THE NATIONAL ACADEMIES

MEETING IN BRIEF

The Big Data Revolution
What Does It Mean for Research?

Government-University-Industry Research Roundtable
October 14-15, 2014

Big data offers a range of new research opportunities and challenges (e.g., workforce and training issues, ethics and privacy concerns, and new chances for public-private partnerships) that impact the scientific research community. On October 14-15, 2014, the Government-University-Industry Research Roundtable held a meeting to explore these issues and discuss how big data is changing research.

The meeting's keynote speech was given by **Chaitan Baru,** Senior Advisor for Data Science in the National Science Foundation's (NSF) Directorate for Computer and Information Science and Engineering. Baru recently joined NSF in this newly created position after 17 years at the San Diego Supercomputer Center.

Big data is well-understood to be a vague term with an evolving definition, said Baru. "We are at a tipping point in terms of the amount and kinds of data coming in and the realization of the kinds of applications we can develop." Big data has created a greater awareness of all aspects of data, including the life cycle of data and the importance of metadata. It has also generated new research energy around data.

Ultimately, we want to go from building data infrastructure to using the data effectively and in a timely way to enable applications, he said. Because data come from the "domains"—the term computer scientists use for the other disciplines, those who are familiar with data in the domains are better equipped to readily produce effective tools and technologies.

Baru gave examples of how big data is generated through NSF's work in the domains. When the Sloan Digital Sky Survey began in 2000, it collected more data in its first few weeks of operation than had been amassed in the entire history of astronomy; within a decade, it collected over 140 terabytes of information. The Large Synoptic Survey Telescope, scheduled to start operating in 2016 in Chile, will amass that quantity of data in 10 days. Earthscope is another NSF project that takes a 4-D seismic snapshot of North America, moving at the rate of plate tectonics. The National Ecological Observatory Network (NEON) studies terrestrial ecology across the continental United States. The data are collected using standardized protocols across many different locations. The data are very heterogeneous, and some are collected by citizen scientists.

Most of these initiatives are focused on major research facilities that are creating new instruments to observe phenomena in order to collect new data and then do new science with it. However, we are at a stage where the data itself could be the instrument, said Baru; there is so much data that researchers can make some discoveries just by examining the data—by building hypotheses and then digging into the data to investigate them. Data science at this point is as much or more about discovery than about typical scientific hypotheses that leads to a proof. Indeed hypothesis discovery may by closer to the truth for much of current data science.

We are currently focused on building data acquisition systems, data repositories, data commons, and other infrastructure, said Baru. Looking ahead, what we really want are software/technology environments that allow the end user—for example, a biologist or ecologist—to use all of this data in an easy, highly facilitated way. The big payoffs in big data lie in integrating data across disciplines and subdisciplines that requires breaking silos, including working across directorates at NSF and across federal agencies.

Privacy and ownership of big data are significant issues, and big data is rife with ethics problems, continued Baru. In March 2014, he attended a US-UK data science summit in Maryland where half of the participants were social scientists and half were computer scientists; the top issue at the end of two days was ethics. Ethics cannot be addressed simply by sending a student in the computer science department to take a class in the philosophy department, said Baru. Rather, it means taking advantage of teachable moments in software engineering. Other big questions surrounding big data are how to value data and when it is acceptable to delete data.

The first presentation the next day was given by **Philip Bourne,** Associate Director for Data Science at the National Institutes of Health (NIH). He noted that the entire data holdings of the National Center for Biotechnology Information in 1993 could fit on a single CD-ROM; 20 years later, the Center has amassed the equivalent of 40 million four-drawer filing cabinets.

Bourne discussed four areas of challenge and promise associated with big data:

1) How data are used actually informs us; unexpected discoveries are made. Currently, we often don't know how the data we have are used, but closely examining usage patterns could yield interesting findings, said Bourne. For example, when he and his colleagues examined access patterns to data on certain proteins in the Protein Databank, they saw that the data correlated with the number of reported cases of H1N1 flu. The question became: Could we have used this as a leading indicator to better understand what's happening with the flu outbreak and act accordingly? Examining and measuring data using this method could have a profound impact and potentially change how we value data—which are currently undervalued in the realm of scholarship.

