Testimony of

Farnam Jahanian, Ph.D.
Assistant Director
Computer and Information Science and Engineering Directorate
National Science Foundation

Before the

Committee on Science, Space, and Technology
Subcommittee on Technology
And the
Subcommittee on Research
U.S. House of Representatives

April 24, 2013

Next Generation Computing and Big Data Analytics

Good morning, Chairman Massie and Chairman Bucshon, Ranking Members Wilson and Lipinski, and members of the Subcommittees. My name is Farnam Jahanian and I am the Assistant Director of the Computer and Information Science and Engineering (CISE) Directorate at the National Science Foundation (NSF). I also serve as the co-chair of the interagency Networking Information Technology Research and Development (NITRD) program, which provides a framework and mechanisms for coordination among 20 Federal agencies that support networking and information technology R&D.

I welcome this opportunity to highlight U.S. investments in advanced computing infrastructure and Big Data research and education – and how they are leading to transformative breakthroughs in all areas of science and engineering, as well as providing enormous societal benefit. The goal is to fund Big Data research at the frontiers of knowledge, to capitalize on the intellectual capacity of both early and experienced investigators in our Nation's academic and research institutions, and to foster partnerships across U.S. government agencies, the private sector and international organizations to effectively leverage these investments.

Overview

Innovative information technologies are transforming the fabric of society, and *data represent a transformative new currency for science, education, government and commerce.* Data are everywhere; they are produced in rapidly increasing volume and variety by virtually all scientific, educational, governmental, societal and commercial enterprises.[1]

Today we live in an "Era of Data and Information." This era is enabled by modern experimental methods and observational studies; large-scale simulations; scientific instruments, such as telescopes and particle accelerators; Internet transactions, email, videos, images, and click streams; and the widespread deployment of sensors everywhere – in the environment, in our critical infrastructure, such as in bridges and smart grids, in our homes, and even on our clothing! Consider this fact: every day, 2.5 quintillion bytes of data are generated – so much that 90% of the data in the world today has been created in the last two years alone[2].

When we talk about Big Data, however, it is important to note that it is not just the enormous volume of data that needs to be emphasized, but also the heterogeneity, velocity, and complexity that collectively create the science and engineering challenges we face today.

In December 2010, the President's Council of Advisors on Science and Technology (PCAST) published a report to the President and Congress entitled, *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*[3]. In that report, PCAST pointed to the research challenges involved in large-scale data management and analysis and the critical role of Networking and Information Technology (NIT) in moving from data to knowledge to action, underpinning the Nation's future prosperity, health and security.

Through long-term, sustained investments in foundational computing, communications and computational research, and the development and deployment of large-scale facilities and cyberinfrastructure, Federal agency R&D investments over the past several decades have both helped generate this explosion of data as well as advance our ability to capture, store, analyze and use these data for societal benefit. More specifically, we have seen fundamental advances in machine learning, knowledge representation, natural language processing, information retrieval and integration, network analytics, computer vision, and data visualization, which together have enabled Big Data applications and systems that have the potential to transform all aspects of our lives.

These investments are already starting to pay off, demonstrating the power of Big Data approaches across science, engineering, medicine, commerce, education, and national security, and laying the foundations for U. S. competitiveness for many decades to come. Let me offer three examples:

---

[1] "Dealing with Data," Science Magazine, Volume 331, February 11, 2011.

[2] See http://www-01.ibm.com/software/data/bigdata/.

[3] For example, some new technologies include smartphones, eBook readers, and game consoles; corporate data-centers, cloud services and scientific supercomputers; digital photography and photo editing, MP3 music players, streaming media, GPS navigation; robot vacuum cleaners, adaptive cruise control in cars and real-time control systems in hybrid vehicles, robot vehicles on and above the battlefield; the Internet and the World Wide Web; email, search engines, eCommerce, and social networks; medical imaging, computer-assisted surgery, and large-scale data analysis enabling evidence-based healthcare and the new biology; and rapidly improving speech recognition. Our world today relies to an astonishing degree on systems, tools, and services that belong to a vast and still growing domain known as NIT.

