FEDERAL SUMMITS

JUNE 2016
FEDERAL BIG DATA SUMMIT REPORT*

August 5, 2016

Christine Harvey, Matt Mickelson, Huang Tang,
Bob Natale, Julie McEwen, Dr. Haleh Vafaie,
The MITRE Corporation†

Tim Harvey and Tom Suder
The Advanced Technology Academic Research Center

August 5, 2016

†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 30, 2016, 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 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: the Intersection of Big Data and the Internet of Things (IoT); Driving Innovation with Big Data; Progress toward Prescriptive Analytics; Data Privacy: Challenges and Solutions; and Using Big Data and Analytics in 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:

- The government is taking on more and more data and new agencies are using data to drive their mission. This increasing dependency means agencies need to be fully prepared to take on data and should be planning for a data-driven future. Agencies need to have big data architectures, development sandboxes, and storage strategies in place.

- Big data is a rapidly expanding field in need of standards and regulations that can keep up with the technology. Standards are necessary for inter-agency communication and to establish trust and confidence in data science.

- Federal agencies also need to be prepared and have standards in place related to data privacy. Agencies are trusted with sensitive data, including health data. Individual identities need to be protected and information should be securely encrypted.

1 INTRODUCTION

During the most recent Federal Big Data Summit, held on June 30, 2016, four 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 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:

- Intersection of Big Data and the Internet of Things (IoT)
- Driving Innovation with Big Data
- Progress Toward Prescriptive Analytics
- Data Privacy: Challenges and Solutions
- Using Big Data and Analytics in Health Care

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

2.1 Intersection of Big Data and IoT

The Intersection of Big Data and IoT session discussed the unique challenges, benefits, and the current state of regulations, provenance, and governance.

The session included discussions of the following:

- What is the Internet of Things (IoT)?
- How is the IoT linked to big data? Does the big data ecosystem need to change?
- Where should data generated by an IoT device be processed?
- Can we trust data generated by the IoT? If so, can we standardize "trust marks"?
- How much of a liability is collecting more data?
- What is the government's role as the IoT evolves? What is Industry's role?

2.1.1 Challenges

- Different definitions of IoT make collaboration between technical engineers and policy-makers difficult.
- IoT devices will further accelerate the accumulation of data that has already proven to be a difficult problem to manage.
- Cybersecurity protections for IoT devices exist mostly in the proprietary firmware provided by the manufacturer further security is lacking.
- IoT device manufacturers are not incentivized economically to include cybersecurity or privacy requirements into mass-market products.
- Data generated by IoT devices increasingly contains sensitive, personal, and behavioral information on individuals who may not have control of the data.
- IoT data can be tagged for access control, but how those tags are implemented into business rules for access typically remains a local project-level decision.

- Implementing “trust marks” on IoT data provides more transparency into the pedigree and provenance of the data, but whether to trust or not remains a personal decision.
- The government is facing a shortage of workers who understand current technology and how it will impact the country.

2.1.2 Discussion Summary

The session began with a definition of the IoT, and a brief discussion to further refine the concept of IoT. Per the discussion, the IoT was defined as a system of devices that bridge the cyber-physical gap. Specifically, IoT devices have three key characteristics: a sensor, a communication link, and an actuator. The sensor collects information. The communication link provides access to the device remotely, and allows the device to share information over any connected links. The actuator takes some physical action based on certain thresholds. Removing any of these three characteristics from a device still leaves a capable device. However, all three characteristics were important for a “smart” device.

Next, the discussion turned to the link between the IoT and big data. Although the IoT and big data are very different, they influence each other considerably. The proliferation of IoT devices, each of which produce a constant stream of machine-generated information, will magnify our existing challenge of making sense of the already enormous volume of data we have. Big data ecosystems have already begun addressing the challenges associated with larger volumes of data, the greater velocity of data production and consumption, and the increasing variety of data. The larger concerns highlighted by the IoT are about privacy, security, and trust.

Privacy best practices dictate encryption of IoT device data immediately. Security best practices dictate inspection and potential filtering of all traffic. Analytics best practices dictate performing computation as close to the data as possible. However, much of the value proposition of the IoT is in the aggregation of data and information sharing between systems. Just as the IoT bridges the cyber and the physical world, IoT data have an unprecedented ability to link digital information to physical behavior patterns. Access must be limited, and any individual should have control over data about themselves

Tagging data with “marks” indicating its trustworthiness, attributes for access control, provenance, and pedigree all will become more important. All these will require public standards to ensure consistency and interoperability across such a wide market. However, the implementation of each of these is a private decision. Local individuals and programs determine how to implement access control tags into their local applications. Individuals

determine how much to trust other entities, but do so differently according to localized context.

