HARNESSING BIG DATA FOR PRECISION MEDICINE: INFRASTRUCTURES AND APPLICATIONS

KUN-HSING YU
Biomedical Informatics Training Program, Stanford University
3165 Porter Dr., Room 2270, Palo Alto, CA 94304
Email: khyu@stanford.edu

STEVEN N. HART
Center for Individualized Medicine, Mayo Clinic
200 First Street SW, Rochester, MN 55905
Email: Hart.Steven@mayo.edu

RACHEL GOLDFEDER
Biomedical Informatics Training Program, Stanford University
870 Quarry Rd, Stanford, CA 94305
Email: rlg2@stanford.edu

QIANGFENG CLIFF ZHANG
School of Life Sciences, Tsinghua University
Medical Science Building, B-1002, Tsinghua University, Beijing, China 100084
Email: zhang.lab@biomed.tsinghua.edu.cn

STEPHEN C. J. PARKER
Computational Medicine and Bioinformatics, University of Michigan
100 Washtenaw Ave, 2049B, Ann Arbor, MI 48109
Email: scjp@umich.edu

MICHAEL SNYDER
Department of Genetics, Stanford University
300 Pasteur Dr., M344 MC 5120, Stanford, CA 94305
Email: mpsnyder@stanford.edu

Precision medicine is a health management approach that accounts for individual differences in genetic backgrounds and environmental exposures. With the recent advancements in high-throughput omics profiling technologies, collections of large study cohorts, and the developments of data mining algorithms, big data in biomedicine is expected to provide novel insights into health and disease states, which can be translated into personalized disease prevention and treatment plans. However, petabytes of biomedical data generated by multiple measurement modalities poses a significant challenge for data analysis, integration, storage, and result interpretation. In addition, patient privacy preservation, coordination between participating medical centers and data analysis working groups, as well as discrepancies in data sharing policies remain important topics of discussion. In this workshop, we invite experts in omics integration, biobank research, and data management to share their perspectives on leveraging big data to enable precision medicine.
Workshop website: http://tinyurl.com/PSB17BigData; HashTag: #PSB17BigData.

1. Introduction

Throughout medicine's history, disease prevention and treatment has been based on the expected outcome of an average patient[1]. Data from patients with the same disease were often pooled together for statistical analysis, and clinical guidelines derived from the aggregated analysis informed health and disease management for billions of patients. Although this approach achieves some success, it ignores important individual differences, which can result in different treatment responses[2].

Precision medicine aims to tailor clinical treatment plans to individual patients, with the goal of delivering the right treatments at the right time to the right patient[3]. Recent advances in omics technologies provide clinicians with more complete patient profiles[4,5]. The decreasing cost of sequencing and associated data storage[6] and the development of effective data analysis methods make it possible to collect and analyze big biomedical data for various human diseases at an unprecedented scale[7]. These advancements can improve the diagnostic accuracy of complex diseases, identify patients who will benefit from targeted therapeutics, and predict diseases before their occurrence[3,8].

Nevertheless, many challenges still remain. Conventional methods for data storage, database management, and computational analysis are insufficient for the petabytes of biomedical data generated every year. In addition, as datasets become larger and more diverse, advanced distributed file storage and computing methods are needed to make the data useful. Furthermore, data-sharing policies and result reproducibility continue to be vigorously debated issues[9-10].

In this workshop, world-renowned experts in personal omics profiling, biobanks, biomedical databases, and medical data analysis will describe recent advancements in these areas and discuss associated challenges and potential solutions.

2. Workshop presentations

This section provides a brief summary for each presentation. The full abstracts could be found at the workshop website http://tinyurl.com/PSB17BigData.

2.1. DeepDive: A Dark Data System

Dr. Christopher Ré, Department of Computer Science, Stanford University, CA, USA

Many pressing questions in science are macroscopic, as they require scientists to integrate information from numerous data sources, often expressed in natural languages or in graphics; these forms of media are fraught with imprecision and ambiguity and so are difficult for machines to understand. Here I describe DeepDive, which is a new type of system designed to cope with these problems. It combines extraction, integration and prediction into one system. For some paleobiology and materials science tasks, DeepDive-based systems have surpassed human volunteers in data quantity and quality (recall and precision). DeepDive is also used by scientists in areas including genomics and drug repurposing, by a number of companies involved in various

forms of search, and by law enforcement in the fight against human trafficking. DeepDive does not allow users to write algorithms; instead, it asks them to write only features. A key technical challenge is scaling up the resulting inference and learning engine, and I will describe our line of work in computing without using traditional synchronization methods including Hogwild! and DimmWitted. DeepDive is open source on github and available from DeepDive.Stanford.Edu.

