to get a PhD in either a technical field (e.g., biomedical engineering) or an applied field (e.g., neuroscience), and augment their coursework in order

to obtain the needed breadth of subject matter. Some universities, such as the University of Michigan, has created a graduate certificate in data science, which can be paired with a PhD in specific discipline. Additionally, coursework needs to be matched with appropriate “hands on” research activities at the graduate and post-doctoral levels. One main challenge facing the next generation of data scientists is to establish the culture of interactions between disciplines. In addition to the challenges common to upcoming biomedical data scientists, these students face the extra barrier of EEG analysis being an emergent application area.

5. Conclusion

The golden era of big data science in biomedicine has just begun[2]. Many fields, such as EMR mining, mobile health and community-based health data mining are very new, and clearly challenges exist. However, the data volume will only increase, thus “more is more, less is bore”. The need for data scientists specialized in bio-medicine will continue to drive the market. On the other hand, while the paradigm shift towards data intensive biomedical science is happening, we must also bring to the attention that the “brain drain” from academia to private sectors is likely, and it is critical for institutions to create tenure-track career paths for the new generation of biomedical data scientists after their training programs end.

Acknowledgments

We would like to thank all BD2K K01 awardees for their support to make this workshop a reality.

References

1. Van Horn JD. Opinion: Big data biomedicine offers big higher education opportunities. Proc Natl Acad Sci U S A. 2016 Jun 7; 113(23):6322–6324. [PubMed: 27274038]
2. Moore JH, Holmes JH. The golden era of biomedical informatics has begun. BioData Min. 2016; 9:15. [PubMed: 27069509]
3. Greene AC, Giffin KA, Greene CS, Moore JH. Adapting bioinformatics curricula for big data. Brief Bioinform. 2016 Jan; 17(1):43–50. [PubMed: 25829469]
4. Huang S, Yee C, Ching T, Yu H, Garmire LX. A novel model to combine clinical and pathway-based transcriptomic information for the prognosis prediction of breast cancer. PLoS Comput Biol. 2014; 10(9):e1003851. [PubMed: 25233347]
5. Huang S, Chong N, Lewis NE, Jia W, Xie G, Garmire LX. Novel personalized pathway-based metabolomics models reveal key metabolic pathways for breast cancer diagnosis. Genome Med. 2016; 8(1):34. [PubMed: 27036109]
6. Ching T, Huang S, Garmire LX. Power analysis and sample size estimation for RNA-Seq differential expression. RNA New York N. 2014 Sep 22.
7. Menor M, Ching T, Zhu X, Garmire D, Garmire LX. mirMark: a site-level and UTR-level classifier for miRNA target prediction. Genome Biol. 2014; 15(10):500. [PubMed: 25344330]
8. Wei R, De Vivo I, Huang S, Zhu X, Risch H, Moore JH, Yu H, Garmire LX. Meta-dimensional data integration identifies critical pathways for susceptibility, tumorigenesis and progression of endometrial cancer. Oncotarget. 2016 Jul.:9.
9. Poirion OB, Zhu X, Ching T, Garmire L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front Genet. 2016; 7:163. [PubMed: 27708664]
10. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? PLoS Med. 2010 Sep.7(9):e1000326. [PubMed: 20877712]

11. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun; 13(6):395–405. [PubMed: 22549152]

12. Jensen AB, Moseley PL, Oprea TI, Ellesøe SG, Eriksson R, Schmock H, Jensen PB, Jensen LJ, Brunak S. Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. Nat Commun. 2014; 5:4022. [PubMed: 24959948]

13. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011 May; 40(5 Suppl 2):S154–158. [PubMed: 21521589]

14. Nguyen QC, Kath S, Meng H-W, Li D, Smith KR, VanDerslice JA, Wen M, Li F. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. Applied Geography. 2016; 73:77–88.

15. Gliske SV, Stacey WC, Lim E, Holman KA, Fink CG. Emergence of Narrowband High Frequency Oscillations from Asynchronous, Uncoupled Neural Firing. Int J Neural Syst. 2016 Jul. 14:1650049.