NITRD BDSI-2015 Workshop Report

Spearheading Innovation in the Face of Massive Data

Primary Authors:
Lisa Singh
David Logston
Susan Nusser
Howard Wactlar

Additional Contributers:
Randy Bryant
Steve Dennis
Erica Rosemond
Ashit Talukder

1 Table of Contents

Executive Summary

The growth in scale, diversity, and complexity of data has increased the demand for understanding large amounts of heterogeneous data. This presents new challenges to the way data and information are managed, processed, and used. A need exists to (1) rethink how we design systems to handle massive data volumes that are disparate and spread across thousands of nodes on the Internet, (2) develop methods that rapidly curate and process available data to create actionable knowledge, (3) tackle privacy and ethics issues that arise when using this massive data for analytics applications, and (4) determine curricula for teaching big data analytics and data science that integrates computer science and statistics into a richer, interdisciplinary framework.

These different challenges and needs exist across different federal agencies. Given the large number of overlapping data challenges these agencies are facing, a National Big Data R&D Initiative was launched in March 2012. A steering group (the BDSSG) is drafting a framework and establishing a set of priorities for a National Big Data R&D Strategic Plan. One part of developing a strategic plan is to collect input from those in academia and industry that are working on big data research and new technologies. To further inform the development of an effective National Big Data R&D Strategic Plan, in January 2015, Georgetown University hosted a one-day workshop that brought together a small interdisciplinary group of academic and industry leaders who discussed the issues surrounding big data, needs that researchers, developers, and practitioners in the field have, the role government can play to help address some of the challenges, and viable areas for spearheading innovation and propelling progress.

This report discusses ideas and themes that surfaced during the workshop. Overarching challenges and concerns included: the widening gap in data science and big data education between new graduates and the current workforce; resource constraints for the small-to-mid sized companies and universities in terms of hardware, system support, and data access; privacy and ethical issues associated with using socially annotated or 'behavioral' data in studies and products; and the lack of principled methods and tools that produce actionable knowledge from available data while considering potential uncertainty, bias, and trust issues. To address these challenges, the workshop attendees proposed a number of initiatives that federal agencies and the community as a whole could pursue. The three that resonated throughout the day were the following. First, the development of a shared data hub that maintains data sets and provides an infrastructure that enables researchers to develop, test, evaluate, and share methods, solutions, challenges, and algorithms, while maintaining security and privacy. Second, support for societal grand challenges that require interdisciplinary teams to work together to make real, measurable progress on the problems. Finally, the need for curricula (at all levels including current workforce training) that teaches the core concepts from computer science and statistics needed for working in this arena, as well as the interdisciplinary concepts needed to analyze large data in different domains.

1 Workshop Objectives

In January, Georgetown University hosted a one-day NITRD sponsored workshop – the Big Data Strategic Initiative Workshop (BDSI-2015).[1] The primary objective of the workshop was to bring together a small number of academics and industry leaders across disciplines to further inform the development of an effective Federal Big Data Research Agenda. More specifically, participants were asked to consider the following types of issues:

- Identify research and development areas that are fundamental in making progress toward understanding complex, big data;
- Specify current and future areas where major breakthroughs appear possible;
- Identify needed collaborations (e.g., inter-disciplinary, academic-industry);
- Identify research initiatives, priorities, and facilities needed to meet current and future challenges;
- Consider privacy and ethical challenges that exist when using data about people and their behaviors;
- Discuss educational directions, needs and challenges for integrating big data into the curricula of different disciplines.

2 Workshop Organizers

The Principal Investigator (PI) for this workshop was Lisa Singh, Georgetown University. Early in the process, she enlisted a three-person steering committee that had expertise in areas fundamental to big data. The steering committee members were David Logsdon (TechAmerica), Sarah Nusser (Iowa State University), and Howard Wactlar (Carnegie Mellon University). This committee helped identify participants across many disciplines, and provided suggestions for speakers. They advised the PI when different issues arose and were active participants during the workshop, serving as breakout session leaders. Most importantly, they helped draft this report. We also enlisted additional comments from the Breakout Session co-leaders: Ashit Talukder (NIST), Erica Rosemond (NIH), Randy Bryant (OSPT), and Steve Dennis (DHS).

Finally, in order to understand the needs of the different NITRD agencies, the PI worked closely with members of the NITRD Big Data Senior Steering Group (BDSSG), receiving advice and guidance during the entire planning process. She also received significant logistical help from Wendy Wigen. It would have been very difficult to organize this workshop in such a short time without all this help. The PI would like to give a special thanks to Wendy Wigen, Chaitan Baru, Fen Zhao, Renata Rawlings-Goss, Sylvia Spengler, and Suzanne Iacono for their support and enthusiasm for the event.

