AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

NEXT GENERATION COMPUTING AND BIG DATA ANALYTICS

JOINT HEARING

BEFORE THE

SUBCOMMITTEE ON RESEARCH &
SUBCOMMITTEE ON TECHNOLOGY

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
HOUSE OF REPRESENTATIVES

ONE HUNDRED THIRTEENTH CONGRESS

FIRST SESSION

WEDNESDAY, APRIL 24, 2013

Serial No. 113–22

Printed for the use of the Committee on Science, Space, and Technology

Available via the World Wide Web: http://science.house.gov

U.S. GOVERNMENT PRINTING OFFICE
80–561PDF WASHINGTON : 2013

For sale by the Superintendent of Documents, U.S. Government Printing Office
Internet: bookstore.gpo.gov Phone: toll free (866) 512–1800; DC area (202) 512–1800
Fax: (202) 512–2104 Mail: Stop IDCC, Washington, DC 20402–0001

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY

HON. LAMAR S. SMITH, Texas, *Chair*

DANA ROHRABACHER, California
RALPH M. HALL, Texas
F. JAMES SENSENBRENNER, JR., Wisconsin
FRANK D. LUCAS, Oklahoma
RANDY NEUGEBAUER, Texas
MICHAEL T. McCAUL, Texas
PAUL C. BROUN, Georgia
STEVEN M. PALAZZO, Mississippi
MO BROOKS, Alabama
RANDY HULTGREN, Illinois
LARRY BUCSHON, Indiana
STEVE STOCKMAN, Texas
BILL POSEY, Florida
CYNTHIA LUMMIS, Wyoming
DAVID SCHWEIKERT, Arizona
THOMAS MASSIE, Kentucky
KEVIN CRAMER, North Dakota
JIM BRIDENSTINE, Oklahoma
RANDY WEBER, Texas
CHRIS STEWART, Utah
VACANCY

EDDIE BERNICE JOHNSON, Texas
ZOE LOFGREN, California
DANIEL LIPINSKI, Illinois
DONNA F. EDWARDS, Maryland
FREDERICA S. WILSON, Florida
SUZANNE BONAMICI, Oregon
ERIC SWALWELL, California
DAN MAFFEI, New York
ALAN GRAYSON, Florida
JOSEPH KENNEDY III, Massachusetts
SCOTT PETERS, California
DEREK KILMER, Washington
AMI BERA, California
ELIZABETH ESTY, Connecticut
MARC VEASEY, Texas
JULIA BROWNLEY, California
MARK TAKANO, California
ROBIN KELLY, Illinois

SUBCOMMITTEE ON RESEARCH

HON. LARRY BUCSHON, Indiana, *Chair*

STEVEN M. PALAZZO, Mississippi
MO BROOKS, Alabama
STEVE STOCKMAN, Texas
CYNTHIA LUMMIS, Wyoming
JIM BRIDENSTINE, Oklahoma
LAMAR S. SMITH, Texas

DANIEL LIPINSKI, Illinois
ZOE LOFGREN, California
AMI BERA, California
ELIZABETH ESTY, Connecticut
EDDIE BERNICE JOHNSON, Texas

SUBCOMMITTEE ON TECHNOLOGY

HON. THOMAS MASSIE, Kentucky, *Chair*

JIM BRIDENSTINE, Oklahoma
RANDY HULTGREN, Illinois
DAVID SCHWEIKERT, Arizona

LAMAR S. SMITH, Texas

FREDERICA S. WILSON, Florida
SCOTT PETERS, California
DEREK KILMER, Washington
EDDIE BERNICE JOHNSON, Texas

CONTENTS

Wednesday, April 24, 2013

	Page
Witness List	2
Hearing Charter	3

Opening Statements

Statement by Representative Larry Bucshon, Chairman, Subcommittee on Research, Committee on Science, Space, and Technology, U.S. House of Representatives	8
Written Statement	9
Statement by Representative Daniel Lipinski, Ranking Minority Member, Subcommittee on Research, Committee on Science, Space, and Technology, U.S. House of Representatives	10
Written Statement	11
Statement by Representative Thomas Massie, Chairman, Subcommittee on Technology, Committee on Science, Space, and Technology, U.S. House of Representatives	12
Written Statement	13
Statement by Representative Frederica S. Wilson, Ranking Minority Member, Subcommittee on Technology, Committee on Science, Space, and Technology, U.S. House of Representatives	13
Written Statement	14

Witnesses:

Dr. David McQueeney, Vice President, Technical Strategy and Worldwide Operations, IBM Research	
Oral Statement	16
Written Statement	18
Dr. Michael Rappa, Director, Institute for Advanced Analytics, Distinguished University Professor, North Carolina State University	
Oral Statement	26
Written Statement	28
Dr. Farnam Jahanian, Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate, National Science Foundation	
Oral Statement	36
Written Statement	38
Discussion	55

Appendix I: Answers to Post-Hearing Questions

Dr. Michael Rappa, Director, Institute for Advanced Analytics, Distinguished University Professor, North Carolina State University	76
Dr. Farnam Jahanian, Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate, National Science Foundation	79

Page

Appendix II: Additional Material for the Record

IDC IVIEW report, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, submitted by Representative Derek Kilmer, Subcommittee on Technology, Committee on Science, Space, and Technology, U.S. House of Representatives 86

NEXT GENERATION COMPUTING AND BIG DATA ANALYTICS

WEDNESDAY, APRIL 24, 2013

HOUSE OF REPRESENTATIVES,
SUBCOMMITTEE ON RESEARCH &
SUBCOMMITTEE TECHNOLOGY
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY,
Washington, D.C.

The Subcommittees met, pursuant to call, at 10:04 a.m., in Room 2318 of the Rayburn House Office Building, Hon. Larry Bucshon [Chairman of the Subcommittee on Research] presiding.

LAMAR S. SMITH, Texas
CHAIRMAN

EDDIE BERNICE JOHNSON, Texas
RANKING MEMBER

Congress of the United States
House of Representatives

COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
2321 RAYBURN HOUSE OFFICE BUILDING
WASHINGTON, DC 20515-6301
(202) 225-6371
www.science.house.gov

Subcommittees on Research and Technology Hearing

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

Witnesses

Dr. David McQueeney, Vice President, Technical Strategy and Worldwide Operations, IBM Research

Dr. Michael Rappa, Executive Director of the Institute for Advanced Analytics, Distinguished University Professor, North Carolina State University

Dr. Farnam Jahanian, Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate, National Science Foundation (NSF)

U.S. HOUSE OF REPRESENTATIVES
COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY
SUBCOMMITTEES ON RESEARCH AND TECHNOLOGY
HEARING CHARTER

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013
10:00 a.m. – 12:00 p.m.
2318 Rayburn House Office Building

Purpose

On Wednesday, April 24, 2013, the House Committee on Science, Space, and Technology's Research and Technology Subcommittees will examine how advancements in information technology and data analytics enable private and public sector organizations to utilize mass volumes of data to provide greater value to their customers and citizens, spurring new product and service innovations. The hearing will focus on innovative data analytics capabilities, research and development efforts, management challenges, and workforce development issues associated with the "Big Data" phenomenon.

Witnesses

- **Dr. David McQueeney,** Vice President, Technical Strategy and Worldwide Operations, IBM Research
- **Dr. Michael Rappa,** Executive Director of the Institute for Advanced Analytics, Distinguished University Professor, North Carolina State University
- **Dr. Farnam Jahanian,** Assistant Director for the Computer and Information Science and Engineering (CISE) Directorate, National Science Foundation (NSF)

Overview

Unprecedented volumes of complex and diverse data sets are being generated daily across a range of industries and public sector organizations. The term "Big Data" encompasses the challenge of collecting, analyzing and disseminating the massive data sets that are currently being generated and stored. Private industry and government officials are seeking ways to harness, analyze, and exploit these data sets in ways that provide greater value to their customers and citizens. While Big Data is a relatively new term, the problem is not. What is changing is both the volume of the data and the pressure to find technological solutions to managing, storing, and utilizing that data.

The McKinsey Global Institute estimated that global enterprises stored more than seven exabytes of new data on disk drives in 2010, and that consumers stored more than six exabytes on personal computers and laptops.[1] An Exabyte is 10^{18} bytes or one billion gigabytes. As a frame

[1] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, May 2011, McKinsey and Company.

of reference, one exabyte of data is more than 4,000 times the amount of data stored in the U.S. Library of Congress.

Given the evolution of computing power and analytical capabilities, overcoming the challenge of data management presents significant need for technological innovation. High performance computing can process mass, complex sets of data at a greater rate; mathematicians and statisticians are developing new algorithms to analyze data; and data analytics professionals are employing new techniques to extract value from data.

Big Data has profound implications for a range of industries. For example, health care data can enable care providers to monitor health trends and evaluate different treatments, energy data can inform power distribution creating greater efficiencies, transportation data can be used to mitigate traffic congestion, and information technology data can identify potential cyber threats.

In addition, technological advances allow scientists to both collect and analyze data at a significantly faster rate. Examples include advancements in human genome sequencing, digital astronomy data, and particle physics.

Industry and Big Data

Big Data represents a significant growth area for private industry. In recent years, industry spending on data analytics and management has increased approximately 10 percent a year.[2]

Companies utilize data analytics to manage supply chains, target marketing based on user preferences, provide airline fare prediction services for consumers, and reduce costs by identifying operating inefficiencies, among a multitude of other uses.

Information Communications and Technology (ICT) companies and management consulting companies are providing a range of Big Data capabilities, including software, hardware, and analytics services. Industry customers of Big Data products and services, including health care, transportation, agriculture, and retail companies are identifying ways to increase yields, cut costs, and increase customer retention. ICT companies also work closely with government and academia to build high performance computers and software systems, which enable cutting edge Big Data scientific research and development initiatives.

Big Data Workforce Development

McKinsey has projected the United States will need an additional 140,000 to 190,000 professionals with significant technical depth in data analytics, and the need for an additional 1.5 million managers and analysts who can work effectively with big data analysis by 2018.[3]

To address anticipated workforce demands, colleges and universities are recognizing the value in providing students with education and training in Big Data-related disciplines. Such programs provide instruction in a broad spectrum of Big Data-related disciplines including data management, mathematical and statistical methods for data modeling, and techniques for data

[2] "*A special report on managing information: Data, data everywhere*," The Economist; February 25, 2010.
[3] *Big data: The next frontier for innovation, competition, and productivity*, McKinsey Global Institute, May 2011, McKinsey and Company.

visualization in support of business decision making. Although some institutions have initiated these types of degree programs, overall they are still relatively rare.

Federal Big Data Research and Development Initiatives

On March 29, 2012, the Obama Administration unveiled its "Big Data Research and Development Initiative,"[4] announcing more than $200 million in new funding to improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of digital data.

Six federal departments and agencies, including the National Science Foundation (NSF), National Institutes of Health (NIH), the Department of Defense (DOD) and Defense Advanced Research Projects Agency (DARPA), Department of Energy (DOE), and the U.S. Geological Survey (USGS) are participating in this initiative.

National Science Foundation and Big Data

The NSF Computer and Information Sciences and Engineering Directorate (CISE) supports investigator-initiated research in all areas of computer and information science and engineering, helps develop and maintain cutting-edge national computing and information infrastructure for research and education generally, and contributes to the education and training of the next generation of computer scientists and engineers.

CISE supports Big Data investments in foundational research, cyberinfrastructure, education and workforce development needs, and in efforts to support interdisciplinary research.

Core Techniques and Technologies for Advancing Big Data Science & Engineering.

As part of the President's Big Data Initiative, the NSF/NIH Core Techniques and Technologies for Advancing Big Data Science & Engineering (BIGDATA) is offering research grants to accelerate the progress of scientific discovery and innovation; lead to new fields of inquiry that would not otherwise be possible; encourage the development of new data analytic tools and algorithms; facilitate scalable, accessible, and sustainable data infrastructure; increase understanding of human and social processes and interactions; and promote economic growth and improved health and quality of life.

Yellowstone, Blue Waters, Gordon, and Stampede Supercomputers

NSF also advances Big Data computational research and development through the Yellowstone, Blue Waters, and Stampede Supercomputers, in partnership with the University of Wyoming, the University of Illinois, the University of California, San Diego, and the University of Texas at Austin, respectively.

Yellowstone is the petascale computing resource in the National Center for Atmospheric Research (NCAR)-Wyoming Supercomputing Center (NWSC), which opened in October 2012.

[4] http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal

(A petascale refers to a computer system capable of reaching performance in excess of one petaflops, i.e. one quadrillion floating point operations per second.) The NWSC provides advanced computing services to scientists studying a broad range of disciplines, including weather, climate, oceanography, air pollution, space weather, computational science, energy production, and carbon sequestration.

The Blue Waters supercomputer provides sustained performance of one petaflop on a range of real-world science and engineering applications. Blue Waters enables scientists and engineers across the country to tackle a wide range of challenging problems, from predicting the behavior of complex biological systems to simulating the evolution of the cosmos.

The Gordon Compute Cluster is a unique data-intensive supercomputer sponsored by NSF, went into production January 1, 2012. Large graph problems, data mining, genome assembly, database applications, and quantum chemistry are some of the fields of research benefitting from Gordon's unique architecture.

Stampede was officially dedicated in March 2013 at the University of Texas at Austin's Advanced Computing Center (TACC). Stampede will have a peak performance of 10 petaflops. Research programs already being conducted at Stampede include seismic hazard mapping, ice sheet modeling, improving the imaging quality of brain tumors, and carbon dioxide capture and conversion.

Department of Energy Scalable Data Management, Analysis and Visualization (SDAV) Institute

On March 29, 2012, the DOE announced $25 million to establish the Scalable Data Management, Analysis and Visualization (SDAV) Institute to extract knowledge and insights from large and complex collections of digital data. Led by the Energy Department's Lawrence Berkeley National Laboratory, the SDAV Institute brings together the expertise of six national laboratories and seven universities to develop new tools to help scientists manage and visualize data on the Department's supercomputers, which will further streamline the processes that lead to discoveries made by scientists using the Department's research facilities.[5]

The SDAV Institute helps scientists extract insights from today's increasingly massive research datasets by assisting researchers in using state-of-the art software tools for data analysis on these supercomputers – ranging from superfast search engines to sophisticated visualization software that enables researchers to literally picture and "see" complex relations among data points. The Energy Department supports some of the world's fastest supercomputers located at Argonne, Oak Ridge, and Lawrence Berkeley National Laboratories, which are used by scientists from a wide range of fields.

National Institute of Standards and Technology (NIST) Big Data Initiatives

The NIST Information Technology Laboratory conducts a number of activities related to Big Data through its Computer Security Division. NIST conducts research on the science behind Big

[5] http://www.sdav-scidac.org/report.html

Data, measurement tools to advance Big Data, privacy, and the security of Big Data infrastructure. This effort includes convening industry and interested stakeholders together to explore the challenges of Big Data, including common terms and taxonomies for use by the field, and identification of areas where research is needed.

Specifically, NIST has worked in areas of convergence between cloud computing and data, particularly on the interoperability of cloud platforms. Although NIST does not play a visible role in the Administration's Big Data Research and Development Initiative, it supports the creation of many of the analytical tools to address the challenges of Big Data in both the public and private sectors.

Areas for Examination

Witnesses have been asked to describe private and public Big Data research and development efforts; applications of Big Data initiatives; and management challenges, including workforce development issues.

In addition, the Committee will seek to determine: how federal big data research projects are coordinated across participating agencies; how the public and private sectors manage privacy concerns as part of Big Data initiatives; how federal privacy laws, such as health privacy laws, affect opportunities to gain information from data; and how Congress should prioritize Big Data research initiatives in federal research budgets.

Chairman BUCSHON. All right. This joint hearing of the Subcommittee on Research and the Subcommittee on Technology will come to order.

Good morning, and welcome to today's joint hearing entitled "Next Generation Computing and Big Data Analytics." In front of you are packets containing the written testimony, biographies and Truth in Testimony disclosures for today's witnesses.

Before I get started, since this is a joint hearing involving two Subcommittees, I want to explain how we will operate procedurally so all Members understand how the question-and-answer period will be handled. As always, we will alternate rounds of questioning between majority and minority Members. The Chairmen and Ranking Members of the Research and Technology Subcommittees will be recognized first. Then we will recognize Members present at the gavel in order of seniority on the full Committee and those coming in after the gavel will be recognized in order of their arrival. I now recognize myself for five minutes for an opening statement.

Again, I would like to welcome everyone to today's hearing where we will examine how advancements in information technology and data analytics enable private and public sector organizations to provide greater value to their customers and citizens. Industry, academia, and government are all interested in determining how to extract value, gain insights, and make better decisions based on the wealth of data that is generated today. In recent years, "big data" has become the popular term used to encompass this phenomenon.

TechAmerica, an information technology trade association, defines big data as "large volumes of high-velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information."

Big data offers a range of opportunities for private industry to reduce costs and increase profitability. It can enable scientists to make discoveries on a previously unreachable scale. And it can allow governments to identify ways to serve its citizens more efficiently.

The McKinsey Global Institute predicts that effective information management can provide $300 billion in annual value to the U.S. health care sector alone. TechAmerica released a report last year highlighting how big data initiatives can improve the efficiency and effectiveness of government services, and through the use of advanced computing power and analytic techniques, universities and Federal laboratories can drive new research initiatives that will significantly increase our scientific knowledge base.

There are also various challenges associated with big data that the Committee will explore today. McKinsey has estimated that the U.S. will face a shortfall of 140,000 to 190,000 professionals with significant technical depth in data analytics, and a further shortfall of an additional 1.5 million managers and analysts who can work effectively with big data analysis by 2018. Committee Members will be interested to learn how industry, academia, and government are addressing this shortfall.

While the term "big data" is relatively new, public and private organizations have been investing in computing power and data

analytics for a number of years. In March of last year, the Obama Administration announced a Big Data Research and Development Initiative, including $200 million in new funding across six different government departments and agencies. I am interested to learn how effectively these programs are being coordinated across the different Federal agencies to ensure that taxpayer dollars are being leveraged effectively. Finally, privacy and security are major concerns when private and public organizations are collecting, analyzing, and disseminating massive data sets.

We have an excellent panel of witnesses ranging across industry, academia, and government. I would like to extend my appreciation to each of our witnesses for taking the time and effort to appear before us today. We look forward to your testimony.

[The prepared statement of Mr. Bucshon follows:]

PREPARED STATEMENT OF SUBCOMMITTEE ON RESEARCH CHAIRMAN LARRY BUCSHON

Good morning, I would like to welcome everyone to today's hearing where we will examine how advancements in information technology and data analytics enable private and public sector organizations to provide greater value to their customers and citizens.

Industry, academia, and government are all interested in determining how to extract value, gain insights, and make better decisions based on the wealth of data that is generated today. In recent years, "Big Data" has become the popular term used to encompass this phenomenon.

TechAmerica, an information technology trade association, defines Big Data as "large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information."

Big Data offers a range of opportunities for private industry to reduce costs and increase profitability. It can enable scientists to make discoveries on a previously unreachable scale. And it can allow governments to identify ways to serve its citizens more efficiently.

The McKinsey Global Institute predicts that effective information management can provide $300 billion in annual value to the US health care sector alone. TechAmerica released a report last year highlighting how Big Data initiatives can improve the efficiency and effectiveness of government services. And, through the use of advanced computing power and analytics techniques, universities and federal laboratories can drive new research initiatives that will significantly increase our scientific knowledge-base.

There are also various challenges associated with Big Data that the Committee will explore today. McKinsey has estimated that the US will face a shortfall of 140,000 to 190,000 professionals with significant technical depth in data analytics, and a further shortfall of an additional 1.5 million managers and analysts who can work effectively with big data analysis by 2018. Committee members will be interested to learn how industry, academia, and government are addressing this shortfall.

While the term Big Data is relatively new, public and private organizations have been investing in computing power and data analytics for a number of years. In March of last year, the Obama Administration announced a "Big Data Research and Development Initiative," including $200 million in new funding across six different federal departments and agencies. I am interested to learn how effectively these programs are being coordinated across the different federal agencies to ensure that taxpayer dollars are being leveraged effectively.

Finally, privacy and security are major concerns when private and public organizations are collecting, analyzing, and disseminating massive data sets. We have an excellent panel of witnesses ranging across industry, academia and government. I'd like to extend my appreciation to each of our witnesses for taking the time and effort to appear before us today. We look forward to your testimony.

Chairman BUCSHON. I will now yield to Mr. Lipinski for his opening statement.

Mr. LIPINSKI. Thank you. I want to thank you, Chairman Bucshon, and I want to thank Chairman Massie for holding this hearing. I want to welcome and thank the witnesses for being here.

Today's hearing gives us an opportunity to talk about the new tools and analytics that are being developed for big data. As Chairman Bucshon stated, big data can be thought of as large volumes of complex and diverse types of data that change rapidly with time.

In basic scientific research in national security as well as in economic sectors ranging from energy to health care, big data challenges are becoming fundamentally important. Effectively dealing with big data can impact how we do business and how we think about the world.

