FEDERAL DATA & ANALYTICS SUMMIT

OCTOBER 23, 2018 | MARRIOTT METRO CENTER | WASHINGTON, D.C.

On behalf of the Advanced Technology Academic Research Center, I am proud to announce the release of a White Paper documenting the MITRE-ATARC Data & Analytics Collaboration Symposium held on October 23, 2018 in Washington, D.C. in conjunction with the ATARC Federal Data & Analytics Summit.

I would like to take this opportunity to recognize the following session leads for their contributions:

MITRE Chairs: Mark Wahnish and Glenda Hayes

Session No. 1: Modernization of Enterprise Data Governance
Industry Chair: Nii-Lante Lamptey, Senior Manager, Deloitte
MITRE Chair: Dr. Scott Renner, Principal Engineer, MITRE

Session No. 2: Maximizing Data Access
Industry Chair: Chris Roberts, Solution Architect, Quest Software
MITRE Chair: Dr. Robert Daniels, Principal Engineer, MITRE

Session No. 3: Shaping Data Strategies to Better Support Decision Making and Accountability
Government Chair: Luwanda Jones, Executive Director, U.S. Department of Veterans Affairs
Industry Chair: Jonathan Flynn, Specialist Solutions Consultant, Hitachi Vantara Federal
MITRE Chair: Dr. Justin Brunelle, Lead Researcher, MITRE

Session No. 4: Maximizing Commercialization, Innovation, and Public Use of Government Data
Industry Chair: Michael Wood, Director, Emerging Technology Practice, CGI
MITRE Chair: Lisa Glikbarg, MITRE Center For Program and Technology Partnerships Lead, MITRE

Session No. 5: Data & Analytics in Health
Government Chair: Yvonne Cole, Enterprise Solutions Specialist, DoD/VA Interagency Program Office
Industry Chair: Rashmi Mathur, Partner, IBM
MITRE Chair: Dr. Cj Rieser University of Virginia Site Partnership Leader, MITRE

Session No. 6: Nature-Inspired Machine Intelligence (NIMI): Harnessing Evolution
Vice Chair: Josh Simkol, Technologist
MITRE Chair: Dr. Ronald Campbell, MITRE

Below is a list of government, academic and industry members who participated in these dialogue sessions:

Challenge Area 1: Modernization of Enterprise Data Governance: Scott Renner, MITRE; Anthony Burley, DOJ; Kyle Carrick, GSA; Mike Ross, Deloitte; Erin Bohannon, Deloitte; Bridget Hilal, SEC; Robert Fahs, ISOO; Sue Bussells, USDA

Challenge Area 2: Maximizing Data Access: Robert Daniell, MITRE; Bryan Roettger, VA; Tina Chang; USDA; Harold Lorton, Army National Guard; Kyle Bausch, Quest; Stella Kwon, Quest; Susan Wong, Quest; Dawn Haney, DHS; Benjamin Fieselmann, VA

Challenge Area 3: Shaping Data Strategies to Better Support Decision Making and Accountability: Kelley Hussey, Deloitte; Jonathan Flynn, Hitachi Bantara Federal; Jacqueline Chen, Deloitte; Carol Bean, VA/DoD IPO; Bryan Van Winkle, Hitachi Vantara Federal; Mike Kirtland, Army National Guard; Ceresh Perry, USACE; Deirdre Coley, DOJ; Anthony Joyce, Navy; Ciro Lopez; Navy; Germaine Dunmore, DOJ; Bernice Lemaire, PBGC; Siobhan Chambers, DHS

Challenge Area 4: Maximizing Commercialization, Innovation, and Public Use of Government Data John Broderick, US BLM; Lily Ho, Clarke Birrell, GSA; Mike Wood, CGI; Ved Malik, USDA; Michael Levinson, IPO

Challenge Area 5: Data & Analytics in Health: Marguerite Pridgen, DC Commission on Aging; Jonny Behrens, IDA STPI; Chikezie Maduka, UMD-CDC; Laura Coombs, American College of Radiology; Kerri Zou, VA; Aubrey Hamilton, MITRE; Paul Donohoe, CMS; Asir sheikh, IPO; Jimmy Chen, VA; John Scott, DHA; John Griffith, MITRE; Cj Rieser, MITRE; Rashmi Mathur, IBM; Henry Ogoe, IPO; Jonny Behrens, IDA STPI

Challenge Area 6: Nature-Inspired Machine Intelligence (NIMI): Harnessing Evolution: Jeffrey Bruck, US Courts; Sally TInkle, IDA/STPI; Heideh Shadmand, IPO; Russ Vane, DHS; Kate Borden, ABS Group; Jack Wang, GAO; Chris Treml, ACR; Josh Simkol; George Crombie, DOJ; Derek Strephenson, Straight Talk; Jim Chen, DOD; Kirk Samuel, NASA

Thank you to everyone who contributed to the MITRE-ATARC Big Data Collaboration Symposium. Without your knowledge and insight, this White Paper would not be possible.

Sincerely,

Tom Suder
President,
Advanced Technology Academic Research Center (ATARC)

FEDERAL IT SUMMIT SERIES

OCTOBER 2018 FEDERAL BIG DATA AND ANALYTICS SUMMIT REPORT*

March 27, 2019

Mark Wahnish, Ronald Campbell, Robert Daniels, Lisa Glikbarg,
Glenda Hayes, Scott Renner, Cj Rieser, Justin F. Brunelle
The MITRE Corporation

Tom Suder
The Advanced Technology Academic Research Center

Contents

1 ABSTRACT

The most recent installment of the Federal Big Data Summit, held on October 23, 2018, included six 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 the following six challenge areas in big data:

1. Modernization of Enterprise Data Governance
2. Maximizing Data Access
3. Shaping Data Strategies to Better Support Decision Making and Accountability
4. Maximizing Commercialization, Innovation, and Public Use of Government Data
5. Data & Analytics in Health
6. Nature-Inspired Machine Intelligence (NIMI): Harnessing Evolution

This white paper summarizes the discussions in the collaboration sessions and presents recommendations for government, academia, and industry while identifying commonality between challenge areas.

As an outcome of these collaboration sessions, summit participants indicated high interest in the potential costs savings, ease of use, and better information sharing that can result from leveraging big data. The imperative for public/private cloud technology adoption to meet big data processing requirements was identified as a goal by the participants in several collaboration sessions. However, security issues, difficulties with developing foundational strategies, and others will require more effort before full adoption and exploitation of these capabilities can be established. These areas all require further research and exploration.

The collaboration sessions identified detailed, actionable recommendations for the government, academia, and industry which are summarized below.

1. Organizations adopting big data should adopt a data strategy. For many organizations, beginning to leverage big data is not a matter of modernizing their data strategy, but developing a data strategy for the first time.