[1] *This workshop was supported by the National Science Foundation, grant - IIS-1522745.*

3 Workshop Overview, Structure & Participation

The workshop took place on Friday, January 23rd, 2015 at Georgetown University. There was also a working reception the evening of January 22nd, 2015 at Rigg's Library at Georgetown University. The evening reception included talks from Bob Groves (Provost, Georgetown University), Tom Kalil (Deputy Director for Policy, White House Office of Science and Technology Policy), and Allen Dearry, (Director of the Office of Scientific Information Management, NIEHS). A central part of the evening was table discussions. Tables were given questions to discuss and debate. Because participants came from a wide range of disciplines, having them begin thinking about the core themes of the workshop in a relaxed atmosphere helped drive conversation the next day.

The workshop itself included a keynote by Andrew Moore (Carnegie Mellon University), an interdisciplinary panel discussing Future Directions for Big Data Research, Development & Education, and break out sessions on the following four predetermined topics:

- Next Generation Capabilities & Infrastructure
- Data to Knowledge to Action - Using Big Data for Trustworthy Decisions & Confident Action
- Privacy, Ethics, Security & Access
- Education: Workforce Development & Training

Breakout sessions were co-led by a member of the steering committee and a federal agency program director. This was done to facilitate discussion and generate debate by looking at issues from multiple perspectives. Each breakout session topic was conducted twice, with different participants in each of the two instances of a session. This allowed attendees to participate in two different areas of discussion. The break out sessions produced slides that were presented at the end of the workshop day. Those slides and the scribe notes were used to develop the ideas presented in this report.

A website with the workshop details, including the agenda, photos, and a webcast of many parts of the workshop, can be found at http://workshops.cs.georgetown.edu/BDSI-2015/.

This event brought together approximately 45 participants across academia, industry (large companies and startups), not-for-profits, national labs, and even data meet-up groups. The group of participants also spanned different disciplines, including statistics, law, medicine, public policy, and different areas of computer science. They were selected because they represent thought leaders in their areas and they interact with big data in different ways. There were also an additional 45 observers, mostly from the federal government, who attended different parts of the workshop.

4 The Current Situation – Data Today

Data is accumulating at a rate that far outpaces our ability to curate it. Between email, online shopping, tweets, social media posts, YouTube videos, smart phones, and even appliance sensors, every sector of the economy is sitting in a data swamp. Processing and making sense of this data has become a large challenge. According to TechCrunch, Facebook processed over 2.5 billion pieces of content per day in 2012 (Constine, 2012). According to both Google and Ebay, they currently process over 100 petabytes of data per day (Huss & Westerberg, 2014). While no one really knows how much data there is, IBM estimates that 2.5 exabytes were generated every day in 2012 and that 75% of that generated content was unstructured (Wall, 2014). A recent Forrester report indicates that the average company stores 125 Terabytes of data (Hopkins et al., 2015).

Given this situation, there is an increased demand for innovative ways to store, access, and extract knowledge from these scattered, heterogeneous, unstructured data in an ethical, privacy-preserving manner. Analyzing these large volumes of data requires access to large computing clusters and infrastructures that could be costly to administer and maintain. New constraints also need to be considered, including security, privacy, sustainability, energy footprints, and bandwidth constraints to name a few.

5 Next Generation Capabilities & Infrastructure

5.1 Current Landscape

Many new advances in computing hardware have occurred over the last two decades, including multi-core chips, graphics processing units (GPUs), and flash memory/solid state drives (SSDs). These mainstream technologies are faster and use less energy than their predecessors. However, they have required fundamental shifts in software design at the operating system level to the application level. These shifts have propelled new programming paradigms and abstractions for large, networked multi-core systems used in data centers.

Cloud computing has become a mainstream computation infrastructure, leading to much research both in industry and in academia. Leading companies that provide cloud services include Amazon, Microsoft, Apple and IBM. Its prevalence results from many factors, including the use of cheap computers, increases in bandwidth, and the development of programming frameworks like Hadoop and programming models like MapReduce. Programming frameworks like Hadoop allow for highly distributed application development that processes large amounts of data.

5.2 Challenges & Needs

In this section we discuss the different challenges associated with next generation capabilities and infrastructure. We group our challenges into five main areas: data provenance, real time analytics, resource sharing as a service, standardization and reproducibility, and data fusion.

5.2.1 Data provenance

Because data can be easily gathered and transformed on the web, it is not always obvious where the data originated from or what its original form was. Broadly, we can think of data provenance as the history of a piece of data – the origin of the data, its movement across different databases and websites, and its transformations from one state to the next. Understanding data provenance is an issue in many domains, including scientific databases, digital libraries (particularly in the context of citations that move), medical databases, and analytic databases generated from social media data.

Recent work in the area has included workflow provenances that store data specifications and executions. They are searchable and queriable – allowing for finding and reusing workflows, understanding the meaning of workflows, debugging and correcting bad specifications, and improving the understanding of the downstream effect of bad data. This area is in its infancy. Other areas that need attention include: searching provenance executions and data while maintaining different levels of privacy, and understanding provenance in open-source, personal data.