2.2 Results of the VariantDB Challenge

Dr. Steven Hart, Department of Health Sciences Research, Mayo College of Medicine, MN, USA

The current standard formats for storing genomics data is the VCF and gVCF, but manipulating these large files is an imperfect and impractical long-term solution. Scalability, availability, consistency, are all important drawbacks to the file-based approach. Multiple pieces of metadata are often required to interpret genomic data, but there is no specification for how to tie sample level data (e.g. smoking status, disease status, age of onset, etc.) with variant-level data. The motive of the VariantDB Challenge is to identify a scalable, robust framework for storing, querying and analyzing genomics data in a biologically relevant context. The contextual focus is a central theme in the challenge since it is relatively easy to optimize simple database lookups, but forming queries with multiple predicates becomes a much more complicated task. The VariantDB_Challenge is a 100% open source project, meaning that all code and solutions used must be made publically available via GitHub. In this session, we will present an overview of the challenge and summarize the results from all submitters.

2.3. ADAM: Fast, Scalable Genome Analysis

Mr. Frank Austin Nothaft, Department of Computer Science, UC Berkeley, Berkeley, CA, USA

The detection and analysis of rare genomic events requires integrative analysis across large cohorts with terabytes to petabytes of genomic data. Contemporary genomic analysis tools have not been designed for this scale of data-intensive computing. This talk presents ADAM, an Apache 2 licensed library built on top of the popular Apache Spark distributed computing framework. ADAM is designed to allow genomic analyses to be seamlessly distributed across large clusters, and presents a clean API for writing parallel genomic analysis algorithms. In this talk, we'll look at how we've used ADAM to achieve a 3.5× improvement in end-to-end variant calling latency and a 66% cost improvement over current toolkits, without sacrificing accuracy. We will also talk about using ADAM alongside Apache Hbase to interactively explore large variant datasets.

2.4. Personalized Medicine: Using Omics Profiling and Big Data to Understand and Manage Health and Disease

Dr. Michael Snyder, Department of Genetics, Stanford University School of Medicine, CA, USA

Understanding health and disease requires a detailed analysis of both our DNA and the molecular events that determine human physiology. We performed an integrated Personal Omics Profiling (iPOP) on 70 healthy and prediabetic human subjects over periods of viral infection as well as during controlled weight gain and loss. Our iPOP integrates multiomics information from the host (genomics, epigenomics, transcriptomics, proteomics and metabolomics) and from the gut microbiome. Longitudinal multiomics profiling reveals extensive dynamic biomolecular changes occur during times of perturbation, and the different perturbations have distinct effects on different biomolecules in terms of the levels and duration of changes that occur. Overall, our results demonstrate a global and system-wide level of biochemical and cellular changes occur during environmental exposures.

2.5. Statistical and Dynamical Systems Modeling of Real-Time Adaptive m-Intervention for Pain

Dr. Jingyi Jessica Li, Departments of Statistics and Human Genetics, University of California, Los Angeles, CA, USA

Nearly a quarter of visits to the Emergency Department are for conditions that could have been managed via outpatient treatment; improvements that allow patients to quickly recognize and receive appropriate treatment are crucial. The growing popularity of mobile technology creates new opportunities for real-time adaptive medical intervention, and the simultaneous growth of "big data" sources allows for preparation of personalized recommendations. We present a new mathematical model for the dynamics of subjective pain that consists of a dynamical systems approach using differential equations to forecast future pain levels, as well as a statistical approach tying system parameters to patient data (both personal characteristics and medication response history). We combine this with a new control and optimization strategy to ultimately make optimized, continuously-updated treatment plans balancing competing demands of pain reduction and medication minimization. A workable hybrid model incorporating both mathematical approaches has been developed. Pilot testing of the new mathematical approach suggests that there is significant potential for (1) quantification of current treatment effectiveness for pain management, (2) forecast of pain crisis events, and (3) overall reduction of pain without increased medication use. Further research is needed to demonstrate the effectiveness of the new approach for each of these purposes.

