DECEMBER 2016
FEDERAL BIG DATA SUMMIT REPORT[*]

March 28, 2017

Christine Harvey, Matt Mickelson, Ron Campbell,
Dr. Mike Richey, Bob Natale, Dr. Haleh Vafaie,
The MITRE Corporation[†]

Tim Harvey and Tom Suder
The Advanced Technology Academic Research Center

March 28, 2017

[*] APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. CASE NUMBER 17-1311. ©2017 THE MITRE CORPORATION. ALL RIGHTS RESERVED.

[†] The authors' affiliation with The MITRE Corporation is provided for identification purposes only, and is not intended to convey or imply MITRE's concurrence with, or support for, the positions, opinions or viewpoints expressed by the authors.

Contents

Executive Summary

The most recent installment of the Advanced Technology Academic Research Center (ATARC) Federal Big Data Summit, held on December 13, 2016, included five MITRE-ATARC Collaboration Sessions. These collaboration sessions allowed industry, academic, government, and MITRE representatives the opportunity to collaborate and discuss challenges the government faces in big data research and technologies. The goal of these sessions was to create a forum to exchange ideas and develop recommendations to further the adoption and advancement of big data techniques and best practices within the government.

Participants representing government, industry, and academia addressed five challenge areas in big data: Big Data for Autonomy and Autonomous Systems; Big Data and Cyber Security; Big Data as a Catalyst; Big Data for Mission Success; and Big Data and Health Care.

This white paper summarizes the discussions in the collaboration sessions and presents recommendations for government and academia while identifying orthogonal points between challenge areas. The sessions identified detailed actionable recommendations for the government and academia which are summarized below:

- Data sharing and collaboration continues to be an important area of development. Policies need to be put in place to allow for simple data sharing and avenues for communication between agencies need to be established for easy access and aggregation of data.

- Concerns with data governance, provenance, and reliability are often mentioned at the Big Data Summits. Organizations recognize the importance of reliable and trustworthy data and need to establish regulations to ensure the integrity of the data.

- Big data is no longer a new field and agencies need to recognize the established benefits of working with big data. A skilled workforce is necessary to continue making progress and to enable big data to be able used as a catalyst for mission success in a variety of fields including autonomous systems, cyber security, and health care.

1 Introduction

During the most recent Advanced Technology Academic Research Center (ATARC) Federal Big Data Summit, held on December 13, 2016, five MITRE-ATARC collaboration sessions gave representatives of industry, academia, government, and MITRE the opportunity to discuss challenges the government faces in big data. Experts who would not otherwise meet or interact used these sessions to identify challenges, best practices, recommendations, success stories, and requirements to advance the state of big data technologies and research in the government.

The MITRE Corporation is a not-for-profit company that operates multiple Federally Funded Research and Development Centers (FFRDCs). ATARC is a non-profit organization that leverages academia to bridge between government and corporate participation in technology. MITRE worked in partnership with ATARC to host these collaborative sessions as part of the Federal Big Data Summit. The invited collaboration session participants across government, industry, and academia worked together to address challenge areas in big data, as well as identify courses of action to be taken to enable government and industry collaboration with academic institutions. Academic participants used the discussions as a way to help guide research efforts, curricula development, and to help produce graduates ready to join the work force and advance the state of big data research and work in the government.

This white paper is a summary of the results of the collaboration sessions and identifies suggestions and recommendations for government, industry, and academia while identifying cross-cutting issues between the challenge areas.

2 Collaboration Session Overview

Each of the five MITRE-ATARC collaboration sessions consisted of a focused and moderated discussion of current problems, gaps in work programs, potential solutions, and ways forward. At this summit, sessions addressed:

- Big Data for Autonomy and Autonomous Systems
- Big Data and Cyber Security
- Big Data as a Catalyst
- Leveraging Big Data for Mission Success
- Big Data and Health Care

This section outlines the challenges, themes, and findings of each of the collaboration sessions.

2.1 Big Data for Autonomy and Autonomous Systems

The Autonomy and Autonomous Systems session discussed machine learning systems and the ability of those systems to allow unprecedented levels of decision-making without human involvement.

The session included discussions of the following:

- What is the difference between automation and autonomy?
- How must policy change to reflect the impact of autonomous systems on our lives?
- What are the legal concerns regarding autonomous systems?
- What research topics need to be funded regarding autonomous systems?
- Are any research areas off-limits?
- How will autonomous systems change the roles of human operators?