As a Member of the Research Subcommittee for several years, I have watched as the amount and complexity of data has grown by leaps and bounds. The field of astronomy is a great example. When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in a few weeks than had been collected in the history of astronomy, and that telescope will be surpassed when the Large Synoptic Survey Telescope begins scientific operations in 2020. LSST will photograph the entire sky every few days, producing data at a rate almost 100 times greater than the Sloan Survey. But data is useless without the means to store and analyze it in an efficient manner.

The types of data are changing as well. Data has gone from being mostly numbers entered into Excel spreadsheets to data coming from sensors, cell phone cameras and millions of email messages. In fact, it is estimated that over 85 percent of data generated today are these kinds of unstructured data, data like videos and emails. The change in the volume and variety of data as well as how fast data is being produced and changed creates almost limitless opportunities. For example, since cybersecurity data is massive, varied, and changing quickly, big data technologies have the potential to detect and prevent cyber attacks before they happen. I know that organizations like IBM are developing technologies to do just that. Additionally, big data could be used to establish new business models, create transparency, improve decision-making and reduce inefficiencies within businesses and government.

But along with the opportunities, there are a number of challenges. We need new tools and software packages to manage, organize, and analyze all these different kinds of data. Additionally, we will need an analytic workforce to ensure the gains of big data. These challenges necessitate involvement from government, academia and the private sector. That is why I am happy to see all those sectors represented here today.

The government has and will continue to play an instrumental role in this area. For instance, the Networking and Information Technology Research and Development program, or NITRD, created an interagency big data group that is coordinating Federal efforts in technologies, research, competitions, and workforce development for big data. We had a hearing on the NITRD program back in February, and I expect that we will be able to take a broader look at many of the same issues in today's hearing.

In some cases, agencies have teamed up to issue joint solicitations. For example, NSF and NIH have a joint big data grant pro-

gram that awarded nearly $15 million of grants to eight teams of researchers last year. These first award grants went to projects focused on designing new tools for big data and new data analytic approaches. We will be hearing more about these and other interagency activities from Dr. Jahanian in his testimony. We will also learn more about specific programs at NSF, one of the leading agencies in Federal big data efforts on both the analytics side and the computational resources side.

As I mentioned before, one of the areas being coordinated through NITRD is workforce development for big data. Several agencies, including NSF, have education activities to support a new generation of big data researchers. As we will likely hear from all of the witnesses, we face a looming shortage of workers with the skills needed to analyze and manage large, complex and high-velocity data sets. There is some overlap with the broader STEM skills we so often speak about in this committee, but there are also unique skills required to address the big challenges of big data. We need to consider how to build those skills into STEM curricula, especially at the undergraduate and graduate levels. I look forward to hearing from our witnesses about the current educational efforts and what additional initiatives may be necessary.

And finally, since big data involves different types of data that can be produced and transferred quickly, there are concerns over privacy. We need to ensure that we strike the right balance between exploring and implementing all of the potential benefits of big data while also protecting individuals' personal information.

I look forward to hearing the witnesses' testimony and our discussion today, and I yield back the balance of my time.

[The prepared statement of Mr. Lipinski follows:]

PREPARED STATEMENT OF SUBCOMMITTEE ON RESEARCH
RANKING MINORITY MEMBER DANIEL LIPINSKI

Thank you, Chairmen Bucshon and Massie for holding this hearing on examining the next generation of computing and big data analytics. I want to welcome and thank the witnesses for being here today.

Today's hearing gives us an opportunity to talk about the new tools and analytics that are being developed for big data. Big data can be thought of as large volumes of complex and diverse types of data that are also high velocity—meaning they change rapidly with time.

As a member of the Research Subcommittee for several years now, I have watched as the amount and complexity of data has grown by leaps and bounds. The field of astronomy is a great example. When the Sloan Digital Sky Survey started work in 2000, its telescope in New Mexico collected more data in a few weeks than had been collected in the history of astronomy. And that telescope will be surpassed when the Large Synoptic Survey Telescope goes online in about 2020. LSST will photograph the entire sky every few days. That's difficult for any of us to wrap our heads around.

The types of data are changing as well. Data has gone from being mostly numbers entered in excel spreadsheets to data coming from sensors, cellphone cameras, and millions of email messages. In fact, it is estimated that over 85 percent of data generated today are these kinds of unstructured data—data like videos or emails.

The change in the volume and variety of data as well as how fast data is being produced and changed creates almost limitless opportunities. For example, since cybersecurity data is massive, varied, and changing quickly, big data technologies have the potential to detect and prevent cyber attacks before they even happen. I know that organizations like IBM are developing technologies to do just that. Additionally, big data could be used to establish new business models, create transparency, improve decision-making, and reduce inefficiencies within businesses and government.

But along with the opportunities, there are a number of challenges. We need new tools and software packages to manage, organize, and analyze all these different kinds of data. Additionally, we will need an analytic workforce to ensure the gains of big data. These challenges necessitate involvement from government, academia, and the private sector. That is why I am happy to see all those sectors represented today.

The government has and will continue to play an instrumental role in this area. For instance, the Networking and Information Technology Research and Development—or NITRD—program created an interagency big data group that is coordinating federal efforts in technologies, research, competitions, and workforce development for big data.

In some cases, agencies have teamed up to issue joint solicitations. For example, NSF and NIH have a joint big data grant program that awarded nearly $15 million of grants to eight teams of researchers last year. These first awarded grants went to projects focused on designing new tools for big data and new data analytic approaches. We will hear more about these and other interagency activities from Dr. Jahanian in his testimony. We will also learn more about specific programs at NSF, one of the leading agencies in federal big data efforts on both the analytics side and the computational resources side.

As I mentioned before, one of the areas being coordinated through NITRD is the workforce development needs for big data. Several agencies, including NSF, have education activities to support a new generation of big data researchers. As you will likely hear from all of the witnesses, we face a looming shortage of workers with the skills needed to analyze and manage large, complex, and high-velocity data sets. There is some overlap with the broader STEM skills we often speak of in this committee. But there are also some unique skills required to address the challenges of big data. We need to consider how to build those skills into STEM curricula, especially at the undergraduate and graduate levels. I look forward to hearing from our witnesses about the current educational efforts and what additional initiatives may be necessary.

Finally, since big data involves different types of data that can be produced and transferred quickly, there are concerns over privacy. We need to ensure that we strike the right balance between exploring and implementing all of the potential benefits of big data while also protecting individuals' personal information.

I look forward to hearing the witnesses' testimonies and to our discussion today.

Chairman BUCSHON. Thank you, Mr. Lipinski. The Chair now recognizes the Chairman of the Subcommittee on Technology, Mr. Massie, for five minutes for his opening statement.

Mr. MASSIE. Thank you, Chairman.

Good morning. Today we are examining an issue that we hear a lot about. "Big data" is a popular new term that can mean a lot of different things. The scientific community, though, has generated and used big data before there was the term "big data." In fact, in 1991 this Committee authored the High Performance Computing Act, which organized the Federal agency research, development, and training efforts in support of advanced computing.

Individual researchers have always been faced with difficult decisions about their data: what to keep, what to toss, what to verify with additional experiments. And as our computing power has increased, so has the luxury of storing more data. Incorporating computer power to process more scientific data is transforming laboratories across the country.

At the same time, the ability to analyze large amounts of data across multiple networked platforms is also transforming the private sector. Through big data applications, businesses have not only revealed previously hidden efficiency improvements in their internal operations, but, more importantly, also uncovered entirely new types of businesses built around data that was previously not accessible due to its size and complexity.

Today's hearing will examine the hype around big data. Is the United States the most innovative Nation in big data? Is our regulatory system creating any burdens on businesses? Could public-private partnerships with the Federal agencies be improved to allow for more data innovations?

I thank our witnesses today for their participation today and I look forward to hearing their testimony. Thank you. I yield back.

[The prepared statement of Mr. Massie follows:]

PREPARED STATEMENT OF SUBCOMMITTEE ON TECHNOLOGY
CHAIRMAN THOMAS MASSIE

Good Morning. Today we are examining an issue that we hear a lot about. "Big Data" is a popular new term that can mean a lot of different things.

The scientific community has generated and used Big Data before there was Big Data. In fact, in 1991 this Committee authored the High Performance Computing Act, which organized the federal agency research, development and training efforts in support of advanced computing.

Individual researchers have always been faced with difficult decisions about their data: what to keep, what to toss, what to verify with additional experiments. As our computing power has increased, so has the luxury of storing more data. Today, managing this data allows for better-informed experiments, more exact metrics, and perhaps significantly longer doctoral theses. Incorporating computer power to process more scientific data is transforming laboratories across the country.

At the same time, the ability to analyze large amounts of data across multiple networked platforms is also transforming the private sector. Through Big Data applications, businesses have not only revealed previously hidden efficiency improvements in their internal operations, but also uncovered entirely new types of business built around data that was previously not accessible due to its size and complexity.

Today's hearing will examine the hype around Big Data. Is the United States the most innovative nation in Big Data? Is our regulatory system creating any burdens on businesses? Could public-private partnerships with the federal agencies be improved to allow for more data innovations?

I thank our witnesses for their participation today and look forward to hearing their testimony.

Chairman BUCSHON. Thank you, Mr. Massie. The Chair now recognizes Ms. Wilson for five minutes for her opening statement.

Ms. WILSON. First of all, I would like to thank both Chairman Bucshon and Chairman Massie for holding this joint hearing, and thank you all to our witnesses for being here today. Welcome.

This morning's hearing provides us with the opportunity to discuss one of the newest buzzwords in Washington, and you know we have many buzzwords here. This one: big data. This buzzword is not an exaggeration. A computer that used to take up the space of this entire room now fits in the palm of your hand. It is remarkable.

Just as computers have gotten immensely smaller, they have also gotten immensely more powerful. Instead of talking about megabytes, we are now talking about petabytes and zettabytes—quadrillions and sextillions of units of information. It boggles the mind. Collecting and storing this huge volume of data would have been impossible just a few years ago.

I am looking forward to your testimony and learning more about the benefits of big data to society. As I understand it, big data has the potential to improve nearly all sectors of society. The National Cancer Institute is funding a prototype in biological big data that could lead to new advances in cancer treatment. Companies and agencies are using big data to run controlled experiments that im-

prove decision-making. Scientists at Florida International University in my district are using big data to advance understanding of topics including cybersecurity, social networks and cloud computing.

But there are challenges. In order to reap all the benefits of complex and broadly available data, we need new technologies and software. We also need a workforce, a workforce with the skills necessary to analyze data of such great volume and complexity. A recent study estimates that the United States is in need of 190,000 additional data scientists.

In thinking about this hearing on big data, I couldn't help but think about the tragic events last week in Boston. The marathon bombings may be one of the most photographed attacks in history. The Massachusetts State Police asked the public to share the photos and videos taken on that awful day. Now all of this digital information has been and is being used by the Boston Police Department and the FBI in their investigation. It appears that this data has been instrumental in helping to identify the individuals who were involved.

Examples like this one demonstrate how important it is that we develop and attain the tools and the skills people need to analyze tremendous amounts of complex data. Big data can not only lead to amazing scientific discoveries; it can also save lives.

As we learn more about these opportunities and challenges today, I hope our witnesses will offer recommendations on how the Federal Government can help create the new tools, software and workforce needed to realize the full potential of big data.

Chairman Bucshon, Chairman Massie, thank you again for holding this hearing, and I yield back the balance of my time.

[The prepared statement of Ms. Wilson follows:]

PREPARED STATEMENT OF SUBCOMMITTEE ON TECHNOLOGY
RANKING MINORITY MEMBER FREDERICA S. WILSON

I'd like to thank both Chairman Bucshon and Chairman Massie for holding this joint hearing. And thank you to all of our witnesses for being here today.

This morning's hearing provides us with the opportunity to discuss one of the newest buzz-words in Washington and around the world—"big data."

This buzz-word is not an exaggeration: A computer that used to take up the space of this entire room now fits in the palm of your hand. It is remarkable.

Just as computers have gotten immensely smaller, they have also gotten immensely more powerful. Instead of talking about megabytes, we are now talking about petabytes and zettabytes—quadrillions and sextillions of units of information. It boggles the mind. Collecting and storing this huge volume of data would have been impossible just a few years ago.

I'm looking forward to the testimony of today's witnesses and learning more about the benefits of "big data" to society.

As I understand it, big data has the potential to improve nearly all sectors of society. The National Cancer Institute is funding a prototype in biological "big data" that could lead to new advances in cancer treatment. Companies and agencies are using "big data" to run controlled experiments that improve decision-making. Scientists at Florida International University—in my district—are using "big data" to advance understanding of topics including cybersecurity, social networks, and cloud computing.

But there are challenges. In order to reap all the benefits of complex and broadly available data, we need new technologies and software. We also need a workforce with the skills necessary to analyze data of such great volume and complexity. A recent study estimates that the United States is in need of 190,000 additional data scientists.

In thinking about this hearing on "big data," I couldn't help but think about the tragic events last week in Boston. The marathon bombings may be one of the most photographed attacks in history. The Massachusetts State Police asked the public to share the photos and videos taken on that awful day. Now, all of this digital information has been and is being used by the Boston Police Department and the FBI in their investigation. It appears that this data has been instrumental in helping to identify the individuals who were involved.

Examples like this one demonstrate how important it is that we develop and attain the tools and the skilled people needed to analyze tremendous amounts of complex data. Big data can not only lead to amazing scientific discoveries—It can also save lives.

As we learn more about these opportunities and challenges today, I hope our witnesses will offer recommendations on how the federal government can help create the new tools, software, and workforce needed to realize the full potential of "big data."

Chairman BUCSHON. Thank you, Ms. Wilson.

If there are Members who wish to submit additional opening statements, your statements will be added to the record at this point.

It is now time to introduce our panel of witnesses. Our first witness is Dr. David McQueeney, the Vice President of Technical Strategy and Worldwide Operations at IBM Research. In this capacity, he is responsible for setting the direction of IBM's overall research strategy across 12 worldwide labs and leading the global operations and information systems teams. Dr. McQueeney's background covers a wide range of disciplines, spending about half of his career as a researcher and research executive and half in IBM's customer-focused areas. He holds an M.S. and Ph.D. in solid-state physics from Cornell University and an A.B. in physics from Dartmouth College. Welcome.

Our second witness is Dr. Michael Rappa, the Executive Director of the Institute for Advanced Analytics and Faculty Member of the Department of Computer Science at North Carolina State University. Dr. Rappa has 25 years of experience as a professor working across academic disciplines at the intersection of management and computing. He began his teaching career at the University of Minnesota where he earned his doctorate degree. Welcome.

And our final witness is Dr. Farnam Jahanian, the Assistant Director for the Computer and Information Science and Engineering Directorate at the National Science Foundation and a frequent visitor to our Subcommittee. He oversees the CISE's mission to uphold the Nation's leadership in computer and information science and engineering. He also serves as Co-chair of the Networking and Information Technology Research and Development, or NITRD, Subcommittee of the National Science and Technology Council Committee on Technology, providing overall coordination for the activities of 14 government agencies. Dr. Jahanian holds a master's degree and a Ph.D. in computer science from the University of Texas at Austin. Welcome again.

As our witnesses should know, spoken testimony is limited to five minutes each after which Members of the Committee have five minutes each to ask questions. Your written testimony will be included in the record of the hearing.

I now recognize our first witness, Dr. McQueeney, for five minutes for his testimony.

TESTIMONY OF DR. DAVID MCQUEENEY, VICE PRESIDENT, TECHNICAL STRATEGY AND WORLDWIDE OPERATIONS, IBM RESEARCH

Dr. MCQUEENEY. Good morning, Mr. Chairman, Ranking Members, Members of the Subcommittees. Thank you for the opportunity to testify today. My written testimony covers next-generation computing, big data and analytics, workforce development and the role of government. In my five minutes, I will focus on areas where I can offer critical insights from my personal experience.

Computing today is undergoing profound change. We are moving from computing based on processors that are programmed to follow a predesigned sequence of instructions to cognitive computing systems based on massive amounts of data evolving into systems that can learn. This new approach will require new strategies in hardware and in software and improved skills to maintain U.S. leadership. Cognitive systems will digest and exploit massive data volumes. Tools such as mobile phones, videos and social networks generate as much data in two days in 2013 as in all of human history prior to 2003.

Advanced analytics can be thought of as tools for infusing all this data to make decisions on facts rather than intuition. The challenge is to transform latent data into actionable information to decide what to do next. For example, the Memphis Police Department is using data analytics to map crime hotspots and find patterns. As a result, they have been able to reduce crime by 30 percent with no increase in overall police manpower.

To run advanced analytics, it is essential to have the most powerful computing systems. However, current supercomputing systems are reaching performance levels that will stagnate without significant innovation. We must move to the next generation of large-scale computing called exascale computing, a thousand times faster than today's petascale machines.

The United States needs to invest now in the research and development for exascale systems to maintain strategic and economic leadership. Government-funded research on domain skills, especially at our national laboratories, should target systems for modeling, simulation, and analytics on big data.

Before 2005, the United States had a clear lead in the global supercomputing race. Today, we are still ahead but the rest of the world is catching up rapidly. To stay ahead will require new skills and knowledge and new types of decision-making. Nearly two million IT jobs will be created by 2015 in the United States to support big data, and the job candidates with analytic skills will get these jobs.

Industry is developing many collaborative skills programs, as enumerated in my testimony. I highlight our announcement today with Rensselaer Polytechnic Institute to offer a graduate degree program in the fall of 2013, the Master of Science in Business Analytics.

Privacy must be considered in the design of big data systems. Big data does not require the sacrifice of personal privacy. When personal information is used, design-in processes such as IBM's Privacy By Design can protect privacy. When people understand how information is used, they have the ability to set data usage policies

and enjoy benefits of the analysis, they tend not to have privacy concerns.

The government's role should focus on research and skills. First, Federal research investment in high-performance computing is critical to big data. Industry needs university-based exploratory research into numerous areas including system design, flexible software defined environments, and IT infrastructure.

Second, IBM strongly supports the reauthorization of the Department of Energy *High End Computing Revitalization Act of 2004* to be offered by Representative Hultgren. This bill will improve high-end computing R&D at the DOE and strengthen government industry partnerships for exascale platforms. IBM has a long history of successful partnerships with DOE. This partnership established computational simulation as an essential tool for scientific inquiry and led to world leadership in the United States in high-performance computing. The challenge ahead is to continue this growth. Past Federal investments in HP-related research, particularly at DOE's national labs, have underpinned mission-critical supercomputers at DOD, NASA, NOAA, and in the intelligence agencies.

Third, the professional science masters program supported by NSF is particularly relevant as it provides advanced training in science or mathematics and develops workplace skills valued by employers. Finally, Congress should reauthorize the Carl D. Perkins Act and the Federal work-study program and restructure them to align workforce needs and big data.

In conclusion, there exists today a tremendous abundance of data about our world. New cognitive computing capabilities will help determine which countries and businesses will thrive. The United States should support advanced computing and build its workforce to seize the future.

Thank you, and I welcome your questions.

[The prepared statement of Dr. McQueeney follows:]

Dr. David F. McQueeney
Vice President, Technical Strategy and Worldwide Operations
IBM Research

House Science Research Subcommittee and
House Science Technology Subcommittee
Next Generation Computing and Big Data Analytics Hearing
April 24, 2013

Good morning, Chairman Bucshon, Chairman Massie, Ranking Member Lipinski, Ranking Member Wilson, and members of the subcommittees. Thank you for the opportunity to speak with you about big data, analytics, and the new opportunities they present.

My name is David McQueeney and I am the Vice President for Technical Strategy and Worldwide Operations at IBM Research. I am responsible for setting the direction of IBM's overall research strategy and I lead the creation of IBM's Global Technology Outlook. This is an annual effort which guides IBM's R&D directions and its acquisition strategies. I have held a variety of senior research and business unit leadership roles throughout my twenty-five years at IBM.

My testimony today draws on this and the wider IBM experience with Big Data and Analytics and associated technology and skills requirements. I will focus on next generation computing, the nature of big data and analytics, opportunities and issues these technologies present, and how best to promote their growth.

What is Big Data and Analytics?

We are entering a new era of computing which is causing profound change in the IT industry. We are moving from computing based on processors that are programmed to follow a pre-designed sequence of instructions, to the cognitive computing era based on massive amounts of data and systems evolving into systems that can "learn". Cognitive systems can modify and optimize projections or weigh the value of information based on experience and results. This new approach to computing requires new strategies and skills to maintain U.S. leadership.

Cognitive systems will digest and exploit the massive data volumes being generated today. The data is coming from the technologies which mark our age: mobile phones, cloud computing, social networks and what we call the "internet of things", including everything from your car to your refrigerator to the thousands of texts your son or daughter sends each month. Imagine the data volumes generated by just two popular sources, the two billion videos watched daily on YouTube and the 293 billion emails sent every single day. As much data is generated in two days in 2013 as in all of human history prior to 2003.

Why is Big Data Analytics important?