2.6. Integrated Database and Knowledge Base for Genomic Prospective Cohort Study: Lessons Learned from the Tohoku Medical Megabank Project

Dr. Soichi Ogishima, Tohoku Medical Megabank Organization, Tohoku University, Japan

The Tohoku Medical Megabank project is a national project to revitalize medical care and to realize personalized medicine in the disaster area of the Great East Japan Earthquake. In our prospective cohort study, we recruited 150,000 people at Tohoku University, its satellites health clinics, and Iwate Medical University. We collected biospecimen, questionnaire, and physical

measurement during baseline and follow-up investigations. Along with prospective genome-cohort studies, we have developed integrated database and knowledge base, which will be the foundation for realizing personalized medicine and disease prevention.

3. Conclusion

Big data in biomedicine presents a great opportunity to understand health and disease states at an unprecedented level. This workshop will highlight landmark achievements in integrative omics studies, biobank research, and novel data mining methods for large datasets. With the growing number and size of biomedical datasets worldwide, we envision that approaches discussed in this workshop will facilitate the development of precision medicine.

4. Acknowledgments

K.-H. Y. is supported by a Howard Hughes Medical Institute (HHMI) International Student Research Fellowship and a Winston Chen Stanford Graduate Fellowship. R.G. is supported by a National Science Foundation (NSF) Graduate Research Fellowship. M.P. is partially supported by National Institutes of Health grants 1U54DE02378901, 5P50HG00773502, and 5U24CA16003605.

5. References

1. Collins FS. Exceptional opportunities in medical science: a view from the National Institutes of Health. *JAMA*. **313**:131-2 (2015).
2. Shastry BS. Pharmacogenetics and the concept of individualized medicine. *Pharmacogenomics J.* **6**:16-21 (2006).
3. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med.* **372**:793-5 (2015).
4. Chen R, Mias GI, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell.* **148**:1293-307 (2012).
5. Yu KH, Snyder M. Omics Profiling in Precision Oncology. *Mol Cell Proteomics.* **15**:2525-36 (2016).
6. Stein LD. The case for cloud computing in genome informatics. *Genome Biol.* **11**:207 (2010).
7. Wichmann HE, Kuhn KA, Waldenberger M, et al. Comprehensive catalog of European biobanks. *Nat Biotechnol.* **29**:795-7 (2011).
8. Ashley EA. The precision medicine initiative: a new national effort. *JAMA.* **313**:2119-20 (2015).
9. Longo DL, Drazen JM. Data Sharing. *N Engl J Med.* **374**:276-7 (2016).
10. Ioannidis JP. Expectations, validity, and reality in omics. *J Clin Epidemiol.* **63**:945-9 (2010).

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], MICHELLE DUNN[10]

[2]*Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI, 96813, USA. Email: LGarmire@Hawaii.edu*
[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 trans-institutional 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

[1*] This work is supported by BD2K K01 program

2†3†4†5†6†7† 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,

2† 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.

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 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.

6. Acknowledgement

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. PMID: 27274038
2. Moore JH, Holmes JH. The golden era of biomedical informatics has begun. BioData Min. 2016;9:15. PMID: 27069509
3. Greene AC, Giffin KA, Greene CS, Moore JH. Adapting bioinformatics curricula for big data. Brief Bioinform. 2016 Jan;17(1):43–50. PMID: 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. PMID: 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. PMID: 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; PMID: 25246651
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. PMID: 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; PMID: 27409342
9. Poirion OB, Zhu X, Ching T, Garmire L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front Genet. 2016;7:163. PMID: 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. PMID: 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. PMID: 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. PMID: 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. PMID: 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. PMID: 27712456

NO-BOUNDARY THINKING IN BIOINFORMATICS

JASON H. MOORE
Institute for Biomedical Informatics
University of Pennsylvania
Philadelphia, PA 19104, USA
Email: jhmoore@upenn.edu

STEVEN F. JENNINGS
Sector3 Informatics
Marana, AZ 85658, USA
Email: drsfjennings@gmail.com

CASEY S. GREENE
Department of Systems Pharmacology and Translational Therapeutics
University of Pennsylvania
Philadelphia, PA 19104, USA
Email: csgreene@upenn.edu