More generally, we must begin building robust evaluation frameworks that consider data provenance and search queries and are well suited for the challenges of the next-generation query processing and search algorithms and applications. There is a growing need for searching, managing, processing, and analyzing large complex information networks that include structured data, text, multimedia, sensory data, biometric data, and other data types in a way that maintains data provenance. Processing a query generically using any combination of underlying, heterogeneous data sources is an open problem, particularly when considering data provenance and data reliability.

5.2.2 Real time analytics

In order to generate timely, accurate real time analytics, systems must be able to rapidly process massive amounts of disparate data and present the resulting analytics in an understandable format. This means that the next generation of computing structures for real time analytics must be flexible, allowing for diverse, multi-granular data to be accessed and combined. Currently, systems that handle complex data with high throughput are in their infancy. Options such as collaborative grids, hypercubes, hierarchical cluster, and other types of distributed and parallel environments need to be advanced. There is also a great deal of data that cannot be accessed because of its format or location. For example, there is a large amount of spatial information in the deep web, but it is not accessible because the techniques for searching have not been developed yet. There are still issues in data collection and delivery related to space, time, accuracy, and source. Developing benchmarks for query and analytics efficiency of heterogeneous data is necessary. Other real time analytic areas that are still in their infancy include approaches for effective spatial crowd sourcing, personalization based on user/social profile for spatial/temporal interpretation, and understanding spatial/temporal trajectories and clustering during events, especially catastrophic events such as natural disasters.

5.2.3 Resource sharing as a service

Given the scale of data that is generated and needs to be processed, discussion is emerging about how those without a large infrastructure can participate in the big data wave. Resource sharing is always a high priority. Universities, not for profits and small businesses are in need of large, secure, scalable platforms to conduct research in the area of big data systems, to develop algorithms for indexing, searching, and analyzing big data, and to advance interdisciplinary research that requires real time analytics on potentially massive data sets.

One approach to resource sharing is to consider developing it as a service. This service would allow for interoperability between different types of shared resources. Similar to web services, there would be a standardized way to integrate resources using open standards. The advantage of a path toward resource sharing as a service is two-fold. First, it provides the needed access to resources in a customizable way. Second, it standardizes the communication between different heterogeneous systems.

5.2.4 Standardization and reproducibility

A perpetual problem in system development is standardization. Currently, standards for systems and analytics related to big data are in their infancy. They are necessary for compatibility between different types of devices and platforms. Standards are beginning to emerge for applications related to the Internet of Things, but with the diverse set of embedded devices, chips, appliances, and physical systems arising, more robust standards need to be developed. Also, given the new and emerging world involving big data, existing standards for operating systems, data formats, large-scale distributed and parallel systems, analytics, and user interfaces need to be revisited.

5.2.5 Data fusion environments

A demand exists for context aware applications and services. These applications require a shift from sensor fusion to more general, heterogeneous data fusion. In this context aware environment, device data, user activity data, local-based data, and current conditions need to be merged to generate meaningful context-aware knowledge, e.g. you need to walk to the meeting because there is too much traffic for a taxi to make it on time.

There are a number of challenges related to developing context aware applications, including (1) no open framework for fusing different data sources, with different types of data in different formats; (2) no solutions for accommodating data arriving at different rates; (3) and no methods for compensating for incomplete data. Intelligent data fusion is a key to smarter applications. Technologies and standards will be a challenge to develop in this arena since both hardware and software developers will need to collaborate for more rapid advancement. However, this type of collaboration is necessary for the next generation smart applications that fuse real time, heterogeneous data.

5.3 Broader Impact Opportunities

Given the mentioned needs and challenges with regards to next generation capabilities and infrastructure, this subsection identifies three opportunities that can have a large impact on improving the current landscape and spearheading innovation in these areas.

5.3.1 National Data Resources

Developing national data resources that allowed for inexpensive uploading, storing, and analyzing of data in a collaborative way would advance both systems and analytics research. These resource centers could each support a different thrust, e.g. science, society, health, and security. They could tackle data privacy and ethical standards related to data sharing and usage. Finally, this resource could provide access to different technologies and platforms.

5.3.2 Grand Challenge Problems

Investing in grand challenge problems that bring together interdisciplinary teams and teams across different parts of computer science will help address large-scale challenges we face with Big Data. One way to do this is to develop a grand challenges program as part of a data 'Center of Excellence'.

5.3.3 A Service-Oriented Model for Shared Resources

In order to facilitate sharing of data sets, analytic capabilities, expertise, and large-scale infrastructure, a service-oriented model can be developed. This will allow for straightforward comparison between different algorithms and computing infrastructures.

6 Data to Knowledge to Action

6.1 Current Landscape

Smart phones, sensors, satellites, social media, etc. – data are gathered from many different devices and applications. It has become commonplace for much of these data to be noisy, incomplete, heterogeneous and multi-modal, continuously changing, and massive. Regardless of their quality, these data are often being used to answer questions that differ from the original purpose of the data collection. This sometimes leads to analytic results that are biased or have high levels of uncertainty, hindering the decision making process and in some cases, leading to decisions and actions based on inadequately (or at a minimum incomplete) evidence.

