FEDERAL SUMMITS

JUNE 2015
FEDERAL BIG DATA SUMMIT REPORT*

September 23, 2015

Zachary Firth, Christine Harvey, Danny Moore,
Jim Soehlke, Irina Vayndiner, and Sanith Wijesinghe
The MITRE Corporation†

Tim Harvey and Tom Suder
The Advanced Technology Academic Research Center

†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 Federal Big Data Summit, held on June 18, 2015, included five MITRE-ATARC (Advanced Technology Academic Research Center) 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. The goal of these sessions is 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 Analytics to Enhance Mission Capabilities, Predictive Analysis, the Future of Big Data and the Internet of Things, the Roadmap to Big Data, and Big Data for Cyber Defense.

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:

- The government needs to establish a strong, supportive foundation for big data, which means informing senior leadership of the technical needs and opportunities that come with big data.
- Federal agencies need to implement formal standards for governance, provenance, and security. Standards are necessary to ensure data is traceable and clearly defined.
- Communication and collaboration are essential to share best practices within and beyond the government. Agencies that have experienced success need to share this knowledge.
- Academia needs to develop programs and supplemental curricula to produce big data talent with skills in analysis, data mining, database management, and related skills.
- Both academia and government organizations need to work to develop tools that can capture, store, and analyze big data.

1 INTRODUCTION

During the most recent Federal Big Data Summit, held on June 18, 2015, five MITRE-ATARC (Advanced Technology Academic Research Center) 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 academic 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 Analytics to Enhance Mission Capabilities
- Predictive Analysis
- The Future of Big Data and the Internet of Things
- The Roadmap to Big Data
- Big Data for Cyber Defense

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

2.1 Big Data Analytics to Enhance Mission Capabilities

The Big Data Analytics to Enhance Mission Capabilities session discussed the unique challenges, benefits and current state of using big data to improve duties and operations within the government. The session included discussions of the following:

- Determine how data and analytics can be used to enhance an agency's mission and deliver more effective results.
- Identify ways sophisticated data and analytics tools be used to harness and review information.
- Identify barriers of adoption for Big Data Analytics in mission capabilities.

2.1.1 Challenges

- Agencies should standardize data representation and develop a collective understanding of the intent of each standard.
  - Varying standards often increase overall processing time, which in turn reduces the time available for analysis.
  - A lack of understanding among data providers of the importance and reasons behind standards adversely affects data quality.
- Organizations need better data traceability and metadata. Agencies must capture the business rules in place at the time the data were gathered. This information prevents comparison of incompatible data and allows analysts to account for historical variations.
- Each agency should hold an open dialog about the risks and benefits of varying data capture and retention schemes. This dialogue should cover:
  - Cost of storing data
  - Risk to the public, private sector, and agency if data are lost, stolen, or otherwise compromised
  - Benefits and risks of providing data as a self service option

- Agencies need to build an understanding of how unfunded mandates drive data collection. Such mandates could lead to the reallocation of analytic staff to tactical vs. strategic tasks that may not directly apply to the agencies mission.

2.1.2 Discussion Summary

While participants intended to outline recommendations for how to leverage big data and analytics to improve mission capabilities, they quickly recognized that many of the challenges agencies face occur well before data analysis can take place. The group believed that in many cases the ability to store data has diminished the perceived need to think critically about value, source, and characteristics of the data captured, even though those factors will ultimately drive analysis.

Participants generally agreed that the first step in leveraging big data consists of determining what data are valuable to the mission. While some believed that agencies should capture and retain as much data as possible, the majority considered this unnecessary. While the financial costs may be low, this practice may create additional risks from a security perspective. Considering the value that each piece of data provides to the mission allows stakeholders to balance the risks and benefits of capturing and retaining datasets for real-time and retrospective analyses.