The government's role in all this is (1) to protect the people's best interests and (2) to ensure that free market economics are not the only forces driving IoT. The government should create standards for IoT commerce, IoT device security requirements, and the usage of a person's behavior data. The government's buying power is staggering, but still is overwhelmed by aggregated consumer demand. The government is uniquely positioned to both drive IoT requirements through its own acquisitions, and aggregate consumer demand through public-private partnerships and education campaigns. The open question is who should lead, and whether new skill sets and organizations are required?

2.1.3 Important Findings

- The IoT will make the data challenges we have today even worse, and today's big data ecosystems will need to adapt.
- Vulnerabilities in IoT devices will be difficult to patch. Standards for embedded system security, long-term operation, and ongoing maintenance updates need to be established and enforced.
- Standards for establishing trust between individual consumers and IoT providers need to be established.
- Standards for implementing access controls to IoT-generated data need to be established.
- Existing federal, state, and local policies regarding data ownership and usage of behavior data need to be updated.

2.2 Driving Innovation with Big Data

The Driving Innovation with Big Data session discussed the best practices and methods for beginning to use big data to drive innovation.

The session included discussions of the following:

- What is big data?
- What is innovation?
- What are the obstacles for driving forward big data innovations at federal agencies?

2.2.1 Challenges

- Data sources may be managed by and distributed among various government agencies, even private institutes. In many applications, to fully utilize big data analytics to achieve deep insights, one needs to rely on data fusion. This distributive ownership of relevant data can be a big challenge.
- For individual agencies, the ever increasing data sources and varieties pose challenges to the data governance policies and standard procedures.

2.2.2 Discussion Summary

Participants in this session focused the initial discussions on defining big data, innovation, and the meaning of combining the two. The session leads and participants covered the important aspects of big data and how the topic is defined across the government and industry. Big data tools that are useful for driving innovation were also covered including Hadoop and MapReduce. Session participants also took time to describe the meaning and concepts behind innovation. Participants in the session described possibilities and limitations in their organizations when it comes to innovation. Success stories of innovation were also described in instances unrelated to big data.

The conversation among participant let to the topic of the disruptive force of big data. Even though the workflow of business process remains intact, the disruption of big data technologies will fundamentally change the extents, depth, speed, and sophistication of the work. A systematic high level approach in data model and data governance policies are necessary to encourage cross-sector, cross-agency big data collaborations, which can be the driving force and foundation for big data innovations in many areas.

The final aspect addressed by participants were the reasons for data safety. Often, data sources are hosted by their owning agency and are not allowed to be transferred off-premise. This is counter intuitive to the innovation drive to fuse various data sources together and obtain a more complete picture on demand. An innovative technology and architecture design is in need to overcome such cross-domain barriers.

2.2.3 Important Findings

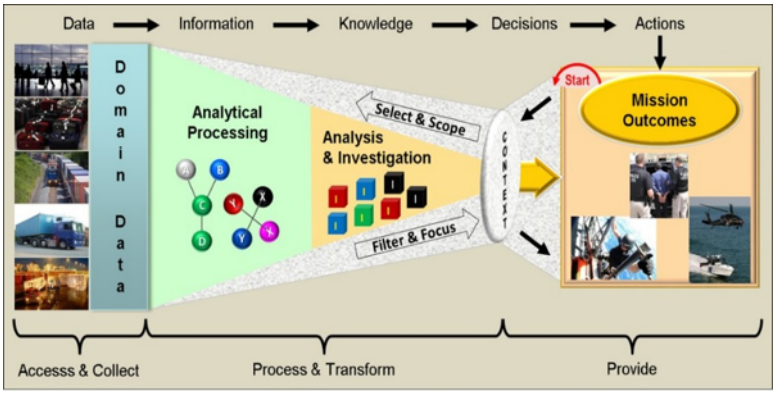
- Due to the increasing quantity of data, the issue of knowledge management is becoming a more prominent issue.

- Data that are available and need to be combined are often from various agencies, this calls for a unified data model and an overarching data governance policy within agencies so data quality and security can be assured. A universal data exchange standard needs to be implemented at the federal level.
- One of the biggest innovations needed in order to facilitate big data research is the ability to directly tap into data sources that are distributed across various public and private domains in real time.
- An all-inclusive data model must be crafted at the very beginning in order to insure compliance, coordination, and sufficient support.