- Today's homes account for more than 20 percent of the total energy consumption in the Nation,[4] and about half of that energy is consumed for heating and cooling[5]. Each degree cooler a house is kept in the winter or each degree warmer in the summer can mean energy savings of 20%, translating to $200 to $300 in lower energy bills per year – not to mention fewer power plants built and lower carbon emissions[6]. A team of researchers has pioneered an intelligent thermostat that uses machine learning to transform home heating and cooling. At first, a person may set the thermostat four times in one day — upon getting up, going to work, getting back from work, and going to bed. The thermostat uses those settings daily, but then adapts to further changes. If a person is out of town every other Monday on business, for instance, the thermostat's sensors coupled with machine learning algorithms detect the lack of activity and switch to an "auto away" setting for lower energy use.
- Collectively, Americans spent nearly 630,000 years – 5.52 billion hours – stuck in traffic in 2011, at a cost of $121.2 billion in fuel, maintenance, and lost productivity[7]. A number of regional ventures – in Los Angeles, the Bay Area, northern New Jersey, and here in the Washington, DC, region – are integrating heterogeneous data sources such as road sensors, traffic cameras, individuals' GPS devices, etc., to develop principles and methods that go beyond real-time traffic data and allow us to do inference over entire cities[8,9,10]. The aim is to identify hotspots and traffic-sensitive directions to drivers well before a potential traffic jam materializes. In Los Angeles, for example, city planners have synchronized every one of the 4,500 traffic signals across 469 square miles of downtown, and they use a system of sensors in the road measuring traffic flow, live traffic cameras, and a centralized computing platform leveraging data mining and machine learning to make constant, automated adjustments and keep cars running as smoothly as possible[11]. Under this system, the average speed of traffic across the city has increased by 16%, with delays at major intersections down 12%[11].
- Breast cancer is the most common cancer among American women, except for skin cancers, and nearly 40,000 women die from the disease each year[12]. By extending image analysis techniques to hundreds of breast cancer biopsy images, researchers were able to identify a small subset of cellular features – out of over 6,000 possible features – that was predictive of survival time among breast cancer patients[13]. This feature set was unique: pathologists had not previously identified it as relevant to cancer prognosis, and the information and insight was above and beyond that of many existing standard measures of cancer severity, such as grade, protein markers, tumor size, and lymph node status. The

[4] U.S. Energy Information Administration (EIA). Annual Energy Outlook 2013: Market Trends — U.S. Energy Demand. See http://www.eia.gov/forecasts/aeo/MT_energydemand.cfm#indus_comm.
[5] See http://www.energystar.gov/index.cfm?c=heat_cool.pr_hvac.
[6] See http://www.nest.com/saving-energy/.
[7] "2012 Urban Mobility Report," Texas A&M Transportation Institute, December 2012, http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-report-2012.pdf.
[8] See http://www-03.ibm.com/press/us/en/pressrelease/34261.wss.
[9] "Trapping 'Big Data' to Fill Potholes: Start-Ups Help States and Municipalities Track Effects of Car Speeds, Other Variables on Traffic," *The Wall Street Journal*, June 12, 2012.
[10] See http://www.cattlab.umd.edu/?portfolio=ritis.
[11] "To Fight Gridlock, Los Angeles Synchronizes Every Red Light," *The New York Times*, April 1, 2013.
[12] See http://www.cancer.org/cancer/breastcancer/detailedguide/breast-cancer-key-statistics.
[13] Beck, A.H., Sangoi, A.R., Leung, S., Marinelli, R.J., Nielsen, T.O., van de Vijver, M., West, R.B., van de Rijn, M., and Koller, D. 2011. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**: 108ra113.

feature set also resulted in an unexpected finding: the features that were the best predictors of patient survival were not from the cancer tissue itself, but rather from adjacent tissue – something that had gone undetected by pathologists and clinicians. These new discoveries will allow clinicians to better understand the genesis and morphology of breast cancer, enabling personalized treatments that aim to improve survival times among patients.

These kinds of breakthroughs are catalyzing a profound transformation in the culture and conduct of scientific research, requiring new methods to derive knowledge from the data; new infrastructure to manage, curate and serve data to communities; new approaches to education and training; and, finally, new types of collaborations of multi-disciplinary teams and communities that have the potential to solve today's most complex science and engineering challenges. Through the NITRD program and its member agencies, the U.S. Government has responded to this Big Data revolution with a bold, sustainable, and comprehensive approach. These agencies support the development of cutting-edge tools and algorithms for all aspects of the data lifecycle, all with the aim of helping scientists to transcend the logistics of data handling and to focus on scientific discovery. Many of these advances leverage capabilities in high performance computing, which shares many of the same technology challenges as data-intensive science and has become a critical tool in analyzing and interpreting scientific data. As part of this overall effort, NSF will also address data access policies to better enable science communities to work together to address science challenges.

Why is Big Data Important?

Big Data is important to all facets of the discovery and innovation ecosystem, including the Nation's academic, government, industrial, entrepreneurial, and investment communities.

First, insights and more accurate predictions from large and complex collections of data have important implications for the economy. Access to information is transforming traditional businesses and is creating opportunities in new markets. Further, Big Data is driving the creation of new IT products and services based on business intelligence and data analytics, and is boosting the productivity of firms that use it to make better decisions and identify new business trends.

Second, advances in our ability to store, integrate, and extract meaning and information from data are critical to accelerate the pace of discovery in almost every science and engineering discipline. From new insights about protein structure, biomedical research and clinical decision-making, and climate modeling, to new ways to mitigate and respond to natural disasters, and develop new strategies for effective learning and education – there are enormous opportunities for data-driven discovery.

Third, Big Data will be a key component to solving the Nation's most pressing challenges – in education, healthcare, medicine, energy, transportation, commerce, disaster prevention and mitigation, and cyber and national security – yielding enormous societal benefit and laying the foundations for U.S. competitiveness.

There are enormous opportunities to harness the increasingly large-scale and diverse data sets, to **extract knowledge** from them, to provide powerful new approaches to **drive discovery** and **decision-making,** and to make increasingly **accurate predictions** and move toward deeper understandings of **causal relationships** based on advanced data analysis. These advances will lead to new innovations, job creation, and long-term economic development and prosperity.