2. Several sessions recommended implementing methods for maintaining consistent, well-engineered, and accessible metadata; facilitating internal and external sharing; improvement of analytical quality through big data principles; and better access control.

3. Big data users must establish understandable, consistent access restrictions and agreements to facilitate data sharing.

4. Organizations dealing in big data must create dedicated roles within their organizations for managing data, technology, and governance, such as Chief Data Officers (CDOs) and data stewards.

5. Organizations should work to remove and streamline the technological and policy barriers preventing private/public partnerships and data sharing agreements.

2 INTRODUCTION

During the most recent Federal Big Data Summit, held on October 23, 2018, six 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 may 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) [6]. ATARC is a non-profit organization that leverages academia to bridge between government and corporate participation in technology[1].

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 workforce 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.

3 COLLABORATION SESSION OVERVIEW

Each of the six 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 the following topics:

1. Modernization of Enterprise Data Governance
2. Maximizing Data Access
3. Shaping Data Strategies to Better Support Decision Making and Accountability

[1] http://www.atarc.org/

4. Maximizing Commercialization, Innovation, and Public Use of Government Data
5. Data & Analytics in Health
6. Nature-Inspired Machine Intelligence (NIMI): Harnessing Evolution

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

3.1 Modernization of Enterprise Data Governance

Governance is a vital component of any data strategy. Data policies, role definitions, and compliance provide a framework upon which successful data strategies are built. One rationally expects the arrival of big data – the huge increase in data volume, variety, and velocity – to present new or increased challenges to enterprise data governance. The purpose of this session was to discuss the data governance challenges that have arrived (and are expected) and the governance changes needed to allow organizations to cope with using big data. The participants identified several challenges, topics, and recommendations that the government faces when improving data governance.

3.1.1 Challenges

This discussion session identified a number of challenges.

- Existing data governance tends to be focused entirely on separate data silos. Governance that is effective across the entire enterprise is lacking or entirely missing.
- Effective change requires a business champion who has the time and energy required to lead change, and who will remain dedicated for the time needed to effect change. Such champions are often unavailable in government organizations.
- It is necessary to have a common group to make enterprise-wide decisions about data, but it is difficult to establish this group in a way that is seen to care for the interests of all the stakeholders.
- It is difficult to define the role of data owners and stewards in a way that creates incentives for doing the role well. Stakeholders may want to “own” data even though they are neither equipped nor motivated to properly curate and maintain it.
- A data governance maturity model seems desirable, but the participants have observed neither success in finding and implementing an acceptable model, nor in creating a new model suited to the enterprise.

- Improvements to data policy are not always carried down to implementation. Often, the high-level language is changed but the actual day-to-day regulations remain the same.
- Different data types have varying requirements which makes managing the data complex. This includes unexpected growth rates of the varying data.
- Participants report a tension within data access policy: how do practitioners increase data sharing while still preventing unauthorized access.

3.1.2 Discussion Summary

The discussion began with the properties of big data, leading to the question "What effect might these properties have on enterprise data governance?" The consensus of the participants – expressed verbatim by more than one person – was: What data governance? The participants said they saw so little in the way of enterprise data governance today that they could not see how big data might make a difference in the future. Big data might make some things harder in the future (and could make other things easier) but for the present, the participants did not expect enterprise data governance to be greatly affected one way or the other. The participants did discuss challenges they saw with enterprise data governance; the challenges were most often related to cultural and procedural challenges rather than being specific to big data.

3.1.3 Recommendations

The participants also proposed ideas and best practices for dealing with those challenges.

- Establish a Chief Data Officer (CDO) to create and implement an enterprise data strategy, with particular emphasis on educating the enterprise about data governance roles.
- Establish a data council to review policies, address data changes, and vet applications and standards.
- Promote data governance learning through storytelling to answer the question "What is data management and data governance?" Ensure these stories are tied to important business needs.
- Use the data strategy to change the culture around data sharing, development processes, and budget control mechanisms.

- Develop and apply a maturity model to measure progress and benefits through each stage of implementing the data strategy.
- Share best practices and lessons learned across agencies.

3.2 Maximizing Data Access

The *Maximizing Data Access* collaboration session focused on the challenges of making data available to the right people, when it is needed, in the format it is needed. The discussion revolved around how data practitioners can best leverage new technologies to maximize non-sensitive data sharing with the public, and increase access to sensitive data while maintaining privacy, security, confidentiality, and respecting the interests of data providers and other stakeholders. The session had the following three goals:

- Discuss what maximizing data access challenges meant for various organizations.
- Identify any best practices for managing, publishing, and sharing data.
- Recommending ways to reduce risk and promote accountability through an organization's governance and strategy.

3.2.1 Challenges

Several challenges emerged from the discussions within the group.

- Data generation is so voluminous that we now call it Big Data. But the quantity of information published is not – in itself – inherently provide useful or usable information. Without context to form correlations to other data sets, its very meaning becomes ambiguous, at best.
- Data is typically stored and published with insufficient context and metadata for the potential consumers to understand how it should (or should not) be used.
- As a data curator or steward of a public dataset there is a significant challenge to provide meaningful, relevant, and timely access to information.
- Sharing and interoperability between organizations raises additional challenges. Progress is often measured in yearly increments.
- Systems can only be changed when resources become available. Multiple governance boards may have to agree on the parameters of data sharing.

- Since public, commercial, and opposition consumers will compare data from multiple sources, if one agency protects certain values while another agency does not, the consumer may be able to perform their own aggregations.

3.2.2 Discussion Summary

The afternoon's discussions began with several participants expressing surprise that the topic was about maximizing public data access rather than intra-agency sharing. They perceived that the two issues were related but quite different. The participants agreed that we could discuss both challenges and apply lessons across the two goals.

To begin the conversation, the discussion leaders presented a review of the use case for GPS today and how it became a public location network. This scenario was used to illustrate how data might be released with less precision and/or with certain values redacted to satisfy a public good without compromising security. The following description was presented:

> The observations of the Sputnik satellite and the doppler effect on its signal provided a mechanism for determining its location. Scientists at JPL then reversed engineered the effect to determine a user's ground-based location. This was the first practical implementation of a global positioning system or GPS. DARPA would later create multiple iterations of the technology for military applications. This was a very protected and guarded network capability for the US military as it was used to guide subs, ships, missiles and manpower around the globe. However, that all changed when the Russians shot down KAL007 in 1983 after the flight drifted off course and into Russian airspace. Ronald Reagan issued an executive order opening the GPS network to commercial users. The military asked that the accuracy of the system for commercial use would be degraded. This provides the textbook case study of how our national data sources must constantly find a balance between public and restricted data access.