2.3 Progress toward Prescriptive Analytics

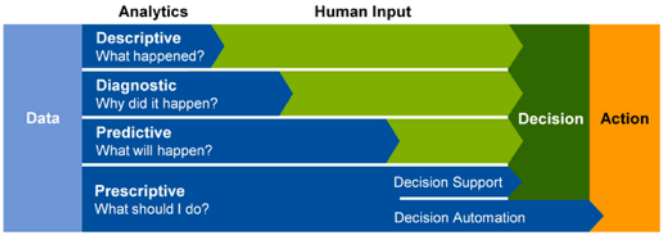
Prescriptive analytics builds off the findings of descriptive analytics and the projections of predictive analytics to deliver the value promised by the data-to-decisions and evidence-based decision-making visions. Prescriptive analytics produces recommendations for courses of action tuned to facts and science to assist humans in making sounder decisions in both routine and novel situations. In concert with implementation of governed automation, prescriptive analytics can even finalize selection and initiate execution of situationally and contextually optimal decisions. This session examined the current state of prescriptive analytics in both research and operational environments, what obstacles to maturation persist, and how progress might be advanced via algorithms, standards, tools, pilots, and other means.

Figure 1: Data analytics operational flow [1]



Per the notional Data Analytics Operational Flow diagram, the ultimate purpose of prescriptive analytics in the mission lifecycle is to facilitate positive mission outcomes under the

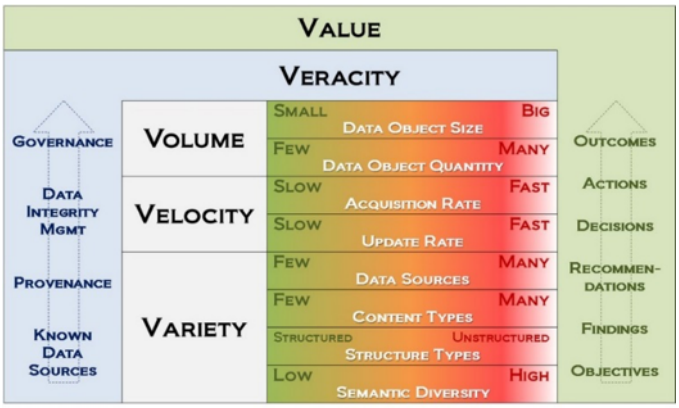
Figure 2: From data to action. [1]



given operating constraints (e.g., resources, time, uncertainty, etc.) by promoting optimal decisions and actions. Successful mission outcomes require optimal decisions realized via effective actions, enhanced via efficient execution via governed automation where feasible and appropriate.

Prescriptive analytics complete the value proposition for advanced analytics relative to mission outcomes. Major applications typically involve many diverse and interacting patterns across the volume, velocity, and variety dimensions âĂŞ very few applications are one-dimensional in this respect. Successful investments in big data analytics focus on the value factor, assured via the veracity factor as applied to the volume, velocity, & variety factors germane to the target application.

Figure 3: Value-driven data analytics [1]



Prescriptive analytics requires a predictive model with two additional components: actionable data and a feedback system that tracks the outcome produced by the action taken. Prescriptive analytics is characterized by techniques such as graph analysis, simulation, complex event processing, neural networks, recommendation engines, heuristics, and machine learning.

The session included discussion of the following questions

- What is Prescriptive analytics? How does it relate to descriptive and predictive analytics?
- What contributions can/should Prescriptive Analytics make in respective stages of the mission lifecycle (i.e., planning, provisioning, execution, others)?
- What kinds of data sources are particularly useful for prescriptive analytics (e.g., mission objectives, priorities, lessons learned, POA&M performance assessments, after-action reports, etc.)?
- Is risk-scoring a prototypical use case for predictive analytics for applications supporting complex life-critical missions?
- What kinds of automated actions via prescriptive analytics can be supported in live operations, and how?
- Is "big data" necessary for prescriptive analytics?
- Beyond statisticians and data scientistscuratorsmanagers, what kinds of human resources are needed for prescriptive analytics?

