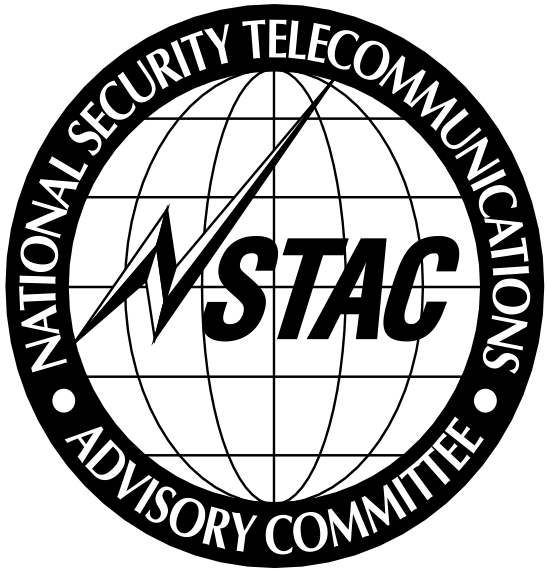
THE PRESIDENT'S
NATIONAL SECURITY TELECOMMUNICATIONS
ADVISORY COMMITTEE



NSTAC Report to the President on Big Data Analytics

May 11, 2016

TABLE OF CONTENTS

EXECUTIVE SUMMARY

While there is no single accepted definition of big data, the National Institute of Standards and Technology (NIST) describes big data as, "the common term used to describe the deluge of data in today's networked, digitized, sensor-laden, and information driven world."

Big data is also known by its characteristics of:

- Volume (i.e., the size of the dataset);
- Variety (i.e., data from multiple repositories, domains, or types);
- Velocity (i.e., rate of flow); and
- Variability (i.e., the change in other characteristics).

Big data analytics (BDA) is a collection of techniques that leverage vast and disparate sets of data to extract knowledge.

In February 2015, the Executive Office of the President (EOP) tasked the National Security Telecommunications Advisory Committee (NSTAC) with identifying specific recommendations regarding how the greater use of BDA could enhance the Government's national security and emergency preparedness (NS/EP) capabilities.

While national security is segmented into military and non-military aspects, this report focuses on the non-military aspects of national security. In response to the EOP's request, the NSTAC also focused on identifying the applications for BDA within the realm of emergency preparedness, which involves a wide array of threat assessment, mitigation, preparedness, response, and recovery activities designed to lessen the impact of natural and man-made disasters.

Through an extensive fact-finding and information gathering process, the NSTAC analyzed how BDA can improve the Nation's response to NS/EP events. In order to better conceptualize how BDA could be widely adopted by the Government, the committee first examined specific areas where BDA has already been successfully deployed and is in use today: (1) healthcare; (2) transportation; and (3) city management. The NSTAC found that certain commonalities existed across each of these domains.

The committee developed three use cases around which to structure this study: (1) a natural disaster; (2) a man-made disaster; and (3) a cyber attack on critical infrastructure. The NSTAC constructed each use case around the steps in the data lifecycle in order to identify specific and key issues related to BDA. The committee then compared results across the use cases and the domains that have successfully deployed BDA to identify issues and highest-priority themes.

Through this study, the NSTAC discovered recurring themes that present a common set of challenges to discovering, sharing, and utilizing data, particularly with regard to how data should be handled and secured in order to maintain its integrity and protect civil liberties.

To address its findings, the NSTAC recommends that the President take the following actions to improve the use of BDA to enhance NS/EP capabilities.

To expand policies, plans, standards, and tools used for BDA that allow data to be utilized more readily during an NS/EP event, with the appropriate protections in place, the President should direct the appropriate Federal Government agencies to:

1. **Collaborate with the private sector to collect, manage, and make available common and adaptable NS/EP ontologies. This includes the use of standard labeling methods, other shareable components, and the development of robust, community-driven standards.**

2. **Study and recommend ways to improve the capacity and robustness of industry-provided services that are necessary for NS/EP capabilities.**

    - Since the private sector is an important partner in responding to NS/EP events, the Government should consider tax incentives for bolstering and building infrastructure for delivering critical services.

3. **Conduct proof-of-concept research and exercises that would be shared across an expanded range of Government agencies to elicit and integrate datasets in anticipation of possible NS/EP events.**

    - The Government should develop and execute exercises that test the efficacy of analytic approaches as new ones emerge and existing ones evolve. At the same time, the Government should develop and sustain technology pilots and proof-of-concept models, as well as share the results across a broader range of Federal agencies. This is designed to facilitate expanded organizational participation and standards development of shareable databases and applications.

    - The Government should incentivize and provide test datasets for consumption in the public sector, private sector, and academic community.

In order to further clarify the availability, access, handling, and protection of data in line with industry's privacy and security best practices to most effectively capture the potential benefits of BDA, the President should direct the appropriate Federal Government agencies to:

4. **Develop a "Good Samaritan" Framework for exchanging information between the Government and consenting private entities during an NS/EP crisis.**

    - The framework should afford standard agreed upon protections to entities sharing data in good faith during an NS/EP event.

    - The development of this framework should be a collaborative effort between the appropriate public sector, private sector, security, and civil liberties stakeholders.

- This framework should pre-establish general rules between the Government and the participating private sector organizations to define the appropriate use of data during an NS/EP event. Specifically, the "Good Samaritan" Framework should clarify rules regarding the protection of privacy, data use, ownership, storage, retention, accidental disclosure, and deletion.

To best minimize and effectively overcome the skills gap that currently exists in BDA, the President should direct the appropriate Federal Government agencies to:

5. **Identify data science and analytics as a key discipline limited by shortages in practitioners, educators, and graduate and undergraduate programs of study within the *Federal Science, Technology, Engineering, and Mathematics (STEM) Education 5-Year Strategic Plan* (2013).**

6. **Direct contracting agencies to add appropriate data science skill and training requirements for all applicable disciplines, as appropriate for Government contracts addressing big data problems.**

    - Data scientists and data ethicists should be identified as key personnel on big data contracts.

    - Contracting firms should develop a training program upon award of the contract and provide notification of this approach to industry to allow sufficient time for training and personnel acquisition.

    - Contracting firms should provide relevant personnel with training covering big data issues of security, privacy, ethics, provenance, and transparency for Government contracts addressing big data problems.

7. **Define and require data assurance plans and programs for Government big data contracts.**

    - When tasking the appropriate agency to define data assurance plans and programs, NIST's Special Publication 800 series of computer security publications could be used as a guide.

    - In order to address the need for expanded data assurance and to ensure that big data is used ethically, the Government should train personnel on the appropriate use of and inherent privacy risks associated with big data.

1.0 INTRODUCTION

The explosive growth of data is a recent phenomenon. In 2010, it was estimated that five exabytes of data, the equivalent of 15,000 times the content in the Library of Congress, was created from the beginning of time up until 2003. By 2010, that same amount of data was created every two days.[1] That estimate was made six years ago when there were two billion people on the Internet. Today, 3.2 billion people, about 40 percent of the world's population, are on the Internet. Additionally, Gartner, Inc., estimates there will be approximately 6.4 billion Internet-connected "things" in 2016, and that those "things" will continue to create exponentially increasing amounts of data.[2] Therefore, it is an inescapable conclusion that society currently finds itself in the age of big data and that big data is, and will continue to be, transformative.

While there is no single accepted definition of big data, the National Institute of Standards and Technology (NIST) describes big data as "the common term used to describe the deluge of data in today's networked, digitized, sensor-laden, and information-driven world."[3]

Big data is also known by its characteristics of:

- Volume (i.e., the size of the dataset);
- Variety (i.e., data from multiple repositories, domains, or types);
- Velocity (i.e., rate of flow); and
- Variability (i.e., the change in other characteristics). [4]

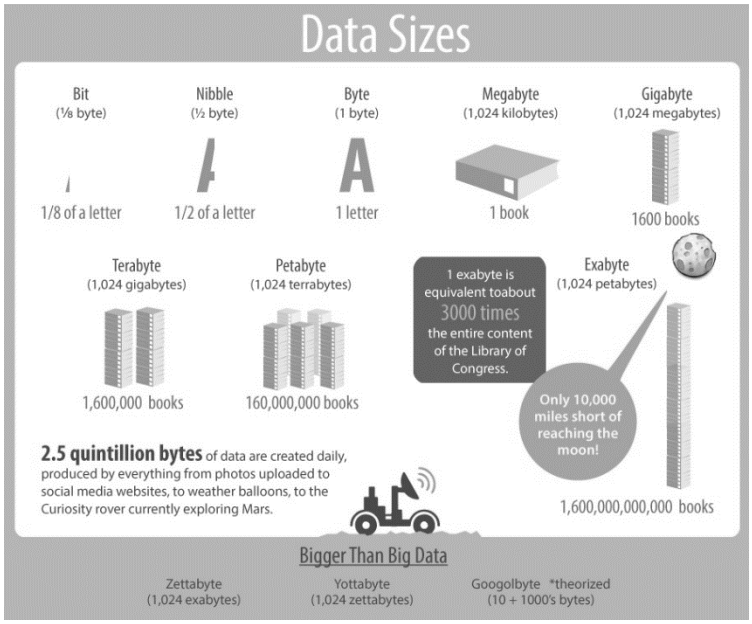


Figure 1.1. Information on data sizes from Adeptia.

The collection of techniques leveraging vast and disparate sets of data to extract hidden knowledge is known as big data analytics (BDA). BDA can be used in a variety of contexts to improve situational awareness and provide accurate predictions. For instance, BDA is used by online retailers to recommend products to consumers, by credit card companies to detect fraud, and by scientists to reveal changes in climate. The use of BDA drives medical advancements in genomics and contributes to the fight against disease. It is also used by local city planners and industry in support of transportation and smart city initiatives.

---

[1] Eric Schmidt. *Remarks at the Techonomy Conference in Lake Tahoe*. August 2010. http://techcrunch.com/2010/08/04/schmidt-data.

[2] Gartner, Inc. "Gartner Says 6.4 Billion Connected 'Things' Will Be in Use in 2016, Up 30 Percent From 2015." November 10, 2015. http://www.gartner.com/newsroom/id/3165317.

[3] National Institute of Standards and Technology (NIST) Big Data Public Working Group Definitions and Taxonomies Subgroup. *NIST Big Data Interoperability Framework, Volume 1: Definitions*. NIST Special Publication (SP) 1500-1. September 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1, page 1.

[4] Ibid, page 4.

1.1 Scoping and Charge

Recognizing the explosive growth of big data, the Executive Office of the President (EOP) requested that the President's National Security Telecommunications Advisory Committee (NSTAC) study how BDA could enhance the Government's national security and emergency preparedness (NS/EP) capabilities. In response, the NSTAC's BDA Scoping Subcommittee was formed in March 2015. The NSTAC concluded its scoping effort and produced its *Big Data Analytics Scoping Report* in August 2015. After completing its scoping phase, the NSTAC then began focusing on developing a report that examines the transformative nature of BDA on public policy activities related to the Government's NS/EP functions. What differentiates BDA from other data-centric paradigms is the need to account for big data's unique character, which is best described as continually changing in volume, variety, velocity, and variability. The implications of this ever evolving landscape have both positive and negative impacts on existing and new policy structures, and may affect the Government's NS/EP capabilities, including those related to communications and information use.

1.2 Approach

Traditionally within Government, some agencies are much more involved in the use and analysis of big data. Agencies such as the National Aeronautics and Space Administration (NASA), the National Oceanic and Atmospheric Administration (NOAA), and others with a strong scientific focus have used large datasets for analysis for many years. Non-profit groups are also beginning to more widely utilize large and big datasets. The NSTAC found it insightful to discuss issues of privacy with appropriate experts, as understanding where data is derived from and how it is to be used must always reflect Americans' basic civil liberties. The committee also found it insightful to speak with other public sector, industry, academic, and non-profit group representatives to learn what common practices are important to successful analytical outcomes and what pitfalls must be avoided.

In the course of performing the research necessary to complete this tasking and develop the resulting recommendations, the NSTAC:

- Received briefings from over 40 subject matter experts (SME) from Government, the private sector, non-profit organizations, civil liberties organizations, and academia;
- Reviewed Government BDA activities and projects, such as those conducted by NASA, NOAA, and the National Science Foundation;
- Studied Governmental and private sector activities regarding standards for BDA;
- Reviewed current industry best practices and capabilities for possible relevance and portability;
- Examined academic literature and current BDA research studies;

- Discussed privacy concerns regarding the collection, use, and manipulation of data with academic and civil liberties experts; and

- Evaluated Government and private sector policy documents and recommendations, including previous NSTAC reports, for applicability.

To better understand the current applications of BDA, the NSTAC first sought to understand the rapidly evolving use of big data throughout society. The NSTAC examined specific non-NS/EP use cases where BDA has been and is currently successfully utilized. The insights derived from those use cases and other briefings and research were then applied around three hypothetical NS/EP use cases, which were chosen in order to best structure the committee's study with regard to NS/EP. These use cases include:

- A natural disaster;

- A man-made disaster; and

- A cyber attack on critical infrastructure.

To provide further structure to its study, the NSTAC reviewed the use cases through the standard paradigm provided by the lifecycle of data. This allowed the committee to examine and glean recommendations from a linear and structured approach. The steps in the lifecycle of data are illustrated in Figure 1.2.[5]

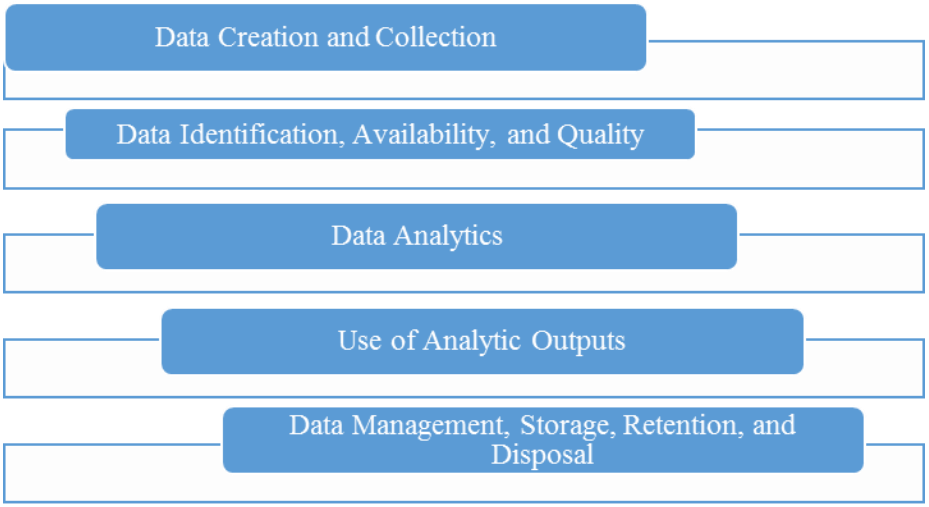


Figure 1.2. Steps in the data lifecycle.

The NSTAC's recommendations are derived from its research into the rapidly evolving topic of BDA and its conclusions regarding how BDA can be used to support the Government's NS/EP mission functions.

[5] Please refer to Appendix D, *Steps in the Data Lifecycle*, for more information.

1.3 Previous NSTAC Reports

The EOP often calls upon the NSTAC to provide thought leadership for emerging, transformational technologies, such as BDA, and to determine how those technologies can support Executive Branch initiatives and the overall public good. Related NSTAC reports include the *NSTAC Report to the President on Cloud Computing* (May 2012) and the *NSTAC Report to the President on the Internet of Things* (November 2014).

There is a symbiotic relationship among BDA capability, inexpensive hardware, open source software, and cloud computing. Cloud computing collocated with storage is able to provide unprecedented efficiencies in managing big data through the implementation of scalable software infrastructure.

In 2012, the NSTAC developed the *NSTAC Report to the President on Cloud Computing*. BDA is closely aligned with the open source software movement and cloud computing. The advent of low cost, open source software has allowed organizations to take advantage of and store high volumes of data. Access to such high-volume data processing and storage technologies has enabled the decoupling of hardware and software in favor of a virtualized cloud, and has led to highly scalable, low cost storage, and the emergence of Infrastructure-as-a-Service (IaaS).

Cloud computing, which NIST defines as, "[a] model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction," can be used to promote BDA solutions.[6] Specifically, businesses and local governments can implement BDA solutions by leveraging Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), and IaaS, which are available for purchase from a growing number of private companies.

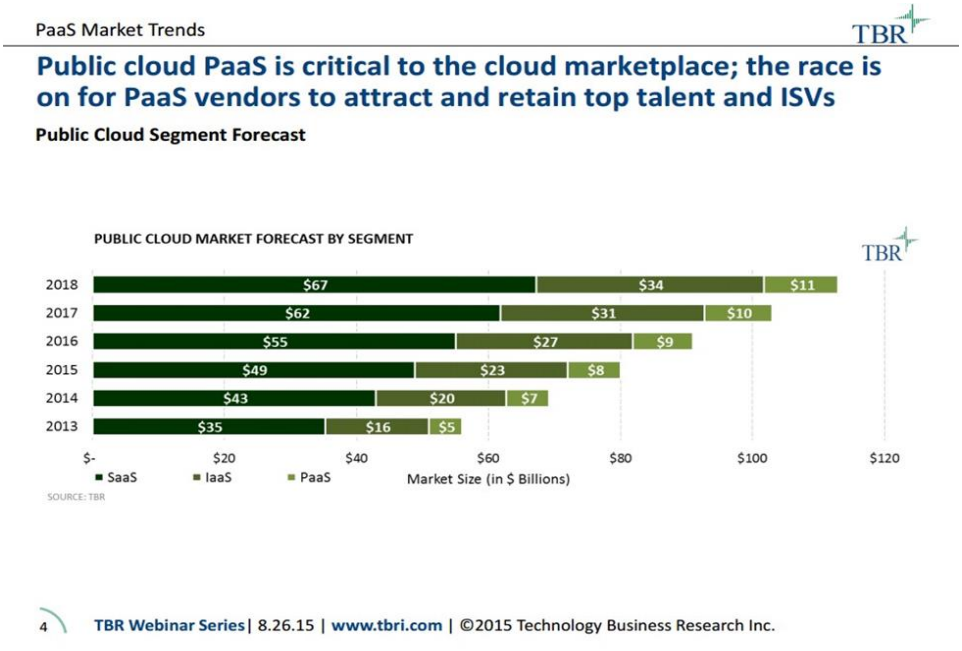


Figure 1.3. Graph on cloud PaaS market trends from Forbes.

[6] President's National Security Telecommunications Advisory Committee (NSTAC). *NSTAC Report to the President on Cloud Computing.* May 15, 2012. https://www.dhs.gov/sites/default/files/publications/2012-05-15-NSTAC-Cloud-Computing.pdf, page 4.

These services allow organizations to develop BDA capabilities without investing in infrastructure or software. All segments of the public cloud are expected to grow in the coming years, which will increase the demand for all of these service models.[7]

The Internet of Things (IoT) is also an emerging and major contributor to the growth of big data. In 2014, the NSTAC developed the *NSTAC Report to the President on the Internet of Things*. The NSTAC report defined the IoT as, "a decentralized network of objects, applications, and services that can sense, log, interpret, communicate, process, and act on a variety of information or control devices in the physical world."[8] The demand for and use of Internet-connected devices is expected to grow in coming years, creating opportunities for BDA techniques to be used to yield greater insights from the large amount of data generated by the IoT.[9]

While estimates of exactly how many IoT devices exist vary, one agreed upon fact is that their introduction and growth in both the industrial and consumer marketplace is fast and steep. In November 2015, a Gartner report predicted that 6.4 billion connected "things" will be in use by 2016, constituting a 30 percent increase from 2015. In addition, Gartner predicted that the IoT will reach 20.8 billion devices by 2020.[10] In July 2015, Juniper Research estimated that there were 13.4 billion devices connected in 2015 and that that figure would rise to 38.5 billion by 2020.[11]

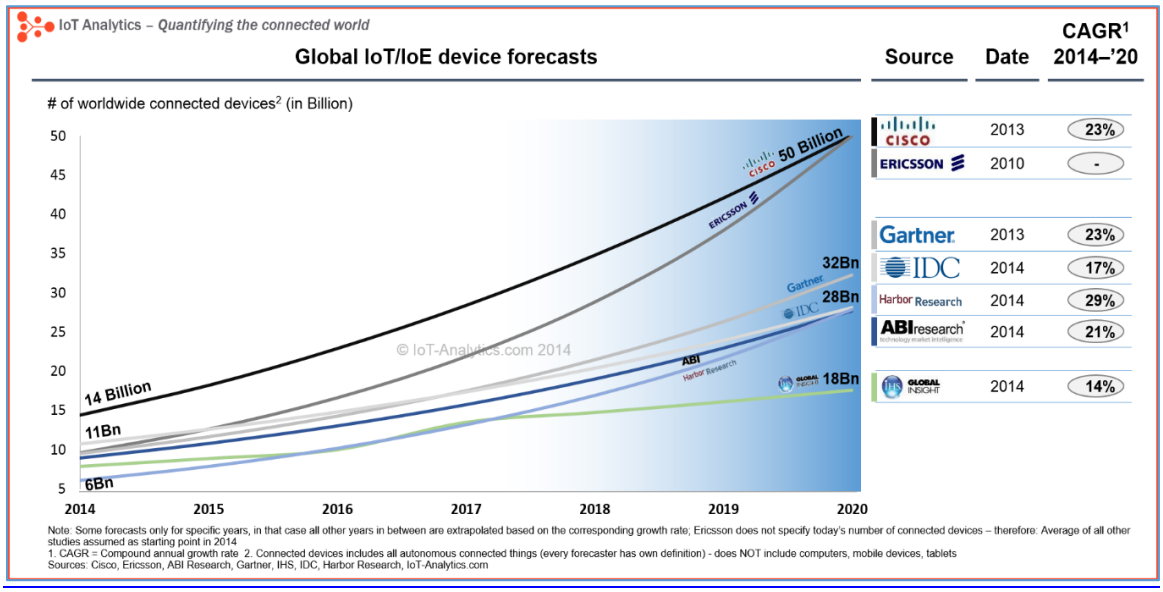


Figure 1.4. Projection of global IoT-device proliferation from Juniper Research.