New Era of Data and Information

We are now in a new era of observation as well as a new era of data and information[14]. Today, our scientific tools provide an unprecedented sophistication, resolution and scope. Within the NSF-supported research context, these tools can reach the outer edges of the universe as well as dig deep into the tiniest phenomenon. They can transcend all scales, from the molecular and genetic to the organismal and social. Our ability to gain new knowledge would be impossible without these capabilities.

At the one extreme, advanced research infrastructure – large-scale facilities, experimental tools, and cyberinfrastructure – enables new knowledge at the far reaches of the cosmos. For example, the Large Synoptic Survey Telescope (LSST), jointly funded by NSF and the Department of Energy (DOE), is expected to enter full operations for a 10-year survey beginning in January 2022. This telescope will probe mysterious dark matter and dark energy, map small objects in the solar system, particularly near-Earth asteroids, and detect transient optical events such as novae or supernovae – generating 30 terabytes of data each night that span billions of light years.

At the other end of the spectrum, we see research that explores phenomena at nano, pico, and femto scales. NSF is supporting IceCube, a particle detector drilling three kilometers deep into the Antarctic in search of interactions of a nearly massless subatomic particle called a neutrino, which could reveal the new physical processes associated with the enigmatic origin of the highest energy particles in nature.

Likewise, we can explore the properties of a single neuron in the brain, in response to different stimuli and stresses – and then integrate data about millions of such neurons to begin to understand not just the underlying biology, but also the interconnections of the brain that give rise to the psychology of the human mind.

We have new opportunities with technology as well. For example, with the advent of the Internet and mobile devices, “citizen science” is increasingly leading to new knowledge. Take the iPad, for example. One can envision a middle-school child anywhere in the world accessing data in real time that comes out of a hundred-million-dollar facility funded by NSF. The child could participate in an experiment in which he or she actually gathers data where he or she lives. In another example, last year, citizens distributed across the U.S. using a software application, FoldIt, together resolved the detailed molecular structure of an enzyme that is believed to play a critical role in the spread of the AIDS virus – a breakthrough that had confounded scientists for decades[15].

These advances are not only experimental in nature, but also, combined with advances in computational hardware and software to capture and make sense of the data, they are equally computational and theoretical.

---

[14] “Vision from the National Science Foundation,” A presentation by Subra Suresh, Director, at the National Academy of Sciences Symposium on Science, Innovation, and Partnerships for Sustainability Solutions, Washington, DC, May 16, 2012: http://www.nsf.gov/news/speeches/suresh/12/ss120516_nas_symposium.jsp.

[15] See http://cosmiclog.nbcnews.com/_news/2011/09/16/7802623-gamers-solve-molecular-puzzle-that-baffled-scientists.

What Does this Mean for Scientific Discovery?

Data are motivating a profound transformation in the culture and conduct of scientific research. Data-driven discovery is revolutionizing scientific exploration and engineering innovations. This approach has been called the "fourth paradigm," in contrast to the three earlier approaches to scientific research: empirical observation and experimentation; analytical/theoretical approaches; and computational science and simulation. Such a data-enabled approach to science complements these other earlier approaches, but has the promise to revolutionize science even further.

Indeed, the fourth paradigm has led to improved hypotheses and faster insights. From new knowledge about protein structure paving the way to advances in biomedical research and clinical decision-making, to new ways to mitigate and respond to natural disasters, to new strategies for effective learning and education, there are enormous opportunities for this new form of scientific discovery called "data-driven discovery"!

Data access and analysis are already having enormous impacts. The opportunities for the future are immense. Imagine, if you can:

- Complete health/disease/genome/environmental knowledge bases that enable biomedical discovery and patient-centered therapy. The data can be mined to spot unwanted drug interactions or to predict onset of diseases.
- Companies that, by linking together finance, human resources, supply chain, customer management systems, can use data mining techniques to get a complete picture of their operations – to identify new business trends, operate more efficiently, and improve forecasting.
- Accurate high-resolution models that support forecasting and management of increasingly stressed watersheds and ecosystems.
- Consumers that all have the information they need to make optimal energy consumption decisions in their homes and cars.
- Accurate predictions of natural disasters, such as earthquakes, hurricanes, and tornadoes, that enable life-saving and cost-saving preventative actions.
- A cyber-enabled world that is safe, secure, and private, enabling assured use of our critical infrastructure and on-line commerce.
- Students and researchers who have access to intuitive tools to view, understand, and learn from publicly-available, large scientific data sets on everything from genome sequences to astronomical star surveys, from public health databases to particle accelerator simulations, and teachers and educators who use student performance analytics to improve learning and enhance assessment.

Many R&D challenges remain. Below is a list of Big Data hard problems that the research community is addressing:

- *Many data sets are too poorly organized to be usable.* Research must come up with new techniques to better organize and retrieve data.
- *Many data sets are comprised of unstructured data.* Research must develop new data mining tools and/or machine learning techniques to make these data usable. Opening government data to all is one of the first steps to spur innovation.
- *Many data sets are heterogeneous in type, structure, semantics, organization, granularity, and*

accessibility. Research must find novel ways to integrate and customize access to federated data; research must find ways to make heterogeneous data more interoperable and usable.