2.3.1 Challenges

The discussion session identified the following major challenges confronting progress toward wider and more effective use of prescriptive analytics:

- Agencies need to identify the concrete factors or elements required for prescriptive analytics to transform the findings of descriptive analytics and the associated projections of prescriptive analytics into usable course of action recommendations for human decision-makers and enable trustable automated actions for human operators.
- Selecting the level of use case specificity is necessary to provide the optimal bounding of the foregoing challenge to a manageable level without unnecessarily limiting the scope of value of the resulting solution.
- The government needs to establish the requisite degree of confidence, control, and trust such that the transition from decision recommendations to automated decision execution via prescriptive analytics can be accepted.
- There is difficulty in instrumenting prescriptive analytics solutions for "dial-ability" in the range of options from decision recommendations to automated decision execution.

- Agencies need to establish the requisite level of context-awareness and situational awareness to accurately "set the dial" for prescriptive analytics in any given operating environment.
- There are a range of barriers âĂŞ organizational, policy, trust of machines, magnitude of potential impact, etc. – to acceptance of prescriptive analytics, particularly in the automated decision execution mode.
- There needs to be a definite way to collect and ensure access to the range of additional data sources needed for prescriptive analytics.

2.3.2 Discussion Summary

The discussion participants recognized that systematically capturing and efficiently structuring the kinds of specialized data needed to establish an information foundation for prescriptive analytics – e.g., mission objectives, priorities, lessons learned, POA&M performance assessments, after-action reports – have received significantly less attention than has been the case for other purposes. Resolving this critical gap will require changes in policy and practice, possibly including new approaches to data preparation and curation, information architecture, and data governance, among other things.

Five categories of information technology big data sources were identified as being generally useful for prescriptive analytics applications:

- Web and social media data
- Machine-to-machine (M2M) data
- "Big Transactions" (very large transactions and/or very large quantities of related transactions)
- Biometric data
- "Human-generated" data (e.g., opinions, behaviors, movements, etc.)

The group consensus was that appropriate use case formulation was critical to the success of efforts to advance prescriptive analytics. Likely domains for productive use cases in this respect were identified, including health care, medical devices, cybersecurity, risk management, threat reduction, benefit assignment, business investment, logistics and resource planning (e.g., FEMA), transportation, and staffing (e.g., placement of firefighting crews in large-scale

forest fires). Categorization within candidate domains was also considered to be an effective way of narrowing in on more actionable use cases.

The challenges facing the "graduation" from having prescriptive analytics simply make recommendations about courses of action for human decision-makers to evaluate and select to actually automating actions to execute analytically established decisions received much attention. The group recognized a difference between automation and autonomics, with the latter requiring either a much reduced scope of action, a much high level of verifiable trust, or both. The premise that all such capabilities must be both governed and "dial-able" (a flexible form of applying such governance was universally acknowledged). Mechanisms such as dense closed-loop feedback paths, possibly continuously improved via machine learning technologies were noted as key enablers. Likewise, empirical and adjustable methods of establishing levels of confidence, via metrics and measurements, were identified as essential. Lastly, the group discussed the possible role of recommendation and reputation systems, and the technologies that underlie them, in progressing prescriptive analytics as those systems and technologies mature via advanced research and development and via marketplace experience.

Recognition of the importance of context and situational awareness (see Figure 1 above in this section) for credible and effective outcomes from Prescriptive Analytics led to a discussion of the range and nature of the data sources that would go into any such analytics. Adequate coverage of the decision space via relevant data sources and measures (yet to be identified) to guard against bias due to skewed or incomplete data were noted as particular concerns. Designation and sustainment of authoritative sources, provenance controls, and data quality and integrity assurances were also discussed. Advanced analytical tools and models, possibly enabled in part by semantic technologies, were seen as necessary to augment the capabilities of human data analysts and data scientists (which were seen to be in dramatically short supply at this time in most government agencies, relative to the need).

Several substantial barriers to progress in prescriptive analytics were discussed, including:

- Organizational challenges
- Acceptance of policy changes
- Trust of "machines"
- The relative paucity of skilled and experienced personnel
- Inadequate tools and technology

- The magnitude of impact of some decision areas in government missions
- Reliable risk-scoring of decision alternatives
- The criticality and number of potential outcomes in complicated scenarios
- Cognitive learning and user interface issues

Time did not permit detailed analysis of options for overcoming these barriers, beyond what is outlined in other parts of this section.

