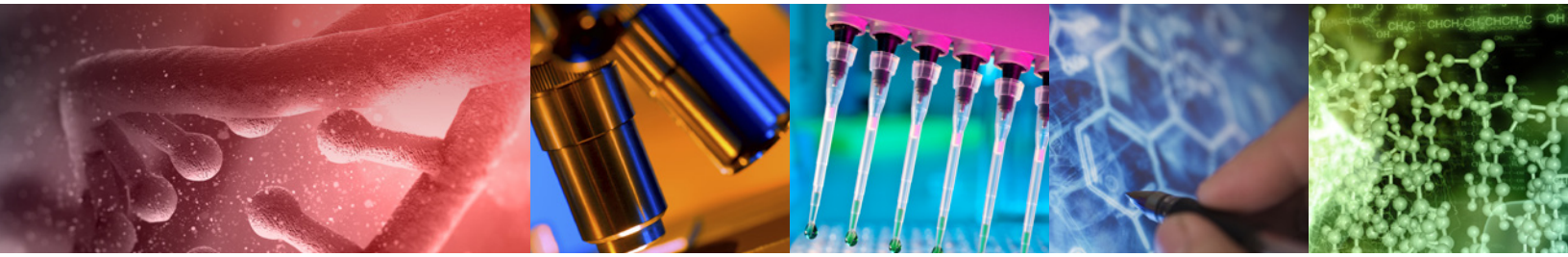


NIH Big Data to Knowledge Program

Advancing Health and Discovery through Big Data

Program Overview

As biomedical tools and technologies rapidly improve, researchers are producing and analyzing an ever-expanding amount of complex biological data called "big data." The Big Data to Knowledge (BD2K) program is a trans-NIH initiative that was launched in 2013 to support the research and development of innovative and transformative approaches and tools to maximize and accelerate the integration of big data and data science into biomedical research.The BD2K Program also supported initial efforts toward making data sets findable and reusable, efforts that have evolved into "FAIR" Findable, Accessible, Interoperable, and Reusable principles. (https://commonfund.nih.gov/bd2k)

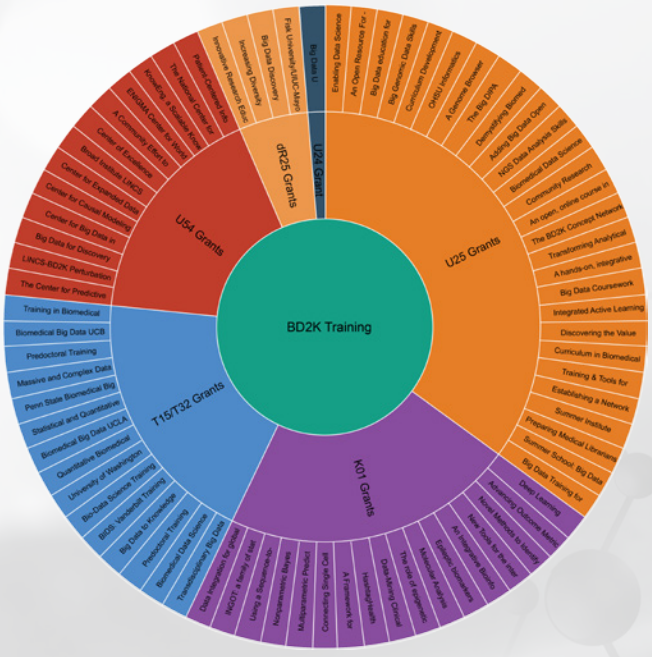
BD2K Phase I & II

The first phase of BD2K focused on facilitating broad use of biomedical big data, developing and disseminating analysis methods and software, enhancing training relevant for large-scale data analysis, and establishing centers of excellence for biomedical big data. A list of projects funded under BD2K Phase I is available on the BD2K website (https://commonfund.nih.gov/bd2k/fundedresearch). BD2K has now entered a second phase that will focus on making the products of research developed in Phase I usable, discoverable, and disseminated to their intended end-users. In addition, the program will support the NIH Data Commons Pilot initiative to test the feasibility of, and develop best practices for, making NIH-funded data sets and computational tools available through communal, collaborative platforms on public clouds. (https://commonfund.nih.gov/bd2k/commons)