2) Researcher X discovers researcher Y as a potential collaborator at the point of data generation, not publication; progress is accelerated. Currently, researchers have to wait until a paper comes out to find out what research is going on, Bourne said. Progress could be accelerated by looking for patterns of similarity in data starting the day it is collected then alerting researchers: "Researcher X, you should be talking to researcher Y." Bourne explained that this process could be aided by gathering data in a single place where it would be easy to see patterns of similarity—a possibility NIH has begun to explore through the Commons, a public-private data-gathering partnership.

3) Demand for data science training is met by supply; workforce development yields economic gains. Currently, the demand for data science training is not being met by the supply, said Bourne. In response, NIH is trying to develop a variety of training initiatives that entail either training data scientists in biomedicine or providing needed analytical training to people already in biomedicine. NIH is also working with colleagues in Europe to come up with standards and metadata representation to describe the fast-growing mass of both virtual and physical data-science courses, to catalog them and make them searchable.

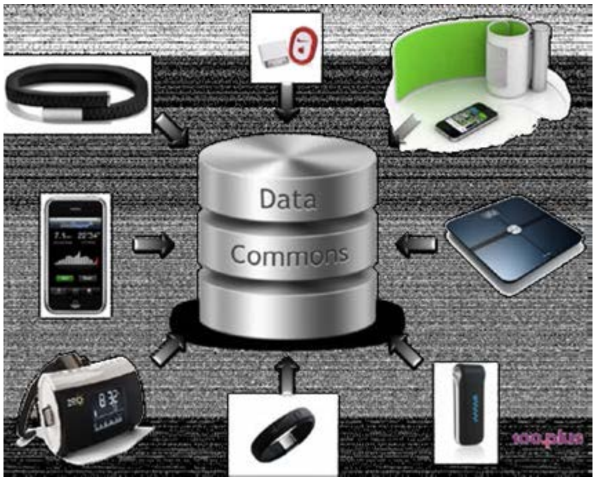


FIGURE 1 NIH Approach to Big Data
SOURCE: Presented by Phil Bourne, October 14-15, 2014.

4) Sustaining the ecosystem amidst rapid data growth and changing usage patterns; discovery continues. From a sustainability standpoint, we need to start thinking about data resources as a business model, said Bourne. Right now, NIH's standard grant awards provide full funding for an unspecified number of years, and then funding stops completely. What we need to do instead, Bourne offered, is offer full funding at the beginning of a project but specify that the project has four years of full funding—after which funding will drop to 50 percent—to develop a self-sustaining business model.

The next speaker, **H.V. Jagadish,** Bernard A. Galler Collegiate Professor of Electrical Engineering and Computer Science at the University of Michigan, discussed some myths and realities associated with big data.

Myth: It's all hype. While there is no question that there is hype around big data, the existence of hype doesn't mean there's no substance and that it's not worth paying attention to it, said Jagadish. Given our ability to collect and store data cheaply, nearly every field of endeavor is transitioning from data-poor to data-rich; whether we can extract data of value from the ocean of data becomes the important question. In addition, a new data-intensive mode of inquiry—one that is not hypothesis-driven—has become a tool in the research arsenal.

Myth: Size is all that matters. It is important not to focus solely on volume, said Jagadish. The Gartner Group and IBM have pushed to have other characteristics also—variety and veracity—included in the definition. People tend to focus on volume and velocity because they are measurable; but the variety and veracity of data are far more challenging and deserve more attention. In the big-data pipeline, important and challenging actions also need to be taken at the front end; and after the analysis, it's necessary to interpret it and ensure that whoever is going to act on the analysis has enough understanding and trust in how it was conducted.

Myth: Data re-use is low-hanging fruit. Instead of simply thinking of a question and then collecting data to answer it, researchers are now thinking, what data already exists? While that is right and appropriate, data-reuse is not low-hanging fruit, said Jagadish. Data are often organized in ways that make it difficult to correlate with other data that are organized differently. Data can also be difficult for a third party to use without metadata that explain what the fields are, what they mean, and where they came from; currently there are not enough incentives to add that metadata.

Myth: Data science is the same as big data. Big data focuses on the means, and data science focuses on the ends. Data science is the use of data to address a question in a domain or field of interest through methodologies that come from those domains.

