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.