Furthermore, the inter-relationships, gaps and challenges in a complex data science workflow, from data acquisition to analytics to visualization and action are not well understood. Fundamentally, foundational improvements in data-to-knowledge-to-action need to be achieved and public awareness of data science needs to be improved.

6.2 Challenges & Needs

In this section we discuss the different challenges associated with data-to-knowledge-to-action. We group our challenges into five main areas: scalability of sophisticated predictive analytics for

working with massive data, understanding and improving data quality, reasoning from data to improve knowledge acquisition, integrative data and visual analytics that support effective knowledge to action, data and knowledge access, and data actionability.

6.2.1 Scalability of sophisticated predictive analytics for working with massive data

Machine learning, statistical relational learning, and statistics continue to develop sophisticated methods and algorithms that support predictive analytics. Unfortunately, only a small percentage of these methods scale to massive data scale. A need exists to build toolkits and environments for scaling these algorithms. While packages like Hadoop-ML are emerging, scalable machine learning methods are still in their infancy. Algorithms in this space also do not take advantage of their domain context. Doing so can improve decisions about the use of different scalable data structures, indexing schemes, and compression techniques.

6.2.2 Reasoning from data to improve knowledge assimilation

In order to effectively reason about data and improve knowledge assimilation, we must first understand and assess the quality of the data associated with the analysis and then improve the methods being used to reason about data.

While everyone complains about the quality of big data (the data are noisy, missing, biased, inconsistent, etc.), very little work has gone into understanding the problems and developing methods for improving the quality of open source big data. The database community has developed a large suite of tools and methods for data cleaning over the past two decades. Those need to be revisited in the context of open source data and augmented to handle less structured data, including image data (assessing image quality), sensor data and social media data. One way to accomplish this is to incorporate metadata that describes the data source, the purpose of the data, and the assessed quality of the different data fields.

In order for the field of data science or big data analytics to gain traction, foundational research into the theoretical underpinning of different methods and their limitations needs to be undertaken. The field needs to develop measures for efficiency, uncertainty, and quality and understand the trade-offs. We need to push methods beyond correlation and better understand causality when data is incomplete and biased. We need to develop models, including semantics representations that incorporate domain knowledge into the analysis process. A need also exists to develop toolkits that capture all the phases of the data science life cycle.

6.2.3 Integrative data and visual analytics that support effective knowledge to action

Researchers have shown that visual information is retained at a higher rate than textual information. After multiple days, users retain only 10-20 percent of written or spoken information. However, they retain approximately 65% of visual information (Dale, 1969). Even though this has been known for decades, the ability to rapidly create visualizations and visual stories has been limited. Given the ease with which they can be created today and the benefits of visual learning, more emphasis needs to be given to incorporating visualization and usability into data analytics tools. This also implies that foundational improvements to visualization tools and

visualization capabilities are necessary. More research needs to focus on usability and human factors that can be improved to support effective knowledge to action. Finally, given the myriad of platforms that we use today, e.g. mobile, workstations, large display environments, immersive environments, we need to develop evolving visual solutions that can effectively adapt to these different environments.

6.2.4 Data and knowledge access

In order to facilitate innovation, a need exists to develop a big data ecosystem that contains access to large data sets, tools, computing infrastructure, and benchmarks. Doing this will not only accelerate innovation, but it will also level the playing field, allowing researchers at smaller institutions and small companies to work on problems in this arena. One can imagine this as a set of interconnected hubs, each specializing in some domain or in some part of the data science process. Some of these hubs could focus on other aspects of data science, including legal frameworks for sharing and data usage, provenance of data ownership, methods for maintaining privacy and/or security, the ethics of data sharing, training modules related to these different ideas, etc.

6.2.5 Data actionability

While research is emerging related to methods and algorithms for handling and analyzing the large volumes of data we have, we do not have adequate methods and tools for understanding their reliability or how 'actionable' they are. More tools need to be developed that take results, make them more accessible to the average decision maker, and clearly identify their limitations. Current tools do not have simple interfaces – they are complicated and do not adequately allow decision makers to change conditions and conduct 'what if' analyzes that allow for adjustment of evidence/domain knowledge if the evidence is insufficient.

6.3 Broader Impact Opportunities

Given the mentioned needs and challenges with regards to data to knowledge to action, this subsection identifies five opportunities that can have a large impact on improving the current landscape and spearheading innovation in these areas.

6.3.1 Development of Collaborative Environments

In Section 5, we suggest creating a national data resource as one of the broader impact opportunities. Here, we extend this notion to the development of collaborative environments. We have seen phase one with collaborative document tools (e.g. Google Docs), collaborative coding (e.g. CodeBunk), collaborative workspaces (e.g. Huddle, InMotion, etc.). However, we need more integrated collaborative environments that allow for more comprehensive collaboration and social interaction.

6.3.2 Developing statistical and analytical tools for different types of data structures

Most current statistical and analytics tools are designed for traditional survey data or relational data. Given the abundance of non-traditional design data, we need to develop new tools that can be used to analyze these more organically generated data. Examples of non-traditional data types

include, text, image data, streaming video data, etc. To accomplish this effectively, statisticians and computer scientists need to collaborate and identify the uses and limitations of different methods.