Myth: The central challenge is computer algorithms (and big data is all in the cloud). This misses the point, said Jagadish, whose definition of the threshold for being "big data" is "more than you know how to handle"—a threshold that depends on the context. If you posted a job advertisement and got five resumes, you would know how to handle them and be able to pick a good candidate to interview or work with. But what if you got 5,000 resumes? That becomes a big data problem at only a few thousand, raising the question: What do computer systems need to do in order to help you deal with those resumes?

The next panel explored ethics and privacy concerns related to big data, with the first presentation offered by **Julia Lane** of the American Institutes for Research, the University of Strasbourg, and the University of Melbourne. There is no canonical definition for big data, said Lane; one can think about big data as a rich and complicated set of characteristics, practices, techniques, ethics, and outcomes. Big data is not so much data as a change in an analytical model. Big data also lets us do fine-grained analyses, she added.

Big data is prompting a fundamental change in the way we need to think about privacy, said Lane. It raises privacy concerns because of the merging and crossing of data in analysis, and there isn't a particularly good roadmap for dealing with these issues. In some ways, said Lane, privacy and big data are incompatible.

The context for her comments is the science of science policy, she continued; we've got to be able to understand who is being funded, what research is being funded, and the results, and currently we do not. Data on federally funded research sits in multiple agencies. An initiative known as STAR METRICS is attempting to pull data sets together so as to create a repository that will be useful in assessing the impact of federal R&D investments.

She and her colleagues wrote a book on privacy issues related to big data, *Privacy, Big Data, and the Public Good.* The first half of the book discusses the legal framework for privacy. The previous model for dealing with privacy depended upon informed consent, and that is not possible in a big data era, she said. The question is: What is the marginal risk associated with research access, and what kind of rules should we have? The second part of the book deals with a practical framework for privacy issues related to big data. It is crucial that we use big data for the public good, noted Lane. The book's third part explores the importance of valid inference and the need for a new analytical framework with a mathematically rigorous theory of privacy and the capacity to measure privacy loss.

A second perspective on privacy issues was offered by **Alvaro Bedoya,** Executive Director of the Center on Privacy and Technology at Georgetown University. Bedoya focused his talk on a set of normative claims about big data that he said have received less attention than they merit.

Much of his presentation focused on privacy issues related to geo-location data. Half of the top apps collect individuals' geographic location and share it profusely, he said; generally, firms can sell people's location data to anyone. Passive information—meaning information that does not require active input from the user—can be resold without consent whereas active information—such as that collected through Google maps—cannot be resold without consent. In other words, he observed, the less you know about your data, the fewer protections there are.

One normative claim related to big data is that we should rely on correlation rather than causation, said Bedoya. The critique from the privacy standpoint is: Yes, correlations are there, but are they fair to use? For example, researchers found that length of commute time is inversely proportionate to how long a person stays with a specific employer. Is it fair to use that correlation? Bedoya posed these questions and others as examples of privacy issues that he thought should be publically debated.

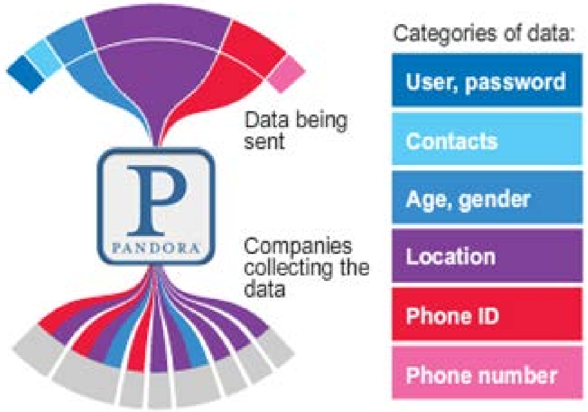


FIGURE 2 How is Location Data Shared?
SOURCE: Presentation by Alvaro Bedoya, October 14-15, 2014.

Other normative claims about big data suggest that algorithmic decision making is less biased than human decision making and that ubiquitous data collection is inevitable. It's not inevitable, said Bedoya. If there is a consensus rule, then, you should get consent to collect data, and as companies break that rule, other companies innovate to protect consumers against ubiquitous data collection. Accepting ubiquitous data collection as inevitable would undercut critical innovation around protections for privacy. To illustrate this point, Bedoya contrasted the privacy consent features provided by Google and Apple. Bedoya showed that Apple gives consumers more options to limit the access to their personal data than Google currently does.