2.1.1 Challenges

- Autonomous systems fall along a continuum between automatic (deterministic) and fully-intelligent (capable of learning and adapting without human modification), and there is no single universal degree of autonomy that is optimal for all systems.
- Maintaining meaningful human control is essential, but difficult to translate into requirements and evaluate in completed systems.
- Determining how an autonomous system learned, and what it learned, is critical. However, producers of autonomous systems have little incentive to provide such information, and consumers typically have little ability to determine such information.
- Government and industry lack a solid framework for testing systems that are capable of learning or self-modification.
- There is no right way to determine liability when autonomous systems cause harm (i.e., fault of the manufacturer, algorithm, user, developer, owner, other system); especially when other autonomous systems are involved.

- Some autonomous systems require ethical consideration to operate, and this would require the ability to imbue ethics into the system.

2.1.2 Discussion Summary

The session began with a definition of autonomy, and a brief discussion of how autonomy contrasts with automation. Per the discussion, autonomy was defined as the degree to which human involvement is no longer required for a particular task. Further, autonomous systems lie on a continuum somewhere between automatic and fully-intelligent.

Next, the discussion turned to the need for meaningful human control in a given system. As systems move toward the fully-intelligent side of the continuum, some degree of human control is still desired. However, there is no consensus on what the appropriate level of human control should be, nor is there a single level of human control that is applicable for all systems. For example, the group expressed greater desires for human control in systems with potential health and safety hazards. Specifically, as the health and safety concerns of a system increased (e.g., vacuum cleaners vs. thermostats vs. self-driving cars vs. weapon systems), the more human control was desired. This tendency was confounded, however, by competitive scenarios where an autonomous system gives an adversary an advantage. This created a game theoretic "arms race" scenario in which there was greater tolerance for autonomy maintain competitive advantage. Examples of these scenarios included military and cyber security uses.

The group then discussed how to apply legal and policy constructs to autonomous systems; specifically, how to hold autonomous systems accountable. The group discussed whether or not autonomous systems have "personhood" in the eyes of the law. Furthermore, autonomous systems are often comprised of multiple subsystems, and the group found it difficult to attribute liability to any specific portion of the system (e.g., the hardware, the algorithm, the manufacturer, the owner, the operator). Finally, the group discussed the potential ways to enforce policy violations in systems with little human-based component and manage accountability.

Government's role in all this is (1) to protect the best interests of the nation's people, and (2) to ensure that meaningful human control exists in all autonomous systems. The government is uniquely positioned to drive standards, policy, law, and research interests regarding autonomous systems. However, there is a lack of clarity in the law's lexicon and the existing policy framework that needs to be resolved. Otherwise, meaningful human control of technology systems is at risk of disappearing. While the use of autonomous systems is still an emerging capability, it is evolving rapidly and is evolving faster than government processes can react.

2.1.3 Important Findings

- Autonomous systems should not make the complete decision. Each system requires a threshold of meaningful human control.
- Autonomous systems coupled with existing human judgment (i.e., augmented intelligence) have the highest likelihood of commercial adoption.
- Reducing the human involvement in a task will contribute to the erosion of skills in that task (e.g., driving skills will erode as human drivers give up the driving experience).
- Autonomous systems will still exhibit bias because of how they were trained.
- There is too much uncertainty in the law regarding autonomous systems, and this must be resolved to properly establish and enforce policies and regulations.
- The general population does not understand enough about how autonomous systems work and the risks such systems pose to make informed purchasing and usage decisions.
- The government should develop a plan and vision for the acceptable usage of autonomous systems.
- The government should invest in appropriate research, influence international policy, and drive commercial efforts appropriately to ensure autonomous systems do not disproportionately disadvantage the general population.

2.2 Big Data and Cyber Security

The Big Data and Cyber Security session discussed the challenges in implementing big data concepts for cyber security. The participants noted there are three major skills needed to address this topic: security information and event management (SIEM) tool skills, cyber security skills (such as the cyber-attack lifecycle) and data analytics skills (such as those used by a data scientist). The participants also noted that these are typically higher-end skills that are not readily available in the job market.

The topics also included a discussion of the Big Data Vs: volume, velocity, variety, veracity, value, validity and visualization.

The session included discussions of the following:

- What are the skills required to develop and sustain big data cyber security capability within an organization?
- How can big data analysis techniques be used to solve cyber security issues?
- What does cyber analytics look like for a given government agency?
- How can the Chief Information Officer (CIO) "pivot" (i.e., change) his/her organization to leverage and use cyber analytics?
- Due to limited budgets how should agencies prioritize their cyber security activities?
- What is the biggest impact of big data for cyber security?
- Can the U.S. Federal Chief Information Officer Council (i.e., CIO Council) be used to share technical knowledge (including best practices, tactics, techniques and procedures (TTPs) and configuration data) across federal agencies?