6.3.3 GitHub for Analytics

GitHub is generally used to host open-source coding projects and maintain version control on them. Having a similar hub for analytics projects would benefit the community immensely. The open source paradigm is ideal. However, to initially encourage participation, some projects could include monetization to allow developers to recoup the cost of sharing their analytics ideas.

6.3.4 Startup ecosystem for big data systems

The cost to engage in big data system development is high for small companies. Having programs that help develop an ecosystem for collaborative development in this arena is a large opportunity.

6.3.5 Datapalooza

The idea behind a datapalooza is to encourage data holders, researchers, and hackers to come together and develop foundational methodologies for analyzing different types of data. Together they could develop foundational methodologies for analysis of a particular domain of data, create benchmarks that are analytic specific, and identify actionable tasks for different types of data.

7 Privacy, Ethics, Security, & Access

7.1 Current Landscape

The amount of sensitive information that people share continues to increase. Cloud computing is now a major force in industry, but data security in this new environment is still in its infancy. Data privacy continues to be a challenge. Currently, there is no single federal law that adequately regulates the collection and use of personal data (Jolly, 2015). While guidelines and best practices exist, privacy laws are inconsistent across states and dated at the federal level when it comes to limiting use and sharing of personal, behavioral data. Therefore, securing data and protecting the privacy of consumer data are large concerns.

While questions are arising about the ethics of using personal data to understand behavioral tendencies (see Facebook contagion experiment (Kramer etal., 2014)), there is still a cultural acceptance of open data use by companies. The public has been trained that once the data is given to a company, the company can use it for purposes beyond the original intended use. This feeling is especially prevalent among middle school, high school and college students using social media sites.

Within computer science, researchers have been investigating ways to protect user data privacy, while still maintaining the utility of the data for statistical analysis and data mining. Recent developments include differential privacy for statistical databases, anonymization techniques for

relational data, and prototypes of user controlled identity management systems. While progress is being made, these methods are largely academic and are not being integrated into real world systems.

7.2 Challenges & Needs

In this section we discuss the different challenges associated with data privacy, ethics of data usage, and data security. We group our challenges into four main areas: data sharing & ownership, data ethics, reactive security mindset, and public awareness.

7.2.1 Data Sharing & Ownership

Every day users share data with companies. They sign off on complicated terms of service agreements that generally allow these companies to use the personal data for purposes other than the original use or share them with other companies. There are a number of challenges and concerns that arise from the current data sharing environment including the following:

- Generally, consent for data usage is given once. However, the data a company collects may change over time. How can we enforce simple, adjustable consent policies that are clearly stated and understandable to consumers?
- Who owns personal data that is shared with companies? What 'inalienable' rights should individuals have with regards to personal data? What rights should companies have?
- Because much data are relational, not individual, data may be shared about an individual without his/her knowledge. How do we keep people informed about data that others are sharing about them?
- From the base data that is shared by users, companies may generate new data that is used for target marketing, recommendations, and more generally predictive analytics. Who owns these newly generated data? How can users access these data and possibly change them if they are inaccurate?

7.2.2 Data Ethics

While segments of the population utilize privacy features offered by social media sites, many Internet users do not. Personal demographic information, as well as ideas and thoughts (tweets or messages) which once would have been shared in a more private setting with groups of friends/acquaintances are now accessible to anyone with a computer. Because these data are public or accessible to a company, many believe that they can be used for whatever purpose researchers and companies decide. However, the ethics of doing this needs to be considered. Currently, there are no universal data ethics standards that companies abide by. As a consequence, there is great variation in these standards depending on the values of the community or company involved. In many cases, computer scientists do not consider ethical implications of using data to understand behavioral trends or security risks. Getting computer scientists to care about data ethics and developing standards which different communities with different norms will adhere to are both large challenges.

7.2.3 Public Awareness

In June 2014, there were more than 2.9 billion Internet users (Internet Live Stat, 2014), 1.5 billion of whom share information on Facebook, 343 million on Google+, 238 million on LinkedIn, and 200 million on Twitter (Social Media, 2013).

While it is likely that many users of these online services understand that they are sharing personal information with strangers, they may not understand the potential risks and implications of doing so. How do we help users understand the dangers of sharing so much private information? What are the best ways to inform children about the dangers and potential consequences of sharing information? What types of tools would be most beneficial?

7.2.4 Reactive Security Mindset

Personal health records, financial data, behavioral data, and consumption data – all these different types of personal information need to be maintained in secure systems. While many data security methods exist, e.g. authorized access, authentication, encryption, disaster recovery, large-scale data breaches continue to occur, e.g. Target and Home Depot credit card breach. Part of the issue is the reactive nature of security. Given the sensitivity of the data that companies, governments, and international organizations are maintaining (and in some cases sharing), a large challenge is finding ways to propel a new generation of secure systems that proactively identify system weaknesses as opposed to reactively controlling damage when a breach occurs. This suggests the need for systems that have the ability to adapt/enhance their security as their computation environment changes.