Advanced analytics can be thought of as tools for using data to make decisions based on facts rather than intuition. These tools are applied to unstructured data from the Web, communications networks, governments, businesses, and tens of millions of sensors, to help us understand how the world works better than ever before. The challenge for business and government alike is to transform latent data into meaningful, actionable information. Analytic tools enable predictive insights and inform decision making to prevent bad outcomes and define helpful courses of action.

For example, the Memphis Police Department is using data analytics to create maps of crime hot spots and find patterns that police could not see themselves. They have correlated crimes committed at night with presence of pay phones outside convenience stores. As a result, the police suggested moving the phones inside the stores. The outcome – crime has fallen. Memphis has increased targeted, effective police presence and eliminated random patrolling. The city has seen a 30 percent reduction in crime with no increase in overall manpower using new knowledge and skills based on data analytics. There are numerous examples around the world from all sectors of society.

In sum, advanced analytics are tools for using data to improve organizational effectiveness and create value. Combining analytics with massive data flows enables organizations to make decisions based on facts rather than intuition. This factual, analytic decision making can revolutionize industries and help society successfully address challenges such as energy conservation, health care, and transportation, as well as rooting out waste, fraud and abuse.

Trends in technology development for Next Generation Computing and the Data Explosion

The most powerful computing systems today are essential in both government and industry. They are used in applications ranging managing the nation's nuclear stockpile to automobile, aircraft and semiconductor design to the development of new tires and race car aerodynamics to oil exploration and oil recovery. However, they are reaching a performance level that will stagnate without significant innovation. We must move to the next generation of large-scale computing, exascale computing – 1000 times faster than today's petascale machines.

An exaflop is a quintillion floating point operations per second. Exascale computers will have performance speeds at or above 10 to the 18th power of floating point operations. That's fast. We are at a major inflection point in developing next generation, high end, exascale systems.

The United States must invest now to maintain its economic leadership and competitiveness. Government funding and domain skills, especially in our national labs, are required. Investment should be targeted at developing innovative and composable systems for modeling, simulation and analytics on big data. The new imperative is to design for data, not for processor performance. New systems and software concepts must be developed to continue U.S. strategic leadership and to support and enhance the economic competitiveness of U.S. Industry.

To put this in context, before 2005, the U.S. led decidedly in the global supercomputing race, with Europe and Japan following. Today, however, and in the future, additional regions are making important sovereign strategic investments to compete aggressively.

The U.S. is still ahead, but others are catching up fast. Japan is continuing its long term effort in high performance computing (HPC) as well, with a focus on building "Big Iron" systems. India is making a $1B+ investment to pursue exascale computing with an emphasis on applications and user capability. And in Europe, which has been a leader in software and applications for industrial use, we now see strong HPC support for small and medium enterprises. There is an aggressive effort to build a European capability in HPC with support for companies such as Bull and in complementary infrastructure efforts with a number of "Big Science" projects ($1B+) including Blue Brain, Graphene, Robotic, and Health Care.

In sum, the global race is on. The necessity for leadership in high end computing for a highly competitive economy has been recognized around the world and regions are stepping up to the plate.

Work Force Development

Skills are a major inhibitor to the growth of next generation computing and Big Data. Big Data requires new skills, knowledge and new types of decision-making. Today's employers are seeking job candidates who can analyze and build strategy around Big Data, or the 2.5 quintillion bytes of information generated daily. Indeed, the number one barrier to the adoption of business analytics technologies is insufficient skills and experience, according to IBM's 2012 Tech Trends Report.

Much has been written about the urgency of this skills gap. McKinsey Global Institute reports that over the next seven years, the need for skilled business intelligence workers in the U.S. alone will dramatically exceed the available workforce -- by as much as 60 percent. There is not a job crisis in the U.S.; there is a skills crisis. Nearly two million information technology jobs will be created by 2015 in the U.S. to support Big Data, according to research firm Gartner, Inc. Analytics skills will be a key differentiator for candidates seeking to fill those jobs.

Further, in our global society, applying analytics to Big Data will be a key factor in determining which countries pull ahead economically and which ones fall behind. Industry is responding to this situation with a variety of programs to build new skills. I will describe several which IBM has created. We are addressing skills requirements from grade nine through graduate school and into the working population around the globe.

In fact, just today IBM is announcing a new partnership to prepare business students for the expanding scope of careers requiring Big Data analytics skills. IBM and Rensselaer Polytechnic Institute are combining forces to offer a new Lally School of Management and Technology graduate degree program in fall 2013: the Master of Science in Business Analytics.

This new Master of Science in Business Analytics is a one year, 30-credit graduate degree program that will provide students and career professionals with the hands-on experience and knowledge required to succeed in analytics jobs. It will feature a three-part curriculum comprised of:

- A business core to ensure students understand where Big Data fits into a business' strategy and operations, as well as how analytics-driven decisions can impact a business' growth, competitive standing and bottom line.
- An analytics core that includes hands-on training in predictive modeling to help businesses identify profitable data patterns, focusing also on data management, statistical analysis and leading-edge techniques in harnessing Big Data.
- An experiential core with project-based courses that allow students to apply their newly gained skills to real-world problems faced by businesses spanning a range of industries.

This new collaboration builds on experience gained from our numerous other engagements. Some highlights of those are below:

Early STEM Development - IBM collaborates with the New York City Department of Education, the City University of New York, and NYC College of Technology to provide a single public school for grades 9-14 and a new public school program to be replicated in other schools. The school, called P-TECH (Pathways in Technology Early College High School) is preparing students to fill entry-level jobs in technology fields or provide them with the foundation for ongoing learning in a four-year college in the STEM disciplines.

Higher Ed - IBM has 200 academic partnerships focused on Big Data analytics with schools such as Yale, Northwestern and Michigan State universities (in addition to our 30,000 academic partnerships overall). The mission: develop Big Data analytics curriculums primarily in their business schools. IBM

works closely with professors in support of curriculum materials and case studies and provides guest speakers and faculty awards to accelerate degree program development. A key example of this curriculum: Working with Fordham University in New York City, IBM partnered with faculty to create a Center for Digital Transformation and two degree programs in analytics: the Masters of Science in Business Analytics and Marketing Intelligence.

Advanced Analytics Center **-** IBM recently unveiled a first of its kind Advanced Analytics Solutions Center in Columbus, Ohio, which aims to create 500 new analytics jobs. The center will serve as an innovation hub to advance analytics skills, drawing on the expertise of local businesses, educational institutions and industry partners. IBM is partnering with Ohio State University to develop new analytics curricula at the undergraduate, graduate and executive education levels. The Columbus center is the most recent addition to our existing centers, which include both Dallas and Austin.

Watson Joins Rensselaer Research Team – IBM is providing a modified version an IBM Watson system to Rensselaer Polytechnic Institute, making it the first university to receive such a system. The arrival of the Watson system will enable new leading-edge research at Rensselaer, and afford faculty and students an opportunity to find new uses for Watson and deepen the systems' cognitive capabilities.

Watson Case Competitions – IBM works with professors and local businesses to create project-focused case studies for students to gain hands-on experience in Big Data, analytics and cognitive computing. For example, the University of Rochester's Simon School of Business launched the first-ever Watson academic case competition where MBA students identified critical areas in which Watson technology could be beneficial such as crisis management, mining and transportation. This was followed by competitions at Cornell University, the University of Connecticut and the University of Southern California, where more than 100 business and engineering students combined their skills to recommend new uses for Watson, including an innovation that helps doctors identify people who may be suffering from Post-Traumatic Stress Disorder.

Interning with Watson - This past summer, IBM brought students into its labs to learn about and develop applications for Watson's ground-breaking analytics technology. IBM Watson interns worked directly with clients and IBMers on real-world projects.

Professional Training - IBM actively trains current workers on Big Data skills. This past year, IBM held 1,200 Big Data boot camps at client, partner and university sites and trained over 2,400 IT professionals and students on the latest data management techniques; this year IBM launched BigDataUniversity.com to help students learn Hadoop, stream computing, and Big Data analytics skills. Over 13,000 students have enrolled over the past six months.

These collaborations between IBM and top U.S. universities are building a workforce of professionals and are creating jobs now. Consider North Carolina State University, whose Master of Science in Analytics program has generated the ultimate outcome – 90 percent of its graduates receiving job offers from data-hungry employers.

Privacy matters

Privacy must be considered in the design of Big Data systems. Importantly, realizing the promise of Big Data does not require the sacrifice of personal privacy. In many cases in Big Data projects, the data being aggregated is non-regulated, de-identified information with no need to re-identify to derive value. When personal information is used, organizational processes and technology can protect privacy.

Organizational practices include a systematic way of thinking and acting proactively and responsibly about the use of data. When organizations using personal information take privacy into account from the start and design in privacy protection practices, they act as better stewards of information and help individuals make more informed choices. For example, IBM practices "privacy by design" and has included privacy considerations as we have developed our new sense-making analytics technology. We welcome the growing ranks of organizations working to adopt this approach.

Furthermore, IBM and other companies are working with the Future of Privacy Forum on a consumer trust seal, authenticated by a third party, to give consumers confidence in smart grids. IBM believes that industry codes of conducts should promote transparency with customers. We have been urging better communication for a long time – as far back as 1999 we decided to withhold advertising dollars from North American websites that did not post their privacy policies.

Although these considerations are critical as technology is changing rapidly, a major trend in consumer expectations is also evolving. Indeed, Americans' views of online privacy are shifting as they use multiple social networking tools every day to describe their personal thoughts, actions and emotions very publicly, such as through online profiles, tweets, blog posts, and photos.

Our experience is that when people understand how information is used, have the ability to set data usage policies, and enjoy the benefits of the analysis, they tend to see a helpful tool rather than a privacy violation. Information – including personal information – is becoming a helpful tool on a grand scale. When combined in great quantities and put through sophisticated analysis, general information can be used to address some of society's most pressing problems. For example, the Center for Medicare and Medicaid Services uses big data in a new program to prevent health care fraud. A sophisticated and evolving set of algorithms is used to highlight problem areas and generate fraud alerts that allow the agency to direct attention and resources, and take appropriate action. Increasingly, CMS will be in a position to interrupt claims that should not be paid. The need to deter and detect is great, with health care improper payments totaling in the hundreds of millions of dollars.

What is the role of government in research?

Federal research investment in high performance computing is critical to Big Data. Research both creates new ideas and insights and trains students with critical skills for later employment. Industry needs fundamental, exploratory research as we push the boundaries of programmable systems with our high performance systems. We need research into numerous areas including: system design for optimized handling of the volume, velocity, and variety of data described earlier; software research to understand how to create dynamic and flexible software-defined environments; IT infrastructure research to build programmable, optimized and automated environments. New knowledge, practices, and infrastructures will enable the discoveries and innovation which are the foundation of U.S. competitiveness.

In addition, IBM strongly supports the reauthorization of the Department of Energy High-End Computing Revitalization Act of 2004 to be offered by Representative Hultgren. This bill will improve high-end computing research and development at the DoE and strengthen government-industry partnerships for integrated research, development and engineering of exascale platforms. IBM has a long history of successful partnership with DoE. Through our joint work, computational simulation has been established as an essential, broadly available tool for scientific inquiry. World leadership for the U.S. in HPC has been grown and sustained, and HPC has become a true engine for innovation. The challenge ahead is to continue this growth, recognizing that Big Data is an intrinsic aspect of high performance computing. There is not an "either/or" choice between HPC and Big Data.

The explosion of data is having a significant impact on the focus of HPC. HPC workloads are evolving from single "heroic" calculations to complex simulations with different time and space scales involving

multiple scientific disciplines and several types of computational algorithms. These complex simulations interact with each other using mountains of data. Research is needed into systems which can integrate computation, data and storage to successfully exploit Big Data to address many enormously important scientific, social and commercial workloads.

Past federal investments in HPC-related research, particularly at DOE's national laboratories in partnership with the industry and academia, have underpinned mission critical supercomputers in many Federal agencies, including the Department of Energy, for both the Office of Science and the National Nuclear Security Administration; the Department of Defense; the National Aeronautics and Space Administration; the National Oceanic and Atmospheric Administration; and U.S. intelligence agencies.

What is the role of government in workforce development?

The Professional Science Masters program supported by National Science Foundation is particularly relevant to the new era of computing. A Professional Science Masters (PSM) degree is a new graduate degree designed to allow students to pursue advanced training in science or mathematics, while simultaneously developing workplace skills valued by employers. According to the Council of Graduate Schools, enrollment in PSM degrees increased 22% between 2010 and 2012 and was dominated by four fields of study: computer/information sciences (21%), biotechnology (16%), environmental sciences and natural resources (14%), or mathematics and statistics (14%).

Additionally, Congress should reauthorize the Carl D. Perkins Act and the Federal Work Study Program to align to labor needs in Big Data. While this is unlikely to fall in the Science Committee jurisdiction, it is important for the growth of STEM (science, technology, engineering and math) education and jobs.

In the U.S., community college graduation rates hover at or about 25 percent. At the same time, there are 28 million middle skill jobs – those requiring postsecondary degrees – currently available in the U.S. which pay close to $40,000 per year on average. Over the next 10 years, 14 million jobs – a 50 percent increase – will be created for students with "middle skills."

With so many high school graduates in need of remediation rather than prepared with the skills and credentials needed to fill these jobs, we must refocus our efforts on strengthening the education system in order to make the U.S. more competitive.

In the U.S., career and technical education (CTE), once called vocational education, is the core program linking school to career. Federal funding under Perkins provides over $1 billion to schools. Unfortunately, Perkins is not currently structured to meaningfully address the skills gap that we face.

Similarly, the federal government invests nearly $1 billion in the Federal Work Study Program (FWSP) every year, providing nearly 800,000 undergraduates the opportunity to gain work experience while earning critical financial aid. To maximize its potential, FWSP should not be viewed solely in the lens of college affordability, but also as a way to prepare students to be successful in a career.

If 10 percent (or $200 million) of the $2 billion invested in these two programs annually were targeted in new ways, CTE programs could be reshaped to help significantly bridge the skills gap. For example, over a 10 year period, a new skills-based apprenticeship program could better prepare more than 1,000,000 young people for the jobs of the 21st Century and reinvigorate the American economy.

IBM supports reauthorizing Perkins funding to align to labor market needs in high-growth industry sectors; improve CTE programs with strong collaborations among secondary, postsecondary institutions and employers (like the P-TECH example cited earlier); and create accountability

measures that provide common definitions and clear metrics for performance of CTE programs to improve academic outcomes while building technical and employability skills of participants.

Conclusion

There exists today an overabundance of data. Leveraging the capabilities presented by this new era of cognitive computing presents us with the opportunity to provide benefit in many areas. It will be a key factor in determining which countries pull ahead economically and which fall behind, which cities attract knowledge workers and business development and which businesses thrive. I believe in the value and power of information and technology. The United States should continue and increase its support of advanced computing and its investment in building its workforce to seize the value that Big Data, Analytics and next generation computing offer.

Thank you for the opportunity to appear before you today to provide this testimony. I welcome your questions.

Dr. David F. McQueeney
Vice President, Technical Strategy & Worldwide Operations
IBM Research
Yorktown Heights, NY
davidmcq@us.ibm.com

Dave McQueeney is Vice President, Technical Strategy & Worldwide Operations at IBM Research. Dave is responsible for setting the direction of IBM's overall Research Strategy across twelve worldwide labs and leading the global operations and information systems teams. In this capacity, Dave leads the annual creation of the Global Technology Outlook, 3 - 10 year technology projections which guide IBM's R&D directions along with its acquisition strategies.

Dave's background covers a wide range of disciplines ranging from solid state Physics, to high-speed interconnect design, to distributed software development tools, to participation in a startup software company in Scientific Data Analysis, to Government-specific industry solutions. Dave has spent about half of his career as a researcher and research executive, and half in IBM's customer-facing units including Global Services, Sales and Distribution and Software Group. Dave has been the CTO for IBM's Federal business, the Global Government Solutions General Manager and leader of the IBM Federal Systems Integration services unit.

Throughout his career, Dave has driven strong connections between IBM Research and IBM's clients, as well as the other units of IBM. Dave was recognized by Consulting Magazine as one of the top 25 consultants for 2002, citing his work to make the innovations of IBM Research directly available to customers via IBM's Global Business Services unit.

Dave has held a number of other significant positions in IBM Research, including Director of the IBM Zurich Research Laboratory, Vice President of Communication Technology, and Vice President of Software.

He joined IBM in the Research Division in 1988. He holds an M.S. and Ph.D. in Solid-State Physics from Cornell University, and an A.B. in Physics from Dartmouth College.

Dave was recognized as one of the "Fed 100" top leaders in the Federal community for 2006 by Federal Computer Week magazine.

Chairman BUCSHON. Thank you, Dr. McQueeney.

I now recognize Dr. Rappa for five minutes for his testimony.

TESTIMONY OF DR. MICHAEL RAPPA, DIRECTOR, INSTITUTE FOR ADVANCED ANALYTICS, DISTINGUISHED UNIVERSITY PROFESSOR, NORTH CAROLINA STATE UNIVERSITY

Dr. RAPPA. Good morning, Chairman Bucshon, Chairman Massie, Ranking Member Lipinski, Ranking Member Wilson and other Members of the Subcommittee. I appreciate the opportunity to be here this morning to speak with you about data analytics and the role institutions of higher learning can play in advancing the field.

I am going to draw this morning's testimony on my own behalf as a professor and director of a research institute, educational institute for over the past 25 years.

I think it is important to start with the fact that the world is changing around data very rapidly and our ability to productively use it becomes a very central part of what we do as a society today, as has been heard already. A generation ago, data was scarce, expensive, time consuming to collect and difficult to analyze. Today, data is everywhere.

Advances in computer technology and powerful analytic tools make it possible not only to collect vast quantities of data but also analyze and draw insights from data to solve pressing problems from increasing operational efficiency to combating fraud, to better health care, to protecting national security. Data is everywhere. The question is, how well are we prepared to use it? We have the data, the technology, the methods and tools, all of which continue to advance. The national challenge, in my view, going forward will be our ability to educate a data-savvy workforce that has the analytical skills to put data into action. Estimates of the talent gap as we have heard are large and growing.

This is a dire but solvable problem. As we have shown at NC State, working closely with employers and focusing on their needs, we can produce the kind of talent that is so desperately needed today. We do it quickly in just 10 months with a domestic student population ranging from their early 20s to their late 50s, many of whom are returning to school. We have done this now for six years economically with consistently high student outcomes using a sustainable and scalable business model based on self-financed tuition.

What it comes down to is creative innovation, how we organize graduate education, allowing us to engage with employers more productively to yield high-quality results in the skills and readiness of our graduates.

I encourage the Committee to focus its attention on workforce needs, to encourage the government to seek out innovation in higher education and to promote new and novel learning models. This is a solvable problem. With the proper incentives, focused resources, open collaboration with industry, we can produce the analytics professionals needed to extract value from big data and to move the economy forward. As I said, we have done this ourselves now for 6 straight years to great effect. We will graduate a class in a matter of another week, 80 students in the Master of Sciences and Analytics Program, with already 95 percent of them placed in

jobs. They are literally the most sought after and highest-paid graduates of the university.

So we can do this. It is a solvable problem. Thank you again for your time. I will be glad to answer any questions.

[The prepared statement of Dr. Rappa follows:]

Testimony of

Michael Rappa, Ph.D.
Executive Director and Distinguished University Professor
Institute for Advanced Analytics
North Carolina State University

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 Bucshon, Chairman Massie, Ranking Member Lipinski, Ranking Member Wilson, and members of the subcommittees. Thank you for the opportunity to speak with you about big data, analytics, and the important role institutions of high education play in advancing the field.

My name is Michael Rappa and I am the Executive Director of the Institute for Advanced Analytics and Distinguished University Professor in the Department of Computer Science at North Carolina State University. I am responsible for overseeing the nation's first graduate degree in Analytics, which was founded in 2007 and currently enrolls 80 students each year. I also co-direct a government sponsored research project that seeks to advance the science of security in cyberspace.

My testimony today draws on my 25 years of experience in graduate education and research and, in particular, my experience over the past seven years leading the Master of Science in Analytics degree program at NC State. I will focus on the role different academic disciplines play in the field of Big Data Analytics, the success of the Institute for Advanced Analytics in partnering with industry to produce analytics professionals and lessons for other universities, and conclude with recommendations for policymakers in their support of Big Data and developing the analytics capabilities of the workforce.

What is the role of different academic disciplines in Big Data and Analytics?

Many of the things we do each day as individuals and organizations generate data. How much data? No sooner would I give you an estimate than that number would be surpassed. The global use of connected digital devices—computers, smart phones, tablets, and the like, that propagate numerical, voice, image and text data—is growing just that fast. With all of this data come both necessities and opportunities, which have given rise to what today is called "Big Data" and the concomitant need for Analytics; namely, the tools, methods and applications for drawing insights from large quantities and varieties of data. Big Data is a relative term that suggests a realm of data that pushes the upper limits of existing capabilities

in computation, storage and analysis. In this sense there always has been and always will be Big Data.

