



Holdren/OSTP memo issued in February 2013 and directing all agencies of the government to ensure the results of federally funded scientific research are made broadly available.

The first BD2K Funding Opportunity was for investigator-initiated Centers of Excellence in Data Science (RFA-HG-13-009); this called for proposals to test and validate novel ideas in data science that not only focused on particular challenges but also had the potential for broad impact.

The second launched BD2K area aimed to enhance the training of methodologists and practitioners in data science. Skills in demand under the data science 'umbrella' include computer science, mathematics and statistics, biomedical informatics, biology and medicine, and others, all incorporated as 'data science.' At the same time, the generation of large amounts of data together with the complex questions being posed, requires interdisciplinary teams to design the studies and perform the subsequent data analyses.

A fundamental question for BD2K is how to enable the identification, access, and citation of (ie, credit for) biomedical data. The DIWG proposal for federated data catalogs, as distinct from data repositories, requires descriptions of and pointers to the data. A necessary first step has been recognition of the need to assemble and validate ideas drawn from the broader scientific community in developing a Data Discovery Index (DDI). The DDI will enable advanced approaches to search, integrate, and facilitate visualisation of data. Central to development of the DDI will be the ability to link data to associated publications to enhance discovery and facilitate better understanding and interpretation of data and associated analyses.

Among the difficult challenges is how best to create greater value from the expanding availability and use of electronic health records (EHRs). Clinical data from EHRs, together with individual health data captured by various personal devices, offer considerable opportunities for advancing clinical and biomedical research. Data from clinical sources could very well provide a cost-effective means to study different health interventions, to conduct large-scale surveillance of disease incidence and progression in real time, to identify patient cohorts for recruitment into clinical trials, and more. However, unlike most other forms of biomedical research data, clinical data are typically captured outside of traditional research settings and must then be re-purposed for research use. Doing so raises important issues of consent and protection of patient privacy. Changes in policies and practices are needed to govern research access to clinical data sources and facilitate their use for evidence-based learning in healthcare.

## Conclusion

Addressing the challenges associated with biomedical big data must of necessity engage all parts of the big data ecosystem. While these challenges are complex, they are also addressable. BD2K is deploying an integrated plan of action that will tackle numerous aspects of the big data challenge, including multiple elements of data science, training, policy, and community behaviour. While the NIH is only one part of the biomedical big data ecosystem, it can provide leadership for convening stakeholders, providing seed funding, developing metrics for success, implementing a working process, establishing and modifying policies, and supporting basic infrastructure that can catalyse long-term solutions for the biomedical research community.

## Reference:

R Margolis, L Derr, M Dunn, M Huerta, J Larkin, J Sheehan, M Guyer, E D Green; J Am Med Inform Assoc  
doi:10.1136/amiajnl-2014-002974

([http://jamia.bmj.com/content/early/2014/07/09/amiajnl-2014-002974.full?g=w\\_jamia\\_open\\_tab](http://jamia.bmj.com/content/early/2014/07/09/amiajnl-2014-002974.full?g=w_jamia_open_tab))

Image credit: Wikimedia Commons

Published on : Mon, 11 Aug 2014