7.3 Broader Impact Opportunities

Given the mentioned needs and challenges with regards to privacy, ethics and security, this subsection focuses on identifying five opportunities that can have large impact on improving the current landscape and spearheading innovation in these areas.

7.3.1 Educating consumers, researchers, and companies about data privacy and data ethics

Companies and researchers see personal data as a gold mine for understanding human behavior. Many times they start using it without thinking about user privacy or the ethics of using certain personal data. There are a number of ways to improve this situation, including:

- Develop online modules that teach people (including minors) about the importance of keeping data private, the implications of sharing data, and proper uses of data, i.e. data analyzes that should not be conducted even if data are available.
- Provide companies and the community with simple consent forms that are understandable. Place them in one or more ‘privacy’ hubs for comment, debate, and use. This allows companies to see high quality forms and adapt them for their needs. It also establishes a forum for debate and feedback about them.

7.3.2 Data ownership and data-use models

What does data ownership mean? We need a clear definition and legal framework that can be used to identify acceptable boundaries of privacy. We should also enforce total privacy and security minimums for certain classes of private data. There are existing frameworks that can be leveraged for data privacy, e.g. HIPAA. Social scientists have been thinking about privacy in other contexts for years and we should leverage what they have already developed.

Another broad impact opportunity is to support the development of software that allows users to clearly see what companies are using their data for and how different companies are using them. One possible implementation is to have a personal data graph that is annotated with information about companies that are using the data and how it is being used. This graph could also be used to give and remove consent for data usage. Having such a framework would allow users to not only see who has their data and how they are being used, but also how the data are mutating and being aggregated. It gives users control of their data. It gives them the ability to decide on how much they are willing to share. It may also be useful for understanding data usage violations if companies use data without 'registering' usage on this data graph. This is a way to promote community engagement in determining adequate standards for data sharing and aggregation.

7.3.3 The new field of data ethics

Data ethics is emerging as a complex issue that has differences from general ethics. Similar to bioethics, data ethics needs to be its own field. Supporting interdisciplinary workshops on data ethics would help with the establishment of this field.

While other scientists are engaged in different ethics debates, computer scientists are less engaged. Getting them involved in the debate is important. One possible way to 'push' them into thinking about data ethics is to require a data ethics statement with funding submissions. NSF has a data management statement. Something similar could be setup for data ethics. These statements could also be added to publications, etc. Some of the specific questions or statement requirements could leverage existing work done by university Integrity Review Boards.

7.3.4 Proactive secure and trustworthy systems

While secure and trustworthy systems have always been a goal, attaining them is non-trivial. The broadest impact opportunity is to actually develop proactively secure and trustworthy systems. Unfortunately, it is unclear how we get there. Promoting security as a fundamental design principle as important as functionality is a good idea, but hard to implement in current curricula. A complimentary idea is to develop a rating system that rates software on different measures of security. The expectation is that all software should receive a rating and that the default rating is one of secure (anything less should be considered bad software).

7.3.5 Development of data privacy and data ethics information hubs

Researchers in different disciplines are discussing privacy and ethics issues in silos. We need a sustained dialog among academia, industry, and the public sector. We also need a place people can go to share stories and lessons. An opportunity for a broad impact is the creation of a

clearinghouse for (1) data that can be used for data privacy and data ethics challenges; (2) tools to improve and assess privacy; and (3) challenges that different interested communities can work on together.

8 Education for All

8.1 Current Landscape

Big data, data science, data analytics – graduate programs that support this emerging new field are being developed all across the country. Two years ago, there were only a handful of programs. Now there are dozens, and the number continues to grow. Still, there is no consensus on what a program should like and what the core competencies should be. For those already in industry, there are online tutorials and courses that teach different topics that broadly fall into the area of Big Data Systems and Analytics. However, curricula for the current workforce are still in their infancy. There has also been little thought put into when data education should begin or what should be taught before college?

8.2 Challenges & Needs

In this section we discuss the different challenges associated with education in this arena. We first consider different levels of education, beginning with education for our youth. We then consider challenges related to women and minorities in the workforce. Finally, we discuss the need for regional education hubs.

8.2.1 What is a data scientist?

Even though the term "data science" has been around for decades (Press, 2013), it is still unclear what data science is. At the workshop, many participants that represented federal agencies were also unclear about the required skill sets and role of a data scientist. While it is unlikely that we can agree on a definition in this report, the components that the workshop attendees felt should be part of this new field included:

- Date science needs to develop a theory and principles of working with different, potentially conflicting, incomplete, dynamic, heterogeneous data.
- Data science applications should involve creating new data to tell a story about the existing data. Communication and story telling are important components of this field.
- Data science needs to confront issues of bias, sample design and noise head on. New data sources require new theories about these ideas.
- Gathering, cleaning, exploring, analyzing and visualizing data all need to be part of the data science process.
- Data science research projects must result in capturing insight into the data's impact as it relates to other disciplines.
- Using data is not data science.