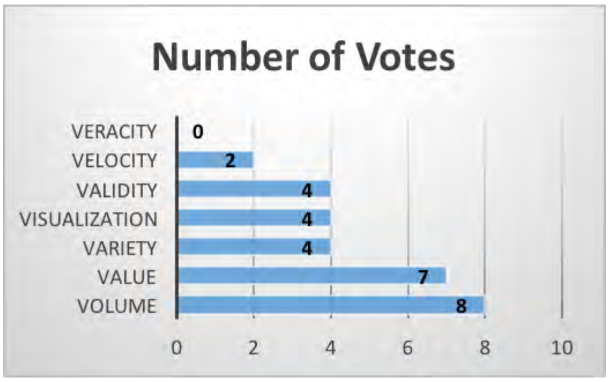
2.2.1 Challenges

- There is a tremendous amount of data to sort through. Typically, early data filtering is based on what is currently known. Filtered data may also provide insight into cyber security threats.
- Given, the large amount of data that is available for processing it is challenging to determine which data elements are actionable.
- Semantic differences between sensors from different vendors makes it difficult to compare and analyze data in a multi-vendor environment.
- Data sharing across multiple agencies can be challenging due to the unique data needs of each agency, the need to analyze data "in context" and a lack of a common lexicon that can be used across government agencies.
- There is currently limited sharing of configuration data across multiple government agencies. While vulnerability data is sometimes shared, the sharing of virtual machine (VM) configuration data is a challenge.

2.2.2 Discussion Summary

The session began with a discuss of the skills needed to leverage big data analytics for cyber security issues. Participants noted that since personnel with big data and cyber security skills are not readily available, a special education tract is needed to sustain big data cyber security capabilities. Personnel with the relevant big data and cyber security skills typically command higher salaries, resulting in higher priced contractors. Government agencies should consider training existing employees in big data analytics and cyber security capabilities in order to develop and sustain an analytical cyber security capability within their organization. CIOs and Chief Information Security Officers (CISOs) should work closely with the business units to assign "value" to assets, starting with the mission and business objectives. CIOs and CISOs should also better inform the business units on what can (and cannot) be done with respect to cyber security data analysis.

Figure 1: Voting results for the definition of big data.



Participants noted that since there are a large number of actions required to protect data and other valuable assets from attack and disruption, agencies must prioritize their cyber defense activities based on the value of the assets (e.g., critical data) and the potential impact of a given threat. There is also a tremendous amount of data to sort through (e.g., multiple system logs, firewall logs, etc.). This data is often filtered based on what is currently known (e.g., known vulnerabilities) while the remaining data is ignored. Valuable information may be included in the data that is filtered out. One participant suggested that filtered data (i.e., data based on known vulnerabilities) could be sent to a SIEM tool while all data could be retained and later analyzed using open source tools.

Data sharing was also discussed. Participants stated that while software vulnerability data is sometimes shared, system configuration data is not often shared among government agencies. This includes the lack of sharing of VM security configuration data that could be used to help agencies when attempting to secure new VM environments. While participants agreed that it is possible to share VM configuration data across government agencies, they also noted that policy issues may limit the sharing of detailed implementations and solutions. Participants would like to know if there exist a data model that can be shared among government agencies regarding how to build, secure, publish and reuse VMs.

The participants recommended putting together a framework for evaluating cyber security activities to determine what can be implemented. The framework would include a listing of the assets, the value of the assets (i.e., the value of the asset if a breach occurred), the devices used for detecting cyber security events (e.g., intrusion detection devices, firewall logs, system logs, etc.) device information (e.g., vendor, data elements collected, etc.) and agency specific data collected. The framework could be used to model the impact of a cyber security event and estimate the overall cost. The framework could also be used to assess the impact and risk of configuration changes (i.e., risk management assessment). Finally, the framework could provide a "single glass view", showing assets, along with the current risk assessment.

The participants discussed the value of incident management and divided incident management data into three categories: hindsight (data from sensors and historical data; this is data that is currently available); insight (e.g., streaming data, social media data, situation awareness information; this is data that is currently available); foresight (converged data that can be used for prediction and prevention; this data is created by converging and analyzing the insight and hindsight data).

At the end of the discussion the participants were asked to rank the Big Data V's based on importance. Approximately 16 people participated in this ranking. The results are shown below and indicate that Volume and Value are the most important to those participating in this survey. The survey indicates that Volume is twice as important as Validity, Visualization and Variety. Government CIOs also put a ton of value in keeping collecting and analyzing data.