Session participants cited the numerous sources and formats in which federal agencies receive data. The group agreed that maintaining data pedigree was critical to ensuring data are used correctly. That pedigree becomes especially important in a regulatory setting where the use of the data may only be permitted in support of a narrowly defined mission. With varying sources also come multiple data formats. While in most cases the data format does not prevent analysis, the format can make data ingestion much more time consuming. Several participants cited instances where disparate formats created a bottleneck that limited the ability to preform novel analytics that would have benefited the agency's mission.

The group agreed that the characteristics of the data ultimately drive the analytics that can be performed. Attributes such as volume, variety, veracity, velocity, variability and volatility often make analysis more difficult and compress the window in which results remain relevant. Participants generally agreed that failing to account for any of these variables puts agencies at risk of making decisions on outdated or inaccurate information.

Despite general consensus that the above factors are important in successfully leveraging big data and analytics to enhance mission capability, several participants believed that these steps are often an afterthought and that agencies place more emphasis on the analytic software that will be used and the reports that will be generated. Several others cited instances

where this type of thinking has limited their agency's ability to fully leverage its data in support of its mission. Participants went on to stress the need for a strong analytic foundation that begins with relevant data that are understood and thoroughly documented. Only then can agencies confidently leverage their data for transformational change.

2.1.3 Important Findings

- Agencies must ensure the data they capture is relevant and aligned with the agency's mission.
- The data must be completely understood and clearly defined. The organization must understand the source of the data as well as the format and interval in which it is received.
- Organizations need to understand the characteristics of the data (e.g., volume, variety, veracity, velocity, variability and volatility).

2.2 Predictive Analysis

The Predictive Analysis session sought to outline recommendations and best practices for harnessing big data in a useful way to perform predictive analysis. Participants hoped to:

- Identify how predictive analysis can help agencies detect adverse behavior.
- Recommend improvements in predictive analysis to detect healthcare anomalies, natural disasters, or other difficult-to-forecast events.
- Provide success stories and best practices for using big data to perform predictive analysis.

2.2.1 Challenges

These discussions identified the following challenges:

- Agencies need to implement and/or improve data governance to ensure better data quality. Participants cited a Master Data Management (MDM) strategy, standardization of common domains, and identified data stewards to manage these domains as components of data governance that are lacking within the Federal data environment.
- Agencies should increase accessibility to essential data stores.

  - The federal government needs to identify ways to simplify sharing of data among governmental agencies/organizations at all levels (city, county, state, and federal).
  - The government must also improve methods and practices for sharing data with non-governmental agencies/organizations.

- The government needs to increase the pool of expertise in Business Intelligence (BI) tools, big data technologies, and software development for data mining and machine learning. Teams need to be made of a diverse group of individuals that are knowledgeable in statistics, analysis, data management, and distributed computing.

- Agencies need to raise senior leadership's awareness of predictive analytics and big data capabilities.

  - Senior leadership must understand that a "data scientist" may not be simply an individual position, but an interdisciplinary combination or team of personnel.
  - Organization need additional assistance in overcoming cultural boundaries that hamper the implementation of predictive analytics, such as predictive analytics that fall outside the organization's direct line of business.

- Organizations need a streamlined process for accessing cloud resources across all federal government organizations.

2.2.2 Discussion Summary

The initial discussion focused on the challenges that agencies face as they seek to implement predictive analytic solutions using big data. Some of these challenges were consistent with those associated with traditional predictive analytics, such as a lack of an effective data governance strategy and difficulty accessing data from internal/external governmental agencies and organizations. However, big data raises additional challenges. This includes a shortage of skilled big data analytic talent, the difficulty some agencies still encounter in obtaining access to cloud computing resources, and the general culture of government organizations that can hamper initiation of new efforts.

With respect to application of predictive analytics and big data to predict adverse behavior, the discussion recognized instances of positive utilization in fraud detection and within the healthcare environment. Predictive analytics in applications such as forecasting natural disasters may be more problematic, because events such as earthquakes and tsunamis are rare, current sources of data provide little ability to predict these disasters.