- *The utility of data is limited by our ability to interpret and use it.* Research must find better usability techniques to extract and visualize actionable information. Research needs to discover new techniques for evaluating and showing results.
- *More data are being collected than we can store.* With the right data infrastructure, practitioners could analyze data as it becomes available; they could immediately decide what to archive and what to discard.
- *Many data sets are too large to download or send over today's Internet.* With the right data infrastructure, practitioners could analyze the data wherever it resides, instead of sending it to data centers.
- *Large and linked datasets may be exploited to identify individuals.* Research on privacy protection and Big Data is critical; new techniques and analysis could have "built-in" privacy preserving characteristics.

The landscape of open research and development challenges is vast. American scientists must rise to the occasion and seize the opportunities afforded by this new, data-driven revolution. The work we do today will lay the groundwork for new enterprises and long-term economic prosperity.

U.S. Government Response to Big Data R&D Challenges

The December 2010 PCAST report – *Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology*[16] – recommended several actions to take advantage of Big Data opportunities.

The potential impacts and outcomes for the Nation are huge – on the economy, the pace of discovery in science and engineering, national security, healthcare, education, energy efficiency, real-time labor market information, to name a few national priorities. The Office of Science and Technology Policy (OSTP) in the Executive Office of the President (EOP) responded to these recommendations, in part, by chartering a Big Data Senior Steering Group (BDSSG) that would focus on Big Data R&D under the NITRD umbrella. Currently, NSF and the National Institutes of Health (NIH) co-chair the SSG, and membership is comprised of representatives from the science agencies, such as NSF, NIH, the National Institute of Standards and Technology (NIST), DOE Office of Science, and National Aeronautic and Space Agency (NASA), as well as Departments of Defense, Health and Human Services, Treasury, and Commerce (National Oceanic and Atmospheric Administration).

Over the course of the year following its establishment, the BDSSG inventoried existing Big Data programs and projects across the agencies and began coordinating their efforts in four main areas: investments in Big Data core techniques and technologies; education and workforce; domain cyberinfrastructure; and challenges and competitions.[17] Other critical areas for Big Data were also identified, including privacy issues, open access to government data, and partnerships with industry and not-for-profits.

---

[16] President's Council of Advisors on Science and Technology (PCAST). *Report to the President and Congress: Designing a Digital Future: Federally Funding Research and Development in Networking and Information Technology.* December 2010. Executive Office of the President. http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf.

[17] See http://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title.

On March 29, 2012, the Administration launched the National Big Data Research & Development Initiative. Led by OSTP, this initiative seeks to greatly improve the tools and techniques used for Big Data analysis and the human capital needed to move data to knowledge to action.

Examples of agency efforts that are well aligned with this initiative include[18]:

- The Department of Defense (DOD) launched "Data to Decisions," a series of programs that are harnessing and utilizing massive data in new ways, and bringing together sensing, perception, and decision support to (a) make truly autonomous systems that can maneuver and make decisions on their own, and (b) improve situational awareness to help warfighters and analysts and provide increased support to operations.
- Defense Advanced Research Projects Agency (DARPA) announced the XDATA program to develop computational techniques and software tools for analyzing large volumes of data, both semi-structured (tabular, relational, categorical, and meta-data) and unstructured (text documents, message traffic), particularly in the context of targeted defense applications.
- NIH made available 200 terabytes of data from the 1000 Genomes Project in the cloud through Amazon Web Services (AWS), constituting the world's largest set of data on human genetic variation and enabling genome-wide association studies to understand the genetic contribution to disease.
- Through its Scientific Discovery Through Advanced Computing (SciDAC) program, the DOE Office of Science unveiled a $25 million Scalable Data Management, Analysis, and Visualization Institute, spanning six national laboratories and seven universities, that is developing "new and improved tools to help scientists manage and visualize data" and supporting the scientists in their use.
- The U.S. Geological Survey (USGS) John Wesley Powell Center for Analysis and Synthesis issued awards focused on improving our understanding of earth system science through Big Data, including "species response to climate change, earthquake recurrence rates, and the next generation of ecological indicators."

In addition, a number of other agencies are participating, including the Office of the Director of National Intelligence (ODNI) through its Intelligence Advanced Research Projects Activity (IARPA), the Department of Homeland Security (DHS), Department of Veterans Affairs (VA), Food and Drug Administration (FDA), National Archives and Records Administration (NARA), and National Security Agency (NSA)[19].

Anchoring this coordinated effort, NSF and NIH released a joint solicitation, "Core Techniques and Technologies for Advancing Big Data Science & Engineering," or "BIGDATA." This program aims to extract and use knowledge from collections of large data sets in order to accelerate progress in science and engineering research. Foundational research advances in data management, analysis and collaboration promise to change paradigms of research and education, and develop new approaches to addressing national priorities. The goal is new capabilities for data-intensive and data-enabled science to create actionable information that leads to timely and more informed decisions and actions. It will both help to accelerate discovery and innovation in all sciences, engineering and education, as well as support their transition into practice to benefit society.