2.2.3 Important Findings

- There is a need for a special education tract to teach unique cyber security skills.
- CIOs and Chief Information Security Officers (CISOs) should work closely with the business units to assign "value" to assets, starting with the mission and business objectives. CIOs and CISOs should also better inform the business units on what can (and cannot) be done with respect to cyber security data analysis.

- Due to the large number of activities that could be performed government agencies need to prioritize their cyber security activities.
- System vulnerabilities include software, hardware and configuration vulnerabilities.
- Data analytics results are at the mercy of the veracity of the data collected at the end points. For example, the accuracy of the log data is very important.
- Sensor data tends to be vendor specific. Similarly, agency data tends to be agency specific.
- Participants stated that a "single glass view", showing assets, along with a current risk assessment, would be helpful.
- Policies should be reviewed to determine if changes are required to allow government agencies to share configuration data used to implement and secure virtual machine environments.
- Participants ranked the 7 Vs and determined that Volume and Value are the most important to their environments.

2.3 Big Data as a Catalyst

The "Big Data as a Catalyst" session discussed how to enable more government and private uses of data so the data can be a catalyst for better effectiveness, knowledge, etc.

The session included discussions of the following:

- What policy and technical capabilities are needed to enable greater amounts of useful data sharing?
- How can the government broker non-government data to facilitate its use?
- What capabilities should automated analytics have?
- What issues arise regarding protection of sensitive information

2.3.1 Challenges

- Tools and resources should be available to potential users to research the existence of data and gain access to the data.

- Users will want unstructured data with great "variety" to be formatted so those data can be searched conveniently.
- To enable greater sharing, data-security policies need to be consistent across agencies, yet still satisfy agency-specific needs.
- After individuals give permission for their data to be used for a specific purpose, a process is needed so other users can get permission to use those data for an alternative purpose. This process is often tedious and must start at the beginning, a streamlined process would be beneficial for widespread and effective access.

2.3.2 Discussion Summary

The group was not concerned about the exact definition of "big data", only that society produces more and more data, so it is desirable to enable the uses of data. Most of the discussion focused on goals, policies, and processes rather than specific technical solutions. The most prominent theme was how to handle data sharing.

Whether data sharing is to be done among government agencies or between the government and the public, potential users of the data need to know that the data exist and where to go to access the data. Government data assets should be registered so users can go to a website to find data of interest. When possible, data should be posted on publish-subscribe systems to help standardize data-access methods. To motivate agencies to make their data available, data management should be included in government-program budgets, in order to cover the cost of data "shipping and handling".

Merely agreeing to post data is not enough to make data accessible; users (government and private) also need to know how to find data of interest among the posted data. Variety of data formats has increased as data have become "bigger". Formats that label their fields enable searching even if the data are not stored in a traditional database. Data recorded in free-text should be stored so they are searchable by using modern text-search methods.

There are similarities between open-source software and the government providing access to "open-source" data. Clearly, each empowers larger user bases than would be possible otherwise. Frequent use of software or data reveals problems, leading to improvements in the form of upgraded software or changes to data-gathering and posting methods. A key difference is that while posted software can be improved by other programmers, data providers still own the data, implying responsibility to ensure the veracity of those data, and for enhancing future data gathering.

NOAA has partnered with several cloud-computing companies to allow them access to NOAA's massive store of weather and other data. When there are not sensitive data to protect, such partnerships allow these companies to extract valuable information which would not exist without the partnerships. The government ensures transparency and equal access; the companies provide public access and provide value-added services (for a profit).

The Departments of Transportation and State also have significant stores of, respectively, vehicle-movement data and social-media data. These data stores record measurements or events that occurred, allowing users to do searches and compute analytics on the data. Such analysis allows the State Department to receive public feedback on its activities. The group did not discuss these departments' use cases in detail.

In several situations, non-government organizations hold data that could benefit the public if data from many organizations could be accessed and analyzed. However typically those organizations do not want their data to be accessible in a way that would reveal information about their organization in particular. Currently the government gathers cyber-incident information to facilitate nationwide cyber defense without endangering the businesses and agencies that report the incidents. Similarly, the Federal Aviation Administration gathers in-flight incidents, such as near misses, for sharing without attribution. Also mentioned was National Science Foundation sharing of researcher data - but researchers want to publish their results ahead of research competitors.