What's interesting about Big Data today is less about its inherent bigness then its ubiquity. A case can be made that Big Data is both a systemic and transformative technological phenomenon—the latest phase in a decades long process that began with the rise of modern digital computing in the 1940s. The computerization of business processes and later personal computing had the effect of creating first large and then highly dispersed stores of data. Subsequently, the rise of the Internet and World Wide Web facilitated the movement of data, as well as created huge amounts of data through its use—dramatically increasing the amount, immediacy, and the interconnectivity of data flowing from an increasing number of users and digital devices that now extends far beyond computing in the conventional sense. Thus, the attention to Big Data and Analytics should come as no surprise. It is part of a natural progression that has grown out of the digital world we live in today. The ability to analyze and draw insights from data is without question important. In the opinion of some observers, it has the potential to become a defining factor in how well organizations perform their missions in the future.

Looking at the academic landscape, certain disciplines play a critical role in pushing forward the research frontier: work in computer science, statistics, operations research, and applied mathematics is particularly relevant. But it doesn't stop there. Just as Big Data and Analytics will touch almost every corner of the economy, so too will it draw in researchers from a wide spectrum of disciplines and sub-disciplines like the Internet did a generation earlier. I encourage the Committee to embrace an expansive perspective in its support of Big Data research in the academic community. When the hype around Big Data eventually dies down, we will nonetheless find ourselves in a world saturated with data. Our ability to draw value from this data will be as important as it ever was to every sector of the economy. Only through sustained investments in the academic community will we ensure the technology and workforce necessary to leverage data to its fullest extent.

Research is needed to continue to advance our understanding of how to collect, store and process extremely large amounts of data, where the definition of "large" is a rapidly expanding quantity that knows no bounds. Many areas of computation and data analytics are relevant, including high performance computing, databases, networking, machine learning, data mining, algorithms, data visualization, natural language processing, optimization, geospatial analytics, remote sensors, data privacy and security, among others. Again, I suggest taking a broad view. It's important to recognize that the academic landscape is fluid with new disciplines and sub-disciplines emerging over time. Under the right conditions these new frontiers of knowledge will become woven into the institutional framework of universities through the establishment of centers, institutes and, over time, possibly as new schools and colleges.

How academic disciplines combine and split would matter little were it not for the organizational boundaries that are created or dissolved in the process. The domain of Big Data and Analytics is not unusual in how it spans disciplines, but we need not be so concerned about crossing disciplinary boundaries with respect to research. Many incentives are in place to overcome the barriers, and the current generation of faculty has grown up in an environment where multidisciplinary work is not uncommon, if not the norm. At NC State, for example, we have a large multidisciplinary project on the analytics needed to advance a science of security. There are natural hurdles researchers will encounter working across disciplines, certainly, but it is something that can be overcome and is overcome on a regular basis.

More attention should be paid to how the interplay of disciplines can facilitate or inhibit the

development of educational programs that fulfill the needs of employers and an evolving workforce. Academic disciplines and the related degree majors don't always align well with occupational roles. There is a need to pull together disparate disciplines in graduate education to produce the kind of technical professionals employers want, such as data scientists. This can be difficult because it goes beyond faculty collaboration to the larger challenge of reorganizing how universities design and deliver degree programs. To overcome the projected scarcity of data scientists, universities need to reduce or eliminate organizational boundaries so that students can acquire the bundle of skills they will need to succeed in the workplace.

To address this challenge, North Carolina State University partnered with SAS in 2006 to create the Institute for Advanced Analytics and embark on a grand experiment in designing a new graduate degree in Analytics, the first of its kind. By every measure, the experiment has been an unambiguous success. The program's enrollment has grown to 80 masters students annually, with over 95-percent self-financed—allowing the program to be self-sustainable based on tuition revenue. As a group they are among the most sought after and highest paid graduates of the university. Students come from a variety of disciplines and range in age from their early twenties to their late fifties. The majority of students return to school from the workplace, and 80- to 90-percent are U.S. citizens or permanent residents. This year the Institute will record its sixth consecutive year in which 90-percent or more of the graduates are placed by graduation in positions as analytics professionals—despite the worst economy since the Great Depression. Key to this success is close collaboration with industry combined with an innovative way of organizing graduate education.

The Institute for Advanced Analytics

When the Institute for Advanced Analytics was proposed in April 2006, it was conceived as an educational initiative intended to address the growing demand for professionals who can draw insights from (what we now call) Big Data. At the cornerstone of the proposed Institute was the idea of creating the nation's first Master of Science in Analytics (MSA) degree. The proposal was unusual in that it was adopted as a greenfield project that allowed us to start with first principles when creating the degree program. Furthermore, the Institute was launched as a university-wide collaboration—organizationally independent of the colleges—to give it maximum flexibility to blend together whatever disciplines were needed in the education without being anchored to a single department or college by default.

The MSA curriculum was designed by a large group of faculty from more than a dozen academic departments across six colleges, in close consultation with colleagues at SAS and a large industry council. We shed conventional thinking about degree programs as much as we possibly could. Instead of starting with a menu of core courses and electives—the basic inputs—the design process began with a clear focus on desired students outcomes, positioning employers as the core customer and seeking to understand what they look for when hiring analytics talent. Analyzing job descriptions and talking with employers led to a balanced perspective embodying five core objectives for student outcomes that guided the MSA:

- Technical skills
- Teamwork skills
- Communication skills
- Tool skills
- Domain knowledge

When it comes to analytics professionals, employers want to hire individuals with the requisite technical skills to clean, analyze and interpret data, but it doesn't end there. Employers also want people who can perform well in multi-functional teams, who have strong verbal and written communication skills, who have experience programming with industry standard software tools and programming languages, and who have prior knowledge of the business domain whenever possible.

The MSA curriculum was laid out with all of these objectives in mind, not as a series of courses, but as a single integrated learning experience. Faculty members are engaged not by the course, but instead in customized lecture streams that align closely with their expertise. This allows us to use a larger number of faculty and to integrate subject matter across different disciplines with greater flexibility. The result is quite different from conventional programs. The MSA is optimized to achieve the five core objectives by leveraging both the content and programmatic structure of the learning experience. While the focus on employers and the inclusion of professional skills development aligns the MSA with the larger Professional Science Masters (PSM) degree movement, it takes the idea a step further than most, given the Institute's flexibility in crafting the curriculum.

The MSA is a cohort-based 30-credit hour curriculum spanning ten months. It uses the intensity of a condensed format to immerse students in the learning process. MSA students pursue the degree full-time—literally. The program calendar runs 9:00 to 5:00, five days a week for most of the ten months (vacation time is more limited than the normal student calendar). When students aren't in class, they are typically working on a long-term team project (the practicum) or in study groups that rotate membership every five weeks. A premium is placed on learning by doing. The result is a highly structured team-based learning experience that has more of the look and feel of a normal workplace.

The MSA places heavy emphasis on "soft" skills. Professional skill training is woven into the learning experience throughout the curriculum. Students receive training in public speaking, technical writing, teamwork, leadership, project management, and conflict resolution, among other areas. The soft skills are not treated as non-credit add-ons or extra-curricula activities. Instead, these skills are taught as an integral component of the curriculum and students are measured and tested against them.

Teamwork is extremely important in most organizations today. Teams form the basis for most of the work students perform in the MSA program. The cohort driven format enables the students to work in teams continually across the entire learning experience. The team structure of the curriculum and diverse student body maximizes peer-to-peer learning, which is known to be very effective and perhaps as important as direct instruction. Students are given guidance on teamwork and undergo frequent iterations of peer evaluation and feedback on their performance. The peer evaluation is both structured and open-ended and forms the basis for personalized coaching to help students improve their ability to function effectively in teams.

At the core of the MSA curriculum is a practicum—a hands-on learning experience that gives students the opportunity to conduct real-world analytics projects using data from sponsoring organizations. Students work in teams of 4-5 members each to understand an actual business problem and then clean and analyze the data. The Practicum spans seven months and culminates with a report and presentation to the sponsor. The teams perform their work under confidentiality agreements and the results are the sole property of the respective sponsors. Past projects span virtually every industry segment—advertising, banking and financial services, consulting, e-commerce, energy, entertainment, healthcare, insurance, retail, and transportation—and a number of state and federal agencies. The Institute currently conducts 17 practicum projects each year. This year's sponsors included:

- Allscripts (Chicago, IL)
- Caterpillar (Peoria, IL)
- Central Intelligence Agency (McLean, VA)
- Federal Communications Commission (Washington, DC)
- GE Energy (Atlanta, GA)
- Global Knowledge (Cary, NC)
- GlaxoSmithKline (Research Triangle Park, NC)
- Hanesbrands (Winston-Salem, NC)
- Houston Astros (Houston, TX)
- Inmar (Winston-Salem, NC)
- Lowe's Home Improvement (Charlotte, NC)
- M&T Bank (Buffalo, NY)
- Monsanto (St. Louis, MO)
- North Carolina Department of State Treasurer (Raleigh, NC)
- Procter & Gamble (Cincinnati, OH)
- U.S. Postal Service (Washington, DC)

The Institute's open solicitation of practicum proposals keeps the curriculum grounded in the kind of problems facing industry, which continue to evolve in character and complexity. Dozens of proposals are received each year from an equally large number of organizations representing almost every industry segment. The proposals inform decisions made about the kinds of methods and tools we choose to emphasize in the curriculum, given the limits of time in a 10-month program cycle.

In both the practicum and class instruction, students learn analytics by using industry standard software tools. The Institute's collaboration with SAS is particularly important in this regard. Enterprise class analytics tools are a challenge to deploy and maintain in university settings, given the lack of standardized platforms and operating systems. Through the generous ongoing technical support of SAS, not only do students have the opportunity to use an industry leading tool set, they can (and frequently do) complete numerous SAS certifications in route to their degree. The significant market value of such industry certifications is well documented and further underscores the readiness of MSA graduates when they land on the job.

The MSA program was designed from top to bottom with the intention of producing the kind of analytics talent employers seek to hire. Since the beginning, the focus remained on how to achieve successful student outcomes by understanding employer needs. It should be no surprise that no sooner did we start to produce graduates from the program than those graduates attracted the strong interest of employers. The Institute has achieved over 90-percent job placement by graduation each year since the first MSA cohort entered in 2007. The current class of 81 students, which will graduate in a few weeks, will enter new positions with over 30 different employers in 12 states and the District of Columbia—at record high salaries. Nearly three-quarters of the class had two or more job offers, and over half had three or more offers of employment. Five-year benchmark studies put MSA student outcomes, in terms of placement and salary, on par with some of the country's most prestigious universities.

Six years later there are now at least two-dozen graduate degree programs in the U.S. that focus explicitly on Big Data and Analytics, and new ones are announced with regular frequency. Some programs, like one at Louisiana State University, have sought to closely replicate our model, while other

programs, like one at Northwestern, have adopted similar components like the practicum and industry standard tools. It's common for schools to work with one or more supportive companies like SAS, IBM and others, to help them deploy industry standard tools or share business problems and data with their students. While many programs are situated in business schools, it is yet to be seen whether organizationally this is the best approach in balancing the disciplinary knowledge needed to produce the kind of data scientists required by industry.

Conclusion

The value of big data lies in our ability to extract insights and make better decisions. The acute shortage of analytics professionals and data-savvy managers will be addressed most successfully through creative partnerships between industry, government and universities. There are several efforts underway, but we must intensify and accelerate the national investment in proven models.

I had the pleasure of serving as academic co-chair on the TechAmerica Foundation's Big Data Commission, along with many representatives of the Foundation's member companies. Among the Commission's findings was a recommendation to continue to invest in research and development of advanced computing technologies that can effectively process not only the vast amounts of data being continually generated but also the various types. Those investments should focus on key government priorities such as education, fraud and abuse, cybersecurity, healthcare and public safety.

The Commission also recommended a strong focus on skill development in the workforce. Public-private partnerships should be strengthened and expanded to invest in skills-building initiatives for the workforce in the area of Big Data. While many of the Commission's recommendations were directed specifically to the federal government's own needs, it also encouraged the development of data-intensive degree programs and scholarships to prepare a new generation of data scientists.

The Institute for Advanced Analytics has a proven track record over the past six years with the Master of Science in Analytics degree program that shows we can succeed in educating a new generation of data savvy professionals to satisfy the needs of employers. We can do it quickly with an intensive and highly targeted educational format that yields consistent student outcomes. The program attracts a diverse, high quality, domestic student population and yet runs with a sustainable, self-financed business model. PSM programs like ours show that graduate education can be designed effectively in collaboration with employers to fill the skills gap in critically important areas like data science and cybersecurity.

Thank you for the opportunity to appear before you today to provide this testimony. I welcome your questions.

Biographical Sketch

MICHAEL RAPPA

Michael Rappa is Executive Director of the Institute for Advanced Analytics and a member of the faculty in the Department of Computer Science at North Carolina State University. As head of the Institute, he leads the nation's first and preeminent Master of Science in Analytics as its founder and principal architect. Before joining NC State as Distinguished University Professor in 1998, for nine years he was a professor at the Massachusetts Institute of Technology.

Dr. Rappa has 25 years of experience as a professor working across academic disciplines at the intersection of management and computing. An accomplished researcher and instructor, his passion is to bring an entrepreneurial and forward-thinking mindset to innovation in higher learning. His current role is to prepare a new generation of data savvy professionals for leadership in a digital world.

In addition to his duties as director, Rappa is co-principal investigator of the Science of Security Lablet, a large multidisciplinary research project sponsored by the U.S. National Security Agency. The project is run in parallel with sister labs at Carnegie Mellon University and the University of Illinois at Urbana-Champaign.

A study published in 2006 in the British journal R&D Management identified Rappa as a leading scholar in the field of technology management, ranking him in the 99th percentile among over 9,000 authors in terms of research productivity in top journals over the past 50 years. His research has been cited on three occasions as an outstanding contribution to the field, and his early work on business models is one of the most often cited and widely read publications on the subject.

Rappa is perhaps best known as the creator of *Managing the Digital Enterprise*, an innovative and award-winning educational Web site devoted to the study of management in the digital world. Launched in 1999, originally as the foundation for a course he taught, the site is a valued resource used by several million learners and hundreds of university instructors from around the world. In 2010, Rappa presided as general co-chair at the *19th International World Wide Web Conference*, the premiere annual gathering of the Web research community.

NC State has recognized Rappa on several occasions for his contributions to teaching and service. He is the recipient of the Outstanding Extension Service Award, the Award for Graduate Teaching Excellence, and the Gertrude Cox Award for Innovative Excellence in Teaching and Learning with Technology. He is also winner of the MERLOT Award for Exemplary Online Learning; a four-time recipient of the IBM Faculty Award; and twice a finalist for the Epton Prize.

Rappa began his teaching career at the University of Minnesota, where he earned his doctorate in 1987.

MICHAEL RAPPA

Biographical Statement

Michael Rappa is Executive Director of the Institute for Advanced Analytics and a member of the faculty in the Department of Computer Science at North Carolina State University. As head of the Institute, he leads the nation's first and preeminent Master of Science in Analytics as its founder and principal architect. Before joining NC State as Distinguished University Professor in 1998, for nine years he was a professor at the Massachusetts Institute of Technology.

Dr. Rappa has 25 years of experience as a professor working across academic disciplines at the intersection of management and computing. An accomplished researcher and instructor, his passion is to bring an entrepreneurial and forward-thinking mindset to innovation in higher learning. His current role is to prepare a new generation of data savvy professionals for leadership in a digital world.

In addition to his duties as director, Rappa is co-principal investigator with Dr. Laurie Williams of the Science of Security Lablet, a large multidisciplinary research project sponsored by the U.S. National Security Agency. The project is run in parallel with sister labs at Carnegie Mellon University and the University of Illinois at Urbana-Champaign.

A study published in 2006 in the British journal *R&D Management* identified Rappa as a leading scholar in the field of technology management, ranking him in the 99th percentile among over 9,000 authors in terms of research productivity in top journals over the past 50 years. His research has been cited on three occasions as an outstanding contribution to the field, and his early work on business models is one of the most often cited and widely read publications on the subject.

Rappa is perhaps best known as the creator of *Managing the Digital Enterprise*, an innovative and award-winning educational Web site devoted to the study of management in the digital world. Launched in 1999, originally as the foundation for a course he taught, the site is a valued resource used by several million learners and hundreds of university instructors from around the world. In 2010, Rappa presided as general co-chair at the 19th International World Wide Web Conference, the premiere annual gathering of the Web research community.

NC State has recognized Rappa on several occasions for his contributions to teaching and service. He is the recipient of the *Outstanding Extension Service Award*, the *Award for Graduate Teaching Excellence*, and the *Gertrude Cox Award for Innovative Excellence in Teaching and Learning with Technology*. He is also winner of the *MERLOT Award for Exemplary Online Learning*; a four-time recipient of the *IBM Faculty Award*; and twice a finalist for the *Epton Prize*.

Rappa began his teaching career at the University of Minnesota, where he earned his doctorate in 1987.

Chairman BUCSHON. Thank you for your testimony.

I now recognize our final witness, Dr. Jahanian, for five minutes for his testimony.

TESTIMONY OF DR. FARNAM JAHANIAN, ASSISTANT DIRECTOR FOR THE COMPUTER AND INFORMATION SCIENCE AND ENGINEERING (CISE) DIRECTORATE, NATIONAL SCIENCE FOUNDATION

Dr. JAHANIAN. Good morning, Chairman Massie, Chairman Bucshon, Ranking Members Wilson and Lipinski, and Members of the Subcommittee. It is my pleasure to be back here to discuss the next generation of computing and big data analytics.

Today we live in an era of data and information enabled by advanced technologies that surround us. Data is generated by modern experimental methods, scientific instruments such as telescopes and particle accelerators, large-scale simulators, Internet transactions, email, video images, clickstreams, and widespread deployment of sensors everywhere. Approximately 90 percent of the data in the world today were created in the last two years alone. However, when we talk about big data, it is important to emphasize not only the enormous volume of data being generated but also the velocity, heterogeneity and complexity of data that now confronts us.

Why is big data important? Several others have alluded to this already. Data represents a transformative new currency. Big data is increasingly important to all facets of our Nation's discovery and innovation ecosystem. First, insights and more accurate predictions from large and complex collections of data are creating opportunities in new markets, driving the creation of IT products and services and boosting the productivity of businesses. Second, advances in our ability to store, integrate, and extract meaning and information from data are accelerating the pace of discovery in almost every science and engineering discipline. Third, big data has the potential to solve many of the Nation's most pressing challenges from health care and education to cybersecurity and public safety, yielding enormous societal benefits and ensuring sustained U.S. competitiveness.

Let me share with you just a few examples of the promise of big data. These are all grounded in research that is funded by the Federal Government or by the private sector, the work that is done in the private sector. By integrating biomedical, clinical and scientific data, we can predict the onset of diseases and identify unwanted drug interactions. By coupling roadway sensors, traffic cameras, individual GPS devices, we can reduce traffic congestion and generate significant savings in time and fuel. By accurately predicting natural disasters such as hurricanes and tornadoes, we can employ lifesaving and preventative measures that mitigate their potential impact. By correlating disparate data streams through text mining, image analysis and face recognition, we can enhance public safety and public security. By integrating emerging technologies such as MOOCs and inverted classrooms with knowledge from research about how people learn, we can transform formal and informal education.

What does this mean for scientific discovery? Data-driven discovery, also called the fourth paradigm, is revolutionizing scientific

exploration and engineering innovations. It enables extraction of new knowledge, provides novel approaches to driving discovery and decision-making, yields increasingly accurate predictions and provides deeper understanding of causal relationship based on advanced data analysis.

What is government doing to ensure we harness this potential? As it was mentioned already, in 2011 U.S. Networking and Information Technology Research and Development Program, also called NITRD, formed a big data senior steering group to identify, initiate and coordinate big data research and development activities across the government to ensure that Federal agencies, the scientific research enterprise, and public maximally benefit from data-driven discovery. In March 2012, the National Big Data R&D Initiative was launched, focusing the steering committee group's focus on the tools, technologies and human capital needed to move from data to knowledge to action. We see exciting new partnership opportunities with the private sector, state and local governments, academia and nonprofits.

At NSF, we have identified four major investment areas that address current challenges and promise to serve as the foundation of comprehensive long-term agenda: first, investment in foundational research to advance big data techniques and technologies; second, support for building new interdisciplinary research communities; third, investment in education and workforce development; and finally, development and deployment of cyber infrastructure to capture, manage, and analyze and share digital data.

I should add that NSF's investment in cyber infrastructure includes advanced computational resources that support data-enabled science. In particular, the newly dedicated Blue Waters, Stampede and Yellowstone supercomputers will expand our Nation's computational capabilities significantly.

In summary, big data represents enormous opportunities for our Nation. Investments in big data research and education will advance the frontier of knowledge, further fostering innovation, creating new economic opportunities, and yielding new approaches to addressing national priorities.

Thank you again for this opportunity. I would be happy to answer any questions.

[The prepared statement of Dr. Jahanian follows:]

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)*.

Dr. Farnam Jahanian, Assistant Director, Computer and Information Science and Engineering (CISE)

Dr. Farnam Jahanian serves as assistant director for the Computer and Information Science and Engineering (CISE) Directorate at the National Science Foundation. He holds the Edward S. Davidson Collegiate Professorship at the University of Michigan, where he served as chair for computer science and engineering from 2007 to 2011.