In some sense, data science can be viewed as a discipline that applies scientific methodology to understand data across disciplines, data related to religion, economics, medicine, history, physics, etc. The core competencies of data scientists are still unclear and need to be fleshed out to truly understand what a comprehensive curriculum would look like.

8.2.2 Traditional education at all levels

Our current education system teaches basic statistics, but it focuses on well-defined, well sampled data sets. It does not consider topics related to data analysis of data that students interact with regularly. In order to train the next generation of students in data-related disciplines stemming from this new world of available data, a need exists to start educating our youth about the data analytic process. While the basics of computer science and statistics exists in high school curricula, little thought has gone into the development of a data science or data analytics curriculum at the high school level. Could it be included within existing math and science courses or does it need to be introduced as a separate class?

At the undergraduate level, it is unclear what an interdisciplinary data science major or minor should include. A need exists to have those in computer science, statistics, social sciences, and humanities come together to ensure that the fundamental concepts of the area and the potential applications of the concepts are integrated into a single curriculum.

If we look at the traditional curricula that undergraduate students take, it is clear that pieces of data science are taught in different departments across universities. However, important concepts and methodologies are being ignored when students do not learn how to think about and evaluate the entire life cycle from data collection to data analysis. Exposing them to the entire lifecycle through project-based courses is an important way to enhance their traditional education and improve their quantitative skills.

Core concepts that are important for any curriculum in this field include: the data science lifecycle, data preparation, data cleaning, data management, exploratory analytic methods, predictive analytic methods, scalable data modeling, domain knowledge (familiarity with problems in different fields), data presentation/visualization, and data ethics. If we think of traditional curricula, this list represents concepts taught in computer science and statistics, but it also highlights the need for domain expertise in the domain of interest. In other words, knowledge from humanities, social science, and other sciences is important for successful studies in this field.

Finally, graduate level education is the current focus of this discipline. Unfortunately, there is no consistent, coherent curriculum. What are the skills and competencies of someone that has a graduate degree in data science or data analytics? What should the core topics include? Do they differ from the undergraduate ones? Are there other degrees that need to better understand analytics to be successful? For example, given today's trends in marketing, does it make sense to have a marketing degree without fundamental knowledge of analytics? Should the industry the

student plans to enter inform the content of the curriculum? Do we create specialists or general practitioners?

A fundamental problem for universities who want to have data science and big data analytics programs at the undergraduate and graduate levels is the need for a large-scale computational infrastructure for educating students with real world data. Universities cannot afford Google-size infrastructures, but still have a need to educate students and enable researchers to advance in this arena. To further exacerbate the problem, even if the infrastructure could be developed, companies hold the data. We need to define and move to a model where resources and data are shared with academia.

8.2.3 Educating (re-educating) the workforce

Today's current work force is ill-prepared to deal with the surge of available data, from data collection (determining which data are representative) to analysis. We need to invest in re-educating the workforce. Traditional education tends to be difficult for those working. Therefore, we need to consider more innovative approaches including:

- Creating interdisciplinary teams that mirror industry (both in academia and government) and give credit (grants) for not only research, but education across fields.
- Incentivize learning of different skills using salary bonuses.
- Encourage educators to break the educational topics into small modules, some of which can be used to educate existing professionals and train new data scientists simultaneously.
- Invest in small grants for boot camps in data science.
- Promote training related to communicating results.

8.3 Broader Impact Opportunities

Given the mentioned needs and challenges with regards to education for all, this subsection identifies three opportunities that can have a large impact on improving the current landscape and spearheading innovation in these areas.

8.3.1 Data science literacy

Because of the influx of data everywhere, data science literacy will be a requirement in five to ten years for those in high school to those in industry. In order to propel literacy, many different forms of education should be developed, including: short week long courses, boot camps, immersion courses that retrain mathematicians to use their learned knowledge in industry, embed analytics courses into MBA, medical school, and law school curricula, professional graduate degrees designed for industry, high school data science literacy programs, graduate curricula that incorporates interdisciplinary teams addressing real world challenges, a community college data science track, and online self learning modules. The goal is to create a population that has a general literacy about data science.

8.3.2 Improving diversity in the workforce, pulling in women and minorities

Given the lack of diversity in computer science, beginning this emerging field with workforce diversity in mind is important. To support this, a social network with specific foci on women and minorities can be developed. Developing mentorship programs (or leveraging ones like MentorNet) to create connections between minority and female professionals and high school and college students pursing data science can encourage and foster diversity from the outset.

8.3.3 Regional educational hubs

Open data policies are a good start, but educators are craving access to more data. More data drives the opportunity for the formulation of better case studies. Hubs would bring together federal/state/local governments, private industry, non-profits, and academia to help tackle issues of national and regional importance. The hubs could be fertile grounds for providing students with timely, relevant case studies to work on. It could also be used to drive and standardize curricula.