[7] Louis Columbus. "Roundup of Cloud Computing Forecasts and Market Estimates Q3 Update, 2015." *Forbes Magazine*. September 27, 2015. http://www.forbes.com/sites/louiscolumbus/2015/09/27/roundup-of-cloud-computing-forecasts-and-market-estimates-q3-update-2015/#7032b29a6c7a.

[8] NSTAC. *NSTAC Report to the President on the Internet of Things*. November 19, 2014. https://www.dhs.gov/sites/default/files/publications/NSTAC%20Report%20to%20the%20President%20on%20the%20Internet%20of%20Things%20Nov%202014%20%28updat%20%20%20.pdf, page 1.

[9] This expanded volume of data, and the subsequent business and mission demands to analyze it, will likely continue to drive a larger usage and evolution of cloud computing technologies.

[10] Gartner, Inc. "Gartner Says 6.4 Billion Connected 'Things' Will Be in Use in 2016, Up 30 Percent From 2015." November 10, 2015. http://www.gartner.com/newsroom/id/3165317.

[11] Juniper Research. "'Internet of Things' Connected Devices to Almost Triple to Over 38 Billion Units by 2020." July 28, 2015. http://www.juniperresearch.com/press/press-releases/iot-connected-devices-to-triple-to-38-bn-by-2020.

The *NSTAC Report to the President on the Internet of Things* recognized the value of big data for improving the efficiency and effectiveness of Government, private sector, and consumer operations, while also citing the increased risks posed by Internet-connected devices. This NSTAC study of the impact of BDA on the Government's NS/EP functions found the same concerns and warned that safeguarding data and securing devices from both intrusion and as an attack vector must remain a critical priority for both the Federal Government and the private sector.

1.4 Recent Executive Branch Actions

As the value of big data becomes more broadly apparent, the Executive Branch has taken several actions to promote the openness and accessibility of data for the public good.[12] These actions include:

- In May 2009, the Government launched data.gov. The Website currently has over 193,000 datasets from Federal, State, and local agencies, amongst others.[13]

- On May 9, 2013, the President issued Executive Order (EO) 13642, *Making Open and Machine Readable the New Default for Government Information*. EO 13642 states that "Government information shall be managed as an asset throughout its lifecycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, [E]xecutive departments and agencies…shall ensure that they safeguard individual privacy, confidentiality, and national security."[14]

- In May 2014, the EOP published the *Big Data: Seizing Opportunities, Preserving Values* report, in which the Administration called for a 90-day review to examine how the public and private sector could maximize the benefits of big data while minimizing its risks, particularly in the area of privacy.[15]

- NIST, within the Department of Commerce, is leading the development of a Big Data Technology Roadmap. This roadmap seeks to define and prioritize requirements for interoperability, portability, reusability, and extensibility for big data analytic techniques and technology infrastructure in order to support secure and effective adoption of big data. The first version of NIST's *Big Data Interoperability Framework* was released in November 2015.[16]

---

[12] Please refer to Appendix E, *Recent Executive Branch Actions*, for more details on these, and other, Government big data initiatives.

[13] General Services Administration (GSA)). "About Data.gov." Office of Citizen Services and Innovative Technologies. Accessed on March 15, 2016. https://www.data.gov/about.

[14] White House Office of the Press Secretary. *Executive Order 13642, Making Open and Machine Readable the New Default for Government Information*. May 9, 2013. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

[15] Executive Office of the President (EOP). *Big Data: Seizing Opportunities, Preserving Values*. May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[16] NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

2.0 DISCUSSION

2.1 Overview of BDA

In order to extract meaningful and valid insights from big data, one must follow a systematic and disciplined approach for understanding and extracting information from raw data. One commonly used description of big data comes from the 2011 McKinsey Global Institute report, *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, which states:

> Big data refers to a dataset whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data. As technology advances over time, the size of datasets that qualify as big data will also increase. The definition can also vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. Big data can range from a few dozen terabytes to multiple petabytes.[17]

Big data is often characterized as structured or unstructured.[18] With the increase of wireless technology and smart devices, most of the growth in data today is from unstructured data (e.g., social media posts, images, videos, emails, and Web pages).[19] Unstructured data and the surrounding digital universe are projected to double in size every two years and will multiply tenfold between 2013 and 2020 – from 4.4 trillion gigabytes to 44 trillion gigabytes.[20] How unstructured data is managed and turned into actionable information is critical for its successful utilization.[21]

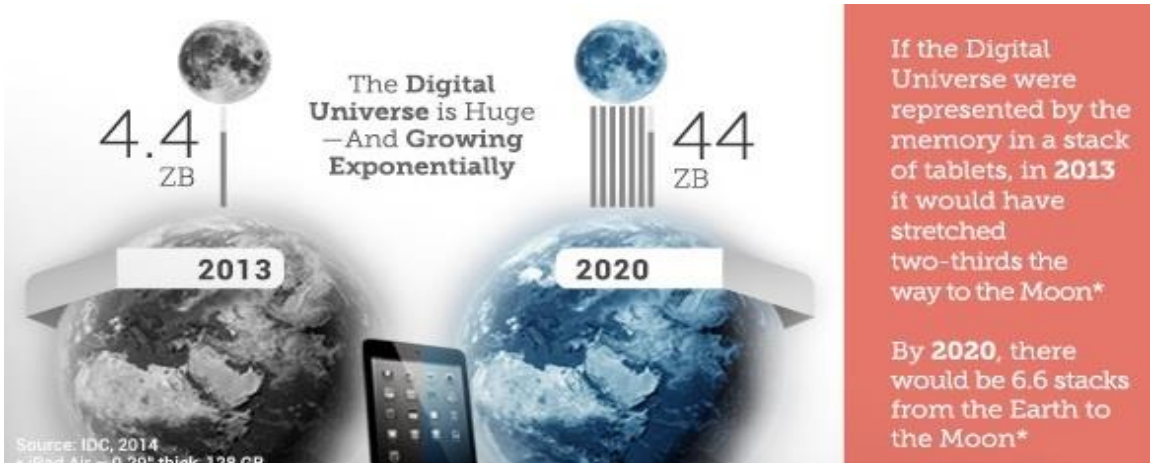


Figure 2.1. Projected growth of the digital universe from EMC Digital Universe with research an analysis by IDC.

[17] McKinsey Global Institute. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

[18] Structured data is easily stored in a relational database because it has defined data fields and types, while unstructured data is information that is not easily interpreted by relational databases.

[19] McKinsey Global Institute. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

[20] EMC Corporation. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. "Executive Summary." April 2014. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

[21] Ibid.

2.1.1 Volume, Velocity, Variety, and Variability

Big data has traditionally been defined using three dimensions: volume, variety, and velocity. This widely used model was first described by a Gartner analyst in a 2001 research report as a way to define the data growth challenges and opportunities with which enterprises and organizations will continue to be faced.[22] The definition of big data has evolved to include other dimensions. For instance, the NIST *Big Data Interoperability Framework* has included variability as an additional dimension. These four "Vs" of big data are the dimensions that will drive the shift to new parallel architectures for data intensive applications.[23] Table 2.1 provides a high-level definition and explanation of each of these dimensions.

DIMENSION	DEFINITION
Volume	The amount of data that is generated, collected, and stored. Volume is the main dimension that makes big data a large challenge, as data is rapidly increasing every year. However, access to greater volumes of data does not necessarily mean access to more actionable information; therefore, users must develop tools that allow them to determine what data is usable.
Variety	The different types of data that are available. Variety is also an increasingly difficult dimension to address. As more and more devices become digital, an increasing diversity of data is being transmitted on these devices. Data can be structured, unstructured, or even semi-structured; however, the majority of data generated is unstructured data.
Velocity	The speed at which data is moving and being processed. This can also refer to the speed at which new data is being created. This is due, in part, to the expanding IoT and the use of social media.
Variability	The meaning of data is changing (i.e., data whose meaning is dependent on context and time). Variability is often called data dynamicity. For example, within the context of natural language processing, a single word may have multiple meanings or may change over time.

Table 2.1. The four "V"s of big data.

2.1.2 The Five Steps of the Big Data Lifecycle

The usage of big data should follow the five steps in the data lifecycle. These guiding principles, along with setting a scope and intended outcomes for the data, help users better utilize these large amounts of information. According to the 2015 NIST *Big Data Interoperability Framework*, the five steps in the data lifecycle are:

1. Data creation and collection;

2. Data identification, availability, and quality;

[22] Doug Laney. "3D Data Management: Controlling Data Volume, Velocity, and Variety." *Gartner.com*. February 6, 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

[23] NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

3. Data analytics;

4. Use of analytic outputs; and

5. Data management, storage, retention, and disposal. [24, 25]

Some big data sources are so large in volume, diverse in variety, moving with such velocity, or so variable, that traditional architectures and modes of data capture and analysis are insufficient. Big data requires scalable architectures that can effectively support the collection and analysis of data, and, through their design, allow for the extraction of value from very large volumes of data of a wide variety. These architectures must also allow for efficient storage and manipulation of data. The development of an architecture capable of handling this amount of data is a multi-layered effort that should be structured around several processes, which include:

- Integrating and using the data in an actionable manner;
- Establishing post-processing reporting and visualization requirements for the data;
- Creating analytic frameworks and engines for processing the data;
- Forming ingest mechanisms and pre-processing of the data;
- Constructing a data storage;
- Developing data warehousing and data modeling techniques; and
- Incorporating a usage scale and elastic forms of infrastructure, including massive compute, storage, and high capacity networks, through which data would be distributed.[26]

Securing the big data and the system architecture is also critical and will require defense-in-depth techniques, including encryption of data at rest and in motion, authentication and authorization of users accessing the system, role-based access controls, and complete monitoring and auditing. Moreover, in order to derive the maximum benefit from the data, one must consider how the data is collected, how it can be used, how to transform it, what analytics should be applied, what the resulting analysis means, and the legal and ethical implications of the results.

2.1.3 The Evolution of Data Science

NIST defines data science as the extraction of actionable knowledge directly from data through the process of discovery, or hypothesis formulation and testing.[27] As big data has evolved, so too has the science around it. There is a movement toward new algorithms and approaches to

[24] NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

[25] Please refer to Appendix D, *Steps in the Data Lifecycle*, for more information.

[26] Richard Puckett. General Electric. *Briefing to the NSTAC Big Data Analytics (BDA) Subcommittee*. March 17, 2015.

[27] NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. *NIST Big Data Interoperability Framework, Volume 1: Definitions*. NIST Special Publication (SP) 1500-1. September 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1, page 7.

develop analytics that are consumable by operational analysts and scientists across many disciplines. As a result, it is now feasible to move toward more complex data science models with open source big data platforms. Having access to these new approaches to data science allows for the creation of new methods for exploiting data.

Traditional approaches to data science tend to be deterministic in nature, following a blueprint or flowchart to ask questions and then take the next step. Deterministic approaches are still relevant and useful for analysts, as they tend to reduce the amount of variability in the answers delivered. On the other hand, non-deterministic approaches to data science allow for variability in the system and approach and may deliver different answers to a question or set of questions an analyst has for a given set of data.

In addition, data visualization techniques have also experienced significant growth. There are a number of very active communities in this area, such as Data-Driven Documents, an open source community based on various types of JavaScript-based visualization techniques.[28] This increased emphasis on data visualization has demonstrated that techniques can be reused and applied across a myriad of domains and datasets. Similarly, the amount of active research and technology development has increased around scaling and sampling of information and measuring the efficacy of the analytics results against the amount of data. Reasons for this are storage requirements, analytic performance, and security. These techniques may allow for collection of much less data while providing comparable results.

2.2 Examples of Big Data Uses

"Information is the oil of the 21st century, and analytics is the combustion engine."[29] When Mr. Peter Sondergaard of Gartner said these words in 2011, few in the room knew, or truly understood, how quickly BDA's impact would be felt. Many experts at the time believed that BDA would have a substantial effect on the Nation's future, but what has surprised the BDA community is how quickly it weaved its way into key aspects of the lives of many Americans. BDA elegantly rides on the coattails of society's last large technical innovations: the Internet and the expansion of available connectivity.

These advancements have created an economy ripe with innovation and have led to an interconnected world. Many sectors of the Nation's economy are deriving value from intelligent sensors and the data they create. The promise of BDA lies in integrating these large amounts of data and using them to deliver previously unknown insights. Forrester Research believes that 80 percent of the value in today's data will come through the integration and correlation offered by BDA techniques, rather than just by collecting the data itself.[30]

The integration of BDA offers organizations the ability to look at problems by identifying patterns that are not immediately obvious to human beings, but can be identified algorithmically. To examine the potential benefits of BDA in the NS/EP context, it helps to first understand areas in which BDA is already making an impact. Three such fields are healthcare, transportation, and

[28] Data-Driven Documents. "About." Accessed on March 15, 2016. https://d3js.org/.

[29] Peter Sondergaard. *Remarks at the Gartner Symposium/ITxpo*. October 2011. http://www.gartner.com/newsroom/id/1824919.

[30] Adam Infante. "Big Data = Big Opportunity? How Does It Work?" PricewaterhouseCoopers. March 3, 2016. http://pwc.blogs.com/analytics_means_business/2016/03/big-data-big-opportunity-how-does-it-work.html.

city management. While each utilizes BDA in different ways, collectively these cases show how powerful BDA can be when utilized correctly, even in its earliest stages.

2.2.1 Healthcare

In many respects, there have already been large institutional shifts in thinking as healthcare providers invest in data-driven diagnostic tools and transition from paper-based record keeping to digital health record databases.

Many believe that tomorrow's cures may already exist today in voluminous patient records. Currently, patient records exist in silos in different corners of the healthcare system. It has been estimated that the Nation's healthcare data consists of more than 150 exabytes, and is growing at nearly 48 percent per year.[31] Often, this data is used once, archived, and never accessed again. This data is comprised of digitized medical images (e.g., X-rays, mammograms, ultrasounds, and magnetic resonance imaging scans), and text data (e.g., demographic information, physician's notes, diagnostic reports, drug information, and financial records).

The clinical system has produced substantial amounts of data to fulfill its bookkeeping needs. In recent years, the healthcare industry has turned to new tools to automate its bookkeeping processes, leading to near instant record retrieval, but little more. Opening this data to analysis, rather than simply retrieval, may unlock a new wave of medical breakthroughs.

Anonymized data analytic techniques aimed at pinpointing trends and insights are helping healthcare providers prevent hospital-acquired infections and reduce hospital re-admissions.

Mr. Bryan Sivak, chief technology officer, U.S. Department of Health and Human Services, believes that while the transition to BDA-driven decisions may prove challenging, this approach has the potential to further improve patient care and minimize costs.[32, 33] Some experts believe that in the future, by using BDA to improve healthcare, doctors will be better able to personalize treatment strategies directly to a patient's genome. In the near term, however, innovative uses of BDA are already being used in a variety of healthcare contexts. For instance, BDA is being used by hospitals to help fight sepsis, which is a secondary condition following an infection in a major organ or as the result of an invasive medical procedure that introduces bacteria into the blood stream.[34] Sepsis is difficult to diagnose, and often the signals that a patient has system inflammatory response syndrome (SIRS), a precursor to sepsis, can mimic symptoms of several other conditions. In addition to the difficulties in diagnosing the disease, the treatment for sepsis is incredibly costly and almost always requires the full attention of medical staff.[35] The Agency for Healthcare Research and Quality found that sepsis was, by a wide margin, the most expensive condition treated in U.S. hospitals in 2011 at more than $20 billion.[36]

---

[31] EMC Corporation. "The Digital Universe Driving Data Growth in Healthcare." 2014. https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf

[32] Mark Headd. "Making the Business Case for Big Data. What Is Big Data, Anyway?" *The Promise of Big Data for the Public Sector*. The Center for Digital Government. 2013. http://media.navigatored.com/documents/CDG13_SPQ1_V.pdf.

[33] RockHealth. "Big Data in Digital Health,"." Accessed on March 31, 2016. http://rockhealth.com/resources/rock-reports/big-data/.

[34] National Institute of General Medical Sciences. *Sepsis Factsheet*. Accessed on March 15, 2016. https://www.nigms.nih.gov/education/pages/factsheet_sePSIs.aspx.

[35] Anne Elizhauser et al. "Septicemia in U.S. Hospitals." *Healthcare Cost and Utilization Project*. October 2011. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb122.pdf.

[36] Stefanie Söhnchen. "Why Big Data Is the Future and the Present of Transportation." *Move Forward.* June 7, 2015. https://www.move-forward.com/news/details/why-big-data-is-the-future-and-the-present-of-transportation/.

Since diagnosing and treating sepsis is challenging, those in the medical technology field have deployed new innovative BDA solutions. Recently, information technology (IT) consulting and managed services companies have combined efforts with data analytics specialists to create real-time clinical monitoring solutions designed to detect SIRS and alert physicians before sepsis occurs. Using a U.S. Food and Drug Administration (FDA)-certified wearable device placed directly on the patient's skin, the BDA technicians are able to design data input systems that monitor a patients key vital signs such as movement, heart rate, breathing patterns, skin temperature, and even posture. This BDA solution integrates real-time patient data with preexisting personal medical data and macro-level medical data made available for research. By combining all three sources of data, data scientists are able to build algorithms that trigger alerts at the onset of SIRS. While each patient can exhibit the early symptoms of SIRS differently, using BDA allows analytical models to vastly improve the rate at which SIRS is detected, thereby preventing the onset of sepsis.

By utilizing BDA, healthcare providers are not only able to greatly reduce the number of sepsis incidences through early detection, but they are also able to develop verified data to understand the intricacies of the condition and create innovative treatments.

2.2.2 Transportation

BDA has started to interact with almost every component of the Nation's transportation network, and plays some role in keeping passengers and freight moving. The American railway system has been one of the early adopters of Internet-connected sensors and large-scale BDA solutions. By deploying Internet-connected wheel and track sensors, speed indicators, visual and acoustic sensors in brakes, rails, switches, and handheld tablets, the railroad industry has begun to amass large quantities of data regarding train movement around the country. This stockpile of data, when combined with powerful analytical tools, can quickly aid in both automated and manual train management. [37] For example, one of America's largest railroads invested in a BDA-driven automated rescheduling system. The automated system analyzes train location and track conditions across the country and manages the scheduling of thousands of trains in real-time. In recent tests, the system was able to route all 8,000 trains across 23 States, in spite of having to deal with multiple track outages without requiring them to stop.[38]

> The railway system's use of BDA is leading to dramatically fewer delays and increased resource efficiencies system-wide allowing freight and passengers to cross the country in a more expeditious manner.

BDA solutions can also be tailored to more complicated environments, with a greater number of actors and more complex rules, such as urban automobile traffic. The country's congested roadways are another prime example of a problem involving a myriad of variables and untapped data. In this context, BDA solutions can be tailored to meet the needs of complicated environments with a large number of actors and more complex scenarios.

For instance, certain areas in New Jersey have deployed BDA to deal with two of the largest contributors to traffic: broken down vehicles and traffic accidents. The system primarily relies on a 22-foot tall sensor screen that collects millions of drivers' mobile and Global Positioning

[37] Stefanie Söhnchen. "Why Big Data Is the Future and the Present of Transportation." *Move Forward.* June 7, 2015. https://www.move-forward.com/news/details/why-big-data-is-the-future-and-the-present-of-transportation/.

[38] Ibid.

System (GPS) data, using BDA to analyze car speed, weather conditions, and community events. The data is then combined with other State and local data on road conditions and traffic cameras. All of this data is analyzed and integrated to form a live data traffic map of more than 2,600 miles of roads in New Jersey. The solution enables the road technicians to detect traffic anomalies in near real-time and has already made a considerable impact to reduce congestion.[39]

One of the most significant components of incident response times has historically been the amount of time between when the incident first takes place and when it is reported. During this time, traffic builds and hinders emergency response efforts. In one instance, the live data traffic map alerted officials to a portion of Interstate 80 that was beginning to inhibit traffic flow, and within 30 minutes technicians were on site and had cleared an overturned car that was causing the delay. Before this BDA traffic system was in place, similar incidents would routinely cause hours of traffic delays.[40]

2.2.3 City Management

The city of Fresno, California has developed several BDA solutions to improve city management. Data collected from a variety of city services, such as water consumption, traffic signals, and electricity usage, enable different types of analysis.

Fresno uses powerful data centers that are able to ingest billions of data points and create visualization models capable of providing city officials previously unknown insights.

Fresno has deployed a network of sensors that detect how its citizens use its roads. The city leverages its sensor network to create a detailed traffic map of its 56-miles of 'smart light' traffic signal-controlled roads, enabling traffic timing specialists to monitor and control hundreds of smart traffic signals at the same time. In most cities, traffic light timing is done by using on-road car counting to form traffic snapshots, which aim to provide technicians a representative sample from which they can extrapolate larger traffic patterns. In practice, these models' broad timing-based structures are inefficient.[41]

Alternatively, by using data collected from Internet-connected traffic sensors, Fresno can visualize weeks of traffic data at once. City planners can run traffic signal simulations against this data and adjust signals accordingly. Furthermore, these models can identify traffic patterns during an exact time of day rather than deploying traditional rush-hour versus non rush-hour models. Officials can even change traffic signals in real-time in response to traffic accidents or evacuation routes.[42] Fresno is also testing automated signal-timing BDA systems, which use BDA algorithms to adjust signal sequences in real-time. Fresno officials believe that soon all of the lights on their busiest roads will be handled using an automated data-driven system.[43]

Fresno's BDA city management solutions go far beyond traffic control. The city has heavily invested in BDA to curb excessive water usage. Due to the installation of digital smart meters citywide, Fresno is able to measure exact water usage rather than estimating consumption. The

[39] State of New Jersey Department of Transportation. "New Jersey Traffic Monitoring Program." June 5, 2014. http://www.state.nj.us/transportation/refdata/roadway/pdf/program.pdf.

[40] Shira Ovide. "Tapping 'Big Data' to Fill Potholes." *The Wall Street Journal*. June 12, 2012. http://www.wsj.com/articles/SB10001424052702303444204577460552615646874.

[41] Ibid.

[42] Ibid.