Over the last two decades, Jahanian has led several large-scale research projects, studying the growth and scalability of the Internet infrastructure that have 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 recently 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 Jahanian co-founded in 2001. Jahanian served as chairman of Arbor Networks until its acquisition by Tektronix Communications, a division of Danaher Corporation, in 2010.

The author of over 90 published research papers, 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, including a National Science Foundation Faculty Early Career Development (CAREER) Program award, the Amoco Teaching Award, an IBM Outstanding Technical Innovation Award and the governor's University Award for Commercialization Excellence. He has been an active advocate for regional economic development efforts over the last decade, working with entrepreneurs, frequently lecturing on the topic, and serving on numerous advisory boards. In 2009, he was named Distinguished University Innovator at the University of Michigan. 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).

Credit: *University of Michigan*

Chairman BUCSHON. Thank you for your testimony. I would like to thank all the witnesses for their testimony. I am reminding Members that Committee rules limit questioning to five minutes, and the Chair at this point will recognize himself for five minutes to start the questions.

First, Dr. Jahanian, the Administration announced their Big Data Research and Development Initiative in March 2012 including $200 million in new commitments for big data research initiatives. However, the National Science Foundation, Department of Defense, Department of Energy, and other agencies have had significant research programs and data analytics that predated the initiative. How has the Administration's initiative changed the ways these agency research programs are coordinated and are we effectively managing and leveraging our research investments across agencies?

Dr. JAHANIAN. Thank you for your question. You are absolutely right that it is not that suddenly last March we woke up and said boy, data is really important, we need to do something about it. There has been significant investment by the Federal sector and private sector in areas having to do with data. The challenges we face are many—stewardship of digital data and software, for example. Many data sets, as was mentioned, are too poorly organized or also unstructured. Many data sets are heterogeneous. The utility of data is also limited by our ability to interpret them. Many data are being collected at a scale that we can't even store them, let alone analyze them. Also, large and linked data sets may be exploited to identify individuals and so there are also the privacy issues. So there are enormous challenges that we face.

As you alluded to, on March 29, 2012, OSTP in concert with a number of Federal agencies launched the national Big Data Research Initiative. It expands the scope of our activities in several directions, for example, state-of-the-art core technologies that we need to collect, store, preserve, manage and analyze data, harnessing these technologies to accelerate pace of discovery, supporting responsible stewardship, for example, and sustainable business models for big data.

There are a number of cross-coordination efforts taking place under NITRD. Let me start with NSF. All NSF directorates, for example, are participating in this. A multidisciplinary panel of experts are making a recommendation on funding of this. Furthermore, big data is being coordinated through a senior steering group that reports to the assistant directors at NSF for all the coordination because it involves every science and engineering discipline.

As far as the Federal Government is concerned, the Big Data R&D Initiative is coordinated through the NITRD Subcommittee. As you know, I Chair the Subcommittee. There is a senior steering group that regularly meets to coordinate the activities on many of the fronts that I alluded to. There are also enormous opportunities not only in terms of joint solicitations but there are a number of workshops that we are holding jointly with other agencies including NIH, NIST, DOE, DOD to advance the frontiers of knowledge and exploration in big data.

I should also mention that when it comes to this initiative, we can't forget that the private sector plays a significant role. When

we think about innovation and discovery ecosystems, not only are we talking about universities, we are talking about scientists and engineers, you know, a rich, talented labor force, investments in research and education, and of course, a vibrant private sector. So there are a number of programs that we have at NSF that attempt to connect the dots when it comes to transfer of knowledge.

Chairman BUCSHON. Thank you. I am glad to hear there is quite a bit of coordination at the Federal level because I think all of us are concerned about that, and again, investing the taxpayer dollar wisely.

Dr. Rappa, I also serve on the Education and Workforce Committee, and I have got children age 9 through 20, four of them, and I have a really strong interest in how we get young people interested in different fields of study, and obviously we have a tremendous challenge not only with this area but many others, and do you think that—what are your ideas on how we engage young people in understanding what opportunities there are in this area and what the jobs of the future might hold? I mean, how do we do that? Because, you know, when you go to a high-school class, and I talk to a lot of high-school class, people say, you know, not many people come up when you ask them what they want to be, you know, they want to analyze big data. So how do you do that? What is your recommendation?

Dr. RAPPA. Well, thank you very much for your question, and I understand exactly what you are saying, and I think that things are changing. You know, I think it is exactly true that your average 8-year-old doesn't say they want to grow up, for example, to be a statistician. It is not common, unless they are really interested in sports. Then you see a sort of nexus there because of the relationship. But I think what is changing is that it is really about producing education, in my case, at the graduate level, reaching further into the pipeline down into undergraduate education and even touching upon high school where people begin—where students begin to understand how data is really used in action. So it is really about creating, not just sort of creating knowledge or understanding but also applying that knowledge. And when our students—our whole education is driven around the application of that knowledge, and so students really understand, and increasingly undergraduates understand that this kind of graduate education is going to lead them to a very interesting, compelling professional life.

Chairman BUCSHON. Well, thank you, because I think that we do—you know, we do need to have this type of information gravitate down, even to middle-school kids to get them interested, and there is a program in Indianapolis called Project Lead the Way who I know very well that is beginning to do that at the high-school level, and it is showing some success.

But my time is expired, so I would love to talk more about that but at this point I am going to yield to Ms. Wilson for five minutes for her questions.

Ms. WILSON. Thank you, Mr. Chair.

Along those lines, can you tell me either one of you what skills are necessary for the big data workforce? I heard you say something about an analytical something. And also as you are speaking,

I would like to hear from you what role can community colleges play in preparing the next-generation workforce for big data.

Dr. RAPPA. Thank you very much for your question. I would like to try my hand at that. So what is sort of interesting and novel about what we have done around the education, we really started from scratch in building an entire new graduate degree program, and we really wanted to address this question of what skills were needed, and we focused ourselves really looking at the employer as the customer in a sense, the person, the individuals who buy our product and the students and really tried to understand the skills that they need, and really where that brings you is that there is these technical skills which are important in programming, in math and statistics, but employers really want much more than that. They want individuals who can work well in teams, who can communicate these insights to decision makers, who can actually use the tools and apply the knowledge in an organizational context, and so we have structured the whole education to build a very balanced set of skills as opposed to what I think is really the conventional approach in graduate education and to some extent undergraduate education to focus on the technical skills almost exclusively. And so really what we need to do is sort of approach the whole student. Now, I think community colleges can play a very important role because you can really begin to channel pipelines where students can go and get the prerequisite knowledge that they need, the early levels of math and statistics, before they go on to graduate education. Thank you.

Dr. MCQUEENEY. I would just like to comment that a lot of the focus in the past has been on the graduate level of education, as Dr. Rappa just talked about, and while we continue to have a strong need for Ph.D.'s and computer science and electric engineering and mathematics, the biggest skill gap that we see is at the masters level, quite frankly, of people who may not have the mathematical skills to create an entire new type of analysis of data but who have more than basic IT skills who actually can understand the implications of using different analytical techniques given a problem, given a data set with certain statistical properties, what would be the appropriate analytical technique to use, and when you apply that technique, how could you be sure that the results would be reliable and proper, and so a lot of our focus has been on creating an intermediate level of skill that has the basic understanding of how to use these tools even if it would fall on someone with more of a Ph.D. level of training to create new analytical approaches.

Dr. JAHANIAN. Representative Wilson, I want to echo something that has been said. If you think about big data, let us just step back. There are three related problems that go beyond big data. It includes all of our IT workforce, computer science, computational science and so on. These problems have to do with underproduction, which everybody recognizes, underrepresentation and then pipeline issues. Chairman Bucshon already alluded to this, that we need to worry about our high schools, we need to worry about the pipeline. I have three kids, and I know where we lose our kids, it is not in masters or Ph.D., we lose the interest of our kids in high schools and middle schools, so that has to be fixed, and there are

a number of programs that we have initiated, pilot programs that try to address that issue.

Let me share with you one anecdotal sort of evidence that provides data on this. Annualized Bureau of Labor Statistics data predicts that each year we need about 140,000 job openings. We will have 140,000 job openings in computing and broadly speaking IT-related jobs but we are only producing about 100,000 qualified individuals including masters, Ph.D., undergraduate and community colleges. In fact, many of these jobs would be available to individuals who have two year or four year degrees.

Another data point that I want to share with you is that 62 percent of all newly created STEM job openings between 2010 and 2020 will be in computing and IT. Let us not forget that. And that includes data, that includes computational skills and many of the other skills that the other witnesses alluded to. Thank you.

Ms. WILSON. Just in my 16—oh, 10, 9, 8—what would you suggest that we begin to—how do we begin to get children interested in these sort of skills? I know every little child has an iPad. They can work these computers better than adults. What do you think we can do to stimulate that all the way from K–12 and into the community colleges so we will have more IT graduates? Do you suggest we buy each one—we outfit classrooms with iPads, or what do you think?

Dr. MCQUEENEY. I think there is an intrinsic curiosity in younger folks about a lot of the tools they use to communicate with each other that have tremendously greater scalability than the tools that I use to communicate with my friends.

Ms. WILSON. Right.

Dr. MCQUEENEY. So the essence of what is a large community's opinion on a topic of interest could involve the opinions of thousands or millions of people and so I think a lot of the young folks I talk to when I visit K–12 programs or, you know, in programs like eWeek, they have an intrinsic sense not only of the device and the technology but they have a sense of the reach of that device and technology which is the beginning of an appreciation of really what we are talking about with big data, that there are trends that they can reach with that device, and I think that fires their imagination in a very powerful way.

Chairman BUCSHON. Thank you. I will now recognize Mr. Massie, Chairman Massie, for his questioning.

Mr. MASSIE. Thank you, Chairman.

So one of the questions that I have as we deal with the interface between government and private industry here is, are you aware of any government data sets that we need to get more into the public domain for usage? For instance, I think we have done a pretty good job about getting some of the mapping stuff out there but some of that map information is old, goes back to the 1940s and 1950s, and I know the government has been paying for LIDAR mapping, which is a high-resolution terrain mapping, and I am kind of concerned that that is not getting out there. Are you aware of that, and are there any other data sets that would be useful to the public that the public has paid for that we might want to work on getting out to the public?

Dr. MCQUEENEY. I think the government has done an excellent job and had many initiatives that were very focused on getting that valuable data out so it could be used. You mentioned LIDAR. I know that one of the uses that is very promising for LIDAR is to do something like an inventory of the forests in the country, to actually be able to conduct a definitive inventory. Right now, the agencies that are responsible for that use a statistical sampling technique but in a world where you can take LIDAR images and process that enormous data volume, you are able to move then from a statistical sampling basis, which is all we could do before, to a more definitive approach to get a very, very good picture of one of the more valuable natural resources that needs tremendous amounts of stewardship. So I think that is an example of a data set that could be extremely valuable. But I think in general, the government is very well and properly focused on getting those valuable data sources out. Weather would be another—basic weather data would be another good example that can be built on to add extra value.

Mr. MASSIE. Are the other witnesses aware of any data sets that we need to promote more?

Dr. JAHANIAN. I want to highlight a couple of things. I am sure you are aware of data.gov, which is a Web site that makes a lot of government data sets available, and the goal here is to increase public access to high-value machine readable data sets that are generated by the government. Hopefully it will create new economic values. There are also a number of activities in encouraging the private sector, entrepreneurs to develop applications on top of that data. It is not just making the data available but also making the data valuable so there are a number of essential activities related to that.

There was a recent Wall Street Journal article actually that highlighted at least a dozen different kind of government data sets that have been made available from labor and health violations to flu incidents, energy prize, offshore activities, solar information, and so on and so on that are interesting. From the National Science Foundation's point of view, I should mention that as you may know, we have a number of large facilities—LSST was mentioned, Neon, which is another facility that collects a lot of data, will be collecting a lot of data. The science and engineering community needs that data, and many Federal agencies are working very hard to make that data available. There are a number of issues having to do with open access that go beyond the scope of this question.

Mr. MASSIE. Let me ask a follow-up question to that. So big data like any other data could be misused, altered, hacked, illegally accessed, and sometimes it may just be an honest mistake. We share data that we probably shouldn't have, for instance, where some farm data that got out there and it could really compromise our food safety if people know where all the food sources are. How do we balance the desire for privacy, actually the constitutional right to privacy, with sharing all of this data now that everybody is under a microscope?

Dr. RAPPA. I thank you for your question, and I would like to sort of just turn it a little bit because we do work—each year we work

with about 16, 17 organizations that share data under a confidentiality agreement including three government agencies in which we put teams of students working on very complex analytics projects, and so while I applaud, and I think it is very important and I do think the government is doing a good job at sharing data openly, it is a very important thing to do, I think there is also an opportunity to engage the academic community in other ways to help understand that data, which might mitigate some of these issues around the privacy element.

Mr. MASSIE. Dr. McQueeney?

Dr. MCQUEENEY. Yes, that is an excellent question. Thank you for that. One of the things that we can do is to get data about the data. We call it metadata. So we analyze the data and we don't just look at what information we can get from the data but we describe the data perhaps in terms of its sensitivity—is this more or less sensitive from a point of view of privacy or security or secrecy—and we can then tag those data sets with metadata that describes the implications of using that data and then we can build into the systems that handle the data policies that look not only at the data but the metadata that describes what are the contents and what are the implications of sharing and combining that data and so we can actually build into the foundation of big data systems the ability to interpret policies that we have set in a very conscious and clear-eyed way and as they process the data they can be respectful of that metadata. The medical community has actually done a lot of very good work around patient confidentiality while still getting very good pattern analysis of different kinds of outcomes.

Mr. MASSIE. Thank you very much. My time expired. I appreciate your answer and concern for that question, Mr. Chairman.

Mr. BUCSHON. Thank you, Mr. Massie. I now recognize Dr. Bera for five minutes for his questions.

Mr. BERA. Thank you, Mr. Chairman, and thank you for the series of hearings that we have had on the Subcommittee. It has been great.

You know, big data is incredibly important and how we manage data and with the rapidity of how the world is changing. I mean, when I think back to being a high-school student, and for me it was going and looking at the index cards, walking down and looking in the encyclopedia. Now, when my daughter, you know, she has vast access, or when I do rounds in the hospital, we would have to race down to the library to get information but now before we are even finished presenting, the medical students or the residents can just look at the latest data on, you know, a device like this and get access to the most accurate and timely information. So it is incredibly important that we make these investments to not only manage the data, to sort that data and then to make sure it is accessible. It is a critical priority that we have that workforce both at the professional level but then also at the management level and I think the number that I read was we need about 1.5 million managers. So there is a huge need but also a huge opportunity.

When I think back to the talent that has been impacted in the last four years in the recession, you know, there are a large number of extremely intelligent and talented individuals in their 30s and 40s who have been hit hard. These are folks like myself that

were trained for a 20th-century workforce but now we find ourselves in a 21st-century economy.

Dr. Rappa, are there some best practices—and these aren't individuals that need to get a graduate degree, you know, they are talented individuals—where we could take them and quickly train them for this new economy? Are there examples?

Dr. RAPPA. Right. So we do offer it as a graduate degree but we do this in 10 months, and indeed, a good, fairly substantial, larger portion of our population are people who are returning from—or coming from the workforce to go through this and some of them are in exactly the position that you say. They were transitioning, their companies were faltering. And so the key really with this is short duration. Ten months is actually a very reasonably good time because you could build the skills that you need. If it is too short, you can't accumulate the skills but the key thing is that you have really demonstrated ROI on that education because that person who is coming in to do that has to know that they have a very high probability of getting a job when they leave and at a particular salary rate so that they can justify the investment and time, and that is really what we have done.

Mr. BERA. Dr. McQueeney, are there potentially any examples—you know, again, a lot of these folks are also paying their mortgage, they have to continue to foot their bills—of possibly even doing an advanced work-study type of program where you recruit this talent and they are getting on-the-job training as opposed to a traditional school model?

Dr. MCQUEENEY. Yes. In fact, there is a related topic here that I think is quite interesting, which is the application of big data and analytics back on to the educational process itself. You have seen the great upsurge in videos that attempt to replace traditional brick-and-mortar classroom attendance, coursework. You have seen a number of startup companies formed in this space. If you look at the education process, each of us really learns quite differently. Some of us may learn more from hearing or from seeing or from working problems, and great teachers, great professors are sensitive to how their different students learn and are capable of presenting material in alternate ways to make sure they reach all the students. With electronic delivery of course materials and monitoring of student progress, we generate digital exhaust, if you will, that describes how that student is learning, how that student responds to the instruction, and for the parts of the instruction that are delivered electronically, we actually have the ability to do analytics and to do an optimization process so that each of us on the panel might not get the same length of lecture on five different topics. It might be adjusted to our historical learning patterns.

So we have worked with a number of universities and other, you know, non-traditional educational institutions to apply the big data and analytics techniques to the education and training process itself.

Mr. BERA. Great. In my last 30 seconds, so we have access to data. I think one element that we should also be conscious of is the quality of the data because there certainly is very good-quality data but at the same time there is very poor-quality data that is out

there and, you know, any of you who want to comment on how we monitor quality?

Dr. RAPPA. I think most data starts off as bad data, for the most part, unless it is being collected in a highly careful way. And so it is, you know—I think just as we hear about big data today, we are going to hear about bad data in the future. Most projects start out where you have enormous front end to them to really understanding cleaning and cultivating that data to make it useful, and that is an important part of the educational process.

Dr. JAHANIAN. I would just add that there are a number of techniques that have been developed and are in development dealing with data exploration, data cleaning and so on. Furthermore, when we talk about large-scale data sets, there are statistical techniques that are being applied that really take care of the noise, take care of some of these inconsistencies, and that is one of the attractions of big data.

Mr. BERA. Great. Thank you.

Chairman MASSIE. [Presiding] Thank you, Mr. Bera. I now recognize Mr. Schweikert from Arizona for five minutes.

Mr. SCHWEIKERT. Thank you, Mr. Chairman.

This is one of those types of conversations, you know, we could all sit around and buy you some well-caffeinated coffee and talk for hours and still have no idea if we made any progress.

Doctor, is it McQueeney?

Dr. MCQUEENEY. Yes.

Mr. SCHWEIKERT. First, you are with IBM?

Dr. MCQUEENEY. Yes.

Mr. SCHWEIKERT. In your testimony, help me do a little ferreting out here. Hardware technology or IT talent, what is your biggest bottleneck right now?

Dr. MCQUEENEY. There are bottlenecks in a number of areas. If I looked at the hardware itself, the biggest challenge getting from the petascale to the exascale is actually the power dissipation of the systems. The new technology work that we are doing is to get the computations more efficient in terms of floating point operations per watt so that if you assembled a system thousand times bigger than today's supercomputers you could house it and cool it.

Mr. SCHWEIKERT. You don't want to take down the power grid?

Dr. MCQUEENEY. The power grid may not in fact be able to supply enough power if we didn't make some innovations. That is a good point.

Mr. SCHWEIKERT. But hasn't your company actually been one of the leaders at producing some of those breakthroughs?

Dr. MCQUEENEY. In fact, we have, and in fact, a lot of that history goes back to work that started with the Department of Energy many years ago, and this bears on an interesting historical point. In a time when we are concerned about making investments efficiently, if I go back to the beginning of the ASCII program with the Department of Energy to do the nuclear weapons stockpile stewardship program, the Department of Energy scientists did a very careful analysis of what were the core algorithms, the core analytics, if you will, in today's language, that needed to be done at a certain level to provide the mission that they needed to provide, and they found that the current path at that time of supercom-

puting was going to take five years to produce a machine that they needed in 1 or two years. The analysis they did was thorough enough to reveal that there weren't bottlenecks everywhere but at that time there were bottlenecks mostly in the inner process or communication. So they made a very thoughtful, very surgical investment in accelerating just the piece that was needed to close their mission gap, which was the beginning of a very long run of government-industry collaboration.

Mr. SCHWEIKERT. But you are in some ways heading towards where my question is. So if that bottleneck, in today's world, do I find the technology if I went out to the private sector around the world that is competing and producing high-end supercomputing or is it coming out of a government lab? And I know the pop culture terminology is "public-private partnership" but the reality, they do operate in pretty substantially different silos.

Dr. MCQUEENEY. The real forcing function for a breakthrough is a critical mission need. So in the case of high-performance computing, it has often been a government agency with a critical mission that——

Mr. SCHWEIKERT. But they were doing a specific request for how they wanted to manage their data?

Dr. MCQUEENEY. That is correct, and once that technology is available, it can be consumed very rapidly in lots of other applications that could take great advantage of it but didn't have a compelling enough need to get over that hurdle. That is when the disbursal of technology starts.

Mr. SCHWEIKERT. Just as an aside, only because I had some acquaintances who were—I used to be an old SQL programmer so I am way out of date now. IBM was actually running a fascinating large data project where they were doing sweeping data sets through the world's social media and gathering it and looking for trends. Can you in 30 seconds or so tell me your knowledge on that?