LAWRENCE E. HUNTER
Computational Bioscience Program,
University of Colorado School of Medicine
Aurora, CO 80045, USA
Email: larry.hunter@ucdenver.edu

ANDY D. PERKINS
Department of Computer Science and Engineering,
Mississippi State University
Jackson, MS 39762, USA
Email: perkins@cse.msstate.edu

CLARLYNDA WILLIAMS-DEVANE
Department of Biological and Biomedical Sciences,
North Carolina Central University
Durham, NC 27707, USA
Email: clarlynda.williams@nccu.edu

DONALD C. WUNSCH
Department of Electrical and Computer Engineering
Missouri Science and Technology University
Rolla, MO 65409, USA
Email: dwunsch@mst.edu

ZHONGMING ZHAO
Center for Precision Health
University of Texas Health Science Center
Houston, TX 77030, USA
Email: zhongming.zhao@uth.tmc.edu

XIUZHEN HUANG
Department of Computer Science
Arkansas State University
Jonesboro, AR 72467, USA
Email: xhuang@astate.edu

1. Bioinformatics is a Mature Discipline

Bioinformatics had its origins in the 1970s with the convergence of DNA sequencing, personal computers, and the internet. The field rapidly evolved as biotechnology improved making it critical to store, process, retrieve, and analyze bigger and bigger data to address important questions in the biological and biomedical sciences. Bioinformaticians throughout the 1980s and 1990s were often seen as consultants that provided a data service that represented one step in the process of asking a question, formulating a hypothesis, carrying out an experiment, analyzing the results, and making an inference. Much of bioinformatics at that time was about developing the capacity for providing this service. As the discipline matured in the 2000s it quickly became apparent that bioinformaticians were needed as collaborators and not just consultants. This facilitated the integration of bioinformatics into every aspect of a research project. We are at yet another turning point in the evolution of bioinformatics that will see in the coming years bioinformaticians transition from collaborators to leaders that bring interdisciplinary teams together to solve a complex problem. In other words, bioinformaticians will be able to ask the questions, define the hypotheses, and orchestrate the scientific study. This is the natural result of interdisciplinary training, the public availability of data, open-source software, the widespread availability of core facilities for conducting experiments and, importantly, the ability to integrate and synthesize knowledge sources to ask more impactful questions.

2. The Golden Era of Bioinformatics Has Begun

The turning point in the maturity of bioinformatics as a discipline has led some to speculate that we are entering a golden where the focus on computational approaches to biomedical research will be front and center[1]. There are several reasons for this speculation. First, big data is now the norm rather than the exception and computational methods are critical for successful storage, management, retrieval, analysis, and interpretation for answering scientific questions. Bioinformatics has never been so important for moving research forward. Bioinformatics areas such as databases, machine learning, and visualization are in high demand. Second, high-performance computing (HPC) is inexpensive and widely available in different technologies such as cloud computing and parallel computing using graphic processing units (GPUs) that bring thousands of compute core to a single desktop computer. Third, artificial intelligence and machine learning have matured and are now routinely being used to solve complex problems in the biomedical sciences. This is the result of decades of research on intelligent algorithms and software and the HPC resources necessary to apply them to big data. Fourth, the power of combining computational intelligence with statistical methods has emerged in the form of data science that allows the integration of different philosophical and quantitative schools of thought to solve biomedical problems. Fifth, visual analytics that brings visualization technology together with data science and human-computer interaction is maturing quickly with areas such as virtual reality, augmented reality, and 3D printing. Visual analytics will be essential for allowing human to interact with and understand data and research results that are too big and too complex to understand. Sixth, data and knowledge integration is maturing quickly as we have seen with electronic health records, data warehouses, and knowledge resources such as PubMed. Seventh, there is an increasing recognition of the importance of bioinformatics by federal funding agencies, biotechnology and pharmaceutical companies, and academic institutions. Investment in bioinformatics personnel and technology has never been greater and is expanding quickly. Now is the time for bioinformatics to have a substantial impact on biological and biomedical research.