Participants suggested that some agencies could use predictive analysis to facilitate better utilization of their facility and process inspectors by identifying sites at higher risk of non-compliance. This would allow agencies to assign their inspectors more effectively.

Finally, the session covered success stories and best practices. The group identified the Obama election campaign as an example of successful utilization of big data and predictive analytics. Another area beginning to emerge as a successful implementation is individualized, precision medicine.

2.2.3 Important Findings

- Agencies lack effective data governance to improve the veracity and accessibility of the data.
- A shortage of available of technical labor to implement solutions exists and hinders progress.

2.3 The Future of Big Data and the Internet of Things

The Future of Big Data and the Internet of Things session was intended to explore the diverse and difficult challenges and current state of the Internet of Things (IoT) and big data within the government. Participants in this joint session sought to:

- Determine how agencies can manage the increasing volume of data, with all of the devices currently contributing to the amount of data.
- Recommend effective ways to manage data privacy and ownership of the increasing amount of data.
- Identify and recommend how to merge historical data with present and future IoT data.

2.3.1 Challenges

- The government should fund infrastructure for storing, processing, and analyzing IoT data. This includes new architectures or upgrades to the existing systems in order to accomodate such data.
- Agencies must define and establish rules and regulations to set standards, enable innovation, and protect data related to IoT. Standards are necessary to manage IoT disclosure and ownership, and the government should put a full suite of policies in place to ensure security and privacy as related to IoT.

- Agencies need to determine how to apply analytics to new streams of data. This includes performing real-time big data analytics.
- Agencies must identify what IoT data has value for the government mission and how frequently it must be updated. Organizations within the government also need to understand the new opportunities and challenges of integrating historical data with IoT data.

2.3.2 Discussion Summary

Session participants recognized that government agencies need to be proactive in adopting big data technologies and developing skills related to the IoT. Billions of internet-connected "things" will generate massive amounts of data, and the government must understand and act upon that data [1]. This requires collaboration among government, industry, and academia to agree on rules and regulations that set standards, enable innovation, and protect data. As in other sessions, participants suggested that agencies with more experience in dealing with streams of data and machine to machine (M2M) communications should advise those newer to the field.

Much of the discussion in this session focused on the need for government agencies to prepare to meet the IoT big data challenge by establishing the infrastructure for storing, processing, and analyzing IoT data. Participants noted that funding the infrastructure presents one of the major challenges for this work. Agencies will have to identify the IoT data that has value for their mission, justify why they collect this new IoT data, and define the questions or problems they plan to address. Furthermore, agencies must recognize that more data does not always mean more value.

Participants noted that the government must be ready to process and store multiple new steams of data and integrate this information into the existing (historical) agency data. This might require building new architectures or upgrading existing ones to handle big data from the IoT in real time. Government agencies also needs to determine how to apply meaningful analytics to the new data streams, including real-time big data analytics. One of the factors in this area that concerned session attendees was the high cost of storing and analyzing IoT data. IoT data storage should be optimized; this includes the frequency of updates and the duration of data storage. Organizations should select big data systems that are most suitable for IoT data or develop such systems themselves.

As in other sessions, the participants discussed ownership and stewardship of IoT data, which is especially important for highly regulated industries as well as government agencies.

Many agencies already have regulatory standards for data access in place, but some agencies have not yet established standards specifically governing IoT data. Obtaining the rights to use the data might add complexity and possibly cost to the process. The also recognized that agencies must address concerns related to the quality and accuracy of IoT data. This includes identifying parties responsible for cleaning the data in situations where the data is collected in one place and then transferred to another organization.

Privacy and security issues raised particular concerns. The government should develop best practices and standards on how to utilize and manage IoT data. Using devices to collect IoT data requires that agencies put new measures in place to ensure privacy. Session participants also noted that as people become more aware of the value that IoT data generates, their desire to become more active participants in decisions about what data is collected and how it is used will also increase.