Examples of government agencies with mandated public access requirements were discussed for topics such as grant information, geospatial data, and images collected by NASA. The need for enhancement/enrichment of many of the raw data formats with semantic and context tags is a challenge. How does one explain how to use the data correctly? How do agencies respond to feedback that "the data is wrong" when the lack of metadata leads to mis-interpretations?

The GPS example was revisited to explain how the ubiquitous and simple notion of a timestamp might be deceptive given the infrastructure and topology involved in the data

generation and collection.

Some examples of popular public datasets that many take for granted were elicited.

- Census data[2]
- US labor statistics[3]
- GAO, OMB, CBO[4]
- State Department country reports[5]
- Postal codes[6]
- NOAA/NWS[7]
- FBI crime statistics[8]
- Municipal public works[9]
- US Geological Survey[10]
- NASA[11]

Some examples of public data being enhanced for commercial use were raised with considerations for "smart home" and Internet of Things (IoT) devices as well as transportation systems. For example, the smart home systems may be linked to weather, occupant behavior patterns, traffic, transportation delays, deliveries, school system, emergency management agencies, or sensors (e.g., thermostat, meters, smoke/carbon dioxide detectors). The widespread use of artificial intelligence (AI) devices (e.g., voice activated remotes), computers, and mobile devices increases demand for situational awareness. The Air Traffic Control system data is often used for information about weather, flight schedules, tarmac operations, congestion, security, Transportation Security Administration (TSA), Customs and Border

[2] https://www.census.gov/
[3] https://www.bls.gov/
[4] https://www.gao.gov/products/131405, https://obamawhitehouse.archives.gov/open/around/eop/omb/datasets, https://www.whitehouse.gov/omb/, https://www.cbo.gov/about/products/budget-economic-data
[5] https://www.state.gov/j/drl/irf/rpt/
[6] https://catalog.data.gov/dataset/zip-code-data
[7] https://www.nws.noaa.gov/climate.php
[8] https://ucr.fbi.gov/crime-in-the-u.s
[9] https://catalog.data.gov/dataset?q=municipal+counties
[10] https://catalog.data.gov/organization/usgs-gov
[11] https://data.nasa.gov/

Protection (CBP), U.S. Citizenship and Immigration Services (USCIS), and even news feeds. Finally, applications such as ParkMobile[12] attempt to correlate data about parking availability, construction and road crews, autopay, and parking rules. Recently, there have even been advertisements linking pizza delivery and pothole tracking.

The need for scalability means that data is being managed by less mature technologies such as open source, distributed, and NoSQL platforms (e.g., MongoDB[13], HBase[14], MySQL[15], SQL Server Azure[16]). Sharing data imposes requirements for both performance and security partitioning. Surprisingly, architecture decisions (such as indexing and replication) may be much harder to change than with traditional data architectures.

The performance impact on internet infrastructure from downloads and live queries were discussed as well as the need for backup/snapshot capability. Anecdotes about public queries crashing web sites are well known. The batch update frequency required for data sources to maintain acceptable freshness will need to be tracked. Any data publishing effort requires a mechanism for a feedback loop and someone with the authority and responsibility to respond to the feedback.

The general problem of leveraging big data was not widely addressed, but participants expect it to make everything more difficult. Likewise, cloud solutions are not seen as silver bullets [2]. However, they may shift more decisions from system specific concerns into the relatively invisible infrastructure category. The Open Systems Interconnection model (OSI model[17]) inspired a framework for our discussion to keep the focus on application/data level issues.

Agencies should have plans for managing Data Delivery Dependencies (DDD) in their architecture:

- Infrastructure capacities (e.g., CPU, disk, memory, or network)
- Services – (e.g., directory or messaging)
- Platforms – (e.g., not only SQL (NoSQL) databases, Software as a Service (SaaS), Oracle[18], SQL Server[19], firewall and port management (such as what is allowed for port 8080))

[12] https://parkmobile.io/
[13] https://www.mongodb.com/
[14] https://hbase.apache.org/
[15] https://www.mysql.com/
[16] https://azure.microsoft.com/en-us/free/sql-database
[17] https://en.wikipedia.org/wiki/OSI_model
[18] https://www.oracle.com/index.html
[19] https://www.microsoft.com/en-us/sql-server/sql-server-2017

- Applications – (e.g., Business Intelligence tools, Command Line Interface tools, Office 365[20], Google Suite tools[21], and browsers)

After an initial discussion of data segmentation strategies for access control and the growing concerns about Personally Identifying Information (PII), the topic of Freedom of Information Act (FOIA)[22] requests were raised. Participants agreed that many of their agencies are asked for information that they should not and will not share with the public about operational matters.

Another general issue is that data posted to sites such as data.gov lack contextual information (such as files in the simple comma delimited value format) and are subject to both accidental and deliberate misuse. Even legitimate FOIA requests raise concerns because the participants felt it is unclear what liability an agency has for the consequences of misused data.

A specific concern is the need to explain how and why answers to questions change over time. The high-profile example is the agency having to explain to congress why an answer seems inconsistent. In general, government systems are not designed to provide the truth as of a given point in time. An archive of periodic snapshots is considered the best available practice to address this need for most organizations.

The use of data dictionaries for internal data management is helpful, but many formats are still focused on data elements rather than entity/object level structures. There is a greater need to address metadata at the business level. This can be related to the data provenance information with answers to basic questions (who, what, why, when, where, etc.) to provide information about "fitness for use".

The ability of formats such as Extensible Markup Language (XML)[23] and JavaScript Object Notation (JSON)[24] to make data more self-descriptive is important. However, the assumption that the rendering (i.e., visualization) of the data is straightforward and predictable is not always justified. There is no guarantee that a service endpoint is not manipulating the data in unexpected ways. Mechanisms such as digital signatures or hash values only work to ensure that the download has not been corrupted. There is little that can be done when data is used out of context.

A focus on providing data via batch download requests rather than live queries may lessen the chance that the data is distorted, or that the requests are disruptive of business due to

[20] `https://outlook.office365.com`
[21] `https://gsuite.google.com`
[22] `https://www.foia.gov/`
[23] `https://en.wikipedia.org/wiki/XML`
[24] `http://json.org/`

performance problems. The practice of presenting data on web pages and via formats such as Portable Data Format (PDF) documents has benefits and draw-backs. The consumer's ability (or lack of ability) to screen scrape web pages, use Optical Character Recognition (OCR), or parse PDF content may lead to data errors. In general, if information such as a list of offices is published in Hypertext Markup Language (HTML), there should be a formatted and self-descriptive download option provided by the data owner.