3. No-Boundary Thinking in Bioinformatics

The purpose of this workshop is to introduce and discuss the future of bioinformatics as a mature discipline. We have previously defined this evolution and its impact as No-Boundary Thinking (NBT) in Bioinformatics[2,3]. The NBT philosophy provides bioinformaticians with the unique opportunity to move past being service providers to asking and answering research questions. This is because they

are in the best position to integrate and synthesize knowledge across many disciplines to articulate a question that might have broader impact than one formulated from the knowledge of a single discipline. This allows them to be an equal contributor to the motivation and design phases of research studies. NBT puts the emphasis on knowledge-based question definition with big data serving a secondary and supporting role. This is counter to the current philosophy of letting big data drive the questions that are asked[3]. The workshop will introduce and define the NBT approach and will provide several scientific examples. An important component the workshop is providing examples of how NBT can be moved into the classroom to prepare bioinformatics students for a future where they are leading scientific studies. Panel discussions around NBT in science and education will allow for a robust discussion about these new ideas.

References

1. J.H. Moore, J.H. Holmes, *BioData Mining* **9**, 15 (2016).
2. X. Huang, B. Bruce, A. Buchan, C.B. Congdon, C.L. Cramer, S.F. Jennings, H. Jiang, Z.Li, G. McClure, R. McMullen, J.H. Moore, N. Nanduri, J. Peckham, A. Perkins, S.W. Polson, B. Rekepalli, S. Salem, J. Specker, D. Wunsch, D. Xiong, S. Zhang, Z. Zhao, *BioData Mining* **6**, 19 (2013).
3. X. Huang, S.F. Jennings, B. Bruce, A. Buchan, L. Cai, P. Chen, C.L. Cramer, W. Guan, U.K. Guan, U.K. Hilgert, H. Jiang, Z. Li, G. McClure, D.F. McMullen, B. Nanduri, A. Perkins, B. Rekepalli, S. Salem, J. Specker, K. Walker, D. Wunsch, D. Xiong, S. Zhang, Z. Zhao, J.H. Moore, *BioData Mining* **8**, 7 (2015).

OPEN DATA FOR DISCOVERY SCIENCE

PHILIP R.O. PAYNE[1], KUN HUANG[2], NIGAM H. SHAH[3], JESSICA TENENBAUM[4]

[1]*Washington University Institute for Informatics, Washington University in St. Louis School of Medicine, St. Louis, MO 63130, United States of America*

[2]*Department of Biomedical Informatics, The Ohio State University College of Medicine, Columbus, OH 43210, United States of America*

[3]*Center for Biomedical Informatics Research, Stanford University, Stanford, CA 94305, United States of America*

[4]*Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, United States of America*

Email: prpayne@wustl.edu, kun.huang@osumc.edu, nigam@stanford.edu, jessie.tenenbaum@duke.edu

The modern healthcare and life sciences ecosystem is moving towards an increasingly open and data-centric approach to discovery science. This evolving paradigm is predicated on a complex set of information needs related to our collective ability to share, discover, reuse, integrate, and analyze open biological, clinical, and population level data resources of varying composition, granularity, and syntactic or semantic consistency. Such an evolution is further impacted by a concomitant growth in the size of data sets that can and should be employed for both hypothesis discovery and testing. When such open data can be accessed and employed for discovery purposes, a broad spectrum of high impact end-points is made possible. These span the spectrum from identification of *de novo* biomarker complexes that can inform precision medicine, to the repositioning or repurposing of extant agents for new and cost-effective therapies, to the assessment of population level influences on disease and wellness. Of note, these types of uses of open data can be either primary, wherein open data is the substantive basis for inquiry, or secondary, wherein open data is used to augment or enrich project-specific or proprietary data that is not open in and of itself. This workshop is concerned with the key challenges, opportunities, and methodological best practices whereby open data can be used to drive the advancement of discovery science in all of the aforementioned capacities.

1. Rationale for Workshop

There are significant realized and potential benefits associated with the use of open data for discovery science. Unfortunately, despite such opportunities, the computational and informatics tools and methods currently used in most investigational settings to enable such efforts are often labor intensive and rely upon technologies that have not be designed to scale and support reasoning across heterogeneous and multi-dimensional data resources (1-3). As a result, there are significant demands from the research community for the creation and delivery of data management and data analytic tools capable of adapting to and supporting heterogeneous analytic workflows and open data sources (4-7). This need is particularly important when researchers seek to focus on the large-scale identification of linkages between bio-molecular and phenotypic data in order inform novel systems-level approaches to understanding disease states. In these types of situations, the scalar nature of such data exacerbates almost all of the aforementioned challenges. In this context, it is of interest to note that while the theoretical basis for the use of knowledge-based systems to overcome such challenges have evolved rapidly, their use in "real world" context remains the domain of experts with specialized training and unique access to such tools (1, 8, 9).