Yet another claim is that ubiquitous data collection is not only inevitable but also good. We need to distinguish between research that benefits all versus research that benefits a narrow few, he said. Advocates point out the economic benefits of collecting data—e.g., that geotargeted ads generate revenue that's five times higher than non-geotargeted ads. Perhaps, Bedoya posited, it is ethically unsavory to allow for privacy infringement when the benefits are limited to the economic gain of a small subgroup. On the other hand, it may be more ethically acceptable to collect information that will ultimately benefit a larger group, especially a group that includes the person whose information is being collected.

In addition to not being inevitable, ubiquitous data collection is also not desirable in that geo-location can "out" vulnerable communities, Bedoya continued. Based on the time of day and your geographic information, it's possible to determine where you live, where you work, and whether you spent any time at an HIV clinic. Our society is historically slow to protect information on vulnerable populations; for example, during World War II, Congress mandated that the Census Bureau provide block-by-block information on Japanese Americans. Eliminating or deemphasizing user controls on data threatens to disproportionately harm vulnerable communities. Bedoya illustrated his point by explaining the highly personal information that could be garnered from time stamped location data on a map of a fictitious person in the District of Columbia.

A third perspective on privacy issues was offered by **Jules Polonetsky,** Executive Director and co-chair of the Future of Privacy Forum, who focused his talk on corporate ethics. The Future of Privacy Forum represents 120 or so of the chief privacy officers of the world—from big tech companies to small startups to big brands—and also includes academics and advocates who provide a healthy skepticism on the claims and want to see progress moving forward, said Polonetsky.

He said that while he agreed with Bedoya's concerns, accepting many of his premises would undermine many of the goals and objectives of using big data. He expressed doubt that many consumers would sit down for an informed consent briefing, fully understanding and agreeing to the implications of an app before using it. "In almost every legal and policy conversation in which I take part, we come to the conclusion that the government has abused and will abuse its powers of data collection and use," said Polonotsky. Given that, do we give people tools that help them hide from the NSA, or do we trust that the right structure and control and oversight and transparent processes can be put in place so that the data can be used while also providing protections for people?

"I fear that we will have less government-industry-academic cooperation if we don't figure out the right ethical model for companies," said Polonetsky For areas where there is no clear societal consensus about data use, Polonetsky suggested that we will need corporate ethics boards who are freed from day-to-day concerns about making money and can ask whether it is appropriate to use data in a certain way, whether it would harm people (even if not a legal harm), whether it will make customers unhappy, and whether it is morally defensible.

Big data is bad data in terms of applying bedrock privacy principles, observed Polonetsky. We have well-accepted data privacy principles—individual control, limiting collection, specifying purpose, and minimizing data—that are the source of privacy law around the world, and these principles are all bumping into big data. Much of big data use is far outside of the context from which it was collected.

We're very good at analyzing risks, he continued, but how does one analyze benefit? Who benefits is relevant; if data collection is going to directly benefit the person from whom it is collected, there may be more license. "If I'm going to help all users of the product, that may be fine." There is a lot of skepticism among privacy advocates that big data is good for anything more than selling ads, concern over whether it is being used legitimately, and concern that it will be used for discrimination and denying benefits, said Polonetsky.

The next presentation, which focused on workforce and training issues, was given by **Raghu Ramakrishnan,** a technical fellow at Microsoft. Big data systems can be transformative, he said; we can do things cost-effectively we couldn't contemplate before, resulting in data-driven science, commerce, government and social programs, manufacturing, and medicine. He gave the example of the field of ecology, where there are some huge questions we can't answer—for example, how is the climate changing, and will forests accelerate or slow climate change?—that a new approach to ecology that generates quantitative predictions might be able to help with, said Ramakrishnan.

The biggest bottleneck in harnessing big data is the lack of trained people, said Ramakrishnan. He described the data analysis pipeline, which begins with asking a general question and then framing it in light of available and relevant data. We then query and transform the data in order to prepare it, and then do an exploratory analysis to get a feel for the data and the insights it could offer. Following that is modeling, interpretation, and reaching actionable insights.