A registry of available, downloadable information is a best practice that can allow data consumers to schedule the processing load to convenient times. This also provides a venue for descriptive metadata (including the schedule for data refresh) and instructions for use. The consumers may be more competent in using a tool such as excel to process a download than formatting the right parameters for a REST[25] query [4]. General syntax options such as SQL[26], SPARQL[27], or XQuery[28] are both error-prone and potentially dangerous. For example, the "SQL injection" technique is a well-known threat where a query can contain commands disguised as search criteria.

There is little potential for preventing deliberate misuse of downloaded published data. Simple manipulations such as filtering the data and misrepresenting the data set will be difficult to recreate for explanation or defense. Critical considerations for maximizing inter-agency data access should also include Service Level Agreements (SLAs) and security boundaries. Providing read only access for shared data via REST Application Programming Interfaces (API) can simplify the recognition and acceptance of credentials.

Data governance and policies must be coordinated. The priorities for the organization should not be undercut by the policies and governing boards. The general trend towards data migrations raises another priority conflict. Often the desire to move from mainframes to newer architectures, or from data centers to cloud may take a large proportion of an organization's IT resources. The need to operate data management systems in parallel and validate the migrations may compete for resources with providing and sharing data outside the organization.

The desire to have agile and responsive capabilities also suggests that data delivery should be shifted from a programming and development task to an administration and configuration task. The goal should be to move away from custom coding to exporting data sets. Health records have become a very visible example of the tension between data sharing and data protection. How does one know what the data release rules should be and if they are working

[25] https://en.wikipedia.org/wiki/Representational_state_transfer
[26] https://en.wikipedia.org/wiki/SQL
[27] https://www.w3.org/TR/rdf-sparql-query
[28] https://en.wikipedia.org/wiki/XQuery

as expected? The mechanism of moving data via cross-domain solutions or outside of the enclave into a "demilitarized zone"[29] will involve rules that must be maintained and validated.

Moving from legacy technologies (such as Fax to modern API) is important. Delivering data with metadata is also key. One must determine if the consumer requires "an answer" or "the answer" from a dataset. The completeness, currency, and usability metadata may prevent inappropriate use. A suggested best practice is counter-intuitive: "Keep the data available as simple as possible." The more data elements are provided, the more questions that the consumers will have in their attempts to understand it. The more questions people have, the more time the organization needs to spend addressing them. It was suggested by the participants that data delivery should "be brief, be good, and be gone". Sharing and interoperability between organizations raises additional challenges. Progress is often measured in yearly increments. Systems can only be changed when resources become available. Multiple governance boards may have to agree. Defining success factors and progress measures becomes a key. Dashboards and visibility mechanisms are necessary to keep momentum and priority.

Since public, commercial, and opposition consumers will compare data from multiple sources, the more that the agencies synchronize themselves, the better. This may involve both conflicting data and inconsistent coverage. If one agency protects certain values while another agency does not, the consumer may be able to perform their own aggregations.

The use of snapshots introduces a problem. One cannot execute unanticipated follow-on questions about snapshot data. The source data will have changed, so asking additional questions may provide inconsistent answers.

The overall challenge of getting a "single version of the truth" within the agency is often too difficult. For example, the definitions of military units in different systems may lead to different answers to similar questions. There is also the lag time problem where data flows from system to system. If this lag is measured in days rather than seconds, it becomes more severe. There are many examples where the equivalent of a commercial entity "closing their books" cannot be implemented in the government.

Another challenge stems from data that is derived via functions, triggers, and other mechanisms. These are often not stable. Furthermore, data models and other descriptive information often ignore these mechanisms.

Another challenge is determining how well data is fit for uses such as aggregation. Can it be summarized and still understood? In general, the job definitions and interests of people working in policy, governance, and IT professionals are disjoint. Policy professionals are rarely

[29] `https://en.wikipedia.org/wiki/DMZ_computing`

interested in enough detail for this problem space. There is also a huge range of technical skills in the ranks of program managers and analysts throughout the agencies. How can you change the roles so that policy professionals are more interested in data? Should privacy and trust be the priority versus data collection (i.e., should data owners require consumer registration to use data)? Should there be a new agency policy on data collection to expand the context and provenance as early as possible?

There needs to be a storyteller role that explains what the results of data science mean to decision makers and champions. Using natural language processing (NLP) technology to analyze and model the contents of policy documents may also help make organizational behavior more consistent. Understanding the pain points may be possible without the skills to fix it. It is still too difficult to find out who are the users, what their use cases are, and what decisions are being made.

The organizations must decide what exactly to secure and how to do it. Candidates include the following:

- Data stored in databases
- Database servers
- Database Management System Software (DBMS)
- Other Database Workflow Applications
- Sources (non-classified versus classified) for PII data markers

The discussion ended with some challenged assumptions about logs. Although most stakeholders may assume that record keeping requirements can be satisfied with logs and archives, the actual use of them is limited. Furthermore, it is not clear who can (or should) use the logs and if they will be able to use them.

3.2.3 Recommendations

The *Maximizing Data Access* collaboration session participants arrived at several recommendations:

- In general, government systems are not designed to provide information at a given point in time. An archive of periodic snapshots is considered the best available practice to address this need for most organizations.

- When information such as a list of offices is published in HTML, there should be a formatted and self-descriptive download option provided. A focus on providing data via batch download requests rather than live queries may lessen the chance that the data is distorted, or that the requests are disruptive of business due to performance problems.
- A registry of available downloadable information is a best practice that can allow one to schedule the processing load to convenient times. This also provides a venue for descriptive metadata (including the schedule for data refresh) and instructions for use.
- Keep the data made available as simple as possible. The more data elements are provided, the more questions that the consumers will have in their attempts to understand it.
- Defining success factors and progress measures becomes a key for sharing and interoperability between organizations. Dashboards and visibility mechanisms are necessary to keep momentum and priority.
- End state guidance must cover a set of new functional responsibilities - to Monitor, Manage, Migrate, and Protect the published data.

Through these recommendations, organizations can begin to formulate data sharing and publishing strategies that better manage the way in which data is prepared and the way in which providers provide contextual metadata.

3.3 Shaping Data Strategies to Better Support Decision Making and Accountability

The *Shaping Data Strategies to Better Support Decision Making and Accountability* collaboration session focused on the intersection between organization data strategy and the ability to use data to support decision making while also enforcing accountability of how the data is used. Timely, high quality information is necessary to enable evidence-based decision making, research for improving future policy-making, and providing accountability and transparency. Organizational data strategies provide the policies and guide the standards for using data effectively, making data strategies a critical enabler for these subject areas. This session discussed the ways in which data can support these tasks, and how these considerations shape the development of a data strategy.