9 Grand Challenges/Priorities

The previous sections have identified a number of needs, challenges and opportunities. Here we integrate the ideas from the different breakout sections to identify three grand challenges that federal agencies can help with. We feel that these are areas of high impact and therefore, should be considered priorities.

9.1 Data science literacy

Because of the interdisciplinary nature of big data and data science, many in the workforce are exposed to concepts related to big data. However, because of the constant innovations in the area, staying current with both theory and practice is a challenge. Data science literacy involves ideas from statistics, computer science, ethics, and specific domains of interest. Community engagement is also vital for the societal grand challenges. Funding interdisciplinary teams to define the field and develop initial curricula will not only help educate people about data science, it will also help the field advance more rapidly. Data science needs a clear identity and clear, foundational theories to survive and grow as a stand-alone discipline.

9.2 Build data ethics as a field

To date, data ethics has been taken very lightly by industry and academia. No clear standards exist and because it is not a distinct field, limited discourse about data ethics takes place. Support for establishing standards and debating data sharing issues needs to be as high a priority as the development of new data analytics techniques. Tutorials and guidelines need to be developed about data usage and the ethics of different types of data usage and sharing. Dialog and standards can also be forced by limiting funding for researchers and businesses that do not adhere to strict data privacy standards and/or incorporate data ethics training into their data science project.

9.3 Data Science Challenge Problems

Formulation of data science challenge problems with appropriate reference data, metrics, and ground-truth that is applicable across multiple domains will help foster interdisciplinary solutions for hard problems. Working on challenge problems will also bring together academic, government, and industry partnerships and result in quantifying, assessing and improving the complete end-to-end data–to-knowledge-to-action workflow. It will also result in better sharing of tools, infrastructure, data, knowledge and resources. These grand challenge problems can serve as the cornerstone for data science hubs with collaborative environments and tools. Having data science hubs will allow researchers to go to easily find measurable characterization of the current state of art, identify critical gaps and needs, and share data, algorithms, and analytics. Not only does working on grand challenge problems improve societal conditions, it will also result in significant improvements to the field of data science, data management, analytics, visualization, usability and decision-making.

10 Lessons Learned

Hindsight is always 20/20. While the workshop was very successful, there are always different ways to improve the workshop. Below is a list of suggestions for future workshops.

1. The process of organizing such a workshop should ideally span six months to a year. This workshop was organized in less than 2 months. With more time, the organization would have been smoother and less stressful. Further, some speakers we invited may have been available with more lead time.

2. Finding the right balance between discussion time and presentations is a challenge. Increasing the overall duration of the workshop is difficult for participants. To really get more complete, written recommendations at the workshop itself, adding a second day would have been useful.

3. Having a social event is very important for facilitating interaction and building bridges among the researchers. Having the evening reception really helped jump start the workshop.

4. Arranging the hotel and air travel for participants requires some organization, but it does reduce the number of receipts and makes the reimbursement process smoother.

5. Providing participants with a survey at the end of the workshop would give organizers and NSF instant feedback concerning the success of the workshop and things that need improving. This would have been a good thing to do.

6. Getting ideas from participants using sticky notes was a nice way to exchange ideas. In the future, a better way would be to do it online so everyone could see the ideas that are being generated.

References:

Constine, J. (2012). How big is Facebook's data? 2.5 Billion pieces of content and 500+ terabytes ingested every day. *TechCrunch*. August 22, 2012. http://techcrunch.com/2012/08/22/how-big-is-facebooks-data-2-5-billion-pieces-of-content-and-500-terabytes-ingested-every-day/.

Dale, E. (1969). Cone of experience, In *Educational Media: Theory into Practice*. Wiman RV (ed.) Charles Merrill: Columbus, Ohio.

Hopkins, B., Owens, L., Goetz, M., Gualtieri, M., and Keenan, J. (2015). Deliver on big data potential with a hub-and-spoke architecture. *Forrester Report*. February 26, 2015. https://www.forrester.com/Deliver+On+Big+Data+Potential+With+A+HubAndSpoke+Architecture/fulltext/-/E-RES83303.

Huss and Westerberg. (2014). Follow the data – Data size estimates. https://followthedata.wordpress.com/2014/06/24/data-size-estimates/.

Internet Live Stats – Internet Usage & Social Media Statistics. (n.d). Retrieved June 6, 2014, from http://www.internetstats.com.

Jolly, I. (2015). *Data protection in United States: Overview.* Retrieved 2015-11-01 from http://us.practicallaw.com/6-502-0467.

Kramer, A., Guillory, J., & Hancock, J. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Science*, 111(42), 8788–8790.

Press, G. (2013). A very short history of data science. *Forbes Magazine*. May 28, 2013. http://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/.

Social Media - How many people use the top social media, apps and services? Retrieved September 2013 from http: //expandedramblings.com/index.php/ resource-how-many-people-use-the-top- social- media/.

Wall, M. (2014) Big data: Are you ready for blast-off? *BBC News.* March 4, 2014. http://www.bbc.com/news/business-26383058.