Ramakrishnan listed some of the skills that need to be taught in order to support that pipeline, including data management, machine learning, parallel computing, security, software engineering, statistical analysis and inference, and visualization. It is also important to ground this learning in the domain in which the skills will be applied—formulating real problems and analyzing specific data in order to reach actionable outcomes. He pointed out some challenges and remaining questions surrounding curricula and programs for big data and data science. Should big data and data science be a new discipline, or should it be seen as a complementary skill set for other disciplines? Do we introduce big data and data science at the undergraduate or graduate level? In addition, we're not going to do this topic justice by stitching together existing courses, said Ramakrishnan.

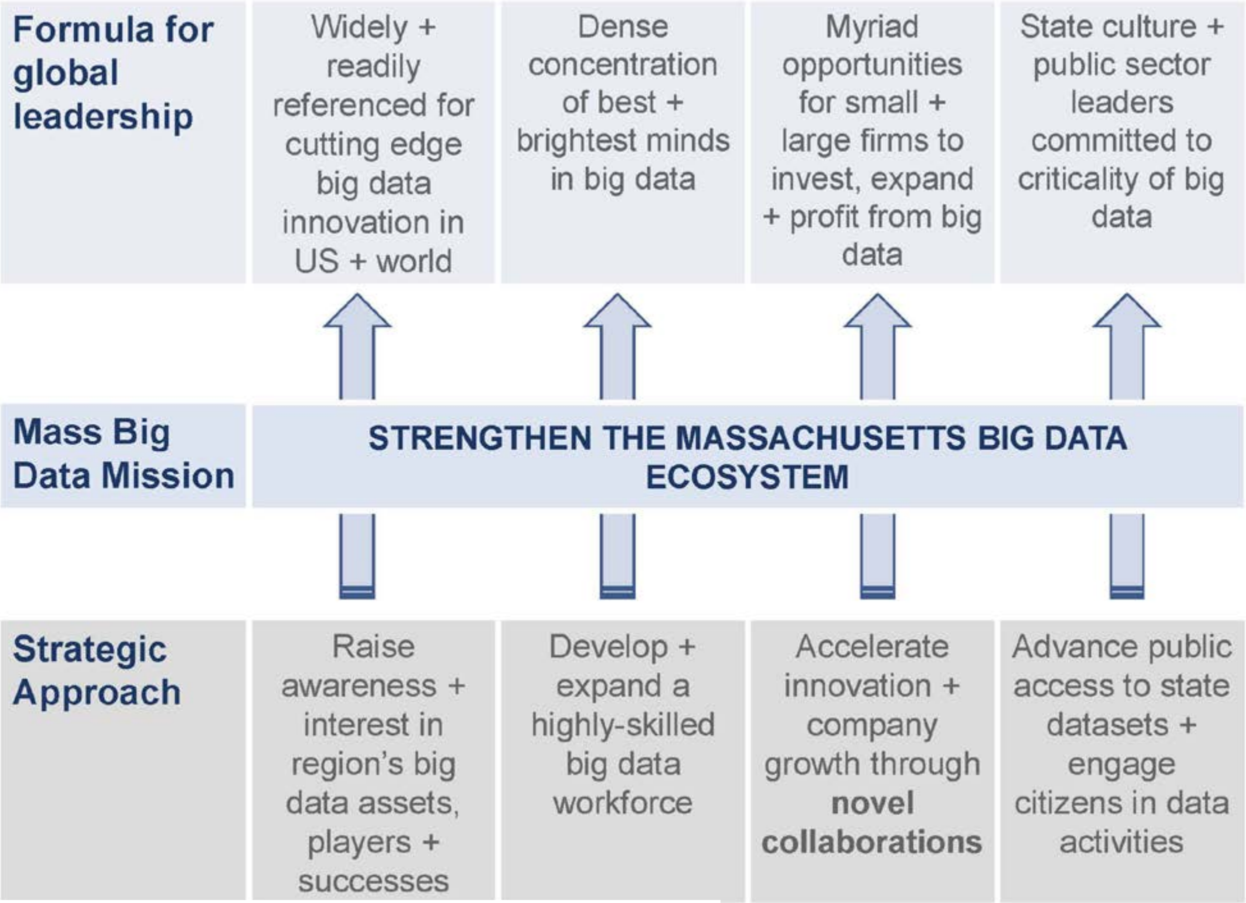
The final set of presentations discussed public-private partnerships around big data. **Valerie Taylor** of the Dwight Look College of Engineering and Texas A&M University (TAMU) described a partnership Texas A&M began with IBM 2 years ago, under the leadership of the university system's chancellor. The partnership includes Texas A&M's 11 universities, whose student body totals 115,000 students. The partnership focuses on research collaborations between IBM staff and faculty in the TAMU system, and particularly on long-term collaborations on fundamental and applied research, Taylor explained. The partnership has "professors of practice"—industry researchers who teach classes.