[43] Ibid.

meters have the ability to record billions of data points, measuring exactly how much water a building is using and when, which allows city officials to see which ones are responsible for the most water usage. Officials discovered water wasters do not fall into geographical categories, as previously hypothesized, but rather are evenly distributed across the city. This data enabled decision makers to reassess and redesign their community communication plans to address a wider audience.

2.2.4 Summary of Successful Examples of BDA

The following table summarizes the above examples of successful BDA deployment.

EXAMPLE	SUMMARY
Healthcare	Using an FDA-certified wearable device placed directly on the skin of the patient, BDA technicians were able to design data input systems that monitor a patients key vital signs such as movement, heart rate, breathing patterns, skin temperature, and even posture. By combining preexisting patient records with this real-time health data, data scientists are able to build algorithms that trigger immediate alerts when a patient shows signs of sepsis.
Transportation	Sensor data on speed, acceleration and deceleration, weather conditions, and community events were combined with other State and local road conditions data and traffic camera footage to detect traffic anomalies in near real-time. Incident response times were dramatically reduced.
City Management	Due to the citywide installation of digital smart meters, Fresno has the ability to know exact water usage metrics rather than estimating usage based on calculated representative samples. After utilizing BDA, officials discovered water wasters do not fall into geographical categories as previously hypothesized, but rather are evenly distributed across the entire city. This enabled decision makers to reconsider and redesign water saving strategies more effectively. Moreover, Fresno's deployment of sensors to monitor traffic signal-controlled roads resulted in better management of traffic throughout the city.

Table 2.2. Summary of the successful BDA use cases.

2.2.5 Commonalities Seen Between Successful Examples of BDA

The following table summarizes common themes seen in each of the above examples.

EXAMPLES	COMMONALITIES SEEN IN EXAMPLES
Healthcare	A. Medical terminology is well-defined B. Some legal parameters around data use (i.e., the *Health Insurance Portability and Accountability Act*) C. Mature pipeline for medical personnel
Transportation	A. Industry terms are known within the community B. Focus is on infrastructure and systems analytics, not personally identifiable information (PII) C. Personnel trained in operational technologies (OT), combined with IT

EXAMPLES	COMMONALITIES SEEN IN EXAMPLES
City Management	A. Industry terms are known to the given community B. Focus is on infrastructure and systems analytics, though there may be challenges associated with using PII (i.e., data on individual homes and cars) C. Personnel in both OT and IT combining their efforts

Table 2.3. Summary of the commonalities seen across the successful implementation of BDA.

3.0 THREE HIGH-LEVEL USE CASES DESIGNED TO SHOW THE USE OF BDA IN SPECIFIC NS/EP SCENARIOS

The sections below outline three hypothetical scenarios: (1) a natural disaster; (2) a man-made disaster; and (3) a cyber attack on critical infrastructure. These use cases demonstrate challenges associated with utilizing BDA in preparation for and response to an NS/EP event. These challenges include the lack of clear rules and policies regarding how big datasets can be made available to first responders, handled responsibly with security and privacy best practices, and utilized to drive response operations.

3.1 Natural Disaster

The objective of this case study is to illustrate an ideal near-term deployment of BDA before, during, and after a critical emergency event, such as a hurricane hitting the gulf coast of Texas. This scenario is based on current technological advances and is constructed using the data lifecycle (as discussed in Section 1.2) which allows a structured approach to the narrative. [44, 45]

3.1.1 Pre-Event

In recent years, a number of State-run transportation efforts have invested in thousands of traffic sensors and the analytical systems to back them up. These sensors, and the data they provide, dramatically improve the way that State officials understand and manage their roadways. They allow officials to have a better understanding of the causes of traffic jams, thereby increasing their ability to react to events. In many instances, by the time transportation crews have arrived on the scene of a traffic incident, the crew has already planned their method of resolving the incident. Furthermore, officials now know where typical anomalies appear and how traffic will typically react. Traffic cameras provide transportation officials with these capabilities; however, the analysis of video recorded by the cameras still needs to be performed by technicians rather than computers. With the myriad of sensors and BDA processing power, computers alert road crews to traffic jams and other traffic related incidents.

When applied to large regions there are secondary system-wide benefits as well, including allowing ambulances and fire crews to reach emergency scenes more quickly and reprogramming traffic signals to keep pedestrians safer. Although these sensors are beneficial in everyday life, these sensors and BDA systems prove invaluable during emergencies.

[44] While the details of this scenario are fictitious, the capabilities described are based on the briefings the NSTAC received.

[45] Each step in the process will help guide the reader's understanding rather than act as a definitive moment when one step ends and another begins. In reality, all stages of the lifecycle are happening simultaneously to different degrees.

3.1.2 Data Creation and Collection

NOAA has exhibited one of the best examples of successful BDA utilization. In addition to being the country's centralized weather-monitoring body, NOAA organizes and processes the ever growing stockpile of environmental data created in the United States. In early 2016, NOAA deployed significant increases in supercomputing capacity designed exclusively to process the massive amount of data created by modern weather forecasting equipment. Since this supercomputing capability came online, it analyzes over a billion weather observations daily. Much like other Federal organizations, NOAA collects its data in many different formats. Data sources can be as varied as a multi-million dollar radar array to a basic Internet-connected weather station at a local high school.

Since its founding, the agency has expanded from collecting basic metrics to a more complex, multi-layered measurement system. Today, each of those basic metrics are saved in at least five different file formats. Interoperability allows NOAA to quickly process data volumes of unimaginable size filled with both structured and semi-structured data. When the agency started using BDA to help process incoming data, scientists spent the majority of their time converting data into workable formats. Now that Federal officials have issued standard application programming interface (API) guidelines for the automated formatting of data received, this time has been dramatically reduced.

Certain sources, such as hurricane hunters used to monitor storms, create massive amounts of critical data. This crucial storm information is only useful if its data can be processed in a timely manner in conjunction with data collected from other sources (e.g., networked weather stations and weather satellites). Interagency data agreements are critical in ensuring the data is sent to processing centers as quickly as possible.

In the future, Government agencies will likely consolidate cloud computing infrastructures to provide lower cost through volume purchase. By utilizing standards such as NIST's *Big Data Interoperability Framework*, agencies will have specific guidance when implementing BDA solutions within the cloud infrastructure. This guidance will reduce errors and improve interoperability and commonality among the agencies' mission specific BDA solutions. This will promote data sharing and allow for easily focusing resources on NS/EP capabilities as required.

Furthermore, this model prevents the creation of data islands where data remains stagnant and decreases in value as it remains unused and becomes outdated. In this natural disaster scenario, the Federal weather and atmospheric community and other emergency response agencies would receive priority access to Federal BDA resources during an NS/EP event.[46]

3.1.3 Data Identification, Availability, and Quality

In this scenario, the storm comes ashore just south of Galveston, Texas. Similar to previous storms, mandatory evacuations begin at least 48 hours before the storm hits. The same road sensors that were installed to alleviate traffic jams and shorten emergency response times in normal weather conditions now become invaluable tools used to support the evacuation. Because of BDA, Texas Department of Transportation officials quickly implement evacuation plans and respond to road side incidents almost immediately. Using the sensors paired with

[46] Wendy Harman. American Red Cross. *Briefing to the NSTAC BDA Subcommittee*. June 9, 2015.

Internet connected signs and signals, officials put a full contraflow strategy in place for all roads within 75 miles of the coast quicker than in the past.

Once again, interoperability dramatically increases the usefulness of data and enables the true power of BDA. For first responders to effectively use BDA, data handlers must first be able to easily ingest incoming data from both private and Federal sources. During an incident of national significance, such as this natural disaster scenario, there is a variety of information sources (e.g., data from mobile phones, social media, electrical outages, traffic and road sensors, water supply, etc.) that, when combined, can provide a rich understanding of the location of the remaining population.[47]

The rapid adoption of smartphones containing advanced sensors and capabilities may provide important new sources of information to emergency responders.

Some residents are unable to evacuate and 9-1-1 calls start to come in before the storm hits, all of which are routed to the local Public Safety Answering Point (PSAP). These PSAPs play a critical role in the Nation's emergency response capabilities. They have answered millions of emergency calls and strategically dispatched emergency personnel and resources accordingly. Historically, they performed this role using basic call data and were often forced to rely on emergency callers alone to describe both their location and situation. This process often left many unanswered questions about what was needed in terms of resource deployment or what emergency responders should expect upon arrival at the scene. During the late 1990s, with the advent of the consumer cellular telephone, PSAPs began using tools that could approximate caller location using data from nearby cell towers. This was a powerful change that improved situational awareness as the decline in landline services reduced the location-based qualities of any specific phone number, but operators were still left with the task of collecting a vast amount of information from the caller.

The hypothetical affected population along the Texas coastline now have the ability to receive important evacuation notices and weather updates through their smartphones and other electronic devices as a result of the consumer electronics revolution. With the rapid adoption of smartphones and household IoT devices, an unprecedented number of networked sensors are connected to the Internet. As the devices that call 9-1-1 become more sophisticated, so too does the data that PSAPs may be able to pull from calls. Many newer devices include advanced sensors such as high definition microphones, text, cameras, accelerometers, barometers, and more. In the future — depending on data agreements between the State government, local law enforcement, and mobile phone carriers — some jurisdictions may have the ability to enable callers to share situational data with PSAP operators when a user places an emergency call. Devices will produce a substantial amount of data exhaust that could be crucial when identifying victims and measuring response effectiveness in an NS/EP event. To have the full potential of these technological advancements, it is critical to develop the needed frameworks for data sharing.

In this scenario, an Emergency Operations Center in Texas could quickly become a large operator of an intelligent sensors network, illustrating how BDA technology can save lives. As storm victims call 9-1-1, they could be prompted with a message on their phone asking if they

[47] Stephen Hayes and Louise Tucker. Ericsson, Inc. *Briefing to the NSTAC BDA Subcommittee*. April 14, 2015.

would like to share crucial information, such as location and video data. By using BDA to process this vast amount of data, first responders can better traverse the storm-damaged region and the time it takes to get to storm victims is dramatically reduced. In the same way that Federal, State, and local networks show increased activity around the storm, so too do private networks. Americans often turn to social media to post about the storm and read updates the second they become available, generating significant amounts of data that can then be analyzed for emergency response purposes.

To first responders, social media feeds and trending hashtags contain critical information regarding the storm, but also are flooded with unnecessary noise created from the numbers of users. Social media companies can make a substantial impact on storm recovery efforts by deploying their "check in" emergency protocols that allows emergency responders to obtain useful information regarding specific locations and individuals.

3.1.4 Data Analytics

The wide variety of data is of little use unless it can be processed and analyzed; the true value of data is rarely in its initial form. As a result of the growing ability to apply BDA to call information, first responders are able to make analytical predictions regarding a large portion of typical emergency incidents. For example, correlating time stamp and location data with distinct incident types has the potential to identify patterns of accidents related to weather. This has been seen in current small-scale BDA deployments and led to critical pre-deployment of ambulances and medical services prior to specific weather patterns. For instance, ahead of substantial rain storms, the city of Los Angeles positions ambulances near intersections that BDA has deemed troublesome in an effort to reduce, if not eliminate, response times.[48]

This heightened ability to predict outcomes is critical in preparing and responding to NS/EP events. However, the true potential of BDA comes from the integration of these disparate datasets. For unstructured text data, Federal data scientists often rely on powerful text analytics approaches, such as natural language processing (NLP).[49] Historically, in the moments following a catastrophe, there is a substantial amount of unstructured data created (e.g., social media, news reports, pictures, emails, text messages, sensors, etc.). The timely analysis of this data by first responders is essential to achieving and maintaining situational awareness in response to natural disasters such as the one laid out in this case study.

3.1.5 Data Analytics Output

The ability to quickly derive connections between systems by recognizing patterns in text will greatly enhance the future of interoperability. As noted above, members of a group of first responders deployed to a location are likely to each describe an event differently. NLP allows officials to combine a large number of written accounts and derive the most important details from a diverse and complex dataset. NLP can also be adapted to align with voice-to-text technologies, as well as to derive insights based on radio communications between first responders.

---

[48] Trey Forgety. National Emergency Number Association. *Briefing to the NSTAC BDA Subcommittee*. December 15, 2015.

[49] NLP is the process of analyzing unstructured text to derive actionable insights. For instance, processing social media data could reveal specific locations requiring the attention of first responders. NLP can be a versatile tool for dealing with a situation rife with high volume, unstructured, or heterogeneous data, such as a hurricane.

In addition, NLP can play a role in storm recovery efforts because the events are associated with large volumes of unstructured text data. The true value of BDA is the potential to draw connections and insights from this type of data far more quickly than humans are capable of doing. BDA can also provide insights into the interconnected nature of critical infrastructure systems so that officials can respond appropriately in order to avoid cascading failures.

> A significant gap exists between the needs of first responders in managing and analyzing big datasets and the numbers of experienced data scientists capable of effectively helping them do so.

During this crucial time of data processing, there is one obstacle that both the Texas and Federal hurricane response teams could encounter – limited staff and SMEs. Arguably, one of the BDA community's largest problems to date has been finding and hiring data scientists who can manage and manipulate complicated data systems. It is easier to invest money in hardware and technology than it is to create a new branch of academia, but what BDA often needs most is simply enough bright minds to interpret the analytic outputs.

3.1.6 Data Analytics Storage and Disposal

After the storm, officials have the opportunity to evaluate how effectively BDA supported preparation and response. By operationalizing BDA, Federal and State officials have a much higher likelihood of having been able to order and execute swift and effective evacuations in most of the affected areas. For those who were unable to evacuate, first responders are more likely to have been able to efficiently route themselves to those in need. Critical infrastructure throughout the region is more likely to have maintained operation despite substantial damage from the storm due to creative problem solving and seamless transfer of functions to redundant systems. Power is more likely to have been restored quickly due to analytical models that prioritize repairs and keep crews working in the most important areas.

The state of emergency will be rescinded after the storm and the personal data given to Government from private individuals and entities to assist in the response and recovery efforts will be deleted to prevent misuse. In the interest of privacy, algorithms used by the Federal Government were designed to keep data anonymous and secure. Careful planning and exercises involving the collection, use and eventual deletion of data had been followed.

Security at all levels is important. Although there are significant benefits associated with the use of BDA, techniques should be deployed with security in mind (e.g., secure communication, storage, and disposal of data used in BDA). While some aspects of this are not new and exclusive to BDA, the scale at which BDA operates brings new caveats to traditional networking security problems. BDA requires the stockpiling of data, which increases the value to an attacker and must be considered when designing security platforms.

Additionally, in the same way individuals are afforded protections by a "Good Samaritan" doctrine when helping someone during a medical emergency, so too should private entities be afforded protections if they consent to sharing data with the Government to assist communities during NS/EP events. In the event of a natural disaster, reluctance to share life critical data could impair rescue efforts and cost lives. By relying on a framework of "Good Samaritan" protections, both Government and private entities would have a clear understanding of rules regarding the protection of privacy, data use, ownership, storage, retention, accidental disclosure, and deletion.

3.1.7 BDA Lifecycle Analysis for Natural Disasters

The table below illustrates how the natural disaster case study can be mapped to the BDA lifecycle.

LIFECYCLE STEP	SUMMARY	FINDINGS
I. Data Creation and Collection	Federal, State, and local agencies, as well as private companies created a substantial amount of data before, during, and after the storm.	Data must be made available to the appropriate entities during natural disasters. Creation and collection should be aided by the standardization of data formats and APIs used to mark data.
II. Data Identification, Availability, and Quality	Data scientists designed interoperable systems to aid in the exchange of data. A central catalog of data was created which didn't share the data itself, but communicated what data was available.	Data interoperability enables greater use of BDA; data can only play a critical role if it is made available to emergency responders in a format consistent with analytic tools or has an API used to mark and label data fields.
III. Data Analytics	A combination of man-made and machine learning algorithms was used to derive previously unknown insights. Federal processing infrastructure was made available to State officials.	The creation of BDA tools designed specifically for NS/EP purposes must be created and tested before a storm. Sample data on which to run is crucial in these tools creation. BDA often requires substantial processing power, network storage, and computer resources. Therefore, the Federal Government should promote the availability of resources required to respond to NS/EP events.
IV. Data Analytics Output	Analysis of data led to actionable insights that allowed for more efficient evacuations, more resilient infrastructure and overall fewer lives lost.	BDA has the ability to dramatically affect the outcome of NS/EP events. Response plans must be designed to incorporate the output of BDA as to not waste its powerful potential.
V. Data Analytics Storage and Disposal	After the storm, Federal networks shutoff connections to private networks and private data was deleted.	It is crucial for there to be a clear separation between Federal and private networks, as well as Federal and private data. Diligent data minimization practices must be put into place to increase privacy concerns, as well as minimize cyber security risk.

Table 3.1. The natural disaster use case mapped against the BDA lifecycle.

3.1.8 Findings

The success seen in this use case can be attributed to several key principles regarding the implementation of BDA in the context of a natural disaster, including:

- The ability for different agencies to seamlessly share information during an NS/EP event is critical. In the above scenario, necessary data sharing was made possible through agreed

upon data formats, searchable APIs, and text analytics. The rapid increase in the amount of data produced will likely continue. Overall, the situational awareness of emergency responders will be significantly improved if this data can be quickly accessed by the appropriate police, fire, and Federal personnel. Making the technical linkage between systems easier will enable officials to use a new set of tools when responding to large NS/EP events. Additionally, BDA will require infrastructure from both the Federal Government and the private sector allowing emergency responders to simply tap into readily-available resources when responding to an event, rather than having to build them.

- To fully capture their potential benefits, all BDA operations rely on human technicians. This use case involved many trained individuals who could operate the BDA systems that powered the response and recovery as well as interpret their outputs. One of the largest commonalities between BDA systems is that no matter how complicated they are, there is no replacement for an intelligent human user. BDA is a tool and as such, Government and industry need skilled technicians who are able to maximize its use. Data scientists need a rich domain understanding, a solid math or statistics background, strong computer science skills, and a holistic philosophy when dealing with problems.

- Achieving the many benefits of BDA requires clarification on the legal availability, access, handling, and protection of data. In the above hypothetical scenario, entities were technically able to share their data, but also possessed the legal protections needed to do so. For BDA to reach its potential, there needs to be clarification around the transfer of data. In natural disasters, unclear legal and statutory environments inhibit timely response and recovery efforts. BDA, by design, brings together multiple data sources and partners. Legal clarity would set the groundwork for cooperation and collaboration between these entities.

A "Good Samaritan" Framework could encourage data sharing in response to a natural disaster, while providing both Government and private entities with a clear understanding of rules regarding the protection of privacy, data use, ownership, storage, retention, accidental disclosure, and deletion of datasets.

3.2 Man-Made Disaster

The purpose of this use case is to discuss how BDA can be applied to prevent, mitigate impact, and assist in the recovery of man-made disasters. In order to be an effective tool, BDA techniques must be integrated into the operational framework of the Nation's NS/EP infrastructure. Ultimately, the ability to deliver actionable, timely and accurate BDA results to Federal, State, and local government agencies determines if a man-made disaster can be prevented and if not, the degree to which it can be mitigated.

Man-made disasters can manifest in multiple ways. Often man-made disasters are associated with acts of terrorism in which the primary objective is to create fear through the ruthless destruction of life and property. More broadly, this use case considers man-made disasters in the current context to include not only acts of terrorism, but also events caused by accidental or negligent actions that result in widespread destruction or threat to life and property, without

> The term **man-made disaster** refers to events that arise either from the malicious or accidental actions of one or more individuals; such that the consequences of those actions threaten national security or warrant "out-of-the-ordinary" emergency response actions that may require collaboration by agencies at the Federal, State, and/or local government levels.

ascribing any particular motive for the perpetrators' actions. For example, the negligent act of an individual responsible for a major power grid may result in a national security breach or warrant an out-of-ordinary response by Federal, State, and local government officials.

As was experienced in the United States on September 11, 2001, Boston in 2013, Paris in 2015, and Brussels in 2016, a common strategy employed by today's terrorists incorporates near-simultaneous and coordinated attacks across a large geography. Common themes from these and similar terrorist events are used in this hypothetical scenario to depict how BDA can be used to thwart such events and/or mitigate the impact to loss of life and property.

3.2.1 Scenario Description

This hypothetical scenario begins five years ago when a team of individuals are tasked to develop a coordinated multi-city attack. The terrorists analyze points of vulnerability, assess timing to ensure maximum impact, and ultimately determine the method and resources for the attack. These individuals frequently communicate with their parent organization through primarily encrypted messages; however, some communication is unencrypted. The use of social media is also commonly used, but with the discussions disguised through cryptic messages and code phrases that are seemingly innocuous and unconnected. They occasionally travel to common domestic and/or international destinations.

The terrorist team begins harvesting public databases for information that supports their planning process. This involves multiple and repeated Web searches into future public events such as concerts and sporting events. The team develops attack options and communicates the details among the team members and their parent organization. The increased chatter includes increased security as the plan is finalized. At this point, there is a significant amount of disjointed information contained in a variety of data sources that point to an eminent threat. The time to process the information into actionable intelligence is short, and coordination between Federal, State, and local authorities is impeded by complex operational procedures. This is the first pivotal point for BDA to play a substantial role in preventing the event.

Finally the day of the attack arrives. The team members travel to their target locations, some arriving by plane, others by ground transportation. BDA processes have identified a potential threat and collection of meaningful data intensifies. Correlation of travel itineraries, car rentals and credit card purchases can possibly pinpoint the targeted locations and the individuals that have been dispatched to each location.

Unfortunately, the relevant preventive steps did not fall into place and the attack occurs. Five cities are struck within one hour and there is widespread destruction. The terrorists begin executing closure procedures to evade detection and escape, but they leave a trail of evidence that correlates with the picture that was emerging from BDA activities. Video streams and individual photographs from private citizens and fixed camera locations stream into a make-shift command center. The terrorists begin to post details on social media and communicate amongst themselves and with their parent organization. Finally, authorities are able to positively identify

them and launch a massive manhunt effort. They provide photographic evidence to the general public over social media and television. Within days, the terrorists are apprehended and a Federal prosecution case ensues.

The use of big data and data analytics have become important tools for law enforcement and intelligence agencies. While technological advances and personnel skills enhancements have helped drive major successes, there are opportunities for improvements, particularly in information access, data formatting, and availability of analysts.