All of the preceding issues are further amplified when considering the nature of modern approaches to hypothesis discovery and testing when exploring biological and clinical open data, which are often based on the intuition of the individual investigator or his/her team to identify a question that is of interest relative to their specific scientific aims, who then carry out hypothesis testing operations to validate or refine that question relative to a targeted data set (10, 11). This approach is feasible when exploring data sets comprised of hundreds of variables, but does not scale to projects involve data sets with magnitudes on the order of thousands or even millions of variables (1, 8). An emerging and increasingly viable solution to this particular challenge is the use of domain knowledge to generate hypotheses relative to the content of such data sets. This type of domain knowledge can be derived from many different sources, such as complementary and contextualizing databases, terminologies, ontologies, and published literature (8). It is important to note, however, that methods and technologies that can allow researchers to access and extract domain knowledge from such sources, and apply resulting knowledge extracts to generate and test hypotheses are largely developmental at the current time (1, 8).

Finally, even when the major hurdles to the regular use of open data for discovery science as noted above are adequately addressed, there remains a substantial reliance on the use of data-analytic "pipelining" tools to ensure the systematic and reproducible nature of such data analysis operations. These types of pipelines are ideally able to support data extraction, integration, and analysis workflows spanning multiple sources, while capturing intermediate data analysis steps and products, and generating actionable output types (12, 13). Using data-analytic pipelines provide a number of potential benefits, including: 1) they support the design and execution of data analysis plans that would not be tractable or feasible using manual methods; and 2) they provide for the capture meta-data describing the steps and intermediate products generated during such data analyses. In the case of the latter benefit, the ability to capture systematic meta-data is critical to ensuring that such *in-silico* research paradigms generate reproducible and high quality results (12, 13). Again, while there are a number of promising technology platforms capable of supporting such data-analytic "pipelining", their widespread use is not robust, largely due to barriers to adoption related to data ownership/security, usability, scalability, and socio-technical factors (7, 14).

Given the aforementioned challenges and opportunities and the current state of knowledge concerning the use of open data across and between types and scales for the purposes of discovery science, this workshop addresses the following major topic areas:

- The state-of-the-art in terms of tools and methods targeting the use of open data for discovery science, including but not limited to syntactic and semantic standards, platforms for data sharing and discovery, and computational workflow orchestration technologies that enable the creation of data analytics "pipelines";
- Practical approaches for the automated and/or semi-automated harmonization, integration, analysis, and presentation of "data products" to enable hypothesis discovery or testing; and
- Frameworks for the application of open data to support or enable hypothesis generation and testing in projects spanning the basic, translational, clinical, and population health research and practice domains (e.g., from molecules to populations).

3. Workshop Speakers

Philip R.O. Payne, PhD: Dr. Payne is the founding Director of the Institute for Informatics (I2) at Washington University in St. Louis, where he also serves as a Professor in the Division of General Medical Sciences. Previously, Dr. Payne was Professor and Chair of the Department of Biomedical Informatics at The Ohio State University. Dr. Payne's research primarily focuses on the use of knowledge-based methods for in silico hypothesis discovery. He received his Ph.D. with distinction in Biomedical Informatics from Columbia University, where his research focused on the use of knowledge engineering and human-computer interaction design principles in order to improve the efficiency of multi-site clinical and translational research programs.

Kun Huang, PhD: Dr. Kun Huang is Professor in Biomedical Informatics, Computer Science and Engineering, and Biostatistics at The Ohio State University. He is also the Division Director for Bioinformatics and Computational Biology in OSU Department of Biomedical Informatics and Associate Dean for Genomic Informatics in the OSU College of Medicine. He has developed many methods for analyzing and integrating various types of high throughput biomedical data including gene expression microarray, next generation sequencing (NGS), qRT-PCR, proteomics and microscopic imaging experiments. Dr. Huang received his BS degree in Biological Sciences from Tsinghua University in 1996 and his MS degrees in Physiology, Electrical Engineering and Mathematics all from the University of Illinois at Urbana-Champaign (UIUC). He then received his PhD in Electrical and Computer Engineering from UIUC in 2004 with a focus on computer vision and machine learning.