[18] See http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf.
[19] See http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final_2.pdf.

As we enter the second year of the Big Data Initiative, the BDSSG is encouraging multiple stakeholders, including federal agencies, private industry, academia, state and local government, non-profits and foundations, to develop and participate in Big Data innovation projects across the country. The BDSSG is planning an event to announce these new projects and to emphasize the importance of building multi-stakeholder partnerships in all areas of Big Data science and engineering across the country.

Collectively, these coordinated activities led PCAST to conclude in a January 2013 update to its 2010 report on Networking and Information Technology R&D, "Federal agencies have made significant progress in supporting R&D for data collection, storage, management, and automated large-scale analysis ('big data')"[20]. PCAST found that Big Data remains a "critical focal point" in 2012 and beyond, and recommended continued emphasis and coordination.

Coordination of Federal Big Data Investments

NSF coordinates its Big Data R&D activities with other Federal agencies, including the NIH, NASA, DOE Office of Science, DARPA, and many others, through the following "mission-bridging" mechanisms:

- The National Science and Technology Council's NITRD Sub-Committee, of which I am co-chair, has played a prominent role in the coordination of the Federal government's Big Data research investments.
- Under the NITRD umbrella, the BDSSG coordinates Big Data R&D across the member agencies by 1) promoting new science and accelerating the progress of discovery through large, heterogeneous data; 2) exploiting the value of Big Data to address areas of national needs, agency missions and societal and economic importance; 3) supporting responsible stewardship and sustainability of Big Data resulting from federally-funded research; and 4) developing and sustaining the infrastructure needed to advance data science.
- Under the auspices of the NITRD program and the BDSSG, various participating agencies collectively sponsor workshops, develop joint programs, and invest in other activities that leverage their complementary missions.

Most multi-disciplinary, cross-agency fields of NIT inquiry in which NSF makes investments are managed in a similar way (cybersecurity, cyberphysical systems, etc.).

A Framework for NSF Investments

At NSF, it is expected that improvements in access, manipulation, data mining, management, analysis, sharing and storing of Big Data will provide new insights, change paradigms of research and education, and create new approaches to addressing national priorities. NSF has identified four major investment areas that address these challenges and promise to serve as the foundations of a comprehensive, long-term agenda. They are:

1. *Foundational Research in all Areas of Science and Engineering*: Advance the core scientific and technological means of managing, analyzing, visualizing, and extracting useful information from large, diverse, distributed and heterogeneous data sets. Facilitate the development of new data analytic tools and algorithms; scalable, accessible, and sustainable data infrastructure; and large-

[20] See http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd2013.pdf.

scale integrated statistical modeling. This research aims to, among other things, advance our knowledge and understanding of mathematical and physical systems, the science of learning, and human and social processes and interactions.

2. *Cyberinfrastructure*: Provide science, engineering, and education with a comprehensive data infrastructure that will enable the capture, management, curation, analysis, interpretation, archiving and sharing of data of unprecedented scale, parallelism, and complexity in a manner that will stimulate discovery in all areas of inquiry, and from all instruments and facilities, ranging from campus- to national-level investments.

3. *Education and Workforce Development*: Ensure that the future, diverse workforce of scientists, engineers, and educators is equipped with the skills to make use of, and build upon, the next generation of data analytics, modeling, and cyberinfrastructure. Support new approaches to K-16 teaching and learning that takes advantage of new cyberinfrastructure and data-driven approaches, leading to a national learning laboratory.

4. *Scientific Community Building and Governance*: Support transformative interdisciplinary and collaborative research in areas of inquiry stimulated by data through the development of robust, shared resources and partnerships across diverse communities. This development must acknowledge the new challenges surrounding reproducibility, storage, curation, and open dissemination of scientific data in all its forms, and recognize its importance for accelerating fundamental discovery, interdisciplinary research, and innovation in society. Open and shared data can enable new approaches for communities to address complex problems in science and engineering.

NSF is developing a bold and comprehensive approach for this new data-centric world – from fundamental mathematical, statistical and computational approaches needed to understand the data, to infrastructure at a national and international level needed to support and serve our communities, to policy enabling rapid dissemination and sharing of knowledge. Together, these activities will accelerate scientific progress, create new possibilities for education, enhance innovation in society and be a driver for job creation. Everyone will benefit from these activities.

*Big Data Research***:** NSF continues to cast a wide net and let the best ideas surface, rather than pursuing a prescriptive research agenda. It engages the Big Data research community in developing new fundamental ideas, which are then evaluated by the best researchers through the peer review process. This process, which supports the vast majority of unclassified researchers in the United States, has led to innovative and transformative results. Indeed, NSF investments today leverage a long history of Foundation-wide support for data analytics and computational science.

In October 2012, just six months after the Big Data Initiative launch, NSF and NIH announced nearly $15 million in new Big Data fundamental research projects, the first step toward realizing the goals to advance the foundational science and engineering of Big Data. We received over 450 proposals in response to the joint solicitation, spanning a broad spectrum of R&D activities from new scientific techniques for Big Data management, to new data analytic approaches, to e-science collaborations with possible future applications in a variety of fields, such as medicine, physics and economics.