Using the 2013 Boston Marathon as an example, the City of Boston Police Department (BPD) did not readily have access to data from the Federal Bureau of Investigation, United States Secret Service, or larger intelligence community. Instead, an interagency meeting was needed on the day of the attack to coordinate each organization's role in the response efforts and how information was to be shared among stakeholders.[50]

The BPD manually gathered social media, photo, and video stream data from witnesses through online searches and by door-to-door canvassing. During this initial data gathering and analysis phase, data governance and ownership were difficult to manage as multiple agencies collected a variety of data in a variety of formats and structures.[51]

In order to address the privacy concerns associated with its use of big data, the BPD continues to work closely with the American Civil Liberties Union (ACLU). For instance, the BPD has previously used license plate readers to track and apprehend suspects in criminal investigations; however, the ACLU successfully petitioned to end this practice on the grounds that the data was retained for too long of a period. The BPD has since altered its policies and procedures to limit the retention of the data it collects.[52]

This scenario illustrates the opportunities to apply distinct BDA techniques to each of FEMA's emergency management phases:

EMERGENCY MANAGEMENT PHASE	BDA APPLICATION
Threat Assessment and Mitigation	**Predictive Analytics** can be used to develop effective *national security* threat analysis that can lead to prevention (e.g., man-made disasters) or in optimizing preparation for eminent natural disasters to mitigate the loss of life and property.
Preparedness	**Vulnerability Analytics** can be used to estimate the impact of particular disasters. For example, analytical processes can be used to estimate the impact of loss of critical infrastructure due to natural or man-made actions. The outcome of these analytical processes can be used to enhance emergency preparedness.

[50] William Evans. Boston Police Department. *Briefing to the NSTAC BDA Subcommittee*. June 30, 2015.
[51] Ibid.
[52] Ibid.

EMERGENCY MANAGEMENT PHASE	BDA APPLICATION
Response	**Operational Analytics** can be used to coordinate emergency response across Federal, State, and local agencies. This involves the timely collection and dissemination of critical real-time data between government agencies and the general public as required, with the goal of minimizing loss of life and property during an on-going emergency situation. This includes on-going damage assessment as the emergency situation evolves. For man-made emergencies, this also includes the collection of evidence.
Recovery	***A Posteriori* Analytics** can be used to measure the impact of the emergency event and assist Federal, State, and local agencies to organize and finance recovery efforts.

Table 3.2. FEMA's four phases of emergency response mapped against BDA applications.

3.2.2 Man-Made Disasters: Challenges for the Future

The ideal outcome of applying BDA to the threat of man-made disasters is prevention; failing that, the key objectives are to: (1) reduce or mitigate the impact when such man-made disasters occur; (2) rapidly apprehend the perpetrator(s); and (3) assist in prosecution in the event of criminal activity.

The challenges and shortcomings in applying BDA to the problem of man-made disasters can be summarized in terms of the common characteristics used to denote big data:

- **Volume:** The universal problem with big data is that datasets continue to get exponentially larger over time. In some application domains, more recent data is more relevant than data that is aged (e.g., weather prediction versus climate change). History has shown that relevant data about malicious individuals may span years or even decades.

- **Variety:** Big data sources that concern individuals are highly varied, and include highly structured transactional datasets (e.g., airline reservations, passport checks, car rentals, etc.) and substantial unstructured datasets (e.g., social media feeds, video surveillance cameras, etc.).

- **Velocity:** In both preventing and responding to disasters, time is frequently of the essence and the ability to respond in a timely manner often determines if prevention is possible, or the extent of the loss of life or property once an event has occurred. Therefore the availability of analytical output from BDA processes in a timely manner frequently dictates its utility.

- **Variability:** Inconsistency between datasets that relate to individuals can result in divergent or inaccurate threat assessments. As described above, relevant data may span a decade or more. Therefore, the high degree of variability in datasets over extended periods of time poses significant challenges in determining relevance. The quality of captured data concerning individuals can vary greatly. For example, structured transactional data is largely factual, unless it has been intentionally altered. Data collected from social media feeds can be noisy and in some cases intentionally deceitful. The quality of highly processed unstructured data such as video streams is subject to the ability of analytics processes to

extract accurate and meaningful data (e.g., such a person is/was at a certain location at a certain time).

Ultimately, effectively utilizing BDA within the context of man-made disasters requires linking, connecting, correlating, and analyzing data from multiple sources. These sources can be found in a variety of domains, including the private domain (e.g., car reservations and telephone records), the Government domain (e.g., passport checks and driver's license applications), and the public domain (e.g., social media feeds).

3.2.3 BDA Lifecycle Analysis for Man-Made Disasters

The table below illustrates how the man-made disaster case study can be mapped to the BDA lifecycle.

LIFECYCLE STEP	SUMMARY	FINDINGS
I. Data Creation and Collection	An impactful application of BDA to the problem of man-made disasters is in the "threat assessment and mitigation" phase. In this phase, the relevant datasets are the large, complex, highly diverse, and in most cases unconnected.	Datasets differ significantly across the emergency response lifecycle. Relevant datasets for "threat assessment and mitigation" include Government-collected data and commercial sources collected in the private sector. Datasets associated with preparedness, response, and recovery are largely within the Government (e.g., Federal, State, and local levels).
II. Data Identification, Availability, and Quality	Rigorous analysis of successful cases where man-made disasters have been thwarted, and extensive post-mortem analysis in cases where a man-made disaster occurred, is necessary to develop a comprehensive inventory of relevant datasets.	A critical aspect of identifying threats and developing risk-mitigation strategies is the need to identify and share information between the public and private sectors. The lifecycle of data relevance is potentially very long (multiple-years), and the type and nature of relevant data will evolve as perpetrators develop more sophisticated methods for planning and executing attacks.
III. Data Analytics	While the relevant data lifecycle is very long, the time to generate actionable BDA output can be extremely short. Failure to produce timely and accurate results will frequently mean the difference between thwarting an event and having it fully materialize.	Both datasets and analytical techniques vary across the emergency lifecycle. Optimizing the BDA approach requires deep insight into the operational nature of each phase. In addition, the BDA output from each phase must be interoperable. For example, BDA associated with "vulnerability analysis" should inform "threat assessment", and in the event a threat materializes into a disaster, both of these should inform "response" and "recovery".

LIFECYCLE STEP	SUMMARY	FINDINGS
IV. Data Analytics Output	BDA output must be tailored for the target government organizations that consume and act on it. Tailoring includes timing, content, dissemination methods, etc.	The utility of BDA output hinges on its timing, accuracy, and specificity; all of which ultimately determine if it is actionable.
V. Data Analytics Storage and Disposal	Storage and disposal of relevant data and BDA output is extremely complex for the early phases of emergency response (threat assessment, mitigation, and preparedness), and by comparison, relatively straightforward in the later phases (response and recovery). Significant research is required to develop comprehensive BDA solutions for these early phases.	Threats from man-made disasters come, go, and evolve. Knowing what state a particular threat is at comprises a significant part of the problem. Hence the storage and disposal of relevant data and BDA output is a non-trivial problem. What might initially seem like a low probability outcome may eventually fully materialize into a significant man-made disaster. Therefore developing a successful strategy to data retention and BDA output storage and retrieval is critical.

Table 3.3. The man-made disaster use case mapped against the BDA lifecycle.

3.2.4 Findings

As indicated previously, the application of BDA to the emergency lifecycle associated with man-made disasters requires integration into the operational capabilities of the Nation's NS/EP infrastructure, beginning with threat assessment and mitigation, and ultimately response and recovery. Specific findings regarding the operational integration of BDA include:

- Predictive analytics for man-made disasters is fragmented across Federal government agencies and between Federal, State, and local agencies.

- There is a need to define, develop, and conduct vulnerability analytics processes on an on-going basis for critical infrastructure that would cause serious damage to national security if it were to be compromised by man-made or natural causes.

- Interoperable operational analytics technologies are not widely in use across Federal, State, and local agencies. These solutions should leverage technologies within the different levels of Government (e.g., Federal, State, and local) and private citizens. Response could be enhanced if conditions to encourage the sharing of information exist without fear of liability.

- The U.S. Federal Emergency Management Agency (FEMA) and other State emergency management agencies lack a common toolkit of *a posteriori* analytics solutions to standardize and optimize the collection, analysis of damage assessment data.

3.3 Cyber Attack on Critical Infrastructure

The security of cyberspace affects everyone. The ability to gain national-level context of cyber attacks targeting critical infrastructure in a timely manner allows the Federal Government to effectively respond. This case outlines how BDA can be applied to mitigating the effects of cyber attacks on the Nation's critical infrastructure. As previously defined, a variety of cybersecurity capabilities can be enabled by leveraging various data sources and types at rapid speed across a wide range of Government and industry organizations. Given the size of cyberspace, the ever increasing sophistication of cyber adversaries, and the speed with which the situation changes, leveraging BDA to promote the cybersecurity of the Nation's critical infrastructure will be a vital component in national security going forward.

Cyber attacks could be utilized by the Nation's adversaries in a variety of ways and could result in significant NS/EP impacts.

There are many ways that an adversary might use cyber capabilities to attack and compromise national critical infrastructure. An adversary might target a pivotal agency to gain intelligence on how critical infrastructure and key resources organizations will plan and respond to a disaster, or launch a distributed denial-of-service (DDoS) attack against emergency call centers to impede the ability for first responders to provide emergency services. There are a range of potential cyber attacks that could compromise NS/EP functions. The challenge in the current environment is that the attack surface in contemporary networking is effectively unlimited. Further, adversaries do not obey local laws, which combined with the asymmetric nature of cyber attacks where a single individual can cause significant, wide ranging damage, leads to a highly scalable, unstable, and rapidly changing problem space. This ever-expanding attack space explains why signature-based approaches (e.g., intrusion detection systems [IDS], intrusion prevention systems [IPS], and anti-virus software) increasingly fail to stop new zero-day attacks and prevent systems compromise.

As thinking about protecting the Nation's national security and enhancing emergency preparedness matures, it is useful to think in terms of adversary groups and their attack playbooks. Adversary playbooks are the generic guidelines that an adversary group develops in order to penetrate a potential victim's network and accomplish its mission – a step-by-step process that outlines the tactics used to attack their targets. Once they decide upon a target or class of targets, they launch a campaign based upon their playbook. While overcoming a particular obstacle may require modifications to the playbook, an adversary campaign is the direct application of the adversary playbook against a specific victim or victims. Network defenders are the people who try to prevent the adversary's campaign from succeeding.

Like network defenders, adversary groups are limited by the resources at their disposal. Specifically, the time-consuming and resource-intensive nature of inventing new ways to attack each new target means that they typically adopt and tailor their already established playbooks to the target at hand.

When they launch their campaigns against specific targets, adversary groups leave behind indicators of compromise.[53] Indicators of compromise are forensic artifacts that describe an

[53] Indicators of compromise are forensic artifacts that describe an adversary's methodology; digital clues left behind by the adversary group as it works its way through the phases of the attack lifecycle.

adversary's methodology; digital clues left behind by the adversary group as it works its way through the phases of the attack lifecycle. The attack lifecycle is a phased model that describes the tasks an adversary group must accomplish in order to complete its mission: reconnaissance for victim weaknesses, delivering the initial attack, compromising victim zero, and installing a command and control channel. From here, what adversaries do next depends on their motivation (i.e., crime, espionage, hacktivism, terrorism, warfare, or mischief). They might spread laterally, compromising as many endpoints as they can for some future task. They might cause damage. They might extract valuable or proprietary information.

Over time, threat researchers, security vendors, and Government intelligence agencies discover new indicators of compromise by observing campaign activity. Because of the adversary group's propensity to reuse their playbooks against multiple victims, it is possible for the network defender community to recognize many, if not most, of the indicators of compromise that adversary groups leave behind as they attack their victims.

Threat prevention is the act of turning known indicators of compromise into one or more deployed prevention controls.[54] Prevention controls are technical safeguards or countermeasures derived from observing adversary group campaign activity that network defenders design and deploy to thwart adversary campaigns at each phase of the attack lifecycle. It is possible to thwart an entire adversary group campaign by deploying the correct prevention control at the precise spot in the attack lifecycle. Moreover, deploying as many prevention controls as possible, at every stage in the attack lifecycle, almost assuredly guarantees that the specific adversary campaign will not succeed.

Threat prevention alone, however, is not sufficient to preventing cyber attacks. If network defenders only use threat prevention to enhance NS/EP functions, they will likely fail for one of two reasons: (1) determined adversary groups will deploy new, unknown playbooks that have not been seen by the network defender community before; or (2) the adversary group may be operating in an area of a network where the network defender team has failed to adequately deploy their threat prevention program. In both cases, more efforts are needed in order to improve the existing methods used to detect indicators of compromise. Threat detection is the act of hunting for "known" indicators of compromise throughout the enterprise at each phase of the attack lifecycle and investigating "unknown," anomalous behavior wherever it is found, deciding what the anomalous behavior is and taking the appropriate actions once discovered.

Preventable attacks are those that are amenable to the signature approach.

Containable attacks are those that are only discovered because of breach/compromise.

As the networking space expands, it may be beneficial to view attacks, exploits, and vulnerabilities in two categories: preventable and containable. Preventable attacks are those that are amenable to the signature approach (e.g., port scanning, known virus/malware, simple denial of service, certain types of DDoS attacks, and known phishing attacks) and containable attacks that are only discovered because of breach/compromise (e.g., backdoors, man-in-the-middle, network pivots, spoofing, botnets, certain types of DDoS, malicious cloud jobs, novel virus/malware, and novel phishing attacks). It is difficult to classify the type of attack, as attacks

[54] Prevention controls are technical safeguards or countermeasures derived from observing adversary group campaign activity that network defenders design and deploy to thwart adversary campaigns at each phase of the attack lifecycle.

are either amenable or non-amenable to signature-based detection. The value of BDA in protecting the Nation's critical infrastructure against cyber attacks is in providing prevention controls for known adversary playbooks to network defenders and for analyzing anomalous behaviors on critical infrastructure networks to detect new unknown attacks.

For containable attacks, it is only by leveraging BDA that the context for these evolving threats can be inferred in a timeframe that allows for prevention, detection, and mitigation of attacks that are not amenable to threat-based detection. Actionable cybersecurity threat intelligence increases the network defender's ability to prevent, detect, and respond to an active cybersecurity event. The methodology for this case study is to discuss the application of BDA to the protect, detect, and respond functions identified in the NIST's 2014 *Framework for Improving Critical Infrastructure Cybersecurity* (NIST Cybersecurity Framework) in the context of a containable, novel attack.[55,56] Much like FEMA's four phases of emergency management (see Appendix F), the NIST Cybersecurity Framework has similar corollaries for mitigation, preparedness, response, and recovery.[57,58]

The ability to detect the event, and determine what type of event is occurring is key to an effective response. Through the use of BDA at the detect stage, the response can be more effective and reduce the consequences that might require recovery. Together, the use of BDA to enhance detection and response is key to minimizing the impact of a cyber event. Thus, the BDA lifecycle can be mapped to the NIST Cybersecurity Framework.

CORRELATING THE FIVE FUNCTIONS OF NIST'S CYBERSECURITY FRAMEWORK TO THE STEPS OF THE BDA LIFECYCLE[59]

FUNCTIONS OF THE NIST CYBERSECURITY FRAMEWORK	BDA LIFECYCLE STEPS
Identify: Develop the organizational understanding to manage cybersecurity risk to systems, assets, data, and capabilities.	**Data Creation and Collection:** *Development of tools and techniques, including exchange syntax, semantic interoperability, and data sharing frameworks to share intelligence.*
Protect: Develop and implement the appropriate safeguards to ensure delivery of critical infrastructure services.	**Data Identification and Availability:** *Development and deployment of tools to populate shared data and threat intelligence.*

55 NIST. *Framework for Improving Critical Infrastructure Cybersecurity*. February 12, 2014. http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf.

56 The 2014 NIST *Framework for Improving Critical Infrastructure Cybersecurity* (NIST Cybersecurity Framework) contains five functions: (1) identify; (2) protect; (3) detect; (4) respond; and (5) recover. For the purposes of this use case and considering how BDA can be deployed during an event, the NSTAC chose to focus on the detect and respond functions.

57 Please refer to Appendix F, *FEMA's Four Phases of Emergency Management*, for more information.

58 The NSTAC chose to focus on NIST Cybersecurity Framework in this case study because the U.S. Federal Emergency Management Agency's (FEMA) framework is specific to emergency management in terms of a wide spread event in nature that requires FEMA to coordinate efforts in different phases of response. While similar, the NIST Cybersecurity Framework is the more authoritative source within the Government specific to cyber incidents, and is specific to the common set of actions used during different phases of the cyber kill chain.

59 NIST. *Framework for Improving Critical Infrastructure Cybersecurity*. February 12, 2014. http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf.

CORRELATING THE FIVE FUNCTIONS OF NIST'S CYBERSECURITY FRAMEWORK TO THE STEPS OF THE BDA LIFECYCLE[59]	
FUNCTIONS OF THE NIST CYBERSECURITY FRAMEWORK	**BDA LIFECYCLE STEPS**
Detect: Develop and implement the appropriate activities to identify the occurrence of a cybersecurity event.	**Data Analytics:** *Application of the existing large suite of analytics tools to consistently collected and available data to create more effective detection techniques.*
Respond: Develop and implement the appropriate activities to take action regarding a detected cybersecurity event.	**Data Analytics Output:** *Use of data visualization and publication techniques to rapidly disseminate detected threat intelligence in an actionable manner.*
Recover: Develop and implement the appropriate activities to maintain plans for resilience and to restore any capabilities or services that were impaired due to a cybersecurity event.	**Data Analytics Storage and Disposal:** *Process and techniques to analyze events post-facto to improve baseline security posture.*

Table 3.4. Correlating the five functions of NIST's Cybersecurity Framework with the BDA lifecycle.

3.3.1 Scenario Description

Cybersecurity events are challenging due to the complex interconnectedness and intangible nature of cyberspace. Adversaries can use and recombine any number of different attacks, so any case study in this space is purely notional. In order to illustrate the shortcomings of current infrastructure, and to illustrate the potential that BDA have to remediate this, the NSTAC proposes the following hypothetical scenario.

A malicious adversary wants to disable a component of U.S.-based critical infrastructure. This could be any of the 16 critical infrastructure sectors, but for the sake of this case study, the focus is on any of the major lifeline sectors such as power, water, communications, and financial services.[60] To facilitate the attack, the adversary must quickly and simultaneously penetrate and re-configure the supervisory control and data acquisition (SCADA) system or other operational support systems supporting the impacted critical infrastructure. An attack of this nature requires years to plan and several months, possibly even years, to execute. In this case, the attack follows a specific planning sequence:[61]

- In order to develop a network pivot point, the adversary takes advantage of the increasing attack surface to insert a backdoor into the networks for the companies operating critical

[60] As outlined in *Presidential Policy Directive 21: Critical Infrastructure Security and Resilience* (2013), there are 16 critical infrastructure sectors "whose assets, systems, and networks, whether physical or virtual, are considered so vital to the United States that their incapacitation or destruction would have a debilitating effect on security, national economic security, national public health or safety, or any combination thereof." These sectors include: (1) chemical; (2) commercial facilities; (3) communications; (4) critical manufacturing; (5) dams; (6) defense industrial base; (7) emergency services; (8) energy; (9) financial services; (10) food and agriculture; (11) Government facilities; (12) healthcare and public health; (13) information technology; (14) nuclear reactors, materials, and waste; (15) transportation systems; and (16) water and wastewater systems. For more information, please visit the Department of Homeland Security's (DHS) Office of Infrastructure Protection Website at: https://www.dhs.gov/critical-infrastructure-sectors.

[61] This scenario is hypothetical. It is unlikely to happen as many of these systems are highly secured and the networks they run on are segregated; however, it illustrates how a BDA approach to cybersecurity might help to detect and intervene early to prevent, or at least mitigate, the effects of such an attack.

infrastructure or their suppliers. There are a wide range of possible attack vectors that could be exploited for this purpose.

- Through this backdoor the adversaries place network "sniffers" inside of the network of the owner or operator of the critical infrastructure, or their suppliers, to learn topology and identify the information assets critical to the operation.

- Simultaneous to collecting network topology information, the adversary engages in a targeted campaign of phishing and social engineering to obtain password and credential information from those with administrative oversight of the critical assets, for use in the preparatory stages of the final operation.

- With topology and administrative credentials known, the adversary commences the first step of the final attack: a large DDoS attack designed to cover the tracks of a massive exfiltration operation of the content necessary to plan the final coordinated attack. The volume of data extracted is massive, and the data exfiltration goes unnoticed by overwhelmed security and Web support personal trying to keep applications available to legitimate users.

- Now, with network access, internal pivot points, known network topologies, administrator credentials and the relevant content data, the adversary has all of the information necessary and plans a swift, simultaneous penetration into the SCADA systems of the critical infrastructure operator, with penetration plans, scripts, and routines written and sequenced in advance. The adversary then uses these scripts to disable critical infrastructure and administrators are locked out of their own system.

This scenario is hypothetical. It is unlikely to happen as many of these systems are highly secured and the networks they run on are segregated; however, it illustrates how a BDA approach to cybersecurity might help to detect and intervene early to prevent, or at least mitigate, the effects of such an attack.

In the above case study, the sequence of events takes place over months, possibly even years. The data contained all of the warning signs existed in the data; therefore it is important to determine how BDA can be used to intervene early.

3.3.2 Identify, Mitigate, Create, and Collect Data

In this scenario, the first step in creating an effective security posture is to develop a better understanding of the network, baseline behavior on the network, the devices and systems on the network, and the structure of critical assets. During a cyber-related NS/EP event, the Government could apply relatively simple BDA and IDS methods to develop a baseline set of behaviors for risk assessment and other methods for improving the security posture for identification, detection, and response, including:

- Understanding devices by reviewing software through resources such as the National Vulnerability Database and understanding what communications protocols they use;

- Sandboxing new devices, collecting and analyzing traffic data before placing them in the network in order to understand the behavior of the device;

- Collecting ongoing data flows to be available for analysis and detection/response tools using a simple collector, such as Cisco NetFlow, allowing security practitioners to study flows on the network;

- Collecting personnel data to understand individuals that may be "high risk" for social engineering;

- Collecting information on email traffic from spam filters, potentially identifying a social engineering campaign; and

- Understanding network topology and behavior.