Since the partnership began, it has centered on four Grand Challenges: sustainable availability of food, disease spread tracking/modeling/prediction, energy resource management, and climate/ocean modeling. Specific areas of research engagement cross three colleges: AgriLife, which is involved in research on plant and animal genomics; the College of Geosciences, which is engaged in research on petroleum, gulf climate modeling, and water modeling; and the College of Engineering, which is doing research on health surveillance, genomic signal processing, and emergency informatics, among other areas.

They are about a year into the collaboration in some specific areas, including manufacturing, said Taylor. Researchers from IBM Watson, Austin, and Almaden are collaborating with TAMU faculty in materials science engineering and industrial systems engineering; they have weekly calls and monthly visits. Taylor described the layered computer architecture that supports the data analysis that is part of the research, as well as the interactions between IBM staff and university staff that happen as part of the collaboration. IBM provides research staff for exploration of technologies, technical and research support, and fellowships and internships for students. It took about nine months to finalize the agreement that created the partnership, and the agreement is now public so that others can see and use it. “We think the partnership is off to a great start,” said Taylor in conclusion.

The final presentation was given jointly by **Pat Larkin,** Executive Director of the Innovation Institute at the Massachusetts Technology Collaborative, and **John Goodhue,** Executive Director of the Massachusetts Green High Performance Computing Center, who spoke about activities underway in Massachusetts related to big data.

Larkin described a big data initiative managed by the Innovation Institute, a quasi-governmental economic development state agency. The initiative was announced by Governor Deval Patrick in 2012, based on the



Formula for global leadership

Widely + readily referenced for cutting edge big data innovation in US + world

Dense concentration of best + brightest minds in big data

Myriad opportunities for small + large firms to invest, expand + profit from big data

State culture + public sector leaders committed to criticality of big data

Mass Big Data Mission

STRENGTHEN THE MASSACHUSETTS BIG DATA ECOSYSTEM

Strategic Approach

Raise awareness + interest in region’s big data assets, players + successes

Develop + expand a highly-skilled big data workforce

Accelerate innovation + company growth through **novel collaborations**

Advance public access to state datasets + engage citizens in data activities

Figure 3 Formula for Global Leadership
SOURCE: Presented bv Pat Larkin. October 14-15. 2014.

recommendations of a technology leadership roundtable, where stakeholders in big data came to the Commonwealth and made the case that the state has many indigenous strengths in big data and needs to focus on that advantage.

Those strengths include a strong critical mass around industry, over $2 billion invested in big data by the venture community, great research institutions, and 5,000 students graduating annually from Massachusetts colleges and universities in data science related programs.

The initiative’s strategy to maintain Massachusetts’ global leadership in big data, formulated with leaders in industry and research institutions, includes four pillars:
1) **Raising awareness**. So much work is going on in big data in Massachusetts that it’s necessary to help connect the component parts; institutions and companies didn’t always know what was going on next door in the Mass innovation ecosystem for big data.
2) Develop and expand a highly skilled big data work force. The initiative is trying to capture more talent and develop opportunities for students to engage with industry through hackathons, Tech Trek’s, internships, and other activities.
3) Novel collaborations as described below.
4) Advance public access to state datasets. “One leverage point we have as a Commonwealth is public datasets, and so we need to think about the regime around that,” said Larkin. “We’re making progress, but it’s not easy.”