2.3.3 Important Findings

- The government needs to collaborate with industry and academia to find workable solutions to IoT data challenges and discover opportunities.
- Agencies have an urgent need to address security and privacy concerns as they relate to the use of IoT data in government.
- Agencies need big data infrastructures that can capture and store IoT data of extremely high volume and velocity.
- Organizations must develop new analytical tools and applications to handle analytics for IoT data.

2.4 The Roadmap to Big Data

The Roadmap to Big Data session reviewed challenges and best practices relevant to developing the workforce and technologies necessary to manage and perform big data research within the government. This joint session sought to outline recommendations for building big data projects from the ground up and through discussion of the following:

- Identify how the government can improve the way big data initiatives are planned, launched, and sustained to realize the most value.
- Recommend the tools and technological developments needed to advance the impact and use of big data.

- Discover ways to counter the workforce shortage among big data and analytics professionals.

2.4.1 Challenges

- The government confronts a shortage of employees and teams skilled in data analytics. Many agencies acknowledge the negative impact of this skill gap and realize that IT employees are not sufficient substitutes for big data analytic talent.
- Agencies must establish clear goals and outcomes of big data projects. Many previous efforts have amounted to a "leap of faith" in the hope of garnering benefits.
- Federal agencies need to identify ways to work with the commercial sector. For example, one agency has developed a relationship where the private sectors use the content in exchange for providing storage.
- The government has many needs in the areas of data management and storage. Data must be cataloged in a format that allows easy identification of existing data and metadata.
- Agencies need to appropriately identify whether data should be stored in house or outsourced. They must also recognize the funding constraints associated with different management techniques.
- A paradigm shift has occurred in thinking about what analytics can do for an organization. To launch a big data initiative, an agency must define what big data means to the organization and to individual components of the organization.
- Organizations with an established history of handling big data should share their lessons, best practices, and knowledge across agencies including how they manage and work with their data. Agencies should also share more of their metadata, algorithms, and automatic performance dashboards both externally and internally.
- Requirements for big data tools and technologies are often overstated, and should be realistic.
- Agencies need to find ways to outsource or take advantage of available tools when they need interface tools for access. Agencies also lack administrative tools for finding "hot spots" in the data. Often organizations only have laundry lists of tools and technologies

relevant to big data; they need a more complete guide to these tools and the advantages and disadvantages of each in order to make a complete decision on the proper path to follow.

- In shaping a big data workforce, agencies need to form teams with a stratification of skill specialization. Not everyone on the team has to be a data scientist; effective teams combine these skills with other specialties.
- Organizations need to look beyond computational science graduates to build the work force. For example, specalists in many areas of science have similar skill sets, such as biologists or physicists.

2.4.2 Discussion Summary

Federal agencies have gone beyond the first stages of adopting big data technologies and skills. Agencies have become more realistic in planning and framing their projects by drawing on resources across other agencies and outside sources that can provide key insights to facilitate the onboarding of big data. Participants in the session devoted most of the time to the first topic, which centered on improving adoption initiatives. The pragmatic discussion focused on data preparation, lack of available expertise, and a renewed emphasis on sharing information about big data use and management. Several participants noted the importance of a data catalog and metadata dictionary. These key references provide an essential compass for adoption of best practices. Roadmaps seem to focus primarily on implementation while maturity models focus more on sustainment. Both are needed for each phase in making optimal use of big data.

When the discussion turned to tools and technology development, the group recognized both definite gaps and opportunities. Some of the foremost needs center on data management and stewardship. The government must make use of some new tools in big data to mitigate some of the more difficult aspects of performing big data research. Acquisition is also an important part of the equation. When agencies outsource services, providers must supply complete solutions and contracts should hold them accountable for meeting their requirements.

Finally, in dealing with the reality of a workforce shortage, not everyone in IT needs to have this big data and data science skills. Agencies should select staff who have the appropriate motivation and aptitude and should invest in their career growth as they demonstrate interest in this area. Agencies should also ensure that these skills are widespread throughout the

workforce, and should nurture teams aligned to mission and business who understand how to use data for discovery and decision making.