All of the above can be enabled by BDA tools. Collection tools such as network sensors, system data loggers, Structured Query Language (SQL), and not-only SQL (noSQL) stores can automate the collection of large volumes of relevant data to support my baseline security posture.

3.3.3 Protect, Prepare, Identify, and Make Data Available

After establishing the necessary infrastructure for data collection, it is critical to store and organize data in ways that make it useful to detect/analyze and respond/analyze output functions. In order to ensure consistency and the usefulness of data, data should be described in a consistent manner to achieve semantic interoperability, and thus made readily available for consumption by BDA toolsets referred to as service-oriented architectures. Some examples of BDA-related issues that could be valuable for this purpose include:

- Ontologies and/or taxonomies that provide the semantic content to ensure the consistent, rigorous description of data and services that enables sharing of data and analysis, both internally and externally;

- Data sharing agreements and protocols that allow critical infrastructure operators to quickly share relevant information;

- Distributed data storage that scales quickly and cheaply, and also supports archiving of data for future research;

- Distributed data processing that allows for the rapid provisioning of data to visualization and analytical toolsets;

- Data services that provide both direct query and pre-configured data feeds to security personnel;

- Extract, Transform, Load (ETL) toolkits that facilitate rapid data transformation for use by both technical and non-technical personnel;[62]

[62] Extract, Transform, Load (ETL) toolkits deliver solutions for the most time- and labor-intensive portion of data warehousing-data staging, or the ETL process. These toolkits also delineate best practices for extracting data from scattered sources, removing redundant and inaccurate data, transforming the remaining data into correctly formatted data structures, and then loading the end product into the data warehouse.

- Analytics toolkits, provided as standard services, that allow for rapid analysis of consistently described data services; and

- Visualization toolkits that support scalable, intuitive visualizations of complex data.

The result is a cyber "data lake" providing security personnel (e.g., researchers and analysts) a broader, more rapid vision into the state of the network.

3.3.4 Detect and Analyze

NIST defines the detect function as "develop[ing] and implement[ing] the appropriate activities to identify the occurrence of a cybersecurity event…to enable timely discovery of cybersecurity events." Examples of outcome categories within this function include: anomalies and events; security continuous monitoring; and detection processes.[63] The NIST Cybersecurity Framework recommends a variety of controls in this space, including: baselining network operations and expected data flows for users; detecting anomalies from the baseline; analyzing detected events to understand attack targets and methods; aggregating event data and correlating data from multiple data sources; and continually monitoring the network to detect potential cybersecurity events.

A DDoS attack is an effective means for compromising networks and disguising other malicious activity. In this case study, early detection and intervention for the DDoS attack would have left the exfiltration traffic exposed and more easily detected, allowing security personnel to block access and remediate the compromised systems before they can be used in the next phase of attack.

BDA can support each of these elements. In this case study, it is possible to baseline network traffic and machine interaction patterns through the use of streaming visualizations, graph analysis, and volumetric data. A baseline flow of traffic against which traffic anomalies can be compared through volumetric analysis on source and destination ports, Internet Protocol (IP) pairings, and on specific protocols (e.g., Transmission Control Protocol, User Datagram Protocol, Internet Control Message Protocol, and Simple Mail Transfer Protocol). Cybersecurity practitioners can also look for changes in flow, packet, and byte volumes looking for significant changes in activity of each parameter relative to expected values. In this case study, BDA toolsets are used in very specific ways to identify the early warning signs that an attack is being prepared or is underway:

- Log and analyze device communications.

- Study flows on the network using streaming analytics, graph analysis, and advanced visualizations to identify anomalous traffic flows through IP and port pairing patterns, such as devices talking to database servers, or clients serving large files on an unknown protocol.

- Employ user data and NLP to correlate suspicious email traffic to critical personnel to identify if a class is being targeted.

[63] NIST. *Framework for Improving Critical Infrastructure Cybersecurity*. February 12, 2014. http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf, page 8.

- Rapidly conduct statistical analysis of traffic patterns to isolate data exfiltration during DDoS or otherwise by discriminating by destination IP, port usage, or packet sequencing patterns.

- Process system and data logs using statistical and NLP tools to identify inconsistent access patterns (e.g., a database being accessed by an administrator who is not otherwise authenticated to other network services, such as email).

- Aggregate classical signature-based detections (e.g., IDS, IPS, next-generation firewalls, etc.) to look for meta-patterns within basic network security traffic.

- Identify anomalous scanning activity that could potentially be a reconnaissance activity preceding an attack or malicious exfiltration activity.

- Detect network scanning via fluctuations in traffic volume.[64]

- Deploy analysis methods that have been developed to correlate the activity of these malicious actors to likely control points and server hubs associated with botnets.[65]

3.3.5 Respond/Analytics Output Function

NIST defines the respond function as "develop[ing] and implement[ing] the appropriate activities to take action regarding a detected cybersecurity event."[66] There are a variety of response activities that can take place depending upon the nature of the DDoS attack. For example, a volumetric DDoS attack typically involves sending enough data packets to overwhelm a target's bandwidth or server capacity. To mitigate an attack, the IP addresses belonging to the hosts sending the data packets must be filtered and black listed, separating malicious traffic from legitimate traffic. In this case, multiple IP addresses must be identified in order to fully mitigate the attack. Other information, such as network flow data, is commonly combined with other sources of security intelligence information regarding botnets, sources of spam, Web threats, and other activities on the Internet. Combining this information to stop attack underscores the power information sharing can have in improving the Nation's cybersecurity posture.

In this case study, BDA would have allowed one to identify the compromised devices early and either sandbox, patch, or disable the compromised hardware. It may have provided the ability to identify phishing patterns and warn key personnel, putting them on a state of heightened alert and reducing the risk of disclosure. It also could have enabled the ability to identify anomalous traffic patterns inside the network, allowing entities to intervene by changing out network topologies to protect critical assets from compromised assets. BDA also has the potential to identify activities in progress (e.g., DDoS-covered exfiltration) to intervene and terminate early, and to remediate before the information can be used in an additional attack.

---

64 Increased activity on a given port or protocol can be indicative of a new exploit in use, exfiltration, or compromised hardware or software that facilitates malicious acts such as a spamming campaign or a DDoS attack.

65 Based on the sources of malicious activity (i.e., scanning for exploits and spam activity), specific information about the botnet such as Internet Protocol addresses and associated domain names used can help detect and block malicious activity.

66 NIST. *Framework for Improving Critical Infrastructure Cybersecurity*. February 12, 2014. http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf.

This case study illustrates an example of how BDA can be used in each of the various phases of the cyber attack lifecycle to identify and mitigate attacks. The lifecycle for an advanced persistent threat (APT) type of malware typically involves the attacker: (1) conducting network reconnaissance identify points of access or penetration; (2) developing custom software to inject the malware; (3) conducting intrusion attempts based on information obtained from spear phishing attacks, to spread the malware and to gain access; (4) establishing communications between the infected network device the attacker; and (5) monitoring activity and extracting data. More effectively mining and correlating NetFlow events using BDA techniques can help identify reconnaissance activity, intrusions and exfiltration events. For example, flow data identifying a communications link between an entity and a foreign third party may be indicative of an actual exfiltration event. These lifecycle activities provide points of reference for the creation of actionable threat intelligence. BDA may hold the most promise for cybersecurity in regards to threat intelligence that captures the tools, tactics, techniques and procedures used by attackers.

There are different categories of threat intelligence. The most basic intelligence is information about adversaries, tools or techniques and applying this data to incoming data in order to identify malicious activity. This is largely technical information such as an IP address, domain name, Uniform Resource Locator (URL), email address, file hash, Secure Socket Layer certificate identifier, Autonomous System Number, service provider organization, operating system registry entries, and other data that could be used to support the detect function. An elevated category consists of more contextual data, which can be used used to provide the technical data with context and enable security professionals to make decisions on actions they can take to remediate the threat. Examples of contextual information may include the attacker group, botnet, threat advisories, exploit examples, attack trends, activity patterns, named actors, and other data points supporting the respond function. The final category of intelligence is the tools, tactics, techniques, and procedures used by attackers as well as operations in progress. Focus on this final category is gleaned in part from the first two categories and is where BDA may hold the most promise for cybersecurity. By combining and analyzing all of these categories of information, there is an opportunity to develop threat intelligence that falls into this third category which may enable the ability to better understand threat actors and get in front of attacks.

3.3.6 Recover/Data Storage and Disposal

Assuming successful discovery and intervention (or even failure to detect and intervene in a timely manner), it is critical that data be stored and shared for research purposes. This data can lead to the next round of attack signatures, training materials, identification of adversaries and compromised hosts, as well as recommendations for improved network engineering to make networks resistant to future attacks.

3.3.7 BDA Lifecycle Analysis for a Cyber Attack on Critical Infrastructure

The table below illustrates how the cyber attack on critical infrastructure case study can be mapped to the BDA lifecycle.

LIFECYCLE STEP	SUMMARY	FINDINGS
I. Data Creation and Collection	Companies and cybersecurity vendors determine what datasets, such as network flow data, are available and can provide insights into cyber attacks.	Supporting the development of common ontologies and taxonomies to support semantic interoperability would make data sharing frameworks valuable.
II. Data Identification, Availability, and Quality	Cybersecurity professionals develop systems that create baselines against which, anomalies that may be indicative of a cyber attack can be detected. Gaining visibility across the entire attack lifecycle will help ensure the best results for data analysis.	A critical aspect of identifying threats is the sharing of information within industry so that a common operating picture begins to emerge which can aid in the overall detection and then mitigation of threats. This requires both common semantics and syntax as well as legal and organizational frameworks for sharing.
III. Data Analytics	When an event occurs, cybersecurity professionals use data to detect and respond to events.	BDA can be used to detect threats but a key element, as noted above, is the availability and standardization of the data so that a common operating picture can emerge across industry. Further, there is a shortage of qualified personnel who have the high technical skills needed to operate the available BDA tools.
IV. Data Analytics Output	Output will aid cybersecurity professionals in detecting threats. For example, in this use case using volumetric data to detect a DDoS attack and planning response activities (mitigating the attack and restoring service to a normal state in the case of a DDoS attack).	BDA has the ability to enhance response and mitigate cyber threats. Cybersecurity response plans should consider the role of information sharing and BDA to aid in those activities including cross-sector information sharing.
V. Data Analytics Storage and Disposal	Data used to detect the threats may be routing network flow information. Any unique information specific to the threat as part of forensics may be used to develop a threat signature or discarded depending upon circumstances.	There are multiple forms of cybersecurity information sharing, including private to private, private to Government, etc. To the extent a threat rises to the level where Government engagement is required, then minimization processes should be implemented to ensure an appropriate balance with privacy.

Table 3.5. The cyber attack on critical infrastructure use case mapped against the BDA lifecycle.

3.3.8 Findings

The threat landscape is large and constantly changing. Since contemporary cybersecurity systems are largely based on manual rules and signatures, it can be challenging for security professionals to respond to threats in a timely manner. Furthermore, current, rule-based systems are designed to look at events in a limited time window and, as such, inhibit the ability to find higher order correlation between events captured across different systems, as shown in the case study. Additionally, rule-based systems can only catch known threats or threat signatures, while APTs by definition are often able to elude detection by mutating or obfuscating activity.

Leveraging BDA technologies provide a potential way for security professionals to keep up with cyber threats by reducing the amount of information that needs to be reviewed manually and train security systems to differentiate between malicious and benign anomalous activity. BDA allows for the integration of algorithms and statistical models with software systems that together accelerate the evaluation and visualization data in order to detect interesting patterns. A large ecosystem exists of software and BDA products that are being developed by open source, academic, and commercial organizations which are intended to support analyzing and visualizing big data. However, in order to successfully implement BDA within the cybersecurity domain, several challenges need to be addressed, including:

- **Exporting interoperable data:** The first challenge is ensuring that data is collected in a consistent manner as part of the existing cybersecurity infrastructure. Aspects as simple as describing where in the network topology data was collected, or how the terms "client" and "server" are being used in documentation hamstring this effort. The current vector of BDA development is built upon a model where users own their data. Therefore the ability to share consistent data is critical for leveraging open source tools and combining multiple datasets. To this end, supporting the creation of a common ontologies and taxonomies for describing data in the cyber domain is extremely valuable.

- **Limited data fidelity and availability:** Another key challenge is to improve network data collection and structural ability to share, including packet-level (on-demand), NetFlow, and domain name systems logs/traffic, to improve cyber defenses. For instance, communication companies today typically collect only sample data events across their networks because it is cost prohibitive to capture the entirety of the relevant data from an infrastructure and storage prospective. Additionally, there are customer privacy concerns associated with such a broad collection of data. Finally, there are security concerns in collecting and sharing this data, such as exposing network topologies. While samples and simulations may have been effective in the past, industry is finding these techniques limit the usefulness of the data for more sophisticated cybersecurity investigations and analytics. The escalation of the cyber threat far outpaces the investments needed to fund the upgrades necessary to improve data capture across large-scale carrier networks. To effectively harness BDA, there is a need to create a legal and regulatory environment that will facilitate, as well as protect, and encourage information sharing.

- **Collaboration:** A similar challenge is associated with collaboration among owners of similar or comparable data. As with some of the data fidelity issues, there will also be customer privacy concerns and organizational security concerns associated with data sharing or collaboration that will need to be addressed. There is a need to create legal and regulatory environments that facilitate, protect, and encourage sharing.

- **Business and infrastructure incentives:** Another challenge is building the infrastructure to apply BDA to cybersecurity. Processing large amounts of data requires a lot of physical infrastructure, this is especially true for any system that will operate in real-time. Additionally, because so much sensitive data will be concentrated on these servers, fault tolerance and security are both concerns. The architecture for any production system will need to have multiple geographic locations, access to the internet backbone, high fault tolerance, and are generally designed to be single purpose. Costs associated with developing BDA infrastructure for cyber security can be addressed through national investment programs and tax incentives to bridge these gaps. In addition, incentives provided to the private sector can include protection from legal and business liabilities, such as through the use of a "Good Samaritan" Framework. Notably, Congress recently passed the *Cybersecurity Information Sharing Act of 2015*, which is intended to facilitate greater information sharing both within the private sector and between the private sector and Government. The legislation is a good first step to building a collaborative big data environment.

- **Workforce issues:** Despite the costs involved, the hardware challenges are not as great as the challenges of finding and attracting great talent. Ideal candidates should have domain knowledge from cybersecurity as well as knowledge of big data tooling, advance mathematics, statistics, and problem solving. Since this is a new field, few people have all the desired skills and expertise; it is especially rare to find people with both cybersecurity and data science backgrounds. Workshops, conferences, and professional development can assist in building a community of talented individuals who understand the same lexicon and have familiarity with the same tools. Nonetheless, there are still too few individuals to meet the current and projected needs. Addressing this data analytics gap will likely require a workforce development strategy comparable to the current effort to address cybersecurity workforce gaps.

3.4 General Themes Within Use Cases

There are several important topics related to BDA that have emerged in examples discussed in Section 2.2 and the NS/EP use cases described in Section 3.0.

3.4.1 Privacy

Data analytics provides tremendous opportunities to derive insight from combining different data fields from individuals, systems, and events to provide a prediction or a situational awareness for an emerging situation. While this may be very helpful within the context of NS/EP, BDA providers must balance this objective with protecting American's civil liberties and Constitutional rights to privacy.

A 2014 White House report, entitled *Big Data: Seizing Opportunities, Preserving Value*s, states that, "[b]ig data analytics have the potential to eclipse longstanding civil rights protections in how personal information is used in housing, credit, employment, health, education and the marketplace."[67] This warning must be heeded when using data analytics to support the Government's NS/EP functions; however, ignoring the potential value of BDA in this context

[67] EOP. *Big Data: Seizing Opportunities, Preserving Values*. May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

would be a mistake. Through its study of BDA, the NSTAC has found that a balanced approach, while challenging, is achievable.

In the United States, specific rules govern privacy in certain sectors, including the *Children's Online Privacy Protection Act* (1998), the *Financial Services Modernization Act* or *Gramm-Leach-Bliley Act* (1999), the *Fair Credit Reporting Act* (1970), and the *Health Insurance Portability and Accounting Act* (2013). More generally, Government and industry have examined data privacy through the Fair Information Practice Principles framework, which includes criteria related to: transparency, individual participation, purpose specification, information minimization, use limitation, information quality, security and accountability.[68] Some of these points could be challenging in a big data environment that is built around machine learning and the integration of multiple data streams.

In order to derive the value from BDA while preserving civil liberties, a provider of BDA for NS/EP should consider the implications for individual privacy when developing any potential new product or service. For example, data collectors and aggregators should respect the context in which the data was collected when drawing conclusions from the analysis performed. In doing so, they should ensure that any error rates are distributed evenly across populations and not concentrated in any specific demographic to prevent the perception that individuals are being targeted based on vulnerability or protected class (i.e., minors, health status). Moreover, data collectors and aggregators should also follow cultural norms when determining what information is acceptable to gather from users.

3.4.2 BDA Education

The key to addressing emerging big data problems is the transformation of BDA from a niche capability to a core competency within Government, academia, and industry. This transformation begins with addressing the BDA skills gap. Many believe that there is a shortage of qualified personnel available to apply to emerging and existing BDA problems. The McKinsey Global Institute estimates a shortage of 140,000 to 190,000 big data scientists by 2018.[69] McKinsey also asserts a shortage of managers and analysts capable of operating companies based on insights from big data.[70]

A big data project will produce a system that leverages software, computers, data science, data, and people to address a set of problems. These projects are typically executed by a multidisciplinary team comprised of managers, architects, programmers, data scientists, technicians, business area experts, and others. Big data problems require atypical solutions; therefore, the big data project team and the resulting project execution will require skills and approaches that are different from traditional projects. It is not sufficient to simply augment project staff with data scientists; rather, specific BDA-related training is required for all staff.

---

[68] Federal Trade Commission. *Fair Information Practice Principles Framework*. March 31, 2009. https://web.archive.org/web/20090331134113/http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

[69] McKinsey Global Institute. Big Data: *The Next Frontier for Innovation, Competition, and Productivity*. May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

[70] Ibid.

As one commentator asserted, "[t]he looming issue in big data isn't technology, but the privacy and ethics decisions associated with how, when, and if results should be provided."[71] The growth of big data and its associated ethics concerns (e.g., privacy, legal, etc.) have created an environment where data ethicists have become a necessity on big data projects. In order to promote public trust, the data ethicist provides transparency to designated overseers, such as Government regulators, industry ethics boards, and others as required. In many instances these overseeing groups do not exist, therefore, the data ethicist must use a surrogate similar to a corporate ombudsman.

Data ethics is an interdisciplinary field that creates best practices for data use, ensures that these practices can be implemented, and provides mechanisms for independent verification. In practice, data ethics addresses a blend of legal, technical, social, and business concerns.

Many if not most organizations face impediments to executing a successful and timely big data project. For instance, one big data marketing expert identifies the top five big data challenges as:

1. Finding and retaining skilled staff;

2. Time from "proof-of-concept" to deployment;

3. Rapidly changing big data technologies;

4. Reduction of realized return on investment due to inaccurate project scoping; and,

5. Understanding the security/privacy requirements and veracity of eclectic datasets.[72]

These challenges when viewed through a skills gap lens reveal not only shortages in data science expertise, but also reinforce the conclusion that organizational skill deficiencies exist across all disciplines with respect to large data projects. The fifth item on the above list reinforces the need for more deliberate attention to data ethics. These challenges also reflect a new domain whose body of knowledge, as well as the requirements dictating the necessary personnel and processes needed to support it, are still evolving.

BDA is a young discipline whose evolutionary path parallels other disciplines such as computer science. Previously, there was a gap in computer science professionals to address the explosion in software development requirements. Academia responded by producing more and more computer science graduates. Concurrently, and as importantly, the software ecosystem evolved. The computer science field now includes a variety of specialties, including software engineering, programming languages, visualization, and databases, among others. Similarly, BDA is an emerging discipline that calls for academia to respond by producing specialized BDA curriculum and graduates.

Big data success stories are often the result of the actions of a few extremely creative people, some of whom may not have been trained in a science, technology, engineering, and

[71] Eric Lundquist. "Why You'll Need a Big Data Ethics Expert." *InformationWeek*. January 3, 2013. http://www.informationweek.com/it-leadership/why-youll-need-a-big-data-ethics-expert/d/d-id/1108001.

[72] John Haddad. "Top 5 Big Data Challenges." *Informatica*. July 10, 2014. http://blogs.informatica.com/2014/07/10/top-5-big-data-challenges/#fbid=uOMCiPkWhS8.

mathematics (STEM) field. In addition to being intuitive problem solvers, data scientists who have a rich domain understanding, a solid math/statistics background, and the computer science skills of programming and data structuring are needed to promote the success of big data projects. While there will never be a substitute for smart, intuitive people, the evolution of the BDA ecosystem will demand many contributors.

This evolution is fueled by education and training. All participants in a big data project need to recognize the differences between a "normal" and a "big data" software development project. It is this lack of appreciation that leads to poor project scoping, delays in the production timeline, vacillation on platform selection, and the unfortunate revelation that privacy protections are inadequate.

In general, development projects are divided into the following concurrent activities:

- **Management:** Budget, staff, control, and plan;
- **Engineering:** Specify, conceptualize, architect, design, development, test, deploy, and operate;
- **Quality Assurance (QA):** Prevent management and engineering mistakes and system defects; and
- **Information Assurance (IA):** Manage information-related risks by protecting and defending the system.

Both QA and IA necessarily are part of an independent authority structure rather than management and engineering. This is to prevent conflicts of interest due to schedule, cost, and implementation unduly influencing QA and IA.

For big data projects, there is a heightened responsibility for data privacy protection, the ethical use of data, and the legitimate access to data. This responsibility falls under the auspices of the data ethicist and is referred to as data assurance (DA). DA is a necessary additional formal activity for big data projects. Like QA and IA, DA should be under a separate line of authority than management and engineering. Unlike QA an IA, no formal models or doctrine for data assurance have been developed.