2.3.3 Important Findings

In addition to the insights outlined in the preceding parts of this section, the group discussion uncovered the following important findings:

- The overarching finding of the group discussion was that Prescriptive Analytics for significant government applications generally is at an embryonic stage of development.
- Prescriptive analytics lacks generally accepted standards and reliably productive tools relative to what is available for descriptive and predictive analytics.
- On the whole, most people and most organizations in the government are at a very early stage on the learning curve concerning prescriptive analytics. At the same time, it might be good to ensure that their knowledge about, understanding of, ability to use, and experience with both descriptive and predictive analytics matures substantially as a prelude to attempting prescriptive analytics. Not doing so could lead to early failures, demoralization, and undermining of confidence in the promise of prescriptive analytics.
- Government decision-makers are likely to accept recommendations from Prescriptive Analytics tools given demonstrated assurances about data adequacy and model and algorithm accuracy. Neither academia nor industry is there yet.
- Despite the relative lack of maturity of prescriptive analytics at this time, the proven and emerging benefits derived from descriptive and predictive analytics paired with the promised benefits to be derived from prescriptive analytics and with the evident trajectories of many of the key technologies needed for prescriptive analytics largely ensures that this discipline will progress, steadily and ever more rapidly in the years to come.

2.4 Data Privacy: Challenges and Solutions

The Data Privacy: Challenges and Solutions session focused on understanding and reviewing the current challenges and potential solutions facing the realm of privacy in big data.

The session included discussions of the following:

- What are the best ways to handle data privacy issues while still maintaining collaboration and data sharing whenever possible?
- How do agencies across the government handle data privacy issues and concerns and what needs to be done in order to improve?

2.4.1 Challenges

- Federal agencies collect large amounts of information about individuals, and there is a shortage of resources available to assist with privacy.
- It is difficult to stay ahead of trends regarding technology use and types of attacks/privacy incidents associated with particular technology.
- Assessing privacy risk can be difficult.

2.4.2 Discussion Summary

Collaboration session participants provided examples of many ways that their organizations work to address data privacy in the use of big data. These activities fell primarily into three main categories:

- Laws and Policy
- Leadership and Culture
- Operations and Tools

Participants discussed how federal agencies have many laws, policies, procedures, regulations, guidance, and standards related to privacy that they must follow. They mentioned that part of the process of using those documents includes having legal review of their privacy activities to ensure that they are compliant, and having independent verification and validation of their activities and formal audits performed. In addition, privacy professionals must process a number of types of agreements, such as data use agreements, as part of

protecting privacy while sharing data. Based on the descriptions provided by the session attendees, ensuring compliance is a large part of what the privacy professionals in the federal government do.

Participants emphasized the importance of instilling a culture of privacy within their organizations in order to be successful. This includes providing appropriate privacy training to new individuals as well as refresher privacy training to those already on board. The group discussed the different perspectives on privacy that are evident in the workplace, including the idea that some individuals may not value privacy as much as others and that there is a perception that this is a generational difference. Participants generally agreed that support for privacy efforts beginning at the very top of an organization is a key to the success of privacy programs. They discussed how increasing engagement with privacy stakeholders working outside the privacy office can help to increase privacy awareness and better integrate privacy activities throughout an organization. Some of the privacy stakeholder organizations that participants mentioned with whom they engage are institutional review boards and data governance boards. One participant noted that her organization has a checklist that they use to ensure that they engage with all of the appropriate privacy stakeholder organizations whenever a new project begins. The participants agreed that engaging with privacy stakeholder groups will also increase privacy visibility to Senior management, resulting in greater leadership attention to privacy and hopefully increased resources for privacy activities.

In terms of operations, participants discussed the importance of attention to risk management for privacy efforts, and that broader privacy harms should be identified as part of assessing risk. Participants mentioned the connection between privacy and security, and how it is important to have appropriate security protection in place in order to protect privacy. In particular, they listed access management and control and encryption as two important security mechanisms to use. Session attendees also discussed the increased use of privacy-enhancing technologies, such as data loss prevention tools, de-identification, and anonymization for protecting privacy. The concept of Privacy by Design was discussed âĂŞ this is the idea that privacy should be fully integrated into systems from the very beginning. Privacy engineering is the actual implementation of Privacy by Design by including privacy activities in every phase of the system engineering process. The group discussed how privacy engineering is a somewhat new concept to some of them, and also noted that the National Institute of Standards and Technology is leading work in the area of privacy engineering to be used throughout the federal government and within private industry.