In the near future, there is huge potential for brokering of medical data, for improved understanding of public health, and for enhancing research into medical diagnoses. However there are significant risks of hacking, and a high financial liability for revealed data. Hence an alternative to the government gathering and storing the data is facilitating the secure sharing of the analytics to be applied to those data, where only the results from the analytics would be returned to the person doing the analysis; that person would not have access to the raw data. This model could be applied to other data that are distributed across many sites.

For big data to be a catalyst for value to government and private entities, enabling access to stores of events is not enough - analytics software also must be available so users can extract useful information from the data. These analytics need to include descriptive methods (say what is), predictive methods (say what will be), and prescriptive methods (say how things should be). For data which will be used by many non-technical users, the organization posting the data should make common analytics available, much as a banking website provides tools for calculating basic metrics of one's financial situation. The group went further than this, proposing easy availability of pipelines of analytic processes, on-demand data fusion, even "Watson cows" to graze through text information.

Typically each government agency has its own policies that restrict data sharing due to various risks, not just for protecting sensitive information. If the government had broad policies used by many agencies, with not too many exceptions and add-ons, users would be able to satisfy policy hurdles more consistently. Even better would be for government policy to include incentives to encourage agencies to knock down barriers to information sharing. That said, different types of data have different data-governance requirements, but several policy templates could be written which, together, would cover most situations. We mentioned the example of fishermen being reluctant to reveal their locations to the government for environmental-analysis purposes.

Often when individuals and businesses provide data to the government, they give consent for those data to be used for specific purposes, but not for other purposes. As data sharing increases, more and more often users will want to use data for purposes other than the original purposes for which the data providers gave permission. The idea of re-doing data-gathering consent was mentioned, but no one suggested how to implement this idea.

Output from big-data analytics must be examined to ensure it does not reveal information about specific individuals or entities. One prominent member of the group stated that current methods of encryption and anonymization are not sufficient to protect against this reverse engineering. The Census Bureau has had similar issues for many years, so their techniques may be applicable for other organizations.

2.3.3 Important Findings

- Much policy and integration work is needed to enable technical implementation of greater data sharing.
- For data which will be used by many non-technical users, the organization posting the data should make common analytics available, in a user-friendly manner.
- While several agencies have implemented data sharing, most of this work has not included cross-agency consistency needed for use cases such as publicizing the existence and formats of data, data search, and automated analytics.
- Partnerships between government and industry have been demonstrated as useful for making government data more available for public use.
- Partnerships between government and industry have been demonstrated as useful for using non-government data for the public good while protecting attribution and other privacy concerns of data providers.

2.4 Leveraging Big Data for Mission Success

The Leveraging Big Data for Mission Success session discussed a wide range of factors that either facilitate or hinder the use of Big Data and the associated data analytics to accomplish mission outcomes.

The session included discussions of the following:

- What is the Big Data landscape (definitions, context, players, technologies, tools, etc.) from which to cultivate mission success?
- What are some exemplar use cases?
- What are the key challenges in employing Big Data for mission success?
- What opportunities exist to resolve those challenges effectively and efficiently?
- What key findings and recommendations emerge from the collaborative discussion of those questions?

2.4.1 Challenges

- Aggregating and combining disparate data source effectively, this includes struggles with collaboration (including providing access to external partners), extracting data from large, unstructured sources, and assessing the relative benefits of centralization versus decentralization of of data, tools, and human resources.
- A novel and pressing issue across government agencies is the reliability and utility of social media and other "open" data sources.
- Across the government, there are issues with budget, funding, and opportunity costs to support Big Data. Additionally, there are concerns with the resistance to change and cultural hurdles within government organizations.
- There is a continued scarcity, relative to growing demand, of trained and skilled personnel.
- Differences in agility are a challenge when leveraging big data for missions success. There is a struggle with the ability to exploit innovations between government and commercial entities.

2.4.2 Discussion Summary

The session participants first discussed the workflow for Big Data analytics versus more transactional technologies, including the need - for most non-trivial government missions - to combine several or all of the major types of data analytics:

- Descriptive analytics – What happened?
- Diagnostic analytics – Why did it happen?
- Predictive analytics – What will happen (trends, projections)?
- Prescriptive analytics – What to do about what did or will happen (optimization)?

Beyond the need to understand how, where, and when to employ those types of data analytics, the participants agreed that trustworthy data is critical to all involved. This is especially important for predictive and prescriptive analytics since they involve probabilities and decisions and sometimes direct actions in the operating environment. The trustworthiness discussion identified the need to have stakeholder buy-in for access control, data sharing, privacy and so forth. This led to consideration of data granularity (e.g., atomic or aggregate) and stakeholder perspective (e.g., producer, consumer, regulator, affected third-party, etc.) - and similar factors - in making specific decisions in specific use cases.

