Published in final edited form as:
*J Biomed Inform*. 2017 May ; 69: 115–117. doi:10.1016/j.jbi.2017.03.017.

# Envisioning the Future of ‘Big Data’ Biomedicine

**Alex A. T. Bui**[1] and **John Darrell Van Horn**[2] **the NIH BD2K Centers Consortium**

[1]BD2K Centers Coordinating Center (BD2K CCC), www.bd2kccc.org, University of California Los Angeles, Los Angeles, CA USA

[2]BD2K Training Coordinating Center (BD2K TCC), www.bigdatau.org, University of Southern California, Los Angeles, CA USA

## Introduction

In our era of digital biomedicine, data take many forms, from “omics” to imaging, mobile health (mHealth), and electronic health records (EHRs). With the availability of more efficient digital collection methods, scientists in many domains now find themselves confronting ever larger sets of data and trying to make sense of it all (1–4). Indeed, data which used to be considered large now seems small as the amount of data now being collected in a single day by an investigator can surpass what might have been generated over his/her career even a decade ago (e.g., (e.g. 5)). This deluge of biomedical information requires new thinking about how data are generated, managed, and ultimately leveraged to further scientific understanding and for improving healthcare. Responding to this challenge, the National Institutes of Health (NIH) has spearheaded the “Big Data to Knowledge” (BD2K) program (6). Data scientists are being engaged through BD2K to guide biomedical researchers through the thickets of data they are producing. NIH Director, Francis Collins, has noted, “Indeed, we are at a point in history where Big Data should not intimidate, but inspire us. We are in the midst of a revolution that is transforming the way we do biomedical research…we just have to devise creative ways to sift through this mountain of data and make sense of it” (7). The NIH is now taking its first major steps toward realizing biomedical science as an interdisciplinary “big data” science.

## Maximizing the potential of data for everyone

The interplay between biomedical science and advancing technology drives a continuous cycle of data growth: as new technologies enable more and different varieties of data to be amassed, scientists exploit the potential of these technologies and the data being produced to uncover knowledge, and pose new questions that require novel technologies to probe further. Inherent to this accelerating cycle is the requirement to handle the growing data complexity and computational analyses. Towards this, the BD2K program involves the efforts of multiple Centers of Excellence (Table 1), two coordinating centers, and a set of focused individual research and training projects, considering the latest approaches in data science and their application to large-scale biomedical data.

For instance, the challenges around the development of scalable computing are being examined by the Big Data for Discovery Science Center (BDDS), advancing highperformance computing and methods for disseminating large datasets (8). More sophisticated data management approaches are taken, capturing not only the data, but its provenance and metadata – to support scientific reproducibility. Here, the Center for Expanded Data Annotation and Retrieval (CEDAR) has developed software for editing and sharing metadata templates describing the information needed to annotate datasets, working with biomedical communities to build a searchable library that reflects their standards (9). Other BD2K initiatives seek to democratize biomedical data usage by making resources and materials "findable, accessible, interoperable, and reusable" (i.e., the so-called "FAIR" scientific guiding principles (10)). The biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) program has launched DataMed (https://datamed.org), a resource analogous to PubMed but for cataloging data resources so they can be easily reused and cited. Heart BD2K is indexing the software tools and platforms which operate on these datasets (https://dev.aztec.io). Collectively, these efforts reframe biomedical data and its use into an open science ecosystem supporting everyday collaboration and discovery, fostering NIH's emergent Commons Framework (https://datascience.nih.gov/commons) to enhance the exchange of scientific data between investigators as well as major data archives.

## A holistic perspective on biology and health

An exciting aspect of biomedical big data comes from the untapped integration of large-scale, heterogeneous information repositories to find new relationships. BD2K's activities pave the way to link -omics, phenotype, behavior, and outcomes into a more complete picture of the human condition. For example, the Library of Integrated Network-based Cellular Signatures (LINCS) Data Coordination and Integration Center has unveiled Harmonizome, a web-based portal aggregating omics resources for knowledge mining of genes and proteins (11). KnowEng has created a network of big data resources embodying community-based data on genes, proteins, functions, and phenotypes (12). Application programming interfaces (APIs) and software tools that integrate genomics into clinical perspectives are pursued by the Center for Big Data in Translational Genomics. Working at the intersection of neuroimaging and genomics, the Enabling Neuroimaging Genetics through Meta-Analysis (ENIGMA) Center has performed some of the world's largest genetic studies of the brain to date – revealing genomic loci affecting brain structure and risk for disease (13). And changing how clinical data is captured for research, PIC-SURE's patient-centric information commons combines genetic, environmental, imaging, and EHRs; this supports the Sync for Science project (http://syncfor.science) to gather health information for the landmark precision medicine research program, All of Us (https://www.nih.gov/research-training/allofus-research-program). These BD2K efforts represent key developments for integrating disparate datasets into a comprehensive view of health, connecting molecular and genomic features to human phenotypes.

