BioOne COMPLETE

Harnessing the Power of Big Data in Biological Research

Author: McCulloch, Eve S.

Source: BioScience, 63(9) : 715-716

Published By: American Institute of Biological Sciences

URL: https://doi.org/10.1525/bio.2013.63.9.4

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non - commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.

Downloaded From: https://complete.bioone.org/journals/BioScience on 25 Apr 2024
Terms of Use: https://complete.bioone.org/terms-of-use

Harnessing the Power of Big Data in Biological Research

EVE S. McCULLOCH

From the dawn of civilization to 2003, humankind generated five exabytes [5 billion gigabytes] of data. Now we produce five exabytes every two days… and the pace is accelerating.

—Eric Schmidt, executive chairman at Google, quoted in R. Smolan and J. Erwitt, eds. 2012. *The Human Face of Big Data.* Against All Odds Productions.

A data revolution is changing the face of science. Scientists are confronting research challenges that require the analysis of large amounts of information on topics ranging from global climate patterns to genetic blueprints. These *big data* challenges are often summarized in four words: volume, variety, velocity, and veracity. Managing these four parameters could unlock revolutionary new applications, tap the potential of crowdsourcing, and produce a new way of doing science.

Scientists struggle to capture, curate, share, analyze, and visualize continuously generated data. In March 2012, the White House announced the Big Data Research and Development Initiative, committing more than $200 million to accelerate scientific discovery, strengthen national security, and transform education. Six federal departments and agencies are participating in the initiative. In addition, the Obama administration released the Open Data Policy, promising to make information generated by the federal government—including health care data (e.g., the Health Data Initiative)—more accessible to innovators, researchers, and the public.

Genetic research is facing data challenges. It is now possible for a single investigator to generate volumes of DNA-sequence data that a decade ago required a network of major sequencing centers. These data hold clues to everything from curing cancer to developing superior crop varieties, but these advances will not be realized without better analytical tools.

Federally funded teams from Iowa State, Stanford, Virginia Tech, and the University of Michigan are among those developing biocomputing toolboxes. The goal, stated in their project summary, "is to empower the broader community to benefit from clever parallel algorithms, highly tuned implementations, and specialized high-performance computing hardware, without requiring expertise in any of these."

Genetic information combined with health care data could revolutionize medicine—reducing costs and improving outcomes through increased treatment efficiency and medical innovation. The variety of health care data, however, is a significant obstacle to integration: Pharmaceutical companies, health care providers, and public stakeholders have huge stores of medical data. However, according to a recent report by McKinsey and Company—a global for-profit consulting firm—the US health care industry could potentially save $300 billion to $450 billion a year (12–17 percent of health care costs) with systemwide integration of health care data.

Other fields are also confronting big data. "[Individual] ecologists are already collectively producing big data… but we are not harnessing its power," stated Stephanie Hampton and her coauthors in a recent publication in *Frontiers in Ecology and the Environment* (doi:10.1890/120103). They posited that data, themselves, are important products of research: "to address major environmental challenges [researchers] will [have to] leverage their expertise by leveraging their data."

The promise of open-access big data is evident in the National Ecological Observatory Network (NEON). Once it is fully operational, NEON will produce colossal data sets, capturing changes in the biosphere, the geosphere, the hydrosphere, and the atmosphere, using measurements of 539 variables taken continuously at 106 locations nationwide from 2017 to 2047. Its potential applications are tremendous.

Researchers are increasingly tapping the potential of big data from scientific collections and other sources. The United States has more than 1600 biological collections and a billion specimens. The Network Integrated Biocollections Alliance is an initiative developed by the scientific community that is focused on mobilizing a sustained, large-scale digitization effort to answer critical questions about the environment, human health, biosecurity, commerce, and the biological sciences. "Data are much more easily accessed through a central portal than through… separate institutions, and this is a huge benefit to scientists," explained Larry Page, project director at iDigBio, a National Science Foundation–funded organization enabling the digitization and sharing of data from all US biological collections.

Big data could change what scientists know and how they do science. Rather than analyzing data to answer a particular question, creative data mining may allow data to inspire questions—opening the door for hypothesis-generating as well as hypothesis-driven science.

Downloaded From: https://complete.bioone.org/journals/BioScience on 25 Apr 2024
Terms of Use: https://complete.bioone.org/terms-of-use

Obstacles to the realization of big data science remain. "For most researchers, there is no clear reward system for sharing data," says Hampton. "It takes a lot of time to prepare data for sharing and a lot of money to archive it well," although, Hampton adds, "there is a very vocal group of mostly early-career scientists who want science to be open… [and who are] reconsidering the paradigm of publishing and knowledge transfer more generally." Ethical questions and technical issues are also challenging. Logistically, however, programmers are producing tools to manage large volumes of rapidly delivered data. Alliances between the public and private sectors may quicken this enterprise.

Big data initiatives also face funding challenges. Budget sequestration is hitting agencies hard. Funding for basic government operations is constrained, so funding for new initiatives may require persistent advocacy from the scientific community and stakeholder groups who will use big data.

Eve S. McCulloch (emcculloch@aibs.org) is a science and public affairs writer at the American Institute of Biological Sciences.

doi:10.1525/bio.2013.63.9.4

Downloaded From: https://complete.bioone.org/journals/BioScience on 25 Apr 2024
Terms of Use: https://complete.bioone.org/terms-of-use