Dr. MCQUEENEY. Yeah, we have analyzed the public social media sources with several of our customers and we can gain a lot of insights. For example, you know, retailers can gain insights about trends and their clients. Transportation agencies can gain insights about likely traffic congestion. There are many sources of public data, both social media and other forms that can be analyzed to reveal patterns about how people conduct their daily activities that are very useful for optimizing the public infrastructure.

Mr. SCHWEIKERT. Forgive me, I am blind as a bat without these. Is it Dr. Rappa?

Dr. RAPPA. Yes.

Mr. SCHWEIKERT. Isn't my single biggest problem in big data right now is noise that when I put data set after data set after data set and build on it, that just small incremental errors actually create really bad decisions on the end?

Dr. RAPPA. Well, I think part of the education around handling big data deals very squarely with the quality of the data and how to clean it and cultivate it to reduce the noise, to——

Mr. SCHWEIKERT. But you and I can go over a long series of public policies, both state, national, you know, military, others, where we built it on really gigantic analyzed data sets and it was wrong.

Dr. RAPPA. Well, I think that, you know, the challenge here is education. So as I alluded to earlier, we have teams of students——

Mr. SCHWEIKERT. Is it education or developing educational skepticism?

Dr. RAPPA. It is developing the education around how to squarely understand the inherent challenges in data. Data is not born clean. It isn't born ready to be analyzed.

Mr. SCHWEIKERT. And when you and I build our model, the way we wait, you know, because we start plugging in human factors that, you know, you and I bring our biases and we——

Dr. RAPPA. And this is why we really need a focused education squarely around how do you draw insights from data because there are these inherent problems in data, especially as you scale them up, as you combine different data sets, as you combine different types of data.

Mr. SCHWEIKERT. Thank you, Doctor, and Mr. Chairman, thank you for tolerating. It is just one of my great fears. And look, I am a data freak. I mean, you have got to see the servers and stuff I have at home. But I have learned when we make big-time public policy on something we all know is right, we keep making huge, very costly mistakes.

Chairman MASSIE. Thank you, Mr. Schweikert. I now recognize Mr. Hultgren from Illinois for five minutes.

Mr. HULTGREN. Thank you, Mr. Chairman. Thank you all for being here. First of all, I just want to thank Dr. McQueeney too. I appreciate your mention and your support for the exascale computing bill I am currently authoring. I am very excited about the potential there and see some huge shift in our national computing capabilities and I am very excited about that, so I appreciate your mention and support of that.

I do have a few questions, and first I guess I would address this one to Dr. McQueeney and also Dr. Jahanian. Is that right? I am sorry. I wonder if you could comment briefly on where the United States stands in your opinion in worldwide computing leadership? I know the metric of the fastest supercomputer is one metric but what do you use as a metric for big data to determine which countries are using it most effectively?

Dr. MCQUEENEY. The common thing that is cited in these discussions is the top 500 supercomputers list. That is something that is compiled twice a year, as you well know, and we have usually been at the top of that list. We have continued to be the majority of the systems on that list but other countries have noticed the success that we had in, you know, government leading the way on high-performance computing breakthroughs. Once those systems are built, they find hundreds and thousands of other applications, each with a client that might not have been able to fund that breakthrough themselves but can certainly utilize it. Other countries have popped up on the top of that list because they are interested in emulating the success we have had in leading the way with innovation and then seeing that innovation used broadly across the commercial sector. So the top 500 list is a very technical, perhaps very geeky measure of who is on top, and I would say that we are still in a leadership position there but it has been stronger in the past than it is today.

If you turn to more of a business view, you would want to look at the companies that were taking the best advantage of data sources, either to drive value in their companies or to provide benefits such as public safety or health benefits, and there again I think we are in a good position but it is a very different kind of skill, a conversation we didn't quite finish before about the skill to build these large systems is a very focused, very large-scale, very capital-intensive activity but the skills to use these systems are more focused on creativity and are actually better done by large groups of small teams. In fact, you know, the NSF has been a leader in fostering that kind of innovation where thousands and thousands of groups can build innovative applications and take advantage of these systems.

Mr. HULTGREN. Thanks. Dr. Jahanian?

Dr. JAHANIAN. Yes, just a couple of quick comments. There is no question that we continue to maintain our leadership worldwide in this area, and there is no doubt that continued investment in this area is extremely important to the future of the country. As I mentioned just a few minutes ago, NSF's investment in Blue Waters, Stampede, as well as the Yellowstone supercomputing centers represent a range of investments that we make in high-performance computing, addressing the needs of not only the top five percent of application that have exceptionally high computational needs but also a broad spectrum of researchers across the country in science and engineering who would need computational resources.

A couple of comments. Just look at Blue Waters, for example, which is at University of Illinois. A couple of data points about it. It has—if you could—just the computing power of it, if you could multiply two numbers together every second, it would take 32 million years to do what Blue Waters does in one second. That is astonishing power, for example, of Blue Waters. In terms of storage capacity, memory capacity and so on, there is a similar kind of scale.

The second point that I want to make is, we view computation and data to be two sides of the same coin. You really need to address both. So when we talk about computational capabilities, we also have to worry about cyber infrastructure to manage, to curate, to serve data to science and engineering community, and the investment in cyber infrastructure has to be balanced between the computation side of it as well as management and curation of data.

Mr. HULTGREN. Let me have—my time is running out but I have a follow-up question to the two of you as well if you could both comment in the time I have. It seems to me that exascale computing is focused on solving discrete problems that necessitate massive computing power and speed. Are these different problems than those we are addressing through big data analytical tools and how do these two terms, how are they different, how are they similar?

Dr. MCQUEENEY. Historically, we have tended to talk about them differently, but as we project how the exascale systems will be built and how they will be used and we look at the growing importance of big data analytic systems, we see that the platforms on which these systems will both depend will be much more common than separate, and in fact, we see that there is no conflict between investments in classically what we have called HPC and what we are

now calling big data analytics, and both are changing actually. The way we use an exascale system will not be the same way that we use a petascale system. There isn't time here to go into it, but it actually morphs into a direction that is much more common with what we will do in big data and analytics.

Dr. JAHANIAN. I would just add that many of the problems that the business community needs, the science and engineering community needs are being addressed today through different kind of computational architectures that doesn't necessarily require today to have exascale computing including weather modeling, a number of other applications that have been mentioned. So it is really important to consider the investment in exascale computing in the spectrum of investment that we make to support computational and data needs of the entire science and engineering community and of course the private sector.

Mr. HULTGREN. Thank you so much. Chairman, thank you. I yield back.

Chairman MASSIE. I now recognize Mr. Lipinski from Illinois for five minutes.

Mr. LIPINSKI. Thank you, Mr. Chairman. I am glad that Dr. Jahanian mentioned Blue Waters there. We were just there not that long ago, but since you covered that, I can move on to a different area.

Dr. McQueeney, in your testimony you talk about how the Federal Government needs to invest in big data if the U.S. is going to maintain its leadership and competitive edge in this area. The needs and potential benefits of big data for the Federal Government align closely with those of private industry in a number of areas. If that is the case, how can the Federal Government more effectively partner with industry to achieve common goals and do you believe that industry has sufficient input in the Federal Government's research agenda as it relates to big data?

Dr. MCQUEENEY. I do think we have sufficient input. I think we have excellent dialogs with the relevant agencies and national laboratories, and I think the roles are complementary. I go back to the story about the early days of the ASCII program where through a collaboration we realized that the key piece of a supercomputing system that needed to be accelerated was not the entire investment. We could ride on the commercial investments for most of the components of the supercomputing systems at that time except for one, which was the high-bandwidth switching between processors. And so that kind of thoughtful connection between the leaders in commercial computing and the leaders on the government side has been able historically to identify which areas are critical to attain government mission imperatives and where we can leverage commercial technology and where we need to accelerate that in a surgical fashion. So it has, in our view, been a very good partnership based on very high-bandwidth technical communications, understanding of applications and knowing when the government should be leveraging commercial investments and when they need to accelerate parts of that investment to attain unique mission goals, and again, as I have said before, once those barriers are crossed in terms of either the scalability of the system or the internal bandwidth of the system, it opens up thousands of new applications

where there were ready problems to be analyzed but those applications weren't large enough to drive that breakthrough. So that is how the effect works of the leadership coming from some of the government agencies and then being realized broadly across industry. That is the essence of where this leadership has come from so successfully over the years.

Mr. LIPINSKI. I want to follow up with Dr. Rappa on that. Dr. Rappa, you discussed the importance of public-private partnerships to realizing the benefits of big data and stated specifically that we must intensify and accelerate the national investment in proven models. What characteristics make a public-private partnership successful and what models should we be investing in? What were you referring to there?

Dr. RAPPA. Well, I think first of all, we have been doing this now for six years and so I think we do have a fairly interesting, novel model for producing talent in this field with a kind of proven track record based on data, based on market value of the graduates, but I think it comes really, you know, partly from the university community, partly from the academic community. Obviously we have a set of missions to educate students but we need to also, I think, do that by trying to really understand the employer, what are they looking for when they hire talent, what are the kinds of skills that they need in order to be effective on the job, and I think employers need to sort of be open to working with the academic community. You know, there is a certain amount of dissidence that naturally occurs because there are two different worlds with different missions but I think it is really—I think we have shown that it is possible with organizational innovation, with a focused effort, with a sense of openness to engage the private sector in a very positive way, not just at NC State but at other universities. There are many, many examples now that I hope we are providing some leadership on but that other universities are working with our model but also pursuing other creative models to do this. There are probably about two dozen around the country already.

Mr. LIPINSKI. Thank you. Dr. Jahanian, anything you want to add about public-private partnerships?

Dr. JAHANIAN. Yes, indeed. There is no question that when we think about the innovation ecosystem in this country, it includes academia, it includes the private sector, it includes government investment and a talent-rich workforce. The private sector is investing heavily in cloud computing, as you know. It is investing heavily in making computational resources also available. I think there are opportunities for the Federal investment to leverage that and make some of that available. Of course that is commercially available today to our researchers, to our scientists and engineers who could rely on those systems. We have announced a number of partnerships, one with IBM and Google, another one with Microsoft that make some of those resources available to the research community.

Dr. McQueeney already mentioned this, that there is high-bandwidth communication between the private sector and various Federal agencies. I can tell you from NSF's perspective, it is a very, very rich collaboration. On my advisory committee, I have a number of the senior leader from the private sector who serve on my advisory committee advising us on our portfolio, on our invest-

ments in addition to academics who serve on my advisory committee.

The final comment that I want to make is, there are a number of programs at NSF, and I know you are familiar with all of them, including SBIR, including I–Corps and so on that focus on transfer of knowledge from lab to practice. Federal Government invests heavily in advancing frontiers of knowledge. For us to accelerate those programs such as I–Corps, SBIR and so on serves a tremendous purpose, and here again, there are opportunities to engage the private sector and accelerate the transfer of knowledge to practice to benefit the Nation. Thank you.

Mr. LIPINSKI. Thank you.

Chairman MASSIE. Thank you, Mr. Lipinski. I now recognize Mr. Bridenstine from Oklahoma for five minutes.

Mr. BRIDENSTINE. Thank you, Mr. Chairman.

I also serve on the House Armed Services Committee, and I am aware that the Department of Defense is moving towards cloud-based computing solutions, and this of course creates some consternation about security issues, cyber hacking, other cyber crimes, and I was wondering if any of your organizations are involved in helping the Department of Defense work through these issues and what those solutions might be, if you could share with us on that?

Dr. MCQUEENEY. Sure, if I could start? You are quite right to raise the concern about security for any systems used by the Defense Department especially, although it would be true for all Federal agencies. And when you move to a cloud computing model, there is an extra imperative to be concerned about security, and if you think of it in terms of the DOD might think of it, if that environment should be compromised by an enemy, it is a bigger piece of resource than an individual machine so it requires special vigilance. Now, the good news technically is, the way we handle virtualization, which is the foundation of how cloud computing is delivered from a compute virtualization point of view, there are actually sophisticated techniques that can provide additional security in a virtualized environment that we can provide even when using things running on bare metal. We have additional abilities to instrument the operation of that cloud and to very rapidly detect any kind of pattern or behavior that is indicative of a threat.

We did a project a number of years ago with the U.S. Air Force and they graciously let us write a short press release on it where we built a cloud computing environment that was at the cutting edge a few years ago. We instrumented it very thoroughly with watching the package flowing on the interconnected network that built the cloud in question and we very carefully isolated it from the rest of the world, introduced known cyber attacks into it and were able to show that if we knew the patterns of command and control, as the defense folks might say, of these cyber attacks, we could actually spot them assembling themselves and interrupt them before they had a chance to launch. So having tremendous control over the environment out of which we were getting compute resources gave us abilities to do additional security and additional monitoring, even if we assumed the security was not perfect and could be breached, could we essentially in real time detect that breach and interrupt it before it stopped. I thought that was a very

forward-looking piece of work that was driven by the Air Force CIO's office.

Mr. BRIDENSTINE. Excellent. Go ahead.

Dr. JAHANIAN. As you alluded to, these new environments, whether it is mobile platforms or cloud computing, are introducing new challenges, and we recognize that attackers and defenders are coevolving and there are enormous challenges to protecting our critical infrastructure and our cyber infrastructure.

I wanted to mention NSF's Secure and Trustworthy Cyberspace program, which is a research program addressing many of the challenges that we alluded to, and this is a research program that addresses not only the technology issues but also transition to practice. Furthermore, the NITRD research and development subcommittee has a working group that focuses on coordination of activity across various agencies on cybersecurity and there is rich dialog involving various agencies on that issue.

Mr. BRIDENSTINE. Excellent. Are there any other things that the Department of Defense could do to help you guys with the objective of securing cloud computing for the Department of Defense?

Dr. RAPPA. So I am currently co-directing a project with a colleague at NC State, which is the science of security project that is done in collaboration with Carnegie-Mellon University and University of Illinois, and we are trying to bring together large groups, multidisciplinary groups of faculty to really try to understand the underpinning of the security problem and how to produce science around it. It is a very long-term challenge but it is one which I think has to start with getting the faculty across different disciplines focused on it and certainly I think it has been a tremendous opportunity and I look forward to moving into the future.

Dr. MCQUEENEY. Yeah, Dr. Rappa makes a very interesting point, to close the loop here. The cybersecurity problem is itself a big data and fast-data problem, and in fact, with some of the advanced persistent threats that we see today, which depend on breaching an infrastructure and then laying dormant for several months, what the attacker is trying to do is to wait out how long you keep your log file data so that when they launch themselves, it is difficult to do forensics, and so what we have learned is that these log files are actually the essence of the big data you need to do pattern analysis, pattern discovery on forensics, you know, should any attack occur. So in fact, most of the science behind big data including data at rest and large-scale computation and fast-data that are eating very high-speed streams is directly relevant to the subject of cyber defense.

Mr. BRIDENSTINE. Thank you.

Chairman MASSIE. Thank you, Mr. Bridenstine. If the Ranking Member is amenable to this, I think we will do another round of questions?

Ms. WILSON. Yes.

Chairman MASSIE. Did you have something to introduce into the record?

Ms. WILSON. I do. Thank you, Mr. Chair. Mr. Kilmer has lots of conflicts. As we saw him come to the meeting, he had to leave, and I want to ask unanimous consent on behalf of Mr. Kilmer to intro-

duce a report on big data from IDC into the record, and then I have a question.

Chairman MASSIE. Without objection, so ordered. It will be set into the record.

[The information appears in Appendix II]

Ms. WILSON. Thank you. This question is for everyone.

We have all had several discussions lately about the value of NSF-funded research to society and how we might certify that value based on the grant proposal. I think we might use big data instructively here. It is an incredibly interdisciplinary field where tools are developed in the pursuit of one narrow research question, let us say in the social sciences might have profound applications across many fields of science and even in many sectors of the economy that can't possibly be anticipated at the time of the proposal. What is the potential for data analytics being developed in one little seemingly irrelevant corner having unintended benefits to other fields and societal applications? And if you have concrete examples, that would be even better for us to understand. Thank you.

Dr. JAHANIAN. Okay. I guess I will start. There is no question there are all sorts of explorations that we are doing in the area of big data that we can't even begin to see the potential impact of it. I will give you an example. NSF has been investing and other agencies with the private sector in what is known as the area of machine learning. These investments have taken place for at least 20 or 30 years. In fact, IBM has also led efforts in this area. I can tell you that it is investments of the last 20 or 30 years that have come to fruition such that these machine learning algorithms essentially allow us to look at these large data sets and identify trends and be able to adapt. Essentially, they have a broad range of applications from weather forecasting to financial modeling to biomedical research and so on that have had tremendous, tremendous impact and now we use these techniques as if they are off-the-shelf solutions available that you can buy. These are through years of investment that we have made that have come to fruition, so that is an example of that.

We are investing in all sorts of areas in natural language understanding, in information retrieval, in various algorithms and approaches to automated scalable approaches to reasoning that could be applied to understanding relationship between gene sequence structure and biological functions. These are all essentially the kinds of investments that we are making that some of us we could see how it comes to fruition. Some of it relies on decades of investment that we have already made in computational techniques and data-intensive techniques.

Dr. MCQUEENEY. If I could offer you an example from the medical world, one of the critical problems in medicine is the loss of premature infants due to infections, and physicians have struggled for a long time with identifying the onset of an infection at a very early point because as these infections can grow exponentially, the earlier you can intercept them, the more likely you are to have a lifesaving benefit for someone who is very vulnerable such as a premature infant. We have done work with the Toronto Hospital for Sick Kids where a physician up there had an idea that all the instrumentation in the NICU that is—you know, you have probably

been in a hospital room or intensive-care room, all the instruments around the bed, someone comes in every half an hour and writes down those numbers but the instruments are producing readings continuously, and this physician had the idea that if we kept all that data and we stored all that data as it came out of the machines in real time, which was a tremendous aggregation from a velocity of data point of view and correlated with the eventual issues that these premature infants had, we might be able to detect patterns using techniques such as machine learning that we were just hearing about that would give us an early identification of an upcoming infection, the ability to treat it before it got out of control, and her theories were absolutely correct. There were signatures in the data that gave up to 24 hours advance notice of an onset of an infection that was time for the doctors to in many cases provide some kind of lifesaving therapy. So there is an example of very, very deep mathematics, computer science being applied to a problem where the data was being produced every day by these instruments and it wasn't being captured and it wasn't being looked at and it wasn't being correlated with results to produce a fantastic outcome.

Dr. RAPPA. I would just sum up by saying that really big data is part of a decades-long process that really started with computerization in the 1940s and 1950s and eventually got interconnected through the Internet in the 1970s, 1980s and 1990s that the world that we are turning into, data is going to be everywhere. It is going to affect exactly what happens here. It is going to affect hospitals, universities, every corner of the economy literally, and so we need to take approaches to that to try to develop understanding around big data, how it is applied, how the tools of analytics are applied across, you know, virtually every sector of the economy, and so I would take a very broad view, not looking at it as specifically, you know, a realm of computer technology or some other sort of isolated realm but looking at it as, you know, unfortunately as the big thing it is.

Dr. JAHANIAN. May I offer another example as I was thinking about it? I am reminded of the work by Daphne Koller and her collaborators at Stanford on classifying breast cancer via image analysis. As you know, 40,000 women die from this disease each year. By extending essentially image analysis techniques to hundreds of, I should say thousands and thousands of biopsy images, they were able to identify a subset of cellular features. Out of 6,000 possible features, they were able to essentially identify a few of them that were predictive of survival time among breast cancer patients. What is really surprising is that the feature that they identified, it wasn't just from—the best feature, I should say, that is a predictor of survival, was not from the cancerous tissue itself but it was from the surrounding tissue, and that has led to new kinds of treatments. It has led to new kinds of diagnosis techniques and also a very personalized treatment that could aim to improve survival times in patients. That is a very, very concrete example.

Another example is the work that Google had done during H1N1 virus. I will be very brief about this. Before they actually discovered a vaccine, we wanted to track the spread of disease. Google engineers used data that had nothing to do with the virus directly

from billions of essentially web searches from around the world together from publicly available, essentially historic data on flu trends, to predict the spread of flu virus down to small regions in the country—or across the world, rather. This is a remarkable essentially application of data that one would have never thought could be applicable to something like H1N1 virus.

Ms. WILSON. Thank you very much.

Chairman MASSIE. Thank you, Ms. Wilson. Thank you for that very excellent example of how we can use—a private company can find information in the data.

We got a little bit out of order so the last question is going to be mine. I reserve five minutes for myself. And the question I want to ask is, we have heard about banks that are too big to fail, and we also know that the Internet is now too big to fail. We recently in the House passed a CISPA bill which is somewhat controversial but some people felt it was necessary to do because the Internet was so big and pervasive in our lives. So my question to you is, are there any big data sets that are too big to fail? In other words, are there ones that are pervasive that we have let through osmosis become—we have become too dependent upon or maybe not too dependent but we are dependent upon these data sets, for instance, weather, you know, and early warning systems? Not all of those, I imagine, are government systems. Some of them are private but possibly the government is relying on these systems and so I would be remiss if I didn't ask this question now before something fails, but tell us what is too big to fail right now? What would we bail out, and is there sufficient redundancy in the collection, storage and access of these data sets? Thank you.