## Enabling learning from the past to predict the future

Data mining and machine learning epitomize modern big data science. BD2K aims to forge contemporary insights by advancing these techniques in biomedicine. Making these and other modern computational methods more accessible, the Mobilize Center's Snorkel application enables users with little-to-no programming background to extract and analyze information from their data. Constructing predictive models elucidating disease trajectories over time are the focus of several BD2K sites, with the objective of using such models with an individual's past history and current context to tailor interventions. The Center for Predictive Computational Phenotyping (CPCP) is building such models with multimodal biomarkers, providing physicians with better prognostic tools (14). The Center for Causal Discovery (CCD) develops algorithms and software which learn causal networks from high-dimensional biomedical datasets (15), resulting, for example, in networks resolved at the image voxel level for the resting state functional dynamics of the human cortex. The Mobile Sensor Data to Knowledge (MD2K) program is pioneering mobile health methods for assessment of human health in everyday settings (16); their mCerebrum platform provides scalable, continuous data collection with real-time analytics. BD2K programs are developing the computing needed to process vast quantities of data used in creating these models; and sets the stage for them to be improved as more data and knowledge become available.

## Training the current and next generation of biomedical data scientists

Importantly, interwoven throughout the NIH BD2K program is a range of data science training programs including undergraduate and graduate education, summer workshops, career path development, and online content, including MOOCs and video lectures (17). Involvement of underserved communities is a major component of these endeavors. The BD2K Training Coordinating Center (TCC), in particular, is working to create an Educational Resource Discovery Index (ERuDIte) of online training course materials for use in crafting personalized training curricula in biomedical data science. Working jointly, the BD2K Centers Coordinating Center (CCC), TCC, and NIH Office of the Associate Director for Data Science are producing an online seminar series to educate and engage the scientific community and general public concerning the data science underlying modern biomedical investigation (http://www.bigdatau.org/data-science-seminars).

## Conclusion

The BD2K program's charge has been to make large-scale data usage commonplace – streamlining its synthesis, exploration, and its ease of analysis. While the term "big data" carries with it some skepticism (18), technical challenges (19), and even ethical concerns (20), BD2K is showing how these issues can be explored and surmounted. The BD2K program's efforts have already made strides towards applying data science and computational methods to a range of modern biomedical research challenges. Though time will tell, having made inroads toward exploring big data biomedicine the BD2K program appears poised to influence the future of biomedical research and clinical care. Indeed, the NIH's investment in BD2K has already been unprecedented, laying the foundation for future advances in precision and participatory medicine (21, 22). Without the support of multiple NIH institutes, a far-reaching strategy to evolve the "discovery science" of biomedical big data from would not be attainable.

## Acknowledgments

This work is supported by the NIH Big Data to Knowledge (BD2K) awards to the BD2K CCC (U54 GM114833; https://bd2kccc.org) and BD2K TCC (U24 ES026465; http://www.bigdatau.org). Further information on the NIH