Nigam Shah, MBBS, PhD: Dr. Nigam Shah is associate professor of Medicine (Biomedical Informatics) at Stanford University, Assistant Director of the Center for Biomedical Informatics Research, and a core member of the Biomedical Informatics Graduate Program. Dr. Shah's research focuses on combining machine learning and prior knowledge in medical ontologies to enable use cases of the learning health system. Dr. Shah was elected into the American College of Medical Informatics (ACMI) in 2015 and to the American Society for Clinical Investigation (ASCI) in 2016. He holds an MBBS from Baroda Medical College, India, a PhD from Penn State University and completed postdoctoral training at Stanford University.

Jessica Tenenbaum, PhD: Dr. Tenenbaum is Assistant Professor in the Division of Translational Biomedical Informatics, Department of Biostatistics and Bioinformatics at Duke University, and Associate Director for Bioinformatics for the Duke Translational Medicine Institute. Her primary areas of research include infrastructure and standards to enable research collaboration and integrative data analysis; informatics to enable precision medicine; and ethical, legal, and social issues that arise in translational research, direct to consumer genetic testing, and data sharing. After earning her bachelor's degree in biology from Harvard, Dr. Tenenbaum worked as a program manager at Microsoft Corporation in Redmond, WA for six years before pursuing a PhD in biomedical informatics at Stanford University.

2. Acknowledgements

The authors wish to acknowledge the contributions of Drs. Gustavo Stolovitzky (IBM) and Josh Swamidass (Washington University in St. Louis) to the preparation of this workshop summary.

References

1.Ruttenberg A, Clark T, Bug W, Samwald M, Bodenreider O, Chen H, et al. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007;8 Suppl 3:S2.
2.Fridsma DB, Evans J, Hastak S, Mead CN. The BRIDG project: a technical report. J Am Med Inform Assoc. 2008;15(2):130-7.
3.Kush RD, Helton E, Rockhold FW, Hardison CD. Electronic health records, medical research, and the Tower of Babel. The New England journal of medicine. 2008;358(16):1738-40.
4.Payne PR, Johnson SB, Starren JB, Tilson HH, Dowdy D. Breaking the translational barriers: the value of integrating biomedical informatics and translational research. J Investig Med. 2005;53(4):192-200.
5.Maojo V, García-Remesal M, Billhardt H, Alonso-Calvo R, Pérez-Rey D, Martín-Sánchez F. Designing new methodologies for integrating biomedical information in clinical trials. Methods of information in medicine. 2006;45(2):180-5.
6.Casey K, Elwell K, Friedman J, Gibbons D, Goggin M, Leshan T, et al. Broken Pipeline: Flat Funding of the NIH Puts a Generation of Science at Risk
. 2008. p. 24.
7.Ash JS, Anderson NR, Tarczy-Hornoch P. People and Organizational Issues in Research Systems Implementation. Journal of the American Medical Informatics Association : JAMIA. 2008.
8.Payne PR, Mendonca EA, Johnson SB, Starren JB. Conceptual knowledge acquisition in biomedicine: A methodological review. J Biomed Inform. 2007;40(5):582-602.
9.Richesson RL, Krischer J. Data standards in clinical research: gaps, overlaps, challenges and future directions. Journal of the American Medical Informatics Association : JAMIA. 2007;14(6):687-96.
10.Erickson J. A decade and more of UML: An overview of UML semantic and structural issues and UML field use. Journal of Database Management. 2008;19(3):I-Vii.
11.Butte AJ. Medicine. The ultimate model organism. Science. 2008;320(5874):325-7.
12.van Bemmel JH, van Mulligen EM, Mons B, van Wijk M, Kors JA, van der Lei J. Databases for knowledge discovery. Examples from biomedicine and health care. International journal of medical informatics. 2006;75(3-4):257-67.
13.Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, et al. caGrid 1.0: An Enterprise Grid Infrastructure for Biomedical Research. Journal of the American Medical Informatics Association : JAMIA. 2008;15(2):138-49.
14.Kukafka R, Johnson SB, Linfante A, Allegrante JP. Grounding a new information technology implementation framework in behavioral science: a systematic analysis of the literature on IT use. J Biomed Inform. 2003;36(3):218-27.