The session had the following three goals:

- Discuss the characteristics of an effective data strategy
- Identify any best practices for establishing a data strategy within an organization
- Recommending ways to reduce risk and promote accountability through an organization's governance and strategy

3.3.1 Challenges

Several challenges emerged from the discussions within the group:

- Data architectures typically have little proven guarantees for preventing data leaks, making risk elimination nearly impossible.
- Few documented use cases are available to inform data strategies.
- Accountability varies based on role and is often not appropriately incentivized.

3.3.2 Discussion Summary

The afternoon's discussions began with the participants identifying the challenges of operating over multiple classification levels and the implications of fusing data from multiple sources. Of particular concern were the regulations that guide handling of healthcare, UNCLASSIFIED, SECRET, or financial data and how the regulations change when combined with other controlled or sensitive datasets. This is particularly worrisome from the practitioners and data managers for which the penalty for mishandling the data potentially includes jail time. As such, the participants noted that data architectures used to process and derive information from data typically do not come with guarantees against data leaks. This makes quickly acquiring and using the architectures and associated data for high-impact decisions a challenge for the practitioners.

Because of a cited lack of data subject matter experts within government, agencies often must rely on industry products for data management and processing. Due – in part – to the lack of guarantees against data leaks, the participants noted that the safest way to prevent data leaks and mitigate the risk of penalty against the data practitioners is to operate in the appropriately protected environments (which, of course, can limit the ways in which data can be shared with external partners).

The tools available for data processing pipelines can assist with the curation, quality improvement, visualization, and analytics. There are methods available for providing information to internal and external stakeholders, but the needs and limitations (including the potential risks) must be considered when evaluating the tools to be used to build out an organization's data pipeline.

This evolved into a discussion thread regarding how data strategies can help guide the decisions about data management and associated risk and accountability. VAUTI[30] (visible, accessible, understandable, trusted, interoperable) was cited as a set of guidelines regarding characteristics of a high-quality data enterprise[31]; VAUTI (or variations on the acronym) is often used in data strategies. When considering a common (or generic) data strategy, it often describes the set of questions that data should help answer, the sources of the data that is used to drive the answer, the impact of being able to answer the questions, and the reason that data and analytics should be used to find the answers (e.g., ability to find more answers or find answers faster). Along with this information, a data strategy should describe how the use of data to answer the questions contributes to the organization's risk assumption. Ideally, the strategy should also identify areas of risk mitigation within the data pipeline (e.g., focus on cleaning data in the data prep phase to reduce risk).

The participants built on the notion of risk identification in a data strategy by discussing how accountability is assigned. There are various stakeholders and roles within the process of making data-based decisions, from the analyst constructing algorithms to the leadership that uses the resulting information to make decisions. Data practitioners should also be appropriately accountable and incentivized to provide high-quality data for use within an organization as well as using the resulting information in the most appropriate manner to inform decision making. Ultimately, this results in a need for the data strategy to balance the business, technical, and risk aspects of a data enterprise.

Finally, the session conversation touched on the importance of multi-disciplinary teams in data science (e.g., computer scientists paired with psychologists and statisticians). Being able to trace how data creates the information for decisions is also a major gap in the current state of the art of machine learning. The ability to have multiple viewpoints to reduce bias in data and algorithms along with the ability to provide introspection on how decisions were made can inform future use cases and decision-making processes using data.

[30] `https://www.army.mil/standto/archive_2016-03-14`

[31] The participants mentioned the importance of robust – yet flexible – ontologies, dictionaries, and standards for data. They also mentioned high levels of interoperability as a characteristic of high-quality data.

3.3.3 Recommendations

The *Shaping Data Strategies to Better Support Decision Making and Accountability* collaboration session participants arrived at several recommendations:

- Use (and contribute to) the Federal Data Strategy use cases to inform data strategies[32].
- Strategies should provide emphasis on data quality, preparation, and handling along with associated consequences and motivations for providing high quality data.
- Use VAUTI and other features (e.g., what questions will the organization answer with what data and to when end?) to begin informing a data strategy.

Through these recommendations, organizations can begin to formulate data strategies that better manage the way in which data is used for decision making and the way in which practitioners are held accountable for ensuring data is made available at the highest quality and decisions are made with the utmost care.

3.4 Maximizing Commercialization, Innovation, and Public Use of Government Data

The *Maximizing Commercialization, Innovation, and Public Use of Government Data* session focused on how the government can incentive use of data to stimulate innovation and economic development in the commercial sector and general public. The session discussion focused on barriers to commercializing government data sets, including privacy, security, access, quality, and incentives.

The goals of the session included the following:

- Identify the barriers that prevent the public and commercial companies access data collected by government agencies.
- Understand the incentives commercial companies and the public need to take on the existing challenges to accessing data to enable their ability to innovate and stimulate their local economies.
- Understand the security and privacy implications of large datasets being available for commercialization, innovation, and public use.

[32] `https://strategy.data.gov/use-cases/`

3.4.1 Challenges

Several challenges emerged from the discussions within the group.

- Government data is typically physically located across several locations that are not connected[33]. Efforts to move data to the cloud and/or interconnected network are labor intensive and substantial in cost [1].
- Data is retained in numerous formats/languages, including legacy formats and languages, that are becoming more difficult to maintain and translate into useful data.
- Quality of the data in the government varies and generally requires a significant amount of curation to become useful to a commercial or public entity to use.
- Data rights and access restrictions vary by department and agency. Data use licenses and agreements are not consistent or understandable hindering a commercial or public entity from understand what their usage options are.

3.4.2 Discussion Summary

The session discussion facilitated the identification of the barriers, best practices, and recommendations for commercialization, innovation, and public use of data. The group focused on a use-case of bringing cellular service to rural America and the challenges for big business to identify gaps in their existing services. The discussion began around how a big business or commercial organization could get access to the data the government has collected about gaps in cellular and broadband services. The initial thoughts were this information would be available via network connection or in the cloud, but as the group deconstructed the use case the challenges mentioned above were identified.

The government agency responsible for collecting and receiving information on broadband and cellular available across the US has geographically dispersed offices. Each office stores their data in on-premise servers. The agency is working to be more interconnected, but does not have a method available to collect regional information regarding gaps in broadband and cellular service. After discussing the barriers to gain access and commercializing this regional broadband and cellular data, the group discussion continued under the assumption this barrier could be overcome.

Once this barrier was removed, the group shifted focus to what is preventing more data from being commercialized with the assumption that commercialization leads to innovation

[33] For example, an agency's data may be distributed across several disconnected regional offices

and more use by the public. The discussion initially focused around incentives and the amount of time and labor a company must put in to understand and curate data. The example of Data.gov was used to discuss how there are hundreds of thousands of datasets available for public use, but the quality of the data is low, the formats can be difficult to ingest and for an outside entity to understand the "jargon" being used within the data only delays an already lengthy process.