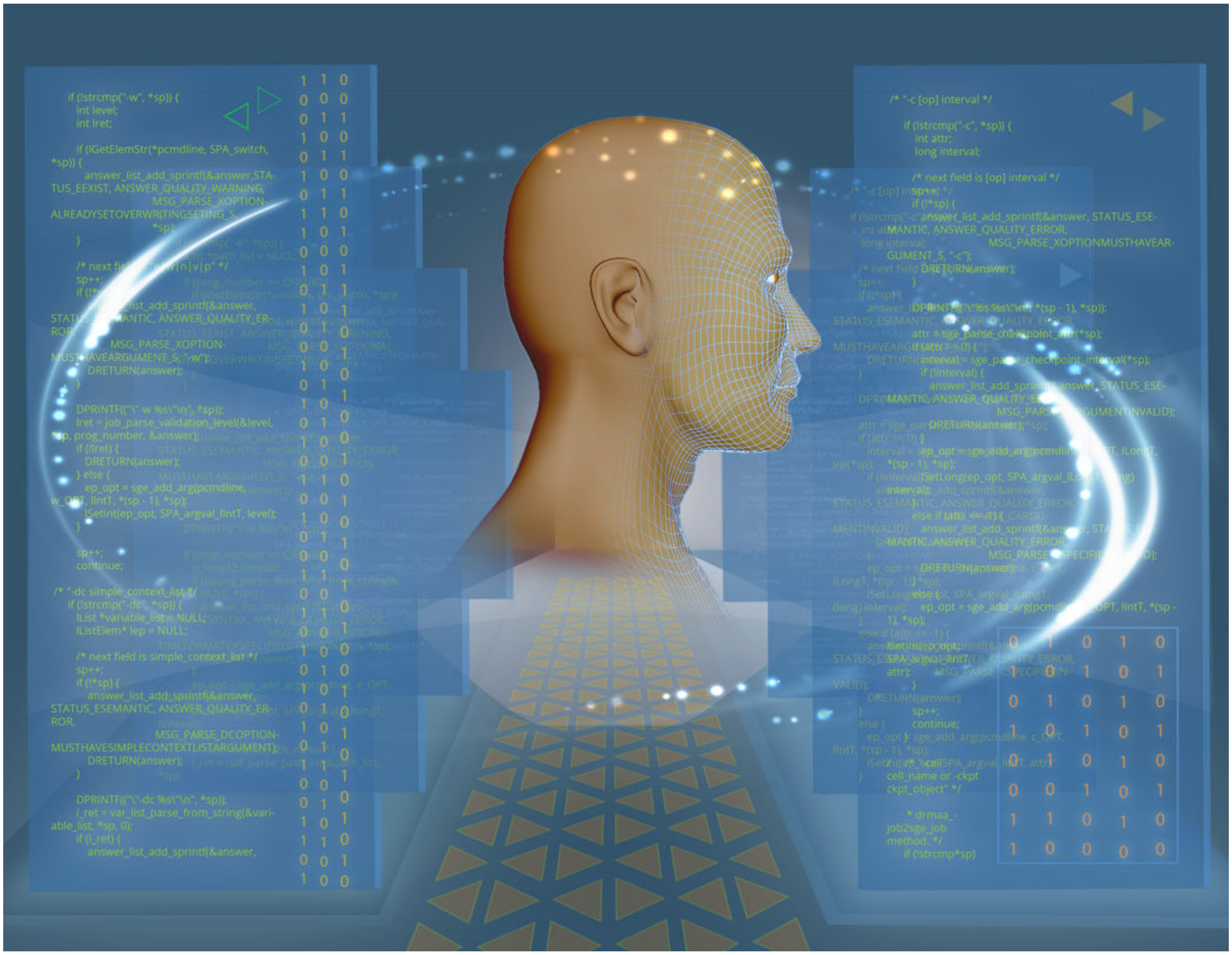
BD2K program, its contributing NIH institutes, and all of its supported projects may be found via the NIH Data Science Office website (https://datascience.nih.gov/). The *NIH BD2K Centers Consortium is comprised of the following NIH BD2K U54 awardees: Gregory Cooper, Ivet Bahar, and Jeremy Berg, University of Pittsburgh; Mark Craven, University of Wisconsin-Madison; Scott Delp, Stanford University; Todd Golub and Aravind Subramanian, Broad Institute; Jiawei Han, Saurabh Sinha, Jun Song, and Richard Weinshilboum, University of Illinois Urbana-Champaign; David Haussler, David Patterson, and Laura Van't Veer, University of California Santa Cruz; Isaac Kohane, Harvard University Medical School; Santosh Kumar, University of Memphis; Avi Ma'ayan, Mario Medvedovic, and Stephan Schurer, Icahn School of Medicine at Mount Sinai; Mark Musen, Stanford University; Peipei Ping, Merry Lindsey, Andrew Su, and Karol Watson, University of California Los Angeles; Paul Thompson, University of Southern California; and Arthur W. Toga, University of Southern California. The authors would also like to acknowledge Clark Glymour, Michelle Kienholtz for their comments and edits, as well as the teams of the BD2K Centers Coordinating Center (CCC) and the BD2K Training Coordinating Center (TCC).