The state of data assurance for many organizations "can best be classified as abysmal."[73] Often neglected DA activities include: establishing formal data stewardship; obtaining formal agreements to access and release data as required; producing a data use plan and policies; and providing a mechanism for independent validation of data use.

As with QA and IA, DA should be part of the engineering and engineering management curriculums for all disciplines. Specialized, detailed instruction for data ethicists is required to produce a generation of data assurance advocates and practitioners. For big data projects, often the conceptualization phase is longer to account for a need for proof-of-concept evaluation and prototyping. Additional time is required in the architecting phase due to the eclectic set of data

[73] David Marco. "Data Assurance Road Map, Part 1." *InformationManagement.com*. December 1, 2014. http://www.information-management.com/issues/20041201/1014518-1.html.

sources, computing techniques, and infrastructure options. According to many in the field, the most underappreciated activity is data curation. Data curation is the collection of activities to: clean data (e.g., remove errors, correct format, etc.); assess suitability (i.e., quality versus use); assess veracity; disambiguate; remove duplication; and identify missing data.

According to NIST, "BDA projects require the coordinated efforts of all team members across all disciplines."[74] The NIST Big Data Reference Architecture (NBDRA), shown in Figure 3.1, represents a big data system comprised of five logical functional components connected by interoperability interfaces (i.e., services).

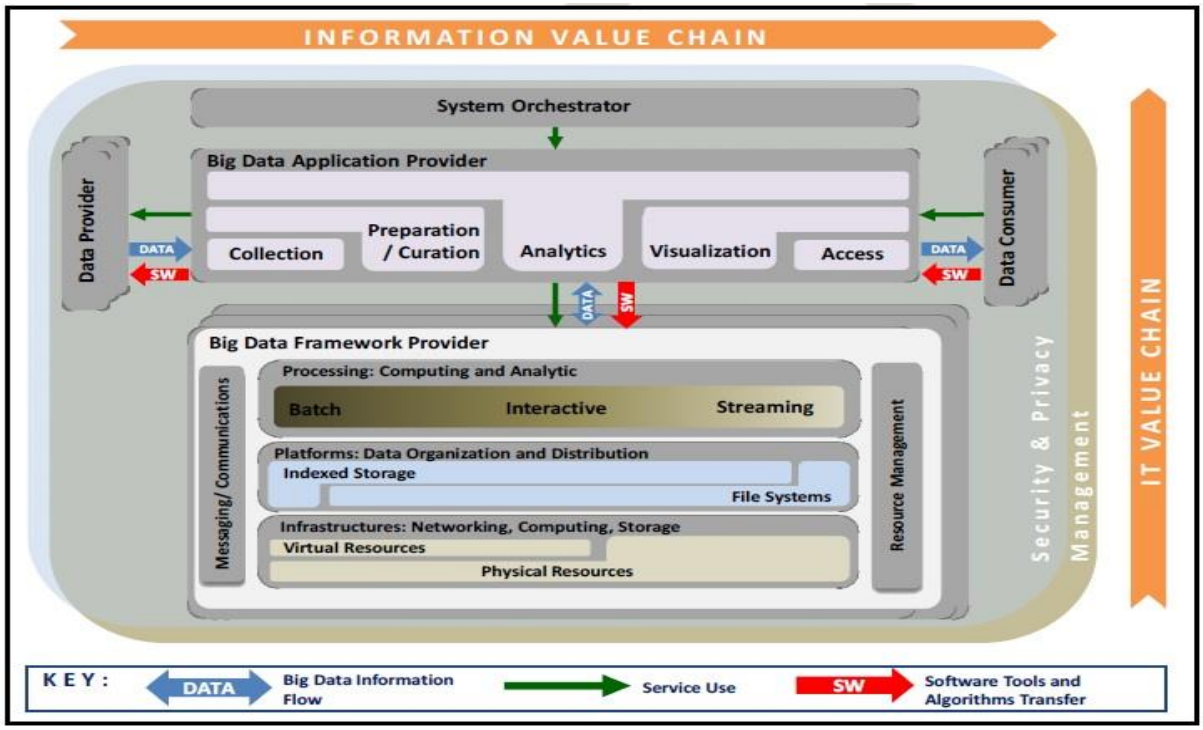


Figure 3.1. The NBDRA: a vendor-neutral, technology- and infrastructure-agnostic, conceptual model of big data architecture.

The Information Value Chain and the IT Value Chain envelop the components, representing the interwoven nature of management, security, and privacy with all five of the components. The NBDRA is intended to "enable system engineers, data scientists, software developers, data architects, and senior decision makers to develop solutions to issues that require diverse approaches due to convergence of big data characteristics within an interoperable big data ecosystem."[75]

Mr. John Eberhardt, a professor at George Mason University and partner at 3E Services, LLC, represents the NBDRA engineering approach as shown in Figure 3.2. This approach highlights specific considerations for each engineering phase (e.g., concept, architect, design, implement, test, deploy, and operate).

[74] NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

[75] Ibid.

Mr. Eberhardt's engineering approach to big data demonstrates how this pattern is repeated at further levels of detail as the project advances from concept to implementation. Most traditional development processes do not adequately address the data curation.[76]

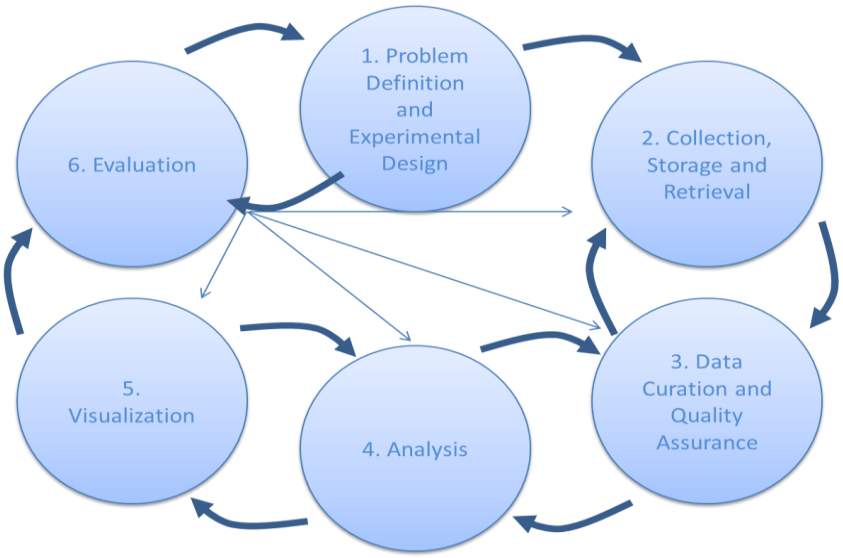


Figure 3.2. Mr. Eberhardt's engineering approach to big data problems.

According to Mr. Eberhardt, the mapping of project roles to project activities is as shown in Figure 3.3. His mapping shows the need for cross-training for the various disciplines, as well as shows the advantage of a diverse team. The mapping also calls out the ethicist's role for supporting the data security, privacy, and ethics activities.[77]

Project Activity \ Project Role	SME/User	Analyst	Statistician	DBA	GUI Designer	Program Manager	Ethicist	Architect	Developer
Define the Problem	■	■				■			
Experimental Design	■	■	■			▒	▒	▒	
Data Collection, Storage, and Retrieval			▒	■	▒			■	■
Data Curation and Quality Assurance	■	■	■	■					■
Analysis	■	■	■	▒		▒	▒	■	■
Visualization and Reporting	■	■			■			■	■
Evaluation	▒	■	■			■			
Security	▒	▒		■	▒	■	■	■	■
Privacy/Ethics	▒	▒	▒		▒	■	■	▒	▒
Effort Initiation						■		▒	▒
System Design			▒	■	■			■	▒
Implementation		▒		■		■		▒	■

■ Principle Contributor ▒ Secondary Contributor □ Minor Contributor

Figure 3.3. Mr. Eberhardt's mapping of activities to roles for big data projects. Successful big data projects require a varied skill set among the participants.

[76] John Eberhardt, III. 3E Services, LLC. *Briefing to the NSTAC BDA Subcommittee*. December 18, 2015.
[77] Ibid.

3.4.3 Data Handling Best Practices

In a big data landscape, there are potentially dozens of mechanisms to directly or indirectly handle data. It becomes very important to consider all aspects of data handling within a BDA construct, not only due to the sheer amount of data that will be collected and processed, but also due to the complexity and amount of potential analytic operations.

When ingesting new data into a big data platform, it is important to understand as many attributes of the data as possible so that the data can be adequately stored and transformed as needed. Once its attributes are correctly identified, the data needs to be appropriately labeled (e.g., by source, sensitivity, etc.) in order to apply the proper protections. This includes information that is aggregated across multiple sources or results. As data is integrated and correlated, it can become difficult to track and maintain "originator control," that is, who owns which data elements, what are the access rights, and what are the handling and retention rules.

Data handling enables the sharing of data, which could range from raw data, transformed or enriched data, analytic results, or any combination of these. Information sharing, both within the Government, as well as between the Government and external partners, could be facilitated by the use of robust standards, such as the National Information Exchange Model. Moreover, data normalization is key to facilitating effective and accurate analytics. When combining information from a variety of sources, following standards will result in the common formatting of the data being shared (e.g., date format, timestamps, etc.).

There is a need to develop additional standards for sharing information across various domains of the public and private sector.

3.4.3 Data Security and Protection in the Big Data Landscape

In big data environments, there are implications to data security and protection that go beyond traditional needs. Due to the distributed nature of the data along with the myriad of methods for storing and querying information, it becomes important to understand all of the components and potential attack surfaces (i.e., in terms of availability, confidentiality, and integrity of the data and analytic results) that constitute any big data platform. Security by design is vital when building these types of systems. Specific security issues related to BDA include:

- Lack of security in depth, which can mean that one security flaw exposes all data;
- Algorithms that aggregate or fuse data often unintentionally reveal "secrets" of the underlying data; and
- Data provenance is often not rigorously verified; data with poor veracity is often incorrectly assigned improper levels of significance.

Organizations that do not adequately address security can damage reputations, negatively impact individuals and companies financially, and put valuable or proprietary information at risk. Alternatively, organizations that are too conservative with respect to security may not share any data that is needed by the appropriate NS/EP organizations. The responsible sharing of quality data is vital to improving capabilities, especially in the NS/EP domain.

The principle of least privilege states that every program and user of a system should operate using the least set of privileges necessary to complete the job. This principle limits the damage that can result from an accident or error, as well as reduces the number of potential interactions among privileged programs to prevent improper uses of privilege.

The principle of least privilege applies to a big data environment and should be adopted to help shape what data is collected, how it is stored, and for how long it is maintained. It is also important to be able to effectively audit access to data and analytics results, as well as the lineage of the data within the BDA platform. This concept extends to the need to continue to properly authenticate and authorize users, preferably following known methodologies (e.g., role-based access control). Full authentication, authorization, and audit systems not only help to protect data, but they also ensure a more stable and available big data platform.

From a data protection perspective, systems need to have the ability to filter, tokenize, and mask data as needed. Before the data is searched or retrieved, it can be labeled and filtered based on authorization policies. Tokenization of data elements is replacing sensitive information with benign values which contains no identifying information. Tokenization can become important for data sharing when there is a desire to protect certain fields or data elements. Masking, in the big data context, is a way to "black out" information that the viewer is not authorized to see.

To effectively secure big data and the system architecture used to analyze it, defense-in-depth techniques will also need to be deployed, such as encryption of data at rest and in motion; authentication and authorization of users accessing the system; role-based access controls; and complete monitoring and auditing. In addition, users must consider how the data is collected, how it can be used, how to transform it, what analytics should be applied, what the resulting analysis means, and the legal and ethical implications of the results.

3.4.4 Data Use and Ownership

Policies related to BDA should be more focused on how big datasets are used, rather than how they are collected. While the process of collecting data does not differentiate between legitimate and malicious purposes, the use of that data does. Instead of limiting the collection and retention of data, data owners should emphasize controlling data at the most important phase of the data lifecycle (i.e., the moment when the data is used).

4.0 CONCLUSIONS

The NSTAC has concluded that there are several challenges that need to be overcome, and standards that need to be set, before the Government can effectively maximize the benefits of using BDA in support of NS/EP functions. These include:

- A need to expand policies, plans, standards, and tools used for BDA that allow data to be utilized more readily during an NS/EP event.
    - In order to enhance the Government's NS/EP functions, the NIST Big Data Public Working Group's efforts could be expanded to facilitate policies that allow for the more widespread use of BDA by Federal, State, local, and non-governmental agencies.

  - There is a lack of common ontologies with regard to BDA used in the NS/EP community. The lack of a central lexicon results in poor communication and information sharing in this domain, especially when compared to other contexts where BDA ontologies are more mature and well-developed (e.g., the healthcare sector).

    - The NS/EP community would benefit from a collection of agreed upon formal names and definitions for the types, properties, and interrelationships between available and emerging datasets. Robust, community-driven standards will facilitate data sharing.

  - There is a need for additional infrastructure investment to provide continuity of operations for critical services during an NS/EP event.

    - The private sector owns and operates much of the Nation's critical infrastructure necessary to respond to an NS/EP event. Since the infrastructure can be overwhelmed during an NS/EP event, steps should be taken to mitigate the risk.

  - There is a need to conduct joint public-private sector exercises based on realistic NS/EP scenarios to regularly test procedures for accessing BDA and enhance coordinated response and recovery efforts. These exercises may reveal gaps in information required to appropriately and effectively respond to NS/EP events.

- A need to clarify the availability, access, handling, and protection of data in line with privacy and security best practices, as organizations may be reluctant to share data due to a variety of concerns (e.g., liability, loss of intellectual property, etc.).

  - During an NS/EP event, there is a need for rapid and perhaps unprecedented response. Important information may be available within data sources that were not included in pre-negotiated data sharing agreements. In addition, the rules governing the release and use of that data may be unclear, constrained by slow procedures, or may not exist at all.

  - Data sharing is more likely to occur if entities are not constrained by traditional risk models regarding liability, loss of intellectual property, and other competitive concerns. Companies that contribute to NS/EP event response should not face penalties for having done so.

- A need to address the BDA skills gap.

  - Since the Government and the private sector compete for the same limited BDA resources, the Government needs to encourage the private sector to educate and cross-train personnel and recruit appropriately trained graduates (e.g., BDA-literate data scientists and data ethicists).

  - Government big data problems are typically addressed by teams; therefore, big data cross-training for all STEM fields is necessary. The successful use of BDA requires key personnel during an NS/EP event to have a common understanding of the

problem, the solution being produced, and the responsible execution of the contract. Cross-training is required in several areas, including: (1) data curation; (2) data analytics; (3) architectures and solutions; and (4) data assurance.

- Rather than waiting for a new generation of data scientists, the existing STEM workforce can be cross-trained. Continuing education in data science for existing personnel can help alleviate the shortage of data scientists.

- Since BDA is an emerging capability, the practices for BDA in the execution of Government contracts are now being defined. The Government would benefit from implementing formal guidelines for defining, planning, using, and enforcing the appropriate and ethical use of big data.

5.0 RECOMMENDATIONS

After careful consideration and study, the NSTAC recommends that the President take the following actions to improve the use of BDA to enhance NS/EP capabilities.

To expand policies, plans, standards, and tools used for BDA that allow data to be utilized more readily during an NS/EP event, with the appropriate protections in place, the President should direct the appropriate Federal Government agencies to:

1. **Collaborate with the private sector to collect, manage, and make available common and adaptable NS/EP ontologies. This includes the use of standard labeling methods, other shareable components, and the development of robust, community-driven standards.**

2. **Study and recommend ways to improve the capacity and robustness of industry-provided services that are necessary for NS/EP capabilities.**

    - Since the private sector is an important partner in responding to NS/EP events, the Government should consider tax incentives for bolstering and building infrastructure for delivering critical services.

3. **Conduct proof-of-concept research and exercises that would be shared across an expanded range of Government agencies to elicit and integrate datasets in anticipation of possible NS/EP events.**

    - The Government should develop and execute exercises that test the efficacy of analytic approaches as new ones emerge and existing ones evolve. At the same time, the Government should develop and sustain technology pilots and proof-of-concept models, as well as share the results across a broader range of Federal agencies. This is designed to facilitate expanded organizational participation and standards development of shareable databases and applications.

- The Government should incentivize and provide test datasets for consumption in the public sector, private sector, and academic community.

In order to further clarify the availability, access, handling, and protection of data in line with industry's privacy and security best practices to most effectively capture the potential benefits of BDA, the President should direct the appropriate Federal Government agencies to:

4. **Develop a "Good Samaritan" Framework for exchanging information between the Government and consenting private entities during an NS/EP crisis.**

    - The framework should afford standard agreed upon protections to entities sharing data in good faith during an NS/EP event.

    - The development of this framework should be a collaborative effort between the appropriate public sector, private sector, security, and civil liberties stakeholders.

    - This framework should pre-establish general rules between the Government and the participating private sector organizations to define the appropriate use of data during an NS/EP event. Specifically, the "Good Samaritan" Framework should clarify rules regarding the protection of privacy, data use, ownership, storage, retention, accidental disclosure, and deletion.

To best minimize and effectively overcome the skills gap that currently exists in BDA, the President should direct the appropriate Federal Government agencies to:

5. **Identify data science and analytics as a key discipline limited by shortages in practitioners, educators, and graduate and undergraduate programs of study within the *Federal Science, Technology, Engineering, and Mathematics (STEM) Education 5-Year Strategic Plan* (2013).**

6. **Direct contracting agencies to add appropriate data science skill and training requirements for all applicable disciplines, as appropriate for Government contracts addressing big data problems.**

    - Data scientists and data ethicists should be identified as key personnel on big data contracts.

    - Contracting firms should develop a training program upon award of the contract and provide notification of this approach to industry to allow sufficient time for training and personnel acquisition.

    - Contracting firms should provide relevant personnel with training covering big data issues of security, privacy, ethics, provenance, and transparency for Government contracts addressing big data problems.

7. **Define and require data assurance plans and programs for Government big data contracts.**

    - When tasking the appropriate agency to define data assurance plans and programs, NIST's Special Publication 800 series of computer security publications could be used as a guide.

    - In order to address the need for expanded data assurance and to ensure that big data is used ethically, the Government should train personnel on the appropriate use of and inherent privacy risks associated with big data.

APPENDIX A: MEMBERSHIP

SUBCOMMITTEE MEMBERS

Mr. William Brown, Harris Corporation and Subcommittee Co-Chair
Ms. Lisa Hook, Neustar, Incorporated, and Subcommittee Co-Chair

Ms. Terri Claffey, Neustar, Incorporated, and BDA Working Group Co-Chair
Mr. Christopher Metts, Harris Corporation and BDA Working Group Co-Chair

Apple, Incorporated	Mr. Walter Kuhn
AT&T, Incorporated	Mr. Elliott Battles Mr. Christopher Boyer Ms. Rosemary Leffler
CenturyLink, Incorporated	Mr. Lucas Budman Ms. Kathryn Condello Mr. Greg Petropoulos
Ciena Corporation	Mr. Robert Tomkins
Communication Technologies, Incorporated	Mr. Milan Vlajnic
CSRA, Incorporated	Mr. Guy Copeland
Department of Commerce (DOC), National Telecommunications and Information Administration (NTIA)	Ms. Regina Harrison
Department of Homeland Security (DHS)	Dr. Bahador Ghahramani Mr. Gabriel Martinez Dr. Paul Ngo
Ericsson, Incorporated	Mr. Stephen Hayes Ms. Louise Tucker
Executive Office of the President (EOP), National Security Council (NSC)	Mr. Daniel Prieto
FireEye, Incorporated	Mr. Travis Rosiek

Harris Corporation	Mr. Christopher Collings Mr. John Farrell Mr. M. Edwin Johnson Dr. Dennis Martinez Ms. Christina Steele
Intel Corporation	Mr. Patrick Flynn Mr. Kent Landfield Mr. Alan Ross Mr. Brian Willis
Microsoft Corporation	Mr. Erin English Mr. Christopher Krebs
Neustar, Incorporated	Mr. Ken Inman Mr. Rodney Joffe Mr. Michael Spencer
Palo Alto Networks, Incorporated	Mr. William Gravell
Raytheon Company	Mr. Michael Daly Mr. Brett Scarborough
Symantec Corporation	Mr. Jeffrey Greene
Verizon Communications, Incorporated	Mr. Michael Woods

BRIEFERS – SUBJECT MATTER EXPERTS

3E Services, LLC	Mr. John Eberhardt, III
American Red Cross	Ms. Wendy Harman
Apple, Incorporated	Mr. Erik Neuenschwander Dr. Guy Tribble
Berkeley Center for Law and Technology	Mr. James Dempsey
Center for Democracy and Technology	Ms. Alethea Lange
Center for Open Data Enterprise	Mr. Joel Gurin
Centers for Disease Control and Prevention	Ms. Angela Schwartz

City of Boston Police Department	Commissioner William Evans
DOC, National Institute of Standards and Technology	Mr. Wo Chang Mr. James St. Pierre
DOC, National Oceanic and Atmospheric Administration	Mr. David Michaud
DOC, NTIA, First Responder Network Authority	Mr. TJ Kennedy Mr. Edward Parkinson
DHS, Federal Emergency Management Agency	Mr. Christopher Shoup
DHS, National Protection and Programs Directorate, Office of Cybersecurity and Communications	Deputy Under Secretary Phyllis Schneck
DHS, Office of the Secretary	Mr. Peter Verga
DHS, Science and Technology Directorate	Mr. Stephen Dennis
Department of Justice, Federal Bureau of Investigation	Mr. Dale Killinger
Dun and Bradstreet Corporation	Dr. Anthony Scriffignano
EOP, NSC	Dr. Christopher Kirchhoff
FireEye, Incorporated	Mr. Peter Silberman
General Electric Corporation	Mr. Richard Puckett
George Mason University	Dr. Kirk Borne
George Washington University	Mr. Orin Kerr
Institute for Defense Analyses	Dr. Arun Maiya
Intel Corporation	Dr. Mark Seager
Massachusetts Institute of Technology	Dr. Alexander Pentland