Several incentives to bring companies to use the data were then discussed, including efficiency in expanding a business. For example, in the use-case of bringing broadband and cellular service to rural America, companies could be more efficient in knowing where to establish new towers or how to network with other companies to increase coverage. Instead of each company researching this issue on their own, the government coverage maps could quickly inform a company saving them time and labor to expand their business. Accuracy of the data was also discussed. While data quality is a recognized issue, with the sharing of data and use by commercial companies and the public the data accuracy would improve over time, benefiting companies by reducing the miss information about gaps in coverage. The last incentive discussed was around creating a market for data. The group discussed companies that are already leveraging government data and reselling in various forms. This type of opportunity, if provided as an example to other business, may further demonstrate the benefits to spending the time to curate data for commercial use.

Following the discussion of incentives, the discussion shifted focus to existing best practices or exemplars of how to commercialize data. The group identified the following potential best practices that are currently being leveraged in portions of the government and could easily expanded:

1. Leveraging the role of the data steward to work with commercial companies and public inquiries

2. Providing real world examples of how the government is using the data and provide the dataset with that example

3. Creating partnerships between academia, non-profits, and the government to build out quality datasets

After the group discussed best practices, the session wrapped up with a final discussion around what would success look like for that use-case in rural America if the government could *Maximize Commercialization, Innovation, and Public Use of Data.* Success would look like broadband and cellular support to being provided to the gaps in rural America, ability for

any group to access information on broadband and cellular gaps and providing opportunities for companies to invest and provide this information to others.

3.4.3 Recommendations

This collaboration session identified the following recommendation:

- Organizations should develop and/or create a role at government agencies as data stewards. These stewards would provide a service to the public and commercial agencies in understanding the dataset and learning how to use it.
- Create more public-private partnerships to ingest, curate, and make available large datasets.
- Standardize across the agencies how and where to host data sets (e.g., cloud vs on premise).
- Create cross-cutting challenge problems for the public and academia to be solved using multiple government datasets.
- Demonstrate the power of using a government data on a real-world problem.

These recommendations focused on removing the barriers to accessing the data and then showing the benefits of working with the data. Throughout the discussion, the overwhelming message was there are too many barriers to accessing good quality data which is a disincentive to commercializing and innovating on data. If the government could remove some of those barriers and showing to companies the possibilities that come from using the data that commercialization and innovation would quickly increase.

3.5 Data & Analytics in Health

Hosted by the Department of Defense/Veterans Affairs Interagency Program Office (DoD/VA IPO), this session focused on how emerging technology standards and related clinical workflows might enable medical data analytic interoperability that could improve health outcomes. Facilitated by emerging shared analytic engineering approaches, the medical workforce can conduct predictive analytics for early detection and mediation of emerging hazards, public health threats, and to promote community well-being and resilience. The session identified challenges and potential solutions related to sharing data analytics of interest amongst the multiple government organizations focused on health care and health care delivery along

with others in industry and academia. The session explored the critical issues that leverage nascent health data analytic technology to revolutionize healthcare and allow better medical understanding via standardized means to share personalized analytics and portable data models that provide integrative care.

During this 2018 health data analytics session, numerous important summary points from the 2017 summit were discussed [5]. In addition, developments post 2017 summit included establishment of the ATARC Data Analytics Working Group as well as a Project Authorization Request (PAR) acceptance for IEEE [3] regarding shared analytics.

In addition, shared data analytics applications were discussed including the following:

- Introduction to sharing analytics, what the approach entails, and why it is needed
- Potential standards for sharing analytics and portable data models
- Use cases for facilitation of sharing analytics
- Potential clinical workflows using shared analytics
- Potential obstacles to implementation of analytic sharing in clinical ecosystems

3.5.1 Challenges

Several challenges including the lack of collaboration across cross-department (i.e., DoD, VA, FDA, etc.) hinder the growth of personalized health analytics:

- Global shifts in law beyond the Health Insurance Portability and Accountability Act (HIPAA) to the new General Data Protection Regulation (GDPR) one of the most important changes in data privacy regulation in 20 years[34].
- Consolidation and interoperability of data, including understanding the differences between fixed analytics vs future machine learning and artificial intelligence health data applications.
- Concerns about consumer and patient centricity – how do government organizations empower patients to share data while keeping privacy and the right to access information in mind without compromising the quality and accuracy of data analysis?

[34] `https://eugdpr.org`

3.5.2 Discussion Summary

Several recommendations were made in the session such as developing shared analytic standards and platforms with variations that allow virtual consolidation and interoperability by combining and accessing data from various sources without moving data to a broker. In addition, the participants discussed the idea of standards for exchange of both analytic and portable data models being used to help build a federated understanding of what is in the network (i.e., understand the data that is out there).

Several efforts in interoperability and standards are led by the Office of the National Coordinator (ONC) nationwide that address emerging data centric needs to facilitate and assess healthcare delivery. In the past year or so, an interest for emerging health standards and technologies focused shared analytics standard has catalyzed the IEEE P2795[35] standard working group from 2018-2022 that will help harmonize and benefit stakeholders going forward. Such shared analytics workflows are envisioned to include a four-part IEEE metadata exchange. In the future P2795 compliant health sensors and information systems may exchange big data definitions that represent the actual trust and meaning of the data to address regulatory topics (policy), clinical issues, wearables, etc. These will be important in emerging technologies in patient engagement. The IEEE standard may improve telemetry, telemedicine, visualization, fraud protection, and more. The four-step IEEE P2795 analytic sharing workflow model develops meta data which characterizes how information about the data model and processing capacity is sent and received in step 1 and 2 then also delineates meta data about how an analytic is prescribed and a vetted privacy preserving result is returned in step 3 and 4.

Additional recommendations were made by the participants to continue work to encourage health data providers to designate a portable data model, pursue analytics, and then share it since such definitions are a derivative of patient and provider data but would not violate privacy and security requirements, as well as defining the algorithms that would be permitted by analytic governance, regulators, providers, and patients.

This October 2018 ATARC health data analytics shared analytics session explored important use cases to help define and drive stakeholder engagement. Building on this, the IEEE will explore how analytic interoperability is possible via standards for sharing analytics and data model designs being envisioned by engineering in medicine sponsors that authorized the P2795 standards working group.

[35] `https://standards.ieee.org/project/2795.html`

3.5.3 Recommendations

This collaboration session identified the following recommendation:

- Develop shared analytic standards and platforms with variations that allow virtual consolidation and interoperability.
- Encourage broad participation in the IEEE P2795 standards activity.
- Encourage health data providers to designate a portable data models and then share them.

3.6 Nature-inspired Machine Intelligence (NIMI): Harnessing Evolution

The Nature-Inspired Machine Intelligence (NIMI) session examined the fundamental thoughts underpinning NIMI and its relationship with related concepts such as artificial neural networks (ANNs), swarm intelligence, artificial immune systems, and autonomous entities.