Consequently, the group outlined several use cases for study against the identified challenges. The first use case discussed was the USDA Risk Management Agency (RMA) assessments of producers (e.g., farmers) and lands for crop insurance and benefits purposes. Producers must be evaluated on factors like efficiency and experience while lands must be evaluated on factors like historical production record and location. Each such factor might have multiple related factors that must be included - e.g., land location might relate to weather, supply chain efficiency, and overall economic viability. Such calculations can, then, require complicated analytics to arrive at accurate, reliable, and comparable assessments. Assessors need appropriate and capable tools to perform these calculations, above and beyond access to current and reliable data.

The second use case involved a government mission to perform rating of financial institutions for consumer protection purposes. This mission is very concerned with imperfections introduced in the data collection process (e.g., biased consumer reports and institutional claims) and the challenge of finding credible patterns in such an environment due to the (potentially high) signal-to-noise ratio. Hence, data quality - from provenance onward - is critical. Ensuring that data analysts and their algorithms are asking the right questions (e.g., from the management versus the consumer perspective, and vice versa) is, therefore, very important in this environment.

The third use case discussed dealt with medical device performance - e.g., the durability and effectiveness of prosthetics used in knee replacements. Useful analytical applications would span the descriptive through predictive space for this mission, with coherence across those interdependent analytics being both difficult to ensure and essential to credible results. Additionally, some potentially critical data - such as patient lifestyle and patient medical conditions over time - could be very difficult to obtain, while other related data - such as the surgeon(s) and hospital(s) involved might be more readily available. Raw availability, however, then leads to issues related to privacy, defamation (libel), culpability, etc., that can have major consequences for various parties. Public clearing houses for relevant data (such as the National Center for Health Statistics, NCHS) that follow professional data curation practices - such as anonymization and masking - can play an important role in leveraging Big Data for mission success in such use cases.

2.4.3 Important Findings

The discussion of the use cases - in the context of the Big Data landscape and the identified challenges for leveraging Big Data for mission success - led to the following findings and (sometimes implicit) recommendations:

- Data availability is a double-edged sword; we often do not have the data and when we often can't trust it when we do have it.
- The empirical demonstration of trustability is critical; open communication and transparency are critical.
- Data governance requires a combination of policy, process, people, and technology.
- To successfully leverage big data for mission success, both data scientist types and data-savvy business types (mission leaders, operations) of personnel are required. Organizations must maximize internal training and information sharing opportunities to realize a complete workforce.
- Organizational and bureaucratic power struggles present major roadblocks to success and must be resolved expeditiously. The perception of favoritism and institutional differences can complicate public-private data sharing (including analytical results). Consequently, new and smaller vendors face steep barriers to entry.

- Some government missions might be so specialized (with small market value) as to preclude vendor investment.
- There is a general lack of big data management and analytics tools on Approved Products Lists (APL) and a general lack of Technology Readiness Level (TRL) and Assessment (TRA) guidance for these tools at this time. Accountability for use of Big Data analytics, the source data, the algorithms used, and the resulting decisions made and outcomes achieved must be assigned and accepted.
- In order to build upon past work and to continue marketing our successes, we need to get the word out about productive solutions, best practices, etc.
- The next Federal Big Data Summit should include a session on practical solutions to specific data processing problems related to Big Data e.g., document decomposition via text analytics for data structuring.

2.5 Big Data and Analytics in Health Care

The Big Data and Health Care session facilitated discussion on big data and analytics' impact on health care. "Big Data" is a broad term that represents the coordinated use of diverse technologies. There are complex challenges that must be prioritized and addressed to effectively embrace Big Data strategies within an organization. This session examined the following characteristic categories defined by the National Institute of Standards and Technology Big Data Public Working Group (NIST BD-PWG) as they relate to Big Data within the health care environment:

- Data Sources (e.g., data size, file formats, rate of growth, at rest or in motion)
- Data Consumer (e.g., processed results in text, table, visual, and other formats).

The participants in this session hoped to identify:

- How helpful is the NIST Big Data Reference Architecture in supporting actual implementation?
- What are the necessary steps for planning and adoption of health data sharing?
- Are common data models being implemented in government organizations? If so, how did you establish the necessary processes to go from conceptual to production?

- What are the pros and cons of the current data models used across organizations?
- How do organizations fulfill requirements from the data consumer perspective?