Michigan State University	Dr. Arun Ross
Microsoft Corporation	Ms. Harmony Mabrey
National Aeronautics and Space Administration	Dr. W. Phillip Webster
National Emergency Number Association	Mr. Trey Forgety
National Science Foundation	Dr. Chaitanya Baru
National Voluntary Organizations Active in Disaster	Mr. James McGowan
Networking and Information Technology Research and Development Program	Dr. Keith Marzullo Ms. Wendy Wigen
Pennsylvania State University	Dr. Adam Smith
Purdue University	Dr. Ninghui Li
Sandia National Laboratories	Mr. Curtis Keliiaa
Splunk, Incorporated	Mr. Stephen Sorkin
University of Maryland	Mr. Frank Pasquale
Verizon Communications, Incorporated	Mr. Andrew Bonillo

SUBCOMMITTEE MANAGEMENT

NSTAC Designated Federal Officer (DFO)	Ms. Helen Jackson
Alternate NSTAC DFO	Ms. Deirdre Gallop-Anderson
Booz Allen Hamilton, Incorporated	Ms. Ursula Arno Ms. Laura Karnas Mr. Ryan Sheehy
Total Systems Technologies Corporation	Mr. Jordon Birman Ms. Sheila Sengupta

APPENDIX B: ACRONYMS

ACLU	American Civil Liberties Union
API	Application Programming Interface
APT	Advanced Persistent Threat
BDA	Big Data Analytics
BPD	City of Boston Police Department
DA	Data Assurance
DDoS	Distributed Denial of Service
EO	Executive Order
EOP	Executive Office of the President
ETL	Extract, Transform, Load
FDA	U.S. Food and Drug Administration
FEMA	Federal Emergency Management Agency
GPS	Global Positioning System
IA	Information Assurance
IaaS	Infrastructure-as-a-Service
IDS	Intrusion Detection Systems
IoT	Internet of Things
IP	Internet Protocol
IPS	Intrusion Prevention Systems
IT	Information Technology
NASA	National Aeronautics and Space Administration
NBDRA	NIST Big Data Reference Architecture
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NOAA	National Oceanic and Atmospheric Administration
noSQL	Not-Only Structured Query Language
NS/EP	National Security and Emergency Preparedness
NSTAC	President's National Security Telecommunications Advisory Committee
OT	Operational Technology
PaaS	Platform-as-a-Service
PII	Personally Identifiable Information
PSAP	Public Safety Answering Point
SaaS	Software-as-a-Service
SCADA	Supervisory Control and Data Acquisition Systems
SIRS	System Inflammatory Response Syndrome
SME	Subject Matter Expert
SQL	Structured Query Language
STEM	Science, Technology, Engineering, and Mathematics Education
QA	Quality Assurance
URL	Uniform Resource Locator

APPENDIX C: GLOSSARY

Advanced Persistent Threat (APT): An adversary that possesses sophisticated levels of expertise and significant resources which allow it to create opportunities to achieve its objectives by using multiple attack vectors (e.g., cyber, physical, and deception). These objectives typically include establishing and extending footholds within the information technology infrastructure of the targeted organizations for purposes of extracting information, undermining or impeding critical aspects of a mission, program, or organization; or positioning itself to carry out these objectives in the future. The APT: (1) pursues its objectives repeatedly over an extended period of time; (2) adapts to defenders' efforts to resist it; and (3) is determined to maintain the level of interaction needed to execute its objectives. (National Institute of Standards and Technology (NIST) Special Publication (SP) 800-39)

Adversary: Any individual, group, organization, or Government that conducts or has the intent to conduct detrimental activities. (NIST SP 800-30)

Application Programming Interface (API): A set of definitions of the ways one piece of computer software communicates with another. API is a method of achieving abstraction, usually (but not necessarily) between higher-level and lower-level software. (*The Data Reference Model*, Federal Enterprise Architecture Program, Office of Management and Budget)

Attack Vector: An avenue or tool that a threat uses in order to gain access to a device, system, or network in order to launch attacks, gather information, or deliver/leave a malicious item or items in those devices, systems, or networks. (Sandia National Laboratories Report 2012-2427)

Autonomous System Number: The number given to one or more routers under a single administration operating the same routing policy. (NIST SP 800-54, Adapted)

Big Data: Extensive datasets – primarily in the characteristics of volume, variety, velocity, and/or variability – that require a scalable architecture for efficient storage, manipulation, and analysis. Big data also refers to a dataset whose size is beyond the ability of typical database software tools to capture, store, manage and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data. As technology advances over time, the size of datasets that qualify as big data will also increase. The definition can also vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. Big data can range from a few dozen terabytes to multiple petabytes. (NIST SP 1500-1, McKinsey Global Institute)

Big Data Analytics: The synthesis of knowledge from information contained in big datasets. (NIST SP 1500-1)

Cloud Computing: A model for enabling on-demand network access to a shared pool of configurable information technology capabilities/resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. It allows users to access technology-based services from the network cloud without knowledge of, expertise with, or control over the

technology infrastructure that supports them. Both the user's data and essential security services may reside in and be managed within the network cloud. (Committee on National Security Systems Instruction (CNSSI) 4009, Adapted)

Critical Infrastructure: System and assets, whether physical or virtual, so vital to the United States that the incapacity or destruction of such systems and assets would have a debilitating impact on security, national economic security, national public health or safety, or any combination of those matters. Critical infrastructure can be owned and operated by both the public and private sector. [*Critical Infrastructures Protection Act of 2001*, 42 U.S.C. 5195c(e)] (CNSSI 4009, Adapted)

Cyber Attack: An attack, via cyberspace, targeting an enterprise's use of cyberspace for the purpose of disrupting, disabling, destroying, or maliciously controlling a computing environment/infrastructure; or destroying the integrity of the data or stealing controlled information. (CNSSI 4009)

Cybersecurity: The ability to protect or defend the use of cyberspace from cyber attacks. (CNSSI 4009)

Data Assurance: The promotion of data privacy protection, the ethical use of data, and legitimate access to data.

Data Ethics: Norms for conduct that distinguish between acceptable and unacceptable behavior in the collection, analysis, and usage of data.

Data Ethicist: One whose judgment on the ethics of data usage has come to be trusted by a specific community, and is expressed in some way that makes it possible for others to mimic or approximate that judgment.

Data Exhaust: Data generated as trails or information byproducts resulting from all digital or online activities. These consist of storable choices, actions and preferences such as log files, cookies, temporary files and even information that is generated for every process or transaction done digitally. (https://www.techopedia.com/definition/30319/data-exhaust)

Data Filtering: A wide range of strategies or solutions for refining datasets to meet the needs of the user without including repetitive, irrelevant, or even sensitive data.

Data Handling: The set of activities around viewing, using, modifying, or deleting data.

Data Lifecycle: The complete set of phases through which data passes, including (1) data creation and collection; (2) data identification, availability, and quality assurance; (3) data analytics; (4) use of analytic outputs; and (5) data management, storage, retention, and disposal.

Data Masking: The process of systematically removing a field or replacing it with a value in a way that does not preserve the analytic utility of the value, such as replacing a phone number with asterisks or a randomly generated pseudonym. (NIST Internal/Interagency Reports (NISTIR) 8053)

Data Science: The extraction of actionable knowledge directly from data through the process of discovery, or hypothesis formulation and hypothesis testing. (NIST SP 1500-1)

Data Scientist: A practitioner who has sufficient knowledge in the overlapping regimes of business needs, domain knowledge, analytical skills, and software and systems engineering to manage the end-to-end data processes in the data lifecycle. (NIST SP 1500-1)

Data Tokenization: A process by which a sensitive data element is replaced with a surrogate value called a "token." (Payment Card Industry Data Security Standard)

Denial of Service Attacks: The prevention of authorized access to resources or the delaying of time-critical operations. Time-critical may be milliseconds or it may be hours, depending upon the service provided. (CNSSI 4009)

Distributed Denial of Service Attacks: A denial of service technique that uses numerous hosts to perform the attack and prevents the authorized access to resources or delays time-critical operations. (NIST Glossary of Information Security Terms – NISTIR 7298 – Revision 2)

Domain Name Systems (DNS): A mechanism used in the internet and on private intranets for translating names of host computers into addresses. DNS allows host computers not directly on the Internet to have registered names in the same style. (Newton's Telecom Dictionary)

"Good Samaritan" Framework: For the purposes of this report, the NSTAC proposes that the Government develop an information sharing framework that will take specific steps to encourage private entities to share data for NS/EP response, while also limiting liability (e.g., risks of civil and criminal lawsuits, loss of intellectual property, and potential loss of competitive advantage). Therefore, the Government should expand the concept of a "Good Samaritan" Framework to encompass a broad set of protections, including those related to privacy, data use, ownership, storage, retention, accidental disclosure, and deletion.

Hurricane Hunters: Airplanes that are operationalized in support of the National Oceanic and Atmospheric Administration's Aircraft Operations Center. These planes are flown across the United States and around the world, acting as airborne platforms that are essential to the gathering of environmental and geographic data for scientific research. (http://www.flightscience.noaa.gov/, Adapted)

Information Assurance: Measures that protect and defend information and information systems by ensuring their availability, integrity, authentication, confidentiality, and non-repudiation. These measures include providing for restoration of information systems by incorporating protection, detection, and reaction capabilities. (NIST SP 800-53, CNSSI 4009)

Information Technology: Equipment, processes, procedures, and systems used to provide and support information systems (computerized and manual) within an organization and those reaching out to customers and suppliers. (Newton's Telecom Dictionary)

Infrastructure-as-a-Service: The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software, which can include operating systems and

applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, and deployed applications; and possibly limited control of select networking components (e.g., host firewalls). (NIST SP 800-145)

Internet Control Message Protocol: One of the core protocols of the Internet Protocol Suite, it is chiefly used by networked computers' operating systems to send error messages – indicating, for instance, that a requested service is not available or that a host or router could not be reached. (International Engineering Task Force (IETF) Request for Comment (RFC) 792)

Internet of Things: The total interconnected collection of device networks. (Newton's Telecom Dictionary)

Internet Protocol (IP): Part of the Transmission Control Protocol/IP family of protocols describing software that tracks the Internet address of nodes, routes outgoing messages, and recognizes incoming messages; also used in gateways to connect networks at Open Systems Interconnection network Level 3 and above. (Newton's Telecom Dictionary)

Intrusion Detection Systems (IDS): The system and/or process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents, which are violations or imminent threats of violation of computer security policies, acceptable use policies, or standard security practices. IDS are the software that automates the intrusion detection process. (NIST SP 800-94 – Revision 1 (Draft), Adapted)

Intrusion Prevention Systems (IPS): The system and/or process of performing intrusion detection and attempting to stop detected possible incidents. IPS are the software that has all the capabilities of an intrusion detection system and can also attempt to stop possible incidents. (NIST SP 800-94 – Revision 1 (Draft), Adapted)

Man-Made Disasters: For the purposes of this report, the NSTAC defines this term as events that arise either from the malicious or accidental actions of one or more individuals. These events threaten national security or warrant "out-of-the-ordinary" emergency response actions that may require collaboration by agencies at the Federal, State, and/or local government levels.

Metadata: Structured information that describes, explains, locates, and otherwise makes it easier to retrieve and use an information resource. (National Information Standards Organization)

National Security/Emergency Preparedness (NS/EP) Communications: Telecommunication services that are used to maintain a state of readiness or to respond to and manage any event or crisis (local, national, or international) which causes or could cause injury or harm to the population, damage to or loss of property, or degrades or threatens the NS/EP posture of the United States (47 Code of Federal Regulations Chapter II, § 201.2(g)). NS/EP communications include primarily those technical capabilities supported by policies and programs that enable the Executive Branch to communicate at all times and under all circumstances to carry out its mission essential functions and to respond to any event or crisis (local, national, or international), to include communicating with itself; the Legislative and Judicial branches; State, territorial, tribal, and local governments; private sector entities; as well as the public, allies, and other nations. NS/EP communications further include those systems and capabilities at all levels of

Government and the private sector that are necessary to ensure national security and to effectively manage incidents and emergencies. (NS/EP Communications Executive Committee definition based on Executive Order (EO) 13618, *Assignment of National Security and Emergency Preparedness Communications Functions* [2012])

National Information Exchange Model (NIEM): A community-driven, standards-based approach to exchanging information. NIEM provides a data model, governance, training, tools, technical support services, and a community-base to assists users in adopting a standards-based approach to exchanging data. (https://www.niem.gov/aboutniem/Pages/niem.aspx, Adapted)

Natural Language Processing: A range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. (Dr. Elizabeth Liddy, "Natural Language Processing", Syracuse University, 2001, http://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub).

Networks: Information system(s) implemented with a collection of interconnected components, which may include routers, hubs, cabling, telecommunications controllers, key distribution centers, and technical control devices. (NIST Glossary of Information Security Terms – NISTIR 7298 – Revision 2)

NIST *Big Data Interoperability Framework* (2015): A proposed reference architecture drafted by the NIST Big Data Public Working Group to develop big data definitions, taxonomies, important requirements for privacy and security protections, and a standards roadmap. (http://www.nist.gov/itl/bigdata/20150406_big_data_framework.cfm)

NIST *Framework for Improving Critical Infrastructure Cybersecurity* (2014): In response to EO 13636, *Improving Critical Infrastructure Cybersecurity* (2013), NIST created voluntary guidance, based on existing standards, guidelines, and practices, for critical infrastructure organizations to better manage and reduce cybersecurity risk. In addition to helping organizations manage and reduce risks, it was designed to foster risk and cybersecurity management communications amongst both internal and external organizational stakeholders. (http://www.nist.gov/cyberframework/cybersecurity-framework-faqs-framework-basics.cfm)

"Not-Only" Standard Query Language Databases: A type of database that provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases.

Operational Technology: Hardware and software that detects or causes a change through the direct monitoring and/or control of physical devices, processes, and events in an enterprise.

Operating System: A software program which manages the basic operations of a computer system. (Newton's Telecom Dictionary)

Platform-as-a-Service: The capability to deploy consumer-created or acquired applications created using programming languages, libraries, services, and tools onto the cloud infrastructure. The consumer does not manage or control the underlying cloud infrastructure including network,

servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. (NIST SP 800-145)

Principle of Least Privilege: The principle that every program and user of a system should operate using the least set of privileges necessary to complete the job. This principle limits the damage that can result from an accident or error, as well as reduces the number of potential interactions among privileged programs to prevent improper uses of privilege. (Dr. Fred B. Schneider, "Least Privilege and More," Cornell University, 2003, http://www.cs.cornell.edu/fbs/publications/leastPrivNeedham.pdf)

Protocol: A set of rules and formats, semantic and syntactic, permitting information systems to exchange information. (NIST Glossary of Information Security Terms – NISTIR 7298 – Revision 2)

Public Safety Answering Point (PSAP): A facility and/or entity equipped and staffed to receive emergency and non-emergency public safety calls for service via telephone and other communication devices. Emergency calls for service are answered, assessed, classified, and prioritized. A PSAP may also be called a 9-1-1 center. (*Federal Communications Commission Final Report of the Task Force on Optimal Public Safety Answering Point Architecture* [2016], *Standard for the Establishment of a Quality Assurance and Quality Improvement Program for Public Safety Answering Points* [2015], and *National Emergency Number Association Master Glossary of 9-1-1 Terminology* [2014], Adapted)

Quality Assurance: The random sampling of a data collection and its measurement against various quality characteristics, such as accuracy, completeness, validity, non-duplication or timeliness to determine its level of quality or reliability. (http://iaidq.org/main/glossary.shtml, Adapted)

Supervisory Control and Data Acquisition (SCADA) Systems: Computerized systems that are capable of gathering and processing data and applying operational controls over long distances. Typical uses include power transmission and distribution, and pipeline systems. SCADA was designed for the unique communication challenges (e.g., delays, data integrity) posed by the various media that must be used, such as phone lines, microwave, and satellite. Usually shared rather than dedicated. (NIST SP 800-82)

Secure Socket Layer (SSL): A protocol used for protecting private information during transmission via the Internet. SSL works by using a public key to encrypt data that's transferred over the SSL connection. Most Web browsers support SSL, and many Websites use the protocol to obtain confidential user information. By convention, Uniform Resource Locators that require an SSL connection start with "https:" instead of "http:". (CNSSI 4009, Adapted)

Semi-Structured Data: Data that does not conform to the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. (Dr. Peter Buneman, "Semi-Structured Data," University of Pennsylvania, 1997, http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf)

Simple Mail Transfer Protocol (SMTP): A set of standards that regulate the transfer of email over networks such as Internet. SMTP transfers mail across transport service environments in plain text, without security support, and from server to server only. (IETF RFC 2821)

Software-as-a-Service: The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a Web browser (e.g., Web-based email), or a program interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings. (NIST SP 800-145)

Standard Query Language: A widely used language for accessing and manipulating relational databases. (NIST SP 800-8)

Structured Data: Data that is organized in a pre-defined manner and can be immediately identified within an electronic structure such as a relational database. For example, to retrieve the name of a city, the "city field" is accessed. (PC Magazine Encyclopedia)

Threat: Any circumstance or event with the potential to adversely impact agency operations (including mission, functions, image, or reputation), agency assets, or individuals through an information system via unauthorized access, destruction, disclosure, modification of information, and/or denial of service. (NIST SP 800-53, CNSSI 4009, Adapted)

Transmission Control Protocol (TCP): TCP is a connection-oriented, end-to-end reliable protocol designed to fit into a layered hierarchy of protocols which support multi-network applications. TCP provides for reliable inter-process communication between pairs of processes in host computers attached to distinct but interconnected computer communication networks. (IETF RFC 761)

Unstructured Data: Information that either does not have a pre-defined data model or is not organized in a pre-defined manner. Unstructured information is typically text-heavy, but may also contain data such as dates, numbers, and facts. (PC Magazine Encyclopedia)

User Datagram Protocol (UDP): This protocol is defined to make available a datagram mode of packet-switched computer communication in the environment of an interconnected set of computer networks. UDP assumes that IP is used as the underlying protocol. It also provides a procedure for application programs to send messages to other programs with a minimum of protocol mechanism. The protocol is transaction oriented, and delivery and duplicate protection are not guaranteed. (IETF RFC 768)

Zero-Day Attacks: Software or hardware vulnerabilities that have been exploited by an attacker where there is no prior knowledge of the flaw in the general information security community, and, therefore, no vendor fix or software patch available for it. (https://www.fireeye.com/blog/executive-perspective/2014/04/zero-day-attacks-are-not-the-same-as-zero-day-vulnerabilities.html)

APPENDIX D: STEPS IN THE DATA LIFECYCLE

According to the National Institute of Standards and Technology's *Big Data Interoperability Framework*, the five steps in the data lifecycle are:

1. Data creation and collection;

2. Data identification, availability, and quality;

3. Data analytics;

4. Use of analytic outputs; and

5. Data management, storage, retention, and disposal.[78]

As outlined in the President's National Security Telecommunications Advisory Committee's (NSTAC) *NSTAC Big Data Analytics Scoping Report* (August 2015), the committee decided to examine its three use cases (i.e., a natural disaster, a man-made disaster, and a cyber attack on critical infrastructure) around the steps of the data lifecycle to "allow for a linear and structured approach that can be applied to [these] big data use cases."[79] The NSTAC also defined each step around a series of questions in order to better steps to illustrate the type of issues that the committee examined each stage. Each of these specific steps and the surrounding questions considered are defined below:[80]

1. **Data creation and collection:**
    a. What type of data is being collected, by whom, and for what purpose?
    b. What rules and regulations constrain original data collection?

2. **Data identification, availability, and quality for big data analytics (BDA):**
    a. How can analysts identify or how do data owners make available public and private datasets for BDA?
    b. How is data transmitted/transferred from its original source?
    c. Should there be different use practices for data collected from private versus public sources?
    d. What are the privacy and security issues related to use of data for BDA?
    e. How does data need to be cleansed or processed to make it usable for BDA?

3. **Data analytics:**
    a. How do constraints on data affect what and how BDA can be conducted?
    b. Who has the rights to conduct BDA on data that was originally collected for other purposes?

---

[78] NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

[79] NSTAC. *NSTAC Big Data Analytics Scoping Report*. August 12, 2015. https://www.dhs.gov/sites/default/files/publications/Final%20NSTAC%20Big%20Data%20Analytics%20Scoping%20Report%20%288-12-15%29_0.pdf.

[80] Ibid.

c. What legal issues arise when BDA is conducted on data originally collected for another purpose?
d. What are the objectives of the analytics (e.g., reactive, preventive, forensic, or predictive)? Do particular objectives create distinct policy issues?
e. What is the operational model for conducting BDA? How do Federal entities build, source, or share analytic platforms for conducting BDA?
f. What are the privacy and security issues that arise from:
    i. The transfer of large datasets from the original data owner to the entity conducting the analytics?
    ii. The integration of data from multiple sources?
    iii. How and by whom the data is analyzed (e.g., on premise, Federal shared-service, or commercial cloud)?

4. **Use of analytic outputs:**
    a. How are BDA activities integrated with existing organizational processes and mission activities to ensure timeliness, impact, and ease of use?
    b. Who owns the rights to BDA-derived data and the analytic outputs?
    c. Have precautions been taken to ensure the analytic products are nondiscriminatory and do not infringe upon individual privacy?
    d. How are analytic results shared and/or published to provide actionable support for decision making?
    e. How are analytic results shared, published, and operationalized both within an organization and with external stakeholders?
    f. What domain models or standards can effectively facilitate the sharing of analytic outputs?

5. **Data management, storage, retention, and disposal:**
    a. How should retention and disposal of the data be handled?
    b. How should retention and disposal of the analysis results be handled?
    c. Who has access to the data and how is it shared?
    d. What type of protection policies should be in place for stored data?
    e. What issues need to be addressed to facilitate data sharing within Government and between Government and the private sector?
    f. What risks are involved with the storage and retention of big datasets?
    g. What are the unique considerations of retaining data for extended periods of time?

APPENDIX E: RECENT EXECUTIVE BRANCH ACTIONS

Recognizing the explosive growth of big data, the current Administration has taken a number of actions to promote the use of open big datasets in order to make them more accessible for the public good.