3.6.1 Challenges

Several challenges emerged from the discussions within the group.

- Learning from nature is not new. However, nature inspired machine intelligence is a relatively new concept. The session participants spent a considerable amount of time discussing the meaning and scope of NIMI. Clear definitions are needed for NIMI and the related concepts.
- NIMI requires a multidiscipline approach. NIMI requires combining knowledge of computing systems, biology, and other disciplines. Consequently, NIMI research and development will require multidiscipline teams including scientists who understand biology and natural phenomena, domain experts (e.g., cybersecurity experts, medical professionals, disaster recovery experts), data scientist who can extract knowledge from the collected data, and computer scientists who can transform the observed concepts into computer algorithms. Determining which disciplines are needed may be challenging and will likely vary depending on the problem domain.
- Detecting system anomalies by monitoring deviations from normal system behavior is challenging since unique situations may appear to be anomalous (e.g., false positives). NIMI anomaly detection methods will need to differentiate between unique system behavior and anomalous behavior.

- NIMI based solutions may have a range of uses from decision aids to fully autonomous systems. For example, NIMI based solutions may be designed to detect anomalous heath conditions using predictive analytics and respond by administering medication. Consequently, NIMI based solutions must be trustworthy if users are to rely on the results provided. NIMI implementors must also ensure the integrity of the data used since false data can lead to false model results. This will require reviewing the integrity of the NIMI algorithms and data. Defining the trustworthiness of NIMI based models and solutions may be challenging.

- Nature has mechanisms that scientists do not yet know or understand. Studying nature inspired concepts will be challenging and will continue to evolve as scientist explore and conceivably modify previous models and concepts.

3.6.2 Discussion Summary

The government representative for this session, a researcher and practitioner of AI systems and expert systems, started the session by discussing his research and knowledge of NIMI. While describing NIMI as a new and emerging concept the government representative discussed examples of how NIMI concepts have been employed to include neural networks and swarm intelligence to address mission needs.

The government representative described NIMI as a family of concepts that are derived from systems found in nature such as neurons (also called nerve cells), the immune system (the biological system that protects against disease) and swarm intelligence (the collective behavior of multiple entities). He stated that while AI and machine learning (ML) approaches use analytical models which are developed based on mathematical equations, NIMI uses computational models which are developed based on observation or simulation.

Examples of nature inspired concepts include the artificial neuron, the immune system, swarm intelligence, and flight control. The artificial neuron was developed (i.e., inspired) based on the fundamental properties of the neuron. Neurons "fire" (i.e., send an impulse) when the input stimuli exceed a given threshold. The model for the artificial neuron was developed by first observing the behavior of the neuron and then imitating this behavior using electronics or a computer program. The model of an artificial neuron consists of a weighted input signal to which a bias value is added to form a net input value. The output value of the artificial neuron is the result of applying a transfer function to the net input value. The transfer function is often a function that maps the input values to an output value ranging between 0 and 1. Like the neuron, a threshold level is often established such that when the

output value exceeds the established threshold the artificial neuron sends an impulse.

A second example is the immune system. The immune system within the human body monitors the body for abnormal conditions including the presence of toxins, foreign substances, or deteriorating cells and acts to remove these abnormalities. Observing and learning about the human immune system can aid in developing cybersecurity monitoring techniques that monitor system behavior, detect abnormal conditions, and act to autonomously isolate, mitigate, and remove threats. Key to this concept is that anomalous conditions are determined not only by detecting known threat signatures (e.g., rule-based detection methods) but also anomaly-based detection methods that detect novel threats (i.e., threats not previously known or encountered) by detecting deviations from normal system behavior.

A third example is swarm intelligence. Examples of swarm intelligence include flocking birds flying in a V formation, ant colonies foraging for food, and schools of fish which are formed by individual fish swimming together. With the fundamental concept that many minds are greater than one, the benefits of swarm intelligence can be applied to many areas including robotics, target surveillance, disaster response, and search and rescue operations.

The Wright brothers' discovery of flight control is another example of nature inspired learning. The Wright brothers' concept of flight control using "wing-warping" was the direct result of their study and observation of birds in flight.

The group discussed different ways in which nature inspired concepts can be used to help solve challenging problems. For example, cybersecurity threat detection may be accomplished by sending many software applications (e.g., a swarm of bots) throughout a computer network to look for anomalies. Or, security software embedded within the hardware could be designed to "sprout" when anomalous conditions are detected. When an anomalous condition is detected, the software application reports the location and a description of the anomalous condition to a system operator.

The group also discussed how situation awareness during disaster events can be improved by sending a team of drones (i.e., a swarm of drones) into the disaster area prior to sending the first responders to determine the geographical boundary of the disaster area, provide video showing the extent of the damage, determine the presence of toxins and other conditions that may be harmful to first responders, and detect the presence of human life to improve search and rescue operations.

With respect to health care, the group discussed how wearable electronic devices can collectively be used to determine the health (e.g., health data analytics) of individuals. This includes continuous real-time monitoring, analysis, and reporting of blood pressure, heart rate, and body temperature. This is like the on-board diagnostics capability found in most

automobiles today. The automobile on-board diagnostic system continuously monitors parameters including fuel and air mixture, emissions controls, engine timing, and vehicle speed to determine the health of the vehicle. This is also similar to the way the immune system monitors the body to detect abnormal conditions and takes action to mitigate the abnormal condition. Wearable devices can also be used to collect and monitor the health trends of groups of people in a given geographical area which could help in detecting and controlling the spread of disease.

The group discussed how ingestible electronic devices can be used to monitor and detect anomalous conditions within the human body. The ingestible devices may be used as an initial diagnostic tool to detect infections such as methicillin-resistant Staphylococcus aureus (MRSA). The ingestible devices may also be used as a continuous and persistent monitoring capability that uses predictive analytics to detect, analyze, and report the presence of an infection before the signs of disease appear.

While learning from nature has great benefits it was also noted that learning from nature must also include learning about the associated side effects. For example, one of the side effects of cancer treatment is that while chemotherapy kills cancer cells it can also damage healthy cells. Similarly, researchers must consider the side effects of sending a team of drones into a disaster area, or releasing cyber threat detection bots throughout a network, or placing ingestible devices within the body.

One participant stated that attention is often given to disrupting natural phenomena such as preventing flooding, but thought should also be given to using or benefiting from the phenomena. For example, while flooding can be destructive, it also has the benefit of redistributing ground nutrients and refreshing wetlands. Preventing or disrupting flooding in some areas could result in long-term negative impacts on the ecosystem. Consequently, learning from nature includes learning about natural phenomena, learning about the side effects of disrupting the phenomena, and learning about the beneficial use of the phenomena.