As an example of the new awards that we made, consider the work of Eli Upfal of Brown University, who is leading a project on data analytics. Dr. Upfal and his team plan to develop mathematically well-

founded algorithmic and statistical techniques for analyzing large-scale, heterogeneous and so-called "noisy" data. The resultant algorithms will be tested on extensive cancer genome data, contributing to better health and the development of new health information technology.

A second example is an award led by Christos Faloutsos of Carnegie Mellon University and Nikolaos Sidiropoulos of the University of Minnesota, aiming to develop theory and algorithms to tackle the complexity of language processing and to develop methods that approximate how the human brain works in processing language. The research also promises better algorithms for search engines, new approaches for understanding brain activity, and better recommendation systems for the retail sector.

NSF also funds center-scale activities. One project announced at the Big Data Initiative launch in March 2012 was a $10 million award to researchers at the University of California, Berkeley, under the NSF Expeditions in Computing program. The research team will integrate algorithms, machines, and people (AMP) to turn data into knowledge and insight. The objective is to develop new scalable machine-learning algorithms and data management tools that can handle large-scale and heterogeneous datasets (spanning data generated by computers, sensors and scientific instruments; media such as images and video; and free-form tweets, text messages, blogs and documents), novel datacenter-friendly programming models, and an improved computational infrastructure. The team is focusing on key applications of societal importance, including cancer genomics and personalized medicine; large-scale sensing for traffic prediction, environmental monitoring, and urban planning; and network security. Although the project is only in its first year, it has already led to significant contributions, including open source high performance machine learning software, called Spark, that was featured on Siliconangle's list of Top 5 Open Source Projects in Big Data[21].

Aside from its investments in fundamental research, NSF also supports development activities beyond the stage of research prototypes through its Small Business Innovative Research (SBIR) and Small Business Technology Transfer (STTR) programs, as well as the Innovation Corps Teams Program (I-Corps), which identifies NSF-funded researchers who will receive additional support – in the form of mentoring and funding – to accelerate innovation that can attract subsequent third-party funding.

Big Data Education and Workforce Development: A report by the McKinsey Global Institute[22] estimated, "By 2018 the United States alone faces a shortage of 140,000 to 190,000 people with analytical expertise and 1.5 million managers and analysts with the skills to understand and make decisions based on the analysis of big data."

At NSF, investments in Big Data research are accompanied by investments in Big Data education and workforce development. Research undertaken in academia not only engages some of our nation's best and brightest researchers, but because these researchers are also teachers, new generations of students are exposed to the latest thinking from the people who understand it best. And when these students graduate and move into the workplace, they are able to take this knowledge and understanding with them. Moreover, faculty members in this dual role of researchers and teachers have incentives to write textbooks and develop other learning materials that allow dissemination of their work to a wide audience, including teachers and students nationwide.

---

[21] See http://siliconangle.com/blog/2013/02/04/top-5-open-source-projects-in-big-data-breaking-analysis/.

[22] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A.H. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute: http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation.

Over the years, NSF has supplemented its investments in Big Data by giving additional funding to researchers who were willing to bring undergraduates into their labs through the Research Experiences for Undergraduates (REU) program. This program gives many undergraduate students their first hands-on experiences with real science and engineering research projects. In addition, NSF funds up and coming young investigators through the prestigious CAREER program that offers awards in support of junior faculty who are exemplary teacher-scholars. These awardees conduct outstanding research, develop and implement excellent education plans, and integrate education and research within the context of the mission of their organizations.

More recently, NSF used its Integrative Graduate Education and Research Traineeship (IGERT), mechanism to educate and train researchers in data-enabled science and engineering, including 1) core techniques and technologies for advancing big data science and engineering; 2) analyzing and dealing with challenging computational and data-enabled science and engineering (CDS&E) problems; and 3) researching, providing, and using the cyberinfrastructure that makes cutting-edge CDS&E research possible in any and all disciplines.

Finally, the move from face-to-face to online and blended learning, which allows for learning anywhere, anytime, and by anyone, is rapidly transforming education into a data-rich domain. By collecting, analyzing, sharing, and managing the data collected through monitoring learners' use of technology, we can begin to understand how people learn. The result is an ability to advance understanding of how to use technologies and integrate them into new learning environments so that their potential is fulfilled. An anticipated cross-disciplinary effort is participation in an Ideas Lab to explore ways to use Big Data to enhance teaching and learning effectiveness[23].

Computational and Data Cyberinfrastructure: NSF has been an international leader in high-performance computing (HPC) deployment, application, research, and education for almost four decades. With the accelerating pace of advances in computing and related technologies, coupled with the exponential growth and complexity of data for the science, engineering, and education enterprise, new approaches are needed to advance and support a comprehensive advanced computing infrastructure that facilitates transformational ideas using new paradigms and approaches. The goal is a complementary, comprehensive, and balanced portfolio of advanced computing infrastructure and programs for research and education to support multidisciplinary computational and data-enabled science and engineering that in turn support the entire scientific, engineering, and education community.