2.5.1 Challenges

These discussions identified the following challenges:

- The NIST Big Data Reference Architecture is an excellent starting point for organizations but is lacks adoption and misses the interoperability component of data.
- When organizations agree to share data, the agreements need to be clear and open about the data elements to be shared, the business requirements, legal and security considerations, and the governance of the data.
- Various organizations have implemented common data models, these models are not the same across government organizations which may lead to difficulties in sharing and coordinating data.
- Some data models are too federated and are not centralized, which leads to difficulties in obtaining real-time data. Other models can be too limited in scope to work across organizations or too broad to meet the specialized needs of other organizations.
- Legacy data is in different information models and lacks documentation for different iterations.
- A standard ontology is needed to overcome issues with varying data definitions, versions and iterations.

2.5.2 Discussion Summary

The Big Data and Health Care Session focused on the challenges big data and analytics implementers have in health care environments. This session mainly focused on characteristic categories defined by the NIST BD-PW as they relate to data sources and data consumers.

The collaboration session opened with a discussion of the NIST Big Data Reference Architecture. Very few participants were aware of the NIST Big Data Reference Architecture and felt that it is not prescriptive and is missing the interoperability component of data. Participants noted that they use Hadoop to implement their big data solutions.

The next topic for discussion was the steps necessary for the planning and adoption of health data sharing. The necessary steps include: developing common metadata to provide consistency in definitions, implementing standard terminologies, and developing common data elements (CDE). In addition, organizations need to develop policies and procedures for implementing data governance, security, privacy, and compliance.

Participants also discusses the use of Common Data Models (CDM) for health care. Many organizations have implemented a CDM. The VA and DoD have implemented the VA Informatics Computing Infrastructure (VINCI), the Veterans Information Systems and Technology Architecture (VistA), and the Observational Medical Outcome Partnership (OMOP). The CMS Common Data Model is intended to chart the location of, and relationships between, common data elements in CMS' various IT systems. The IBM Unified Data Model for health care is also used by many commercial entities. Organizations were able to move from conceptualization to production by adhering to the following steps:

- Requirements gathering
- Map requirements to data
- Define a data dictionary, data governance, and procedures
- Establish processes and policies for access, data security, and privacy
- Develop the logical design
- Develop the physical design
- Perform data masking and scrubbing

The discussion also covered the benefits and shortcomings of the data models discussed. According to the session participants, the DoD and VA implementations is very federated and is not centralized enough to provide access to real-time data. The OMOP CDM is limited in scope but it has the capability to add extensions. Finally, participants remarked that the IBM Unified Data Model for health care is encompassing but too broad, it needs to be narrowed down to meet business requirements.

From the data consumer perspective, organizations recognized that they still face challenges in making sense of the data, working with legacy data, and developing reports in a timely manner. The participants recommended that organizations develop a standard ontology to overcome issues with definitions, versions, and iterations.

2.5.3 Important Findings

- Not many participants were aware of the NIST Big Data Reference Architecture and felt that it is not prescriptive and is missing interoperability component of data
- Development of a standard ontology can overcome some issues with different data definitions, versions, and iterations
- The major challenge in fulfilling the data consumer requirements is developing reports in a timely manner, and getting the right data to the users when needed.

3 Summit Recommendations

Across all of the challenge areas, participants noted several important challenges to the use and adoption of big data technologies: the need for information and data sharing in a secure, trusted, open manner, the struggle to provide reliability in data sources including liability and ethical considerations, the importance of testing frameworks, and the desire for organizations to recognize the importance of big data and prioritize training work through the resistance to change and cultural hurdles within the government.

Every collaboration session recognized the importance of data and information sharing. This has consistently been considered a challenge in these summits and is a complicated area. One of the biggest challenges is sharing information across government organizations. Standard practices, behaviors, and policies help to ease the process, but the issue has not been resolved. Organizations need to set up tools and resources for users and provide consistent data use policies across agencies to reduce the hassle of data management for users. When organizations agree to share data, the agreements need to be clear and open about the data elements to be shared, the business requirements, legal and security considerations, and the governance of the data. Agencies need to work towards standard ontologies for data definitions to bring clarity to the field. Government organizations will continue to struggle to collaborate on data-centric projects until aggregating and combining disparate data sources becomes simplifies. For topics, such as cyber security, there is a limited amount of sharing across agencies, vulnerability data is occasionally shared, but useful configuration data is rarely communicated. Finally, government agencies need to recognize and plan for future difficulties in utilizing and assessing the reliability of social media and other open data sources. Several collaboration groups also called out the need for data sources to be reliably managed. This includes the need for management, governance, privacy, security, and ethical considerations of the data. The discussion on autonomous systems and autonomy had particular concerns about the proper way to determine liability in the case that an autonomous system causes harm. Autonomous systems, as well as health care data require the governance to make ethical considerations about the use of the data. Data governance is one of the biggest responsibilities when it comes to big data management for government organizations, and this directly relates to the data use and sharing concerns previously discussed.