During the final moments of the session, the group acknowledged there is much to learn about NIMI and that most of the time spent during this session was focused on the "nature inspired" part of NIMI. Not much time was spent defining what is meant by machine intelligence.

3.6.3 Recommendations

This collaboration session identified the following recommendation:

- The key end-user benefits for NIMI should be defined, documented, and shared with

potential researchers, users, and stakeholders.

- An ontology may be useful to define, understand, and categorize the NIMI concepts and data elements.
- A research agenda and roadmap for NIMI should be established to guide NIMI research, develop NIMI concepts, and implement NIMI based solutions. The session included a discussion on short-term and long-term goals for NIMI. The short-term NIMI goals discussed included developing methods to mimic the way the cerebral cortex works (i.e., performing perception, memory, and cognitive functions). The long-term NIMI goals discussed included teaching machines to learn for themselves (i.e., "singularity", machines with "self-consciousness"). Researchers should evaluate these goals and develop additional goals as necessary to explore this research topic.
- Participants of the multidiscipline teams will, and should, challenge the concepts that are discussed by their peers. While this may appear to be adversarial, this is a natural part of the discovery process. Participants should challenge each other and perform the required research to ensure the concepts that are discussed and implemented are sound.
- NIMI research should implement a combination of basic and applied research. That is, researchers should explore the theoretical aspects of NIMI but also focus on the benefits for end-users.
- Practitioners can begin using NIMI concepts by leveraging currently available data to explore and demonstrate potential NIMI capabilities.

4 SUMMIT RECOMMENDATIONS

Throughout the collaboration sessions, several common themes emerged from discussion.

- Organizations adopting big data should adopt a data strategy. For many organizations, beginning to leverage big data is not a matter of modernizing their data strategy, but developing a data strategy for the first time.
- Several sessions recommended implementing methods for maintaining consistent, well-engineered, and accessible metadata; facilitating internal and external sharing; improvement of analytical quality through big data principles; and better access control.

- Big data users must establish understandable, consistent access restrictions and agreements to facilitate data sharing.
- Organizations dealing in big data must create dedicated roles within their organizations for managing data, technology, and governance, such as CDOs and data stewards.
- Organizations should work to remove and streamline the technological and policy barriers preventing private/public partnerships and data sharing agreements.

These problems identified in these sessions are often difficult but are solvable. Solving these problems will require consistent and deep organization. The creation of these structures and processes can be facilitated by creating dedicated roles in the organization's command structure. Organizations dealing (or intending to deal) in big data must create dedicated roles within their organizations for managing data, technology, and governance, such as CDO Data Councils, and Data Stewards. These roles are crucial for guiding the organization of data efforts and the move towards big data, as well as helping customers understand the data and how to use it.

One key component in organizing data efforts is standardization. Standardization of various aspects of data handling appeared multiple times during these sessions. In particular, how and where data sets are hosted, methods for analyzing data, sharing methods, sharing agreements, and usage restrictions should be standardized across the organization. Additionally, information about using and participating in these standards should be readily available within the organization.

Key to these organizational efforts are the creation of data strategies. Data strategies define and describe the organization's approach to their data and are foundational for synchronizing data architecture, building management structures, and motivating organization-wide efforts. A good data strategy focuses on methods for ensuring data quality, details data preparation efforts, and describes how data should be handled. Crucially, it also describes the technical and business motivations and benefits for strategy it prescribes, as well as the consequences to the organization if incorrectly implemented.

Several discussions identified the usefulness of information sharing when shifting an organization towards big data. Many organizations have already moved to a big data paradigm, and few use cases are truly unique. Organizations planning to leverage big data stand to gain from openly available industry best practices and use case. Venues such as the ATARC Federal Big Data Summit provide a venue for sharing these insights, and many case studies can be found online.

5 CONCLUSIONS

The October 2018 Federal Data and Analytics Summit highlighted several challenges facing the Federal Government's adoption of data and analytics. The challenges were not compartmentalized based on the challenge areas at the Summit, but span across the discussions by Government cloud practitioners. Specifically, insufficient or non-existent data governance and data strategies, insufficient or non-existent context and metadata to enable big data analysis, and the variance in file formats and insufficiency of sharing agreements that prevent data aggregation remain perennial difficulties to overcome. Creating dedicated roles within organizations for managing data, building unifying data strategies, standardization of technologies, methods, and access, and the sharing of best practices and use cases can help mitigate the identified challenges.

While the October 2018 Federal Data and Analytics Summit highlighted areas of continued challenges and barriers to adoption, the Summit also cited notable advances in mitigating these perennial challenges. Particularly, more standards are be developed to address data storage and sharing, especially within the healthcare field.

From the recommendations made in the collaboration sessions, government practitioners (at all levels of government) should participate in special interest groups or working groups to increase collaboration; continue to influence standards development within the discipline; and continue to partner with academia to leverage cross-cutting research and to help train the government workforce. These activities will further mitigate the perennial big data adoption challenges cited by the participating big data practitioners.

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.

The authors would also like to thank the session leads and participants that helped make the collaborations and discussions possible. A full participant list is maintained and published by ATARC on its web site[36].

[36] https://www.atarc.org/

REFERENCES

[1] J. F. Brunelle, S. Anand, G. Barmine, M. Spina, K. Warren, A. Winston, M. Javid, A. Kemmer, C. Kim, S. Masoud, T. Harvey, and T. Suder. August 2017 ATARC Federal Cloud & Data Center Summit Report. Technical report, The MITRE Corporation; The Advanced Technology Academic Research Center, 2017.

[2] J. F. Brunelle, S. Anand, R. Cagle, C. Kim, M. Kristan, M. Spina, K. Warren, T. Harvey, and T. Suder. February 2017 ATARC Federal Cloud & Data Center Summit Report. Technical report, The MITRE Corporation; The Advanced Technology Academic Research Center, 2017.

[3] EMB/Stds Com - Standards Committee. P2795 - Standard for Shared Analytics Across Secure and Unsecured Networks. `https://standards.ieee.org/project/2795.html`, 2018.

[4] R. T. Fielding and R. N. Taylor. Principled design of the modern web architecture. *ACM Transactions on Internet Technology*, 2:115–150, May 2002.

[5] C. Harvey, J. Brunelle, R. Campbell, R. Eng, A. Tall, H. Vafaie, A. Verma, T. Harvey, and T. Suder. DECEMBER 2017 Federal Big Data Summit Report. Technical report, The MITRE Corporation; The Advanced Technology Academic Research Center, 2017.

[6] The MITRE Corporation. FFRDCs – A Primer. `http://www.mitre.org/sites/default/files/publications/ffrdc-primer-april-2015.pdf`, 2015.