Larkin then elaborated on the “novel collaborations” pillar, providing an overview of two key partnerships. The Mass Open Data Laboratory emerged through collaboration with the Massachusetts Department of Transportation, whose secretary believed that opening this data set could help them improve the department’s budgeting, infrastructure investments, and service. The Mass Open Data Laboratory will give capability in three areas: 1) the ability to capture the data sets; 2) the tools needed to analyze and extract value from the data, and 3) the resources necessary to support and use the data and tools to prepare the data for wider access when needed (e.g. anonymization). The Laboratory acts as a neutral repository for data and will be a safe harbor for people to be able to work on internal implementation to meet agency needs; over time, they want to open it up to the public and industry.

The other collaborative project is the Mass Open Cloud, which is envisioned as a shared infrastructure that stakeholders in government, industry, and academia can use to develop and run their computationally intensive datasets. The state has made a $3 million investment in the project, which has an opportunity to revolutionize cloud computing, said Larkin. Software and hardware and service companies will serve as partners in a federated model; they will help design and operate the system, and all stakeholders will have access to the data in the cloud.

The Commonwealth will benefit from using this platform and environment to move data in and out and to arrive at societal benefits from that work, said Larkin. Industry can demonstrate technologies in the cloud in a way they have difficulty doing today; and academia will gain a unique platform on cloud computing that will increase competitiveness for federal funding and other external research support. “We think the initiative is relatively low risk in the end,” said Larkin. ”Even if the marketplace doesn’t transform in the ways we’re discussing now, the infrastructure we’ve created will allow our research institutions to collaborate in unique ways.”

John Goodhue then described the Massachusetts Green High Performance Computing Center, a collaboration of five research institutions—Harvard, MIT, Boston University, Northeastern, and the University of Massachusetts. The collaboration evolved from a conversation in 2008 among Governor Deval Patrick, university leaders, and the CEOs of Cisco and EMC, who decided to try to increase dialog among the state government and universities and high-tech companies. That dialog evolved into recognition that computing—both simulation and the emergence of extraordinarily large data sets—was fundamentally altering research.

This university-industry-government research collaboration has been a massive trust-building exercise, and one that has worked out extraordinarily well, said Goodhue. The collaboration started with a $90M investment in a data center that has 10 megawatts of capacity and 40 megawatts available for expansion. This was followed by a 2-year, $2M seed fund for research projects that are too nascent for federal funding. Every project funded by the program involved faculty from at least two universities, a requirement that has seeded new friendships and collaborations. The initiative sponsored about 15 projects and many of them are moving into the grant funding cycle.

Planning Committee for The Big Data Revolution: What Does it Mean For Research?: John Mason (Chair), Auburn University; **Tilak Agerwala**, IBM Thomas J. Watson Research Center; **Jack Wenstrand**, Agilent Technologies.

Staff: Susan Sauer Sloan, Director, GUIRR; **Kristina Thorsell**, Associate Program Officer; **Laurena Mostella**, Administrative Assistant; **Claudette Baylor-Fleming**, Administrative Coordinator; **Cynthia Getner**, Financial Associate.

DISCLAIMER: This meeting summary has been prepared by Sara Frueh as a factual summary of what occurred at the meeting. The committee's role was limited to planning the meeting. The statements made are those of the author or individual meeting participants and do not necessarily represent the views of all meeting participants, the planning committee, GUIRR, or the National Academies.

The summary was reviewed in draft form by David Fenske, Drexel University; Sallie Keller, Virginia Tech; and Paul Zimmerman, Intel Corporation to ensure that it meets institutional standards for quality and objectivity. The review comments and draft manuscript remain confidential to protect the integrity of the process.

About the Government-University-Industry Research Roundtable (GUIRR)

GUIRR's formal mission is to convene senior-most representatives from government, universities, and industry to define and explore critical issues related to the national and global science and technology agenda that are of shared interest; to frame the next critical question stemming from current debate and analysis; and to incubate activities of on-going value to the stakeholders. The forum is designed to facilitate candid dialogue among participants, foster self-implementing activities, and, where appropriate, carry awareness of consequences to the wider public.

Government | University | Industry
RESEARCH ROUNDTABLE

For more information about GUIRR visit our web site at http://www.nas.edu/guirr
500 Fifth Street, N.W., Washington, D.C. 20001
guirr@nas.edu