Dr. MCQUEENEY. Well, first, I would just like to say that we were delighted to support that cyber bill, and I congratulate you on such broad bipartisan support in the House for getting that acted upon.

Data sets have the property that they can often be subdivided and often be replicated, and so we have a lot of techniques by which we can assure the continuity of data if we take the time to do it, and if there were very valuable historical records on things like long-term weather trends that were only stored in one place, that actually could be a concern because that is literally irreplaceable data. But essentially all of the IT techniques needed to take those large data sets and segment them and replicate them in different secure places so they could be re-created do exist but I think you raise an interesting point, that it is worthwhile to periodically check that we are being appropriately vigilant with the digital archives that are so valuable.

Chairman MASSIE. Dr. Jahanian?

Dr. JAHANIAN. I don't have a specific example. What I can tell you is that similar to the issue of cybersecurity, as Nation's critical infrastructure and more generally the Internet is playing a vital role in integrating the economic, you know, political, societal fabric of our society, we are going to become more and more dependent on data, and data is going to play an increasingly significant role in our day-to-day lives, and for that reason, I think the same sort of issues that apply to all sorts of IT solutions that we take for granted will increasingly be applied to data.

From a research and engineering community's point of view, it is not just failure of the data but making that data accessible and also making the data accessible to broad community of scientists and engineers is an issue that we are quite concerned about.

Chairman. MASSIE. Thank you very much. I was part of the bipartisan on CISPA, opposing CISPA actually, but that is okay.

I want to thank the witnesses for their valuable testimony and the Members for their questions today. The Members in the Committee may have additional questions for you, and we will ask that you respond to those in writing. The record will remain open for two weeks for additional comments and written questions from the Members.

The witnesses are excused and this hearing is adjourned.

[Whereupon, at 11:35 a.m., the Subcommittees were adjourned.]

Appendix I

Answers to Post-Hearing Questions

ANSWERS TO POST-HEARING QUESTIONS

Responses by Dr. Michael Rappa

QUESTIONS FOR THE RECORD
THE HONORABLE DEREK KILMER (D-WA)
U.S. House Committee on Science, Space, and Technology

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013

There have been a number of big data reports generated recently by a number of industry leaders. I'm proud to say that companies, EMC and Isilon, which is headquartered in Washington State, have done a lot of great work on big data. EMC recently released their latest "Digital Universe" study, conducted by IDC. Amazingly, this study projects that the digital universe will reach 40 Zettabytes by 2020.

One of the issues I have been passionate about, both in the state legislature and in my first few months in Congress, is STEM education. It seems to be that many of these reports make a compelling case that there is a dire need for more data scientists.

I have two questions:

1. *How are your organizations specifically addressing the need for more data scientists and employees with STEM backgrounds?*

The Institute for Advanced Analytics at North Carolina State University has been at the forefront of educating a new generation of data savvy professionals to address precisely this need for data scientists. Launched in 2007, our Master of Science in Analytics (MSA) degree program is the first of its kind in the nation and has become a template for programs at other universities. The MSA blends together statistics, mathematics, and computer science and business topics into an interdisciplinary curriculum and is classified as a STEM graduate degree. The Institute works closely with employers, professional societies, government agencies, and academic institutions to enhance the pipeline of data science professionals.

As I mentioned in my testimony, the Institute has a proven six-year track record that shows we can succeed in producing the kind of talent employers need. We can do it quickly with an intensive and highly targeted educational format that yields consistent student outcomes. The program attracts a diverse, high quality, domestic student population and yet runs with a sustainable, cost-effective, and self-financed tuition business model. We have demonstrated the ability to produce 80 graduates annually, and I am confident we could produce tenfold with the necessary upfront investment in facilities and personnel.

2. *In your testimony, you both discuss how our nation is facing a data scientist shortage. What policies would you recommend Congress consider to address that shortage?*

The Institute has shown this is a problem we can solve through close collaboration with the private sector. The task now is to scale quickly to meet demand. I believe markets will eventually adjust, but we could accelerate the process with a better flow of data.

Recommendation 1: Provide guidance and incentives to degree granting institutions to make public standardized reports of student employment outcomes, such as job placement and salary data, by degree majors.

Accurate and up-to-date information is key to achieving market equilibrium. Prospective students who are contemplating the time-consuming and financially burdensome decision to invest in their education must have the data needed to calculate the expected return on investment.

There is nothing particularly novel about this recommendation. Already many of the nation's top business schools make public each year comprehensive employment reports for their MBA programs (as does UW's Foster School of Business). This data is absolutely essential to prospective students in making their decision to pursue the MBA degree (typically two years of study, and an average tuition cost of $80,000 at the best schools). Before making a decision about whether or where to attend, individuals can calculate an expected ROI based on recent placement rates and salaries.

Similar employment reports should be produced for STEM graduate programs. Once we have timely and accurate employment data, the market place will adjust quickly to the opportunity in data science. We have seen exactly this kind of reaction in the marketplace for our Master of Science in Analytics degree. The Institute has published detailed employment reports since the inception of the program. Our applicants clearly make decisions based on data driven ROI calculations of expected employment outcomes.

We know from employers the demand for data science talent is there and growing. We also know there are large numbers of students coming out of school or already in the workforce who are underemployed, if not unemployed. There is simply poor alignment between skill training and skill needs. The key to achieving a closer alignment is up-to-date and accurate degree outcomes data that would allow students to make sound educational investments. STEM employment reports need to become as commonplace in universities as they are today with MBA programs.

Recommendation 2: Provide incentives for college and university students to pursue fields of study in data science contingent on institutional performance metrics for student outcomes that meet threshold levels.

Congress may wish to create or target existing incentives for students to pursue STEM degrees. I recommend tying incentives to institutional performance metrics such as graduation rates and employment outcomes. This will ensure students are making sound investment decisions in furthering their education based on expected outcomes.

The current system of subsidized student loans would be on a more solid footing financially if students taking loans knew in advance the likelihood of their ability to repay loans. Perhaps such data could be included in the loan application process. Again, it comes down to more timely and accurate public data about educational outcomes so the marketplace can make better decisions.

Recommendation 3: Provide guidance to Federal agencies to seek-out and take advantage of educational opportunities for existing employees to pursue fields of study in data science. Furthermore, encourage the creation of job categories within the Federal workforce to employ and promote data scientists.

Existing employee training programs within government agencies should be encouraged to take advantage of data science related educational programs and to use institutional performance metrics to determine what programs qualify. Sometimes existing program classifications within agencies can be slow to change and fail to include in a timely manner new areas of studies like Analytics.

I have worked with one Federal agency to help define a new job category of "data scientist." Congress may wish to encourage other agencies to move in a similar direction. By providing leadership, the Federal government will also help the private sector as it, too, seeks to define job categories for data science.

Clearly, the educational community needs to accelerate its efforts to increase the supply of talent if there is any hope of meeting the demand over the next few years. Working with employers and government, I am confident that universities can and will step-up to the challenge.

Once again, thank you for this opportunity. Please do not hesitate to contact me as needed. If I can be of any further assistance to the Committee or to the citizens and employers of Washington State, I will be glad to help.

Sincerely,

Michael Rappa, Ph.D.
Executive Director of the Institute for Advanced Analytics
and Distinguished University Professor
North Carolina State University
920 Main Campus Drive, Suite 530
Raleigh, NC 27606

Responses by Dr. Farnam Jahanian

QUESTIONS FOR THE RECORD
THE HONORABLE CYNTHIA LUMMIS (R-WY)
U.S. House Committee on Science, Space, and Technology

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013

1. The massive volumes of data generated daily across a range of industries and public sector organizations necessitate new methods to store and manage the data. The National Science Foundation (NSF) Computer and Information Sciences and Engineering Directorate (CISE) helps develop and maintain cutting-edge national computing and information infrastructure for research and education. This data must be analyzed to extract knowledge and promote discovery. Often this data resides in scattered locations.

For the nation to take advantage of the discovery that can be derived from big data, please explain how an effective infrastructure can be constructed to connect the entities developing and using Big Data to drive discovery. Additionally, please describe how the infrastructure, connections, and broadband would be developed to enable the entire community of research universities, in particular those like the University of Wyoming from EPSCoR states.

The Division of Advanced Cyberinfrstructure (ACI) in NSF/CISE supports three major programs that emphasize the development of computational infrastructure and participation in Big Data activities: The first program is Data Infrastructure Building Blocks (DIBBS); the second is Campus Cyberinfrastructure - Network Infrastructure and Engineering (CC-NIE); and the third is Cyberinfrastructure Framework for 21st Century Science and Engineering (CIF21). All three programs support research and discovery efforts in data as well as helping campuses to obtain the infrastructure connections and facilities required to participate in Big Data. They are discussed below.

The DIBBS Program focuses on how to develop, implement, and support the new methods, management structures and technologies to store and manage the diversity, size, and complexity of current and future data sets and data streams. DIBBS has three types of awards:

- Conceptualization awards support design specifications for creating a sustainable data infrastructure that will be discoverable, searchable, accessible, and usable to the entire research and education community;
- Implementation awards support development and implementation of technologies and infrastructure that addresses elements of the data preservation and access; and
- Interoperability awards develop frameworks that provide consistency or commonality of design across communities and implementation for data acquisition, management, preservation, sharing, and dissemination.

The CC-NIE Program invests in improving and re-engineering networks at the campus level to support a range of data transfers supporting computational science and computer networks and systems research. CC-NIE has two major types of awards:

- Data Driven Networking and Infrastructure for the Campus and Researcher; and
- Network Integration and Applied Innovation awards.

The CIF21 effort has participation from every NSF Directorate. CIF21 focuses on foundational research, infrastructure support and deployment, and community building. Since CIF21 supports the entire cyberinfrastructure eco-system, it also supports projects involving data, computational science and building research communities.

NSF EPSCoR supports programs that focus on connectivity and cyberinfrastructure for Big Data. These are discussed below, specifically focusing on EPSCoR activities in Wyoming.

Connectivity: Wyoming is a founding member of the Front Range Gigapop (FRGP) in Denver, which provides 10Gbit/sec connectivity between the University of Wyoming and institutions in Colorado, including NCAR, as well as connectivity to the Abilene Network and National LambdaRail. A significant amount of the nation's long-haul telecommunications fiber transits through Wyoming's southern quarter along the mainline of the Union Pacific railroad and Interstate 80. Major telecommunications centers as well as the National Center for Atmospheric Research (NCAR) – Wyoming Supercomputing Center are located in Cheyenne. Fiber connectivity along with the availability of electrical power and favorable climate for data center operation is making southeastern Wyoming an important IT hub.

Managing Big Data: Wyoming has an NSF EPSCoR award that pilots an effective cyberinfrastructure that connects EPSCoR entities developing and using Big Data to drive discovery. The RII Track-2, CI-Water, allows a consortium of Utah and Wyoming researchers to acquire and develop hardware and software cyberinfrastructure to support the development and use of large-scale, high-resolution computational water resources models to enable comprehensive examination of integrated system behavior through physically-based, data-driven simulation. Successful integration requires data, software, hardware, simulation models, tools to visualize and disseminate results, and outreach to engage stakeholders and impart science into policy, management, and decisions. The computational requirements of stochastic methods to consider uncertainties, fine spatial and temporal resolutions to improve accuracy, and representation of dynamic processes that include feedbacks among system components demand use of state-of-the-art high-performance computing (HPC). CI-WATER is working to develop a robust and distributed CI consisting of integrated data services, modeling and visualization tools, and a comprehensive education and outreach program that will revolutionize how computer models are used to support water resources research in the Intermountain West and beyond.

2. Within NSF, the Computer and Information Sciences and Engineering Directorate (CISE) helps develop and maintain cutting-edge national computing and information infrastructure for research and education. NSF has significant investment in computing infrastructure, including the NCAR-Wyoming Supercomputing Center, among others. These high performance computers are capable of processing complex data sets at a greater rate, enabling scientific research and discoveries.

The ability to analyze and utilize information from increasing quantities of data sets is crucial to advancing knowledge. Please describe the contributions these facilities are expected to make to the development and use of Big Data over the next three to five years.

ACI supports national efforts in advanced and cutting edge computational facilities including the recently announced facilities in Texas (Stampede) and Illinois (Blue Waters). While both of these facilities support very high performance and complex data problems, the Blue Waters facility has the largest data storage and management system in the world. When these facilities are in full production, they will permit investigators across the country to engage in innovative research demanding petascale capabilities.

ACI also supports the XSEDE project, which manages and operates 17 different high performance systems across the nation with a common interface to ensure that researchers can get what they need without having to contact each site. XSEDE also manages the allocation process that provides researchers with the resources they need. Usage of these facilities is done via peer review so that the best research is supported.

The NCAR Wyoming Supercomputing Center (NWSC) provides high-performance CI that will enable researchers to perform high-resolution simulations of weather phenomena, global and regional climate, coastal oceans, sunspots, subsurface flow, and more. Earth System research and education will be transformed by the NWSC, as the next generation of Earth science researchers and computational scientists will be attracted by the importance of the problem and the scale of the facilities available to them. Current and planned education, outreach, and training programs built around the facility will help to broaden the impact of the NWSC project on both regional and national scales. Integration of the NWSC with other NSF high-performance CI will provide important linkages with other resource providers and will directly support NSF's vision of a transformative national petascale cyberinfrastructure for science and engineering. Finally, the NWSC has the potential to contribute to economic development in the State of Wyoming in the form of well-paying jobs, workforce training opportunities, and in the transformation of the state into a destination of choice for other high-technology enterprises. Through the facility partnership with Wyoming, these benefits can be extended to other EPSCoR states as well.

NCAR aims to improve researchers' abilities to analyze and utilize information via various efforts focused on data manipulation and visualization (e.g., Globally Accessible Data Environment, GLADE, http://www2.cisl.ucar.edu/resources/glade; data analysis and visualization, http://www2.cisl.ucar.edu/resources/software/dav).

QUESTIONS FOR THE RECORD
THE HONORABLE DEREK KILMER (D-WA)
U.S. House Committee on Science, Space, and Technology

Next Generation Computing and Big Data Analytics

Wednesday, April 24, 2013

There have been a number of big data reports generated recently by a number of industry leaders. I'm proud to say that companies, EMC and Isilon, which is headquartered in Washington State, have done a lot of great work on big data. EMC recently released their latest "Digital Universe" study, conducted by IDC. Amazingly, this study projects that the digital universe will reach 40 Zettabytes by 2020.

One of the issues I have been passionate about, both in the state legislature and in my first few months in Congress, is STEM education. It seems to be that many of these reports make a compelling case that there is a dire need for more data scientists.

I have two questions:

1. *How are your organizations specifically addressing the need for more data scientists and employees with STEM backgrounds?*

NSF has focused for many years on developing the STEM workforce through investments in its research and education programs and projects. Increasingly, the development of skills in the use of large data sets is a critical part of the training needed for the STEM workforce. Collectively, STEM programs support, for example, curriculum development, strategies to increase student retention and success in STEM, and student support through scholarships and fellowships. As part of the merit review of these projects, they have to show evidence that the measures taken will ensure effective learning.

Many of these programs focus on undergraduate and graduate students in formal and informal education settings. In addition, across NSF – in all the science directorates – research projects support graduate students as research assistants. Increasingly, these assistantships require data-intensive research, often involving large-scale data sets. These hands-on learning opportunities are critical in helping to develop a workforce with sophisticated and real experience in deploying these skills.

2. In the FY14 Budget Request, NSF proposes STEM-C Partnerships (i.e., STEM with an emphasis on computing) as one of its primary approaches to advance K-12 teacher and student development of computational skills. NSF also supports research that develops and evolves the knowledge base that informs improvements in the preparation of the workforce. (See http://www.nsf.gov/about/budget/fy2014/pdf/25_fy2014.pdf.)

3. *In your testimony, you both discuss how our nation is facing a data scientist shortage. What policies would you recommend Congress consider to address that shortage?*

Congress should continue to support STEM education at all levels – from kindergarten through lifelong learning. In particular, NSF is looking to invest in evidence-based and evidence-generating approaches to achieve specific educational outcomes. While anecdotal evidence may point to a variety of policy options, NSF, working in partnership with private and public sector stakeholders, is laying the foundation for policies and programs that are rooted in empirical evidence. In particular, retraining efforts, and initiatives that are aligned with the changing needs of business and industry, may be promising areas for strategic investment.

Appendix II

Additional Material for the Record

IDC IVIEW, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, submitted by Representative Derek Kilmer

IDC
Analyze the Future

I D C I V I E W

THE DIGITAL UNIVERSE IN 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East

December 2012

By John Gantz and David Reinsel

Sponsored by EMC Corporation

Content for this paper is excerpted directly from the IDC iView "Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East," December 2012, sponsored by EMC. The multimedia content can be viewed at ***www.emc.com/leadership/digital-universe/index.htm***.

Executive Summary: A Universe of Opportunities and Challenges

Welcome to the "digital universe" — a measure of all the digital data created, replicated, and consumed in a single year. It's also a projection of the size of that universe to the end of the decade. The digital universe is made up of images and videos on mobile phones uploaded to YouTube, digital movies populating the pixels of our high-definition TVs, banking data swiped in an ATM, security footage at airports and major events such as the Olympic Games, subatomic collisions recorded by the Large Hadron Collider at CERN, transponders recording highway tolls, voice calls zipping through digital phone lines, and texting as a widespread means of communications.

With the rise of Big Data awareness and analytics technology, the digital universe in 2012 has taken on the feel of a tangible geography — a vast, barely charted place full of promise and danger. The digital universe lives increasingly in a computing cloud, above terra firma of vast hardware datacenters linked to billions of distributed devices, all governed and defined by increasingly intelligent software.

In this context, at the midpoint of a longitudinal study starting with data collected in 2005[1] and extending to 2020, our analysis shows a continuously expanding, increasingly complex, and ever more interesting digital universe. This is IDC's sixth annual study of the digital universe, and it's chock-full of new findings:

- From 2005 to 2020, the digital universe will grow by a factor of 300, from 130 exabytes to 40,000 exabytes, or 40 trillion gigabytes (more than 5,200 gigabytes for every man, woman, and child in 2020). From now until 2020, the digital universe will about double every two years.
- The investment in spending on IT hardware, software, services, telecommunications and staff that could be considered the "infrastructure" of the digital universe and telecommunications will grow by 40% between 2012 and 2020. As a result, the investment per gigabyte (GB) during that same period will drop from $2.00 to $0.20. Of course, investment in targeted areas like storage management, security, big data, and cloud computing will grow considerably faster.

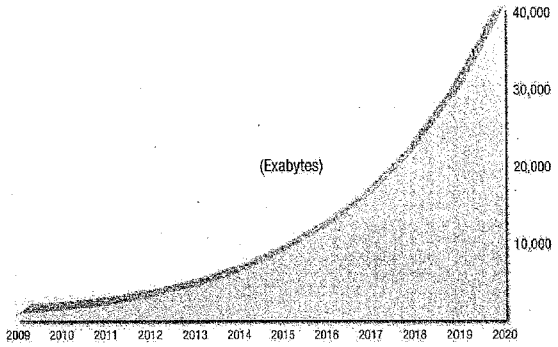
[1] The first *Digital Universe Study* was published in 2007 (see http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf). At that time, IDC's forecast for the digital universe in 2010 was 988 exabytes. Based on actuals, it was later revised to 1,227 exabytes.

- Between 2012 and 2020, emerging markets' share of the expanding digital universe will grow from 36% to 62%.
- A majority of the information in the digital universe, 68% in 2012, is created and consumed by consumers — watching digital TV, interacting with social media, sending camera phone images and videos between devices and around the Internet, and so on. Yet enterprises have liability or responsibility for nearly 80% of the information in the digital universe. They deal with issues of copyright, privacy, and compliance with regulations even when the data zipping through their networks and server farms is created and consumed by consumers.
- Only a tiny fraction of the digital universe has been explored for analytic value. IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed.
- By 2020, nearly 40% of the information in the digital universe will be "touched" by cloud computing providers — meaning that a byte will be stored or processed in a cloud somewhere in its journey from originator to disposal.
- The proportion of data in the digital universe that requires protection is growing faster than the digital universe itself, from less than a third in 2010 to more than 40% in 2020.
- The amount of information individuals create themselves — writing documents, taking pictures, downloading music, etc. — is far less than the amount of information being created *about them* in the digital universe.
- Much of the digital universe is transient — phone calls that are not recorded, digital TV images that are watched (or "consumed") that are not saved, packets temporarily stored in routers, digital surveillance images purged from memory when new images come in, and so on. Unused storage bits installed throughout the digital universe will grow by a factor of 8 between 2012 and 2020 but will still be less than a quarter of the total digital universe in 2020.

©2012 IDC

Figure 1

The Digital Universe: 50-fold Growth from the Beginning of 2010 to the End of 2020



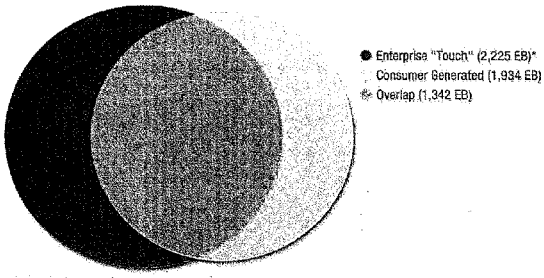
Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Within these broad outlines of the digital universe are some singularities worth noting.