Below, I give a few examples, illustrating the range and scope of today's computational and data cyberinfrastructure.

Advanced Computational Infrastructure. Last October, NSF inaugurated "Yellowstone," one of the world's most powerful supercomputers, based at the National Center for Atmospheric Research (NCAR) in Cheyenne, WY, and this past March, NSF dedicated two advanced computational facilities, "Blue Waters," located at the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, and "Stampede," headquartered at the Texas Advanced Computing Center at the University of Texas at Austin.

The three systems will provide the nation's research community with unprecedented computational

[23] See http://www.nsf.gov/pubs/2012/nsf12060/nsf12060.jsp.

capabilities, further enhancing the already potent union between technology and the human mind, offering the opportunity to better test and advance great scientific ideas. Each strengthens the other.

Consider, for instance, the extraordinary capability of just one of these systems, Blue Waters. This computing system is equipped with more than 1.5 petabytes of memory, enough to store 300 million images from your digital camera; more than 25 petabytes of disk storage, enough to store all of the printed documents in all of the world's libraries; and up to 500 petabytes of tape storage, enough to store 10% of all the words spoken in the existence of humankind. If you could multiply two numbers together every second, it would take you 32 million years to do what Blue Waters does each second.

By all measures, these computers, with their high speed and storage capacity, and the ability to produce high-resolution simulations, will have a significant impact on the pace of scientific progress. They will expand the range of data-intensive computationally-challenging science and engineering applications that can be tackled with current national resources. They will allow today's scientists to better understand the workings of Earth and beyond, for example, by helping to trace the evolution of distant galaxies, by providing data that contribute to the design of new materials, and by supporting researchers trying to forecast tornadoes, hurricanes and other severe storms, and even space weather, such as solar eruptions.

Sloan Digital Sky Survey. The Sloan Digital Sky Survey (SDSS) – one of the most ambitious and influential surveys in the history of astronomy – was launched in 2000. It collected more data in its first few weeks of operation than had been amassed in the entire previous history of astronomy. Within a decade, over 140 terabytes of information were collected, representing 35% of the sky. The final dataset includes 230 million celestial objects detected in 8,400 square degrees of imaging, and spectra of 930,000 galaxies, 120,000 quasars, and 225,000 stars.

NSF and NASA jointly support this project, together with the Alfred P. Sloan Foundation, DOE, and international partners in Japan and Germany. The protocols developed in this cyberinfrastructure underpin astronomical archives the world over, including the Panoramic Survey Telescope and Rapid Response System project, now about to issue its first data release, and the planned LSST, which will produce approximately the same amount of data as the first decade of SDSS, every single night of its operation. A recent survey of literature citations has listed SDSS as the most influential, most cited observatory.

iPlant. iPlant, a plant science cyberinfrastructure collaborative led by the University of Arizona, utilizes new computer, computational science and cyberinfrastructure solutions to address an evolving array of grand challenges in the plant sciences. This center is a community-driven effort, involving plant biologists, computer and information scientists and engineers, as well as experts from other disciplines, all working in integrated teams.

An important grand challenge that iPlant is attempting to address (i.e., bridging the divide between genotype and phenotype) involves integrating many types of data, including DNA sequences, trait data, and geographical occurrence information. The latter is particularly useful, as a large proportion of variation in phenotype is due to environmental influences. iPlant's computational capabilities have enabled species range models for over 88,000 species across North and South America; this will help set the baseline for biodiversity. Catalyzed in part by iPlant efforts, large agricultural datasets have been released from Monsanto and Syngenta for use in modeling crop performance under existing and predicted climate regimes.

NEON. A Major Research Equipment and Facilities Construction (MREFC) project, the National Ecological Observatory Network (NEON) is a continental-scale observatory designed to gather and provide 30 years of ecological data[24]. By making all its data freely available, NEON is providing infrastructure to facilitate hypothesis-driven basic biological and ecological research, enabling the development of a predictive understanding of the direct effects and feedbacks between environmental change and biological processes.

NEON is unique in its continental reach and longitudinal data collection over several decades, delivering and curating a multimodal stream of never-before-available regional and continental scale ecological datasets to the scientific community and the Nation. Just as NEON researchers will benefit from access to data from federal agency networks, federal agencies will benefit from the techniques, sensors and knowledge gained through NEON-enabled activities. NEON's systems engineering-guided design, construction and operations plans, and formalized transition to operations are defining a new standard for research infrastructure deployment and operations. Other Federal Agencies (e.g., US Department of Agriculture, NASA) and international groups (e.g., European Union, Australian Terrestrial Observing Network) are emulating the standards established by NEON, Inc.

Big Data Community Building and Partnerships: NSF seeks to enable research communities to develop new visions, teams, and capabilities dedicated to creating new, large-scale, next-generation data resources and relevant analytic techniques to advance fundamental research across all areas of science and engineering as well as to transition discoveries into practice.