## References

1. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. Brain Imaging Behav. 2014; 8:323–331. [PubMed: 24113873]
2. Brunk E, et al. Systems biology of the structural proteome. BMC Syst Biol. 2016; 1026
3. Dean DA 2nd, et al. Scaling Up Scientific Discovery in Sleep Medicine: The National Sleep Research Resource. Sleep. 2016; 39:1151–1164. [PubMed: 27070134]
4. Gange SJ, Golub ET. From Smallpox to Big Data: The Next 100 Years of Epidemiologic Methods. Am J Epidemiol. 2015
5. Van Horn JD, Toga AW. Is it time to re-prioritize neuroimaging databases and digital repositories? Neuroimage. 2009; 47:1720–1734. [PubMed: 19371790]
6. Bourne PE, et al. The NIH Big Data to Knowledge (BD2K) initiative. J Am Med Inform Assoc. 2015; 221114
7. Collins, FS. NIH Director's Blog. Morgan, K., editor. The National Institutes of Health; Bethesda, MD, USA: 2014.
8. Toga AW, et al. Big biomedical data as the key resource for discovery science. J Am Med Inform Assoc. 2015; 22:1126–1131. [PubMed: 26198305]
9. Musen MA, et al. The center for expanded data annotation and retrieval. J Am Med Inform Assoc. 2015; 22:1148–1152. [PubMed: 26112029]
10. Wilkinson MD, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016; 3(160018)
11. Rouillard AD, et al. The Harmonizome: A collection of processed datasets gathered to serve and mine knowledge about genes and proteins. Database (Oxford). 2016; 2016
12. Sinha S, Song J, Weinshilboum R, Jongeneel V, Han J. KnowEnG: A knowledge engine for genomics. J Am Med Inform Assoc. 2015; 22:1115–1119. [PubMed: 26205246]
13. Hibar DP, et al. Common genetic variants influence human subcortical brain structures. Nature. 2015; 520:224–229. [PubMed: 25607358]
14. Craven M, Page CD. Big Data in Healthcare: Opportunities and Challenges. Big Data. 2015; 3:209–210. [PubMed: 27441403]
15. Cooper GF, et al. The Center for Causal Discovery of biomedical knowledge from big data. J Am Med Inform Assoc. 2015; 22:1132–1136. [PubMed: 26138794]
16. Kumar S, et al. Center of excellence for mobile sensor data-to-knowledge (MD2K). J Am Med Inform Assoc. 2015; 22:1137–1142. [PubMed: 26555017]
17. Van Horn JD. Opinion: Big data biomedicine offers big higher education opportunities. Proceedings of the National Academy of Sciences. 2016; 113:6322–6324.
18. Patel AA, Singh K, Nunley RM, Minhas SV. Administrative Databases in Orthopaedic Research: Pearls and Pitfalls of Big Data. J Am Acad Orthop Surg. 2016; 24:172–179. [PubMed: 26836377]
19. Topol EJ. The big medical data miss: challenges in establishing an open medical resource. Nat Rev Genet. 2015; 16:253–254. [PubMed: 26065035]
20. Rothstein MA. Ethical Issues in Big Data Health Research: Currents in Contemporary Bioethics. J Law Med Ethics. 2015; 43:425–429. [PubMed: 26242964]

21. Xie L, et al. Towards structural systems pharmacology to study complex diseases and personalized medicine. PLoS Comput Biol. 2014; 10:e1003554. [PubMed: 24830652]