2.4.3 Important Findings

- Roadmaps visualize and communicate action plans for using big data across the organization. These roadmaps must include milestones and descriptions to communicate progress and status.
- Agencies need maturity models of their strategy to sustain the effort. Organizations must be able to reflect on research progress to help senior management govern and engage in the adoption of big data technologies.
- Each agency should seek out success stories and a range of resources and tools that will better facilitate their success in using and sustaining big data.

2.5 Big Data for Cyber Defense

The Big Data for Cyber Defense session examined the unique difficulties and benefits of using big data to improve and manage cyber defense within the government. Participants sought to outline recommendations for using big data to manage cyber security risks and to:

- Identify the cyber security risks that affect an organization's mission objectives.
- Determine how big data could help address these risks.
- Recognize the factors that limit the application of big data for this purpose.

2.5.1 Challenges

- Agencies need updated security guidelines to evaluate and accredit:
  - Cloud-based services, especially to disentangle the architectural layers across which services are deployed.
  - Combinations of technologies that are integrated for purposes beyond their original individual scope.
  - Novel technologies that aggregate multiple disparate datasets to evaluate the potential for leaks of personally identifiable information.

  - Cyber-physical systems such as control systems that protect critical infrastructure, building sensors, and power plants.
  - Tradeoffs between application goals and security needs.

- The government needs to establish norms to share threat intelligence:
  - Across governmental organizational boundaries.
  - Within and between private sector industries/corporations.
  - Through additional technical standards such as STIX [5] and TAXII [6] type ontologies.

- Researchers must improve the maturity of cyber intrusion detection algorithms to:
  - Reduce false positives associated with larger datasets.
  - Adapt to evolving threat patterns.
  - Better identify sources of anomalies (software, user, system) to improve classification of data/analytical results.
  - Allow broader deployment/applicability across disparate operational environments.
  - Better integrate with other forensic tools/applications.

- Agencies need scalable infrastructures to harness high-velocity data (obtained via social networks, Twitter, or physical sensors) in real time to help develop and improve an agile response to threats and to ensure storage and bandwidth that can scale cost effectively.

- Organizations must make better use of big data for staff training applications. Agencies can use these technologies to simulate adverse incidents and evaluate the operational readiness of systems and staff.

- Agencies need staff with agile skills/tools who can apply novel technologies constructively and without the burden of historical organizational inertia.

2.5.2 Discussion Summary

The session identified the potential of big data to enhance analytical techniques and algorithms to detect malware, anomalies (from both external and internal sources), and the

activities of Advanced Persistent Threats (APTs). The resulting cause-effect insight can provide useful actionable intelligence to mitigate and/or eradicate a malicious cyber threat. Several technical and organizational factors currently limit the use of big data in this context.

While big data can provide a broader and deeper view of enterprise-wide events in terms of both variety and velocity, without a commensurate improvement in algorithmic fidelity it may cause high false alarm rates. Operational settings have minimal tolerance for false alarms and the associated impact on time and cost could lead operators to ignore alerts from automated algorithms. Agencies also need to ensure that their storage/bandwidth infrastructure for hosting big data can scale economically and does not represent a bottleneck in practical data access and analysis. Participants discussed the use of cloud-based services as a viable approach in this regard, but highlighted the need to ensure that cloud-based security protocols are adequate to avoid data and privacy breaches. The increasing interconnectedness of cyber-to-physical systems was cited as another source of big data that agencies could leverage to provide additional system situational awareness. However, participants commented that the associated sensors lack data broadcasting capabilities and that this presents a limitation for greater adoption.

Ironically, organization's hesitation to adopt big-data technologies for cyber defense arises in part from the lack of standards for evaluating their own security readiness. New application technologies have outpaced the required security guideline publications. Moreover, it appears that organizations lack knowledge about how best to evaluate the security risks that result from using a combination of software tools for purposes beyond their original scope. Participants also noted organizational factors limit the ability to share threat intelligence. For example, the lack of a mechanism to disseminate information about known security risks both within and outside of organizations limits the overall ability of agencies to combat security risks. The session participants emphasized the need to consider not only big data-based tools but also the combination of people and processes needed to provide an effective cyber defense posture. Organizations whose staff lack the required skills and resources to adapt to emergent threats are unlikely to survive the onslaughts from increasingly creative adversaries.