2.4.3 Important Findings

Effective ways to handle data privacy issues while still maintaining collaboration and data sharing whenever possible include:

- Use tools, such as encryption, and privacy-enhancing technologies including data loss prevention tools, anonymization, and de-identification.
- Leverage existing laws and policies and identify best practices across different agencies, including by working as a member of the Federal Privacy Council. Develop a federal privacy dashboard to show effective practices in action.
- Provide privacy training, especially at an early age so that people learn about privacy long before they enter the workforce. Also provide privacy training for different roles within organizations.

Key activities that agencies should adopt that will improve privacy related to the use of big data and data sharing include:

- Fully implement the National Institute of Standards and Technology's Risk Management Framework for privacy and leverage what cybersecurity has done in terms of building security in from the beginning to do the same for privacy âĂŞ this is implementing privacy engineering.
- Get Senior Leadership buy-in for privacy.
- Improve data auditing tools.

2.5 Using Big Data and Analytics in Health Care

The Using Big Data and Analytics in Health Care session facilitated discussion on big data and analytics' impact on health care. Other industries have benefited from the scale and flexibility of "big data", however health care is just starting to gain traction. Despite the potential benefits of mining data to identify epidemics, cure disease, improve quality of life and avoid preventable deaths, many challenges specific to health care hinder rapid adoption. The focus of this session was on how big data pertains to government run health care organizations such as the U.S. Department of Veterans Affairs, and the Defense Health Agency. The participants in this session hoped to identify:

- What are the unique challenges faced by big data implementers in health care environments?

- How do these challenges impede adoption of big data strategies?
- What are the best clinical challenges to start with when first adopting big data technologies?
- What is the best approach to protect patient privacy while uniquely matching patients and providers?

2.5.1 Challenges

These discussions identified the following challenges:

- Health care data is limited in scope, medical data is a small part of a person's personal information. Electronic Health Records (EHR) only contain a person's clinical encounters.
- There are no standards for capturing and storing health data, different providers compile different types and formats of data records. There is a lack of common data elements and variations in the interpretation of the available data.
- Adequate metadata is absent from health records, agencies often receive data from external providers and additional information is often needed to clear inconsistencies in definitions.
- Two of the main barriers to adopting big data technologies is the quality of the data and the absence of real-time data delivery. Accurate data should be available to clinicians in real-time.
- Patient privacy is at risk, tokenization needs to be used along with secure data encryption and processing standards to protect patients.

2.5.2 Discussion Summary

The Using Big Data and Analytics in Health Care session focused on the challenges of using big data and analytics in health care environments. Although electronic medical records have helped to streamline patient data, they are still very limited in scope. Individual medical data only captures information about a patient's clinical encounters, a cell phone or fitness tracking device may contain more information on a person than their physician. The ultimate goal is to provide personalized health care, where the health of an individual is continuously monitored and analyzed. Government agencies are focused on this broader vision. Despite

the wealth of data contained in electronic medical records, the clinician's notes are subjective and there is a lack of common data elements across health care providers and organizations.

Session participants discussed how challenges in health care impede the adoption of big data strategies in health care. One of the major difficulties is the lack of standards for capturing and storing the data. Health care providers with use varying methods and there is a definite lack of common data elements. The common data elements are necessary to link data sets from multiple sources and improve data quality and sharing. Better, standardized metadata is necessary to clear inconsistencies and provide transparency. Another challenge in managing health care data is the backlog of historical paper records. Current tools used in big data analysis for health care are not optimal to handle unclean, noisy data.

Following the discussion of challenges hindering the adoption of big data in health care, participants described the best clinical challenges to approach when first adopting big data technologies. Participants noted that the main barriers to the adoption of big data technologies are the data quality and the speed to market. In order to address this problem, it was suggested that when first adopting big data technologies, start with a sandbox where the implemented system can be tested before transitioning to the production system. Organizations then need to ensure the implementations are FISMA compliant and information is shared through standards such as Health Level Seven International (HL7). Sharing information from multiple data sets using standards lowers the cost and improves the efficiency, quality, and patient safety in health care.

When it comes to working with big data, researchers need to develop prototypes with a specific purpose, and not simply investigate interesting areas or topics as a science project. Participants also noted that big data tools should be tailored to unique health care requirements.

Finally, session participants reviewed the best approaches to protecting patient privacy. Consensus was that tokenization should be used, which allows for matching patients without direct use of patient identification numbers. In tokenization, a patient's identification is anonymized and the information can be matched across datasets. Data provisioning and encrypting patient information were also discussed as other approaches for protecting patient privacy. Personal Identifying Information (PII) and Protected Health Information (PHI) should be protected through encryption.