22. Hood L, Auffray C. Participatory medicine: a driving force for revolutionizing healthcare. Genome Med. 2013; 5(110)



```
      if (!strcmp("-w", *sp)) {
         int level;
         int lret;

         if (lGetElemStr(*pcmdline, SPA_switch,
*sp)) {
            answer_list_add_sprintf(&answer,STA-
TUS_EEXIST, ANSWER_QUALITY_WARNING,
                              MSG_PARSE_XOPTION-
ALREADYSETOVERWRITINGSETING_S,
                              *sp);
         }

         /* next fie    e|w|n|v|p" */
         sp++;
         if (!*
                 _list_add_sprintf(&answer,
STATU       MANTIC, ANSWER_QUALITY_ER-
ROR,
                     MSG_PARSE_XOPTION-
MUST HAVEARGUMENT_S, "-w");
            DRETURN(answer);
         }

         DPRINTF(("\"-w %s\"\n", *sp));
         lret = job_parse_validation_level(&level,
*sp, prog_number, &answer);
         if (!lret) {
            DRETURN(answer);
         } else {
            ep_opt = sge_add_arg(pcmdline,
w_OPT, lIntT, *(sp - 1), *sp);
            lSetInt(ep_opt, SPA_argval_lIntT, level);
         }

         sp++;
         continue;

   /* "-dc simple_context_list */
      if (!strcmp("-dc", *sp)) {
         lList *variable_list = NULL;
         lListElem* lep = NULL;

         /* next field is simple_context_list */
         sp++;
         if (!*sp) {
            answer_list_add_sprintf(&answer,
STATUS_ESEMANTIC, ANSWER_QUALITY_ER-
ROR,
                              MSG_PARSE_DCOPTION-
MUSTHAVESIMPLECONTEXTLISTARGUMENT);
            DRETURN(answer);
         }

         DPRINTF(("\"-dc %s\"\n", *sp));
         i_ret = var_list_parse_from_string(&vari-
able_list, *sp, 0);
         if (i_ret) {
            answer_list_add_sprintf(&answer,
```

```
      /* "-c [op] interval */

         if (!strcmp("-c", *sp)) {
            int attr;
            long interval;

            /* next field is [op] interval */
            sp++;
            if (!*sp) {
               answer_list_add_sprintf(&answer, STATUS_ESE-
         MANTIC, ANSWER_QUALITY_ERROR,
                              MSG_PARSE_XOPTIONMUSTHAVEAR-
         GUMENT_S, "-c");
               DRETURN(answer);
            }
```

## Table 1

NIH "Big Data to Knowledge" (BD2K) Centers Programs

| BD2K Center of Excellence | Institution | Lead PI(s) |
|---|---|---|
| ***Big Data for Discovery Science (BDDS)*** | University of Southern California (USC) | Arthur W. Toga |
| ***Center for Causal Discovery (CCD)*** | University of Pittsburgh | Gregory F. Cooper, Ivet Bahar, Jeremy M. Berg |
| ***Center for Expanded Data Annotation and Retrieval (CEDAR)*** | Stanford University | Mark A. Musen |
| ***Center for Predictive Computational Phenotyping (CPCP)*** | University of Wisconsin-Madison | Mark W. Craven |
| ***ENIGMA Center for Worldwide Medicine, Imaging, and Genomics*** | University of Southern California (USC) | Paul M. Thompson |
| ***Heart BD2K*** | University of California Los Angeles (UCLA) | Peipei Ping, Merry Lindsey, Andrew Su, Karol Watson |
| ***KnowEnG*** | University of Illinois at Urbana-Champaign | Jiawei Han, Saurabh Sinha, Jun S. Song, Richard M. Weinshilboum |
| ***BD2K-LINCS DCIC*** | Icahn School of Medicine at Mount Sinai | Avi Ma'ayan, Mario Medvedovic, Stephan C. Schurer |
| ***Mobile Sensor Data to Knowledge (MD2K)*** | University of Memphis | Santosh Kumar |
| ***Mobilize*** | Stanford University | Scott Delp |
| ***Center for Big Data in Translational Genomics*** | University of California Santa Cruz (UCSC) | David Haussler, David Patterson, Laura J. van 't Veer |
| ***PIC-SURE*** | Harvard University | Isaac S. Kohane |
| ***Broad Institute LINCS Center for Transcriptomics and Toxicology*** | Broad Institute | Todd Golub, Aravind Subramanian |
| ***bioCADDIE*** | University of California San Diego (UCSD) | Lucilla Ohno-Machado |
| **BD2K Coordinating Centers** | | |
| ***BD2K Centers Coordinating Center (CCC)*** | University of California Los Angeles (UCLA) | Peipei Ping, Alex Bui, Wei Wang |
| ***BD2K Training Coordinating Center (TCC)*** | University of Southern California (USC) | John D. Van Horn |

For full descriptions of BD2K U01, R25, T32/T15, and other awards, see https://datascience.nih.gov/bd2k/funded-proarams