Testing frameworks are extremely important to big data, these frameworks allow users of big data to investigate how processes will work and determine the strengths and weaknesses of their products. Participants recognized that although many organizations have implemented common data models, these models are not "one size fits all" and organizations need to be experiment with what works best for their organizations. Testing frameworks allow organizations to experiment with prioritizing data which is incredibly important with the tremendous amounts of data many organizations need to sort through. The ability to examine important data elements and proper data formatting is necessary to work quickly and efficiently with big data. Finally, for autonomous systems, the government and industries are lacking a solid framework that allows for testing systems capable of learning and self-modification.

The cyber security and missions success sessions also recognize the need for proper training and education about big data technologies. Skilled and adequately trained staff are necessary to work with these large volumes of data and to make continues progress. Specialized data science training should be provided and supported by government agencies.

The important findings in the sessions closely aligned with the challenges. The session leads recognized imperative guidance is needed in regards to data sharing, data governance and reliability, and the need for education and cultural changes.

All sessions provided guidance for the needs concerning data and information sharing across the government. The autonomy and autonomous systems sessions determined that the government needs plans and visions in place as soon as possible for the acceptable use of autonomous systems. The cyber security sessions reported agencies should review their poliies to determine if changes are required to allow government agencies to share configuration data used to implement and secure virtual machine environments. Policy changes are a driving force behind big data as a catalyst. Policy and integration work is needed to enable technical implementation of greater data sharing. While several agencies have implemented data sharing, most of this work has not included cross-agency consistency needed for use cases such as publicizing the existence and formats of data, data search, and automated analytics. This session also noted that partnerships between government and industry have proven useful for making government data available for public use. This public data needs to be accessible and available for non-technical users. The Big Data for Mission Success session recommended that the next Big Data summit include a session on practical solutions to specific data processing problems related to big data.

Government agencies need clear plans and visions for upcoming areas in data science including autonomous systems, managing and maintaining the trustability of data, and health care data. Autonomous systems are becoming more and more prevalent in research and policies are needed to ensure ethical and safe standards are in place for this growing field. With the variety and veracity of data currently available, it is becoming increasingly difficult to trust incoming data. Data governance requires a combination of policy, process, people, and technology. Cross-agency standards need to be in place for health care systems to ensure collaboration and compatibility in the future.

Several sessions also noted that government agencies need to begin prioritizing the need for training and skilled staff in the workforce. Concerning autonomous systems, the government should invest in appropriate research, influence international policy, and drive commercial efforts appropriately to ensure autonomous systems do not disproportionately disadvantage the general population. There is also a need for specialized cyber security education programs to teach these unique skills. Many of these programs are already in place and government agencies need to recognize the value of these skills and encourage employees to participate. Finally, in order to successfully leverage big data for mission success, both data scientist types and data-savvy business types of personnel are required. Organizations must maximize internal training and information sharing opportunities to realize a complete workforce.

4 CONCLUSIONS

The December 2016 ATARC Federal Big Data Summit reviewed many challenges facing the federal government's adoption of big data technologies and the progress in this area. These challenges spanned multiple collaboration areas and were widely discusses by all groups, as well as during the morning's panel sessions. Specifically, information and data sharing is lagging behind the current needs of government agencies, organizations struggle to provide reliability in data sources, the importance of testing frameworks, and the desire for organizations to provide training for employees were all noted as continuing challenges in big data.

the importance of providing structure and guidance for data sharing, data governance and reliability, and the need for education and cultural changes were discussed by a majority of participants.

While the December 2016 Federal Big Data Summit highlighted areas of continued challenges and barriers to progress, the Summit also cited provided guidance for what changes need to be made in the government to improve how we work with big data. Government organizations need to recognize he importance of providing structure and guidance for data sharing and put policies in place to allow this to happen quickly and efficiently. Organizations need to establish data governance policies that ensure trust and reliability. Finally, organizations need to have a culture in tune with big data technologies and should provide education to crete a skilled and knowledgeable staff all the way from data scientists to the CIOs.

ACKNOWLEDGMENTS

The authors of this paper would like to thank ATARC and The MITRE Corporation for their support and organization of the summit. Special thanks are in order to Justin Brunelle and Patrick Benito for their contributions in reviewing the paper.

We would also like to thank the session leads and participants that helped make the collaborations and discussions possible.