Data.gov was launched in May 2009 under the guidance of Mr. Vivek Kundra, then U.S. chief information officer. The Website currently hosts over 193,000 datasets from Federal, State, and local agencies, amongst others. The majority of the data is accessible in .html, .xml, or .pdf formats, and is searchable by source organization and topic.[81]

The White House Office of Science and Technology Policy (OSTP) launched the *Big Data Research and Development Initiative* in March 2012. As part of the initiative, six Federal departments and agencies announced more than $200 million in research and development (R&D) investments designed to, "greatly improve the tools and techniques needed to access, organize, and glean discoveries from huge volumes of data."[82] Participating Federal departments and agencies, and their financial commitments in support of the initiative, include the following:

- **Department of Defense (DOD):** The DOD has invested $250 million annually across the military departments in a series of programs that will improve situational awareness of warfighters, and harness and utilize data in new ways. The Defense Advanced Research Projects Agency also launched the XDATA program, investing $25 million annually to develop computational techniques and software tools for analyzing large volumes of semi-structured and unstructured data.[83]

- **Department of Energy (DOE):** The DOE's Scientific Discovery Through Advanced Computing program provides $25 million in funding for the Scalable Data Management, Analysis and Visualization Institute. The institute brings together six National Laboratories and seven universities, "with the goal of developing new and improved tools to help scientists manage and visualize data."[84]

- **Department of Health and Human Services and the National Institutes of Health (NIH):** The NIH has made the world's largest set of data on human genetic variation, produced by the 1000 Genomes Project, publicly available on the Amazon Web Services cloud. This public-private collaboration allows the data, which consists of over 200 terabytes, to be more easily accessed and analyzed in the cloud as a public dataset.[85]

---

[81] GSA. "About Data.gov." Office of Citizen Services and Innovative Technologies. Accessed on March 15, 2016. https://www.data.gov/about.

[82] White House Office of Science and Technology Policy. "Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments." March 29, 2012. https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.

[83] Ibid.

[84] Ibid.

[85] National Institutes of Health. "1000 Genomes Project Data Available on Amazon Cloud." March 29, 2012. http://www.nih.gov/news-events/news-releases/1000-genomes-project-data-available-amazon-cloud.

- **National Science Foundation (NSF):** The NSF released a "Core Techniques and Technologies for Advancing Big Data Science and Engineering" solicitation to "fund research to develop and evaluate new algorithms, statistical methods, technologies, and tools for improved data collection and management, data analytics and e-science collaboration environments."[86]

- **United States Geological Survey (USGS):** The USGS's John Wesley Powell Center for Analysis and Synthesis has promoted the use of big data in earth system science by improving and increasing the availability of data manipulation and management capabilities to USGS scientists.[87]

On May 9, 2013, President Obama issued Executive Order (EO) 13642, *Making Open and Machine Readable the New Default for Government Information*. The President ordered that, "Government information shall be managed as an asset throughout its lifecycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable. In making this the new default state, [E]xecutive departments and agencies…shall ensure that they safeguard individual privacy, confidentiality, and national security."[88]

Pursuant to the above referenced EO, the Office of Management and Budget (OMB) and OSTP released a framework in May 2013 for agencies to manage information as an asset throughout its lifecycle. The Open Data Policy (OMB M-13-13) requires agencies to:

> Collect or create information in a way that supports downstream information processing and dissemination activities. This includes using machine-readable and open formats, data standards, and common core and extensible metadata for all new information creation and collection efforts. It also includes agencies ensuring information stewardship through the use of open licenses and review of information for privacy, confidentiality, security, or other restrictions to release. Additionally, it involves agencies building or modernizing information systems in a way that maximizes interoperability and information accessibility, maintains internal and external data asset inventories, enhances information safeguards, and clarifies information management responsibilities.[89]

This guidance is now available on Project Open Data Metadata Schema v1.1.[90]

Additionally, the Administration called for a 90-day review to examine how the public and private sector could maximize the benefits of big data while minimizing its risks, particularly in

86 National Science Foundation. "NSF Leads Federal Efforts in Big Data." *Press Release 12-060*. March 29, 2012. http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607.

87 United States Geological Service. "About the Powell Center." Accessed on March 23, 2016. https://powellcenter.usgs.gov/about.

88 White House Office of the Press Secretary, *Executive Order 13642, Making Open and Machine Readable the New Default for Government Information*. May 9, 2013. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

89 EOP. "Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy – Managing Information as an Asset." May 9, 2013. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

90 Office of the United States Chief Information Officer. *Project Open Data Metadata Schema v1.1*. Accessed on March 15, 2016. https://project-open-data.cio.gov/v1.1/schema/.

the area of privacy. A working group of senior Administration officials issued a set of policy recommendations, including to:

- **Advance the *Consumer Privacy Bill of Rights*:** The Department of Commerce (DOC) should take appropriate consultative steps to seek stakeholder and public comment on big data developments and how they impact the *Consumer Privacy Bill of Rights* and then devise draft legislative text for consideration by stakeholders and submission by the President to Congress.

- **Pass National Data Breach Legislation:** Congress should pass legislation that provides for a single national data breach standard along the lines of the Administration's May 2011, cybersecurity legislative proposal.

- **Extend Privacy Protections to Non-U.S. Persons:** OMB should work with departments and agencies to apply the *Privacy Act of 1974* to non-U.S. persons where practicable, or to establish alternative privacy policies that apply appropriate and meaningful protections to personal information regardless of a person's nationality.

- **Ensure Data Collected on Students in School is Used for Educational Purposes:** The Federal Government must ensure that privacy regulations protect students against having their data being shared or used inappropriately, especially when the data is gathered in an educational context.

- **Expand Technical Expertise to Stop Discrimination:** The Federal Government's lead civil rights and consumer protection agencies should expand their technical expertise to be able to identify practices and outcomes facilitated by big data analytics that have a discriminatory impact on protected classes, and develop a plan for investigating and resolving violations of law.

- **Amend the *Electronic Communications Privacy Act (ECPA)*:** Congress should amend the ECPA to ensure the standard of protection for online, digital content is consistent with that afforded in the physical world — including removing archaic distinctions between emails left unread or over a certain age.[91]

At the same time the above policy-focused report was written, the President's Council of Advisors on Science and Technology issued a complimentary report examining the current technologies for managing and analyzing big data and for preserving privacy. The report concluded that, "policy [should] focus primarily on whether specific uses of people affect privacy adversely. It also recommends that policy focuses on outcomes, in the 'what' rather than the 'how' to avoid becoming obsolete as technology advances."[92]

[91] EOP. *Big Data: Seizing Opportunities, Preserving Values*. May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[92] President's Council of Advisors on Science and Technology. *Big Data and Privacy: A Technological Perspective*. May 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

The Networking and Information Technology Research and Development (NITRD) Program is a collaboration of Federal research and development agencies working to meet the needs of the Federal government in the areas of advanced networking, computing systems, software, and associated information technologies.[93] To further promote the Government's use of big data, NITRD's Big Data Senior Steering Group has spearheaded a variety of efforts, such as Big Data Regional Innovation Hubs, the National Big Data R&D Initiative, and other measures to expand the workforce needed to develop and use big data technologies.[94]

The Department of Homeland Security (DHS) has also sought ways to use big data in support of its emergency response functions. For example, DHS recently released a request for proposal entitled, "Using Social Media to Support Timely and Targeted Emergency Response Actions," in which the Department solicited input from the private sector on how the Government can best gather and operationalize unstructured social media data to improve first responders' real-time situational awareness during national-level emergencies.[95] Moreover, DHS' Science and Technology (S&T) Directorate has funded the development of a variety of big data projects, such as the Social Media Analytics and Reporting Toolkit (SMART), "a social media analysis system that provides analysts with scalable analysis and visualization of social media posts."[96] Created by the Center for Visualization and Data Analytics at Purdue University, a DHS S&T Center of Excellence, the SMART system, "uses topic extraction, combinations of key word filters, word cluster examination, and unusual event detection to provide situational awareness and improve decision-making for time-critical tasks."[97]

Within the DOC, the National Institute of Standards and Technology (NIST) is leading the development of a Big Data Technology Roadmap. This roadmap seeks to define and prioritize requirements for interoperability, portability, reusability, and extensibility for big data analytic techniques and technology infrastructure in order to support secure and effective adoption of big data. To help develop the ideas in the Big Data Technology Roadmap, NIST created the Big Data Public Working Group (NBD-PWG), the charge of which is the following:

> The focus of the NBD-PWG is to form a community of interest from industry, academia, and [G]overnment, with the goal of developing a consensus definitions, taxonomies, secure reference architectures, and technology roadmap. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable [b]ig [d]ata stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from Big Data service providers and flow of data between the stakeholders in a cohesive and secure manner.[98]

---

[93] Networking and Information Technology Research and Development (NITRD) Program. "About the NITRD Program." Accessed on March 15, 2016. https://www.nitrd.gov/about/about_nitrd.aspx.

[94] NITRD Program. "Big Data Senior Steering Group (SSG)." Accessed on March 15, 2016. https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title.

[95] DHS. "Using Social Media to Support Timely and Targeted Emergency Response Actions." Small Business Innovation Research Program. Accessed on March 23, 2016. https://www.sbir.gov/sbirsearch/detail/867811.

[96] DHS. "Improving Disaster Response and Recovery: Social Media Analytics and Reporting Toolkit." January 2014. Accessed on March 23, 2016. https://www.dhs.gov/sites/default/files/publications/Improving%20Disaster%20Response%20and%20Recovery-Social%20Media%20Analytics%20and%20Reporting%20Toolkit-CVADA-SMART-Jan2014_2.pdf.

[97] Ibid.

[98] NIST. *NIST Big Data Public Working Group*. Accessed on March 15, 2016. http://bigdatawg.nist.gov/home.php.

The working group plans to release three versions of the NIST *Big Data Interoperability Framework*, which seek to:

- **Stage 1:** Identify the high level big data reference architecture key components, which are technology, infrastructure, and vendor-agnostic.

- **Stage 2:** Define general interfaces between the NIST Big Data Reference Architecture (NBDRA) components.

- **Stage 3:** Validate the NBDRA by building big data general applications through general interfaces. [99]

The first of seven volumes the NBD-PWG plans to develop was released in November 2015, and can be found on the NBD-PWG Website.[100]

[99] NIST Big Data Working Group. *NIST Big Data Interoperability Framework*. November 2015.http://bigdatawg.nist.gov/V1_output_docs.php.
[100] Ibid.

APPENDIX F: FEMA'S FOUR PHASES OF EMERGENCY MANAGEMENT

The U.S. Federal Emergency Management Agency (FEMA) provides multiple definitions of disasters, one of which is:

> An occurrence of a natural catastrophe, technological accident, or human-caused event that has resulted in severe property damage, deaths, and/or multiple injuries. As used in this Guide, a 'large-scale disaster' is one that exceeds the response capability of the local jurisdiction and requires State, and potentially Federal, involvement. As used in the *Stafford Act*, a 'major disaster' is 'any natural catastrophe [...] or, regardless of cause, any fire, flood, or explosion, in any part of the United States, which in the determination of the President causes damage of sufficient severity and magnitude to warrant major disaster assistance under [the] Act to supplement the efforts and available resources or States, local governments, and disaster relief organizations in alleviating the damage, loss, hardship, or suffering caused thereby.[101]

FEMA formalized an approach to describing emergency management that provides a useful framework for developing policy recommendations for big data analytics in relation to both man-made and natural disasters.

FEMA'S FOUR PHASES OF EMERGENCY MANAGEMENT	
Mitigation *Preventing future emergencies or minimizing their effects*	• Includes any activities that prevent an emergency, reduce the chance of an emergency happening, or reduce the damaging effects of unavoidable emergencies. • Buying flood and fire insurance for your home is a mitigation activity. • Mitigation activities take place **before** and **after** emergencies.
Preparedness *Preparing to handle an emergency*	• Includes plans or preparations made to save lives and to help response and rescue operations. • Evacuation plans and stocking food and water are both examples of preparedness. • Preparedness activities take place **before** an emergency occurs.
Response *Responding safely to an emergency*	• Includes actions taken to save lives and prevent further property damage in an emergency situation. Response is putting your preparedness plans into action. • Seeking shelter from a tornado or turning off gas valves in an earthquake are both response activities. • Response activities take place **during** an emergency.
Recovery *Recovering from an emergency*	• Includes actions taken to return to a normal or an even safer situation following an emergency. • Recovery includes getting financial assistance to help pay for the repairs. • Recovery activities take place **after** an emergency.

Table F.1. FEMA's four phases of emergency management.

[101] DHS. FEMA. *Guide For All-Hazard Emergency Operations Planning* (State and Local Guide 101). http://www.fema.gov/pdf/plan/slg101.pdf, page GLO-1.

APPENDIX G: BIBLIOGRAPHY

Columbus, Louis. "Roundup of Cloud Computing Forecasts and Market Estimates Q3 Update, 2015." *Forbes Magazine*. September 27, 2015. http://www.forbes.com/sites/louiscolumbus/2015/09/27/roundup-of-cloud-computing-forecasts-and-market-estimates-q3-update-2015/#7032b29a6c7a.

Data-Driven Documents. "About." Accessed on March 15, 2016. https://d3js.org/.

Department of Homeland Security (DHS). Federal Emergency Management Agency. *Guide For All-Hazard Emergency Operations Planning* (State and Local Guide 101). http://www.fema.gov/pdf/plan/slg101.pdf.

DHS. "Improving Disaster Response and Recovery: Social Media Analytics and Reporting Toolkit." January 2014. Accessed on March 23, 2016. https://www.dhs.gov/sites/default/files/publications/Improving%20Disaster%20Response%20and%20Recovery-Social%20Media%20Analytics%20and%20Reporting%20Toolkit-CVADA-SMART-Jan2014_2.pdf.

DHS. "Using Social Media to Support Timely and Targeted Emergency Response Actions." Small Business Innovation Research Program. Accessed on March 23, 2016. https://www.sbir.gov/sbirsearch/detail/867811.

Eberhardt, John, III. 3E Services, LLC. *Briefing to the President's National Security Telecommunications Advisory Committee (NSTAC) Big Data Analytics (BDA) Subcommittee*. December 18, 2015.

Elizhauser, Anne et al. "Septicemia in U.S. Hospitals." *Healthcare Cost and Utilization Project*. October 2011. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb122.pdf.

EMC Corporation. *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*. "Executive Summary." April 2014. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm.

EMC Corporation. *The Digital Universe: Driving Data Growth in Healthcare*. 2014. https://www.emc.com/analyst-report/digital-universe-healthcare-vertical-report-ar.pdf.

Evans, William. Boston Police Department. *Briefing to the NSTAC BDA Subcommittee*. June 30, 2015.

Executive Office of the President (EOP). *Big Data: Seizing Opportunities, Preserving Values*. May 2014. https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

EOP. "Memorandum for the Heads of Executive Departments and Agencies: Open Data Policy – Managing Information as an Asset." May 9, 2013. https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf.

Federal Trade Commission. *Fair Information Practice Principles Framework*. March 31, 2009. https://web.archive.org/web/20090331134113/http://www.ftc.gov/reports/privacy3/fairinfo.shtm.

Forgety, Trey. National Emergency Number Association. *Briefing to the NSTAC BDA Subcommittee*. December 15, 2015.

Gartner, Inc. "Gartner Says 6.4 Billion Connected 'Things' Will Be in Use in 2016, Up 30 Percent From 2015." November 10, 2015. http://www.gartner.com/newsroom/id/3165317.

General Services Administration. "About Data.gov." Office of Citizen Services and Innovative Technologies. Accessed on March 15, 2016. https://www.data.gov/about.

Haddad, John. "Top 5 Big Data Challenges." *Informatica*. July 10, 2014. http://blogs.informatica.com/2014/07/10/top-5-big-data-challenges/#fbid=uOMCiPkWhS8.

Harman, Wendy. American Red Cross. *Briefing to the NSTAC BDA Subcommittee*. June 9, 2015.

Hayes, Stephen and Louise Tucker. Ericsson, Inc. *Briefing to the NSTAC BDA Subcommittee.* April 14, 2015.

Headd, Mark. "Making the Business Case for Big Data. What Is Big Data, Anyway?" *The Promise of Big Data for the Public Sector*. The Center for Digital Government. 2013. http://media.navigatored.com/documents/CDG13_SPQ1_V.pdf.

Hess, Jeffrey. "From Police to Pipes: Fresno Leveraging 'Big Data' To Improve City Functions." *Valley Public Radio*. May 26, 2015. http://kvpr.org/post/police-pipes-fresno-leveraging-big-data-improve-city-functions.

Infante, Adam. "Big Data = Big Opportunity? How Does It Work?" PricewaterhouseCoopers. March 3, 2016. http://pwc.blogs.com/analytics_means_business/2016/03/big-data-big-opportunity-how-does-it-work.html.

Juniper Research. "'Internet of Things' Connected Devices to Almost Triple to Over 38 Billion Units by 2020." July 28, 2015. http://www.juniperresearch.com/press/press-releases/iot-connected-devices-to-triple-to-38-bn-by-2020.

Laney, Doug. "3D Data Management: Controlling Data Volume, Velocity, and Variety." *Gartner.com*. February 6, 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lundquist, Eric. "Why You'll Need a Big Data Ethics Expert." *InformationWeek*. January 3, 2013. http://www.informationweek.com/it-leadership/why-youll-need-a-big-data-ethics-expert/d/d-id/1108001.

Marco, David. "Data Assurance Road Map, Part 1." *InformationManagement.com*. December 1, 2014. http://www.information-management.com/issues/20041201/1014518-1.html.

McKinsey Global Institute. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. May 2011. http://www.mckinsey.com/business-functions/business-technology/our-insights/big-data-the-next-frontier-for-innovation.

Mundie, Craig. "Privacy Pragmatism: Focus on Data Use, Not Data Collection." *Council for Foreign Affairs*. March/April 2014. http://www.cfr.org/intelligence/privacy-pragmatism/p32837.

National Institute of General Medical Sciences. *Sepsis Factsheet*. Accessed on March 15, 2016. https://www.nigms.nih.gov/education/pages/factsheet_sePSIs.aspx.

National Institutes of Health. "1000 Genomes Project Data Available on Amazon Cloud." March 29, 2012. http://www.nih.gov/news-events/news-releases/1000-genomes-project-data-available-amazon-cloud.

National Institute of Standards and Technology (NIST). *NIST Big Data Public Working Group.* Accessed on March 15, 2016. http://bigdatawg.nist.gov/home.php.

NIST Big Data Public Working Group. *NIST Big Data Interoperability Framework*. November 2015. http://bigdatawg.nist.gov/V1_output_docs.php.

NIST Big Data Public Working Group Definitions and Taxonomies Subgroup. *NIST Big Data Interoperability Framework, Volume 1: Definitions*. NIST Special Publication (SP) 1500-1. September 2015. http://dx.doi.org/10.6028/NIST.SP.1500-1.

NIST. *Framework for Improving Critical Infrastructure Cybersecurity*. February 12, 2014. http://www.nist.gov/cyberframework/upload/cybersecurity-framework-021214.pdf.

National Science Foundation. "NSF Leads Federal Efforts in Big Data." *Press Release 12-060*. March 29, 2012. http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607.

Networking and Information Technology Research and Development (NITRD) Program. "About the NITRD Program." Accessed on March 15, 2016. https://www.nitrd.gov/about/about_nitrd.aspx.

NITRD Program. "Big Data Senior Steering Group (SSG)." Accessed on March 15, 2016. https://www.nitrd.gov/nitrdgroups/index.php?title=Big_Data_(BD_SSG)#title.

NSTAC. *NSTAC Big Data Analytics Scoping Report*. August 12, 2015. https://www.dhs.gov/sites/default/files/publications/Final%20NSTAC%20Big%20Data%20Analytics%20Scoping%20Report%20%288-12-15%29_0.pdf.

NSTAC. *NSTAC Report to the President on Cloud Computing.* May 15, 2012. https://www.dhs.gov/sites/default/files/publications/2012-05-15-NSTAC-Cloud-Computing.pdf.

NSTAC. *NSTAC Report to the President on the Internet of Things*. November 19, 2014. https://www.dhs.gov/sites/default/files/publications/NSTAC%20Report%20to%20the%20President%20on%20the%20Internet%20of%20Things%20Nov%202014%20%28updat%20%20%20.pdf.

Office of the United States Chief Information Officer. *Project Open Data Metadata Schema v1.1*. Accessed on March 15, 2016. https://project-open-data.cio.gov/v1.1/schema/.

Ovide, Shira. "Tapping 'Big Data' to Fill Potholes." *The Wall Street Journal*. June 12, 2012. http://www.wsj.com/articles/SB10001424052702303444204577460552615646874.

President's Council of Advisors on Science and Technology. *Big Data and Privacy: A Technological Perspective*. May 2014. https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

Puckett, Richard. General Electric. *Briefing to the NSTAC BDA Subcommittee*. March 17, 2015.

RockHealth. "Big Data in Digital Health." Accessed on March 31, 2016. http://rockhealth.com/resources/rock-reports/big-data/.

Schmidt, Eric. *Remarks at the Techonomy Conference in Lake Tahoe*. August 2010. http://techcrunch.com/2010/08/04/schmidt-data.

Söhnchen, Stefanie. "Why Big Data Is the Future and the Present of Transportation." *Move Forward*. June 7, 2015. https://www.move-forward.com/news/details/why-big-data-is-the-future-and-the-present-of-transportation/.

Sondergaard, Peter. *Remarks at the Gartner Symposium/ITxpo*. October 2011. http://www.gartner.com/newsroom/id/1824919.

State of New Jersey Department of Transportation. "New Jersey Traffic Monitoring Program." June 5, 2014. http://www.state.nj.us/transportation/refdata/roadway/pdf/program.pdf.

United States Geological Service. "About the Powell Center." Accessed on March 23, 2016. https://powellcenter.usgs.gov/about.

White House Office of the Press Secretary. *Executive Order 13642, Making Open and Machine Readable the New Default for Government Information*. May 9, 2013. https://www.whitehouse.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-.

White House Office of Science and Technology Policy. "Obama Administration Unveils 'Big Data' Initiative: Announces $200 Million in New R&D Investments." March 29, 2012. https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf.