HARNESSING BIG DATA WITHIN THE FEDERAL GOVERNMENT

FINDINGS AND RECOMMENDATIONS OF ATARC'S BIG DATA INNOVATION LAB
DECEMBER, 2015

# ATARC Big Data Innovation Lab Sponsors

ATARC Big Data Innovation Lab

Objective

- Define big data
- Identify the challenges Big Data presents
- Develop strategies to surmount challenges
- Harness Big Data to make better decisions

Critical issues addressed in the report

- Data as an asset
- Data as a catalyst
- Extracting value from data
- Role of the Chief Data Officer (CDO)
- Data infrastructure
- Safeguarding federal data
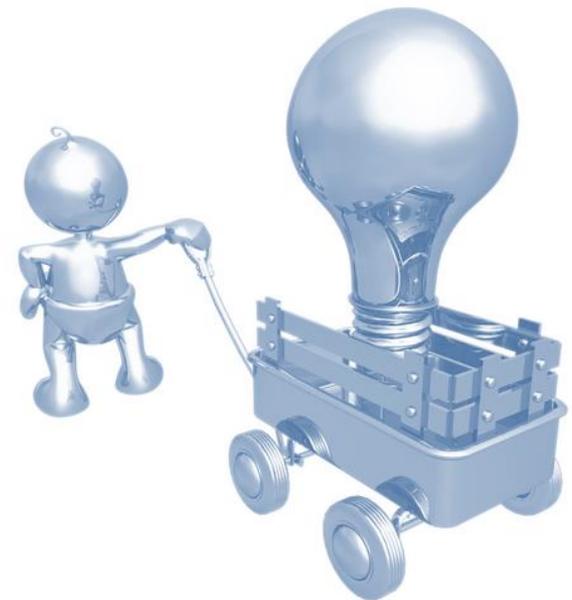
Big Data 101

- Defining "Big Data"

  - Many definitions exist. For our purposes, we focus on "big" datasets that:

    1. Are larger, more complex and multi-faceted which present different challenges and opportunities than those of less volume and variability

    2. Contain value which can become actionable information by leveraging advanced analytics tools

Data Growth:

x50 by 2020

Big Data is Transformational
Driving the growth of big data are technologies that enable enhanced collection of data. This combined with advanced analytics tools turn massive data sets into useful, actionable information that help create a more efficient and effective government

atarc

Why Focus on Data?

Critical Considerations on the Role of Data

Data as an Asset

- Definition
  - An approach to managing, ensuring the quality of, and sharing data that an agency collects, creates or maintains in a manner that delivers value to improve internal operations or delivery of services to citizens & other constituencies

- Impact
  - Making quality data available and usable can promote operational efficiency and improve service effectiveness & quality of life for citizens

- Discussion
  - According to Project Open Data, "Data is a valuable national resource and a strategic asset to the U.S. Government, its partners and the public. Managing this data as an asset and making it available, discoverable, and usable – in a word, open – not only strengthens our democracy and promotes efficiency and effectiveness in government, but also has the potential to create economic opportunity and improve citizens' quality of life."
  - Data can:
    - Help governent officials understand how their agency functions and serves the citizenry
    - Help agencies find new opportunities, expose problems & improve citizen services
    - Help an agency predict and anticipate events and problems
  - Agencies must measure and report on these uses of data
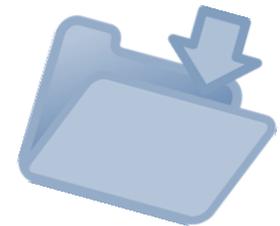
Data as a Catalyst

- Definition
  - An approach to managing, ensuring the quality of, and making data that an agency collects, creates or maintains open, machine-readable and publicly available in a manner that recognizes its potential value as a catalyst for economic activity and maximizes its utility to external organizations and entrepreneurs

- Impact
  - Making government data open and publicly available can spur the creation new, billion-dollar industries that fuel the economy through increases in employment, spending, etc.

- Discussion
  - Making agency data publicly available has high potential:
    - NOAA's weather & GPS data led to the creation of a multibillion-dollar industry, creating new jobs and improving the lives of citizens
    - CMS' release of Medicare provider utilization and payment information offers both providers and consumers with insight into who provides which services at what cost, making the entire healthcare system more accountable
  - Agencies must continue to innovate and invest in developer programs and hubs and Application Programming Interfaces for key data sets
  - Agencies must continue to explore ways to measure use and impact of open data

atarc

Extracting Value

- Definition
  - Value = Usefulness or Importance

- Impact
  - The intrinsic value of data is enhanced by the application of analytics. What we do with the data once we've collected it yields information and leads to data-driven decision making that can improve the efficiency and effectiveness of government

- Discussion
  - How are we making data useful and meaningful?
    - Ex: DEA uses data from state, local and national law enforcement agencies to reduce drug trafficking across the border
  - Making raw data available is a start, but interactive applications & visualizations encourage exploration & findings
    - Ex: CMS Open Payments **raw data** vs. **visualization**

atarc

Data in Action

The Emergence of the CDO and Changes to Our Data Infrastructure

atarc

The Role of the Chief Data Officer

- CDO is a new & evolving role w/ varied responsibilities, including:
  - Advocating for open data & advising data programs on efficiency and effectiveness
  - Guiding the big data journey through tech, ops, analytics & application
  - Leading data governance & internal data quality efforts
- Agencies can set the stage for success by:
  - *Clearly defining the mandate* for your agency (not one-size-fits-all)
  - *Empowering the CDO* to guide agency programs; do not pigeon-hole into an IT role
  - Carefully *thinking through the reporting structure* implications:
    - CIO — supported by existing law, but can limit agility and investment if competing with traditional CIO priorities (like acquisition, capital planning, information security)
    - Agency Head — elevates CDO exposure & aids strategic focus, but could create tensions with the CIO
- CDOs can ensure that they're effective by:
  - Working collaboratively to define their program, including any delegations of statutory authorities
  - Identifying where their engagement will have the most impact
  - NOT trying to be involved with all things "data"
  - Focusing on the value *you* bring to the agency

Data Infrastructure

- Agencies must work in both the traditional and the big data space

  - Agencies need to develop architectures that support both approaches, with a focus on:

    - Discoverability

    - Interoperability

    - Scalability

    - Federation

  - Agencies need to focus on agility and tech insertion to achieve these goals.

    - Big data projects tend to be more fluid and dynamic than traditional projects.

    - Open source software and cloud solutions must be accessible to enable rapid experimentation

- Impact

  - Data are accessible to the right users, at the right time, and in the right format

  - Cloud services, agile project management, and digital services capabilities converge to extract value from agency data assets

- Discussion

  - Agencies must line up a number of technology roadmap items to build the data infrastructure of the future. Addressing these interdependencies is key

  - The Digital Services Playbook applies just as much to big data projects as it does to customer-facing Web sites and projects

Security & Metrics
Safeguarding Data and Measuring Success

Safeguarding Federal Data

- Definition
  - Efficiently applying appropriate information security levels across government

- Impact
  - Efforts to safeguard data continue to grow in complexity and cost as agencies adopt cloud computing, Federal information expands, and the Internet of Things expands
  - Each agency has taken its own approach to applying Federal security standards
  - We have an opportunity to normalize protections (e.g. access controls, data masking, and data usage) and gain efficiency & increased security in the process

- Discussion
  - Agencies have varying risk thresholds
  - Organizations have adopted different processes & tools, with varying levels of success
  - Agreeing on the information to be safeguarded is the first step in normalizing and cost-effectively applying cross-agency protections
  - This will ease cloud adoption & support rapid provisioning of information to the public

atarc

Other Key Considerations

- Select and implement the correct **techniques** to apply to specific analytical problems

- Establish data **interoperability** (metadata, collection frequency & quality) to combine data from multiple agencies

- Take advantage of **shared services** (e.g., cloud storage, analytical platforms) to move fast and reduce cost

- Support **acquisition improvement** efforts

Metrics of Success

- Definition
  - Data governance is the key to taking advantage of these trends.
  - Traditional, quantitative measures like accuracy, completeness and consistency are important
  - Qualitative measures matter, as well – compliance, productivity, customer satisfaction, and innovation opportunity
- Impact
  - Measuring for governance outcomes moves the conversation from how to why
  - Qualitative metrics involve collaboration with data users from inside and outside the agency
- Discussion
  - Develop a metrics program that focuses on making the agency and the public successful with your data
  - Lead by example – your metrics program is data, too!

Use Case: Environmental Protection Agency

- EPA looks for patterns in air quality data

- Description
  - Apply anomaly detection algorithm to air quality data.
  - Analyze if anomalies occur at the exact time for nearby monitors over 6 years.
  - Use wind speed and direction to see if anomalies are correlated to them.
  - Visualize the results on graphs and maps.

- Challenges: Configuring the analytical tools to work together

- Next Steps: Incorporate the new insights into business operations

atarc

Recommendations

Recommendation 1: Create a data-centric culture

- **Build a teaming environment** among all CDOs (Federal CDO and agency CDOs)

- **Build a sharing culture**—focus on end-user benefits, set expectations, establish internal champions

Recommendation 2: Empower agency CDOs to:

- Develop a strategic plan to **capitalize on respective agency data** that includes governance, security, data models, access and analytic tools

- Define an agile program to **bring value to the agency**

- Implement **cross-agency data safeguarding standards**

atarc

Recommendations, continued

Recommendation 3: Enable Data-Driven Decision-Making

- Ensure the available data **add value to the decisions being made**

- Ensure the data are of **appropriate quality to the decisions being made**

- Develop **metrics for assessing quality and value**

Recommendation 4: Focus on improving services, creating efficiencies and meeting mission

- Determine *what* **agency data** to use and *how* **to use it**

- Identify the **economic impact** of opening specific datasets

- Use data to **improve citizens' online service experience**

Recommendation 5: Use the right technology

- Automate data collection and aggregation to reduce costly manual workload, eliminate errors & ensure trust

Methodology

- Over the course of 12 weeks, beginning in August 2015, ATARC convened the Big Data Innovation Lab consisting of more than 20 representatives from government, industry, and academia. The ATARC Big Data Innovation Lab was led by Tom Woteki of Maximus (Chair), Graham Morehead of Basis Technology (Chief Technology Officer) and Heather Gittings of Cliq (Vice Chair).

- The group met more than five times to debate issues related to emerging trends in big data and its application to government.

- The results of the findings are summarized in the preceding report.

About ATARC

The **Advanced Technology Academic Research Center** is a 501(c)(3) non-profit organization that provides a collaborative forum for Federal government (Intel, DoD, Civilian), academia and industry to work together to identify, discuss and resolve emerging technology challenges.

atarc

Acknowledgements

Thanks to the following for their contributions and participation in the ATARC Big Data Innovation Lab:

- Dr. Thomas Woteki, MAXIMUS, Industry Chair

- Heather Gittings, Qlik, Industry Vice Chair

- Graham Morehead, Basis Technology, Innovation Lab Chief Technology Officer

- Tom Suder, ATARC

- Mike Hettinger, ATARC

- Tim Harvey, ATARC