First, while the portion of the digital universe holding potential analytic value is growing, only a tiny fraction of territory has been explored. IDC estimates that by 2020, as much as 33% of the digital universe will contain information that might be valuable if analyzed, compared with 25% today. This untapped value could be found in patterns in social media usage, correlations in scientific data from discrete studies, medical information intersected with sociological data, faces in security footage, and so on. However, even with a generous estimate, the amount of information in the digital universe that is "tagged" accounts for only about 3% of the digital universe in 2012, and that which is analyzed is half a percent of the digital universe. Herein is the promise of "Big Data" technology — the extraction of value from the large untapped pools of data in the digital universe.

Figure 2

The Impact of Consumers (2012)



* Enterprise has some liability or responsibility

Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

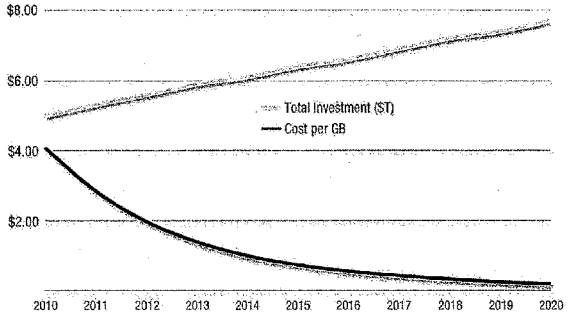
Moreover, IDC believes that much of the digital universe is unprotected. Our estimate is that about a third of the data in the digital universe requires some type of protection — to protect privacy, adhere to regulations, or prevent digital snooping or theft. However, currently, only about 20% of the digital universe actually has these protections. The level of protection varies by region, with much less protection in emerging markets.

Therefore, like our own physical universe, the digital universe is rapidly expanding and incredibly diverse, with vast regions that are unexplored and some that are, frankly, scary.

©2012 IDC

Figure 3

The Digital Universe Paradox:
Falling Costs and Rising Investment



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

However, the digital universe astronauts among us — the CIOs, data scientists, digital entrepreneurs — already know the value that can be found in this ever-expanding collection of digital bits. Hence, there is excitement about Big Data technologies, automatic tagging algorithms, real-time analytics, social media data mining, and myriad new storage technologies.

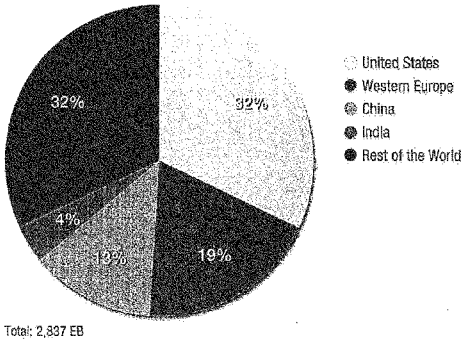
The Geography of the Digital Universe

Although the bits of the digital universe may travel at Internet speeds around the globe, it is possible to assign a place of origin to them and chart the map of the digital universe.

In this year's study, for the first time, we have managed to determine where the information in the digital universe was either generated, first captured, or consumed. This geography of the digital universe maps to the users of the devices or applications that pump bits into the digital universe or pull bits into one's own personal digital solar system for the purpose of consuming information — Internet users, digital TV watchers, structures hosting surveillance cameras, sensors on plant floors, and so on.

Figure 4

The Geography of the Digital Universe (2012)

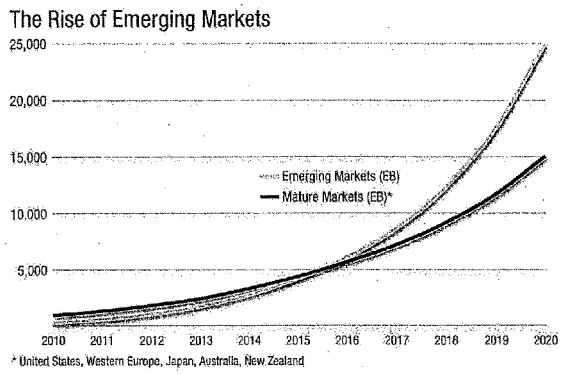


Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

In the early days, the digital universe was a developed world phenomenon, with 48% of the digital universe in 2005 springing forth from just the United States and Western Europe. Emerging markets accounted for less than 20%. However, the share of the digital universe attributable to emerging markets is up to 36% in 2012 and will be 62% by 2020. By then, China alone will generate 21% of the bit stream entering the digital universe.

©2012 IDC

Figure 5



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

It stands to reason. Even though China accounts for only 11% of global GDP today, by 2020 it will account for 40% of the PCs, nearly 30% of smartphones, and nearly 30% of Internet users on the planet — not to mention 20% of the world population.

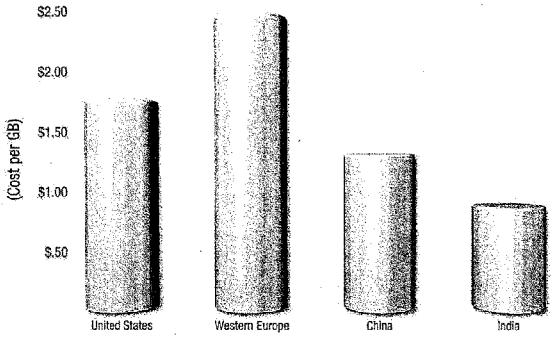
At the same time, the money invested by the regions in creating, managing, and storing their portions of the digital universe will vary wildly — in real dollar terms and as a cost per gigabyte.

This disparity in investment per gigabyte represents to some extent differing economic conditions — such as the cost of labor — and to some extent a difference in the types of information created, replicated, or consumed. The cost per gigabyte from bits generated by surveillance cameras will be different from the cost per gigabyte from bits generated by camera phones.

However, to *another* extent, this disparity also represents differences in the sophistication of the underlying IT, content, and information industries — and may represent a challenge for emerging markets when it comes to managing, securing, and analyzing their respective portions of the digital universe.

Figure 6

Investment in Managing the Digital Universe by Region (2012)



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

This might not be a major issue if the geography of the digital universe were as stable and fixed as, say, the geography of countries. However, bits created in one part of the physical world can easily find themselves elsewhere, and if they come with malware attached or leaky privacy protections, it's a problem. The digital universe is like a digital commons, with all countries sharing some responsibility for it.

The installed base of unused storage bits introduces an interesting geographic twist that establishes a new dynamic by which to understand our digital universe. While emerging markets may indeed grow as a percentage of the digital universe, remember that much of the digital universe is a result of massive consumption on mobile and personal devices, digital televisions, and cloud-connected applications on PCs. As ownership of smartphones and tablets (that have relatively low internal storage and rely heavily on consuming information from "the cloud") increases exponentially within emerging markets, information consumption grows at an even faster pace. Given the connected infrastructure of our digital universe, information does not need to (and in fact will not) reside within the region where the information is consumed. Hence, today's well-running datacenters will continue to expand and to fulfill an increasing number of requests — both local and from halfway across the globe — for information.

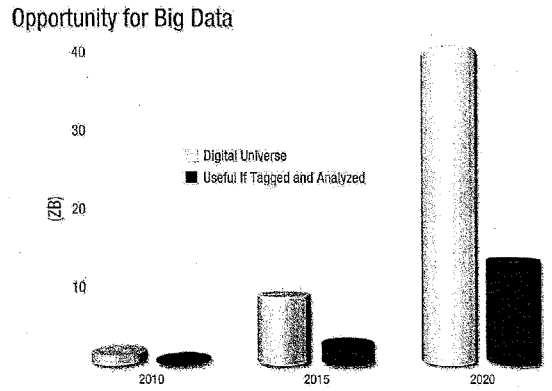
©2012 IDC

Big Data in 2020

Last year, Big Data became a big topic across nearly every area of IT. IDC defines Big Data technologies as *a new generation of technologies and architectures, designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery, and/or analysis*. There are three main characteristics of Big Data: the data itself, the analytics of the data, and the presentation of the results of the analytics. Then there are the products and services that can be wrapped around one or all of these Big Data elements.

The digital universe itself, of course, comprises data — all kinds of data. However, the vast majority of new data being generated is unstructured. This means that more often than not, we know little about the data, unless it is somehow characterized or tagged — a practice that results in metadata. Metadata is one of the fastest-growing subsegments of the digital universe (though metadata itself is a small part of the digital universe overall). We believe that by 2020, a third of the data in the digital universe (more than 13,000 exabytes) will have Big Data value, but only if it is tagged and analyzed (see "Opportunity for Big Data).

Figure 7



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Not all data is necessarily useful for Big Data analytics. However, some data types are particularly ripe for analysis, such as:

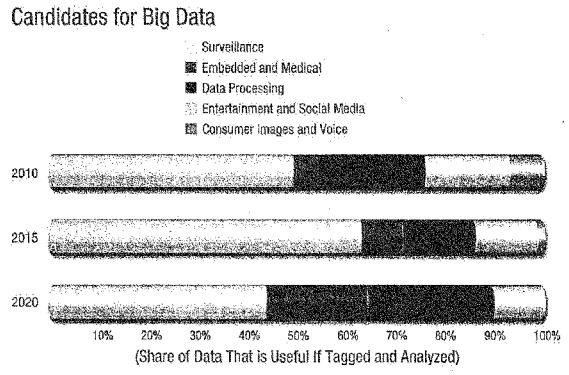
- **Surveillance footage.** Typically, generic metadata (date, time, location, etc.) is automatically attached to a video file. However, as IP cameras continue to proliferate, there is greater

opportunity to embed more intelligence into the camera (on the edge) so that footage can be captured, analyzed, and tagged in real time. This type of tagging can expedite crime investigations, enhance retail analytics for consumer traffic patterns, and, of course, improve military intelligence as videos from drones across multiple geographies are compared for pattern correlations, crowd emergence and response, or measuring the effectiveness of counterinsurgency.

- **Embedded and medical devices.** In the future, sensors of all types (including those that may be implanted into the body) will capture vital and nonvital biometrics, track medicine effectiveness, correlate bodily activity with health, monitor potential outbreaks of viruses, etc. — all in real time.
- **Entertainment and social media.** Trends based on crowds or massive groups of individuals can be a great source of Big Data to help bring to market the "next big thing," help pick winners and losers in the stock market, and yes, even predict the outcome of elections — all based on information users freely publish through social outlets.
- **Consumer images.** We say a lot about ourselves when we post pictures of ourselves or our families or friends. A picture used to be worth a thousand words, but the advent of Big Data has introduced a significant multiplier. The key will be the introduction of sophisticated tagging algorithms that can analyze images either in real time when pictures are taken or uploaded or en masse after they are aggregated from various Web sites.

These are in addition, of course, to the normal transactional data running through enterprise computers in the course of normal data processing today. "Candidates for Big Data" illustrates the opportunity for Big Data analytics in just these areas alone.

Figure 8



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

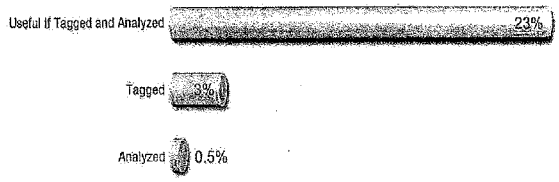
©2012 IDC

All in all, in 2012, we believe 23% of the information in the digital universe (or 643 exabytes) would be useful for Big Data if it were tagged and analyzed. However, technology is far from where it needs to be, and in practice, we think only 3% of the potentially useful data is tagged, and even less is analyzed.

Call this the Big Data gap — information that is untapped, ready for enterprising digital explorers to extract the hidden value in the data. The bad news: This will take hard work and significant investment. The good news: As the digital universe expands, so does the amount of useful data within it.

Figure 9

The Untapped Big Data Gap (2012)



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Information Security in 2020

The rise in mobility and participation in social networks, the increasing willingness to share more and more data, new technology that captures more data about data, and the growing business around Big Data all have at least one assured outcome — the need for information security.

However, the news from the digital universe is as follows:

- The proportion of data in the digital universe that requires protection is growing faster than the digital universe itself, from less than a third in 2010 to more than 40% in 2020.
- Only about half the information that *needs* protection *has* protection. That may improve slightly by 2020, as some of the better-secured information categories will grow faster than the digital universe itself, but it still means that the amount of unprotected data will grow by a factor of 26.
- Emerging markets have even less protection than mature markets.

In our annual studies, we have defined, for the sake of analysis, five levels of security that can be associated with data having some level of sensitivity:

1. **Privacy** only — an email address on a YouTube upload
2. **Compliance** driven — emails that might be discoverable in litigation or subject to retention rules
3. **Custodial** — account information, a breach of which could lead to or aid in identity theft

4. **Confidential** — information the originator wants to protect, such as trade secrets, customer lists, confidential memos, etc.

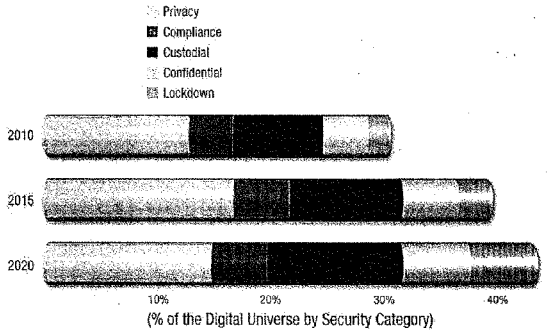
5. **Lockdown** — information requiring the highest security, such as financial transactions, personnel files, medical records, military intelligence, etc.

The tables and charts illustrate the scope of the security challenge but not the solution. While information security technology keeps getting better, so do the skills and tools of those trying to circumvent these protections. Just follow the news on groups such as Anonymous and the discussions of cyberwarfare.

However, for enterprises and, for that matter, consumers, the issues may be more sociological or organizational than technological — data that is not backed up, two-phase security that is ignored, and corporate policies that are overlooked. Technological solutions will improve, but they will be ineffective if consumer and corporate behavior doesn't change.

Figure 10

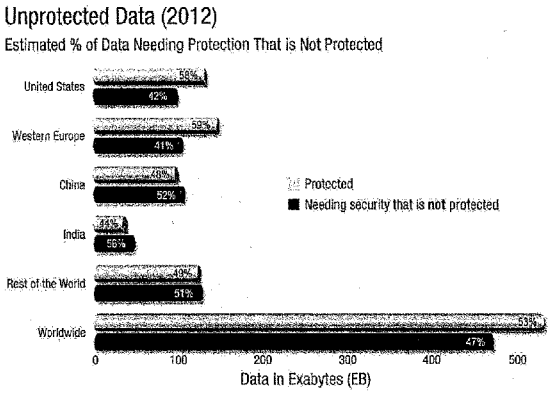
The Need for Information Security



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

©2012 IDC

Figure 11



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

Big Data is of particular concern when it comes to information security. The lack of standards among ecommerce sites, the openness of customers, the sophistication of phishers, and the tenacity of hackers place considerable private information at risk. For example, what one retailer may keep private about your purchase, such as your transaction and customer profile data, another company may not and instead may have other data hidden. Yet intersecting these data sets with other seemingly disparate data sets may open up wide security holes and make public what should be private information.

There is a huge need for standardization among retail and financial Web sites as well as any other type of Web site that may save, collect, and gather private information so that individuals' private information is kept that way.

Cloud Computing in 2020

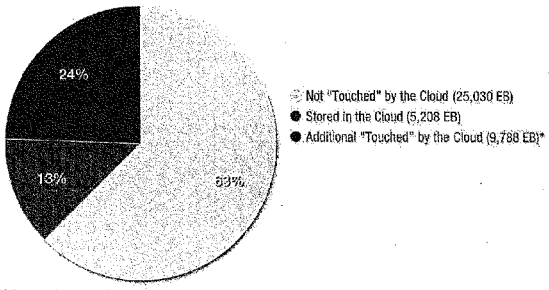
Between 2012 and 2020, the patch of the digital universe that CIOs and their IT staffs need to manage will become not just bigger but also more complex. The skills, experience, and resources to manage all these bits of data will become scarcer and more specialized, requiring a new, flexible, and scalable IT infrastructure that extends beyond the enterprise: cloud computing.

To this end, the number of servers (virtual and physical) worldwide will grow by a factor of 10 and the amount of information managed directly by enterprise datacenters will grow by a factor of 14. Meanwhile, the number of IT professionals in the world will grow by less than a factor of 1.5.

In addition, while spending on public and private cloud computing accounts for less than 5% of total IT spending today, IDC estimates that by 2020, nearly 40% of the information in the digital universe will be "touched" by cloud computing — meaning that a byte will be stored or processed in a cloud somewhere in its journey from originator to disposal. Perhaps as much as 15% will be *maintained* in a cloud.

Figure 12

The Digital Universe and the Cloud (2020)



* Processed or transmitted by the cloud, but not stored

Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

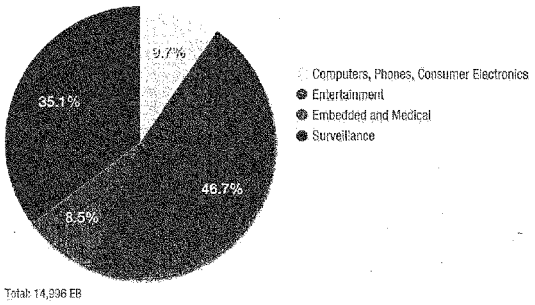
Of course, cloud services come in various flavors — public, private, and hybrid. For organizations to offer their own cloud services, they have to do more than just run virtual servers. They must also allow for virtualized storage and networking, self-provisioning, and self-service and provide information security and billing.

Part of the real genesis of this conversion to the cloud will be a migration to converged infrastructures, where servers, storage, and networks are integrated together, sold, and installed as a unit of IT infrastructure. Few enterprises are at this point yet, so the impact of private clouds in the digital universe today is small.

©2012 IDC

Figure 13

Type of Information in the Cloud in 2020



Source: IDC's Digital Universe Study, sponsored by EMC, December 2012

However, by 2020, it seems likely that private clouds and public clouds will be commonplace, exchanging data seamlessly. There won't be one cloud; rather, there will be many clouds, bounded by geography, technology, different standards, industry, and perhaps even vendor. We may still call it cloud computing, but it will be an interconnected ether, easy to traverse but difficult to protect or manage.

Call to Action

Our digital universe in 2020 will be bigger than ever, more valuable than ever, and more volatile than ever.

By 2020, we'll also be storing a smaller and smaller percentage of our expanding digital universe; yet our digital shadows will be larger than life and on the move given the increase in mobility, and they will require more protection than ever before. IT managers will be responsible not only for ensuring that proper security surrounds our digital lives but also for managing the storing, analyzing, and delivery of zettabytes of content ...no easy task.

Requests for data could come from a faraway jungle, across a mashup of connected devices and network points, to a device that has an obtuse screen. The delivery of the requested data must happen in an acceptable amount of time, guaranteeing that it is consumed flawlessly; if not, then a business may lose a customer. Consider this:

- The network is growing in importance. Latencies must get shorter, not longer. Data must be analyzed, security applied, and authentication verified — all in real time and in levels yet to be seen. Network infrastructure must be a key investment to prepare for our 2020 digital universe.
- Big Data is going to be a big boon for the IT industry. Web sites that gather significant data need to find ways to monetize this asset. Data scientists must be absolutely sure that the intersection of disparate data sets yields repeatable results if new businesses are going to emerge and thrive. Further, companies that deliver the most creative and meaningful ways to display the results of Big Data analytics will be coveted and sought after.
- The laws and regulations governing information security must harmonize around the globe, though differences (or absences) will certainly exist. IT managers must realize that data will be requested outside geographic boundaries, and a global knowledge of information security may be the difference between approval and denial of a data request.
- IT managers must find ways to drive more efficiency in their infrastructures so that IT administrators can focus on more value-add initiatives such as "bring your own device" (BYOD) policies, Big Data analytics, customer onboarding efficiency, security, etc. One way this is likely to happen is through converged infrastructures, which integrate storage, servers, and networks.

Are you ready to create, consume, and manage 40 trillion gigabytes of data?

ABOUT THIS PUBLICATION

This publication was produced by IDC Go-to-Market Services. The opinion, analysis, and research results presented herein are drawn from more detailed research and analysis independently conducted and published by IDC, unless specific vendor sponsorship is noted. IDC Go-to-Market Services makes IDC content available in a wide range of formats for distribution by various companies. A license to distribute IDC content does not imply endorsement of or opinion about the licensee.

COPYRIGHT AND RESTRICTIONS

Any IDC information or reference to IDC that is to be used in advertising, press releases, or promotional materials requires prior written approval from IDC. For permission requests, contact the GMS information line at 508-988-7610 or gms@idc.com. Translation and/or localization of this document requires an additional license from IDC.

For more information on IDC, visit www.idc.com. For more information on IDC GMS, visit www.idc.com/gms.

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

©2012 IDC