2.5.3 Important Findings

- Big data provides an excellent opportunity to improve situational awareness of an organization's operations and cyber threat landscape.
- Algorithms based on big data must consider the adaptive strategies of adversaries and

be designed to reduce overall false positive rates.

- Improved standards and security guidelines for big data technologies as well as to protocols to share threat intelligence among government agencies, can help improve the adoption rate of big data-based cyber defense solutions.

3 SUMMIT RECOMMENDATIONS

Several common themes recurred across all or many of the challenge areas. Participants noted three topics as having particular importance: the need for an engaged culture and leadership, standards, and collaboration in sharing data and best practices across organizations.

Big data offers many new opportunities to government organizations that can require changes in the way an organization operates or shares information. Effectively using big data has the potential to increase overall situational awareness, improve operational efficiency, and enhance mission capabilities. To take advantage of these opportunities, leadership should embrace these possibilities and changes. The culture and leaders within federal agencies must provide a strong and supportive foundation for use of big data. Yet the general culture of organizations can make it difficult to initiate new efforts, particularly in emerging areas such as big data. Senior leadership within the agency must recognize the technical needs associated with performing big data research as well as the opportunities.

Now that big data technologies have become established in many agencies, organizations must put strong standards for governance, provenance, and security into place. Rules and regulations are needed to ensure quality, enable innovation, and protect data. One of the main concerns with regard to standards was the need to adequately ensure data provenance, by making sure data is completely traceable, is clearly defined, has the necessary metadata, and has a proper pedigree. This becomes especially important when data is collected and shared across agencies and even within different sections of the same agency. Agencies must also have suitable standards for security of big data. They need appropriate guidelines for protecting the data and must understand the risks if data is lost, stolen, or otherwise compromised.

All five of the collaboration sessions at the summit recognized the need for more communication, collaboration, and sharing of best practices both across the government and beyond the government. Agencies can cut down on valuable research time and funding by communicating with other organizations and sharing algorithms and tools. Agencies need to seek out success stories and the types of tools and resources that help to facilitate their

use and sustainment of big data and associated best practices. Organizations with a long and successful history of handling big data should to share their roadmaps, lessons learned, best practices, and knowledge with other agencies. In addition to seeking information within the government, organizations should look outward to ongoing efforts and best practices in academia and industry.

In addition to sharing expertise, government organizations should seek to simplify the process for sharing data among agencies at all levels. Improved methods for collaboration and sharing will result in more efficient joint work and easier integration of data sources between organizations. This communication is especially important in sharing situational awareness and threat intelligence between organizations.

Several collaboration sessions discussed the need for big data analytic talent. The government has a severe shortage of the skills and expertise required to perform big data research and implement the necessary solutions to technical problems associated with mastering big data. The government needs an expanded group of researchers with skills and expertise in analysis, big data technologies, and software development for data mining and machine learning. Agencies must ensure that they form teams with the proper skills, including trained data scientists to perform research and analysis. The government needs to recruit and select staff for training who have the motivation and aptitude to become data scientists. Academia should adjust and refine curricula to produce graduates with experience in skills related to big data such as statistics, analysis, data management, and distributed computing.

Participants also recognized the importance of identifying the right data that will provide value to the organization. Government agencies need to determine which data will help them accomplish their mission and goals. The data captured must be correct, aligned with the mission, and relevant to the organization's goals. Some practices in big data include capturing and storing as much data as possible, and while many agencies have this ability, this does not necessarily represent the best way forward. An excess of irrelevant and unnecessary data can distract from the value provided in other data and can clog storage and processing technologies. Agencies should be able to justify why they collect particular data and which problems or questions they want to address with this information.