Program Initiatives

Training and Workforce Development

BD2K has issued a number of awards in biomedical big data training with the goals of improving the big data skills of biomedical scientists, and increasing the number of biomedical data scientists. Biomedical big data training resources developed by BD2K grantees are available for use by the research community, and can be discovered through ERuDIte, an educational resource index (https://bigdatau.ini.usc.edu/about_erudite) which allows interested parties to identify educational resources by topic, format, level of difficulty, and more. ERuDIte was developed and is maintained by the BD2K Training Coordination Center (TCC).TheTCC helps to promote and support training and educational activities across the collection of NIH-funded Big Data to Knowledge (BD2K) program http://www.bigdatau.org. To access the training and learn about additional TCC activities, please visit https://commonfund.nih.gov/sites/default/files/TCC_2017Flyer.pdf.



BD2K Centers

BD2K funded 13 Centers of Excellence, which are large-scale projects developing new approaches, methods, software tools, and related resources, and are also providing training to advance big data science in the context of their biomedical area of focus. A list of resources developed by the Centers is maintained by the BD2K Centers Coordination Center (CCC). The CCC helps to promote collaboration among the Centers and across the BD2K program (https://bd2kccc.org). The Centers "Wow Stories," which capture the Centers' achievements and contribution to data science are available on the BD2K website (https://commonfund.nih.gov/sites/default/files/wowstories.pdf).

Resource Indexing

The BD2K indexing initiative focuses on approaches to find and access biomedical data sets. The bioCADDIE (biomedical and healthCAre Data Discovery Index Ecosystem, https://biocaddie.org) has developed a data discovery index prototype, DataMed (https://datamed.org), designed to do for biomedical data sets what PubMed has done for research articles: make them findable and accessible by more researchers.

Software and Analysis

BD2K issued awards to develop analysis methods and software in targeted areas of high need. These areas included data compression/reduction, data visualization, data provenance, data wrangling, data privacy, data repurposing, and applying metadata. Additional awards were made to support the development of interactive digital media tools for analyzing biomedical data via crowdsourcing. A series of Innovation Labs were supported as a joint NSF/NIH collaboration, providing an intensive mentored collaboration-building opportunity for junior investigators from the biomedical and quantitative fields to work together to bring new computational approaches to biomedical big data.

Standards

The BD2K Standards initiative seeks to promote and improve the management, development, and use of data-related standards by the biomedical research community. BD2K issued awards to provide catalytic support for activities necessary to develop, extend or refine data and metadata standards and/or related tools in areas relevant to the NIH basic, translational, and clinical research mission.

The NIH Data Commons and Data Commons Pilot Phase

In BD2K Phase I, the Data Commons was envisioned as a shared virtual space where scientists could find, deposit, manage, share and reuse data, software, metadata, and workflows—the digital objects of biomedical research. A number of supplements were issued to explore the feasibility of a commons framework and facilitate connectivity, interoperability and access to digital objects through pilot projects. In Phase II, BD2K is implementing the NIH Data Commons Pilot Phase to test ways to store, access and share biomedical data and associated tools in the cloud (https://commonfund.nih.gov/bd2k/commons). The platform will consist of

- A compute platform based on cloud technologies
- Reference data sets
- Software in the form of services and tools needed to make use of the data

Connect With Us!

NIH Common Fund BD2K website:

https://commonfund.nih.gov/bd2k

BD2K Listserv:
https://list.nih.gov/cgi-bin/wa.exe?SUBED1=bd2kupdates&A=1

Common Fund Listserv:
https://list.nih.gov/cgi-bin/wa.exe?SUBED1=NIH-OSC-L&A=1

@NIH_CommonFund

www.facebook.com/NIHCommonFund

QR code for BD2K Program

QR Code for Common Fund General Information