Fact Sheet: Data to Knowledge to Action
Progress by Federal Agencies
November 12, 2013

Last year, the Obama Administration announced a series of new investments by Federal agencies in support of its Big Data Initiative. Following are some notable accomplishments to date.

National Science Foundation (NSF) and the Defense Advanced Research Projects Agency (DARPA): NSF funded a $10 million "Expeditions in Computing" project, and DARPA funded an XDATA grant to the University of California, Berkeley's "Algorithms, Machines, and People Laboratory" (AMPLab). Genomics projects in AMPLab have developed the SNAP alignment algorithm, which reduces processing costs from about $200 per genome to about $5 per genome, and have made advances in data storage formats and performance benchmarking for genomic data processing pipelines. The lab has also released Carat, a novel collaborative application to improve Smartphone battery life using machine-learning techniques. AMPLab is also supported by a number of industry sponsors.

Department of Energy (DOE): The DOE's Office of Science announced a $25 million award over 5 years to establish the Scalable Data Management, Analysis, and Visualization Institute (SDAV), which has already led to several important advances involving data management and indexing techniques for large and complex datasets. In the climate data domain, for example, these techniques have led to increases in the accuracy of seasonal hurricane predictions by more than 25 percent for peak-year storms originating off the continent of Africa.

National Aeronautics and Space Administration (NASA): Over the last two years, NASA has invested $9 million in data and information system research and development focusing on big-data management and data-mining algorithms. This investment has resulted in the development of several new technology platforms, including an Apache OpenClimate Workbench for data-model corroboration, an Amazon Web Services cloud-enabled collaboration environment for satellite radar data processing and mining, and a parallel web service for climate model output data analysis.

U.S. Geological Survey (USGS): The USGS, in collaboration with NSF, awarded more than $1 million through competitive proposals at the John Wesley Powell Center for Earth System Analysis and Synthesis in research areas ranging from understanding fluid-induced seismicity in

oil and gas production to developing a 4-dimensional digital crust of the North American continent.

National Institutes of Health (NIH): NIH's Big Data to Knowledge (BD2K) initiative is a multiyear, multimillion investment that is developing new approaches, standards, methods, tools, software, and competencies necessary for biomedical scientists to capitalize more fully on the big data being generated by the research community, including the large data sets used to understand the molecular and cellular changes that cause disease. For example, The Cancer Genome Atlas (TCGA), an NIH-funded program, is using large genomic data sets to map the genetic changes in more than 20 cancer types. Recently, TCGA researchers revealed that breast and ovarian cancers have genomic similarities that may have implications for treating these diseases. Researchers are also using big data to accrue patients to clinical trials. NIH's National Heart Lung Blood Institute-sponsored Cardiovascular Research Network (CVRN) is collaborating among its member health care delivery systems and associated electronic health records to assemble the largest contemporary cohort of more than 34,000 patients with atrial fibrillation. Through analysis of this very large cohort, CVRN researchers recently developed and validated a new stroke risk score that is superior to other scores and improves the selection and use of stroke-prevention strategies for the growing number of Americans with atrial fibrillation. They have also demonstrated the effectiveness and safety of the blood thinner warfarin, as well as medications used to control heart rate for atrial fibrillation.