Several recommendations for the government centered on tools and technologies central to capturing, storing, and analyzing big data. The government needs tools and technologies that can process the data, identify "hot spots," and manage data. When government agencies plan to develop or acquire such tools, they must state their requirements clearly and require that providers deliver complete solutions. Engaged leadership and experienced data scientists can help contribute to the development and acquisitions of the right tools for the organization.

Participants noted a specific need for tools in the areas of cyber defense and the IoT.

One of the greatest challenges in data integration involves dealing with multiple data formats. The variety of formats makes integration extremely time consuming and can limit analytic capabilities. Agencies need the right tools and resources to work with data from multiple sources. Integrating historical data sources with new sources or combining data from multiple organizations can lead to deeper understanding in many areas.

Beyond specific tools and technologies, government organizations must be willing to fund the necessary infrastructure and resources to work with big data. Agencies need enough storage and bandwidth to host big data in a scalable way that provides practical access to the data. Agencies working with big data need infrastructure and resources that can handle enormous amoungt of data, that varies in accuracy, type, and volatility. Some organizations must capture, store, and analyze big data in real time. All agencies must determine whether to maintain their data in house or outsource to the cloud. Agencies that identify the cloud as the best resource need a streamlined process for accessing cloud resources.

4 CONCLUSIONS

The June 2015 Federal Big Data Summit reviewed many challenges facing the Federal Government's adoption of big data technologies and techniques. Agencies must overcome problems with engaging the culture and leadership within organizations, collaborating internally and externally, and developing complete standards that address data management, provenance, and security. Agencies must also adopt the right tools and resources and identify, hire, and train the right people to perform the work.

Although the summit recognized the various challenges that the government faces and the barriers to big data research, participants also highlighted recommendations and ways forward. Expanded collaboration within the government and sharing of best practices and roadmaps from established big data programs will assist other organizations to develop their own standards and environments for research and analysis. Developing standards across organization will help agencies handle their data and information in a secure and reliable manner, while maintaining the integrity of the data.

ACKNOWLEDGMENTS

The authors of this paper would like to thank The Advanced Technology Academic Research Center and The MITRE Corporation for their support and organization of the summit. Special

thanks are in order to Justin Brunelle, Patrick Benito, and Margaret MacDonald 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.

REFERENCES

[1] Middleton P., Kjeldsen, P., & Tully, J. (2013). *Forecast: The Internet of Things, Worldwide, 2013*.Retrieved July 15, 2015, from Gartner: https://www.gartner.com/doc/2625419/forecast-internet-things-worldwide-.

[2] President's Council of Advisers on Science and Technology, (2014, May). *Big Data and Privacy: A Technological Perspective* [Report to the President]. Retrieved July 15, 2015, from the White House Website: https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.

[3] Podesta, J., Pritzker P., Moniz, E., Holdren, J., & Zientz J. (2014, May). *Big Data: Seizing Opportunities, Preserving Values* [Report to the President]. Retrieved July 15, 2015, from the White House Website: https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf.

[4] The Federal Trade Commission. (2015, January). *Internet of Things: Privacy and Security in a Connected World* [Staff Report]. Retrieved July 16, 2015, from The Federal Trade Commission: https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled_internet-things-privacy/150127iotrpt.pdf.

[5] Barnum, S. (2014, February 20). *Standardizing Cyber Threat Intelligence Information with the Structured Threat Information eXpression (STIX™)* [White Paper]. Retrieved July 20, 2015, from The MITRE Corporation: https://stix.mitre.org/about/documents/STIX_Whitepaper_v1.1.pdf.

[6] Connolly, J., Davidson, M., & Schmidt, C. (2014, May 2). *The Trusted Automated eXchange of Indicator Information (TAXII™)* [White Paper]. Retrieved July 20, 2015, from The MITRE Corporation: https://taxii.mitre.org/about/documents/Introduction_to_TAXII_White_Paper_May_2014.pdf.