2.5.3 Important Findings

- Common data elements are necessary for research and big data integration across providers and government agencies.

- Researchers need to develop prototypes with a specific purpose.
- Big data adopters should begin research by creating and using a FISMA-compliant sandbox as a development environment, that can easily transition to production.
- Tokenization is needed to protect patient's identities in electronic health records and data should be protected by securing PII and PHI through encryption.

3 SUMMIT RECOMMENDATIONS

The challenge areas covered at the Federal Big Data Summit covered multiple topics and various areas of concern within the government and the approach to handling big data in the future. Although the major discussion topics focused on different aspects, several common themes recurred across many of the challenge areas. Participants noted three topics as being of major importance in multiple areas: organization need to be prepared for the growth and expansion of big data, standards are crucial to managing this growth and collaborating between organizations, and individual privacy needs to be protected including device users and patients.

Every collaboration session at the summit discussed the need to prepare to for data in the future. The Intersection of Big Data and the IoT session discussed the rapid growth of device-generated information. In recent years the amount of information and usable data produced by smart devices has expanded along with the desire to collect this information. Government agencies wanting to make use of IoT information need to plan ahead and build data models suitable for the velocity and volume of this data. The Driving Innovation with Big Data session also recognized the need to build data models that are account for all desired information and data available. The Using Big Data and Analytics in Health Care session focused on the need for a FISMA-compliant sandbox for organizations to perform research and testing in a development environment before moving on to production. Complete preparation is key to performing quick, innovative research with minimal setbacks.

The second common theme covered in all sessions is the need for standards, which covers all aspects of working with big data. Standards are needed in prescriptive analytics to ensure quality work is performed, these standards need to match or exceed what is currently available for descriptive and predictive analytics. The innovation and health care sessions both noted the importance of standards in order to perform exchanges of information between various government agencies. Common data elements and standardized metadata formats will allow agencies and organization to share information, leading to greater shared knowledge. The

data privacy collaboration session reviewed the need for adopting current standards from the National Institute for Standards and Technology Risk Management Framework for privacy. Standards are especially important with the IoT, the need for device security, access controls, and agreements between consumers and providers.

The need for standards is critical to ensure there is trust and buy in from senior leadership in all aspects of big data research. Without standards and a common level of understanding of what is possible, it is extremely difficult to have confidence in big data and the outcomes of the work performed.

The mention of standards often included a discussion on the need for reliable and suitable tools for managing and analyzing data. As a new field, prescriptive analytics needs tools that are productive and dependable to establish trust and provide reliable guidance. In the field of privacy, data auditing tools are necessary to manage information flow and security. Health care needs tools that are result and purpose driven to handle patient information and big data analysis.

Multiple sessions also discussed the need for privacy. This was especially important in the privacy-focused session as well as the discussions on health care and the IoT. The data privacy session examined the need for encryption, anonymization, and de-identification of data. This process needs to be done in order to protect individuals as well as organizations. Protecting patient information is highly important in health care, tokenization of patient identifiers along with removing protected health information and personally identifiable information is necessary to ensure patient privacy. Finally, privacy was discussed in the IoT session where participants debated the rights and needs of patients and manufacturers of smart devices. The rules of data ownership and privacy vary across companies and organization and can have a major impact on the privacy of users. Privacy training is also important beginning at an early age as well as specified training within the organization drawing focus to important privacy needs.

4 CONCLUSIONS

The June 2016 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, planning and preparing for big data, instituting and managing standards for big data, and maintaining and securing data privacy were recognized as common challenge areas.

While the June 2016 Federal Big Data Summit highlighted areas of continued challenges and barriers to progress, the Summit also cited notable advances in mitigating these perennial challenges. Progress has been made when it comes to using the technologies and building a level of trust and familiarity with big data. Now, agencies are beginning to inquire as to how to improve their work and functionality using standards and proper data preparation and planning. With the increasing comfort level with big data and acceptance of data science as a reliable field, agencies are looking for ways to standardize and solidify their work.

REFERENCES

[1] Bob Natale, MITRE, 2013.

[2] Gartner Report September 2015, Extend Your Portfolio of Analytics Capabilities https://www.gartner.com/doc/3054119/predictive-analytics-transforming-bb-selling.

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.