An example of successful community building is EarthCube, which focuses on the development of community-guided cyberinfrastructure to integrate big data across geosciences and ultimately change how geosciences research is conducted. Integrating data from disparate locations and sources with eclectic structures and formats that has been stored as well as captured in real time will expedite the delivery of geoscience knowledge.

In 2013 EarthCube released a solicitation to engage all stakeholders, from geoscientists to computer scientists, industry, academia and government, to build on the momentum and enthusiasm generated in the past year. The different components of the call will allow these stakeholders to participate in developing the next stage of EarthCube. This includes coordination networks to help geoscientists develop standards and policies, demonstrations of promising technologies for integrating across geosciences data, and activities to plan innovative architectures across the whole enterprise.

In recent years, the transition of Big Data research results to the private sector has helped bring innovative Big Data solutions and technologies to the marketplace, fuel job growth, and promote economic growth and improved health and quality of life. Some of the examples I noted earlier in this testimony speak to this transition of discoveries into practice. By promoting strong connections between academia and industry, NSF further enhances its research portfolio in Big Data with foundational concepts and new ideas that are directly relevant to the commercial sector.

Data Access Policy

As investments in research, education, and cyberinfrastructure further Big Data science and engineering, there is also recognition of the importance of enabling rapid dissemination and sharing of new

[24] See http://www.neoninc.org/.

knowledge, tools, and expertise. In February 2013, OSTP issued a memo directing Federal agencies to develop plans to support increased access to results from federally-funded research[25]. The memo focuses on two particular elements. First, peer-reviewed publications should be stored for "long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the Federal research investment." Second, digitally formatted scientific data resulting from unclassified research "should be stored and publicly accessible to search, retrieve and analyze."

Summary

In my testimony today, I have tried to illustrate how we find ourselves in the midst of a new era of data and information, driven by innovative information technologies that are at the center of an ongoing societal transformation in terms of how we live, work, learn, play, and communicate. I have outlined the enormous volume, velocity, heterogeneity, and complexity of data being generated through modern experimental methods and observational studies, large-scale simulations, Internet transactions, and the pervasive use of sensor-based technologies. I have indicated how the U.S. government has responded to this new era through a coordinated, multi-agency National Big Data Research & Development Initiative, and, in particular, described NSF's role in support of Big Data research, education, and cyberinfrastructure. Finally, I have shared with you how these investments are starting to pay off; much progress has been made and, in turn, the power of Big Data approaches is evident in nearly all sectors of society and across all national priority areas. With robust sustained support for fundamental research, education, and infrastructure in the area of Big Data in both the executive and legislative branches of our government, there is a unique and enormous opportunity to position the Nation at the forefront of advances in science and engineering, job creation, and economic development for decades to come. This concludes my remarks. I appreciate the opportunity to have this dialogue with members of the Subcommittees on these very important topics, and I would be happy to answer any questions at this time.

[25] See http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

Biographical Sketch

FARNAM JAHANIAN

Farnam Jahanian is the NSF Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate. Dr. Jahanian also serves as Co-Chair of the NITRD Subcommittee of the NSTC Committee on Technology, providing overall coordination for the NIT activities of 20 Federal agencies.

At NSF, Dr. Jahanian guides the CISE Directorate in its mission to uphold the Nation's leadership in computer and information science and engineering through its support for foundational and transformative advances that are key drivers of economic competitiveness and critical in achieving our national priorities. CISE supports ambitious long-term research and innovation, the creation and provisioning of cutting-edge cyberinfrastructure and tools, broad interdisciplinary collaborations, and education and training of the next generation of scientists and information technology professionals with skills essential to success in the increasingly competitive, global market.

Dr. Jahanian is on leave from the University of Michigan, where he holds the Edward S. Davidson Collegiate Professorship in Electrical Engineering and Computer Science. Previously, he served as Chair for Computer Science and Engineering from 2007 – 2011 and as Director of the Software Systems Laboratory from 1997 – 2000. Over the last two decades at the University of Michigan, Dr. Jahanian led several large-scale research projects that studied the growth and scalability of the Internet infrastructure and which ultimately transformed how cyber threats are addressed by Internet Service Providers. His work on Internet routing stability and convergence has been highly influential within both the network research and the Internet operational communities. This work was recognized with an ACM SIGCOMM Test of Time Award in 2008. His research on Internet infrastructure security formed the basis for the successful Internet security services company Arbor Networks, which he co-founded in 2001. He served as Chairman of Arbor Networks until its acquisition in 2010.

The author of over 100 published research papers, Dr. Jahanian has served on dozens of national advisory boards and government panels. He has received numerous awards for his research, teaching, and technology commercialization activities. He has been an active advocate for economic development efforts over the last decade, working with entrepreneurs, and frequently lecturing on how basic research can be uniquely central to an innovation ecosystem that drives economic growth and global competitiveness. In 2009, he was named Distinguished University Innovator at the University of Michigan.

Dr. Jahanian holds a master's degree and a Ph.D. in Computer Science from the University of Texas at Austin. He is a Fellow of the *American Association for the Advancement of Science (AAAS)*, the *Association for Computing Machinery (ACM)*, and the *Institute of Electrical and Electronic Engineers (IEEE)*.